Continual Training, Reasoning, and Evaluation in Multilingual LLMs

Empirical Insights from Massively Multilingual Setting with 500+ Languages

Shaoxiong Ji

Principal Investigator, ELLIS Institute Finland
Assistant Professor, University of Turku
shaoxiong.ji@utu.fi

TruX Distinguished Seminar Series

June 29, 2026

https://olaresearch.org

Outline

Part I: Continual Pretraining

0. **MixCPT**
   Rethinking CPT Configurations
1. **MaLA Corpus**
   Corpus Curation and Mixing
2. **EMMA-500**
   Bilingual vs. Monolingual

Part II: Reasoning & Quality in MT

3. **RM4MT**
   Scaling Test-Time Compute
4. **Linguistic Reasoning**
   Synthetic Reasoning in MT
5. **FineOPUS**
   Massively Translation Corpus

Introduction: The Multilingual NLP Landscape

The Promise & Opportunities

- **Global Accessibility**: Enabling NLP technologies for the world's 7,000+ languages to bridge the digital divide.
- **Cross-Lingual Transfer**: Representations learned from high-resource languages transfer to benefit low-resource languages.
- **Generalization**: Multilingual training improves the model's overall performance.

The Hard Challenges

- **Data Scarcity**: Massive resource disparity between high- and low-resource web corpora.
- **Curse of Multilinguality**: Adding more languages to a fixed-size model degrades performance on individual languages due to capacity constraints.
- **Evaluation Bias**: Standard benchmarks are heavily skewed towards English-centric settings and tasks.

Introduction: Building Multilingual LLMs

Multilingual Pretraining

- **Scale & Curation**: Pretraining from scratch on massive web-scale corpora.

Adaptation & Fine-Tuning

- **Continual Pretraining (CPT)**: Adapting existing models to target languages.
- **Instruction Alignment**: Tuning on multilingual datasets for cross-lingual followability.

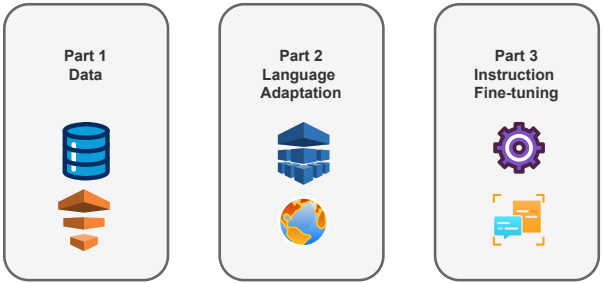


Figure: Three pillars of pretraining and adaptation in multilingual LLMs.

Rethinking Multilingual Continual Pretraining[1]

The Problem

- ⚖ **Disparity**: Pretrained LLMs favor high-resource languages, marginalizing underrepresented ones.
- ❓ **CPT Mixes**: Relative impact of monolingual, bilingual, and code mixes is poorly understood.
- ↔ **Transfer**: Are target languages altruistic, selfish, or stagnant in adaptation?

Experimental Setup

- 🧪 **36 configurations**: Evaluated Llama-3.1-8B, Llama-2-7B, and Viking-7B.
- 🔠 **30+ languages**: High-, medium-, and low-resource categories.
- 📊 **Benchmarks**: SIB-200 (classification), FLORES-200 (translation), MaLA (NLL).

[1] Z. Li, S. Ji, H. Luo, and J. Tiedemann. "Rethinking Multilingual Continual Pretraining: Data Mixing for Adapting LLMs Across Languages and Resources". In: *Conference on Language Modeling (COLM)*. 2025.

MixCPT: Data Composition & Training

Three Key Data Types Evaluated:

- **Monolingual**: Strictly filtered via GlotLID on MADLAD-400.
- **Bilingual**: Translation pairs from Lego-MT/NLLB formatted as: `[src_lang]: [src] [tgt_lang]: [tgt]`
- **Code**: Source files from The Stack (33% of mixture) to study scaffolding effects.

MixCPT Key Findings: Code Scaffolding & Categorization

Code as Scaffolding

- **Structural Help**: Adding code improves downstream classification & NLL across all levels.
- **Low-Resource Gain**: Tail languages benefit the most (with a slight generation trade-off).

Transfer Dynamics

- **↔ No Generalization**: The altruistic/selfish/stagnant division is highly configuration-dependent.

