# One Size Does Not Fit All: Why EU Legislative Translation Demands Domain-Specific Fine-Tuning of LLMs

**Valerio Lorini*** **Paula Vlaic*** **Ulascan Akbulut*** **Daniele Marcoaldi†**

*{valerio.lorini, paula.vlaic, ulascan.akbulut}@europarl.europa.eu
†daniele.marcoaldi@ext.europarl.europa.eu

European Parliament, DG for Translation and Clear Language
Luxembourg

## Abstract

EU legislation is equally authentic and legally binding in all 24 official languages, rendering high-quality translation a legal obligation rather than a mere choice. Therefore, high-quality language technology supporting translation processes in all EU languages is essential for language professionals at the European Parliament (EP). This paper investigates whether domain-specific fine-tuning of an open-weight Large Language Model (LLM) yields consistently larger quality gains on legislative text compared to generic text, in all 23 EU target languages from English. We evaluate ten experimental conditions: base model, in-domain and cross-domain fine-tuning, sequential generic-then-legislative fine-tuning, and zero-shot Claude Sonnet 4.6 as a proprietary reference. We analyse BLEU, chrF, TER, and COMET metrics on nearly 700,000 segments. Results confirm the hypothesis for all 23 languages: legislative fine-tuning enhances BLEU by +12.30 compared to +7.10 for generic fine-tuning, demonstrating a consistent advantage of +5.20 BLEU in all the metrics. The fine-tuned EuroLLM-22B decisively outperforms Claude Sonnet 4.6, Anthropic's latest frontier model, on both domains, highlighting that targeted adaptation of a smaller open-weight model can surpass a state-of-the-art proprietary system. Cross-domain transfer within the institutional domain is positive for all languages, with no catastrophic forgetting. Low-resource languages such as Irish and Maltese benefit the most from fine-tuning, while a divergence between BLEU and COMET rankings for some languages underlines the need for evaluation metrics alongside traditional measures.

## 1 Introduction

The European Union operates under the principle of full multilingualism, with 24 official languages serving as the linguistic foundation of its democratic legitimacy. Every piece of EU legislation is equally authentic in all official languages, creating a translation challenge in terms of both volumes and precision. The Directorate-General for Translation of the European Parliament alone processes millions of pages annually. Ensuring consistency, legal equivalence, and terminological precision for 24 language versions is a continuous institutional priority.

Recent advances in large language models have brought new possibilities to institutional translation workflows. Models such as EuroLLM [1], specifically designed with European multilingual capabilities, offer promising baselines for translation tasks. However, the question of whether general-purpose LLMs, even those trained on multilingual European data, can adequately handle the specific demands of legislative translation without domain adaptation remains open.

Legislative language presents unique challenges that distinguish it from general text. EU legal texts follow rigid structural conventions (preamble, enacting terms, annexes), employ highly spe-



[1] https://eurollm.io

cific and often language-pair-dependent terminology, and must maintain precise cross-references [2]. A subtle mistranslation can alter legal obligations, create loopholes, or lead to conflicting interpretations between language versions. These characteristics suggest that domain-specific fine-tuning may be particularly impactful for legislative translation, more so than for translating general-purpose text.

This paper presents a systematic experimental framework to test the hypothesis that domain-specific fine-tuning with multilingual parallel legislative data produces larger improvements in legislative translation quality compared to the effect of generic fine-tuning on generic text. We use EuroLLM-22B (Ramos et al., 2026) as our base model and design eight experimental runs that compare: (1) baseline translation quality on legislative and generic text; (2) the effect of legislative fine-tuning on both text types; (3) the effect of generic fine-tuning on both text types; and (4) the effect of sequential fine-tuning (generic followed by legislative data) on both text types.

The Directorate General for Translation and Clear Language [3] (DG TRAD) of the European Parliament set up the AI4TRAD project with the aim of establishing a framework for testing and assessing the capabilities and scalability of AI tools and functionalities, with particular emphasis on Large Language Models, focusing on drafting, editing, and DG TRAD workflows. Our contributions are threefold. First, we provide a rigorous experimental design for measuring the differential impact of domain adaptation in a multilingual setting. Second, we produce parallel evaluation datasets in 24 EU languages for both legislative and generic domains, directly contributing to the AI4TRAD project's goals of quality assessment. Third, we offer practical guidance for institutional translation services considering the integration of fine-tuned LLMs into their workflows.

