

The Croatian Dataset Seed Submission to the WMT26 Open Language Data Initiative Shared Task

Dragana Zundanović Hrvoje Leventić Dragana Božić Lenard Krešimir Romić

Faculty of Electrical Engineering, Computer Science and Information Technology Osijek

J. J. Strossmayer University of Osijek

{dragana.zundanovic, hrvoje.leventic}@ferit.hr

Abstract

We present a new English–Croatian parallel corpus for the WMT26 Open Language Data Initiative (OLDI) Seed shared task, consisting of 6,193 sentence pairs. To ensure high quality, the resource was produced via machine translation followed by a two-stage verification process (post-editing and independent revision) conducted by professional translators. We validate the corpus by fine-tuning NLLB-200, TranslateGemma, and Gemma-4-31B. The dataset improves every dedicated translation model in both directions and the strongest open LLMs in the Croatian→English direction. Through a training-data ablation, we demonstrate that the benefit of human verification increases with model capacity and is visible only to neural metrics, not to surface-overlap metrics such as chrF++. Finally, we analyze the quality ceiling for the strongest models, identifying a trade-off between naturalness and surface overlap in the English→Croatian direction.

1 Introduction

Machine translation has undergone substantial development in recent years. However, these improvements have predominantly benefited high-resource languages, while low- and medium-resourced languages progress at a slower pace. Simultaneously, the number of available parallel datasets also differs across languages, with low- and medium-resourced languages at a disadvantage. Croatian is variously classified as a low- or a medium-resourced language.

The Croatian dataset we submit to the WMT26 Open Language Data Initiative (OLDI) shared task consists of 6,193 sentence pairs covering a broad range of domains of human knowledge. The Croatian translations were generated using machine translation, which was subsequently followed by a two-stage verification process. To ensure terminological consistency, grammatical, semantic and lex-

ical correctness, machine translation post-editing and revision were performed.

The contributions of this paper are threefold: (1) a human-validated, domain-diverse Croatian dataset of 6,193 sentence pairs for the WMT26 OLDI shared task; (2) a documented two-stage quality-assurance workflow combining machine translation post-editing and independent revision; and (3) experimental results showing that the dataset improves every dedicated translation model in both directions and the strongest open LLMs in the Croatian→English direction. This includes an ablation study that specifically quantifies how much the human verification stage contributes to the final model quality, and an analysis of when fine-tuning on such data stops helping.

2 Related Work

Modern language technologies do not preserve language diversity and promote multilingualism equally across all languages. Many high-resource languages, such as English, Spanish or German, benefit from extensively developed parallel web corpora. On the other hand, low- and medium-resourced languages suffer from data scarcity on the Internet. Although the situation is improving to some extent, the gap between such languages still exists and needs to be addressed (Joshi et al., 2020). Croatian is one such language. This language has been classified as a mid-resource language (Thakkar et al., 2024) in some works, while others argue that it belongs to the group of low-resourced languages (Bago et al., 2022; Koloski et al., 2022). The Seed corpus itself was introduced by Maillard et al. (2023), who showed that even a few thousand professionally translated sentence pairs measurably improve machine translation for low-resource languages. The Open Language Data Initiative has collected such datasets through a WMT shared task since 2024 (Dale et al.,

2025).

Croatian is currently spoken by approximately 5 million people and has been an official language of the EU since 2013 (Tadić, 2023). It belongs to the group of South Slavic languages. As such, it is a morphologically rich language that differs substantially from English because of its inflectional system, which includes grammatical case, number, person and verbal aspect. These differences may create a challenge for quality machine translation (Popović et al., 2020).

While parallel English-Croatian corpora, such as ParaCrawl (Bañón et al., 2020), hrenWaC (Ljubešić et al., 2016), MARCELL (Váradi et al., 2020) and MaCoCu (Bañón et al., 2023), do exist, detailed, domain-specific corpora remain scarce and thus require resource-intensive collection efforts. Besides that, the existence of large-scale web-crawled corpora does not guarantee uniformly high-quality data. Such resources often contain ungrammatical output, unnatural phrasing and literal translations (van Noord et al., 2025). Human translation and evaluation play a key role in ensuring high-quality resources for machine translation model training.

Furthermore, correct translation into Croatian requires context beyond the single sentence to keep gender, case, and verb forms consistent across a text, which sentence-level benchmarks cannot fully capture (Oliver and Álvarez-Vidal, 2026). In addition to domain-specific datasets, broad-coverage datasets spanning a wide variety of domains, such as Wikipedia articles, also represent good evaluation benchmarks (Goyal et al., 2022). As machine translation has seen significant improvements, translators and other language experts often engage in machine translation post-editing (Brockmann et al., 2022). Additionally, machine-translated data that undergo human verification and post-editing can also serve as a valuable resource for machine translation research (Popović and Arčan, 2016).

3 Data Creation

We compiled the Croatian portion of the OLDI Seed dataset, consisting of 6,193 sentence pairs and covering a wide range of domains, such as history, biology, arts, physics, architecture and many other areas of human knowledge, present on Wikipedia. This process aimed to create a high-quality English-Croatian parallel corpus that can support the training and development of machine translation mod-

els. Instead of relying solely on machine translation output, we subjected the translations to two steps of human verification:

1. Machine translation post-editing
2. Revision

This combined the speed of machine translation with human linguistic expertise.

The Croatian translation of the OLDI Seed dataset was curated using Matecat, a free and open-source computer-assisted translation tool (CAT tool) developed as part of a project funded by the European Commission. The tool’s integrated machine translation engine, ModernMT Lite, supports over 200 languages, including Croatian (Matecat, 2026).

