MultiSynt/MT: Trillion-Token Multi-Parallel Pre-Training Data Translated Across 36 Languages

Anonymous ACL submission

Abstract

Open web-scale pre-training corpora remain concentrated in English, limiting multilingual LLM development. We introduce MultiSynt/MT, an open synthetic parallel corpus with approximately 4.8 trillion target-language tokens across 36 languages, produced by translating 100 billion high-quality Nemotron-CC tokens with TOWER+ and OPUS-MT/HPLT-MT systems. For many medium- and lower-resource European languages, this is the largest openly available pre-training resource. On a broad multilingual benchmark suite, reference LLMs trained on MultiSynt/MT reach the final score of HPLT 2.0, a native-data baseline, using roughly 72% fewer pre-training tokens, and outperform it by approximately 15% relative at a matched 100B-token training budget. Our analyses also identify evaluation blind spots: standard multiple-choice benchmarks miss translation-quality differences that a fluency-sensitive LLM-as-judge evaluation cleanly recovers on the trained LLMs (with no fluency deficit in MultiSynt itself), and Norwegian idiomatic and culturally grounded tasks, for example, remain better served by native data. We release the corpus, including row-aligned translations from multiple systems, to support controlled research on multilingual pre-training data and evaluation.

1 Introduction

Openly available pre-training data at scale exist primarily for English (Penedo et al., 2024; Su et al., 2025), with most other languages lacking volume or quality (see Section 2 for details). This shortfall has motivated growing interest in machine translation as a means of producing multilingual pre-training data at scale, but the practice raises legitimate concerns: translated text can carry stylistic artifacts known as "translationese" and inherits the cultural reference frame of the source language, with names, places, idioms and culturally grounded knowledge in the English source remaining English-anchored after translation (Gellerstam, 1986; Riley et al., 2020).

We address this resource limitation by introducing **MultiSynt/MT**, an open multilingual synthetic parallel corpus of approximately **4.8 trillion target-language tokens covering 36 languages**, produced by translating a 100B-token sample of high-quality web data from Nemotron-CC with open translation models, including TOWER+ 9B and 72B (Rei et al., 2026) and OPUS-MT (Tiedemann et al., 2024). MultiSynt/MT is the largest openly available pre-training corpus to date for many European low- and medium-resource languages, exceeding the largest comparable native resource (HPLT 3.0; Oepen et al., 2025) by more than an order of magnitude in the lowest-resource ones (Figure 1). For a subset of languages, we release row-aligned parallel translations from multiple systems, enabling controlled comparison under matched source data, and the corpus is released under a permissive open license.

Alongside the corpus, we present a balanced empirical characterization reporting both where reference LLMs trained on MultiSynt/MT outperform native multilingual baselines and where they fall short, and drawing attention to evaluation practices that obscure these differences. On standard multilingual base-model benchmarks, reference LLMs pre-trained on MultiSynt/MT reach the final score of a native-data baseline (HPLT 2.0[1]), using approximately 72% fewer training tokens, and improve over it by approximately 15% relative at a matched 100B-token training budget (Section 4). A fluency-sensitive LLM-as-judge evaluation on the trained LLMs cleanly recovers an

[1] Our pre-training experiments began before the more recent HPLT 3.0 release was available, so we use HPLT 2.0 as the native baseline throughout Section 4. While the 3.0 version is larger, the distributions of the two are closely similar (Oepen et al., 2025); see also Figure 6.

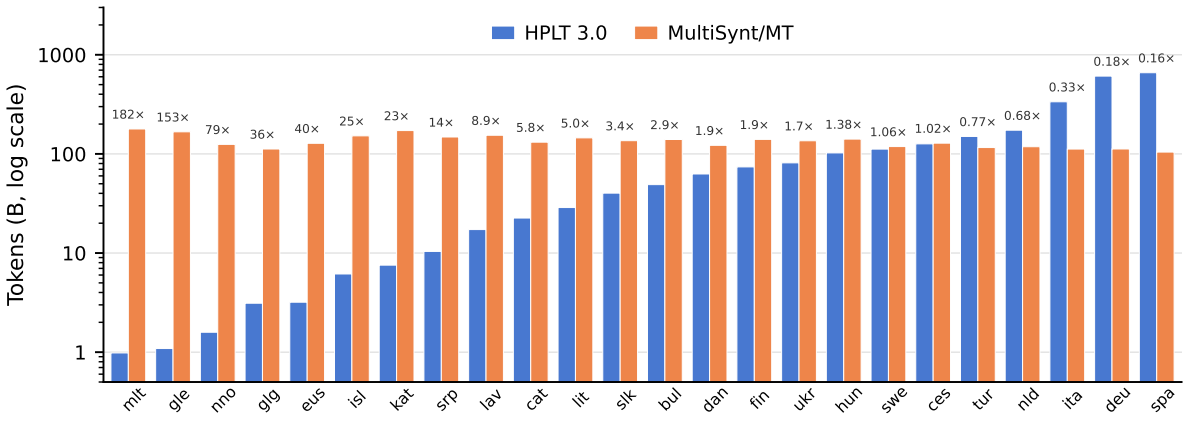

Figure 1: Per-language token counts in MultiSynt/MT and HPLT 3.0 (log scale, Gemma-3 tokenization, 24-language representative subset). MultiSynt/MT is larger for 19 of 24 languages, by over an order of magnitude for the lowest-resource ones (e.g. mlt 182×, gle 153×, nno 79×); HPLT 3.0 is larger only on deu and spa (5–6×). The value above each group gives the MultiSynt/MT-to-HPLT 3.0 token ratio. Full statistics in Appendix D.

MT-system quality ranking that standard multiple-choice benchmarks fail to discriminate, exposing a blind spot in standard evaluation rather than a fluency deficit in MultiSynt (Section 5.1); a complementary embedding-space diagnostic finds that standard benchmark items tend to have more MultiSynt/MT than HPLT 2.0 documents among their nearest neighbors (Section 5.2). On Norwegian tasks targeting idiomatic and culturally grounded knowledge, models trained on native data outperform translated-data models throughout training, while on Norwegian commonsense reasoning the translated data closes the gap and eventually overtakes the native baseline; direct evidence that translated text is not a substitute for native corpora on all phenomena (Section 5.3).

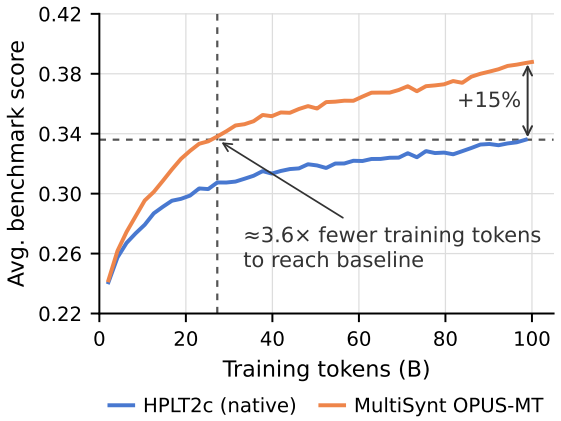

Figure 2: HPLT 2.0 (native) vs MultiSynt/MT (OPUS-MT) on the multilingual benchmark suite, averaged over 5 languages (dan, nld, ita, por, swe). MultiSynt/MT crosses the HPLT 2.0 endpoint at ∼28B training tokens and continues to improve through 100B tokens. Per-language/benchmark breakdowns are provided as supplementary data.

2 Related Work

Multilingual pre-training corpora. Open multilingual pre-training corpora have evolved from filtered Common Crawl pipelines (CCNet, mC4, OSCAR, ROOTS) to larger and better-filtered resources including MADLAD-400, CulturaX, Glot500, HPLT, FineWeb2 and EMMA-style adaptation (Wenzek et al., 2020; Conneau et al., 2020; Xue et al., 2021; Abadji et al., 2022; Laurençon et al., 2022; BigScience Workshop, 2022; Kudugunta et al., 2023; Nguyen et al., 2024; Imani-Googhari et al., 2023; de Gibert et al., 2024; Burchell et al., 2025; Penedo et al., 2025; Ji et al., 2024a), but a persistent volume and quality imbalance remains for lower-resource languages (Kreutzer et al., 2022). MultiSynt/MT targets that part of the language distribution as a complement to these corpora; we use HPLT (Oepen et al., 2025), the strongest openly available native multilingual baseline at our scale, as our principal point of comparison.

High-quality English source data. Translation-based corpora inherit the quality distribution of their source: recent reference-baseline work shows that Nemotron-CC HQ (Su et al., 2025) produces the strongest downstream performance across model and token scales, ahead of established English corpora including C4, The Pile, RefinedWeb, FineWeb-Edu and DataComp-LM (Raffel et al.,

2020; Gao et al., 2021; Penedo et al., 2023, 2024; Li et al., 2024a; Nezhurina et al., 2025), motivating its use as the MultiSynt/MT source distribution.

Translated corpora for language-model pre-training. Studies on Basque, Indian languages, Arabic, Indonesian and Tamil consistently find that machine-translated pre-training data is competitive with native data on standard NLU but underperforms on culturally nuanced tasks (Urbizu et al., 2023; Doshi et al., 2024; Boughorbel et al., 2024; Velasco and Roque, 2025). Closest in scale, TransWeb-Edu / CuatroLLM and its extension TransWebLLM translate up to 1.7T tokens from FineWeb-Edu into nine languages (Wang et al., 2024a, 2025), and the contemporaneous FineTranslations reverses the direction by translating 500+ non-English languages into English (Penedo et al., 2026). A parallel line treats translation as an explicit pre-training objective rather than a corpus-construction mechanism (Ji et al., 2024b; Li et al., 2024b; Ji et al., 2025; Li et al., 2025); we instead train standard next-token language models on translated target-language documents. MultiSynt/MT extends this line in three ways: 36-language coverage at approximately 4.8T target-language tokens, a Nemotron-CC HQ rather than FineWeb-Edu source, and row-aligned outputs from multiple open MT systems (translation-specialized LLMs in the Tower family (Alves et al., 2024; Rei et al., 2026) and classical NMT systems (Tiedemann, 2012; Tiedemann et al., 2024; Junczys-Dowmunt et al., 2018) supporting controlled comparison under matched source data.