## 2 Related Work

### 2.1 Domain Adaptation for Machine Translation

Domain adaptation has been a central concern in Neural Machine Translation (NMT) since it first started. Saunders (2023) provides a comprehensive survey categorising approaches into data-centric methods (data selection, augmentation, and synthesis), model-centric methods (architecture modifications, parameter-efficient tuning), and inference-time methods (ensemble decoding, terminology constraints). A recurring finding in the literature is that fine-tuning a general-purpose NMT model on in-domain data yields substantial quality improvements, but risks catastrophic forgetting of general translation capabilities. Shahnazaryan (2024) investigates the impact of domain specification on cross-language transfer, demonstrating that well defined domain boundaries significantly improve in-domain transfer learning effectiveness. Their findings are particularly relevant to our work, as legislative text constitutes one of the most clearly constrained domains, with distinctive syntactic, lexical, and structural properties. The question of whether LLM-based translation systems benefit from the same domain adaptation dynamics as traditional NMT models has gained attention with the rise of models such as LLaMA (Touvron et al., 2023), Mistral (Chaplot, 2023), and EuroLLM. Parameter-efficient fine-tuning (PEFT) methods, particularly LoRA (Low-Rank Adaptation) (Hu et al., 2022) and QLoRA (Quantized Low-Rank Adaptation)(Dettmers et al., 2023), have made domain adaptation of large models computationally feasible. These approaches update only a small fraction of the model's parameters (typically 0.2–0.3%), while achieving competitive performance with full fine-tuning. For a 22-billion-parameter model like EuroLLM, PEFT methods are essential to make fine-tuning practical for the scale of experiments we propose.

### 2.2 Legislative and Legal Translation

The translation of legal and legislative texts has long been recognised as one of the most demanding specialisations in both human and machine translation. EU legislative translation in particular operates under the dual constraint of producing texts that are equivalent in all language versions while adhering to the drafting conventions and legal traditions of each target system. Corpus-based analyses of EU legal texts have revealed significant terminological variation in multi-word term translation, with strong source-language interference patterns that deviate from target-language legal conventions (Seracini,

[2] Interinstitutional Style Guide, https://style-guide.europa.eu/en/

[3] https://the-secretary-general.europarl.europa.eu/en/directorates-general/trad

2021).

Barman et al. (2025) present a comparative study of transformer training strategies for legal machine translation, finding that fine-tuned models significantly outperform models trained from scratch on legal text. Their evaluation with multiple metrics (BLEU, chrF++, TER, ROUGE, BERTScore, and COMET) demonstrates that pre-training provides a strong foundation upon which domain-specific fine-tuning can build effectively. Studies on terminology translation errors in NMT for legal text have further emphasised the critical importance of consistent terminology, which directly affects legislative harmonisation between EU jurisdictions.

The EUR-Lex corpus, containing the entire body of EU law in 24 languages, represents one of the largest and most consistently aligned parallel corpora available for any specialised domain. Its use for training and evaluating legislative translation systems is well established, though the specific application to LLM fine-tuning in all 24 EU languages remains underexplored.

### 2.3 EU Multilingual NLP and Translation Tools

Several EU-funded initiatives have sought to advance multilingual Natural Language Processing (NLP) capabilities for the Union's official languages. The Neural Translation for the European Union (NTEU) project (Bié et al., 2020) aimed to build 550 direct translation engines covering all EU language pairs, leveraging data from DGT translation memories, Paracrawl [4], and other EU sources. The AI-Based Multilingual Services platform [5] provides a range of NLP tools, including automatic translation, text classification, named entity recognition, and anonymisation, for public administrations in the EU. The Europarl corpus, derived from European Parliament proceedings, has been a foundational resource for multilingual NLP research, with expanded versions (Europarl-ST) providing aligned audio-text samples in multiple European languages. More recently, the AI4TRAD project in the European Parliament has focused on developing and evaluating AI-based translation technologies specifically tailored to legislative and parliamentary contexts, with an emphasis on quality assessment and bias detection in all EU official languages.

### 2.4 Catastrophic Forgetting and Sequential Fine-Tuning

A well documented challenge in domain-specific fine-tuning is catastrophic forgetting, whereby a model's performance on previously learned tasks degrades after adaptation to new data. Luo et al. (2023) provide an empirical study of this phenomenon in LLMs ranging from 1B to 14B parameters, finding that the severity of forgetting intensifies with model scale, though general instruction tuning can help alleviate it. Huang et al. (2024) propose a Self-Synthesized Rehearsal (SSR) framework that uses the LLM itself to generate synthetic training instances from its general knowledge, which are then mixed with domain-specific data during fine-tuning. This approach achieves performance comparable to conventional rehearsal methods, while being more data-efficient. The question of the fine-tuning order and whether to adapt first to general data, then to domain-specific data, or vice versa has been shown to significantly impact both domain performance and retention of general capabilities. Our experimental design explicitly addresses this question through its sequential fine-tuning condition.