As mentioned previously, the 6,193 English sentences were translated into Croatian using Matecat’s integrated machine translation system. The outputs served as the initial version of the Croatian dataset. However, pure machine translation still contains errors, and human involvement is necessary to bridge the gap between machine-generated content and natural language suitable for real-world use (Marashian et al., 2025). We had each sentence manually post-edited by a native speaker of Croatian with formal education in translation studies. This stage included comparing the target text with the source text to detect potential errors. The error types ranged from mistranslations, inadequate lexical choices in cases of polysemy and literal translations, omissions or additions of elements in the sentences, to grammatical and typographical errors or lexemes from other languages within the same language family (e.g., Serbian) and terminological inconsistencies. Representative examples of problematic machine translation outputs, along with their corrections, are provided in Table 1 and briefly characterized below:

- 57: unnatural use of passive voice in Croatian (*viđena je*), prepositional error (*do*)
- 112: incorrect gender pronoun (masculine instead of feminine; context from previous segments indicated the correct gender)
- 5411: literal translations that were not suitable in the context of philosophy of mathematics (*structures* as *građevine* instead of *strukture*; *concrete* as *betonski* instead of *konkretni*)

Throughout this stage, several additional tools were consulted to ensure natural and appropriate vocabulary choices. These included the bilingual corpus Glosbe, the Dictionary of Medicine (Školska knjiga), the English-Croatian Economic Dictionary (Mate), the Croatian-English and English-Croatian Dictionary of Idioms (Ljevak), the Pravopis.hr website and Struna (an online terminology database for specific fields developed by the Croatian Institute for Language). Particular attention was paid to the correct terminology for each of the domains included in the dataset, grammatical accuracy regarding cases, gender, number, as well as agreement and word order. Another important aspect was the translation of idiomatic expressions, which often display a different form in the respective languages.

Following the machine translation post-editing (MTPE) phase, the entire Croatian dataset was subsequently independently revised by a second editor, also a native speaker of Croatian with an academic background in linguistics and translation studies. The aim of this second phase was to identify any errors that remained after the post-editing stage, thus adding another layer of quality assurance. The reviser compared the source and target segments and made corrections where necessary, thus ensuring consistency, grammatical correctness, terminology appropriateness, semantic adequacy and fluency in Croatian. This second phase of linguistic validation was also important since it contributed to stronger uniformity, especially given the broad scope of domains and topics across the dataset.

Appendix A shows a verbatim excerpt of the released data.

4 Experimental Validation

To confirm the practical value of the dataset, we performed a series of fine-tuning and evaluation experiments spanning three model families, along with corpus-level quality estimation and a training-data ablation that isolates the contribution of human verification.

4.1 Experimental setup

Models. We fine-tuned three model families to assess the transferability of the results among different model types:

- NLLB-200 (NLLB Team et al., 2022), representing the dedicated multilingual MT systems that seed data primarily targets;

- TranslateGemma 4B and 27B (Finkelstein et al., 2026), open translation-specialized LLMs for which Croatian is among the 55 languages evaluated in the accompanying technical report;
- Gemma-4-31B-it (Gemma Team, 2026), a strong general-purpose instruction-tuned LLM, which showed strong Croatian fluency in our internal evaluations.

In addition to the fine-tuned models, we evaluate DeepSeek-V4-Flash (DeepSeek-AI, 2026), a 284B-parameter mixture-of-experts model with 13B active parameters, as an external zero-shot reference point; it is not fine-tuned.

NLLB models are fully fine-tuned, while the LLMs are fine-tuned with LoRA (Hu et al., 2022; $r=32$, $\alpha=64$) applied to the language-model projections, with prompt tokens masked from the loss. TranslateGemma is trained and evaluated using its native prompt format (Appendix B).

The native prompt formats of NLLB and TranslateGemma accept only the sentence to be translated, not wider context.

Fine-tuning and evaluating Gemma-4-31B, as a general-purpose instruction-tuned model, allows for more freedom in prompt format selection. Thus, Gemma-4-31B is trained and evaluated in a document-context prompt in which the sentence to be translated is stated first, followed by its full same-article source-side context (its best zero-shot configuration, selected in a pilot).

For TranslateGemma, we additionally report a full fine-tune of the 4B model and, for both sizes, a single bidirectional model trained on both directions at once; for Gemma-4-31B, the reported model is bidirectional. Checkpoint selection uses the FLORES+ dev set exclusively (chrF++ for NLLB, development loss for the LLMs, with early stopping); dev data is never used for reporting.

All fine-tuning was performed on RTX 6000 Pro Blackwell 96 GB GPUs, where full fine-tuning of 27B and 31B models was not feasible and only LoRA adaptation was performed. Learning rates were swept for the LLMs (Table 5).

Evaluation. We report chrF++ (Popović, 2017) as the primary surface metric and spBLEU as secondary, computed with sacreBLEU (Post, 2018), along with two architecturally independent neural metrics: COMET (wmt22-comet-da; Rei et al., 2022) and MetricX-24-XXL (Juraska et al., 2024).

Sentence ID	Source text	Machine translation	Target text
57	Buster Keaton’s wife Eleanor also was seen in the series (notably as Juliet to Buster’s Romeo in a little-theater vignette).	U seriji je viđena i supruga Bustera Keatona Eleanor (osobito kao Julija do Busterovog Romea u maloj kazališnoj vinjeti).	U seriji se pojavljuje i supruga Bustera Keatona, Eleanor (posebno kao Julija Busterovu Romeu u maloj kazališnoj vinjeti).
112	Davis died from a heart attack in 1969, two years before Lloyd’s death.	Davis je umro od srčanog udara 1969. godine, dvije godine prije Lloydove smrti.	Davis je umrla od srčanog udara 1969. godine, dvije godine prije Lloydove smrti.
5411	Structures are held to exist inas-much as some concrete system ex-emplifies them.	Smatra se da građevine postoje u mjeri u kojoj im neki betonski sus-tav daje primjer.	Smatra se da strukture postoje u mjeri u kojoj ih neki konkretni sustav utjelovljuje.

Table 1: Examples of problematic machine translation outputs and their post-edited corrections

Test sets are the FLORES+ devtest (in-domain for seed-style text) and NTREX-128 (Federmann et al., 2022) as an out-of-domain check.