Translationese, cultural grounding, and native-authored evaluation. Translated text differs systematically from natively authored text (Koppel and Ordan, 2011; Volansky et al., 2015; Vanmassenhove et al., 2021; Bizzoni et al., 2020), with downstream effects on cross-lingual benchmark interpretation (Graham et al., 2019; Freitag et al., 2020; Artetxe et al., 2020; Riley et al., 2020, 2021), and multilingual models in turn often lack the cultural and idiomatic knowledge of the target-language community (Hershcovich et al., 2022; Naous et al., 2024; Liu et al., 2024; Pawar et al., 2025; Yao et al., 2024b; Etxaniz et al., 2024). Multilingual evaluation is itself often translation-derived (Belebele, MMLU-ProX, XNLI, XTREME) (Conneau et al., 2018; Hu et al., 2020; Bandarkar et al., 2024; Xuan et al., 2025); native-authored alternatives such as TyDi QA, Global MMLU and NorEval are an essential complement (Clark et al., 2020; Singh et al., 2025; Mikhailov et al., 2025; Romanou et al., 2025; Wu et al., 2025), and LLM-as-a-judge protocols (Zheng et al., 2023; Liu et al., 2023; Verga et al., 2024; Gu et al., 2026) provide a fluency-sensitive complement that we exploit in Section 5.

3 Building MultiSynt/MT

3.1 Source data: high-quality English web text

We translate from the *actual high-quality* split of Nemotron-CC (Su et al., 2025), which contains non-synthetic Common Crawl documents receiving the highest scores from its ensemble quality classifier. Recent reference-baseline work (Nezhurina et al., 2025) finds that Nemotron-CC HQ produces the strongest average downstream performance among established open English pre-training corpora across multiple model and token scales. We draw a uniform random sample of approximately 155 million documents (on the order of 100 billion English tokens), sized so that each per-language translation yields outputs at the same order of magnitude, and apply no topical, length, or quality stratification beyond the underlying quality classifier. Translations preserve a stable mapping to their source documents, so the resulting corpus is usable both as per-language pre-training data and as a multi-way parallel resource for cross-language and cross-system experiments (Section 3.4).

3.2 Translation system pool and selection

We require open translation systems that combine strong target-language quality with throughput feasible at the 100B-token scale, and that together cover the languages targeted by MultiSynt/MT, including medium- and lower-resource European languages where openly available native pre-training data is scarce. No single family meets all three requirements: translation-specialized LLMs in the TOWER family (Alves et al., 2024; Rei et al., 2026) and general-purpose open instruction-tuned LLMs (Martins et al., 2025; Yang et al., 2025; Gemma Team, 2025) concentrate on higher-resource European languages, while OPUS-MT and HPLT-MT (Tiedemann et al., 2024; de Gibert et al., 2024) extend coverage to lower-resource languages. We therefore use LLM-based translation and OPUS-MT/HPLT-MT jointly as the production pool.

To rank candidate systems, we ran a small human-judged quality study covering seven lan-

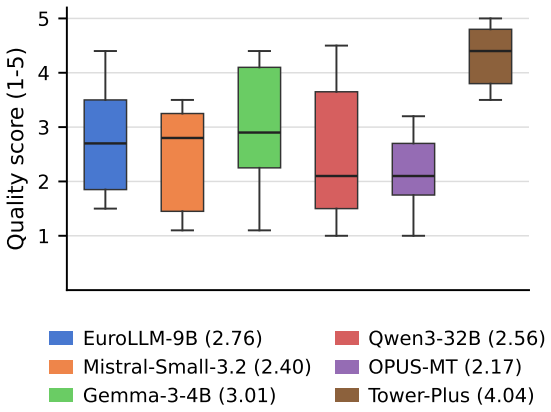

Figure 3: Mean human quality rating per candidate MT system, averaged over the seven evaluated languages on a 1–5 Likert-style scale. TOWER+ is rated highest overall. See Appendix E for the protocol and Figure 8 for the per-language breakdown.

guages (Czech, French, Finnish, German, Italian, Polish, Spanish). For each language, a single native speaker rated translations of the same 10 English source documents from each of six candidate systems on a fluency-focused 1–5 scale, blind to system identity, additionally penalizing hallucinations, repetitions, and extra or missing text. The full protocol, KEOPS annotation interface, and per-language results are in Appendix E. TOWER+ (we use the 72B model) is rated highest in every language, the general-purpose LLMs (EuroLLM-9B-Instruct, Qwen3-32B, Gemma-3-4b-it, Mistral-Small-3.2-24B-Instruct) bunch well below it, and OPUS-MT is close to the general-purpose LLMs on average (Figure 3); the ranking is consistent with the public TOWER MT leaderboard and with per-language Flores-200 BLEU on the OPUS-MT dashboard.[2] Based on this ranking the general-purpose LLMs were dropped from the production pool.

We ship translations from TOWER+ 9B (Gemma-based, 16 languages)[3] and TOWER+ 72B (Qwen-based, 5 languages: German, Finnish, Italian, Spanish, Swedish).[4] A reference-LLM comparison between the two on the 5-language overlap (Section 4) showed a small enough downstream gap that we treat 9B as the default LLM-based system and retain 72B on the high-resource subset where any residual gains can be isolated. OPUS-MT supplies the production models for 31 of our 36 target languages from the Tatoeba Translation Challenge collection (Tiedemann, 2020), and HPLT-MT supplies the remaining 5 (Basque, Irish, Icelandic, Maltese, Albanian) where no comparable OPUS-MT model is available. These NMT systems are orders of magnitude smaller than the LLM-based candidates (∼75M–230M parameters) with corresponding throughput advantages, at the cost of a sentence-level constraint addressed by the pipeline of Section 3.3. The released corpus therefore contains translations from TOWER+ 9B (16 languages), TOWER+ 72B (5 languages), and OPUS-MT/HPLT-MT (36 languages); for the languages where multiple systems are shipped, the corpus supports controlled comparison of translation choices under matched source data, which we exploit in Sections 4 and 5.

3.3 Translation pipeline at scale

We run two parallel translation pipelines on the same 155 million source documents, both supporting per-shard resumption and emitting Parquet files that preserve source document identifiers and order so that the released corpus remains a synchronized multi-parallel resource. The *LLM pipeline* deploys TOWER+ 9B and 72B on Leonardo HPC cluster (CINECA, NVIDIA A100 nodes) with vLLM under a Slurm-native orchestration toolkit released as open-source code,[5] consuming approximately 3.1 million A100 GPU-hours in aggregate. The *NMT pipeline* decodes OPUS-MT and HPLT-MT models with Marian-NMT (Junczys-Dowmunt et al., 2018) on LUMI cluster (AMD MI250x), with sentence-splitting pre-processing and document-reassembly post-processing that preserve both per-sentence and per-document alignments; the pipeline is released as open-source code at `github.com/anonymous-org/translate-fineweb`. Per-shard throughput, decoding settings, energy, and carbon-footprint estimates for both pipelines are reported in Appendix A.

3.4 Resource overview

MultiSynt/MT contains approximately 4.8 trillion target-language tokens across 36 languages, derived from the single shared sample of 155M English documents introduced in Section 3.1, and is the largest openly available pre-training cor-

[2] `https://opus.nlpl.eu`
[3] `https://huggingface.co/Unbabel/Tower-Plus-9B`
[4] `https://huggingface.co/Unbabel/Tower-Plus-72B`
[5] `github.com/anonymous-org/distributed-inference`; URL to be replaced at camera-ready.

pus to date for many European low- and medium-resource languages (Figure 1). Counting each translation system's output separately, the released artifact totals approximately 7.1T tokens (1.7T from TOWER+ 9B, 0.5T from TOWER+ 72B, 4.8T from OPUS-MT/HPLT-MT, all under Gemma-3 tokenization); per-language and per-system breakdowns are in Appendix D.

Splits and schema. For each language we release three splits: The *parallel* split contains the documents translated by all systems released for that language, row-aligned by index so that row i of any language file refers to the same source document, supporting cross-language and cross-system controlled experiments. The *additional* split contains further translated documents that are not present for every language, enabling larger language-specific volumes at the cost of cross-language alignment, and the *all* split is the union of the two. Each row contains the source `warc_record_id`, the translated text, its token count, and the source partition identifier.

Availability and license. MultiSynt/MT is released under a Creative Commons Zero (CC0) license applied to the corpus as a database (underlying text content remains subject to Common Crawl source-content terms), at `hf.co/anonymous`. The OPUS-MT/HPLT-MT subset is additionally released as a sentence-aligned variant.[6]

4 Effectiveness for Multilingual Pre-training

4.1 Reference model setup

We evaluate MultiSynt/MT as pre-training data by training reference language models on each MultiSynt/MT translation variant and on a native multilingual baseline dataset. Across variants, we hold the model architecture, tokenizer, optimizer settings, and total token budget fixed; the variants differ in the pre-training corpus itself—specifically, the source data and the translation system (if any)—and in the set of languages each corpus covers.

Our reference architecture follows that of Penedo et al. (2024): we use a dense Llama-like transformer with 24 layers, hidden size 2048, 32 attention heads, RoPE positional embeddings, and RMSNorm, with 1.61B parameters in transformer layers. We use the `google/gemma-3-27b-pt` tokenizer (Gemma Team, 2025) with a vocabulary size of 262144 for its broad multilingual coverage across the languages in MultiSynt/MT. We tie word embeddings, giving a total of 2.15B parameters. The optimization recipe uses Adam with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.05, gradient clipping at 1, global batch size 1024, and a linear WSD learning-rate schedule with 10% warmup and 20% cooldown. Full hyperparameters are reported in Appendix B, Table 1.