MixCPT: Scaffolding Effect of Code Data

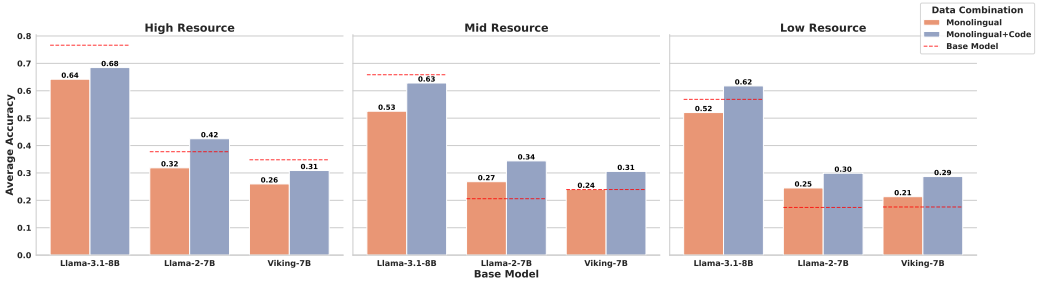


Figure: Scaffolding effect of code data during CPT on SIB-200 accuracy.

MixCPT Key Findings: The Language Mixing Issue

Bilingual CPT Generation Side-Effect

- Bilingual CPT improves classification accuracy but **degrades translation generation**.
- The model learns to pair languages in CPT, leading it to append wrong-language tokens or nonsense fragments at test time.
- Monolingual CPT maintains cleaner, more stable language-specific generation.

“我们现在有了非糖尿病的 4 个月小鼠，它们原本是患有糖尿病的。 Marathi: त्यांना मधुमेह होता. आता, आमचे चार महिन्याचे उंदर आहेत. ज्याला आधी मधुमेह झालेला होता. पण आता नाही. कारण यातील एक गोष्ट म्हणजे, साखर कमी करणे.....”

Figure: Examples of language mixing in bilingual CPT translation outputs.

MixCPT: Monolingual vs. Bilingual CPT Translation Comparison

Figure: Monolingual vs. Bilingual CPT translation comparison.

MaLA: Massively Multilingual Corpus and Model[2]

Motivation

To curate a clean massively multilingual corpus for LLM adaptation.

The MaLA Corpus

- **939 Languages**: 74B tokens
- **Data Mix**: Curated text, Aya instructions, and code (136B tokens).

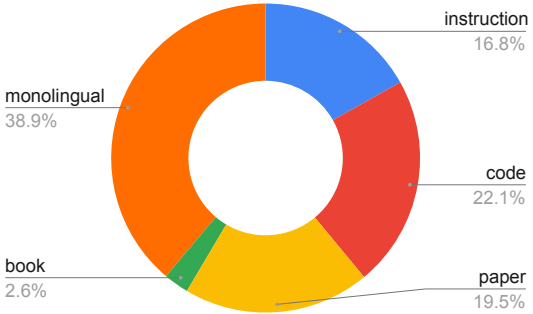


Figure: Monolingual data mixing distribution in MaLA data mix.

[2] S. Ji, Z. Li, J. Paavola, P. Lin, P. Chen, D. O'Brien, H. Luo, H. Schütze, J. Tiedemann, and B. Haddow. *EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models*. 2024. arXiv: 2409.17892 [cs.CL]. URL: https://arxiv.org/abs/2409.17892.

MaLA: Data Mix Statistics

Upsampling & Balancing:

- Aggressively upsamples low-resource languages and downsamples high-resource ones.
- Compares favorably against Glot500-c, MADLAD, CulturaX, and CC100 in language diversity.

EMMA-500 Model Training:

- Adapted from **Llama 2 7B** via CPT for 12,000 steps.
- Processed **200B Llama 2 tokens** using 256 Nvidia A100 GPUs (GPT-NeoX framework).
- Retains the original vocabulary.