NLLB decodes with beam size 4; the Gemma-family LLMs use a fixed prompt and greedy decoding; while the external DeepSeek-V4-Flash reference uses its model-card sampling parameters (Appendix B). All system comparisons use paired bootstrap resampling (Koehn, 2004) with 2,000 resamples, applied to chrF++/spBLEU and to per-segment COMET scores. Two caveats apply throughout: (1) Croatian was only recently added to metric evaluation sets and the neural metrics are largely untested on it (Deutsch et al., 2025) and it is unclear how well the learned metrics correlate with human judgments (Singh et al., 2024); and (2) reference translations exhibit natural variability (e.g., inconsistent renderings of the same named entity within one article), which adds noise to reference-based metrics and provides further motivation for the multi-metric evaluation. The NLLB English→Croatian training runs are replicated with three random seeds, while the LLMs are single-seed runs. Seed standard deviations (≤ 0.1 chrF++/COMET) are well below the main effects reported in almost all cases.

4.2 Post-editing effort and corpus-level quality estimation

The two-stage human verification described in Section 3 shows that the Editors modified 71.5% of all segments. The corpus-level TER between raw machine translation and the final text is 16.27, i.e., roughly one in six machine-translated words was changed by a human. The mean normalized character edit distance is 0.099 (0.139 over edited segments only) and the median editing time is 32.3 seconds per segment. Matecat records editing time

as the interval a segment remains open, so segments left open between sessions accumulate durations that do not correspond to work. We therefore cap per-segment time at 30 minutes, which affects 65 segments (1.05% of the corpus) and gives a tracked editing time of 184.8 hours. The distribution stays heavily skewed after capping: the median segment takes 32.3 seconds, while the slowest 10% of segments account for 62.3% of the total. Most segments needed only quick fixes, and a minority of long sentences required terminology research.

Reference-free quality estimation confirms that this effort improved the text. Scoring the whole corpus with MetricX-24-XXL in QE mode (lower is better), the raw machine translation scores 1.644 and the final human-verified text 1.473. The paired per-segment improvement is +0.171 with a 95% bootstrap confidence interval of [0.146, 0.196].

4.3 Does the data help? Fine-tuning results

Table 2 shows the effect of fine-tuning NLLB models on the dataset. Fine-tuning improves every model size, in both directions, on both test sets, on both surface and neural metrics. All gains are significant at $p < 0.05$, most at $p = 0.0005$. The exception is COMET on FLORES+ (in-domain data, $p = 0.15$) for the NLLB 600M model. Gains on English→Croatian direction rise with model capacity, and more importantly, gains persist on out-of-domain data, indicating genuine translation improvement rather than adaptation to the test-set style.

Table 3 reports the LLM results on the FLORES+ devtest, with each fine-tuned row using the best learning rate from that model’s sweep (full sweeps in Appendix D). For TranslateGemma, fine-tuning yields significant surface-metric gains in both directions (up to +4.96 chrF++ on NTREX;

Model	Direction	FLORES+ chrF++	FLORES+ COMET	NTREX chrF++	NTREX COMET
NLLB-600M	en→hr	51.94→53.08	87.43→87.59	53.70→54.62	85.45→85.87
NLLB-600M	hr→en	59.56→60.49	86.66→86.87	61.30→61.38	86.04→86.26
NLLB-1.3B	en→hr	54.41→55.86	89.46→90.10	56.18→57.42	87.54→88.54
NLLB-1.3B	hr→en	61.67→62.65	87.72→87.93	63.35→63.91	87.12→87.43
NLLB-3.3B	en→hr	56.02→57.87	90.07→90.84	57.44→58.90	87.93→89.26
NLLB-3.3B	hr→en	62.38→62.94	88.02→88.24	64.47→64.58	87.49→87.68

Table 2: NLLB-200 zero-shot → fine-tuned on the Croatian Seed data (chrF++ / COMET).

Appendix D). We find that a single bidirectional model matches the performance of two direction-specific models (deltas ≤ 0.28 chrF++, n.s.; Appendix D), allowing one model to serve both directions without a loss in quality. For the 4B model, the choice between LoRA and full fine-tuning is less critical than the learning rate: at the gentler rates the two methods differ only marginally and with mixed signs (≤ 0.6 chrF++, ≤ 0.35 COMET), though full fine-tuning overfits more quickly and trails LoRA at 10^{-5} . Due to hardware constraints, the larger models were adapted using LoRA only; however, the 4B results suggest that this choice entails little to no quality penalty. All gains in this section were obtained using standard recipes and a limited learning-rate sweep, representing a lower bound on the dataset’s value for models below the quality ceiling (Section 4.5). Fine-tuned Gemma-4-31B produces the strongest Croatian→English system in this study (66.10 chrF++), while the best English→Croatian chrF++ score is achieved by an external zero-shot reference: DeepSeek-V4-Flash (Section 4.1), the largest model we can run on our infrastructure. The English→Croatian behavior of these strongest models on the neural metrics is analyzed in Section 4.5.

4.4 Isolating the human contribution: raw-MT ablation

The Matecat export preserves the raw machine translations from before post-editing, allowing us to directly test what the human verification contributes. For each NLLB size we fine-tune four models that differ only in their training data: (1) original English → raw MT Croatian; (2) original English → verified Croatian; (3) raw MT Croatian → original English; and (4) verified Croatian → original English. Comparing models (1) and (2) measures the value of the verified text as the translation target. While surface metrics show the raw MT and human-verified text are statistically similar (chrF++ deltas are within ± 0.4), both neural met-

rics systematically prefer the human-verified text. Table 4 shows the COMET results (mean $\pm$ sd over three seeds). The model trained on the verified text scores higher in all six model and test-set comparisons, while the seed variation is an order of magnitude smaller than most of the differences. Paired bootstrap tests confirm the advantage ($p < 0.05$) in every comparison except the NLLB 600M on FLORES+. MetricX independently prefers the verified text in five of the six comparisons, with the same single exception. Notably, the advantage of human verification increases with model capacity and is largest on the out-of-domain test (NTREX): from +0.10 COMET (600M, FLORES+) to +0.98 (3.3B, NTREX).

Comparing models (3) and (4) isolates a second contribution. Here, the Croatian side serves as the source text and the target, the original English, is identical in both conditions, so any difference between them measures the value of human verification purely as comprehension input. Results show that the human-verified text as input wins on COMET in all six model $\times$ test-set cells (+0.11 to +0.22, $p \leq 0.023$; seed 42), while the chrF++ differences are again small and inconsistent in sign (-0.21 to $+0.26$).

