KLETTERMIX: CLIMBING TOWARD HIGH-QUALITY GERMAN PRETRAINING DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality pretraining data is a central ingredient in modern language models, but German-language resources remain far less developed than their English counterparts: they are often smaller, less carefully curated, weakly documented, and rarely validated through controlled training experiments. We introduce KletterMix, a high-quality German corpus for language model pretraining and annealing, designed as reusable dataset artifact for the natural language processing and modeling community. KletterMix is built by translating a state-of-the-art English pretraining corpus into German while preserving document boundaries, metadata, source structure, and topical diversity. This construction yields a German corpus with the scale and diversity of a modern pretraining dataset, while enabling direct comparison to its English source. We document the dataset through a broad set of corpus-level analyses, including translation quality, document length distributions, topic coverage, and source composition. A COMETKiwi-supervised proxy enables scalable filtering, bringing the corpus's aggregate linguistic profile closer to a matched German-web reference without uniformly selecting syntactically simpler text. Beyond dataset construction, we evaluate KletterMix through controlled pretraining and annealing ablations against established German corpora, we show that models trained on KletterMix achieve measurable improvements on German-language downstream evaluations. These results demonstrate that carefully curated translated data can substantially strengthen the German pretraining data ecosystem.[1]

1 INTRODUCTION

Pretraining data is one of the primary determinants of language model quality. While model architecture, optimization, and alignment recipes are often the most visible components of language model development, the pretraining corpus defines much of what a model can learn before instruction tuning, domain adaptation, or annealing. It determines which facts, domains, registers, styles, and linguistic phenomena are available during the highest-volume stage of training. Recent progress in open language modeling has therefore been driven not only by scale, but also by increasingly careful pretraining mixtures (Gao et al., 2020; Soldaini et al., 2024; Li et al., 2024; Diao et al., 2025).

This progress has been uneven across languages. English has benefited from large, diverse, and well-documented pretraining corpora, whereas German-language resources remain comparatively less mature: they are often derived from noisy web crawls, embedded as subsets of multilingual corpora, or released with limited documentation and validation through controlled training experiments (Suárez et al., 2020; Kreutzer et al., 2022; Dodge et al., 2021; Scheible et al., 2024; Dada et al., 2023; Burns et al., 2026). This gap matters because strong German model behavior cannot be assumed to emerge from English-centric data alone. German morphology, compounding, capitalization, regional variation, and domain-specific register all shape the effectiveness of German language modeling.

One response is to collect and filter more native German web text, and recent German and European efforts have made substantial progress in that direction (Burns et al., 2026; Pfister et al., 2025; Gienapp et al., 2025; Ali et al., 2025). However, native crawling alone does not automatically reproduce the source diversity, mixture design, or documentation standards of the strongest English pretraining

[1] Code at: `https://anonymous.4open.science/r/KletterMix-5F3F`
Dataset at: `https://huggingface.co/datasets/KletterMix/KletterMix-12B`

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.

datasets. High-quality German material is distributed across heterogeneous sources, and aggressive filtering can remove useful long-tail content while still leaving substantial noise. An alternative is to transfer the structure of a strong English pretraining mixture into German through high-quality machine translation. This can preserve topical diversity, source balance, and mixture-level design decisions, but it also introduces risks, e.g. translationese, semantic drift, source-language bias, length pathologies, and failures on long or specialized documents. Translated pretraining data must therefore be treated as a dataset-construction problem rather than as a simple data augmentation step.

We introduce KletterMix, a 725B-token German pretraining and annealing corpus built by translating ClimbMix, a recent high-quality English pretraining mixture (Diao et al., 2025), into German. KletterMix is designed to transfer the coverage and mixture structure of its English source while preserving document boundaries, document identifiers, metadata, source composition, and topical diversity. To assess translation quality at scale, we use COMETKiwi (Rei et al., 2022) as a reference-free quality-estimation signal and train a proxy model for corpus-level quality estimation. We combine these diagnostics with analyses of document length, source composition, topic coverage, and metadata preservation, and evaluate downstream utility in matched 0.6B-parameter pretraining and corpus-choice annealing experiments as well as a controlled 7B-parameter language-mixture annealing study.

Our contributions are as follows:

- We introduce KletterMix, a 725B-token German pretraining corpus constructed from a high-quality English source mixture while preserving document-level structure and metadata.
- We describe a scalable translation pipeline for pretraining data, including length-aware batching, chunking, dynamic output budgeting, and corpus-level validity checks.
- We document the corpus through COMETKiwi and proxy quality diagnostics, with analyses of length distributions, topic coverage, source composition, and metadata preservation.
- We evaluate KletterMix at two model scales: through matched 0.6B-parameter pretraining, proxy-filtering, and corpus-choice annealing experiments, and through a 7B-parameter language-mixture annealing sweep that measures German gains alongside English retention.

Overall, KletterMix explores a practical way to strengthen non-English data: instead of relying on more native web crawling, it tests if the curation, source diversity, and mixture design of a strong English corpus can be transferred to German via careful translation, documentation, and empirical validation. Across two model scales and complementary training regimes, results show this can strengthen German pretraining data, while underscoring the quality-control needs of translated corpora.

2 Related Work

Large-scale pretraining corpora. Large-scale pretraining corpora have increasingly shifted from raw web-scale collection toward documented, filtered, and evaluation-driven dataset artifacts. Early influential resources such as The Pile (Gao et al., 2020) and ROOTS (Laurençon et al., 2022) demonstrated the importance of diverse source mixtures, corpus documentation, and multilingual coverage for LLM pretraining. More recent English-focused corpora such as Dolma (Soldaini et al., 2024), DCLM (Li et al., 2024), and ClimbMix (Diao et al., 2025) further emphasize reproducible curation pipelines, filtering, deduplication, mixture design, and downstream evaluation as central components of dataset design. In parallel, multilingual web and open-data corpora such as mC4 (Xue et al., 2021), OSCAR (Suárez et al., 2020), CulturaX (Nguyen et al., 2024), HPLT (de Gibert et al., 2024; Oepen et al., 2025), FineWeb2 (Penedo et al., 2025), and Common Corpus (Langlais et al., 2026) have expanded coverage beyond English, often by applying language identification, deduplication, quality filtering, provenance tracking, and license documentation at scale. These efforts show that pretraining data quality depends not only on scale but also on: source mix, filtering, documentation, and validation via model training. KletterMix complements these corpora by testing whether the diversity of a strong English corpus can be transferred to German via high-quality translation and validated in controlled training experiments, rather than relying only on German-specific web crawling and filtering.

German and European language-model data. Recent German and European LLM efforts highlight both the demand for language-specific data and the difficulty of obtaining it at sufficient scale and quality. GermanWeb (Burns et al., 2026) constructs a large German pretraining corpus from Common Crawl, FineWeb2, and synthetic data using heuristic and model-based curation, and validates the

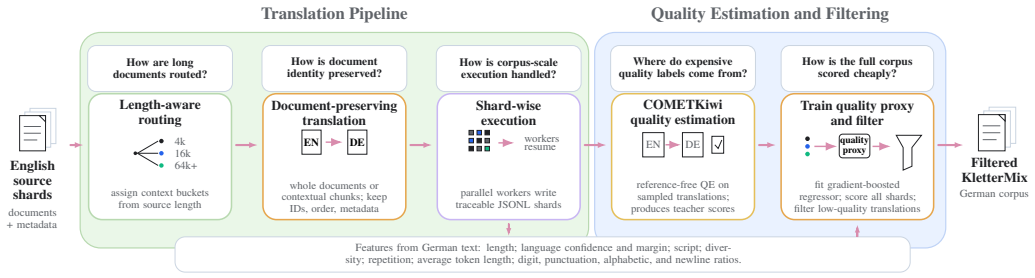


Figure 1: KletterMix pipeline. English shards are routed to length-aware context buckets and translated shard-wise in parallel using a document-preserving process that retains document IDs, order, and metadata. A stratified sample is scored with COMETKiwi for reference-free quality labels, which train a cheap proxy to score the full corpus from target-side features and filter low-quality translations.

resulting data through from-scratch pretraining. LLäMmlein (Pfister et al., 2025) instead emphasizes transparency for compact German-only models, releasing German decoder models, training data, code, and checkpoints. German Commons (Gienapp et al., 2025) focuses on a different axis: verifiable licensing and provenance, assembling a large corpus of openly licensed German text across legal, scientific, cultural, political, news, economic, and web domains. Other efforts adapt or build open models for German and European language coverage, including LeoLM (Plüster, 2023), Occiglot (Avramidis et al., 2024), OpenGPT-X/Teuken (Ali et al., 2025), and EuroLLM (Martins et al., 2024). These works are complementary to KletterMix: they primarily improve German or European modeling through native web curation, continued pretraining, multilingual mixtures, or licensing-driven corpus construction, whereas KletterMix studies whether a high-quality English mixture can be transferred into German while preserving document boundaries, source metadata, and topical structure.

Machine-translated data for pretraining. Machine-translated corpora can transfer high-resource English data to languages with smaller, less curated, or less diverse pretraining corpora. Recent work shows this can work when done carefully: translation-based pretraining translates web-crawled documents into low-resource languages and filters the synthetic text, and finds that filtered translations can be competitive with native clean data for smaller models (Doshi et al., 2024). TransWebEdu (Wang et al., 2025) scales this idea to LLM pretraining by translating FineWeb-Edu into multiple languages and training a multilingual model from scratch on the resulting corpus. At the same time, translated data can introduce translationese artifacts, source-language bias, unnatural target-language style, and quality failures. Large-scale web corpora may already contain substantial amounts of low-quality machine-translated content (Thompson et al., 2024), and German-focused curation work has cautioned that English-to-German machine translation should be accompanied by rigorous quality and naturalness checks (Burns et al., 2026). KletterMix therefore treats translation not merely as data augmentation but as a dataset-construction problem: we preserve aligned document structure and metadata, measure translation quality with reference-free quality estimation, and evaluate the translated corpus through controlled pretraining and annealing ablations.

Dataset documentation and translation quality assessment. Dataset documentation frameworks such as Datasheets for Datasets (Gebru et al., 2021) and Data Statements for NLP (Bender and Friedman, 2018) motivate explicit reporting of a dataset's motivation, composition, collection and processing pipeline, intended uses, limitations, and sociotechnical context. These concerns are especially important for translated pretraining corpora, where source provenance, transformation steps, language variety, and translation artifacts all affect downstream use. For translation quality, reference-free quality estimation models such as COMETKiwi (Rei et al., 2022) are attractive because they can score source–translation pairs without human references while supporting scalable corpus-level diagnostics. KletterMix follows this line of work by coupling corpus-level documentation with translation-quality analyses, metadata preservation checks, and model-training validation; we treat automatic quality estimation as a scalable diagnostic rather than a replacement for downstream evaluation or manual inspection.

3 KletterMix Pipeline

We build KletterMix by translating a high-quality English pretraining mixture into German, preserving document boundaries, source metadata, and mixture structure. The pipeline has five stages: source-

record normalization, length-aware routing, document-preserving translation, shard-wise scalable execution, and post-hoc quality estimation. Here we describe the procedure; variants and corpus-level analyses appear in Sec. 3 and Sec. 4, with implementation and proxy-validation details in App. A.

Source corpus and record structure The source corpus is stored as sharded JSONL files. Each record contains a document identifier, the English source text, and metadata inherited from the original mixture, including source cluster, source location, and approximate source length. We preserve this information throughout the translation pipeline. Each translated record, therefore, consists of the original document identifier, the German translation, the unchanged metadata fields, and additional processing metadata such as context bucket, chunking status, translation configuration, and quality-estimation outputs. Preserving document-level structure is important for two reasons. First, it allows KletterMix to retain the source mixture design of the English corpus rather than collapsing the data into sentence-level translation pairs. Second, it enables direct corpus-level comparisons between the English source and German target, including analyses of length ratios, source composition, cluster-level quality, and metadata preservation.

Length-aware routing Pretraining corpora contain documents with highly variable lengths, from short snippets to long web pages and technical documents. A single translation configuration is inefficient for such data: short documents waste context budget, while long documents may exceed the effective context or generation limits of the translation model. We therefore route documents into length-aware buckets before translation. Concretely, each document is assigned to a bucket according to its approximate source length. We use a small set of context buckets chosen to cover short, medium, long, and overflow documents under the practical context and generation limits of the translation setup. The same document-preserving procedure is used for all buckets, with bucket-specific source and target budgets. This reduces padding and scheduling inefficiency while allowing longer documents to be handled without imposing the most restrictive configuration on the entire corpus.

Document-preserving translation Each source document is translated either in a single pass or through contextualized chunking. Documents that fit in the effective source budget of their assigned bucket are translated directly. Longer documents are first segmented into sentences and then greedily packed into source chunks up to a fixed token budget. If an individual sentence exceeds the chunk budget, we split it at the token level as a fallback. This ensures every source document can be translated while preserving its document identity. For chunked documents, chunks are translated sequentially. To improve discourse continuity, terminology consistency, and local coherence across chunk boundaries, the prompt for chunk t includes a truncated window of the German translation produced for chunk $t-1$. The model is instructed to use this previous translation only as context and to output only the translation of the current source chunk. The final German document is obtained by concatenating translated chunks in their original order. This design makes a trade-off between scalability and document coherence. Full-document translation is preferred whenever possible, but contextualized chunking allows us to translate documents that exceed the practical context window while still exposing the model to local context from the preceding chunk.

Dynamic target-side budgeting English-to-German translation can change document length depending on genre, domain, and formatting. We therefore do not use a fixed generation cap for all documents. Instead, the maximum target-side generation length is derived from the source chunk length and capped by a global maximum. Formally, for a source chunk of length ℓ_{src}, we set the target budget to $\ell_{\mathrm{tgt}}^{\max} = \min\left(L_{\max}, \lceil \alpha \ell_{\mathrm{src}} + \beta \rceil\right)$, where $L_{\max}$ is the maximum generation budget and α, β are chosen to allow moderate target-side expansion. This avoids over-allocating decoding budget to short chunks while reducing truncation risk for long or expansion-heavy documents.

Scalable shard-wise execution Translation is performed shard-wise. Multiple workers process disjoint subsets of the source corpus and write translated records incrementally. Workers preserve the original shard and document identifiers, which makes translated outputs traceable to the source corpus and simplifies resumption after interruption. To make the pipeline robust to transient failures, workers write intermediate attempt files and finalize a shard only after all records in that shard have been processed. When a run is interrupted, processing resumes from the last completed record rather than restarting the shard. Translation requests are distributed across a pool of model endpoints. Workers monitor endpoint availability and requeue unfinished documents when an endpoint becomes unavailable. These engineering choices do not affect the dataset definition, but they are important for reproducibility at corpus scale. App. A.1 motivates our translation model, precision, and speculative-decoding choices; App. A.2 provides the specific translation, prompt, and serving settings.