Training is conducted with Megatron-LM (Shoeybi et al., 2019) on the LUMI supercomputer, using 16 nodes with 4 AMD MI250x GPUs per node. For every pre-training data variant, we train a separate model on a budget of 100 billion tokens, saving checkpoints every 1,000 steps. Each variant is trained once; we do not report variance over training seeds, and the curves shown in Section 4.2 should be read as single-seed point estimates.

We compare four pre-training data variants under this setup: HPLT 2.0 (Burchell et al., 2025), a strong openly available native multilingual baseline at the relevant scale, and MultiSynt/MT translated by TOWER+ 9B (16 languages), TOWER+ 72B (5 languages), and OPUS-MT/HPLT-MT (36 languages). Holding architecture, optimizer, and budget fixed lets us attribute downstream differences to the pre-training corpus. Within the MultiSynt/MT family, this isolates the choice of translation system; between MultiSynt/MT and HPLT 2.0 the comparison additionally varies the source corpus and its quality-filtering pipeline, a confounding factor we discuss in Section 6.

Evaluation. Downstream evaluation uses two open frameworks, LightEval (Habib et al., 2023) and LM-evaluation-harness (Gao et al., 2024). We evaluate each checkpoint on a broad multilingual benchmark suite: belebele (Bandarkar et al., 2024), arc:challenge (Clark et al., 2018; Dac Lai et al., 2023), hellaswag (Zellers et al., 2019; Dac Lai et al., 2023), goldenswag (Chizhov et al., 2025), xstory-cloze (Lin et al., 2021), global-mmlu (Singh et al., 2025), mmlu (Hendrycks et al., 2021; Dac Lai et al., 2023), exams (Hardalov et al., 2020), xcodah (Chen et al., 2019), xcsqa (Talmor et al., 2019), MultiBlimp (Jumelet et al., 2026), enem (Nunes et al., 2023), faquad (Sayama et al., 2019), oab-exams (Pires et al., 2025), glianorex (Griot et al., 2025), xnli (Conneau et al., 2018), pawsx (Yang et al., 2019), Finbench (Kytöniemi et al., 2025). We appy the same evaluation protocol across all variants.

[6]The URL is omitted for anonymity.

Per-language scores are aggregated into a single mean across the languages covered by every variant under comparison; per-language and per-benchmark breakdowns are provided as supplementary data.

Decontamination check. A controlled training experiment with vs. without n-gram decontamination of the Nemotron-CC source rules out source-document contamination as an explanation for the Section 4.2 gains (the released MultiSynt/MT corpus is itself not decontaminated due to the low risk of actual contamination; see Appendix C).

4.2 Gains over native baselines

Figure 2 reports aggregate benchmark performance over training tokens, comparing reference LLMs pre-trained on the HPLT 2.0 native baseline against reference LLMs pre-trained on MultiSynt/MT OPUS-MT, averaged over the five languages where both baselines have evaluation data (Danish, Dutch, Italian, Portuguese, Swedish). Models pre-trained on MultiSynt/MT reach the HPLT 2.0 endpoint score at approximately 28B training tokens, roughly 72% fewer than the native baseline; at the matched 100B-token training budget, MultiSynt/MT exceeds the HPLT 2.0 endpoint by approximately 15% relative.

Figure 4 compares the three MultiSynt/MT variants (OPUS-MT, TOWER+ 9B, TOWER+ 72B) on Swedish, the single language for which a reference LLM was trained on every variant. The three MultiSynt/MT curves cluster tightly together throughout training and all sit well above the HPLT 2.0 baseline, indicating that on these standard multilingual benchmarks the gain over native data is largely independent of the choice of MT system within MultiSynt/MT. The downstream gap among TOWER+ 9B, TOWER+ 72B, and OPUS-MT is therefore small enough that we treat the 9B variant as our default LLM-based system and retain 72B only on the high-resource subset where any residual gains can be isolated (Section 4.1).

Per-language and per-benchmark breakdowns are reported in Appendix D.

5 Interpreting the Gains: Fluency, Benchmark Alignment, and Native-Authored Tasks

Section 4 showed that training on MultiSynt/MT yields large gains over native data on standard multilingual benchmarks. We now interpret those

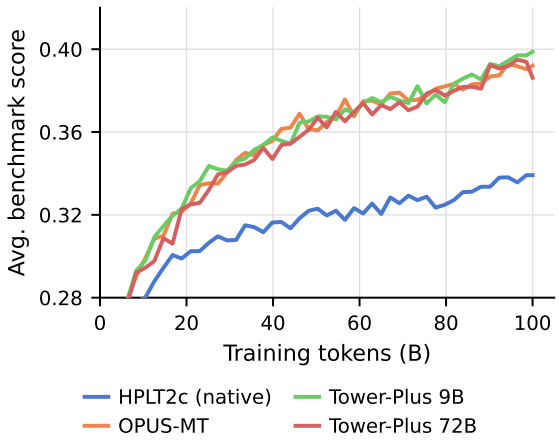

Figure 4: MultiSynt/MT variants (OPUS-MT, TOWER+ 9B, TOWER+ 72B) vs the HPLT 2.0 native baseline on the Swedish multilingual benchmark suite (the single language for which a reference LLM was trained on every variant). The three MultiSynt/MT curves sit indistinguishably above HPLT 2.0 throughout training: translated vs native dominates MT-system choice.

gains along three axes the standard benchmarks do not surface: an LLM-judge fluency probe that exposes MT-system differences flattened by multiple-choice scoring (Section 5.1), an embedding-space diagnostic of benchmark-corpus alignment (Section 5.2), and a Norwegian case study on native-authored idiomatic and culturally grounded tasks (Section 5.3). We draw the methodological consequences in Section 6.

5.1 What standard benchmarks miss about MT-system choice

The MT-variant indistinguishability of Figure 4 is surprising: the same three translation systems are clearly separated by human raters (Figure 3) and by the public TOWER MT leaderboard. A plausible explanation is that translationese artifacts in the MT output propagate to the trained LLM in a form the standard multilingual benchmarks of Section 4, predominantly multiple-choice QA and selection, cannot detect: those benchmarks reward picking the right token under strong exploitable cues, while the surface fluency of the model's own free-form generations does not enter the score.

To probe this hypothesis, we evaluate the pre-trained models with an LLM-as-a-judge protocol that scores fluency of free-form generations. For each of German, Spanish, Finnish, Italian and Swedish we collect 100 natural-language cut-out sentences (e.g., *"Das Grundgesetz der Bundesre-*

publik Deutschland garantiert… ”). Each candidate pre-trained model produces a continuation that DeepSeek V3.1 then compares against a continuation from a monolingual HPLT 2.0 1.7B native-data baseline; we report the candidate’s *winrate* averaged over prompts and languages.

As a sanity check, the off-the-shelf Qwen 2.5 series at five sizes from 0.5B to 14B parameters shows winrate rising monotonically with model size against the HPLT 2.0 baseline (Appendix F, Figure 9), confirming the judge responds to generation quality. The reference LLMs pre-trained on MultiSynt/MT with matched data and budget produce a clear ranking, TOWER+ 72B $\geq$ TOWER+ 9B $>$ OPUS-MT (Figure 5), consistent with the human evaluation of the underlying MT outputs (Figure 3) and the public TOWER MT leaderboard. All three MultiSynt/MT variants match or exceed the HPLT 2.0 baseline on the judge, so MultiSynt/MT is not fluency-deficient; the finding is that standard benchmarks are blind to translation-quality differences the trained LLMs nonetheless inherit.

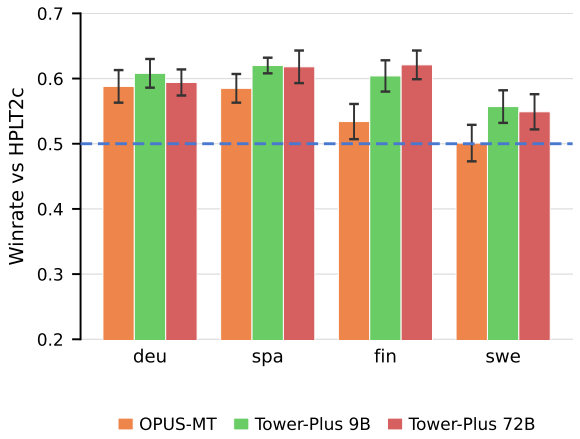

Figure 5: Fluency LLM-judge winrate of our reference 1.7B LLMs pre-trained on MultiSynt/MT against a monolingual HPLT 2.0 1.7B baseline. The ranking TOWER+ 72B $\geq$ TOWER+ 9B $>$ OPUS-MT matches the human MT evaluation (Figure 3) and the public Tower MT leaderboard.

5.2 Benchmark-corpus alignment in embedding space

As a complementary diagnostic, we compare the embedding neighborhoods of standard benchmark items against MultiSynt/MT and HPLT 2.0 documents. Across most benchmarks in the suite, benchmark items have more MultiSynt/MT documents among their nearest embedding neighbors, suggesting that translated high-quality English-source data provides strong coverage of many benchmark-adjacent regions; this is not a causal test of the gains, but it indicates that benchmark-corpus alignment is one reason those gains should be interpreted with care (Appendix G).

5.3 Where translated data falls short: a Norwegian case study

Named entities, idiomatic constructions, and culturally specific knowledge in English text remain English-anchored after translation, so we expect translated data to underperform native data on target-language tasks probing local idiomaticity or culture even when the translations are fluent. We test this by training four reference LLMs sharing the architecture and recipe of Section 4 and differing only in Norwegian pre-training data: HPLT 2.0 (native), HPLT 3.0 (native), MultiSynt/MT translated by TOWER+ 9B, and MultiSynt/MT translated by OPUS-MT, evaluated across 11 training checkpoints on two Bokmål tasks from NorEval[7] (Mikhailov et al., 2025): *NorIdiom* (idiomatic expressions) and *NorCommonsenseQA* (commonsense reasoning), selected as illustrative examples following the criteria in Oepen et al. (2025).