Unlike multi-stage post-training pipelines such as Tower (2024) and TowerPlus (2025), which build broad multilingual and translation-adjacent competence through continued pretraining, instruction-tuning, and preference optimisation, our approach is narrower and more targeted: LoRA-based adaptation directly on domain-specific parallel data held within the institution, optimising for a single constrained domain.

## 3 Experiment

This section describes the data, data filtering, models, experimental conditions, and evaluation methodology. The experimental design is structured to describe the effect of domain-specific fine-tuning on translation quality in a highly regulated domain (EU legislation) against generic institutional text, while also benchmarking against a state-of-the-art proprietary model.

[4] https://paracrawl.eu

[5] https://language-tools.ec.europa.eu/

### 3.1 Data

The experiments require two parallel corpora covering all 24 EU official languages, with English as the source language: one for legislative text and one for non-legislative generic institutional text. Both corpora are built with the same structure and quality checks, so any differences in how they perform can be linked to the nature of the domain rather than the data itself. The data has been sampled from translations produced by the European Parliament's Directorate-General for Translation (DG TRAD) and has been extracted from the European Advanced Multilingual Information System (EUramis), the EU interinstitutional repository of translations memories, where DG TRAD stores its segmented translations in multilingual and multidirectional memories [6].

#### 3.1.1 Legislative Text Corpus

The legislative corpus is drawn from 112 consolidated legislative documents adopted by the European Parliament. We selected documents originally drafted in English, reflecting the institutional reality that 83–87% of documents processed by DG TRAD have English as a source language. Content is restricted to a single document type: consolidated legislative documents; amendments, plenary documents, and other procedural texts are excluded.

#### 3.1.2 Non-Legislative (Generic) Corpus

The non-legislative corpus comprises 3,113 documents from the European Parliament, spanning a range of institutional document types: general documents (DV, 18,271 segments), resolutions (RE, 3,548; RC, 446), texts for the Citizens'App (CI, 2,835), briefings (BR, 153), and others. The corpus represents the work of diverse requesting services, with the largest contributors being COMM (28.2%), GREF-PRES (16.0%), and EPRS (12.0%), ensuring broad coverage across institutional functions. Reports on legislative acts (regulations, directives, decisions), amendments, and plenary legislative documents are strictly excluded. As with the legislative corpus, only documents originally drafted in English are included. A particular risk to be managed is terminological contamination: even non-legislative institutional texts frequently reference policy frameworks and legislation, potentially replicating specialised jargon. The selection of communicative and analytical content types (reports, briefings, studies) rather than policy-prescriptive documents mitigates this risk, though we acknowledge that some degree of shared institutional vocabulary is unavoidable and discuss this as a limitation. Additionally, we replicated our experiment using the Flores+ benchmark (NLLB Team et al., 2024) to compare its generic data quality with that of our dataset.

[6] https://www.europarl.europa.eu/translation/en/translation-at-the-european-parliament/technology-to-support-translation

#### 3.1.3 Data Extraction

To ensure that any performance differences between domains can be attributed to content rather than segment length artefacts, the non-legislative corpus is resampled to match the legislative character length distribution. The legislative distribution is computed from pairs across 36 buckets (10 characters wide, spanning 40–400 characters). Each non-legislative language pool (drawn from an initial candidate set of 100,000 segments per language) is sampled proportionally per bucket with 15% oversampling headroom, followed by similarity deduplication, then trimmed to exact targets. A small shortfall of approximately 450 segments per language (1.7%) in the longest buckets (350–400 characters) is backfilled from adjacent length ranges, reflecting the inherently shorter nature of non-legislative texts.

After deduplication and distribution matching, segments are randomly shuffled and split into train (20,000), evaluation (2,000), and test (4,000) per language pair, with a fixed random seed for reproducibility. Each segment pair is stored in JSONL format with the following metadata fields to support reproducibility, stratified analysis, and reuse in downstream AI4TRAD tasks:

**Table 1:** JSONL segment pair metadata schema

| Field | Description | Presence |
|---|---|---|
| `doc_id` | Document identifier | Both |
| `doc_type` | Type (Consol. Text, DV…) | Both |
| `year` | Publication / transl. year | Both |
| `en_segment` | EN source segment | Both |
| `xx_segment` | Target language segment | Both |
| `lang` | ISO 639-1 target code | Both |
| `segment_id` | Index within document | Both |
| `obs` | Observations / notes | Both |
| `req_serv` | Requesting service | Non-leg. |

#### 3.1.4 Data Analysis

Table 2 provides an overview of the two corpora. Both are identical in size and split structure, with

598,000 total segment pairs each (26,000 per target language).