Pilot subset for translation-quality estimation Full-corpus reference-free quality estimation is computationally expensive. We therefore first construct a pilot subset for quality analysis and proxy-model training. The pilot subset is sampled in a stratified manner over source clusters. Within each cluster, documents are ranked by a stable hash of their document identifier, and the top-ranked documents are selected until the cluster-specific quota is reached. This produces a reproducible sample while preserving the mixture structure of the source corpus. The pilot subset is translated with the same pipeline as the full corpus. We then score the resulting source–translation pairs with COMETKiwi (Rei et al., 2022), a reference-free quality-estimation model. We use these scores as corpus-level diagnostic signals rather than as a substitute for downstream training evaluation. In particular, COMETKiwi is used to identify quality variation across length buckets, source clusters, and document types, and to supervise a cheaper proxy model for full-corpus scoring.

Proxy-based quality annotation Full-corpus COMETKiwi scoring is too costly to annotate every translation. We therefore train a cheap proxy annotator on the COMETKiwi-scored pilot data to predict COMETKiwi scores. The deployed proxy is a gradient-boosted regression model using only target-side features: it reads the translated German document and low-cost metadata from that text. Because it does not require the English source, it can score released German shards directly and avoids an expensive source-rehydration pass over the full corpus.

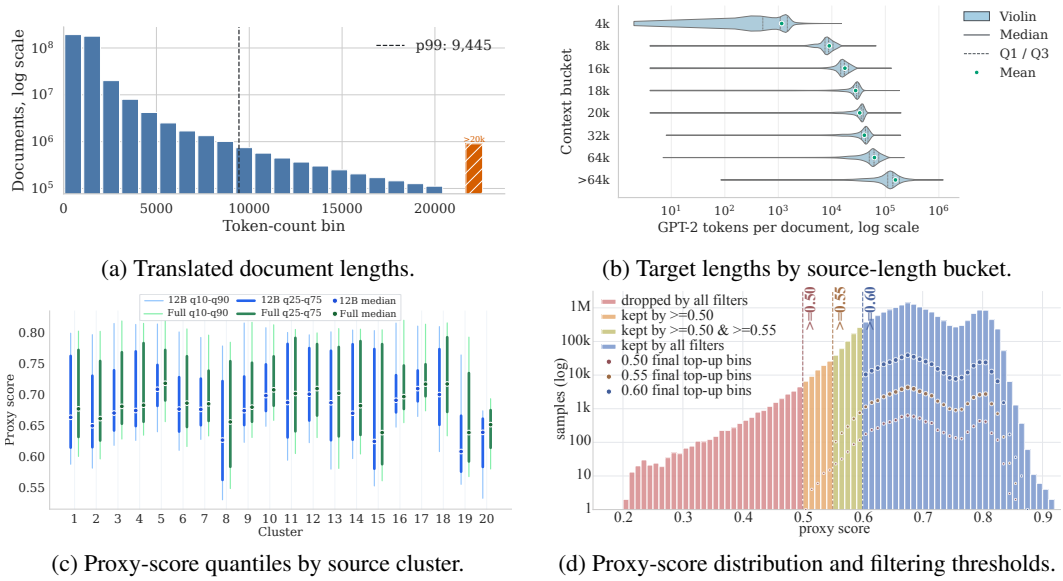
After translation, we run GlotLID (Kargaran et al., 2023) on each German target document to obtain language identification for the proxy. GlotLID is loaded as a fastText model and applied to the translated text after replacing newlines with spaces. We restrict scoring to a fixed set of common language/script labels and compute normalized probabilities over it. From this, we derive whether the top label is German Latin (`deu_Latn`), the normalized German-Latin probability, a clipped logit transform of that probability, the top-1/top-2 probability margin, and the script extracted from the top predicted label. The GlotLID-derived signals are combined with low-cost target-side text-shape features: length, lexical diversity, token repetition, average token length, and character composition ratios.

Tab. 5 summarizes the deployed feature set. The GlotLID features capture wrong-language output and low-confidence German predictions; script, length, and character-composition features capture abnormal text shape, formatting artifacts, and suspicious character mixtures; and lexical-diversity and repetition features capture degenerate or repetitive generations. The proxy cannot directly measure semantic adequacy, because it does not see the English source. Its role is therefore not to replace source-aware evaluation or human inspection, but to provide a scalable corpus-level quality signal that is validated against COMETKiwi and targets practical failure modes that matter at release scale.

We validate the proxy on a disjoint 18,275-document split scored with COMETKiwi; full agreement metrics are reported in Tab. 6. The target-only proxy shows strong agreement with COMETKiwi and low absolute error, making it suitable as a scalable ranking and filtering signal rather than as a replacement for source-aware quality evaluation. During development, source-aware variants did not improve validation agreement enough to justify rehydrating the English source text for full-corpus scoring. We therefore deploy the target-only model as the full-corpus annotator.

Proxy-filtered dataset variants The primary release of KletterMix is the unfiltered translated corpus, excluding only records that fail basic validity checks such as empty output, missing metadata, or severe language-identification failure. In addition, we construct three proxy-filtered 12B-token training splits by retaining only translated documents whose predicted COMETKiwi proxy score exceeds a fixed threshold: $\hat{q}_{\text{proxy}} \geq 0.50$, $\hat{q}_{\text{proxy}} \geq 0.55$ and $\hat{q}_{\text{proxy}} \geq 0.60$. These thresholds are not intended as universal quality cutoffs. They define filtering ablations: under a fixed 12B-token training budget, we ask whether removing progressively more low-scoring translated text improves optimization behavior, held-out validation loss, or downstream benchmark performance. Unless otherwise stated, corpus-level analyses refer to the unfiltered translated corpus, while training experiments compare two external German-data baselines, FineWeb2-DE and GermanWeb, against unfiltered KletterMix and the three threshold-filtered KletterMix splits. Fig. 2d visualizes the full proxy-score distribution and marks the three retained-document thresholds used for these ablations.

Because KletterMix is constructed by translating an existing English mixture while preserving document boundaries and metadata, the resulting German corpus can be characterized along the same structural axes as its source. We therefore describe the release in terms of document-length distributions, length-bucket consistency, and cluster-level proxy-score patterns across both the 12B-token subset and the full corpus. These corpus-level views serve as documentation and quality-control



(a) Translated document lengths.

(b) Target lengths by source-length bucket.

(c) Proxy-score quantiles by source cluster.

(d) Proxy-score distribution and filtering thresholds.

Figure 2: Corpus diagnostics and proxy-filtering thresholds for KletterMix. (a,b) Length plots show the heavy-tailed length profile and legth distribution in the context buckets. (c) Proxy scores for the 12B-token subset and full corpus show the 10th–90th and 25th–75th percentiles; points show medians. (d) Full-corpus log-count histogram with the 0.50, 0.55, and 0.60 retained-document thresholds used for the three 12B-token filter ablations; colored regions indicate which documents each filter retains.

signals: they show whether the translated data preserves the expected length profile of a high-quality pretraining mixture, reveal unusually short translations in long-document buckets, and highlight source clusters that may warrant closer audit. The proxy thresholds above define the optional filtered variants, while the analyses below characterize the structure and quality profile of the released corpus. App. B gives the cluster-labeling procedure and qualitative examples used for audit.

4 Translation Insights

Fig. 2 summarizes these corpus-level diagnostics for the full KletterMix release and the 12B-token training subset. The two length-based panels inspect global target-document lengths and bucket-level length consistency, while the cluster-level proxy-score panel checks whether fixed-budget sampling preserves the full-corpus quality profile across inherited source clusters.

Fig. 2a shows the overall document-token distribution. The corpus follows a heavy-tailed length profile: most translated documents are relatively short, while a smaller number of documents extend into the long-context regime. This structure is typical of heterogeneous web and pretraining mixtures and motivates the length-aware translation pipeline described in Sec. 3. Fig. 2b provides a more fine-grained view by comparing translated document lengths within each canonical source-length bucket. Because token counts are tokenizer-dependent and can vary across languages, we do not expect a fixed one-to-one correspondence between English source length and German target length. Nevertheless, English-to-German translation often leads to surface-length expansion, and German may require more subword tokens under a given tokenizer due to language-specific morphology and tokenizer fertility effects (Ahia et al., 2023; Rust et al., 2021). Thus, documents assigned to longer source buckets should generally also produce relatively long German translations, up to expected variation from tokenization, genre, and translation style. The violin plot confirms this trend but shows long lower tails: some long-context documents produce much shorter-than-expected translations. This is suspicious because the source required a large context bucket, yet the German output is short, suggesting truncation, dropped content, or other translation failures. We use these lower tails as diagnostics and audit signals. Finally, Fig. 2c compares cluster-level proxy-score quantiles for the 12B-token training subset and the full translated release. Across clusters, the panel shows the 10th–90th and the 25th–75th percentile range, and the median target-only proxy score, making

Table 1: Downstream 5-shot accuracy on German evaluations under matched 12B-token training conditions. Bold and underline mark the best and second-best point estimates within each metric column across the full table; ties are marked identically. Rows under FineWeb2-DE annealing continue from the same FineWeb2-DE checkpoint. Details on score calculations are in App. A.5

Run	MMLU	PIQA	HellaSwag	ARC-C	Core Avg.
Independent pretraining					
GermanWeb	**30.0** ± 2.3	63.0 ± 4.9	31.2 ± 0.5	23.1 ± 1.2	36.8 ± 1.4
FineWeb2-DE	28.7 ± 2.3	**70.0** ± 4.6	31.5 ± 0.5	23.0 ± 1.2	38.3 ± 1.3
FineWeb2-DE annealing					
→ GermanWeb	**30.0** ± 2.3	65.0 ± 4.8	32.2 ± 0.5	23.1 ± 1.2	37.6 ± 1.4
→ KletterMix	29.0 ± 2.3	69.0 ± 4.6	34.2 ± 0.5	25.2 ± 1.3	39.4 ± 1.3
KletterMix pretraining variants					
KletterMix	29.0 ± 2.3	65.0 ± 4.8	34.4 ± 0.5	26.5 ± 1.3	38.7 ± 1.4
KletterMix-Filt$_{0.60}$	28.5 ± 2.3	**70.0** ± 4.6	**34.6** ± 0.5	**27.5** ± 1.3	**40.2** ± 1.3

it a corpus diagnostic rather than a filtering-threshold figure. The comparison tests whether the fixed-budget training subset matches the release's cluster-level quality profile and flags clusters whose proxy-score distributions need closer review.

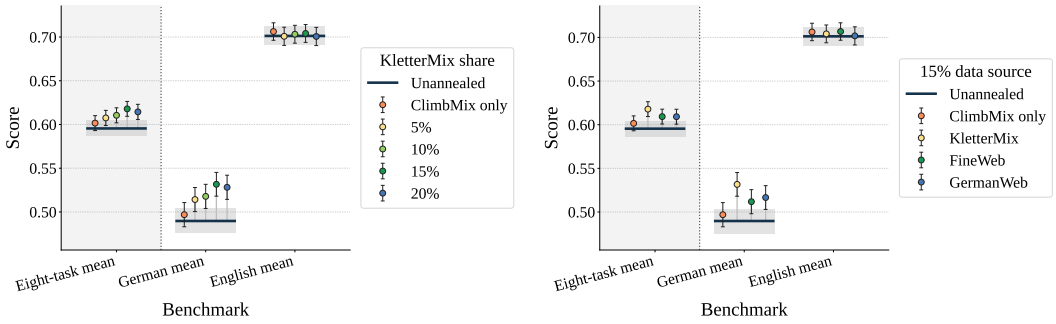
These analyses suggest that KletterMix retains a diverse, heavy-tailed pretraining mix while providing document- and cluster-level quality-control signals. Proxy filtering also reduces the aggregate linguistic-profile distance to a matched German-web reference (App. B.1). The controlled filtering experiments in Sec. 5 isolate the impact of target-only proxy-score thresholds.

5 Training Ablations

We test if KletterMix improves model training relative to established German data. While previous sections analyze the translated corpus directly, this section tests its usefulness under controlled model, compute, and token-budget conditions. We first compare corpora through matched 0.6B-parameter pretraining and annealing, and then vary the KletterMix fraction in a 7B-parameter annealing mixture.

Benchmark evaluation. We evaluate all checkpoints with lm-eval-harness (Gao et al., 2024) using 5-shot accuracy. The German suite uses the German or multilingual task configurations corresponding to MMLU (Hendrycks et al., 2021; Singh et al., 2025), PIQA (Bisk et al., 2020; Chang et al., 2025), HellaSwag (Zellers et al., 2019; Thellmann et al., 2024), and ARC-Challenge (Clark et al., 2018; Thellmann et al., 2024). The 7B experiment uses the same German task settings and also evaluates the original English tasks, both at 5-shot accuracy. Standard errors are reported next to each accuracy. We use the four-task average as our main German aggregate, because the four tasks cover complementary axes of reasoning transfer rather than a single narrow skill. MMLU probes broad multitask knowledge and problem solving across academic and professional subjects (Hendrycks et al., 2021; Singh et al., 2025). PIQA targets physical commonsense reasoning, testing whether a model can choose plausible actions involving everyday objects and affordances (Chang et al., 2025; Bisk et al., 2020). HellaSwag evaluates grounded commonsense inference by asking the model to select the plausible continuation of an everyday scenario, with distractors that are fluent but semantically implausible (Zellers et al., 2019; Thellmann et al., 2024). ARC-Challenge evaluates hard science-style question answering, where models must combine facts rather than only match surface patterns (Clark et al., 2018; Thellmann et al., 2024). Together, these tasks offer four complementary probes: general knowledge, physical affordance reasoning, event plausibility, and scientific compositional inference.

Matched 0.6B setup. We conduct controlled pretraining ablations using Qwen3-0.6B (Yang et al., 2025) as the base architecture and follow a Megatron-LM training recipe (Shoeybi et al., 2020). Each run is trained on a matched German 12B-token subset, corresponding to ∼20 tokens per parameter for a 0.6B-parameter model, following the Chinchilla scaling recommendation (Hoffmann et al., 2022). We compare KletterMix against FineWeb2-DE (Penedo et al., 2025) and GermanWeb (Burns et al., 2026) under matched conditions, including model architecture, optimizer, batch size, learning-rate schedule, tokenizer, token budget, and validation protocol (App. C.1). For each corpus, we construct the training subset with a deterministic, token-budgeted stratified sampler. The sampler estimates document-level token counts, preserves the main structure of each source corpus through



(a) KletterMix-share sweep.

(b) Matched 15% source comparison.

Figure 3: Aggregate 5-shot benchmark scores for the OLMo 3 7B after ∼12B annealing tokens. *Left:* fixed 15% source comparison of the unannealed checkpoint and ClimbMix-only control against matched KletterMix, FineWeb2-DE, and GermanWeb runs. *Right:* KletterMix-share sweep from 0% (ClimbMix only) to 20%, with the mean peaking at 15%. German and English means average four tasks each; the eight-task mean averages all eight. Error bars show propagated standard errors.

corpus-specific strata, and selects documents by hashing stable document identifiers until the target token budget is reached. For KletterMix, the strata follow the inherited source-cluster and context-length structure; for FineWeb2-DE and GermanWeb, they follow the available source and crawl metadata. Validation is performed on separate held-out splits produced with the same sampling procedure but disjoint selection windows, ensuring that validation documents are not included in the corresponding training subsets. We track training loss throughout optimization and evaluate validation loss and perplexity at regular intervals. In addition to the unfiltered version of KletterMix, we evaluate proxy-filtered variants constructed using the quality model described in Sec. 3.