On *NorIdiom* (Figure 6, left), both native-data models outperform both translated-data models with a stable gap throughout training, consistent with idiomatic constructions being a long-tail phenomenon of the target language that the translated corpus does not generate at the rate native text does. On *NorCommonsenseQA* (Figure 6, right), native data leads in early-to-mid training but the translated-data models close the gap and match or exceed the native models in later checkpoints; this is a complementary signal that translated data lacks local cultural anchoring but preserves the structured reasoning patterns of the high-quality English source (Kabra et al., 2026). Together, the two tasks suggest a content-dependent picture: native data wins on tasks probing culturally or idiomatically local knowledge; on tasks probing transferable reasoning structure, models trained on MultiSynt/MT match or exceed those trained on native data at matched training budget.

[7] NorEval comprises 24 native-authored Norwegian tasks covering both Bokmål and Nynorsk, the two written variants of Norwegian. Aggregate benchmark results, obtained by normalizing and averaging scores across all tasks, show a clear advantage for models trained on native Norwegian data.

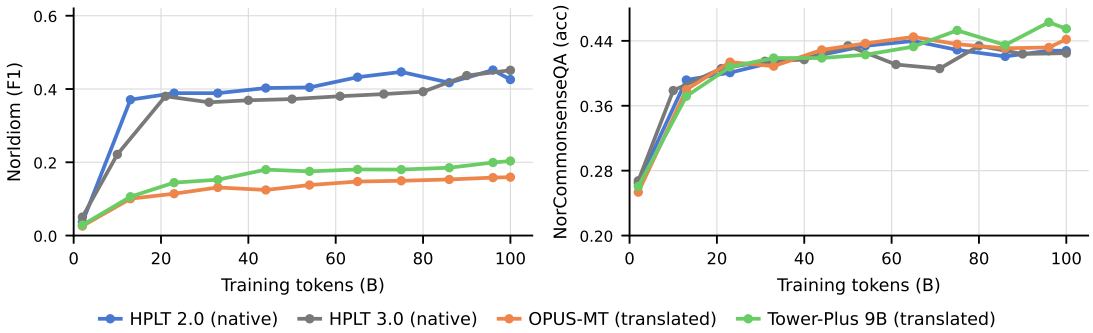

Figure 6: NorEval performance across 11 training checkpoints for reference LLMs pre-trained on native Norwegian (HPLT 2.0, HPLT 3.0) vs MultiSynt/MT translated (OPUS-MT, TOWER+ 9B) data. Left: NorIdiom-nob (Bokmål idiom completion, F1); native data dominate translated data throughout training. Right: NorCommonsenseQA-nob (Bokmål commonsense MCQ, accuracy); the four curves overlap, translated data is slightly ahead at 100B tokens.

6 Discussion and Recommendations

Treat translated data as a complement to native data, not a replacement. For high-resource languages, where native corpora are already abundant, MultiSynt/MT is most useful as a domain-coverage extension. For medium- and lower-resource languages, translating a high-quality source corpus is the primary vehicle for closing the high-quality data-scale gap, but it should still be paired with whatever native corpus is available to preserve idiomatic and cultural anchoring (Section 5.3).

The headline gain may reflect source quality alone, not translation as a beneficial operation. The Section 4.2 comparison varies both the source corpus (Nemotron-CC HQ vs. native HPLT 2.0 web text) and the translation step, so the gain cannot be attributed to translation *per se*: it is consistent with the net effect arising entirely from the higher quality of the English source, with translation contributing nothing of its own. Disentangling the two would require a controlled source-corpus comparison that we leave to future work. The practical counterfactual is nonetheless not arbitrary translated versus arbitrary native text, but translated high-quality source text at 100B-token scale versus the much smaller native pool that would survive equivalently stringent quality filtering, which for medium- and lower-resource languages would shrink the already-scarce native data to a small fraction of itself.

Pair translation-derived benchmarks with native-authored ones, and complement multiple-choice with quality-sensitive measures. A large fraction of widely used multilingual benchmarks, including Belebele (Bandarkar et al., 2024), m-ArenaHard (Dang et al., 2024), MMMLU (OpenAI, 2024) and MT-Bench-X (Weber et al., 2024), are themselves produced by translating English sources, and our embedding-space diagnostic shows that they tend to sit in neighborhoods with more MultiSynt/MT than HPLT 2.0 documents (Section 5.2). A model trained on translated data shares the same source-language frame as these benchmarks and is, in part, being evaluated on how well it has absorbed that frame; standard multiple-choice scoring is additionally blind to fluency-quality differences that the resulting LLMs nevertheless inherit (Section 5.1). Native-authored benchmarks such as NorEval (Mikhailov et al., 2025) expose gaps that translation-derived benchmarks fail to surface (Section 5.3); free-form quality-sensitive measures such as the LLM-judge protocol of Section 5.1 expose differences that discriminative ones flatten.

7 Conclusion

MultiSynt/MT provides approximately 4.8T target-language tokens across 36 languages, the largest openly available pre-training corpus to date for many European low- and medium-resource languages, and reference LLMs trained on it match a native-data baseline (HPLT 2.0) using roughly 72% fewer pre-training tokens while exceeding it by approximately 15% relative at a matched 100B-token budget. The translated corpus nonetheless falls short on idiomatic and culturally grounded Norwegian tasks and is best understood as a complement to native data; we release it under CC0 with row-aligned translations from multiple MT systems to support controlled follow-up on translation-system choice, scaling, native-benchmark coverage, and source-quality vs. translation attribution.

Limitations

Corpus and translation biases. MultiSynt/MT inherits the topical, stylistic, and English-centric biases of its Nemotron-CC HQ source: a quality-filtered Common Crawl pool that favours factual prose and formal registers over oral, vernacular, dialectal, and culturally local text, and that anchors named entities, idioms, and culturally specific knowledge to an English reference frame. The chosen MT systems further introduce systematic translationese artifacts (reduced lexical variety, calque-driven syntax, lower idiomatic density) that the standard multilingual benchmarks of Section 4 flatten but that surface under fluency-sensitive evaluation (Section 5.1) and on natively authored idiomatic and culturally grounded tasks (Section 5.3). The headline comparison against HPLT 2.0 (Section 4.2) additionally varies the source distribution and its quality-filtering pipeline, so the gain cannot be attributed to translation alone; an isolating comparison against an equivalently stringently filtered native baseline is left to future work.

Evaluation coverage, scale, and feedback risks. Our empirical analyses are concentrated on European languages in the high- and medium-resource range and are conducted at a single 1.7B-parameter scale with single-seed pre-training runs (Section 4.1), so behaviour on the lowest-resource and non-European languages in the released corpus, scaling trends to larger models, and seed-level variance are not directly verified by our experiments. The embedding-space alignment diagnostic (Section 5.2) is descriptive rather than causal, the fluency LLM-judge protocol (Section 5.1) inherits the biases of its judge and prompt set, and the human MT evaluation (Section 3.2) covers only seven languages with one annotator each. Finally, releasing MultiSynt/MT openly creates a feedback path into future open MT systems trained on web crawls that will then ingest it, with the attendant long-tail erosion risks of recursive synthetic-data training (Shumailov et al., 2024); we mark the corpus as synthetic in the released metadata and recommend treating it as a complement to native data rather than a replacement (Section 6).

References

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. 2022. Towards a cleaner document-oriented multilingual crawled corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France. European Language Resources Association.

Duarte Miguel Alves, José Pombal, Nuno M. Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and André Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation artifacts in cross-lingual transfer learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7674–7684, Online. Association for Computational Linguistics.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

BigScience Workshop. 2022. BLOOM: A 176b-parameter open-access multilingual language model. *CoRR*, abs/2211.05100.

Yuri Bizzoni, Tom S Juzek, Cristina España-Bonet, Koel Dutta Chowdhury, Josef van Genabith, and Elke Teich. 2020. How human is machine translationese? comparing human and machine translations of text and speech. In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 280–290, Online. Association for Computational Linguistics.

Sabri Boughorbel, Md Rizwan Parvez, and Majd Hawasly. 2024. Improving language models trained on translated data with continual pre-training and dictionary learning analysis. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 73–88, Bangkok, Thailand. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Laurie Burchell, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, Liane Guillou, Barry Haddow, Jan Hajič, Jindřich Helcl, Erik Henriksson, Mateusz Klimaszewski, Ville Komulainen, Andrey Kutuzov,

Joona Kytöniemi, Veronika Laippala, Petter Mæhlum, Bhavitvya Malik, and 16 others. 2025. An expanded massive multilingual dataset for high-performance language technologies (HPLT). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17452–17485, Vienna, Austria. Association for Computational Linguistics.

Michael Chen, Mike D'Arcy, Alisa Liu, Jared Fernandez, and Doug Downey. 2019. Codah: An adversarially-authored question answering dataset for common sense. In *Proceedings of the 3rd Workshop on Evaluating Vector Space Representations for NLP*, pages 63–69, Minneapolis, USA. Association for Computational Linguistics.

Pavel Chizhov, Mattia Nee, Pierre-Carl Langlais, and Ivan P. Yamshchikov. 2025. What the hellaswag? on the validity of common-sense reasoning benchmarks. *Preprint*, arXiv:2504.07825.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

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

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

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. 2024. A new massive multilingual dataset for high-performance language technologies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 1116–1128, Torino, Italia. ELRA and ICCL.

Meet Doshi, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pretraining language models using translationese. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5843–5862, Miami, Florida, USA. Association for Computational Linguistics.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lopez de Lacalle, and Mikel Artetxe. 2024. BertaQA: How much do language models know about local culture? In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Markus Freitag, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Online. Association for Computational Linguistics.

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. 2021. The Pile: An 800GB dataset of diverse text for language modeling. *CoRR*, abs/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, and 5 others. 2024. The language model evaluation harness.

Martin Gellerstam. 1986. Translationese in swedish novels translated from english. *Translation studies in Scandinavia*, 1:88–95.

Gemma Team. 2025. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in machine translation evaluation. In *Proceedings of the Fourth Conference on Machine Translation*, pages 72–81, Florence, Italy. Association for Computational Linguistics.