Three conclusions follow. First, the human post-editing and revision effort is precisely where the dataset’s value over freely producible machine translation lies, and its value increases with the capacity of the model being trained. Second, that value is two-sided: as target text, the verified Croatian supplies better style and content, and as source text it trains measurably better comprehension. Third, surface metrics are blind to both effects: an evaluation based on chrF++ alone would have concluded that human verification adds nothing.

4.5 Behavior at the quality ceiling

We use the term quality ceiling for the point at which a model’s zero-shot output already matches or exceeds the register and quality of the corpus’s

System	en→hr chrF++	en→hr COMET	hr→en chrF++	hr→en COMET
TranslateGemma-4B zero-shot	49.96	89.47	58.96	87.55
TranslateGemma-4B LoRA (10^{-6})	52.66	89.03	61.71	87.94
TranslateGemma-4B full FT ($3 \cdot 10^{-6}$)	53.02	88.92	61.79	87.82
TranslateGemma-27B zero-shot	56.19	92.15	62.16	88.62
TranslateGemma-27B LoRA (10^{-5})	58.88	91.59	64.50	88.79
Gemma-4-31B zero-shot (ctx)	58.64	91.76	64.81	89.17
Gemma-4-31B zero-shot (ctx, thinking)	59.13	92.02	64.77	89.18
Gemma-4-31B ctx-LoRA (10^{-5})	59.65	91.81	66.10	89.26
DeepSeek-V4-Flash zero-shot (ctx; thinking)	60.08	92.05	64.58	89.02

Table 3: LLM results on FLORES+ devtest (chrF++ / COMET per direction). Models show the best adapter/full-FT learning rate in parentheses. All fine-tuned LLM rows are the bidirectionally trained models. Bold marks the column best.

Model	FLORES+ verified	FLORES+ raw MT	NTREX verified	NTREX raw MT
NLLB-600M	87.65 $\pm$ 0.05	87.55 $\pm$ 0.02	85.85 $\pm$ 0.03	85.39 $\pm$ 0.04
NLLB-1.3B	90.14 $\pm$ 0.03	89.65 $\pm$ 0.01	88.53 $\pm$ 0.01	87.78 $\pm$ 0.02
NLLB-3.3B	90.78 $\pm$ 0.05	90.23 $\pm$ 0.02	89.24 $\pm$ 0.02	88.26 $\pm$ 0.02

Table 4: Training-data ablation, en→hr, COMET (mean $\pm$ sd over 3 seeds).

own Croatian side. By register we mean the characteristic style of the text rather than its correctness; for this corpus, that is the style of post-edited machine translation, which stays close to the source structure and phrasing. The NLLB models sit below this point, while the LLMs in Table 3 sit at or above it. Zero-shot DeepSeek-V4-Flash serves as an external reference point and is not fine-tuned. Notably, the strongest models in this study run on accessible hardware: Gemma-4-31B fits on a single 96 GB workstation GPU, while DeepSeek-V4-Flash was served on two DGX Spark units.

All fine-tuned models in this section are bidirectionally trained and use the learning rate reported in Table 3 (best COMET in each sweep). Appendix D shows that unidirectional vs. bidirectional training results in no practical difference at this model tier. Regardless, the two translation directions still behave differently. In the Croatian→English direction, fine-tuning improves both metric families for all three LLMs, in- and out-of-domain (Gemma-4-31B: +1.29/+3.07 chrF++ and +0.10/+0.28 COMET on FLORES+/NTREX, $p \leq 0.034$; TranslateGemma-27B: +2.34/+4.96 and +0.17/+0.56, $p \leq 0.0055$; TranslateGemma-4B: +2.75/+3.92 and +0.39/+0.79, $p = 0.0005$). The fine-tuned Gemma-4-31B is the strongest Croatian→English system in this study (66.10 chrF++). In the English→Croatian direction, fine-tuning yields comparable surface gains (+1.0 to +4.0 chrF++, $p = 0.0005$), but the effect on COMET depends on the model. For Gemma-4-31B, the dif-

ference from zero-shot is statistically zero (+0.05 in-domain, -0.14 out-of-domain, both n.s.), while both TranslateGemma sizes remain significantly below their zero-shot COMET (27B: -0.56 on both test sets, $p = 0.0005$; 4B: $-0.44/-0.66$, $p \leq 0.007$). MetricX ranks the fine-tuned English→Croatian systems in the same order and mildly prefers zero-shot output in the Croatian→English direction (Appendix D).

This English→Croatian cost is not fixed; it shrinks as the learning rate decreases. On TranslateGemma-4B, a four-point sweep (10^{-4} to 10^{-6}) reduces the COMET gap to zero-shot from $-1.37/-1.98$ (FLORES+/NTREX) to $-0.44/-0.66$. The reduction is monotone; every gentler rate is significantly better than 10^{-4} (the usual LoRA learning-rate recommendation), and the gap never reaches zero. TranslateGemma-27B shows the same pattern ($-0.98/-1.19$ at 10^{-4} , -0.56 at 10^{-5}), while on Gemma-4-31B the gap closes entirely (from -0.42 at 10^{-4} to the statistical zero above at 10^{-5}). Since every curve is monotone across its swept range, we report the best swept rates without claiming they are optimal. Notably, on the 4B model the gentler rates also give up some surface overlap (-0.5 chrF++ from 10^{-4} to 10^{-6}), so the learning rate trades one metric family against the other, consistent with the accuracy-naturalness trade-off formalized by Flamich et al. (2025). On the two larger models, the gentlest rate is best on both families.