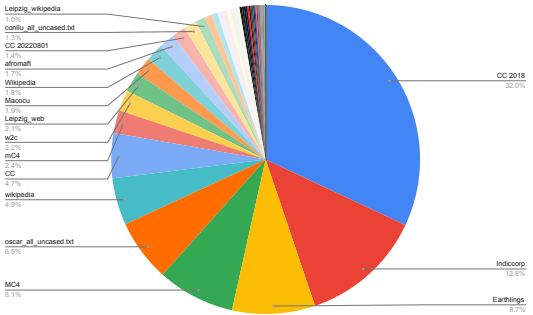
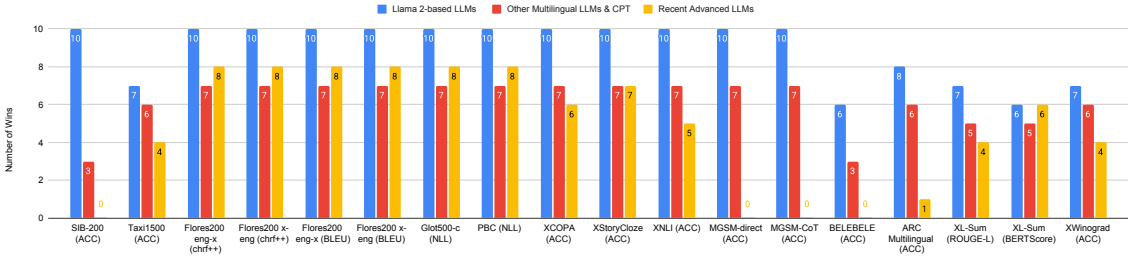


Figure: Distribution of data sources in the MaLA corpus mix.

EMMA-500: Downstream Results

Evaluation across 9 tasks and 15 benchmarks

- EMMA-500 consistently outperforms Llama 2-based CPT adaptions (LLaMAX, MaLA-500).
- Achieves competitive results compared to native multilingual LLMs (Qwen 1.5, Gemma 2).
- Successfully avoids catastrophic forgetting of coding and general reasoning.



Data-Centric Bilingual CPT at Scale[3]

Core Question

- **Transfer**: Does bilingual parallel data enhance cross-lingual transfer for low-resource languages?
- **Scale**: Massively evaluated across 500+ languages.

The MaLA Translation Corpus

- **426B Tokens**: 2,507 language pairs from OPUS, NLLB, and Tatoeba.
- **Format**: Pseudo-document concatenations.

[3] S. Ji, Z. Li, J. Paavola, H. Luo, and J. Tiedemann. "Data-Centric Continual Pre-training for 500+ Languages: A New Bilingual Translation Corpus and Multilingual Models". In: *Findings of ACL*. 2026.

EMMA-500 Suite & Training

EMMA-500 Suite:

- **Mono** (419B tokens, 25k steps)
- **Bi** (671B tokens, 40k steps)

Training Details:

- Trained on LUMI supercomputer using **256 AMD MI250x GPUs**.
- Global batch size 2048, sequence length 8192.
- Retains base model (Llama 3) vocabulary without expansion.

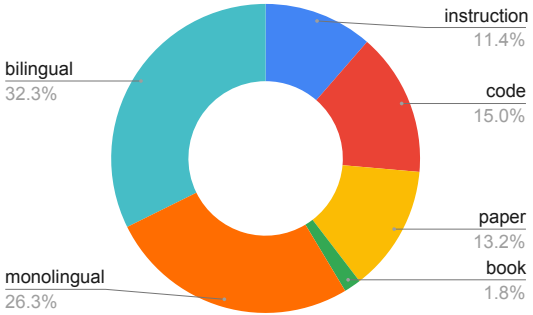
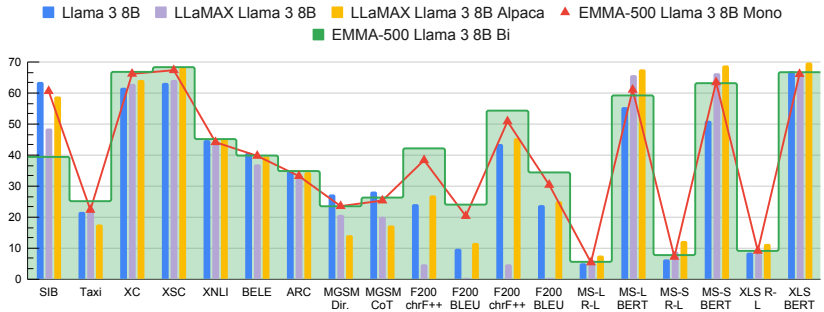


Figure: Data mix configuration for bilingual CPT.

EMMA-500 Suite: Monolingual vs. Bilingual CPT Results

Bilingual CPT Advantages

- Bilingual CPT consistently outperforms monolingual CPT, especially in translation (**+9% to +140% BLEU/chrF++**) and low-resource language classification.
- Parallel data significantly boosts downstream performance on tail languages.