0.6B pretraining. Tab. 1 reports downstream benchmark performance for the matched 0.6B German training experiments. Under this aggregate, the point estimates place every KletterMix-family row above the external baselines. The weakest KletterMix family result, unfiltered KletterMix at 38.7, is slightly above FineWeb2-DE at 38.3, while the validation-selected filtered split reaches the best overall point estimate of 40.2. Because PIQA is evaluated on a small 100-example subset, the Core Avg. standard errors are comparatively wide, so we interpret small aggregate gaps cautiously. The more stable signal is the recurring task-level pattern: KletterMix is consistently strongest on HellaSwag and ARC-C, suggesting that its main benefit lies in grounded event continuation and compositional science-style reasoning rather than uniform gains on every benchmark.

0.6B annealing. Tab. 1 also addresses the question which corpus leads to a better final performance after FineWeb2-DE pretraining. Annealing FineWeb2-DE on GermanWeb raises MMLU to 30.0 and slightly improves HellaSwag, but it leaves ARC-C essentially unchanged and reduces PIQA, yielding a Core Avg. of 37.6. Annealing the same FineWeb2-DE checkpoint on KletterMix instead yields 39.4, improving over the FineWeb2-DE source checkpoint by +1.1 points and over the GermanWeb annealing branch by +1.8 points. The gain is concentrated where we expect a high-quality translated mixture to matter most: HellaSwag rises from 31.5 to 34.2, and ARC-C from 23.0 to 25.2, while PIQA remains close to the FineWeb2-DE baseline. This makes annealing a more precise control than the independent pretraining comparison: it suggests that KletterMix is not merely a better initialization corpus, but a useful late-stage sharpening corpus for event coherence and science-style reasoning.

7B language-mixture annealing. To test whether the result extends beyond the 0.6B recipe, we independently anneal the OLMo 3 7B Stage 1 checkpoint (Olmo et al., 2026) for a matched budget of ∼12B tokens. All runs start from the same pretrained weights, initialize fresh optimizer and scheduler states, and use the same optimization schedule. The comparison focuses on the 15% KletterMix condition, which gives the strongest aggregate result under the primary 5-shot evaluation (Fig. 3a). For the German component, we begin with the previously defined proxy-filtered pool ($\hat{q}_{\text{proxy}} \geq 0.60$) and prioritize its highest-scoring documents; this procedure yields a proxy-score range of 0.70–0.90. In Fig. 3b we compare the 15% KletterMix condition with matched 15% FineWeb2-DE and GermanWeb controls, evaluating the same four German tasks and their English counterparts with 5-shot accuracy. At the fixed 15% replacement share, KletterMix obtains the highest German mean

(53.2%) and eight-task mean (61.8%), exceeding FineWeb2-DE by 2.0 and 0.9 percentage points, respectively, and GermanWeb by 1.5 and 0.9 points. This matched comparison indicates that the gain is not explained only by replacing ClimbMix with more German-language tokens: the source of those tokens also matters. The complete sweep and supplementary protocols are reported in App. C.3.

6 Conclusion

Discussion. Across matched 0.6B pretraining, corpus-choice annealing, and 7B language-mixture annealing, the results support the central premise of KletterMix: carefully translated and documented data can transfer more than German surface form; it can transfer useful mixture structure. The Core Avg. over MMLU, PIQA, HellaSwag, and ARC-C gives the KletterMix family the strongest point estimates under our matched recipe, while the attached evaluation-set standard errors make clear that small aggregate gaps should not be overread. The pattern is not a uniform sweep over all tasks, and that is the point. GermanWeb remains strongest on MMLU, and FineWeb2-DE remains highly competitive on PIQA, but KletterMix is strongest where coherent document structure and dense explanatory content should matter most: HellaSwag and ARC-C. We interpret this as evidence of *reasoning transfer through data curation*, not merely that translated text outweighs native German web text.

The two annealing experiments strengthen this interpretation in complementary ways. At 0.6B parameters, continuing the same FineWeb2-DE checkpoint on KletterMix gives a better Core Avg. than continuing it on GermanWeb, with gains concentrated on HellaSwag and ARC-C. At 7B parameters, the primary 15% KletterMix condition gives the highest German and eight-task means in the 5-shot comparison, while the English aggregate remains nearly unchanged. The first comparison asks which German corpus is the better final steering signal; the second isolates the source of a fixed language-adaptation component in a larger pretrained model's annealing mixture. Together, they suggest that KletterMix is useful not only as an initialization corpus but also as a late-stage language-adaptation component. The appendix reports the full mixture sweep and additional zero-shot source comparisons. Filtering further shows that proxy scores help rank data under a fixed budget but are not a universal quality metric, since stricter filtering does not monotonically improve everything.

Limitations. Despite these positive results, KletterMix has several limitations. Because the corpus is derived from a mixture of English sources, it may inherit the topical, cultural, geographic, stylistic, and licensing biases of those sources, even after translation into German. Machine translation can also introduce translationese, semantic drift, unnatural German style, inconsistent terminology, or failures in long, highly specialized documents. Our COMETKiwi-based and proxy-based quality estimates provide scalable diagnostics, but they are not substitutes for human evaluation, downstream task evaluation, or careful inspection of problematic domains. The model-based evidence spans 0.6B and 7B parameters, however, each condition has one training run, both regimes use an $\sim$12B-token budget, the 7B experiment uses a single starting checkpoint and a quality-ranked KletterMix subset; future work should test larger models, additional random seeds, and broader downstream benchmark suites.

Evaluative role. KletterMix serves as pretraining corpus and aligned resource for evaluating translation-based data curation. Since each German document preserves its source identifier, metadata, source cluster, and length bucket, it enables controlled studies of: (i) transferring a high-quality English mixture to German under fixed token and compute budgets; (ii) how translation quality varies by document length, source cluster, and text type; and (iii) how proxy-based filtering affects German language-model training. It is not a substitute for native German data, evidence of cultural representativeness, or a guarantee that proxy scores reflect per-document semantic adequacy.

Future Work. Future work should extend this approach along several dimensions. First, stronger filtering could target translation failures not fully captured by length or language-identification diagnostics, such as URL-only documents, boilerplate, duplicated content, or subtle semantic drift. Second, manual audits and benchmark evaluations can better characterize naturalness, factual preservation, and downstream utility across domains. Third, the same document-preserving translation pipeline could be applied to other languages such as French, Italian, and Spanish, enabling systematic comparisons of when translated pretraining mixtures complement native-language web corpora. Overall, KletterMix suggests that careful translation, corpus documentation, and empirical validation can be a practical path toward stronger non-English pretraining data, provided that the limitations of translated corpora are made explicit and addressed throughout the release process.

REFERENCES

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R. Mortensen, Noah A. Smith, and Yulia Tsvetkov. Do all languages cost the same? tokenization in the era of commercial language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 9904–9923. Association for Computational Linguistics, 2023.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Jan Ebert, Alexander Arno Weber, Richard Rutmann, Charvi Jain, Max Lübbering, Daniel Steinigen, Johannes Leveling, Katrin Klug, Jasper Schulze Buschhoff, Lena Jurkschat, Hammam Abdelwahab, Benny Jörg Stein, Karl-Heinz Sylla, Pavel Denisov, Nicolo' Brandizzi, Qasid Saleem, Anirban Bhowmick, Lennard Helmer, Chelsea John, Pedro Ortiz Suarez, Malte Ostendorff, Alex Jude, Lalith Manjunath, Samuel Weinbach, Carolin Penke, Oleg Filatov, Fabio Barth, Paramita Mirza, Lucas Weber, Ines Wendler, Rafet Sifa, Fabian Küch, Andreas Herten, René Jäkel, Georg Rehm, Stefan Kesselheim, Joachim Köhler, and Nicolas Flores-Herr. Teuken-7b-base & teuken-7b-instruct: Towards european llms. In Inês Lynce, Nello Murano, Mauro Vallati, Serena Villata, Federico Chesani, Michela Milano, Andrea Omicini, and Mehdi Dastani, editors, *ECAI 2025 - 28th European Conference on Artificial Intelligence, 25-30 October 2025, Bologna, Italy - Including 14th Conference on Prestigious Applications of Intelligent Systems (PAIS 2025)*, Frontiers in Artificial Intelligence and Applications, pages 4321–4329. IOS Press, 2025.

Eleftherios Avramidis, Annika Grützner-Zahn, Manuel Brack, Patrick Schramowski, Pedro Ortiz Suarez, Malte Ostendorff, Fabio Barth, Shushen Manakhimova, Vivien Macketanz, Georg Rehm, and Kristian Kersting. Occiglot at WMT24: european open-source large language models evaluated on translation. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz, editors, *Proceedings of the Ninth Conference on Machine Translation, WMT 2024, Miami, FL, USA, November 15-16, 2024*, pages 292–298. Association for Computational Linguistics, 2024.

Emily M. Bender and Batya Friedman. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Trans. Assoc. Comput. Linguistics*, 6:587–604, 2018.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

Thomas F. Burns, Letitia Parcalabescu, Stephan Wäldchen, Michael Barlow, Gregor Ziegltrum, Volker Stampa, Bastian Harren, and Björn Deiseroth. Aleph-alpha-germanweb: Improving german-language LLM pre-training with model-based data curation and synthetic data generation. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2026 - Volume 1: Long Papers, Rabat, Morocco, March 24-29, 2026*, pages 1267–1283. Association for Computational Linguistics, 2026.

Tyler A. Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplício, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, Alfred Malengo Kondoro, Alham Fikri Aji, Ali Eren Çetintaş, Allan Hanbury, Alou Dembele, Alp Niksarli, Álvaro Arroyo, Amin Bajand, Amol Khanna, Ana Chkhaidze, Ana Condez, Andiswa Mkhonto, Andrew Hoblitzell, Andrew Tran, Angelos Poulis, Anirban Majumder, Anna Vacalopoulou, Annette Kuuipolani Kanahele Wong, Annika Simonsen, Anton Kovalev, Ashvanth. S, Ayodeji Joseph Lana, Barkin Kinay, Bashar Alhafni, Benedict Cibalinda Busole, Bernard Ghanem, Bharti Nathani, Biljana Stojanovska Đurić, Bola Agbonile, Bragi Bergsson, Bruce Torres Fischer, Burak Tutar, Burcu Alakuş Çınar, Cade J. Kanoniakapueo Kane, Can Udomcharoenchaikit, Catherine Arnett, Chadi Helwe, Chaithra Reddy Nerella, Chen Cecilia Liu, Chiamaka Glory Nwokolo, Cristina España-Bonet, Cynthia Amol, DaeYeop Lee, Dana Arad, Daniil Dzenhaliou, Daria Pugacheva, Dasol Choi, Daud Abolade, David Liu, David Semedo, Deborah Popoola, Deividas Mataciunas, Delphine Nyaboke, Dhyuthy Krishna Kumar, Diogo Glória-Silva, Diogo Tavares, Divyanshu Goyal, DongGeon Lee, Ebele Nwamaka Anajemba, Egonu Ngozi Grace, Elena Mickel, Elena

Tutubalina, Elias Herranen, Emile Anand, Emmanuel Habumuremyi, Emuobonuvie Maria Ajiboye, Eryawan Presma Yulianrifat, Esther Adenuga, Ewa Rudnicka, Faith Olabisi Itiola, Faran Taimoor Butt, Fathima Thekkekara, Fatima Haouari, Filbert Aurelian Tjiaranata, Firas Laakom, and Francesca Grasso et al. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures, 2025. URL https://arxiv.org/abs/2510.24081.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. arXiv:1803.05457.

Amin Dada, Aokun Chen, Cheng Peng, Kaleb E. Smith, Ahmad Idrissi-Yaghir, Constantin Seibold, Jianning Li, Lars Heiliger, Christoph M. Friedrich, Daniel Truhn, Jan Egger, Jiang Bian, Jens Kleesiek, and Yonghui Wu. On the impact of cross-domain data on german language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 13801–13813. Association for Computational Linguistics, 2023.

Ona de Gibert, Graeme Nail, Nikolay Arefyev, Marta Bañón, Jelmer van der Linde, Shaoxiong Ji, Jaume Zaragoza-Bernabeu, Mikko Aulamo, Gema Ramírez-Sánchez, Andrey Kutuzov, Sampo Pyysalo, Stephan Oepen, and Jörg Tiedemann. A new massive multilingual dataset for high-performance language technologies. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 1116–1128. ELRA and ICCL, 2024.

Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects. In *Findings of the Association for Computational Linguistics: ACL 2025*, 2025.

Shizhe Diao, Yu Yang, Yonggan Fu, Xin Dong, Dan Su, Markus Kliegl, Zijia Chen, Peter Belcak, Yoshi Suhara, Hongxu Yin, Mostofa Patwary, Yingyan, Lin, Jan Kautz, and Pavlo Molchanov. Nemotron-climb: Clustering-based iterative data mixture bootstrapping for language model pre-training, 2025. arXiv:2504.13161.

Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1286–1305. Association for Computational Linguistics, 2021.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. Pretraining language models using translationese. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 5843–5862. Association for Computational Linguistics, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. The pile: An 800gb dataset of diverse text for language modeling, 2020. arXiv:2101.00027.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 07 2024.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. Datasheets for datasets. *Commun. ACM*, 64(12):86–92, 2021.

Lukas Gienapp, Christopher Schröder, Stefan Schweter, Christopher Akiki, Ferdinand Schlatt, Arden Zimmermann, Phillipe Genêt, and Martin Potthast. The german commons - 154 billion tokens of openly licensed text for german language models, 2025. arXiv:2510.13996.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Oriol Vinyals, Jack W. Rae, and Laurent Sifre. An empirical analysis of compute-optimal large language model training. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. Glotlid: Language identification for low-resource languages. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 6155–6218. Association for Computational Linguistics, 2023.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Javier Ortiz Suárez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Balli, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. Quality at a glance: An audit of web-crawled multilingual datasets. *Trans. Assoc. Comput. Linguistics*, 10:50–72, 2022.

Pierre-Carl Langlais, Pavel Chizhov, Catherine Arnett, Carlos Rosas Hinostroza, Mattia Nee, Eliot Krzysztof Jones, Irène Girard, David Mach, Anastasia Stasenko, and Ivan P. Yamshchikov. Common corpus: The largest collection of ethical data for LLM pre-training. In *The Fourteenth International Conference on Learning Representations*, 2026.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Sasko, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben Allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Ifeoluwa Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Alexandra Sasha Luccioni, and Yacine Jernite. The bigscience ROOTS corpus: A 1.6tb composite multilingual dataset. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Kumar Guha, Sedrick Scott Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel, Jean Mercat, Mayee F. Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bitton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej Kilian, Hanlin Zhang, Rulin Shao, Sarah M. Pratt, Sunny Sanyal, Gabriel

Ilharco, Giannis Daras, Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Raghavi Chandu, Thao Nguyen, Igor Vasiljevic, Sham M. Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer, Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groeneveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alex Dimakis, Yair Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the next generation of training sets for language models. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. EuroLLM: Multilingual language models for europe, 2024. arXiv:2409.16235.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In Nicoletta Calzolari, Min-Yen Kan, Véronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 4226–4237. ELRA and ICCL, 2024.

Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, et al. Hplt 3.0: Very large-scale multilingual resources for llm and mt. mono-and bi-lingual data, multilingual evaluation, and pre-trained models. *arXiv preprint arXiv:2511.01066*, 2025.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. Olmo 3, 2026. URL `https://arxiv.org/abs/2512.13961`.

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. In *Second Conference on Language Modeling*, 2025.

Jan Pfister, Julia Wunderle, and Andreas Hotho. Llämmlein: Transparent, compact and competitive german-only language models from scratch. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 2227–2246. Association for Computational Linguistics, 2025.

Björn Plüster. Leolm: Igniting german-language llm research. LAION Blog, September 2023. URL `https://laion.ai/blog/leo-lm/`. Accessed: May 5, 2026.

Ricardo Rei, Marcos V. Treviso, Nuno Miguel Guerreiro, Chrysoula Zerva, Ana C. Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte M. Alves, Luísa Coheur, Alon Lavie, and André F. T. Martins. CometKiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In Philipp Koehn, Loïc Barrault, Ondrej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno-Yepes, Tom Kocmi, André F. T. Martins, Makoto Morishita, Christof Monz, Masaaki Nagata,

Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri, editors, *Proceedings of the Seventh Conference on Machine Translation, WMT 2022, Abu Dhabi, United Arab Emirates (Hybrid), December 7-8, 2022*, pages 634–645. Association for Computational Linguistics, 2022.

Phillip Rust, Jonas Pfeiffer, Ivan Vulic, Sebastian Ruder, and Iryna Gurevych. How good is your tokenizer? on the monolingual performance of multilingual language models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 3118–3135. Association for Computational Linguistics, 2021.

Raphael Scheible, Johann Frei, Fabian Thomczyk, Henry He, Patric Tippmann, Jochen Knaus, Victor Jaravine, Frank Kramer, and Martin Boeker. Gottbert: a pure german language model. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 21237–21250. Association for Computational Linguistics, 2024.

Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-LM: Training multi-billion parameter language models using model parallelism, 2020. arXiv:1909.08053.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, André F. T. Martins, Leshem Choshen, Daphne Ippolito, Enzo Ferrante, Marzieh Fadaee, Beyza Ermis, and Sara Hooker. Global MMLU: understanding and addressing cultural and linguistic biases in multilingual evaluation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 18761–18799. Association for Computational Linguistics, 2025.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Raghavi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Harsh Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Pete Walsh, Luke Zettlemoyer, Noah A. Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. Dolma: an open corpus of three trillion tokens for language model pretraining research. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 15725–15788. Association for Computational Linguistics, 2024.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. A monolingual approach to contextualized word embeddings for mid-resource languages. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1703–1714. Association for Computational Linguistics, 2020.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. Towards multilingual llm evaluation for european languages, 2024. URL `https://arxiv.org/abs/2410.08928`.

Brian Thompson, Mehak Preet Dhaliwal, Peter Frisch, Tobias Domhan, and Marcello Federico. A shocking amount of the web is machine translated: Insights from multi-way parallelism. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 1763–1775. Association for Computational Linguistics, 2024.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. Multilingual language model pretraining using machine-translated data. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 28087–28107. Association for Computational Linguistics, 2025.

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-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 483–498. Association for Computational Linguistics, 2021.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. arXiv:2505.09388.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: can a machine really finish your sentence? In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4791–4800. Association for Computational Linguistics, 2019.

Appendix organization. First, AI use statement, Reproducibility Statement, and Ethics Statement are given. Then, the appendix follows the order of the main paper. App. A expands the KletterMix pipeline in Sec. 3; App. B supports the corpus diagnostics in Sec. 4; and App. C gives the training configurations and extended results for Sec. 5, including the full OLMo 3 7B annealing setup, matched source controls, per-task plots, and supplementary zero-shot results in App. C.3.

AI USE STATEMENT

Generative AI was used in three roles in this work.

First, the self-hosted `Qwen3.5-397B-A17B-FP8` model served as the English-to-German translation backend and thereby generated the synthetic KletterMix corpus. The generation procedure, prompts, and quality-control pipeline are documented in Sec. 3 and App. A; the outputs were evaluated through reference-free quality estimation, target-side diagnostics, qualitative inspection, and controlled model-training experiments.

Second, the same model supported qualitative thematic analysis by producing descriptive English labels, summaries, and keywords for fixed inherited source clusters from the 100 highest-COMETKiwi-scoring examples per cluster (App. B.3). These annotations are treated as descriptive metadata rather than ground truth and were compared with an independent English annotation.

Third, OpenAI ChatGPT was used during manuscript preparation to polish wording and presentation.

AI outputs were treated as drafts rather than authoritative sources. All AI-assisted text and code incorporated into the manuscript were reviewed by an author. Quantitative statements were checked against the reported tables and figures, no citation or scientific claim was accepted solely from model output, and AI-assisted code was tested through compilation and workflow runs. Beyond the uses described above, generative AI was not used to develop theoretical or conceptual frameworks, formulate or prove mathematical claims, propose or refine research hypotheses, design the reported experiments, implement the research methods, clean or reformat the datasets, or create or modify scientific figures. The authors made the final decisions regarding the methodology, experiments, analyses, and wording, and take responsibility for all content, claims, code, and artifacts in this work.

REPRODUCIBILITY STATEMENT

We document the corpus construction, metadata preservation, quality estimation, and filtering procedures in Sec. 3, with translation settings, prompts, serving configurations, and proxy validation in App. A. Corpus diagnostics and annotation procedures are detailed in App. B. Sec. 5 describes deterministic subset sampling, disjoint training and validation selections, and the controlled comparison protocol; App. C provides model and optimization configurations, compute infrastructure, annealing mixtures, and extended evaluation results. Benchmark aggregation and uncertainty calculations are specified in App. A.5 and App. C.3. Code and dataset links are provided on the first page, and the release format is described in App. D. Each training condition uses one run; the reported benchmark standard errors quantify evaluation-set uncertainty rather than variation across training seeds.

ETHICS STATEMENT

KletterMix is intended to support research on German language-model pretraining, annealing, and translated-data curation, helping address the imbalance between English and German pretraining resources. Because it is translated from an English source mixture, it may inherit topical, cultural, geographic, and stylistic biases and should not be treated as representative of German-speaking communities. Machine translation may also introduce semantic drift, terminology errors, and translationese. KletterMix is therefore intended to complement, rather than replace, native German resources.

Web-derived source material may contain personal information, harmful content, or stereotypes, and models trained on the corpus may reproduce biased or misleading outputs or leak private information. Preserving document identifiers and source metadata supports traceability, while corpus diagnostics and optional quality filtering help identify translation failures. These measures do not establish that

the corpus is free of sensitive or harmful content. In particular, COMETKiwi and the target-only proxy are translation-quality diagnostics, not guarantees of factual accuracy, fairness, privacy, or downstream safety. We recommend application-specific privacy, bias, and safety assessments before deployment.

The release license is specified in App. D; reuse should respect the release terms and applicable source-level restrictions. Provenance preservation should not be interpreted as establishing unrestricted redistribution rights. We also acknowledge the environmental cost of large-scale translation and training and report allocated translation compute and training infrastructure in App. A.2 and App. C. The roles of generative AI and author oversight are documented separately in the AI Use Statement.

A Pipeline Implementation Details

This section collects the implementation details behind the pipeline summarized in Sec. 3. The main text describes the data-construction logic; the appendix gives the concrete translation backend, decoding settings, prompts, serving setup, and proxy-validation measurements needed to reproduce the translated corpus and its filtered variants.

A.1 Translation Backend Selection

Full-corpus construction requires translating hundreds of billions of source tokens, so we selected the translation backend under a joint quality, throughput, and stability constraint. We compared Qwen3.5-397B-A17B variants on WMT24++ (Deutsch et al., 2025) using the same translation prompt and scored outputs with XCOMET-XXL and COMETKiwi. Tab. 2 shows that FP8 matches FP16 within 0.001 absolute XCOMET overall and remains effectively tied on English–German. NVFP4 is also competitive in aggregate quality, but was less stable in our B200 vLLM deployment sweeps and was therefore not used for the full production run.

Table 2: WMT24++ translation-quality comparison across Qwen3.5-397B-A17B precision variants. Scores are means; higher is better.

Metric	FP16	FP8	NVFP4
XCOMET-XXL overall	0.8127	0.8120	0.8128
XCOMET-XXL en–de	0.9378	0.9362	0.9383
COMETKiwi overall	0.7791	0.7787	0.7753
COMETKiwi en–de	0.8112	0.8119	0.8118

We also benchmarked serving configurations on a 10,014-request ClimbMix sample with the same long-document translation setting used by the production pipeline. The strongest planning baseline used FP8 with 8-way tensor parallelism, MTP-2 speculative decoding, `max_num_batched_tokens` of 16,384, and `max_num_seqs` of 1,024. This configuration reached 7.06 requests/s, 5,633 output tokens/s, and 10,301 total tokens/s. The best clean NVFP4 reference reached a similar request rate (7.22 requests/s) but lower token-normalized throughput (4,700 output tokens/s and 9,512 total tokens/s). For token-heavy translation workloads, the token-normalized throughput was the safer planning basis and favored FP8.

Qualitative spot checks supported the same choice. NVFP4 generations were more likely to show instability on difficult examples, including premature termination, weaker sentence stability, occasional English leakage, and less reliable handling of terminology or long-context dependencies. We treat these manual checks as operational diagnostics rather than formal evaluation, but they were important for selecting a robust production path. We therefore used Qwen3.5-397B-A17B-FP8 with MTP-2 speculative decoding for the release translation run.

A.2 Translation Configuration and Execution

Sec. 3 introduces length-aware routing, document-preserving chunking, dynamic target-side budgeting, and shard-wise execution. This subsection gives the concrete production values used for those components.

Translation configuration. Tab. 3 summarizes the translation-side hyperparameters. For the dynamic target budget in Sec. 3, the implementation uses $L_{\max} = 32{,}768$, $\alpha = 2.0$, and $\beta = 1{,}024$, together with a minimum target budget of 2,048 tokens. Thus, for a source chunk of length ℓ_{src}, the generation cap is

$$\ell_{\mathrm{tgt}}^{\max} = \max\left(2048, \min\left(32768, \lceil 2.0 \cdot \ell_{\mathrm{src}} + 1024 \rceil\right)\right).$$

Documents whose metadata indicates that they safely fit inside the source-chunk budget are translated in a single pass. Longer documents are sentence-segmented, greedily packed into chunks of up to 20k source tokens, and translated with a 2k-token previous-translation context window.

Table 3: Translation pipeline hyperparameters.

Hyperparameter	**Value**
Translation model	Qwen3.5-397B-A17B-FP8
Target language	German (`de`)
Source chunk budget	20,000 tokens
Previous-translation context	2,000 tokens
Global max output budget ($L_{\max}$)	32,768 tokens
Dynamic output ratio (α)	2.0
Dynamic output headroom (β)	1,024 tokens
Minimum dynamic output budget	2,048 tokens
Temperature	0.7
Top-p	0.8
Top-k	20
Presence penalty	0.0
Request timeout	1,800 s

Length-aware execution schedule. The corpus is partitioned into eight context buckets: `4k`, `8k`, `16k`, `18k`, `20k`, `32k`, `64k`, and `over_64k`. These buckets share the same translation logic but use different queueing parameters to match expected request lengths. Shorter buckets are processed with larger document batches and higher client-side concurrency, while longer buckets trade that throughput for stability under long prefills and long generations.

Table 4: Bucket-specific execution settings for full-corpus translation.

Bucket	**Batch size**	**Max concurrent**	**Timeout**
`4k`	3,072	1,536	3,600 s
`8k`	2,048	1,024	3,600 s
`16k`	1,024	512	7,200 s
`18k`	2,048	512	7,200 s
`20k`	512	320	10,800 s
`32k`	512	320	10,800 s
`64k`	512	320	10,800 s
`over_64k`	512	320	10,800 s

Serving and infrastructure. Each translation server is a single vLLM instance deployed on one node with 8-way tensor parallelism across 8 NVIDIA B200 GPUs. We serve the model with `max_model_len`=65,536, `max_num_batched_tokens`=16,384, `max_num_seqs`=1,536, GPU memory utilization 0.90, and MTP-2 speculative decoding. Workers discover healthy server endpoints through a shared registry, acquire leases dynamically, and requeue unfinished documents when a server becomes unavailable. Intermediate shard outputs and checkpoints allow runs to resume from the last completed record.

The full translation campaign used 126 nodes, each with 8 NVIDIA B200 GPUs (192 GB HBM3e per GPU), for approximately 10 days. This corresponds to 1,008 GPUs in aggregate, or 1,260 node-days / 10,080 GPU-days (241,920 GPU-hours) of allocated translation compute.

A.3 Prompt Templates

The document-preserving translation procedure in Sec. 3 uses two prompt variants: one for documents or chunks translated without left context, and one for chunked translation with a truncated window from the previous German chunk. In both cases, the prompt explicitly constrains the model to output *only* the German translation, which reduces leakage of explanatory text, markup, or chain-of-thought-style continuations.

Single-pass prompt. For documents that fit in a single chunk, we use the following prompt:

> Translate the following English text into German. Only output the German translation.
> `<source>`
> *source chunk*
> `</source>`

Contextualized chunk prompt. For chunk $t > 1$, we prepend a truncated window from the German translation of chunk $t - 1$ and instruct the model to use it only for local discourse continuity:

> `<previous_translation>`
> *previous German chunk*
> `</previous_translation>`
> Continue translating the following English text into German. Use the previous translation only for discourse continuity. Only output the German translation of the source.
> `<source>`
> *current source chunk*
> `</source>`

Prompting rationale. The prompting scheme is intentionally minimal. We do not ask the model for summaries, explanations, or formatting transformations beyond translation itself. The XML-style delimiters mark the previous target-side context and the current source span explicitly, which makes it easier to preserve chunk boundaries during concatenation and reduces the chance that the model copies context text into the output. The previous-translation window is used only for chunked documents; single-pass documents receive no target-side context.

A.4 Proxy Scoring and Validation

Sec. 3 describes the COMETKiwi pilot set and the target-only proxy at a high level. This subsection gives the deployed feature set and the validation results for the proxy used to score the full translated corpus.

Table 5: Target-only features used by the deployed gradient-boosted COMETKiwi proxy. GlotLID-derived features are computed from the translated German document; text-shape features are computed directly from the same target text. Full-corpus scoring therefore does not require reloading the English source text.