We see a straightforward explanation for this

asymmetry. The Croatian side of the corpus descends from machine translation, so when used as the training target, it teaches the register of post-edited machine translation. For models whose zero-shot output is less natural than that register (the NLLB family), this is an upgrade on both metric families. For LLMs whose output is already more natural, it trades naturalness for surface overlap. Interestingly, the cost is predicted by the model’s output register rather than by its measured quality: TranslateGemma-4B pays the largest penalty despite having the lowest zero-shot COMET of the three LLMs. When the same text is used as the input, with the original English as the target, it helps every model at no measured cost. Manual inspection of the outputs supports this account: fine-tuned models translate more literally and closer to the references (occasionally calquing idioms), while zero-shot outputs are freer and score higher on the neural metrics.

Prompt-level results support this explanation. Injecting the full translation guidelines into the prompt changes nothing measurably (58.76/91.71 vs. 58.64/91.76 chrF++/COMET). Enabling the model’s reasoning mode, on the other hand, improves both chrF++ (+0.49, $p=0.016$) and COMET (+0.27, $p=0.005$), making it the only prompt-level intervention in this study that improves both metric families at once.

Finally, zero-shot DeepSeek-V4-Flash achieves the best English→Croatian chrF++ in the study (60.08) and matches the best Gemma-4-31B configuration on COMET (92.05 vs. 92.02 for the thinking zero-shot), trailing only zero-shot TranslateGemma-27B (92.15). In the Croatian→English direction, however, it trails the fine-tuned Gemma-4-31B (64.58 vs. 66.10 chrF++; 89.02 vs. 89.26 COMET). The two systems differ in prompt, decoding and reasoning mode, so this is a reference point rather than a controlled comparison.

5 Discussion

What the seed data is worth. For dedicated translation models, the class that seed data primarily targets, the result is simple: fine-tuning on the corpus improves every NLLB size, in both directions, on both test sets, on both metric families. The ablation shows where this value lies. Since raw machine translation of the English side can be produced for free, the dataset’s contribution over

such a baseline is the human post-editing and revision. This contribution grows with the capacity of the model being trained, from +0.10 COMET at 600M to +0.98 at 3.3B out of domain. We consider this an encouraging property of seed-style resources: the better the models become, the more the human verification matters.

Where the value ends. The quality-ceiling experiments show the limits of the dataset. For models whose zero-shot output already matches or exceeds the register of post-edited machine translation, English→Croatian fine-tuning buys surface overlap at a cost on the neural metrics. The cost shrinks as the learning rate decreases, but for the translation-specialised models it never reaches zero within the swept range: statistically zero for Gemma-4-31B, −0.56 COMET for TranslateGemma-27B and −0.44 for TranslateGemma-4B at the gentlest rates (Section 4.5). The asymmetry between the two directions, observed within a single bidirectional training run, points to the mechanism. A post-edited corpus carries its machine-translation ancestry on the target side of one direction only. Used as the training target, the Croatian side teaches the register of post-edited machine translation, which is an upgrade for models below that register and a downgrade for models above it. Used as the input, with the original English as the target, the same corpus helps models at every scale at no measured cost, including through the verification of the source-side text itself (Section 4.4). This gives simple guidance for using post-editing-derived seed corpora. Below the ceiling, the data can be used freely in both directions. At the ceiling, prefer the direction in which the post-edited text serves as the input; for the opposite direction, the natural complement is back-translation from native target-language text (Sennrich et al., 2016), which places the derived text on the input side by construction. We leave this controlled comparison to future work.

What this means for evaluation. The ablation would be invisible to an evaluation based on surface metrics alone: chrF++ cannot separate the raw-MT and human-verified training conditions, while both neural metrics separate them consistently. At the quality ceiling, the two metric families can even move in opposite directions under fine-tuning. Neither family can be dropped. chrF++ measures closeness to the reference, the neural metrics measure quality as judged by a learned model, and the two disagree exactly where the interesting effects are.

This is expected in theory as well, since no single score can reward accuracy and naturalness at the same time (Flamich et al., 2025). The prompt-level results fit the same picture: reasoning mode was the only intervention that improved both families at once, while adding the translation guidelines to the prompt changed nothing. In settings like ours, where the evaluated language is zero-shot for every neural metric and reference sets are small, we recommend reporting both metric families together with paired significance tests.

6 Conclusion

We presented the Croatian contribution to the OLDI Seed corpus: 6,193 English–Croatian sentence pairs produced by machine translation and post-edited and verified in two independent stages by professional translators. Beyond describing the resource, we validated it experimentally. Fine-tuning NLLB-200, TranslateGemma, and Gemma-4-31B shows that the corpus improves every model below the quality ceiling in both directions, and the strongest models in the Croatian→English direction. A single bidirectional model matches two direction-specific ones, so one model can serve both directions. Through a training-data ablation, we showed that the human verification itself carries measurable value: it grows with the size of the model being trained, it helps on both the target and the source side of training, and surface metrics cannot detect it. For the strongest models, we traced the limits of the corpus to the fact that its Croatian side originates from machine translation, and we discussed how corpora built this way are best used. The dataset is released to the community through the Open Language Data Initiative. As follow-up work, we plan a further revision pass over the corpus and a controlled comparison against back-translation from native Croatian text.

7 Data Availability

Our Croatian Seed contribution will be made publicly available as part of the OLDI Seed dataset (https://huggingface.co/datasets/openlanguagedata/oldi_seed) under the CC BY-SA 4.0 license.

Limitations

All findings concern a single language pair, and the neural metrics are largely untested on Croatian,

which makes the neural-metric separations conservative, but less calibrated than for high-resource languages (Deutsch et al., 2025; Singh et al., 2024). The quality-ceiling analysis rests on automatic metrics and manual inspection; a human MQM-style evaluation was not possible with the available annotators. The English→Croatian fine-tuning effects replicate across three random seeds and three model families, while the Croatian→English ablation cells and the individual LLM configurations are single-seed. The learning-rate sweeps cover two points for TranslateGemma-27B and four for TranslateGemma-4B, and the reported rates are the best swept rates, not a located optimum. Finally, the corpus inherits the domain profile of its Wikipedia source; validation of the strongest models on specialized domain material is left for follow-up work.

Acknowledgment

This work has been supported in part by a grant from “Integrated Territorial Programme 2021-2027 (European Regional Development Fund)” under the project IP.1.1.03.0081.