Adaptation Resistance in Heavily Pretrained LLMs

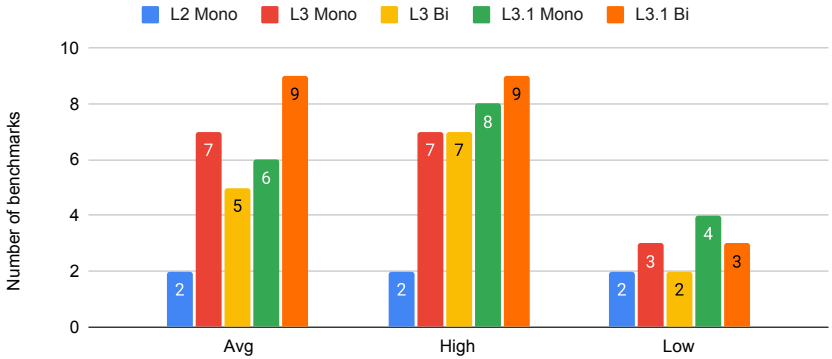


Figure: Model adaptability measured by the number of benchmarks on which CPT models are worse than the base model. CPT on LLaMA 2 (L2 Mono) shows a negligible BERTScore drop on MassiveSumm. More highly optimized models such as LLaMA 3 and 3.1 present greater challenges for effective continual pre-training compared to LLaMA 2, especially for high-resource languages, while the situation slightly eases for low-resource languages.

RM4MT: Motivation[4]

Motivation: Thinking in Translation

- **Test-Time Scaling (TTS)**: Chain-of-Thought (CoT) has shown breakthroughs in math/coding (e.g., DeepSeek-R1, OpenAI o-series).
- **Subjectivity of MT**: Unlike math/code with objective correctness, MT requires cultural adaptation, stylistic nuance, and context.
- **Inference Scaling**: Does more inference compute improve translation?

[4] Z. Li, S. Ji, and J. Tiedemann. "Test-Time Scaling of Reasoning Models for Machine Translation". In: *Proceedings of EACL*. 2026.

RM4MT: Scaling Workflows

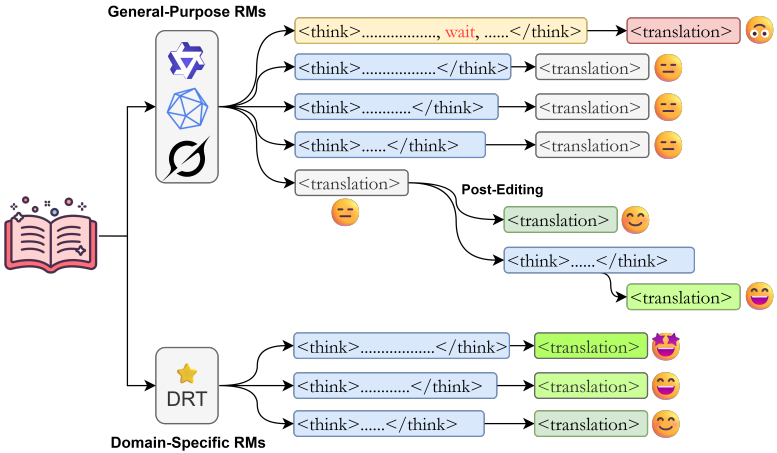


Figure: Direct Translation vs. Post-Editing workflows.

RM4MT: Models & Benchmarks

12 Reasoning Models (RMs)

- **General-Purpose RMs**:
  Qwen-3 (0.6B to 32B), Cogito (3B, 8B).
- **Domain-Specific RMs**:
  DRT (7B, 8B, 14B) (fine-tuned).
- **Proprietary RM**:
  Grok-3-Mini (low vs. high effort).

8 Translation Benchmarks

- **Literature**:
  WMT24-Literary, LitEval, MetaphorTrans.
- **Specialized/Reasoning**:
  CAMT, Commonsense-MT, Biomedical, RTT, RAGTrans.

RM4MT: Budget Forcing

Logits-Based Budget Forcing

- **Logits Processor**: Regulates reasoning inside the `<think>` block.
- **Soft Closure**: Upweights newline and `</think>` near 95% of budget.
- **Forced Termination**: Emits final tokens deterministically at limit.