**Table 2:** Corpus overview

| Metric | Legislative | Non-Legislative |
|---|---|---|
| Source documents | 112 | 3,113 |
| Total segment pairs | 598,000 | 598,000 |
| Segments / language | 26,000 | 26,000 |
| Train / Eval / Test | 20k / 2k / 4k | 20k / 2k / 4k |
| Unique EN segments | 30,861 | 105,794 |
| Cross-corpus overlap | 0 (0.00%) | 0 (0.00%) |
| Disk size | 358.7 MB | 346.3 MB |

Table 3 reports EN source segment length statistics (computed from French pairs as representative). After distribution matching, the mean character length differs by only 4.2 characters (2.2%) and the mean word count by 0.4 words (1.2%). The total absolute distribution difference between all 36 fine-grained buckets is 3.7%, with most of the buckets differing by less than 0.1 percentage point.

**Table 3:** EN source segment length statistics.

| Metric | Legis. | Non-L. | Diff. |
|---|---|---|---|
| Mean (chars) | 186.0 | 181.8 | $-4.2$ ($-2.2\%$) |
| Median (chars) | 176 | 173 | $-3$ ($-1.7\%$) |
| Std dev (chars) | 89.8 | 86.7 | $-3.0$ ($-3.4\%$) |
| Mean (words) | 29.2 | 28.9 | $-0.4$ ($-1.2\%$) |
| Median (words) | 28 | 27 | $-1$ ($-3.6\%$) |
| Min / Max (chars) | 40 / 400 | 40 / 400 | 0 / 0 |

#### 3.1.5 Data Filtering

Both corpora underwent a four-stage filtering pipeline to ensure clean, balanced, and independent datasets. Stage 0 removes segments containing hyperlinks, which were found to produce unrelated translations (this alone improved TER from 54.03 to 52.25 on non-legislative data). Stage 1 performs exact deduplication via MD5 hashing [7] on normalised text and removes misaligned pairs whose source–target character length ratio falls outside the $[0.3, 3.0]$ range. Stage 2 detects near-duplicates using a combined score of normalised Levenshtein similarity and Jaccard similarity over character 5-grams (Levenshtein and others, 1966; Jaccard, 1912; Broder, 1997), flagging cross-split pairs above a 0.75 threshold. Stage 3 identifies semantic near-duplicates via cosine similarity on sentence embeddings (paraphrase-multilingual-MiniLM-L12-v2), flagging pairs exceeding a 0.90 similarity threshold as potential data leakage.

[7] www.rfc-editor.org/rfc/rfc1321

After filtering, the legislative corpus retains 588,610 of 598,000 segments (98.4%) and the non-legislative corpus retains 573,095 (95.8%). The higher removal rate for non-legislative data reflects greater internal redundancy across the 3,113 source documents. Segment length distributions remain stable after filtering.

### 3.2 Model

#### 3.2.1 EuroLLM-22B (Open-Weight Base Model)

The primary model for all fine-tuning experiments is EuroLLM-22B, a 22 billion parameter dense Transformer with grouped query attention (GQA), trained on over 4 trillion tokens in 35 languages with particular emphasis on the 24 EU official languages. Its 128,000-token vocabulary provides strong subword coverage for European languages, and its 32K context window accommodates the longer sentence structures typical of legislative text. EuroLLM's open-weight release enables the fine-tuning experiments central to this study. Fine-tuning is performed using LoRA (Low-Rank Adaptation) to enable efficient parameter updates while maintaining the model's broad multilingual capabilities. We apply LoRA adapters to the attention projection matrices (Q, K, V, O) with a rank of r=32 and a scaling factor of $alpha$ =64. This configuration updates approximately 0.3% of the total model parameters, making the fine-tuning procedure computationally tractable across the multiple experimental conditions. Training is conducted on 4x NVIDIA RTX Pro 6000 Blackwell 96GB, with a learning rate of 1.998e-4, a batch size of 384, and 5 epochs per fine-tuning run.

#### 3.2.2 Claude Sonnet (Proprietary Reference)

As a proprietary reference, the closed-weight Claude Sonnet 4.6 is evaluated in a zero-shot translation setting on the same test sets, accessed via the Anthropic API with temperature=0. Its inclusion tests whether fine-tuning an open-weight European model can match or exceed a leading proprietary system, which is especially relevant in the legislative domain where data sovereignty constraints may preclude reliance on external API providers.