References

- Petra Bago, Sheila Castilho, Jane Dunne, Federico Gaspari, Andre Kåsen, Gauti Kristmannsson, Jon Arild Olsen, Natalia Resende, Níels Rúnar Gíslason, Dana D. Sheridan, Páraic Sheridan, John Tinsley, and Andy Way. 2022. *Achievements of the PRINCIPLE project: Promoting MT for Croatian, Icelandic, Irish and Norwegian*. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 349–350, Ghent, Belgium. European Association for Machine Translation.
- Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarriás, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. *ParaCrawl: Web-scale acquisition of parallel corpora*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567, Online. Association for Computational Linguistics.
- Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L. Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. *Croatian-English parallel corpus MaCoCu-hr-en 2.0*. Version 2.0. Distributed via CLARIN.SI.

- Justus Brockmann, Claudia Wiesinger, and Dragoş Ciobanu. 2022. [Error annotation in post-editing machine translation: Investigating the impact of text-to-speech technology](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 251–259, Ghent, Belgium. European Association for Machine Translation.
- David Dale, Laurie Burchell, Jean Maillard, Idris Abdulmumin, Antonios Anastasopoulos, Isaac Caswell, and Philipp Koehn. 2025. [Findings of the WMT 2025 shared task of the Open Language Data Initiative](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 495–502, Suzhou, China. Association for Computational Linguistics.
- DeepSeek-AI. 2026. [DeepSeek-V4: Towards highly efficient million-token context intelligence](#). *Preprint*, arXiv:2606.19348.
- Daniel Deutsch, Eleftheria Briakou, Isaac Rayburn Caswell, Mara Finkelstein, Rebecca Galor, Juraj Juraska, Geza Kovacs, Alison Lui, Ricardo Rei, Jason Riesa, Shruti Rijhwani, Parker Riley, Elizabeth Salesky, Firas Trabelsi, Stephanie Winkler, Biao Zhang, and Markus Freitag. 2025. [WMT24++: Expanding the language coverage of WMT24 to 55 languages & dialects](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12257–12284, Vienna, Austria. Association for Computational Linguistics.
- Christian Federmann, Tom Kocmi, and Ying Xin. 2022. [NTREX-128 – news test references for MT evaluation of 128 languages](#). In *Proceedings of the First Workshop on Scaling Up Multilingual Evaluation*, pages 21–24, Online. Association for Computational Linguistics.
- Mara Finkelstein, Isaac Caswell, Tobias Domhan, Jan-Thorsten Peter, Juraj Juraska, Parker Riley, Daniel Deutsch, Geza Kovacs, Cole Dilanni, Colin Cherry, Eleftheria Briakou, Elizabeth Nielsen, Jiaming Luo, Kat Black, Ryan Mullins, Sweta Agrawal, Wenda Xu, Erin Kats, Stephane Jaskiewicz, and 2 others. 2026. [TranslateGemma technical report](#). *Preprint*, arXiv:2601.09012.
- Gergely Flamich, David Vilar, Jan-Thorsten Peter, and Markus Freitag. 2025. [You cannot feed two birds with one score: the accuracy-naturalness tradeoff in translation](#). In *Second Conference on Language Modeling*. ArXiv:2503.24013 [cs.CL].
- Gemma Team. 2026. [Gemma 4 Technical Report](#). *Preprint*, arXiv:2607.02770.
- Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc’Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. [The Flores-101 evaluation benchmark for low-resource and multilingual machine translation](#). *Transactions of the Association for Computational Linguistics*, 10:522–538.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [LoRA: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. [The state and fate of linguistic diversity and inclusion in the NLP world](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.
- Philipp Koehn. 2004. [Statistical significance tests for machine translation evaluation](#). In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.
- Boshko Koloski, Senja Pollak, Blaž Škrlić, and Matej Martinc. 2022. [Out of thin air: Is zero-shot cross-lingual keyword detection better than unsupervised?](#) In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 400–409, Marseille, France. European Language Resources Association.
- Nikola Ljubešić, Miquel Esplà-Gomis, Sergio Ortiz Rojas, Filip Klubička, and Antonio Toral. 2016. [Croatian-English parallel corpus hrenWaC 2.0](#). Distributed via CLARIN.SI, ISLRN 969-131-780-609-0.
- Jean Maillard, Cynthia Gao, Elahe Kalbassi, Kaushik Ram Sadagopan, Vedanuj Goswami, Philipp Koehn, Angela Fan, and Francisco Guzman. 2023. [Small data, big impact: Leveraging minimal data for effective machine translation](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2740–2756, Toronto, Canada. Association for Computational Linguistics.
- Ali Marashian, Enora Rice, Luke Gessler, Alexis Palmer, and Katharina von der Wense. 2025. [From priest to doctor: Domain adaptation for low-resource neural machine translation](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7087–7098, Abu Dhabi, UAE. Association for Computational Linguistics.
- Matecat. 2026. [Matecat: Free and open source online CAT tool](#). Accessed August 2026.
- NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht,

- Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. [No language left behind: Scaling human-centered machine translation](#). *Preprint*, arXiv:2207.04672.
- Antoni Oliver and Sergi Álvarez-Vidal. 2026. [Croatian language in the transition from neural machine translation to large language models](#). *Suvremena lingvistika*, 52(101):101–120.
- Maja Popović. 2017. [chrF++: words helping character n-grams](#). In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.
- Maja Popović and Mihael Arčan. 2016. [PE2rr corpus: Manual error annotation of automatically pre-annotated MT post-edits](#). In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 27–32, Portorož, Slovenia. European Language Resources Association (ELRA).
- Maja Popović, Alberto Poncelas, Marija Brkić, and Andy Way. 2020. [Neural machine translation for translating into Croatian and Serbian](#). In *Proceedings of the 7th Workshop on NLP for Similar Languages, Varieties and Dialects*, pages 102–113, Barcelona, Spain (Online). International Committee on Computational Linguistics (ICCL).
- Matt Post. 2018. [A call for clarity in reporting BLEU scores](#). In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Anushka Singh, Ananya Sai, Raj Dabre, Ratish Pudupully, Anoop Kunchukuttan, and Mitesh Khapra. 2024. [How good is zero-shot MT evaluation for low resource Indian languages?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 640–649, Bangkok, Thailand. Association for Computational Linguistics.
- Marko Tadić. 2023. [Language report Croatian](#). In Georg Rehm and Andy Way, editors, *European Language Equality: A Strategic Agenda for Digital Language Equality*, Cognitive Technologies, pages 111–114. Springer International Publishing, Cham.
- Gaurish Thakkar, Vanja Štefanec, Daša Farkaš, and Marko Tadić. 2024. [Building a large language model for moderately resourced language: A case of Croatian](#). In *Proceedings of the 35th Central European Conference on Information and Intelligent Systems (CECIS 2024)*, pages 225–232, Varaždin, Croatia.
- Rik van Noord, Miquel Esplà-Gomis, Malina Chichirau, Gema Ramírez-Sánchez, and Antonio Toral. 2025. [Quality beyond a glance: Revealing large quality differences between web-crawled parallel corpora](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1824–1838, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pėzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. [The MAR-CELL legislative corpus](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France. European Language Resources Association.