RM4MT: Post-Editing & Extrapolation

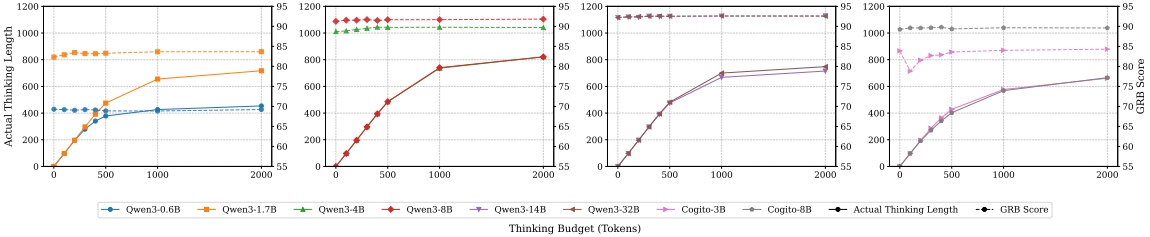
Post-Editing & Extrapolation

- ⏭ **Forced Extrapolation**: Overrides natural stop with a ‘wait‘ token to extend reasoning.
- 🔁 **Post-Editing (PE)**: Two-stage translation comparing draft quality feedback (QS vs. No QS).

RM4MT: General-Purpose RMs & The Reasoning Ceiling

Diminishing Gains & Reasoning Ceiling

- General RMs (out-of-the-box Qwen-3, Cogito) exhibit flat performance plateaus after minimal budget.
- Fail to utilize large budgets; actual thinking tokens autonomously saturate around **600 tokens**.
- More compute/steps alone does not improve translation without domain/task knowledge.

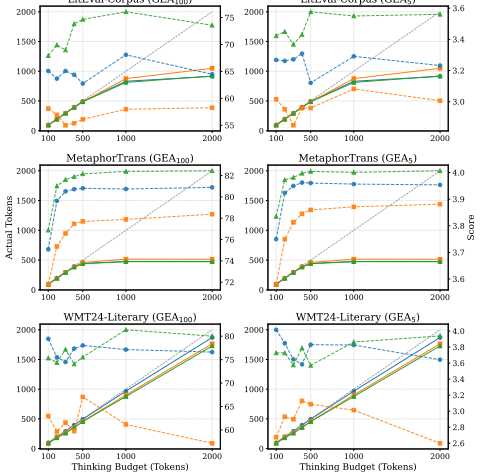


Average GRB scores and actual thinking tokens vs. budget.

RM4MT: Fine-Tuning Unlocks Test-Time Scaling

In-Domain: Monotonic Scaling

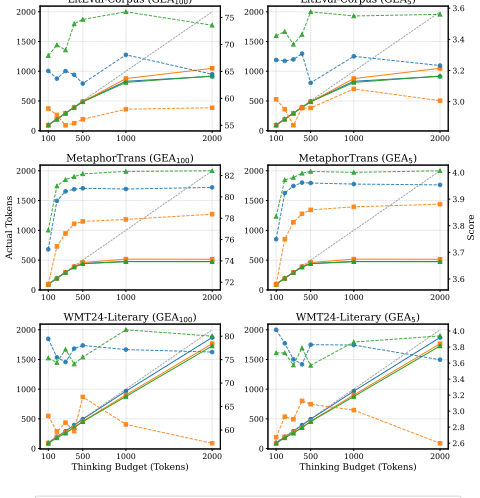
- DRT models (fine-tuned on MetaphorTrans) show clear quality gains (GEA score) with larger budgets.
- Performance improves steadily up to the model's natural stopping point (500 tokens).



RM4MT: Fine-Tuning Unlocks Test-Time Scaling

Emergent Alignment & Out-of-Domain Disconnect

- In-domain: model naturally stops thinking as quality plateaus (efficient deliberation alignment).
- Out-of-domain (WMT24-Literary): models use the extra budget ("think longer") but performance remains erratic



RM4MT: Forced Extrapolation Degrades Performance

Forced Extrapolation via ‘wait‘

- Override </think> with a single ‘wait‘ token to force continued reasoning.
- Successfully extends the thinking chain (by **100–200 tokens** on average).
- Consistently degrades quality: **55 out of 64 metric scores** across 8 models dropped.