Maxime Griot, Jean Vanderdonckt, Demet Yuksel, and Coralie Hemptinne. 2025. Pattern recognition or medical knowledge? the problem with multiple-choice questions in medicine. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 5321–5341. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. 2026. A survey on LLM-as-a-judge. *The Innovation*, page 101253.

Nathan Habib, Clémentine Fourrier, Hynek Kydlíček, Thomas Wolf, and Lewis Tunstall. 2023. Lighteval: A lightweight framework for llm evaluation.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5427–5444, Online. Association for Computational Linguistics.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations (ICLR)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, pages 4411–4421. PMLR.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Shaoxiong Ji, Zihao Li, Jaakko Paavola, Indraneil Paul, Hengyu Luo, and Jörg Tiedemann. 2025. Massively multilingual adaptation of large language models using bilingual translation data. *Preprint*, arXiv:2506.00469.

Shaoxiong Ji, Zihao Li, Indraneil Paul, Jaakko Paavola, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Hengyu Luo, Hinrich Schütze, Jörg Tiedemann, and Barry Haddow. 2024a. EMMA-500: Enhancing massively multilingual adaptation of large language models. *Preprint*, arXiv:2409.17892.

Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024b. Can machine translation bridge multilingual pretraining and cross-lingual transfer learning? In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 2809–2818, Torino, Italy. ELRA and ICCL.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

Anmol Kabra, Yilun Yin, Albert Gong, Kamilė Stankevičiūtė, Dongyoung Go, Johann Lee, Katie Z. Luo, Carla P. Gomes, and Kilian Q. Weinberger. 2026. Learning from synthetic data improves multi-hop reasoning. In *The Fourteenth International Conference on Learning Representations*.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Natalia Fedorova, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakougna, Jessica Lundin, Kenton Murray, Masaaki Nagata, and 9 others. 2025. Preliminary ranking of WMT25 general machine translation systems. *Preprint*, arXiv:2508.14909.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Auguste Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot,

Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, and 33 others. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Advances in Neural Information Processing Systems*, volume 36, pages 67284–67296.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP '23)*.

Joona Kytöniemi, Jousia Piha, Akseli Reunamo, Fedor Vitiugin, Farrokh Mehryary, and Sampo Pyysalo. 2025. Fin-bench-v2: A unified and robust benchmark suite for evaluating finnish large language models. *Preprint*, arXiv:2512.13330.

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, Jorg Frohberg, and BigScience Workshop. 2022. The BigScience ROOTS corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445, Dublin, Ireland. Association for Computational Linguistics.

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, and 40 others. 2024a. DataComp-LM: In search of the next generation of training sets for language models. In *Advances in Neural Information Processing Systems*.

Zihao Li, Shaoxiong Ji, Hengyu Luo, and Jörg Tiedemann. 2025. Rethinking multilingual continual pretraining: Data mixing for adapting LLMs across languages and resources. *Preprint*, arXiv:2504.04152.

Zihao Li, Shaoxiong Ji, Timothee Mickus, Vincent Segonne, and Jörg Tiedemann. 2024b. A comparison of language modeling and translation as multilingual pretraining objectives. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15882–15894, Miami, Florida, USA. Association for Computational Linguistics.

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

Chen Cecilia Liu, Fajri Koto, Timothy Baldwin, and Iryna Gurevych. 2024. Are multilingual LLMs culturally-diverse reasoners? an investigation into multicultural proverbs and sayings. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2016–2039, Mexico City, Mexico. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno Miguel Guerreiro, Ricardo Rei, M. Amin Farajian, Mateusz Klimaszewski, Duarte M. Alves, José Pombal, Nicolas Boizard, Manuel Faysse, Pierre Colombo, François Yvon, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2025. EuroLLM-9B: Technical report. *CoRR*, abs/2506.04079.

Vladislav Mikhailov, Tita Enstad, David Samuel, Hans Christian Farsethås, Andrey Kutuzov, Erik Velldal, and Lilja Øvrelid. 2025. NorEval: A Norwegian language understanding and generation evaluation benchmark. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3495–3541, Vienna, Austria. Association for Computational Linguistics.

Tarek Naous, Michael J. Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16366–16393, Bangkok, Thailand. Association for Computational Linguistics.

Marianna Nezhurina, Joerg Franke, Taishi Nakamura, Timur Carstensen, Niccolò Ajroldi, Ville Komulainen, David Salinas, and Jenia Jitsev. 2025. Open-sci-ref-0.01: Open and reproducible reference baselines for language model and dataset comparison. *CoRR*, abs/2509.09009.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.

Desnes Nunes, Ricardo Primi, Ramon Pires, Roberto Lotufo, and Rodrigo Nogueira. 2023. Evaluating gpt-3.5 and gpt-4 models on brazilian university admission exams. *Preprint*, arXiv:2303.17003.

Stephan Oepen, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajič, Jindřich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkonen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. HPLT 3.0: Very large-scale multilingual resources for LLMs and MT. mono- and bilingual data, multilingual evaluation, and pre-trained models. *CoRR*, abs/2511.01066.

OpenAI. 2024. MMMLU: Multilingual massive multitask language understanding. https://huggingface.co/datasets/openai/MMMLU. Professionally translated test split of MMLU (Hendrycks et al., 2021) across 14 languages.

Siddhesh Pawar, Junyeong Park, Jiho Jin, Arnav Arora, Junho Myung, Srishti Yadav, Faiz Ghifari Haznitrama, Inhwa Song, Alice Oh, and Isabelle Augenstein. 2025. Survey of cultural awareness in language models: Text and beyond. *Computational Linguistics*, 51(3):907–1004.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin A. Raffel, Leandro von Werra, and Thomas Wolf. 2024. The FineWeb datasets: Decanting the web for the finest text data at scale. In *Advances in Neural Information Processing Systems*, volume 37, pages 30811–30849.

Guilherme Penedo, Hynek Kydlíček, Amir Hossein Kargaran, and Leandro von Werra. 2026. Finetranslations. https://huggingface.co/datasets/HuggingFaceFW/finetranslations. Dataset; non-English FineWeb2 documents translated into English with Gemma3 27B; directionally opposite of MultiSynt/MT.

Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro von Werra, and Thomas Wolf. 2025. FineWeb2: One pipeline to scale them all: Adapting pre-training data processing to every language. *CoRR*, abs/2506.20920.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: Outperforming curated corpora with web data only. In *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172.

Ramon Pires, Roseval Malaquias Junior, and Rodrigo Nogueira. 2025. Automatic legal writing evaluation of llms. In *Proceedings of the International Conference on Artificial Intelligence and Law (ICAIL)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Ricardo Rei, Nuno M. Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André F. T. Martins. 2026. Tower+: Bridging generality and translation specialization in multilingual LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. To appear.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a language in "multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7737–7746, Online. Association for Computational Linguistics.

Parker Riley, Aaron Gokaslan, and Noah A. Smith. 2021. Sometimes we want translationese. *CoRR*, abs/2104.07623.

Angelika Romanou, Negar Foroutan, Anna Sotnikova, Zeming Chen, Sree Harsha Nelaturu, Shivalika Singh, Rishabh Maheshwary, Micol Altomare, Mohamed A. Haggag, Imanol Schlag, Marzieh Fadaee, Sara Hooker, Antoine Bosselut, Snegha A, Alfonso Amayuelas, Azril Hafizi Amirudin, Viraat Aryabumi, Danylo Boiko, Michael Chang, and 40 others. 2025. INCLUDE: Evaluating multilingual language understanding with regional knowledge. In *International Conference on Learning Representations*.

Hélio Fonseca Sayama, Anderson Viçoso Araujo, and Eraldo Rezende Fernandes. 2019. Faquad: Reading comprehension dataset in the domain of brazilian higher education. In *2019 8th Brazilian Conference on Intelligent Systems (BRACIS)*, pages 443–448.

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

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Aaditya K. Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. 2024. Evaluation data contamination in LLMs: How do we measure it and when does it matter? *Preprint*, arXiv:2411.03923.

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, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Lrec*, volume 2012, pages 2214–2218.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann, Mikko Aulamo, Daria Bakshandaeva, Michele Boggia, Stig-Arne Grönroos, Tommi Nieminen, Alessandro Raganato, Yves Scherrer, Raúl Vázquez, and Sami Virpioja. 2024. Democratizing neural machine translation with OPUS-MT. *Language Resources and Evaluation*, 58:713–755.

Gorka Urbizu, Iñaki San Vicente, Xabier Saralegi, and Ander Corral. 2023. Not enough data to pre-train your language model? MT to the rescue! In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 3826–3836, Toronto, Canada. Association for Computational Linguistics.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2021. Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2203–2213, Online. Association for Computational Linguistics.

Dan John Velasco and Matthew Theodore Roque. 2025. Scaling, simplification, and adaptation: Lessons from pretraining on machine-translated text. In *Proceedings of the 5th Workshop on Multilingual Representation Learning*, pages 612–630, Suzhou, China. Association for Computational Linguistics.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *CoRR*, abs/2404.18796.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, David Ifeoluwa Adelani, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2025. Multilingual language model pretraining using machine-translated data. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28087–28107, Suzhou, China. Association for Computational Linguistics. TransWebLLM / TransWeb-Edu in 9 languages; EMNLP publication of arXiv:2502.13252.

Jiayi Wang, Yao Lu, Maurice Weber, Max Ryabinin, Yihong Chen, Raphael Tang, and Pontus Stenetorp. 2024a. Multilingual pretraining using a large corpus machine-translated from a single source language. *CoRR*, abs/2410.23956.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Alexander Arno Weber, Klaudia Thellmann, Jan Ebert, Nicolas Flores-Herr, Jens Lehmann, Michael Fromm, and Mehdi Ali. 2024. Investigating multilingual instruction-tuning: Do polyglot models demand for multilingual instructions? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20829–20855, Miami, Florida, USA. Association for Computational Linguistics.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet:

Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Minghao Wu, Weixuan Wang, Sinuo Liu, Huifeng Yin, Xintong Wang, Yu Zhao, Chenyang Lyu, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. The bitter lesson learned from 2,000+ multilingual benchmarks. *CoRR*, abs/2504.15521.