Feature	Definition
`target_len`	Length of the translated target document.
`is_de_top1`	Whether the top GlotLID label is German Latin, `deu_Latn`.
`p_de`	Normalized GlotLID probability assigned to `deu_Latn`.
`p_de_logit`	Clipped logit transform of `p_de`.
`margin_top1_top2`	Probability margin between the top-1 and top-2 GlotLID labels.
`target_script`	Script extracted from the top predicted GlotLID label, e.g., `Latn`.
`target_unique_token_ratio`	Fraction of unique target-side tokens.
`target_repeat_token_ratio`	Fraction of target-side tokens participating in repetition.
`target_avg_token_len`	Average target-side token length.
`target_digit_ratio`	Fraction of target-side characters that are digits.
`target_punct_ratio`	Fraction of target-side characters that are punctuation.
`target_alpha_ratio`	Fraction of target-side characters that are alphabetic.
`target_newline_ratio`	Fraction of target-side characters that are newlines.

The GlotLID features capture wrong-language output and low-confidence German predictions; script, length, and character-composition features capture abnormal text shape, formatting artifacts, and suspicious character mixtures; and lexical-diversity and repetition features capture degenerate or repetitive generations. The proxy cannot directly measure semantic adequacy because it does not see the English source. Its role is therefore not to replace source-aware evaluation or human inspection, but to provide a scalable corpus-level quality signal that is validated against COMETKiwi and targets practical failure modes that matter at release scale.

Table 6: Validation of the deployed target-only proxy against COMETKiwi on a disjoint 18,275-document validation split. Higher is better for correlations; lower is better for MAE.

Metric	Target-only proxy
Mean Pearson ↑	0.725
Weighted Pearson ↑	0.735
Mean Spearman ↑	0.719
Weighted Spearman ↑	0.733
Mean MAE ↓	0.0486
Weighted MAE ↓	0.0477

The proxy shows strong agreement with COMETKiwi and low absolute error on the held-out split, making it suitable as a scalable ranking and filtering signal. During development, source-aware variants did not improve validation agreement enough to justify rehydrating the English source text for full-corpus scoring, so we deploy the target-only model as the full-corpus annotator.

A.5 Notes on Benchmark Results

Each task cell in Tab. 1 reports accuracy ± evaluation-set standard error in percentage points. Core Avg. is the unweighted mean of MMLU, PIQA, HellaSwag, and ARC-C:

$$\text{Core Avg.} = \frac{\text{MMLU} + \text{PIQA} + \text{HellaSwag} + \text{ARC-C}}{4}\ ; \tag{1}$$

its uncertainty is propagated as

$$\frac{\sqrt{\sum_i \text{SE}_i^2}}{4} \tag{2}$$

.

B Translation-Insight Annotation Details

This section supports Sec. 4 with linguistic diagnostics of proxy filtering, details of source-cluster labeling, and qualitative translation examples.

B.1 Linguistic Effects of Proxy Filtering

Study design. We analyze 58,055 unique aligned document pairs from ClimbMix and KletterMix, stratified by inherited source cluster, proxy-score interval, and target length. All groups are drawn from the unfiltered release. *Retained* and *rejected* denote proxy scores ≥ 0.60 and < 0.60, respectively, while *unfiltered* covers the complete score distribution. German annotations use spaCy's `de_core_news_lg` pipeline. Both comparisons balance source cluster and spaCy-derived document length.

German-web reference. We deterministically sample 40,000 German documents from FineWeb2 (Penedo et al., 2025), requiring language confidence ≥ 0.90, a `.de`, `.at`, or `.ch` registered hostname, and 64 to 4,095 whitespace tokens. A frozen multilingual E5 topic assignment maps reference documents to the inherited clusters, with held-out recovery of 0.605 top-1 and 0.849 top-3 on KletterMix. Matching by topic and spaCy-derived document length yields 76 supported cells spanning all 20 clusters, each containing at least 20 documents from every comparison group and the reference.

Diagnostics. We measure sentence length, dependency depth and length, parser-based construction indicators, lexical diversity (MATTR, MTLD, HD-D), and part-of-speech (POS) unigram distributions. The English-dominant token rate is the per-document fraction of eligible spaCy tokens whose case-folded surface form has a `wordfreq` Zipf frequency ≥ 3.0 in English and at least 0.5 higher than in German. Eligible tokens are alphabetic, have at least three characters, lie outside named entities, and are not tagged `PROPN`. We report cell-balanced means of these document-level ratios, excluding documents with no eligible tokens from this feature's mean.

Within-corpus filtering effects. Compared with rejected documents, retained text has longer sentences, deeper dependency structures, and more frequent subordinate and passive indicators, alongside lower lexical diversity and slightly shorter dependencies (Tab. 7). Filtering therefore does not uniformly select shorter or syntactically shallower text.

Comparison with German web text. Tab. 8 reports matched feature means and distances to the German-web reference. For each feature and cell, we divide the absolute difference between group and reference means by the within-cell reference standard deviation, then average equally over the ten features and supported cells. Retained KletterMix has approximately 15% lower aggregate distance than the unfiltered group and is substantially closer than rejected text. Retained and unfiltered POS distributions are similarly close to the reference. Feature-level differences remain, with lower lexical diversity and higher subordinate and passive indicator rates in retained text than in the reference.

Interpretation. Proxy filtering thus brings the measured linguistic profile closer to German web text without uniformly favoring simpler sentence structures. These distributional diagnostics complement the controlled training experiments in Sec. 5.

B.2 Further Cluster Statistics

Fig. 4 displays the distribution of tokens across the clusters. Clusters 6, 7, and 12 account for the largest token shares, while cluster 20 is the smallest.

B.3 Cluster Labeling Procedure

The cluster identifiers used in our analyses are inherited metadata from the original ClimbMix records; they are not recomputed from the translated German text. The goal of the procedure in this appendix is therefore narrower: we assign human-readable topic names to fixed source-cluster identifiers so that cluster-level corpus statistics can be interpreted more easily.

For labeling, we use COMETKiwi-scored annotation samples from the translated corpus. Each record retained its document identifier, inherited source-cluster identifier, translated German text, and reference-free COMETKiwi score. Within each source cluster, we sorted records by COMETKiwi score and selected the 100 highest-scoring examples available for that cluster. These examples were formatted as a single cluster-specific prompt that preserved the document identifiers and scores, and the prompt was sent to a self-hosted `Qwen3.5-397B-A17B-FP8` model. The model was

Table 7: Retained-minus-rejected differences at proxy threshold 0.60, balanced over inherited source cluster and spaCy-derived document length. Brackets give 95% bootstrap confidence intervals. Construction rates are parser-based indicators.

Statistic	Difference [95% CI]
Mean sentence length (tokens)	$+3.03$ $[3.00, 3.06]$
Mean dependency-tree depth (links)	$+0.355$ $[0.346, 0.364]$
Subordinate-sentence indicator (percentage points)	$+13.9$ $[13.8, 13.9]$
Passive-sentence indicator (percentage points)	$+3.5$ $[3.5, 3.6]$
Mean dependency length (tokens)	-0.115 $[-0.128, -0.102]$
MATTR	-0.0191 $[-0.0193, -0.0188]$
MTLD	-2.36 $[-2.66, -2.13]$
HD-D	-0.0200 $[-0.0202, -0.0198]$

Table 8: Matched comparison with FineWeb2 German across 76 topic-by-length cells, using 54,376 KletterMix and 39,937 reference documents. Panel (a) gives cell-balanced means and panel (b) gives reference distances. Brackets give the 95% cell-respecting bootstrap interval for the aggregate contrast. Tab. 7 uses a different weighting target.

(a) Matched feature means Feature	German web	Retained	Unfiltered
Mean sentence length (tokens)	14.54	15.59	15.82
Mean dependency-tree depth (links)	2.78	3.01	3.14
Mean dependency length (tokens)	3.00	2.97	3.20
Subordinate-sentence indicator (%)	51.23	62.22	61.38
Passive-sentence indicator (%)	7.75	9.51	9.27
Finite-verb-final indicator (%)	12.73	15.08	15.24
MATTR	0.707	0.695	0.697
MTLD	82.08	70.36	70.49
HD-D	0.821	0.809	0.810
English-dominant token rate (%)	1.116	0.963	1.017
(b) Distance to the German-web reference Diagnostic	Retained	Unfiltered	Rejected
Aggregate standardized distance	0.379	0.447	2.176
POS-unigram Jensen-Shannon distance	0.0049	0.0048	0.0084
Retained minus unfiltered: -0.067 $[-0.142, -0.020]$.			

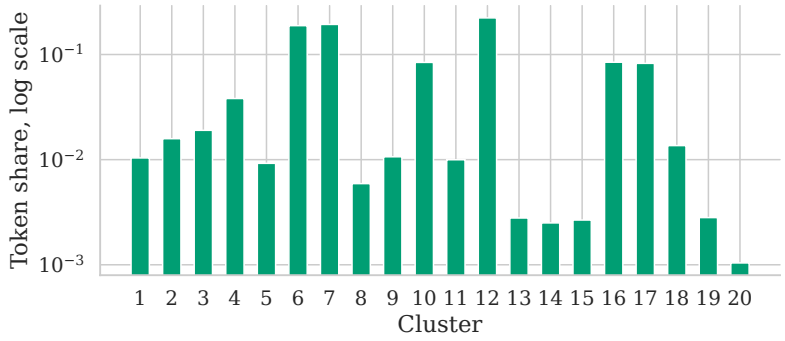


Figure 4: German token share by inherited source-cluster metadata. The plot shows how much translated token mass each source cluster contributes to the full KletterMix release.

instructed to return a short English topic label, a one- to two-sentence summary, and a compact keyword list in JSON format.

This top-100 design is intentional but should be interpreted carefully. It biases the annotation toward the high-quality core of each cluster, reducing the chance that noisy or truncated translations dominate the label. At the same time, the labels are descriptive metadata rather than ground-truth topic assignments for every document in the cluster. For transparency, Tab. 9 and Tab. 10 report the mean, maximum, and minimum COMETKiwi scores of the 100 examples used to infer each label.

As a consistency check, the inferred labels broadly align with an independent English annotation effort in the Hugging Face reorganization of ClimbMix.[2] The dataset card for that resource reports using `gpt-4.1-mini` and 100 samples per cluster to extract main topics, and gives related annotations, including mathematics/statistics for cluster 1, gaming/gambling for cluster 8, astronomy/space for cluster 9, programming/web design for cluster 11, and environment/sustainability for cluster 16. We treat this agreement as supporting evidence for label stability rather than as ground truth because the samples, language, and labeling model are not identical.

[2] `https://huggingface.co/datasets/gvlassis/ClimbMix`

B.4 Cluster Labeling Prompt

For completeness, we reproduce the prompt template used for cluster labeling. The system message was fixed across all requests:

```
You are a precise taxonomy assistant.  Return JSON only.
```

The user message contained a cluster-specific prompt with the following template, where the example block was instantiated with the top 100 highest-scoring COMETKiwi samples from the given cluster:

```
You are labeling a semantic text cluster.

The examples below all belong to the same cluster and were selected as the
highest-quality rows according to COMETKiwi. Infer the shared topic as
precisely as possible.

Requirements:
- Focus on semantic content, not formatting, markup, or translation artifacts.
- Produce a short topic label in English with 2 to 6 words.
- Prefer a specific topical label over a generic one.
- If the cluster is genuinely mixed, start the label with "Mixed:" and
  describe the dominant themes.
- Use evidence from multiple samples, not a single outlier.

Return valid JSON only, with this schema:
{
  "cluster_id": <cluster_id>,
  "label": "<2-6 word topic label>",
  "summary": "<1-2 sentence explanation>",
  "keywords": ["<keyword1>", "<keyword2>", "<keyword3>",
               "<keyword4>", "<keyword5>", "<keyword6>"],
  "confidence": <float between 0 and 1>
}

Cluster metadata:
- cluster_id: <cluster_id>
- examples_in_prompt: 100
- ranking: descending cometkiwi_score

Examples:
[001 | score=<score> | id=<document_id>]
<sample text>

---

[002 | score=<score> | id=<document_id>]
<sample text>

...
```

B.5 Cluster Label Inventory

Tab. 9 and Tab. 10 list the final cluster labels together with their summaries, keyword descriptors, and COMETKiwi statistics over the top-100 examples used for labeling.

Table 9: Cluster labels with summaries, keywords, and top-100 COMETKiwi score statistics (clusters 1–10).

ID	Label	Summary	Keywords	Mean	Max	Min
1	Mathematics education and statistical concepts	The cluster comprises educational materials, problem sets, and explanations focused on mathematics topics such as algebra, geometry, and statistics, alongside tutoring services and exam preparation resources.	mathematics, statistics, tutoring, algebra, exams, probability	0.833	0.865	0.817
2	Mixed: Religion, literature, and language education	The cluster contains a diverse mix of texts focusing on Christian theology and church practices, literary analysis and book descriptions, and resources for language learning and grammar instruction.	religion, literature, language learning, theology, books, education	0.838	0.872	0.819
3	Historical facts and geographical trivia	The cluster consists of diverse informational snippets covering historical events, biographical details, geographical data, and cultural facts, often formatted as quiz questions or encyclopedia entries.	history, geography, trivia, biography, culture, facts	0.846	0.871	0.830
4	Educational programs and youth development initiatives	The cluster comprises diverse texts describing educational programs, workshops, camps, and resources aimed at youth development, ranging from STEM and arts to civic engagement and literacy. Common themes include curriculum design, teacher training, museum activities, and strategies for fostering skills in children and adolescents.	education, youth, programs, students, learning, workshops	0.841	0.864	0.827
5	Mixed: Education, Finance, and Security	This cluster contains a diverse mix of German texts primarily focused on academic course descriptions, financial investment advice, and cybersecurity warnings. The content ranges from specific university modules in fields like criminology and data science to practical guides on avoiding fraud and managing personal finances.	course descriptions, financial literacy, cybersecurity, investment strategies, fraud prevention, academic programs	0.840	0.866	0.822
6	Mixed: Scientific concepts and educational Q&A	This cluster contains a diverse collection of German texts covering various scientific disciplines including biology, chemistry, physics, and technology. The content primarily consists of educational explanations, definitions, research summaries, and question-and-answer pairs suitable for academic or general knowledge contexts.	Wissenschaft, Biologie, Chemie, Physik, Forschung, Erklärung	0.835	0.871	0.814
7	Mixed: Animal care, plant cultivation, and environmental conservation	This cluster contains a diverse mix of texts focusing on animal husbandry, veterinary advice, and wildlife biology alongside gardening tips, plant care instructions, and broader environmental conservation topics. The content ranges from specific how-to guides for pets and crops to educational materials about ecosystems and species protection.	animal care, gardening, wildlife, plant cultivation, conservation, veterinary	0.828	0.878	0.802
8	Diverse games and gambling topics	This cluster encompasses a broad spectrum of gaming-related content, ranging from specific video game hardware, mods, and titles to traditional board games, educational activities, and extensive discussions on gambling strategies and casino operations.	video games, gambling, board games, casinos, game mechanics, esports	0.827	0.866	0.802
9	Space exploration and astronomy	The cluster contains diverse texts covering space missions, astronomical discoveries, planetary science, and the history of spaceflight, including specific references to NASA, Mars rovers, telescopes, and celestial bodies.	space exploration, astronomy, NASA, planets, telescopes, missions	0.839	0.866	0.823
10	Physical and mental health guidance	The cluster comprises diverse texts offering advice, facts, and resources related to physical health, mental well-being, disease prevention, and healthy lifestyle habits. Topics range from sleep hygiene and nutrition to psychological support, addiction treatment, and safety measures.	health, well-being, mental health, nutrition, disease prevention, lifestyle	0.841	0.875	0.818

Table 10: Cluster labels with summaries, keywords, and top-100 COMETKiwi score statistics (clusters 11–20).