### 3.3 Experimental Design

Before full fine-tuning, hyperparameters are optimised using Optuna (Akiba et al., 2019) with Tree-structured Parzen Estimation (TPE) on a subset of

the data, minimising validation cross-entropy loss with early stopping (patience 3, threshold 0.001). The search covers learning rate ($[10^{-5}, 2 \times 10^{-4}]$, log-scale), warmup ratio ($[0.03, 0.15]$), and LoRA dropout ($[0.01, 0.1]$). Fixed design choices include LoRA rank $r = 32$, scaling factor $\alpha = 64$, and a cosine decay learning rate schedule.

#### 3.3.1 Training Runs and Seed Validation

The experiment comprises 10 runs in total. The EuroLLM-22B base model is evaluated once with deterministic inference (temperature=0), since no variation is introduced. Each of the three fine-tuning regimes, legislative, generic, and sequential (generic-then-legislative), is further trained with 2 different random seeds, using the same data split, but varying initialisation, data shuffling order, and dropout masks. This yields 3 checkpoints per regime (9 fine-tuned checkpoints total), allowing us to report means and standard deviations and to assess whether observed differences are robust to training noise.

#### 3.3.2 Evaluation Conditions

Each checkpoint (or model, in the case of the base model and Claude Sonnet) is evaluated on both the generic and the legislative test sets, yielding 10 evaluation conditions. Every EuroLLM-based inference is performed with temperature=0 to ensure deterministic output. Table 5 provides a complete overview of the evaluation matrix.

#### 3.3.3 Inference Protocol

All models are evaluated using the same prompt template to ensure comparability. The prompt instructs the model to translate a given English segment into the specified target language, without additional context, few-shot examples, or domain-specific instructions, a deliberate choice to isolate the effect of fine-tuning from prompt engineering. The temperature is set to 0 for all runs (both EuroLLM and Claude Sonnet) to eliminate sampling variance and ensure fully deterministic outputs.

#### 3.3.4 Hypotheses and Key Comparison

The central hypothesis is that the improvement from C2 to C4 (legislative fine-tuning on legislative text) will be significantly larger than the improvement from C1 to C5 (generic fine-tuning on generic text). Secondary comparisons assess cross-domain transfer (C3 vs. C1, C6 vs. C2), sequential fine-tuning benefits (C8 vs. C4, C7 vs. C5), and the proprietary gap (C4/C8 vs. C10, C5 vs. C9).

### 3.4 Evaluation

Translation quality is assessed using four complementary automatic metrics: Bilingual Evaluation Understudy (BLEU) for n-gram precision, Character-level F-score (chrF) for tokenisation-independent character-level evaluation, Translation Edit Rate (TER), and COMET v22 [8] for neural quality estimation trained on human judgements. All metrics are computed at the corpus level for each of the 23 translation pairs (EN→XX). For fine-tuned conditions (C3–C8), we report the mean and standard deviation across seeds.

## 4 Results

This section analyses the results along four dimensions: baseline and reference performance, single-domain fine-tuning effects, sequential fine-tuning, and comparison with Claude Sonnet 4.6.

Table 5 reports the aggregate results for all 23 language pairs for each of the 10 evaluation conditions (+1, C0 run for comparing a generic non-institutional text).

### 4.1 Baseline and Reference Performance (C0–C2)

The EuroLLM-22B base model exhibits a substantial performance gap between the two domains. On legislative text (C2), the base model achieves 49.66 BLEU, compared to 37.85 on generic institutional text (C1), a difference of 11.81 BLEU points. This asymmetry likely reflects the presence of legislative parallel data in EuroLLM's pre-training corpus, which gives the model a head start on legislative language patterns.

The Flores+ external benchmark (C0) yields 36.33 BLEU, 1.52 points below the generic EP test set (C1). This confirms that the European Parliament's non-legislative corpus represents a comparable level of translation difficulty to the established multilingual benchmark, granting credibility to the generic baseline as a fair comparator.