Weihao Xuan, Rui Yang, Heli Qi, Qingcheng Zeng, Yunze Xiao, Aosong Feng, Dairui Liu, Yun Xing, Junjue Wang, Fan Gao, Jinghui Lu, Yuang Jiang, Huitao Li, Xin Li, Kunyu Yu, Ruihai Dong, Shangding Gu, Yuekang Li, Xiaofei Xie, and 13 others. 2025. MMLU-ProX: A multilingual benchmark for advanced large language model evaluation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 1513–1532, Suzhou, China. Association for Computational Linguistics.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

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, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. PAWS-X: A Cross-lingual Adversarial Dataset for Paraphrase Identification. In *Proc. of EMNLP*.

Bingsheng Yao, Yufan Jiang, Diyi Yang, and Junjie Hu. 2024a. Data contamination can cross language barriers. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17864–17875, Miami, Florida, USA. Association for Computational Linguistics.

Bingsheng Yao, Xiaoyuan Yi Li, Niklas Muennighoff, Dragomir Radev, Pramod Viswanath, and Douwe Kiela. 2024b. Benchmarking machine translation with cultural awareness. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 13078–13096, Miami, Florida, USA. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

A Translation Pipeline Details

Orchestration toolkit (LLM pipeline). For the LLM-based translation runs of Section 3.3, we use an open-source Slurm-native orchestration toolkit[8] that scales inference workloads across hundreds to thousands of GPUs while maintaining linear throughput. The toolkit deploys multiple distributed inference servers in parallel: a single YAML configuration file specifies the cluster, inference-server, and data settings, and the toolkit generates the required Slurm job scripts, manages resource allocation, and coordinates distributed execution. It provides error handling with automatic retries and resumption, progress and throughput monitoring, built-in support for large-scale loading of Parquet input files, and a flexible inference server backend. We ran all inference with vLLM (Kwon et al., 2023).

LLM runs: per-model throughput. Across batched production runs (4 GPUs per inference server), TOWER+ 9B sustained approximately 1,000 output tokens per second per GPU, and TOWER+ 72B approximately 400 output tokens per second per GPU. The 9B variant therefore produces output tokens approximately $2.6\times$ faster than the 72B variant per GPU-second; the per-request advantage is wider (approximately $3.2\times$) because the 72B model supports a longer maximum sequence length, allowing us to feed longer source documents per request and yielding roughly 26% longer translated outputs on average.

LLM runs: aggregate compute, energy and carbon footprint. The LLM-based translation runs were carried out on the Leonardo (CINECA) supercomputer. Aggregate GPU consumption, is approximately 3.1 million A100 GPU-hours. We estimate the corresponding energy and carbon footprint following the methodology used

[8] `github.com/anonymous-org/distributed-inference`; URL to be replaced at camera-ready.

in CINECA project reporting: each A100 on Leonardo draws approximately 440 W under load, giving 1,379,638 kWh of consumed energy; applying the 2025 Italian electricity-grid carbon intensity of 358 g CO_2-eq/kWh yields approximately 493.9 t CO_2-eq. For reference, this energy budget corresponds to roughly 16% of the annual generation of an average wind turbine (3 MW nameplate capacity, 2,800 equivalent full-load hours per year, ≈8.4 GWh/year). These figures cover only the LLM-based translation pipeline; the corresponding energy and carbon estimates for the OPUS-MT/HPLT-MT runs on LUMI (AMD MI250x) are reported alongside the NMT shard statistics below.

NMT models: training data and inference setup. The OPUS-MT and HPLT-MT models we use are decoded with Marian-NMT (Junczys-Dowmunt et al., 2018) at beam size 4 on single-node jobs allocating two AMD MI250x GPUs each, on the LUMI supercomputer. All models are trained on OPUS data (Tiedemann, 2009, 2012) using the Tatoeba Translation Challenge compilation (Tiedemann, 2020), augmented with back-translations of Wikimedia content. None of the NMT models are fine-tuned for specific tasks; we use them off-the-shelf without further data normalisation or filtering. The final list of selected NMT models is published together with the dataset card. The sentence-splitter pre-processing step greedily merges segments shorter than 256 characters and splits sentences longer than 1024 characters before feeding them to the decoder; this length normalisation matters because NMT batching benefits sharply from uniform input length.

NMT runs: sharding, runtimes, energy and per-GPU power. The corpus was split into 100 shards per language of approximately 30 million sentence-aligned lines (roughly 700 million space-separated tokens each). Per-shard runtime averaged about 23 hours for transformer-base models (∼75M parameters) and about 30 hours for transformer-big models (∼230M parameters), with sustained GPU utilisation between 85% and 95%. Total energy consumption for one shard translated with transformer-base models is roughly 12 kWh, at an average of 260–270 W per GPU. For transformer-big models, total energy consumption is roughly double at 24–26 kWh, at an average of 360–370 W per GPU. The pattern suggests that further batching and inference optimisation could improve GPU utilisation for the smaller models and yield additional speed-ups.

B Reference Model Hyperparameters

Compute setup. All reference-model training and evaluation runs were carried out on the LUMI supercomputer (16 nodes of 4 AMD MI250x GPUs per node), with one exception: the Nemotron-CC decontamination control experiment of Appendix C was run on Leonardo (CINECA) with NVIDIA A100 nodes.

Hyperparameter	Value
Number of layers	24
Hidden size	2048
Attention heads	32
Initialization std	0.02
Sequence length	2048
Rotary base	10000
Optimizer	Adam
Adam β_1, β_2	0.9, 0.95
Adam ϵ	1.0×10^{-8}
Learning rate	3.0×10^{-4}
Minimum learning rate	0
Learning-rate decay	linear WSD
Warmup fraction	0.1
Cooldown fraction	0.2
Weight decay	0.05
Gradient clipping	1
Global batch size	1024
Micro batch size	4
Position embeddings	RoPE
Normalization	RMSNorm
Precision	bfloat16
Data parallel size	128

Table 1: Reference-model hyperparameters used in the main pre-training runs.

C Decontamination: Methodology and Full Tables

Motivation and prior work. Large web-derived corpora can overlap with evaluation benchmarks through exact duplicates, near-duplicates, paraphrases, or translated variants. The GPT-3 family established n-gram-based decontamination as a standard pre-training practice (Brown et al., 2020), and deduplication has been shown to reduce memorization and improve downstream model quality (Lee et al., 2022). Recent work emphasizes that contamination is difficult to define uniformly and that its impact depends on task, model, and training recipe (Singh et al., 2024). For multilingual settings the problem is sharper: contamination can cross language boundaries, and translated benchmark items can evade simple text-overlap detection while still inflating evaluation scores (Yao et al., 2024a). We therefore decontaminate the English source documents before translation, report over-

lap statistics per benchmark, and run a controlled training comparison to verify that removing flagged documents has no measurable downstream effect.

Released corpus. The released MultiSynt/MT corpus itself is not decontaminated; the decontamination described here is applied to the English source documents used as input to the controlled training comparison below.

Pipeline stages. We use the NeMo Curator decontamination pipeline, which constructs n-gram representations of downstream task datasets and removes any matching content from the training corpus in three stages. **Indexing:** each benchmark dataset is normalised and tokenised, and n-grams of size 8–13 are extracted from the test split (or the validation split, where no test split is available), following the text formatting in lm-evaluation-harness, to build a compact n-gram index. **Matching:** each source document is processed through the same normalisation/tokenisation/n-gram-extraction pipeline and its n-grams are compared against the benchmark index to identify overlaps. **Removal:** source documents whose matched n-grams occur fewer than 10 times in the corpus are treated as contamination signals and removed in full.

The indexed evaluation suite (see Table 2 for details) spans code, common sense reasoning, instruction following, language understanding, linguistic competence, mathematics, reasoning, translation, and world knowledge.

Match counts and rates. After applying the pipeline to the source Nemotron-CC documents used as input to translation, a total of 34,969 documents per language (0.0002%) are flagged as contaminated. Per-benchmark match counts and contamination rates are reported in Table 2. Contamination is unevenly distributed: the largest absolute contribution comes from BoolQ, followed at lower rates by HellaSwag, MMLU, and MTBench. Even where the absolute number of matches is large, the corresponding percentages remain very low, indicating that contamination is sparse relative to corpus size.

Controlled training experiment. To verify that the residual contamination has no measurable downstream effect, we train a 0.6B-parameter reference model on the chosen Nemotron-CC source slice with and without decontamination, under identical hyperparameters and a 100B-token budget, and evaluate 0-shot on the Finetasks suite (ARC-easy, ARC-challenge, PIQA, HellaSwag, OpenBookQA). The two learning curves track each other throughout training (Figure 7), with no consistent advantage for either setting.

D Per-Language Resource Statistics

Table 3 reports per-language Gemma-3 token counts for each MultiSynt/MT translation system and for the HPLT 3.0 native baseline, covering all 36 languages of the released corpus.[9] Figure 1 in Section 3.4 visualizes a 24-language representative subset drawn from this table; the table itself gives the full picture, including languages where Tower+ 9B or 72B was not run and languages without an HPLT 3.0 baseline. For consistency with Figure 1, the MultiSynt/MT-to-HPLT 3.0 ratio is computed against the per-language maximum across MultiSynt/MT systems (effectively OPUS-MT/HPLT-MT, whose output exceeds Tower+ on every language). Sums of the per-system columns recover the corpus totals reported in Section 3.4: 1.7T from Tower+ 9B (16 languages), 0.5T from Tower+ 72B (5 languages), 4.8T from OPUS-MT/HPLT-MT (36 languages), for approximately 7.1T tokens across all system outputs and approximately 4.8T unique target-language tokens.

E Human MT evaluation: per-language breakdown

Figure 8 shows the per-language breakdown of the human evaluation summarized in Section 3.2 and Figure 3.