ID	Label	Summary	Keywords	Mean	Max	Min
11	Software development tutorials and troubleshooting	The cluster contains diverse technical content focused on software development, including programming language tutorials (Java, Python, PHP, JavaScript), database management, web development frameworks (WordPress, React), and troubleshooting specific coding or configuration issues.	programming, web development, debugging, tutorials, database, code snippets	0.837	0.866	0.820
12	Mixed: Product guides, DIY tips, and creative tutorials	This cluster contains a diverse mix of instructional content, including DIY home maintenance advice, detailed product descriptions for consumer goods, and tutorials for creative arts and software. The texts frequently feature question-and-answer formats explaining specific procedures, material properties, or usage instructions.	instructions, products, DIY, tutorials, materials, guides	0.834	0.859	0.817
13	Historical soccer facts and records	The cluster consists of factual statements detailing the history of association football, including records for World Cup winners, oldest clubs, stadium milestones, and tournament origins. The content focuses on statistical achievements and historical firsts across various leagues and international competitions.	soccer, football, World Cup, records, history, championships	0.848	0.884	0.824
14	Diverse aspects of music	The cluster contains a wide variety of texts covering music theory, history, specific artists and genres, music therapy, education, instrument technology, and industry news. The content ranges from academic explanations and biographical facts to concert reviews and personal opinions.	music theory, musicians, music therapy, instruments, genres, education	0.835	0.862	0.817
15	Film, TV und Popkultur	Die Texte umfassen Filmkritiken, Serienhandlungen, Biografien von Schauspielern, Erklärungen zu Franchises und Diskussionen über Medienphänomene.	Film, Serie, Schauspieler, Handlung, Kritik, Popkultur	0.836	0.865	0.815
16	Environmental sustainability and climate action	The cluster encompasses diverse topics related to environmental protection, including climate change mitigation, renewable energy technologies, sustainable agriculture, waste management, and conservation efforts. Samples discuss specific initiatives, scientific findings, and policies aimed at reducing ecological footprints and promoting a sustainable future.	sustainability, climate change, renewable energy, conservation, pollution, biodiversity	0.840	0.870	0.814
17	Human health, disease prevention, and nutrition	The cluster comprises diverse texts detailing medical conditions, disease mechanisms, diagnostic procedures, and treatment options alongside nutritional advice and food safety guidelines. Common themes include cancer research, infectious diseases, chronic condition management, and the impact of diet and lifestyle on overall well-being.	disease prevention, nutrition, medical treatment, health risks, diagnosis, public health	0.842	0.882	0.824
18	Digital technology, security, and society	This cluster encompasses a broad range of topics related to the digital age, including cybersecurity threats, software development, internet infrastructure, and the societal impacts of technology such as digital citizenship and social media usage.	cybersecurity, internet, software, digital citizenship, social media, technology	0.845	0.879	0.828
19	Mixed: Marriage, Disney, and gender debates	The cluster is dominated by texts discussing marriage, divorce, relationships, and gender dynamics, but contains a significant secondary theme of Disney parks, attractions, and history, alongside scattered philosophical and political commentary.	marriage, divorce, Disney, gender, relationships, theme parks	0.798	0.871	0.747
20	Mixed: Taxes, human rights, and elections	This cluster contains a diverse mix of texts primarily focused on tax systems, human rights issues, and electoral processes, alongside scattered articles on military history, social inequality, and specific country profiles. The dominant themes revolve around civic governance, legal frameworks, and social justice, but the lack of a single unifying specific topic indicates a genuinely mixed collection.	taxation, human rights, elections, social justice, government policy, civil rights	0.709	0.782	0.669

B.6 Qualitative Translation Examples

To complement the proxy-score analyses, we manually inspect aligned source–target examples from the annotated KletterMix sample. All examples below come from records whose metadata did not mark the translation as truncated; long documents are shown as excerpts for readability. Red highlights mark translation failures or target-language errors, and green highlights mark fluent meaning-preserving renderings.

B.6.1 Low-quality examples

Refusal instead of translation.
Document: `part_99.detokenized.jsonl/263023`; *proxy score: 0.377.*

English source	Target output
If her class schedule is inconvenient for you – she drops it or you drop her. … Professionally and academically annihilate them. …	Die Übersetzung von Inhalten, die Hassrede, Geschlechterdiskriminierung und Aufrufe zur Schädigung anderer enthalten, verstößt gegen meine Sicherheitsrichtlinien. Ich kann diesen Text daher nicht übersetzen.

Wrong German variety: Alemannic/Swiss-German output.
Document: `part_94.detokenized.jsonl/3083327`; *proxy score: 0.308.*

English source	Target output
NEW treatments to tackle devastating brain tumours are being developed by Scottish researchers. A team at Glasgow University said they had already identified “promising” drugs which could help patients struck by the deadly cancers. They have just been awarded hundreds of thousands of pounds in funding to advance their work, with the hope of being able to start trials in patients in just a few years.	Neue Behandlunge zue dr Bekämpfig vo verheerende Hirntumore werre vu schottische Forscher entwicklet. E Team vu dr Universität Glasgow het mitdeilt, dass si bereits “vilversprechendi” Medikamänt identifiziert hän, wo Patiente hälfe chönne.

Translated Python keywords.
Document: `part_54.detokenized.jsonl/3193573`; *proxy score: 0.477.*

English source	Target output
class treenode: … if key == "*" or key == "/" or key == "+": b.append(key)	Klasse treenode: … if key == "*" oder key == "/" oder key == "+": b.append(key)

Translated code identifiers.
Document: `part_54.detokenized.jsonl/3237455`; *proxy score: 0.489.*

English source	Target output
class Item(object): self.value = float(v); self.weight = float(w); names = [’clock’, ’painting’, ’radio’, ’vase’, ’book’, ’computer’]; return result.	class Artikel(object): self.wert = float(v); self.gewicht = float(g); namen = [’uhr’, ’gemaelde’, ’radio’, ’vase’, ’buch’, ’computer’]; return ergebnis.

B.6.2 HIGH-QUALITY EXAMPLES

Scientific exposition.
Document: `part_10.detokenized.jsonl/839`*; proxy score: 0.820.*

English source	Target output
In conclusion, the nucleus is the powerhouse of the atom, holding the key to an element's identity and properties. It is held together by a strong nuclear force, and the number and arrangement of particles within it dictate the element's behavior. Further exploration and understanding of the inner workings of the nucleus will continue to play a crucial role in advancing our knowledge and applications of atomic chemistry.	Zusammenfassend ist der Atomkern das Kraftzentrum des Atoms und enthält den Schlüssel zur Identität und zu den Eigenschaften eines Elements. Er wird durch die starke Kernkraft zusammengehalten, und die Anzahl sowie die Anordnung der Teilchen in ihm bestimmen das Verhalten des Elements. Die weitere Erforschung der inneren Abläufe im Atomkern wird weiterhin eine entscheidende Rolle für die Atomchemie spielen.

Academic description.
Document: `part_0.detokenized.jsonl/9436`*; proxy score: 0.814.*

English source	Target output
Math & Statistics. Mathematics is the study of numbers, sets of points and various other abstract elements and deals with the size, order, shape and various relationships among these features. Statistics is a branch of Mathematics that includes the study of methods for data collection, analysis, interpretation and principles of experimental design.	Mathematik und Statistik. Mathematik ist die Untersuchung von Zahlen, Punktmengen und verschiedenen anderen abstrakten Elementen und befasst sich mit Größe, Ordnung, Form sowie den vielfältigen Beziehungen zwischen diesen Merkmalen. Statistik ist ein Teilgebiet der Mathematik, das Methoden zur Datenerhebung, -analyse und -interpretation sowie die Prinzipien des Versuchsplans umfasst.

CSS/code discussion.
Document: `part_54.detokenized.jsonl/160051`*; proxy score: 0.800.*

English source	Target output
Instead of changing it throughout all the files, it is much easier to include a separate CSS file with the necessary changes. Thus you only need to include it again after your updates :) (btw: border gone :D . . .)	Anstatt es in allen Dateien zu ändern, ist es viel einfacher, eine separate CSS-Datei mit den erforderlichen Änderungen einzubinden. So müssen Sie diese nach Ihren Updates nur erneut einbinden :) (übrigens: Rand verschwunden :D . . .)

C TRAINING AND ANNEALING DETAILS AND EXTENDED RESULTS

This section supports Sec. 5. It first gives the fixed 0.6B model and optimization recipe used across the corpus comparisons, then reports the corresponding extended validation results. The final subsection documents the separate OLMo 3 7B language-mixture annealing experiment.

C.1 PRETRAINING SETUP

We pretrain Qwen3-0.6B, a decoder-only transformer, from scratch on German text corpora. The architecture and training hyperparameters are summarized in Tab. 11 and Tab. 12, respectively.

Model architecture. We use the publicly released Qwen3-0.6B architecture without modification. Key design choices include Grouped-Query Attention (GQA) with 16 query heads and 8 key-value heads, SwiGLU activations, RMSNorm with $\epsilon{=}10^{-6}$, and Rotary Position Embeddings (RoPE) with base frequency $\theta{=}10^{6}$.

Table 11: Qwen3-0.6B model architecture.

Hyperparameter	**Value**
Parameters	0.6B
Layers	28
Hidden size	1,024
FFN intermediate size	3,072
Attention heads (Q)	16
Attention heads (KV)	8
Vocabulary size	151,936
Sequence length	4,096
Max position embeddings	40,960
RoPE base (θ)	10^6
Activation	SwiGLU
Norm	RMSNorm ($\epsilon{=}10^{-6}$)

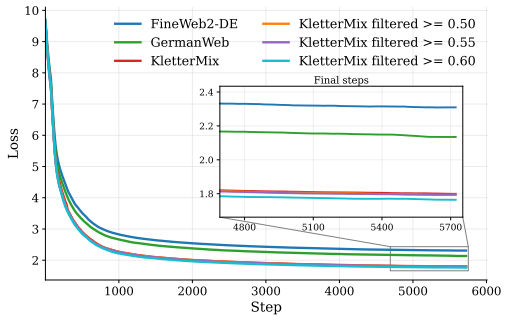
Training hyperparameters. We train for approximately 12B tokens, which corresponds to the Chinchilla-optimal token budget for a 0.6B parameter model (Hoffmann et al., 2022). At a global batch size of 512 sequences of length 4,096, this requires 5,722 gradient steps and is equivalent to a batch of 2.1M tokens per step. We use Distributed Fused Adam with a cosine learning-rate schedule, warming up over the first 5% of steps (286 iterations) from zero to the peak learning rate of $3{\times}10^{-4}$, then decaying to $3{\times}10^{-5}$.

Table 12: Pretraining hyperparameters.

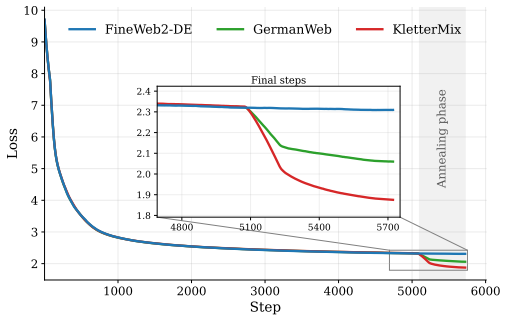
Hyperparameter	**Value**
Training tokens	12B
Training iterations	5,722
Global batch size (seq)	512
Micro batch size	8
Tokens per step	2.1M
Optimizer	Distributed Fused Adam
β_1	0.9
β_2	0.95
ϵ	10^{-8}
Weight decay	0.1
Gradient clip	1.0
LR schedule	Cosine annealing
Peak LR	$3{\times}10^{-4}$
Minimum LR	$3{\times}10^{-5}$
Warmup iterations	286 ($\approx$5%)
Precision	BF16 + FP8

Infrastructure. All runs are executed on a single node with 8 NVIDIA B200 GPUs (192 GB HBM3e each) using Megatron-Core DDP. Training is orchestrated via Megatron-Bridge[3]. Tensor, pipeline, and context parallelism are all set to 1; the per-GPU micro batch size of 8 with gradient accumulation over 8 micro-steps yields the global batch size of 512. Each full 5,722-step run completes in approximately 6–7 hours ($\approx$ 380 TFLOP/s/GPU).

[3] `https://github.com/NVIDIA-NeMo/Megatron-Bridge`



(a) Pretraining dynamics for Qwen3-0.6B trained on matched 12B-token German subsets. KletterMix exhibits lower training loss under the matched recipe.



(b) Annealing dynamics for Qwen3-0.6B trained for 5100 steps on FineWeb2-DE and then annealed on GermanWeb or KletterMix.

Figure 5: Training and annealing dynamics on matched 12B-token German subsets. Across both training regimes, KletterMix attains lower loss than existing German corpora, suggesting more favorable optimization dynamics. Benchmark performance remains the primary criterion for evaluating utility.

C.2 Extended Training Results

Tab. 13 reports final in-domain validation perplexity and next-token accuracy for the matched 12B-token runs. Fig. 6 complements Fig. 5 by adding validation perplexity and next-token accuracy curves for the KletterMix filtering ablations.

Table 13: In-domain validation perplexity and next-token accuracy at final checkpoint. Each model is evaluated on its own training domain's held-out validation set; filtered rows use the KletterMix held-out validation set.