A Data sample

The dataset is released as a JSON Lines file with one record per sentence. Each record carries the sentence text, the OLDI language identification (ISO 639-3 code, ISO 15924 script code, Glottocode), and the URL of the source Wikipedia article; records pair with the English side of the Seed corpus through the shared id. Figure 1 shows two consecutive records from the released file. The corresponding English records read: “Inside one of the chambers, near the ceiling, is a painted hand print.” (5208) and “The location of the cave has been well known in modern times.” (5209).

B Prompts and decoding

All Gemma-4-31B systems use a modular zero-shot prompt assembled from the following templates (verbatim; {sl}/{tl} are language names, {slc}/{tlc} ISO codes):

- Role (always present): You are a professional {sl} ({slc}) to {tl} ({tlc}) translator.
- Task (no-context conditions): Produce only the {tl} translation, without

```

{
  "id": 5208,
  "iso_639_3": "hrv",
  "iso_15924": "Latn",
  "glottocode": "croal245",
  "text": "Unutar jedne od komora, blizu stropa, nalazi se oslikani otisak ruke.",
  "url": "https://en.wikipedia.org/wiki/Chichen_Itza",
  "last_updated": "1.0"
}
{
  "id": 5209,
  "iso_639_3": "hrv",
  "iso_15924": "Latn",
  "glottocode": "croal245",
  "text": "Lokacija špilje u današnje je vrijeme dobro poznata.",
  "url": "https://en.wikipedia.org/wiki/Chichen_Itza",
  "last_updated": "1.0"
}

```

Figure 1: Two consecutive records from the released Croatian Seed file (pretty-printed; the released file stores one record per line).

any additional explanations or commentary. Please translate the following {sl} text into {tl}: \n\n{text}

- Document context (ctx conditions, “sandwich” layout): You are translating this sentence: \n{text}\n\nHere is the document it comes from, for context: \n{ctx}\n\nProduce only the {tl} translation, without any additional explanations or commentary. Translate the sentence given above.
- Guideline injection (guide conditions): Follow these translation guidelines strictly: \n\n{guidelines}, where {guidelines} is the full text of the OLDI translation guidelines.

The context block contains the complete source text from the same article (FLORES+ lines filtered by metadata; for NTREX, see Appendix E) without any markers or target-side texts, so no reference leakage is possible. The sentence-first layout was chosen over a context-first layout in a 200-sentence test. TranslateGemma is prompted exclusively through its native chat template (technical report, Figure 3), with source and target language codes set per direction. DeepSeek-V4-Flash uses the same modular prompt as Gemma with the model-card recommended parameters (temperature 1.0, top-p 1.0, reasoning effort *high*, 16,384-token budget with automatic doubling on truncation).

Decoding is greedy for all Gemma-family systems and beam-4 for NLLB; DeepSeek-V4-Flash used model-card recommendations. Google’s recommended sampling for Gemma (temperature 1.0, top-p 0.95, top-k 64) was evaluated on all four prompt conditions and scored 0.2–0.5 chrF++ below greedy in English→Croatian and on par in Croatian→English, validating the greedy choice. “Thinking” rows enable the model’s native reasoning mode with the same prompts and a 32k token budget; only the final answer channel is scored.

C Training hyperparameters

All training uses AdamW, bf16, warmup ratio 0.05, and checkpoint selection on the FLORES+ dev set exclusively (chrF++ for NLLB, loss for the LLMs; early stopping with patience 2). Prompt tokens are masked from the loss in all LLM runs.

Bidirectional runs mix both directions in one training set (the prompt encodes the direction). LoRA adapters are merged before evaluation; all LLM inference uses vLLM.

D Full sweep and MetricX results

Gemma-4-31B learning-rate sweep, English→Croatian, FLORES+ devtest (chrF++ / COMET / MetricX):

Factorial contrasts isolate learning rate as the driver: at fixed rate, the direction mix changes chrF++ by ≤ 0.14 and COMET by ≤ 0.09 (n.s.), while the $10^{-4} \rightarrow 10^{-5}$ rate contrast is significant on both families. The hr→en cells of the non-headline bidirectional 31B configuration (LoRA 10^{-4}) are 65.93 chrF++ / 89.23 COMET.

Family	Recipe	LR	Effective batch	Epochs (max)	Notes
NLLB-200 (all sizes)	full FT	$3 \cdot 10^{-5}$	32	12	max length 256; seeds 42/1/2 (en→hr)
TranslateGemma-4B	LoRA r=32, $\alpha=64$, dropout 0.05	10^{-4} , 10^{-5} , $3 \cdot 10^{-6}$, 10^{-6}	16	6	targets: LM attention+MLP projections only
TranslateGemma-4B	full FT (LM + head; vision frozen)	10^{-5} , $3 \cdot 10^{-6}$, 10^{-6}	16	6	gradient checkpointing
TranslateGemma-27B	LoRA r=32, $\alpha=64$	10^{-4} , 10^{-5}	16	6	gradient checkpointing
Gemma-4-31B	LoRA r=32, $\alpha=64$	10^{-4} , 10^{-5}	12 (3×GPU DDP)	3	document-context prompt with ± 4 -sentence source window; max length 2048

Table 5: Training hyperparameters per model family.