### 4.2 Single-Domain Fine-Tuning (C3–C6)

#### 4.2.1 In-Domain Gains

The central finding of this study is confirmed: domain-specific fine-tuning yields signif-

[8] Computed using unbabel-comet v2.2.7 aclanthology.org/2022.wmt-1.52/

**Table 5:** Aggregate results for all language pairs ($N$ = number of evaluated segments).

| Cond. | Model | Fine-Tuning | Test | BLEU | chrF | TER | COMET | $N$ |
|---|---|---|---|---|---|---|---|---|
| C0 | EuroLLM-22B | None (base) | Flores+ | 36.33 | 63.04 | 51.98 | 89.50 | 23,276 |
| C1 | EuroLLM-22B | None (base) | Generic | 37.85 | 64.42 | 52.25 | 90.42 | 86,529 |
| C2 | EuroLLM-22B | None (base) | Legis. | 49.66 | 72.57 | 41.41 | 90.53 | 86,883 |
| C3 | EuroLLM-22B | Leg. FT | Generic | 41.32 | 66.56 | 49.61 | 90.69 | 86,529 |
| C4 | EuroLLM-22B | Leg. FT | Legis. | 61.96 | 80.27 | 31.00 | 92.16 | 86,883 |
| C5 | EuroLLM-22B | Gen. FT | Generic | 44.95 | 68.89 | 46.29 | 91.56 | 86,529 |
| C6 | EuroLLM-22B | Gen. FT | Legis. | 52.90 | 74.15 | 39.08 | 90.81 | 86,883 |
| C7 | EuroLLM-22B | Gen.→Leg. FT | Generic | 43.46 | 68.14 | 47.63 | 91.14 | 86,529 |
| C8 | EuroLLM-22B | Gen.→Leg. FT | Legis. | 63.17 | 81.08 | 29.77 | 92.32 | 86,883 |
| C9 | Claude Sonnet 4.6 | Zero-shot | Generic | 39.65 | 66.20 | 49.96 | 90.92 | 86,529 |
| C10 | Claude Sonnet 4.6 | Zero-shot | Legis. | 53.90 | 75.85 | 37.14 | 91.25 | 86,883 |

icantly larger gains on legislative text than generic fine-tuning produces on generic text. Legislative fine-tuning (C2→C4) improves BLEU by +12.30 points (49.66→61.96), while generic fine-tuning (C1→C5) improves BLEU by +7.10 points (37.85→44.95). The legislative advantage of +5.20 BLEU is consistent across all four metrics: chrF (+3.23), TER (−4.45), and COMET (+0.49). Table 6 summarises these gains.

**Table 6:** In-domain fine-tuning gains (Δ over base). Arrows indicate the desired direction.

| | BLEU ↑ | chrF ↑ | TER ↓ | COMET ↑ |
|---|---|---|---|---|
| $\mathrm{Leg}_{\mathrm{Base}} \rightarrow \mathrm{Leg}_{\mathrm{FT}}$ | +12.30 | +7.70 | −10.41 | +1.63 |
| $\mathrm{Gen}_{\mathrm{Base}} \rightarrow \mathrm{Gen}_{\mathrm{FT}}$ | +7.10 | +4.47 | −5.96 | +1.14 |
| $\Delta_{\mathrm{leg}} - \Delta_{\mathrm{gen}}$ | **+5.20** | **+3.23** | **−4.45** | **+0.49** |

The magnitude of the legislative gain is notable, given that the base model already starts from a higher baseline on legislative text (49.66 vs. 37.85 BLEU). Fine-tuning thus delivers larger absolute improvements on top of a stronger starting point, suggesting that the structured, rigorous nature of legislative language is particularly responsive to adaptation through domain-specific parallel data.

#### 4.2.2 Cross-Domain Transfer

Cross-domain evaluation reveals that fine-tuning on one domain does not degrade performance on the other; rather, it produces modest but consistent improvements. Legislative fine-tuning evaluated on generic text (C3 vs. C1) yields +3.47 BLEU, while generic fine-tuning evaluated on legislative text (C6 vs. C2) yields +3.24 BLEU. The near-symmetry of these cross-domain gains (+3.47 vs. +3.24) is notable: regardless of which domain the model is fine-tuned on, it gains approximately 3.3–3.5 BLEU points on the other domain. Importantly, however, these cross-domain gains represent only 26–49% of the corresponding in-domain gains, confirming that domain-matched fine-tuning is substantially more effective than mismatched fine-tuning.

These results provide evidence against catastrophic forgetting within the institutional domain in the fine-tuning regime: fine-tuning on legislative data does not harm generic translation (and vice-versa), but rather produces a modest general improvement, likely through exposure to additional high-quality parallel data that reinforces the model's general translation capabilities.