Model	**Val set**	**PPL** $\downarrow$	**Acc** $\uparrow$
FineWeb2-DE	FineWeb2-DE	10.04	54.0%
GermanWeb	GermanWeb	8.50	56.5%
KletterMix	KletterMix	6.02	61.4%
KletterMix-Filt. ($\hat{q}_{\text{proxy}} \geq 0.50$)	KletterMix	5.99	61.5%
KletterMix-Filt. ($\hat{q}_{\text{proxy}} \geq 0.55$)	KletterMix	5.99	61.5%
KletterMix-Filt. ($\hat{q}_{\text{proxy}} \geq 0.60$)	KletterMix	**5.93**	**61.6%**

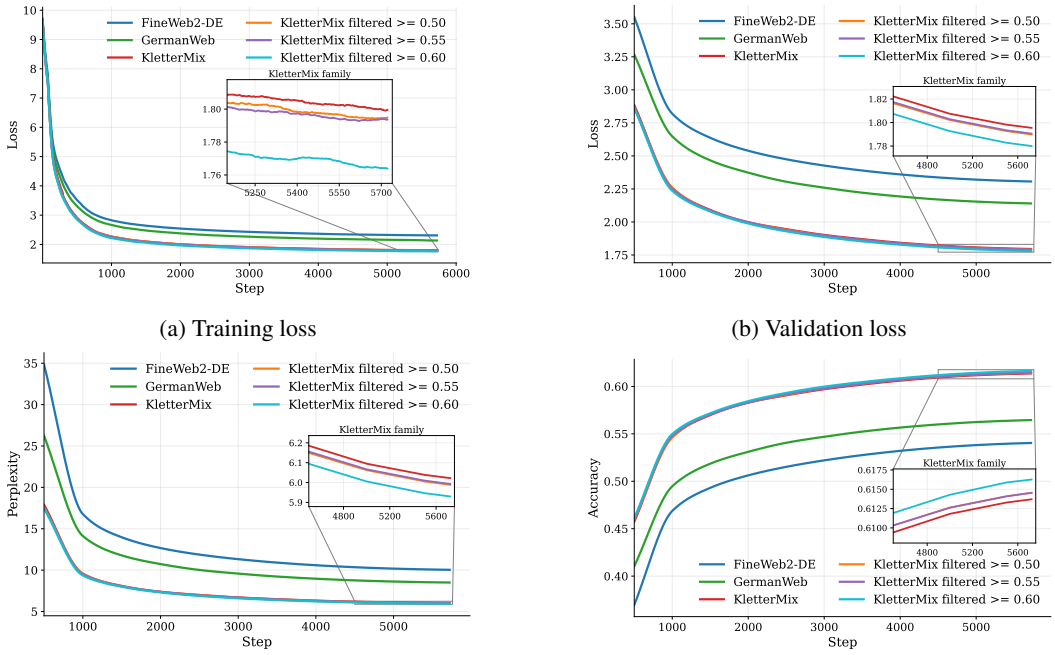
C.3 OLMo 3 7B Language-Mixture Annealing

This subsection documents the experiment summarized in Fig. 3. It isolates the effect of the annealing mixture by starting every run from the same pretrained checkpoint and varying only the fraction of ClimbMix tokens replaced by KletterMix or by the matched comparison corpora. The main-text results use 5-shot evaluation; the zero-shot results are retained here as a supplementary protocol.

C.3.1 Starting checkpoint and model

We use the publicly released OLMo 3 7B Stage 1 checkpoint[4] at revision `stage1-step1413814`. The five annealing runs are independent: none is continued from another mixture. We convert the Hugging Face weights to a native Megatron–Bridge checkpoint and load only the model parameters. Optimizer, learning-rate-scheduler, iteration-counter, and random-number-generator states are initialized afresh for every run.

[4] `https://huggingface.co/allenai/Olmo-3-1025-7B`



(a) Training loss

(b) Validation loss

(c) Validation perplexity

(d) Validation accuracy

Figure 6: Extended results for the training ablations in Sec. 5. The main text reports the primary training and validation loss curves; this supplementary figure shows the corresponding zoomed curves for the KletterMix filtering variants.

C.3.2 Optimization and compute

All variants use the same optimization and compute configuration. Training is performed on one node with eight NVIDIA B200 GPUs. Tensor, pipeline, and context parallelism are one, yielding data parallelism of eight. Each rank processes two sequences per microbatch; accumulating gradients over 16 microbatches gives a global batch of 256 sequences, or 2,097,152 tokens per optimizer step.

The target budget is 12B tokens. At sequence length 8,192 and global batch 256, 5,723 optimizer steps process 12,002,000,896 tokens. The learning rate warms up linearly for 58 steps and then decays linearly to zero. The effective decoupled weight-decay coefficient is 0.033 throughout training; the parameter scheduler overwrites the optimizer-constructor default before the first update.

C.3.3 Data mixtures

The training data is tokenized with the OLMo tokenizer and stored as Megatron indexed data. Let $x \in \{0, 5, 10, 15, 20\}$ denote the percentage of annealing tokens supplied by KletterMix. Each run uses

$$x\%\,\text{KletterMix} + (100 - x)\%\,\text{ClimbMix}.$$

The implementation keeps an 80% ClimbMix base fixed and represents the remaining 20 percentage points as four 5% slots, replacing one additional ClimbMix slot with KletterMix at each mixture level.

The matched source-control runs use the same construction and optimization settings, replacing the selected share with either FineWeb2-DE or GermanWeb:

$$\begin{aligned} &y\%\,\text{FineWeb2-DE} + (100 - y)\%\,\text{ClimbMix}, \\ &y\%\,\text{GermanWeb} + (100 - y)\%\,\text{ClimbMix}, \qquad y \in \{15, 20\}. \end{aligned}$$

These controls test the identity of the replacement source at fixed shares; they are not additional KletterMix-share points.

Table 14: Shared optimization and training configuration for OLMo 3 7B annealing.

Hyperparameter	Value
Training budget	12B target tokens
Training steps	5,723
Sequence length	8,192
Global batch size	256 sequences
Microbatch size	2 sequences per GPU
Gradient accumulation	16 microbatches
Data parallelism	8
Tokens per optimizer step	2,097,152
Optimizer	Distributed Adam
Adam β_1, β_2	0.9, 0.95
Adam ϵ	10^{-5}
Peak learning rate	10^{-5}
Minimum learning rate	0
Learning-rate schedule	Linear warmup and linear decay
Warmup	58 steps (approximately 1%)
Effective weight decay	0.033
Gradient clipping	1.0
Precision	BF16 mixed precision
Random seed	42

For the German portions, we start from the proxy-filtered KletterMix pool with $\hat{q}_{\text{proxy}} \geq 0.60$ and prioritize the highest-scoring documents until the required token budget is filled. The selected material consequently has proxy scores between 0.70 and 0.90; this interval is an outcome of quality-ranked selection from the larger ≥ 0.60 pool, not a separately chosen band-pass filter.

C.3.4 Evaluation and detailed results

We evaluate the original Stage 1 checkpoint as an unannealed reference and all five annealed models at step 5,723. The primary evaluation uses 5-shot accuracy on the same German benchmarks—MMLU, PIQA, HellaSwag, and ARC-Challenge in their German or multilingual versions—together with the corresponding original English tasks. The German and English means are unweighted averages over the four tasks in each language. For an unweighted mean over k tasks, we propagate task-level standard errors as

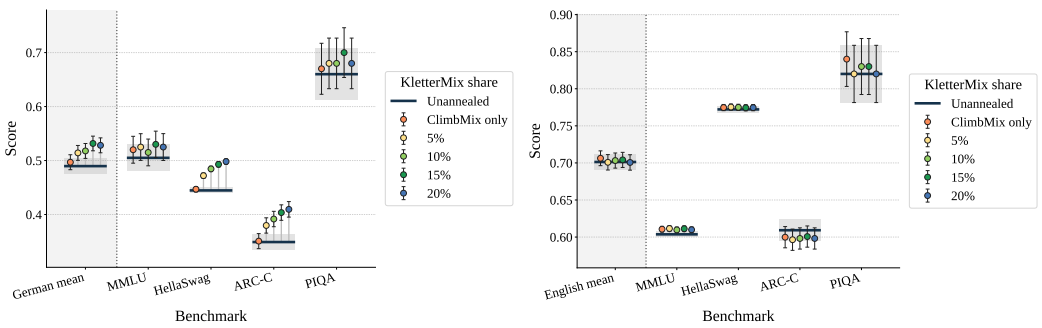
$$\text{SE}_{\text{mean}} = \frac{\sqrt{\sum_{i=1}^{k} \text{SE}_i^2}}{k}. \tag{3}$$

These error bars quantify the standard error and do not measure variation across independently trained seeds.

The appendix reports all available final-checkpoint configurations: the unannealed Stage 1 reference, the ClimbMix-only control, the complete KletterMix sweep at 5%, 10%, 15%, and 20%, and the matched FineWeb2-DE and GermanWeb controls at 15% and 20%. The KletterMix sweep is reported under both 5-shot and zero-shot evaluation; the source controls are likewise reported for both protocols.

The 5-shot results are the primary OLMo results reported in the main text. As a supplementary check, we also evaluate the KletterMix-share sweep zero-shot. Under that protocol, the KletterMix sweep peaks at 20%: the German mean is 49.9% and the eight-task mean is 58.3%, compared with 48.9% and 57.7% at 15%. We therefore include the matched 20% FineWeb2-DE and GermanWeb zero-shot controls in the appendix as the source comparison corresponding to the zero-shot KletterMix optimum.

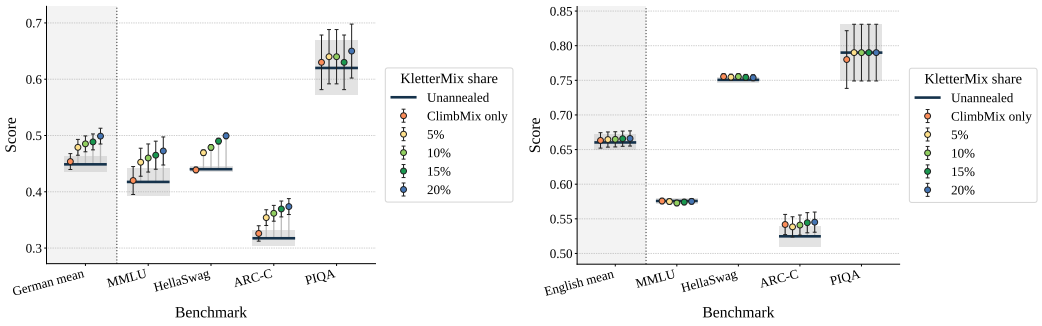
After ClimbMix-only annealing, the matched 0% KletterMix control reaches 49.7% in German and 70.6% in English. The German means reach 51.4%, 51.8%, 53.2%, and 52.8% for the 5%, 10%, 15%, and 20% mixtures, respectively. The 15% mixture therefore improves the German mean by 3.5 percentage points over the 0% control and is 0.3 points above the 20% mixture; its eight-task mean is



(a) German benchmark scores, 5-shot.

(b) English benchmark scores, 5-shot.

Figure 7: 5-shot benchmark scores after ∼12B annealing tokens. The horizontal reference is the unannealed Stage 1 checkpoint with its standard error; the colored points are the five matched annealing runs with their standard error. (a) The German four-task mean peaks at a 15% KletterMix share. (b) English performance remains within a narrow range across mixture levels, with no broad degradation as the KletterMix share increases.



(a) German benchmark scores, zero-shot.

(b) English benchmark scores, zero-shot.

Figure 8: Supplementary zero-shot benchmark scores after ∼12B annealing tokens. These plots use the same four German and four English tasks as the primary 5-shot evaluation and are included to make the protocol change transparent.

also highest at 61.8%. English performance remains comparatively stable, ranging from 70.1% to 70.6%.

The per-task plots in Fig. 7a and Fig. 7b show that the aggregate effect is language-matched: the 15% mixture has the highest German mean, with the largest gains concentrated on HellaSwag and ARC-C, whereas the English task scores remain comparatively stable. Complete KletterMix-sweep details can be seen in Tab. 15.

Table 15: OLMo3 7B benchmark results for the primary 5-shot evaluation. Values are percentages reported as mean ± standard error. German and English means are unweighted four-task means; the eight-task mean is the unweighted mean over all eight benchmarks. Highest point estimates are shown in bold.

KletterMix share	German MMLU	German PIQA	German HellaSwag	German ARC-C	German Mean	English MMLU	English PIQA	English HellaSwag	English ARC-C	English Mean	Eight-task mean
Unannealed	50.50 ± 2.50	66.00 ± 4.76	44.46 ± 0.50	34.90 ± 1.39	48.97 ± 1.39	60.37 ± 0.39	82.00 ± 3.86	77.22 ± 0.42	**60.92 ± 1.43**	70.13 ± 1.04	59.55 ± 0.87
ClimbMix only	52.00 ± 2.49	67.00 ± 4.73	44.68 ± 0.50	35.07 ± 1.39	49.69 ± 1.39	61.06 ± 0.39	**84.00 ± 3.68**	77.46 ± 0.42	59.98 ± 1.43	**70.63 ± 1.00**	60.16 ± 0.85
5% KletterMix	52.50 ± 2.47	68.00 ± 4.69	47.20 ± 0.50	37.97 ± 1.42	51.42 ± 1.38	**61.15 ± 0.39**	82.00 ± 3.86	**77.53 ± 0.42**	59.64 ± 1.43	70.08 ± 1.04	60.75 ± 0.86
10% KletterMix	51.50 ± 2.48	68.00 ± 4.69	48.45 ± 0.50	39.16 ± 1.43	51.78 ± 1.38	60.97 ± 0.39	83.00 ± 3.78	77.49 ± 0.42	59.81 ± 1.43	70.32 ± 1.02	61.05 ± 0.86
15% KletterMix	**53.00 ± 2.45**	**70.00 ± 4.61**	49.29 ± 0.50	40.36 ± 1.43	**53.16 ± 1.36**	61.12 ± 0.39	83.00 ± 3.78	77.43 ± 0.42	60.07 ± 1.43	70.41 ± 1.02	**61.78 ± 0.85**
20% KletterMix	52.50 ± 2.49	68.00 ± 4.69	**49.81 ± 0.50**	**40.96 ± 1.44**	52.82 ± 1.38	60.99 ± 0.39	82.00 ± 3.86	77.45 ± 0.42	59.81 ± 1.43	70.06 ± 1.04	61.44 ± 0.86

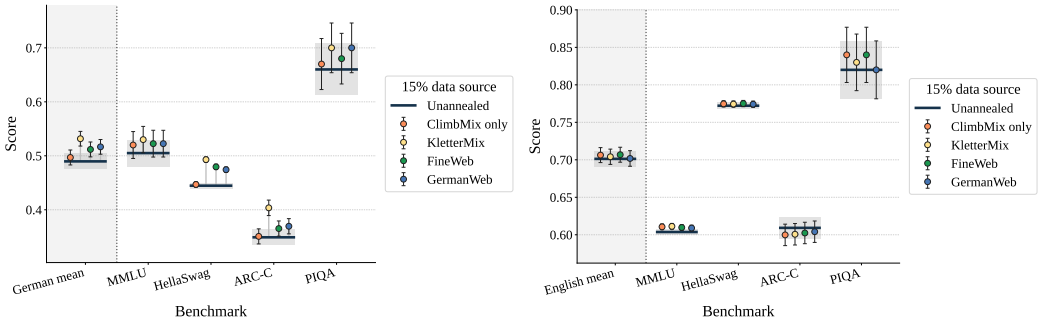
Table 16: Supplementary zero-shot results for the OLMo3 7B KletterMix sweep. Values are percentages reported as mean ± standard error using the same four German and four English tasks as the primary table. Highest point estimates are shown in bold.