Model	Budget	COMET	GRB
Qwen3-1.7B	1000	**0.769** / 0.748	**83.6** / 83.5
Qwen3-4B	1000	**0.791** / 0.774	**89.8** / 89.6
Qwen3-8B	1000	**0.798** / 0.787	91.7 / **91.8**
Qwen3-32B	1000	**0.803** / 0.783	**92.6** / 92.6
Cogito-3B	1000	**0.707** / 0.706	**84.0** / 83.8
Cogito-8B	1000	**0.768** / 0.766	**89.7** / 89.4

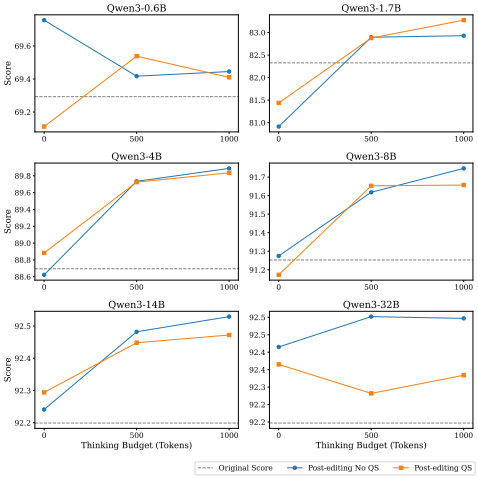
Table: Before vs. After forced extrapolation (1000 budget).

Intrinsic stopping points are meaningful; forcing more reasoning introduces noise and repetition.

RM4MT: Test-Time Scaling is Effective for Post-Editing

Compute Scaling in Post-Editing

- **Two-Stage PE**: Stage 1 generates draft (zero CoT) → Stage 2 refines draft using CoT budget (0, 500, or 1000 tokens).
- **Highly Effective**: Unlike direct translation, scaling compute in PE reliably boosts quality.
- **Mid-sized Models (1.7B–14B)**: 500/1000-token budgets turn PE into a reliably beneficial process.



Post-editing GRB scores across Qwen-3 models.

Reasoning over Grammar for Low-Resource MT[5]

The Problem

- ⚠ **Data Sparsity**: Extreme low-resource translation is limited by scarce parallel data.
- ❓ **Syntactic Bottleneck**: LLMs fail to apply grammatical rules effectively during in-context learning.

Core Idea: Syntax CoT

- 🧠 **Linguistic Traces**: Structured intermediate steps of syntactic/grammatical analysis to guide translation.
- 🔀 **Compositionality**: Progressively derive meaning from small lexical units.

[5] R. Pei, Y. Liu, S. Pyysalo, H. Schütze, and S. Ji. "Reasoning over Grammar: Can Synthetic Linguistic Reasoning Traces Enhance Low-Resource Machine Translation?" In: *arXiv preprint arXiv:2606.03782* (2026).

Pipeline: Synthetic Trace Generation

Trace Structure Generation:

1. **Bottom-up traversal** of the Universal Dependencies (UD) tree.
2. Steps ordered via **post-order traversal** (child nodes before parents), aligning with semantic compositionality.
3. For each step: verbalize POS tags, lemma, features, triggered grammar rules, head-dependent relationships, and insert `[Lexical]` and `[Phrasal]` placeholders.
4. Fill placeholders via LLMs using gold references and dictionary cues.

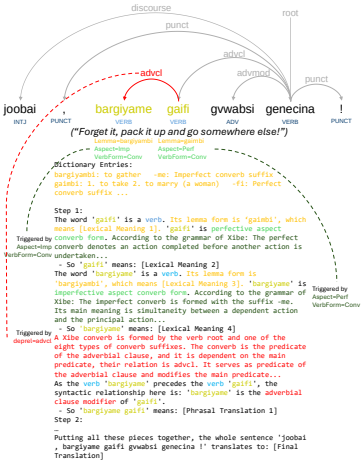


Figure: Visual traversal & trace structure from a Xibe UD tree.

Linguistic Reasoning: Experimental Settings

ICL & SFT Settings

- **In-Context Learning (ICL)**: Prompted with dictionaries and rules; resolves placeholder traces sequentially.
- **Supervised Fine-Tuning (SFT)**: Fine-tunes models to generate trace in `<think>` and answer in `<answer>`.

Reinforcement Fine-Tuning (RFT)

- **GRPO Optimization**: Continues training SFT adapters with multi-criteria rewards:
  - **Translation** (0.75): chrF, BLEU, SBERT.
  - **Format** (0.10): Tag layout.
  - **Process** (0.15): Phrasal recall.