### 4.3 Sequential Fine-Tuning (C7–C8)

The sequential fine-tuning strategy (generic→legislative) emerges as the best-performing configuration overall. On legislative text, the sequential model (C8) achieves 63.17 BLEU, surpassing the single-stage legislative model (C4, 61.96) by +1.21 BLEU, with consistent improvements across all metrics (chrF +0.81, TER −1.23, COMET +0.16). The sequential model thus achieves 110% of the single-stage legislative gain over the base model, indicating that the generic pre-fine-tuning stage provides a beneficial foundation for subsequent legislative specialisation.

On generic text, the sequential model (C7) achieves 43.46 BLEU, which is 1.49 BLEU below the single-stage generic model (C5, 44.95) but still 5.61 BLEU above the base model (C1, 37.85). The sequential model thus retains 79% of the generic fine-tuning improvement, demonstrating that the legislative fine-tuning stage partially erodes but does not eliminate the gains from the generic stage.

### 4.4 Comparison with Claude Sonnet 4.6 (C9–C10)

The fine-tuned EuroLLM-22B models decisively outperform Claude Sonnet 4.6 across both domains. On legislative text, the single-stage legislative model (C4) exceeds Claude (C10) by +8.06 BLEU (61.96 vs. 53.90), and the sequential model (C8) extends this margin to +9.27 BLEU (63.17 vs. 53.90). The advantage is consistent across all metrics: chrF +4.42/+5.23, TER $-6.14/-7.37$, and COMET +0.91/+1.07 for C4/C8, respectively.

On generic text, the generic fine-tuned model (C5) outperforms Claude (C9) by +5.30 BLEU (44.95 vs. 39.65), with gains across all metrics (chrF +2.69, TER $-3.67$, COMET +0.64). Claude Sonnet's zero-shot performance on generic text (39.65 BLEU) sits between the base EuroLLM (37.85) and the legislative fine-tuned model evaluated cross-domain (C3, 41.32), indicating that even mismatched domain fine-tuning of an open-weight model can rival a frontier proprietary system.

These results carry significant practical implications: a 22-billion-parameter open-weight model, when fine-tuned on institutional translation memories already held within the organisation, can substantially outperform a frontier proprietary model, while preserving full data sovereignty and eliminating dependency on external API providers.

Claude Sonnet is evaluated in a zero-shot setting because the study isolates the effect of fine-tuning, and introducing few-shot examples would add a prompt engineering variable dependent on sample selection, quantity, and context window size, which is outside the scope of the research question.

### 4.5 Per-Language Analysis

Table 7 presents per-language results on legislative text for the best model (C8, sequential fine-tuning) alongside the base model (C2). The sequential model improves BLEU for all 23 languages, with gains ranging from +9.4 (DE) to +19.4 (GA and MT).

A clear typological hierarchy emerges. Romance languages occupy the top tier (RO 72.0, ES 70.6, FR 70.1, PT 70.8, IT 67.9 BLEU), followed by Hellenic (EL 68.1), Slavic (58–67 BLEU), Germanic (55–66), Baltic (52–56), and Finno-Ugric languages at the bottom (ET 48.5, FI 48.6, HU 50.8). This ranking largely reflects typological distance from English: Romance languages share substantial lexical and syntactic overlap with the English source, facilitating higher n-gram overlap scores, while Finno-Ugric languages express equivalent content through morphologically complex forms that BLEU penalises.

However, BLEU rankings diverge sharply from COMET rankings. Estonian and Finnish rank 22nd and 23rd on BLEU, but are top on COMET (94.5 and 94.4, respectively), while Spanish and French rank among the top five on BLEU, but 20th and 19th on COMET (90.7 and 91.0). This divergence is systematic: Finno-Ugric and Slavic languages consistently score higher on COMET than their BLEU rank would suggest, while Romance languages show the opposite pattern. COMET, as a learned neural metric trained on human quality judgements, captures semantic adequacy independently of surface form, making it more robust to the morphological variation inherent in agglutinative languages. This finding has important implications for multilingual evaluation: relying on BLEU alone would severely underestimate translation quality for roughly one-third of EU official languages and could lead to misguided resource allocation decisions.

It is worth noting that Irish (GA) and Maltese (MT) show the largest fine-tuning gains (+19.4 BLEU each), nearly double the average. As lower-resource languages, they benefit disproportionately from domain-specific training data, making fine-tuning a particularly high-value investment for these languages. Maltese also achieves the lowest TER score (19.6), indicating that the model learns to reproduce Maltese legislative phrasing with very high fidelity.

## 5 Discussion

The experimental results provide strong evidence for the central hypothesis: domain-specific fine-tuning is disproportionately impactful for legislative translation. The legislative test advantage (+5.20 BLEU, +3.23 chrF, $-4.45$ TER, +0.49 COMET) is consistent for all four evaluation metrics, indicating that this is not an artifact of any single metric's sensitivity, but a robust finding about the nature of domain adaptation for highly regulated text.