Protocol. For each language, a single native speaker of the target language rated translations of the same 10 English source documents from each of the six candidate systems, blind to which system produced each translation; because the source documents are held constant across systems within a language, the within-language comparison among systems is paired. Annotators were instructed to focus primarily on target-language fluency and to additionally penalize specific model failures (hallucinations, repetitions, extra text not present in the source, and missing text); machine-translation

[9] Albanian is listed as `sqi` in MultiSynt/MT (the ISO 639-3 macrolanguage code, matching OPUS-MT/HPLT-MT's output language label) and as `als_Latn` in HPLT 3.0 (Tosk Albanian, the basis of Standard Albanian and the FLORES/NLLB convention for the language). The two refer to the same standard written language in practice.

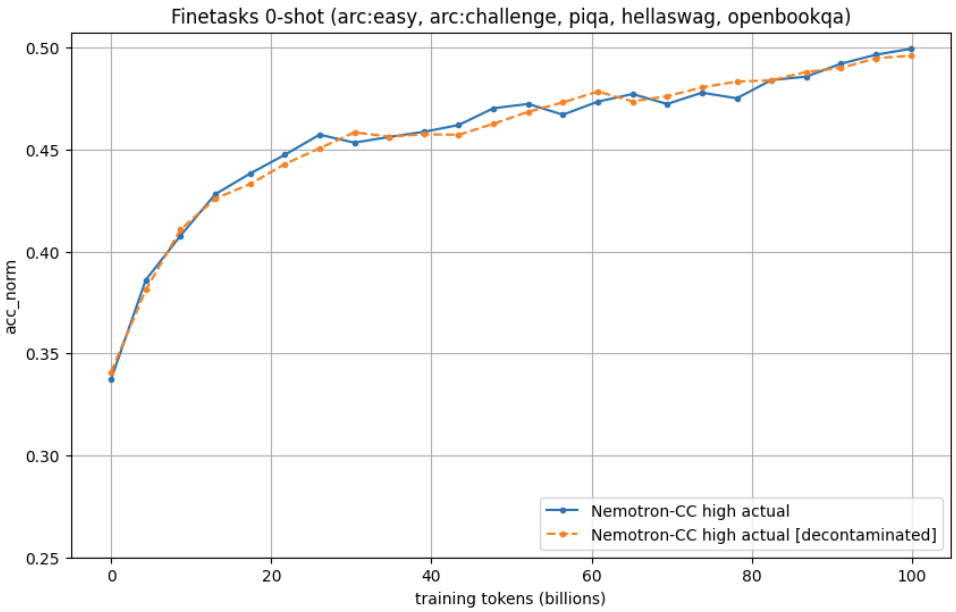

Figure 7: Controlled training experiment on the Nemotron-CC source data, with and without decontamination. A 0.6B-parameter reference model is trained for 100B tokens and evaluated 0-shot on the Finetasks suite (ARC-easy, ARC-challenge, PIQA, HellaSwag, OpenBookQA). Curves track each other throughout training, indicating no measurable downstream effect from decontamination.

outputs that were truncated due to inference length constraints were flagged in the guidelines and were exempted from penalty. Annotators recorded each judgment on a continuous slider mapped to a 0–100 scale, with the endpoints anchored as *very poor translation* (0) and *excellent translation* (100); the recorded scores are aggregated into five 20-point bins (0–20, 21–40, 41–60, 61–80, 81–100) and reported on a 1–5 Likert-style scale (1 = bad, 5 = excellent).

RESULTS. TOWER+ is rated highest in each of the seven evaluated languages; the relative ordering of the other systems varies by language. Czech and Finnish, the two morphologically richest languages in the evaluated set, receive the lowest scores across all systems, including TOWER+; on the higher-resource Romance languages (French, Italian, Spanish) the gap between TOWER+ and the general-purpose LLMs narrows substantially, while OPUS-MT remains a clear step below the LLM-based systems.

Comparison to WMT25. Recent WMT25 rankings (Kocmi et al., 2025) place TOWER+ 9B and 72B mid-pack against frontier proprietary systems (e.g., Gemini-2.5-Pro, GPT-4.1, DeepSeek-V3) and against the strongest constrained submissions. That evaluation, however, operates on paragraph-sized segments of approximately 100 words drawn from news, social media, speech transcripts, and literary text, rather than on the full web documents that constitute the Nemotron-CC HQ source distribution; several of the higher-ranked systems are also closed-weights or otherwise infeasible to deploy at the 100B-token source scale on the open compute available for this work.

F Fluency LLM-judge: sanity check

Figure 9 reports the sanity check referenced in Section 5.1: the off-the-shelf Qwen 2.5 series at five sizes from 0.5B to 14B parameters, evaluated by the same LLM-judge protocol against the monolingual HPLT 2.0 1.7B baseline. Winrate against the baseline rises monotonically with Qwen model size across all evaluated languages, confirming that the judge tracks generation quality and that the MT-variant ranking reported in Figure 5 reflects real differences rather than judge noise.

G Embedding-space coverage analysis

This appendix reports the embedding-space benchmark-alignment diagnostic summarized in Section 5.2. We use it as a descriptive check on benchmark-adjacent coverage rather than as a causal test of the gains in Section 4.2. The analysis measures which data source aligns more closely with benchmark regions in a shared embedding

Task	Category	Passages	Size (MB)	Matches	Contamination%
Humaneval	Code	328	10.6419	20	1.3e-05
MBPP	Code	878	4.8529	32	2.1e-05
ARC Challenge	Common sense	3,516	1.1415	109	7.0e-05
ARC Easy	Common sense	7,128	1.9972	157	1.0e-04
BoolQ	Common sense	3,270	2.0204	23275	1.5e-02
CommonsenseQA	Common sense	1,221	0.2635	0	0
COPA	Common sense	1,100	0.1422	0	0
HellaSwag	Common sense	10,042	10.6699	3034	2.0e-03
Lambada	Common sense	5,153	1.6316	28	1.8e-05
OpenBookQA	Common sense	2,000	0.3851	1	6.5e-07
PIQA	Common sense	5,512	1.4174	2	1.3e-06
XStoryCloze	Common sense	7,555	2.5407	0	0
XWinograd	Common sense	2,671	0.3142	31	2.0e-05
AlpacaEval	Instruction	1,593	0.1720	170	1.1e-04
do_not_answer	Instruction	939	3.8161	1	6.5e-07
IFEval	Instruction	541	0.2743	59	3.8e-05
m-ArenaHard	Instruction	6,000	3.3170	14	9.0e-06
MT-Bench-X	Instruction	400	0.2720	44	2.8e-05
MTBench	Instruction	1,388	1.8000	3127	2.0e-03
Multi-IFEval	Instruction	22,505	34.3774	2	1.3e-06
Toxigen	Instruction	940	0.3500	16	1.0e-05
PAWS-X	Lang. underst.	14,000	3.4758	136	8.8e-05
MultiBlimp	Linguistic	60,693	92.1014	135	8.7e-05
AIME 25	Math	30	0.0118	0	0
GSM8K	Math	1,319	0.6815	11	7.1e-06
Hendrycks_MATH	Math	5,003	2.0000	1065	6.9e-04
PolyMath	Math	3,000	0.6709	64	4.1e-05
XNLI	Reasoning	27,449	5.2555	31	2.0e-05
FLORES-200	Translation	37,444	10.3435	51	3.3e-05
Belebele	World Knowl.	29,700	27.5175	513	3.3e-04
GPQA	World Knowl.	546	4.0179	0	0
MMLU	World Knowl.	14,042	6.6514	2841	1.8e-03

Table 2: Benchmark evaluation suite grouped by capability category, showing approximate passage count and size of the splits used for decontamination, total number of contaminated documents containing at least one benchmark n-gram that failed the frequency threshold and ratio of contaminated documents to the full source corpus per benchmark.

space.

Embedding space. We map each document into an embedding space using `intfloat/multilingual-e5-large-instruct` (Wang et al., 2024b), producing one vector per document.

Figure 10 illustrates benchmark-adjacent coverage for Spanish ARC-Challenge – the benchmark that deviates most from parity in the main kNN figure (Figure 11) – by partitioning the joint pre-training embedding space into $k = 2{,}000$ clusters using k-means and assigning each benchmark example to its nearest cluster. The two distributions partially overlap, but MultiSynt/MT covers several regions that are essentially unpopulated by HPLT 2.0 (and vice versa); clusters that contain a benchmark item tend to lie in MultiSynt/MT-denser regions, with a mean MultiSynt/MT share of 0.72 versus 0.47 for clusters that do not. This is consistent with the per-language kNN ratio of 0.80 for Spanish ARC-Challenge in Figure 11.

kNN coverage. To quantify this effect across benchmarks and languages, for each benchmark example, we retrieve its top-k nearest neighbors from the union of native (HPLT 2.0) and synthetic (MultiSynt/MT) pre-training documents under cosine similarity, and report the average fraction $k_{\text{synthetic}}/k$ across benchmark items for $k = 20$. A value of 0.5 indicates equal local density from the two sources; values above 0.5 indicate denser synthetic coverage of benchmark-adjacent regions.