Results: ICL vs. Fine-Tuning (Chintang ctn)

Setting & Model	BLEU	chrF	SBERT
gemma-4-E4B-it			
Baseline (Dict+Rules)	1.39	17.86	53.72
+ ICL Reasoning Trace	6.96 (+5.57)	29.75 (+11.89)	65.04 (+11.32)
+ SFT w/ reasoning	3.39 (+2.00)	24.25 (+6.39)	50.30 (-3.42)
+ RFT (GRPO)	3.41 (+0.02)	23.01 (-1.24)	48.29 (-2.01)
Qwen3-4B-Thinking			
Baseline (Dict+Rules)	0.22	5.86	44.03
+ ICL Reasoning Trace	1.11 (+0.89)	13.12 (+7.26)	63.77 (+19.74)
+ SFT w/ reasoning	3.23 (+3.01)	24.22 (+18.36)	53.12 (+9.09)
+ RFT (GRPO)	3.48 (+0.25)	24.15 (-0.07)	52.23 (-0.89)

Table: Translation scores on Chintang. Deltas for RFT are relative to SFT; other deltas are relative to pretrained baselines.

Key Findings & Discussion

- **ICL is highly effective**: Gold UD traces act as clean, reliable inference guidance.
- **Fine-tuning bottleneck**: Models learn format readily but generate **erroneous grammar analyses**.
- **Error Propagation**: Incorrect syntax parsing in `<think>` degrades final translation.

The Mission: FineOPUS[6]

Mission & Scale

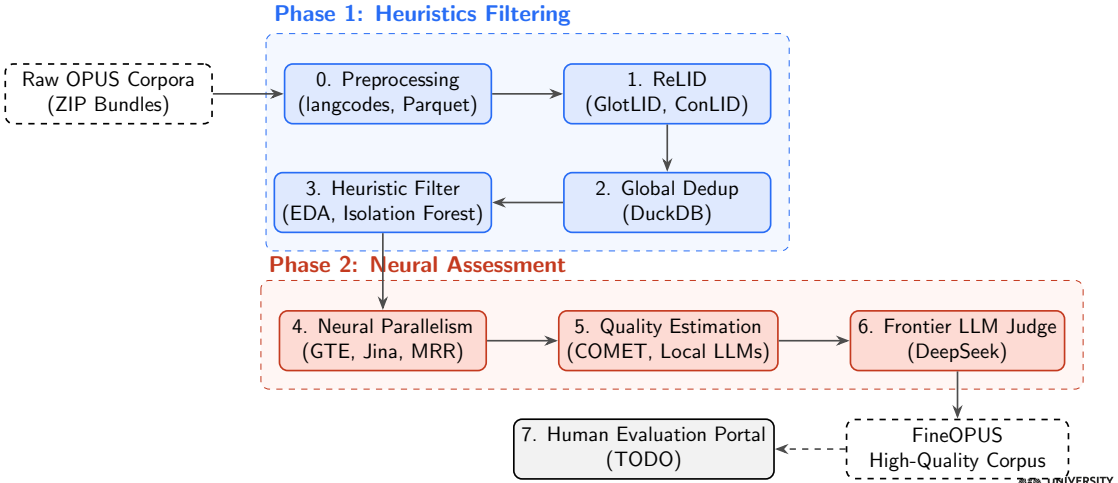
- **Central Mission**: Refine the OPUS dataset to produce high-quality parallel data.
- **Massive Scale**: Supporting **500+ languages** and 9,000+ language pairs.