Configuration	chrF++	COMET	MetricX ↓
zero-shot (ctx)	58.64	91.76	1.512
zero-shot (ctx, thinking)	59.13	92.02	1.335
LoRA 10^{-4} , bidir.	58.45	91.34	1.544
LoRA 10^{-4} , en→hr	58.31	91.37	1.620
LoRA $3 \cdot 10^{-5}$, en→hr	59.21	91.58	1.532
LoRA $3 \cdot 10^{-5}$, bidir.	59.24	91.67	1.495
LoRA 10^{-5} , en→hr	59.57	91.83	1.437
LoRA 10^{-5} , bidir.	59.65	91.81	1.455

Table 6: Gemma-4-31B learning-rate sweep, en→hr, FLORES+ devtest. Bold marks the column best.

TranslateGemma learning-rate sweeps (chrF++ / COMET; bidir. LoRA and full FT, both test sets):

Direction-specific (unidirectional) LoRA at 10^{-4} , FLORES+ devtest (chrF++ / COMET): TranslateGemma-4B 53.17 / 88.29 (en→hr) and 61.61 / 87.83 (hr→en); TranslateGemma-27B 58.58 / 91.38 and 64.34 / 88.73 — within 0.28 chrF++ of the corresponding bidirectional models (n.s.).

MetricX (FLORES+ devtest, lower is better) for the main LLM configurations, including those summarised only in this appendix:

MetricX agrees with COMET on every learning-rate and recipe ordering except one (TranslateGemma-4B full FT, $3 \cdot 10^{-6}$ vs. 10^{-6} , 0.04 COMET apart); the remaining disagreements are between systems within 0.03 COMET of each other. Its systematic divergence is a preference for zero-shot output over fine-tuned systems that shrinks monotonically with the rate in English→Croatian and persists mildly in Croatian→English, consistent with the accuracy–naturalness trade-off (Flamich et al., 2025).

E Document context for NTREX

FLORES+ metadata provides source URLs for each line, which enable grouping and providing the source-side context block directly. NTREX does not have such labels. We reconstruct context by letting the base Gemma-4-31B judge, for each sentence, which of its ± 10 neighbours belong to the same document, and use the accepted neighbours as the context block. This model-inferred context was validated against no-context prompting on NTREX: it improves Croatian→English by +0.81 chrF++ / +0.54 COMET (both $p < 0.05$) and English→Croatian by +0.30 COMET ($p = 0.0005$) with flat chrF++ — the method improves neural-judged quality even where surface metrics are blind to it. All NTREX ctx rows in the paper use this context.

Configuration	en→hr FLORES+	en→hr NTREX	hr→en FLORES+	hr→en NTREX
4B zero-shot	49.96 / 89.47	49.70 / 88.22	58.96 / 87.55	58.21 / 86.41
4B bidir. LoRA 10^{-4}	53.19 / 88.10	53.16 / 86.24	61.89 / 87.75	61.97 / 86.84
4B bidir. LoRA 10^{-5}	53.03 / 88.52	52.76 / 86.61	61.98 / 87.92	62.00 / 87.04
4B bidir. LoRA $3 \cdot 10^{-6}$	52.86 / 88.85	52.55 / 87.07	61.90 / 87.94	62.01 / 87.16
4B bidir. LoRA 10^{-6}	52.66 / 89.03	52.64 / 87.56	61.71 / 87.94	62.13 / 87.20
4B full FT 10^{-5}	52.87 / 87.97	52.14 / 86.10	60.26 / 87.15	60.64 / 86.11
4B full FT $3 \cdot 10^{-6}$	53.02 / 88.92	53.06 / 87.42	61.79 / 87.82	62.02 / 87.00
4B full FT 10^{-6}	52.78 / 88.88	52.89 / 87.60	61.53 / 87.88	61.72 / 87.10
27B zero-shot	56.19 / 92.15	56.26 / 91.26	62.16 / 88.62	61.17 / 87.75
27B bidir. LoRA 10^{-4}	58.32 / 91.17	59.40 / 90.07	64.35 / 88.79	65.35 / 88.00
27B bidir. LoRA 10^{-5}	58.88 / 91.59	60.26 / 90.70	64.50 / 88.79	66.13 / 88.31

Table 7: TranslateGemma learning-rate sweeps (chrF++ / COMET).

System	en→hr	hr→en
TranslateGemma-4B zero-shot	2.035	2.185
TranslateGemma-4B LoRA (10^{-4})	2.486	2.474
TranslateGemma-4B bidir. LoRA (10^{-4})	2.482	2.485
TranslateGemma-4B bidir. LoRA (10^{-5})	2.352	2.414
TranslateGemma-4B bidir. LoRA ($3 \cdot 10^{-6}$)	2.293	2.411
TranslateGemma-4B bidir. LoRA (10^{-6})	2.252	2.347
TranslateGemma-4B full FT (10^{-5})	2.438	2.755
TranslateGemma-4B full FT ($3 \cdot 10^{-6}$)	2.269	2.457
TranslateGemma-4B full FT (10^{-6})	2.252	2.406
TranslateGemma-27B zero-shot	1.176	1.783
TranslateGemma-27B LoRA (10^{-4})	1.579	2.025
TranslateGemma-27B bidir. LoRA (10^{-4})	1.573	1.996
TranslateGemma-27B bidir. LoRA (10^{-5})	1.496	1.994
Gemma-4-31B zero-shot (ctx)	1.512	1.777
Gemma-4-31B zero-shot (ctx, thinking)	1.335	1.712
Gemma-4-31B ctx-LoRA bidir. (10^{-4})	1.544	1.867
Gemma-4-31B ctx-LoRA bidir. (10^{-5})	1.455	1.837
DeepSeek-V4-Flash zero-shot (ctx)	1.406	1.862

Table 8: MetricX (FLORES+ devtest, lower is better) for the main LLM configurations. Bold marks the column best.