The detailed results are shown in Figure 11. Across all benchmarks, the mean values over the available languages indicate that the nearest documents are more often from MultiSynt/MT than from the native HPLT 2.0 data. HellaSwag is the only benchmark for which some languages show a higher fraction of adjacent native documents, though its mean MultiSynt/MT fraction is still 0.53. The other five benchmarks show a higher MultiSynt/MT fraction for all available languages. ARC-Challenge deviates most from the parity line, with

Code	Language	Tower+ 9B	Tower+ 72B	OPUS-MT	HPLT 3.0	MS/HPLT
mlt	Maltese	–	–	178.4	1.0	182×
gle	Irish	–	–	167.0	1.1	153×
nno	Norwegian Nynorsk	107.6	–	124.8	1.6	79×
glg	Galician	–	–	112.0	3.1	36×
eus	Basque	–	–	127.9	3.2	40×
mkd	Macedonian	–	–	154.6	5.9	26×
isl	Icelandic	136.8	–	152.1	6.1	25×
kat	Georgian	–	–	172.0	7.5	23×
sqi	Albanian	–	–	161.8	10.1	16×
srp	Serbian	–	–	148.1	10.4	14×
lav	Latvian	–	–	154.2	17.2	8.9×
est	Estonian	–	–	132.7	20.6	6.4×
slv	Slovenian	–	–	133.0	20.9	6.4×
cat	Catalan	–	–	131.3	22.5	5.8×
lit	Lithuanian	–	–	145.1	28.8	5.0×
bos	Bosnian	–	–	113.8	32.0	3.6×
hrv	Croatian	–	–	125.5	35.1	3.6×
slk	Slovak	–	–	135.9	40.2	3.4×
bul	Bulgarian	–	–	139.7	49.0	2.9×
nob	Norwegian Bokmål	103.6	–	119.7	51.2	2.3×
dan	Danish	106.1	–	122.0	62.7	1.9×
fin	Finnish	120.0	121.4	140.1	73.9	1.9×
ukr	Ukrainian	116.0	–	135.6	81.2	1.7×
hun	Hungarian	122.3	–	141.0	102.3	1.4×
ron	Romanian	112.8	–	131.8	102.5	1.3×
swe	Swedish	103.3	109.0	118.6	111.8	1.06×
ell	Greek	–	–	174.2	115.6	1.5×
ces	Czech	–	–	128.4	126.2	1.02×
tur	Turkish	–	–	115.9	150.0	0.77×
nld	Dutch	102.5	–	118.1	173.4	0.68×
pol	Polish	110.2	–	126.8	270.1	0.47×
por	Portuguese	93.5	–	107.5	318.9	0.34×
ita	Italian	97.2	102.3	111.9	335.5	0.33×
fra	French	102.2	–	115.0	585.0	0.20×
deu	German	98.7	100.2	112.1	609.3	0.18×
spa	Spanish	91.3	97.8	104.3	659.0	0.16×
Total		1724.0	530.7	4833.1	4245.0	1.14×

Table 3: Per-language Gemma-3 token counts (billions) for each MultiSynt/MT translation system and the HPLT 3.0 native baseline, for the 36 languages of MultiSynt/MT. Rows sorted by HPLT 3.0 token count ascending. The MS/HPLT column gives the ratio of the per-language MultiSynt/MT maximum (always OPUS-MT/HPLT-MT, whose output exceeds Tower+ on every language) to HPLT 3.0. The Total row sums each column; the Total MS/HPLT ratio is the sum of MultiSynt/MT maxima divided by the sum of HPLT 3.0 tokens. Tower+ 9B was run on 16 languages, Tower+ 72B on a 5-language high-resource subset, and OPUS-MT/HPLT-MT on all 36 languages.

a MultiSynt/MT fraction of 0.82. These results suggest that translated data increases the coverage of benchmark-adjacent regions, potentially bringing in content that is otherwise underrepresented in the native datasets. However, we only conducted a descriptive study and did not account for causality of any potential gains.

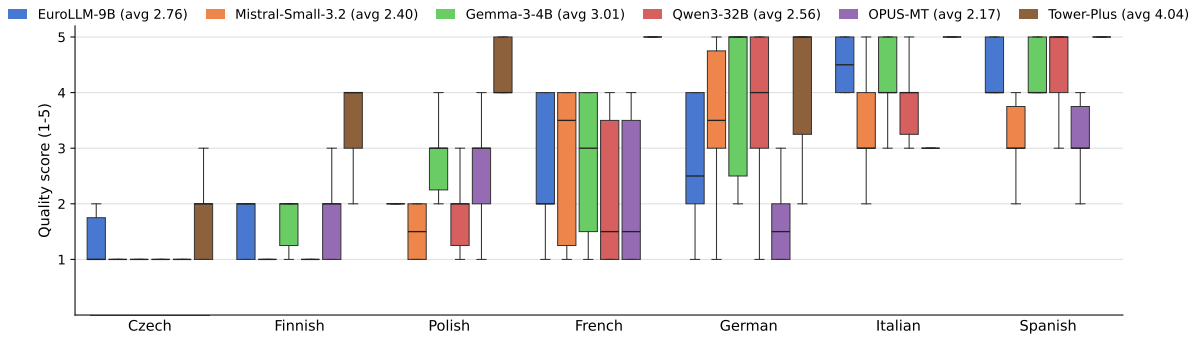

Figure 8: Per-language breakdown of the human evaluation of MT quality (cf. Figure 3). Languages are arranged along the x-axis by mean quality across all six systems (hardest on the left, easiest on the right); the per-system overall mean across the seven languages is shown next to each legend entry. TOWER+ is rated highest in every language; per-language deviations among the other systems are detailed in the bars.

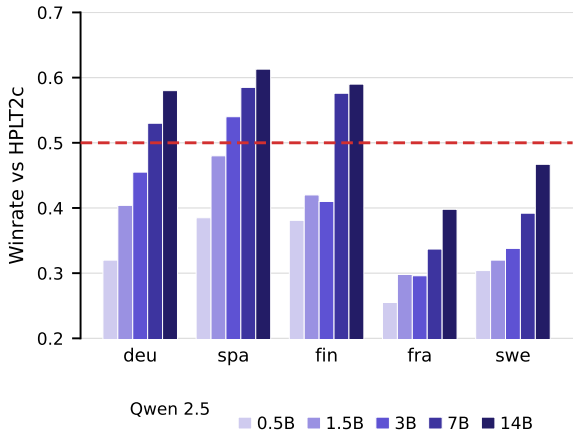

Figure 9: Sanity check: fluency LLM-judge winrate of the Qwen 2.5 model series (0.5B–14B parameters) against monolingual HPLT 2.0 1.7B baselines trained on native data. Winrate rises monotonically with model size across all evaluated languages, confirming that the judge tracks generation quality. The dashed line at 0.5 marks parity with the native baseline.

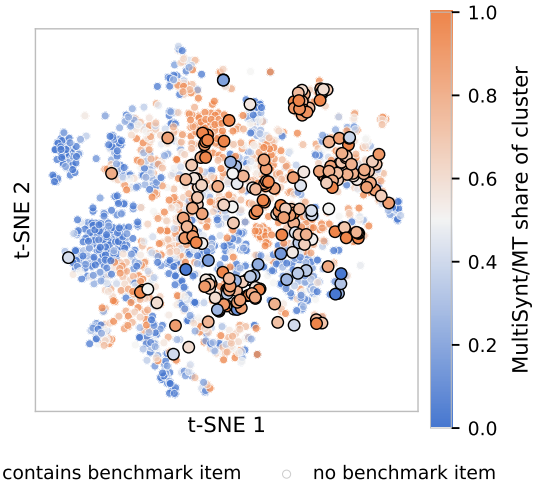

Figure 10: t-SNE projection of 2,000 k-means cluster centroids of the joint HPLT 2.0 + MultiSynt/MT pre-training embedding space for Spanish, colored by the per-cluster share of MultiSynt/MT documents (blue: HPLT 2.0-dominated; orange: MultiSynt/MT-dominated). Clusters that contain a Spanish ARC-Challenge benchmark item are outlined in black and tend to concentrate in MultiSynt/MT-denser regions of the space.

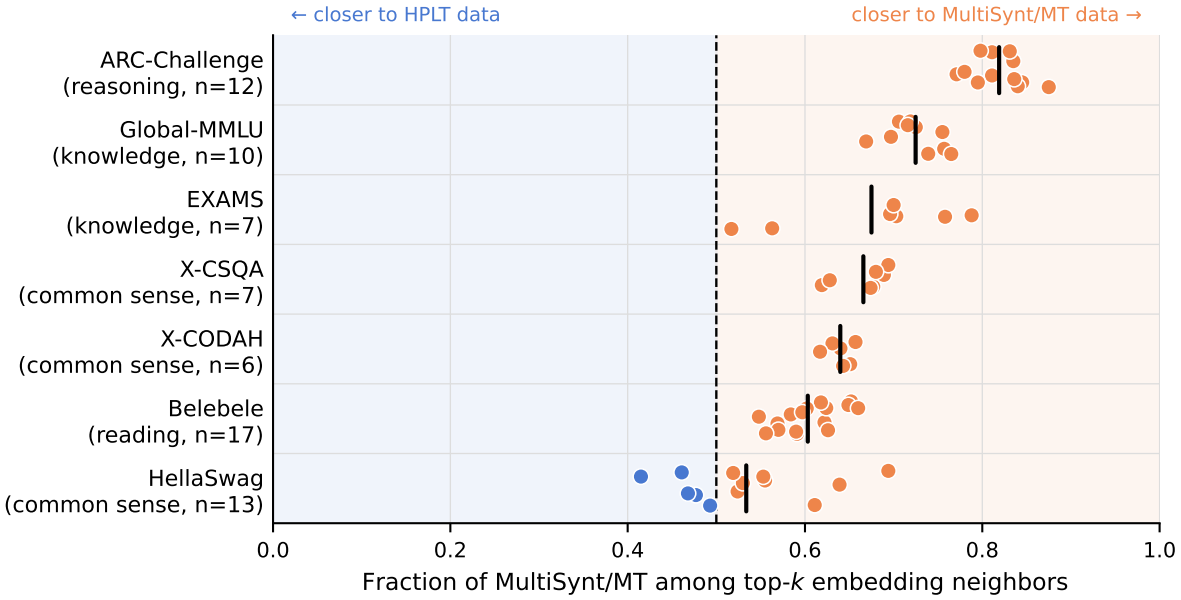

Figure 11: Embedding-space coverage analysis. For each benchmark, the fraction of MultiSynt/MT documents among the top-k embedding neighbors of benchmark items is shown for every available language (one dot per language; n in row labels). Black vertical ticks mark the per-benchmark mean across languages; the dashed line is parity at 0.5. Six of seven benchmarks sit fully right of parity: their benchmark items have more MultiSynt/MT documents than HPLT 2.0 documents among their nearest embedding neighbors. HellaSwag is the lone exception, with five Latin-script languages on the HPLT 2.0 side.