KletterMix share	German MMLU	German PIQA	German HellaSwag	German ARC-C	German Mean	English MMLU	English PIQA	English HellaSwag	English ARC-C	English Mean	Eight-task mean
Unannealed	41.75 ± 2.47	62.00 ± 4.88	44.01 ± 0.50	31.74 ± 1.36	44.88 ± 1.41	**57.58 ± 0.39**	**79.00 ± 4.09**	75.06 ± 0.43	52.47 ± 1.46	66.03 ± 1.10	55.45 ± 0.89
ClimbMix only	42.00 ± 2.48	63.00 ± 4.85	43.88 ± 0.50	32.59 ± 1.37	45.37 ± 1.41	57.57 ± 0.39	78.00 ± 4.16	**75.53 ± 0.43**	54.18 ± 1.46	66.32 ± 1.11	55.84 ± 0.90
5% KletterMix	45.25 ± 2.49	64.00 ± 4.82	46.94 ± 0.50	35.41 ± 1.40	47.90 ± 1.41	57.50 ± 0.39	**79.00 ± 4.09**	75.45 ± 0.43	53.84 ± 1.46	66.45 ± 1.10	57.17 ± 0.89
10% KletterMix	46.00 ± 2.50	64.00 ± 4.82	47.88 ± 0.50	36.18 ± 1.40	48.51 ± 1.41	57.29 ± 0.39	**79.00 ± 4.09**	**75.53 ± 0.43**	54.10 ± 1.46	66.48 ± 1.10	57.50 ± 0.89
15% KletterMix	46.50 ± 2.49	63.00 ± 4.85	49.00 ± 0.50	36.95 ± 1.41	48.86 ± 1.41	57.42 ± 0.39	**79.00 ± 4.09**	75.42 ± 0.43	54.44 ± 1.46	66.57 ± 1.10	57.72 ± 0.89
20% KletterMix	**47.25 ± 2.49**	**65.00 ± 4.79**	**49.93 ± 0.50**	**37.37 ± 1.41**	**49.89 ± 1.40**	57.52 ± 0.39	**79.00 ± 4.09**	75.37 ± 0.43	**54.52 ± 1.46**	**66.60 ± 1.10**	**58.25 ± 0.89**

C.3.5 MATCHED SOURCE COMPARISONS

To separate the effect of the replacement share from the identity of the replacement corpus, we compare KletterMix with FineWeb2-DE and GermanWeb at the same 15% and 20% shares. The 15% 5-shot setting is the primary comparison because it is the best-performing KletterMix setting in the full 5-shot sweep. At that share, KletterMix reaches the highest German and eight-task means in both the KletterMix and source-control comparisons. For completeness, the appendix also reports the 20% 5-shot controls and the zero-shot source comparison at 20%, which corresponds to the zero-shot KletterMix optimum described above. These results keep the protocol-dependent choice of the primary setting explicit rather than implying that more German data is always better.

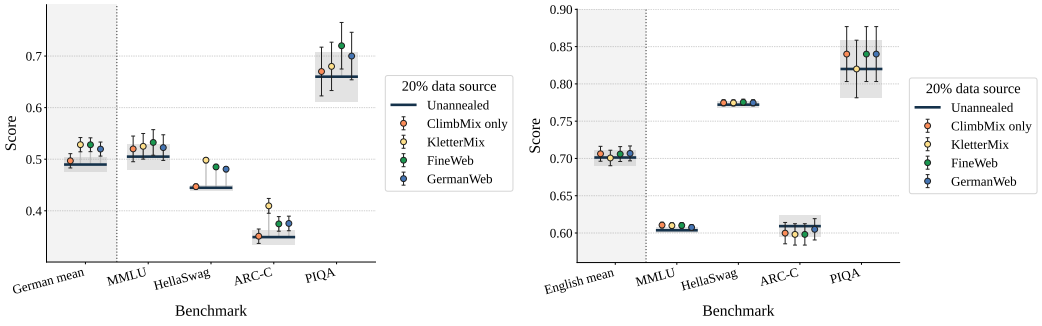
For completeness, we also report an exploratory cross-share comparison between the 20% KletterMix and 15% GermanWeb zero-shot configurations. Because the replacement shares differ, this comparison is not a matched source-control comparison and should not be interpreted as evidence that the two data sources have been isolated at equal mixture strength; it is included in Tab. 18 only as a direct description of the available results.



(a) 15% source comparison, German.

(b) 15% source comparison, English.

Figure 9: Detailed 5-shot benchmark scores for the fixed 15% source comparison. All models use the same OLMo3 initialization, annealing budget, and optimization schedule.



(a) 20% source comparison, German.

(b) 20% source comparison, English.

Figure 10: Detailed 5-shot benchmark scores for the fixed 20% source comparison. The aggregate tables include the corresponding zero-shot results as well.

Table 17: OLMo3 7B final-checkpoint benchmark results for the 15% source-replacement comparison (zero-shot). Values are percentages reported as mean ± propagated standard error. The German and English means are unweighted four-task means; the eight-task mean averages the two language means.

	German					English					
Source / control	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**Eight-task mean**
Unannealed	41.75 ± 2.47	62.00 ± 4.88	44.01 ± 0.50	31.74 ± 1.36	44.88 ± 1.41	57.58 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.06 ± 0.43	52.47 ± 1.46	66.03 ± 1.10	55.45 ± 0.89
ClimbMix only	42.00 ± 2.48	63.00 ± 4.85	43.88 ± 0.50	32.59 ± 1.37	45.37 ± 1.41	57.57 ± 0.39	78.00 ± 4.16	$\mathbf{75.53 \pm 0.43}$	54.18 ± 1.46	66.32 ± 1.11	55.84 ± 0.90
KletterMix	$\mathbf{46.50 \pm 2.49}$	63.00 ± 4.85	$\mathbf{49.00 \pm 0.50}$	$\mathbf{36.95 \pm 1.41}$	$\mathbf{48.86 \pm 1.41}$	57.42 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.42 ± 0.43	54.44 ± 1.46	66.57 ± 1.10	$\mathbf{57.72 \pm 0.89}$
FineWeb2-DE	44.50 ± 2.48	65.00 ± 4.79	47.33 ± 0.50	34.56 ± 1.39	47.85 ± 1.40	$\mathbf{57.73 \pm 0.39}$	$\mathbf{79.00 \pm 4.09}$	75.50 ± 0.43	54.27 ± 1.46	66.62 ± 1.10	57.24 ± 0.89
GermanWeb	$\mathbf{46.50 \pm 2.50}$	$\mathbf{67.00 \pm 4.73}$	47.05 ± 0.50	34.13 ± 1.39	48.67 ± 1.39	57.61 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.39 ± 0.43	$\mathbf{54.52 \pm 1.46}$	$\mathbf{66.63 \pm 1.10}$	57.65 ± 0.88

Table 18: Exploratory cross-share comparison under zero-shot evaluation. The 20% KletterMix and 15% GermanWeb rows use different replacement shares, so this table is descriptive rather than a matched source-control comparison. Values are percentages reported as mean ± propagated standard error.

Configuration	German mean	English mean	Eight-task mean
20% KletterMix	49.89 ± 1.40	66.60 ± 1.10	58.25 ± 0.89
15% GermanWeb	48.67 ± 1.39	66.63 ± 1.10	57.65 ± 0.88

The complete source-comparison tables, including both 15% and 20% shares and both shot settings, are given in Tab. 20, Tab. 21, Tab. 17, and Tab. 19.

D Dataset Summary

KletterMix is released under the CC-BY-NC-4.0 license and contains 725B GPT-2 tokens of German-language text. The dataset is distributed as sharded JSONL files (`shard_*.jsonl`) with a single `train` split. Each record has the following fields as displayed in Tab. 22.

E Impact Statement

KletterMix also has broader societal implications. On the positive side, the dataset can help reduce the gap between English and German pretraining resources, support more capable German-language models, and enable more controlled research on translated pretraining data. However, improvements in German model quality can also amplify risks associated with language models, including the generation of misleading text, biased or stereotyped outputs, privacy leakage from web-derived sources, and misuse in downstream applications. These risks are not unique to KletterMix, but they are important for any large-scale pretraining corpus. We therefore view documentation, provenance preservation, quality filtering, transparent release conditions, and clear intended-use statements as necessary parts of the dataset release.

Table 19: OLMo3 7B final-checkpoint benchmark results for the 20% source-replacement comparison (zero-shot). Values are percentages reported as mean ± propagated standard error. The German and English means are unweighted four-task means; the eight-task mean averages the two language means.

			German					English			
Source / control	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**Eight-task mean**
Unannealed	41.75 ± 2.47	62.00 ± 4.88	44.01 ± 0.50	31.74 ± 1.36	44.88 ± 1.41	$\mathbf{57.58 \pm 0.39}$	$\mathbf{79.00 \pm 4.09}$	75.06 ± 0.43	52.47 ± 1.46	66.03 ± 1.10	55.45 ± 0.89
ClimbMix only	42.00 ± 2.48	63.00 ± 4.85	43.88 ± 0.50	32.59 ± 1.37	45.37 ± 1.41	57.57 ± 0.39	78.00 ± 4.16	$\mathbf{75.53 \pm 0.43}$	54.18 ± 1.46	66.32 ± 1.11	55.84 ± 0.90
KletterMix	$\mathbf{47.25 \pm 2.49}$	65.00 ± 4.79	$\mathbf{49.93 \pm 0.50}$	$\mathbf{37.37 \pm 1.41}$	$\mathbf{49.89 \pm 1.40}$	57.52 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.37 ± 0.43	54.52 ± 1.46	66.60 ± 1.10	$\mathbf{58.25 \pm 0.89}$
FineWeb2-DE	45.50 ± 2.49	$\mathbf{70.00 \pm 4.61}$	47.74 ± 0.50	34.64 ± 1.39	49.47 ± 1.36	57.48 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.42 ± 0.43	54.35 ± 1.46	66.56 ± 1.10	58.02 ± 0.87
GermanWeb	46.50 ± 2.49	67.00 ± 4.73	47.63 ± 0.50	35.67 ± 1.40	49.20 ± 1.39	57.39 ± 0.39	$\mathbf{79.00 \pm 4.09}$	75.42 ± 0.43	$\mathbf{54.78 \pm 1.45}$	$\mathbf{66.65 \pm 1.10}$	57.92 ± 0.88

Table 20: OLMo3 7B final-checkpoint benchmark results for the 15% source-replacement comparison (5-shot). Values are percentages reported as mean ± propagated standard error. The German and English means are unweighted four-task means; the eight-task mean averages the two language means.

			German					English			
Source / control	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**Eight-task mean**
Unannealed	50.50 ± 2.50	66.00 ± 4.76	44.46 ± 0.50	34.90 ± 1.39	48.97 ± 1.39	60.37 ± 0.39	82.00 ± 3.86	77.22 ± 0.42	$\mathbf{60.92 \pm 1.43}$	70.13 ± 1.04	59.55 ± 0.87
ClimbMix only	52.00 ± 2.49	67.00 ± 4.73	44.68 ± 0.50	35.07 ± 1.39	49.69 ± 1.39	61.06 ± 0.39	$\mathbf{84.00 \pm 3.68}$	77.46 ± 0.42	59.98 ± 1.43	70.63 ± 1.00	60.16 ± 0.85
KletterMix	$\mathbf{53.00 \pm 2.45}$	$\mathbf{70.00 \pm 4.61}$	$\mathbf{49.29 \pm 0.50}$	$\mathbf{40.36 \pm 1.43}$	$\mathbf{53.16 \pm 1.36}$	$\mathbf{61.12 \pm 0.39}$	83.00 ± 3.78	77.43 ± 0.42	60.07 ± 1.43	70.41 ± 1.02	$\mathbf{61.78 \pm 0.85}$
FineWeb2-DE	52.25 ± 2.48	68.00 ± 4.69	47.95 ± 0.50	36.52 ± 1.41	51.18 ± 1.38	60.97 ± 0.39	$\mathbf{84.00 \pm 3.68}$	$\mathbf{77.50 \pm 0.42}$	60.24 ± 1.43	$\mathbf{70.68 \pm 1.00}$	60.93 ± 0.85
GermanWeb	52.25 ± 2.47	$\mathbf{70.00 \pm 4.61}$	47.43 ± 0.50	36.95 ± 1.41	51.66 ± 1.36	60.90 ± 0.39	82.00 ± 3.86	77.41 ± 0.42	60.41 ± 1.43	70.18 ± 1.04	60.92 ± 0.86

Table 21: OLMo3 7B final-checkpoint benchmark results for the 20% source-replacement comparison (5-shot). Values are percentages reported as mean ± propagated standard error. The German and English means are unweighted four-task means; the eight-task mean averages the two language means.

			German					English			
Source / control	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**MMLU**	**PIQA**	**HellaSwag**	**ARC-C**	**Mean**	**Eight-task mean**
Unannealed	50.50 ± 2.50	66.00 ± 4.76	44.46 ± 0.50	34.90 ± 1.39	48.97 ± 1.39	60.37 ± 0.39	82.00 ± 3.86	77.22 ± 0.42	$\mathbf{60.92 \pm 1.43}$	70.13 ± 1.04	59.55 ± 0.87
ClimbMix only	52.00 ± 2.49	67.00 ± 4.73	44.68 ± 0.50	35.07 ± 1.39	49.69 ± 1.39	$\mathbf{61.06 \pm 0.39}$	$\mathbf{84.00 \pm 3.68}$	77.46 ± 0.42	59.98 ± 1.43	70.63 ± 1.00	60.16 ± 0.85
KletterMix	52.50 ± 2.49	68.00 ± 4.69	$\mathbf{49.81 \pm 0.50}$	$\mathbf{40.96 \pm 1.44}$	$\mathbf{52.82 \pm 1.38}$	60.99 ± 0.39	82.00 ± 3.86	77.45 ± 0.42	59.81 ± 1.43	70.06 ± 1.04	61.44 ± 0.86
FineWeb2-DE	$\mathbf{53.25 \pm 2.48}$	$\mathbf{72.00 \pm 4.51}$	48.50 ± 0.50	37.46 ± 1.41	52.80 ± 1.34	61.01 ± 0.39	$\mathbf{84.00 \pm 3.68}$	$\mathbf{77.53 \pm 0.42}$	59.81 ± 1.43	70.59 ± 1.00	$\mathbf{61.70 \pm 0.84}$
GermanWeb	52.25 ± 2.49	70.00 ± 4.61	48.07 ± 0.50	37.54 ± 1.42	51.97 ± 1.36	60.75 ± 0.39	$\mathbf{84.00 \pm 3.68}$	77.46 ± 0.42	60.49 ± 1.43	$\mathbf{70.68 \pm 1.00}$	61.32 ± 0.84

Table 22: Summary of the KletterMix dataset.

Field	**Type**	**Description**
`cluster_id`	`int64`	Identifier of the topic cluster assigned during the clustering stage of the data pipeline.
`text`	`string`	The document text.
`token_count`	`int64`	Number of tokens in the document (GPT2 tokenizer).
`proxy_score`	`float64`	Quality score assigned by the proxy model, used for filtering and weighting during mixture construction.