Philosophy & Fairness

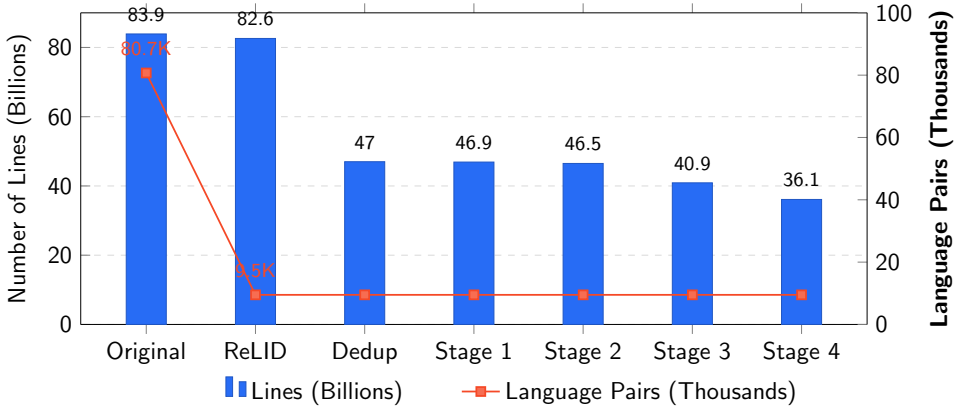
- **FineWeb Philosophy**: Adapting web-refinement approaches to parallel data.
- **Fair Representation**: Prioritizing high-fidelity data for tail languages.

[6] MaLA-LM Team. *FineOPUS: A Massively Multilingual Translation Corpus and Pipeline for Refining Parallel Texts in Many Languages*. 2025. URL: https://mala-lm.github.io/FineOPUS.html.

The FineOPUS Pipeline



Data Filtering Pipeline: Quality over Quantity



The Scale of FineOPUS

"Final" High-Quality Dataset Yield:

1.65 Trillion

Tokens

Token Metrics

- **Exact Count**: 1,654,022,968,970
- **Tokenizer**: DeepSeek V4

Corpus Metrics

- **Final Lines**: 36.1 Billion
- **Language Pairs**: 9,532

Future Work: Synthetic Augmentation & Modeling

Data Refinement

- **LLM Refinement**: Rewriting low-quality pairs via LLM contextual reasoning.
- **Back-Translation**: Translating monolingual data through a pivot language.

Quality & Validation

- **Cycle Consistency**: Strict semantic filter by back-and-forth translation.
- **Empirical Validation**: Training and evaluating LLMs/MT on the refined corpus.

Concluding Takeaways

Pretraining & Scale

- **Data Curation**: High-quality datasets (FineOPUS, MaLA) are vital for tail languages.
- **CPT Mixtures**: Diverse data mix helps; Code acts as scaffolding, while bilingual data enhances translation but requires careful tuning.

Reasoning & Quality

- **Test-Time Compute**: Effective for Post-Editing, but general reasoning hits a plateau without fine-tuning.
- **Syntax CoT**: Improves low-resource MT, but is constrained in different settings.

https://olaresearch.org/

Join us on Discord

Thank You!

Collaborators & Partners:

ELLIS Institute Finland University of Turku University of Helsinki LMU Munich University of Edinburgh

Zihao Li, Jaakko Paavola, Hengyu Luo, Renhao Pei, Abdelaziz M. A. Ibrahim, Yihong Liu, Sampo Pyysalo, Peiqin Lin, Pinzhen Chen, Dayyán O'Brien, Barry Haddow, Hinrich Schütze, Jörg Tiedemann

References I

Ji, S., Z. Li, J. Paavola, P. Lin, P. Chen, D. O'Brien, H. Luo, H. Schütze, J. Tiedemann, and B. Haddow. *EMMA-500: Enhancing Massively Multilingual Adaptation of Large Language Models*. 2024. arXiv: 2409.17892 [cs.CL]. URL: https://arxiv.org/abs/2409.17892.

Ji, S., Z. Li, J. Paavola, H. Luo, and J. Tiedemann. "Data-Centric Continual Pre-training for 500+ Languages: A New Bilingual Translation Corpus and Multilingual Models". In: *Findings of ACL*. 2026.

Li, Z., S. Ji, H. Luo, and J. Tiedemann. "Rethinking Multilingual Continual Pretraining: Data Mixing for Adapting LLMs Across Languages and Resources". In: *Conference on Language Modeling (COLM)*. 2025.

Li, Z., S. Ji, and J. Tiedemann. "Test-Time Scaling of Reasoning Models for Machine Translation". In: *Proceedings of EACL*. 2026.

MaLA-LM Team. *FineOPUS: A Massively Multilingual Translation Corpus and Pipeline for Refining Parallel Texts in Many Languages*. 2025. URL: https://mala-lm.github.io/FineOPUS.html.

Pei, R., Y. Liu, S. Pyysalo, H. Schütze, and S. Ji. "Reasoning over Grammar: Can Synthetic Linguistic Reasoning Traces Enhance Low-Resource Machine Translation?" In: *arXiv preprint arXiv:2606.03782* (2026).