The magnitude of the legislative fine-tuning gain is remarkable when contextualised against the base model's already elevated performance on legislative text. Typically, improvements from a high

**Table 7:** Per-language results on legislative text for the best model (C8: sequential fine-tuning) vs. the base model (C2). Languages ranked by C8 BLEU.

| Language | Family | C2 BLEU | C8 BLEU | Δ BLEU | C8 chrF | C8 TER | C8 COMET | $N$ |
|---|---|---|---|---|---|---|---|---|
| RO | Romance | 60.1 | 72.0 | +11.9 | 85.2 | 22.3 | 94.0 | 3,784 |
| MT | Semitic | 52.2 | 71.6 | +19.4 | 89.2 | 19.6 | 81.5 | 3,761 |
| PT | Romance | 54.3 | 70.8 | +16.6 | 84.3 | 22.9 | 91.3 | 3,771 |
| ES | Romance | 58.2 | 70.6 | +12.4 | 84.0 | 22.6 | 90.7 | 3,777 |
| FR | Romance | 55.6 | 70.1 | +14.5 | 84.4 | 23.2 | 91.0 | 3,799 |
| EL | Hellenic | 54.4 | 68.1 | +13.7 | 82.2 | 24.9 | 93.6 | 3,752 |
| IT | Romance | 55.7 | 67.9 | +12.2 | 83.5 | 24.9 | 92.4 | 3,777 |
| BG | Slavic | 53.8 | 66.7 | +12.8 | 82.4 | 26.1 | 93.8 | 3,782 |
| DA | Germanic | 52.4 | 66.3 | +13.9 | 82.6 | 28.2 | 93.1 | 3,734 |
| SV | Germanic | 52.4 | 64.6 | +12.2 | 82.3 | 27.7 | 93.2 | 3,764 |
| SL | Slavic | 50.0 | 63.5 | +13.5 | 80.6 | 30.6 | 93.9 | 3,808 |
| GA | Celtic | 43.9 | 63.3 | +19.4 | 81.8 | 29.7 | 88.2 | 3,781 |
| SK | Slavic | 51.3 | 62.9 | +11.6 | 80.2 | 29.2 | 94.3 | 3,776 |
| NL | Germanic | 51.4 | 62.9 | +11.5 | 81.2 | 30.2 | 91.4 | 3,837 |
| CS | Slavic | 49.9 | 62.0 | +12.1 | 79.2 | 30.5 | 94.4 | 3,765 |
| HR | Slavic | 44.8 | 61.8 | +17.0 | 79.9 | 30.7 | 94.3 | 3,780 |
| PL | Slavic | 48.6 | 59.5 | +10.9 | 78.0 | 33.5 | 93.6 | 3,779 |
| LV | Baltic | 45.1 | 56.1 | +11.0 | 78.1 | 38.2 | 93.3 | 3,759 |
| DE | Germanic | 46.1 | 55.5 | +9.4 | 77.8 | 37.0 | 89.8 | 3,777 |
| LT | Baltic | 40.0 | 52.5 | +12.5 | 77.1 | 40.4 | 93.5 | 3,759 |
| HU | Finno-Ugric | 35.5 | 50.8 | +15.3 | 75.9 | 42.2 | 92.9 | 3,765 |
| FI | Finno-Ugric | 34.0 | 48.6 | +14.6 | 76.6 | 44.4 | 94.4 | 3,805 |
| ET | Finno-Ugric | 35.8 | 48.5 | +12.7 | 76.3 | 44.1 | 94.5 | 3,791 |

baseline are harder to achieve due to diminishing returns. That legislative fine-tuning delivers larger absolute gains from a stronger starting point, highlighting the responsiveness of legislative text to domain adaptation. We attribute this to the highly formulaic and terminologically consistent nature of EU legislative text: standardised phrasing such as recital structures, article numbering conventions, and cross-references create recurring patterns that LoRA adapters can efficiently capture with a small number of updated parameters.

The legislative advantage is most pronounced for TER, which is reduced by 25.1% on legislative text versus 11.4% for generic fine-tuning on generic text. Since TER measures the edit operations needed to match the reference, this reduction could directly translate into post-editing productivity gains. Notably, cross-domain evaluation shows no catastrophic forgetting: both fine-tuned models improve on the other domain by approximately 3.5 BLEU, suggesting that LoRA adaptation on high-quality parallel data produces general improvements alongside domain-specific gains.

The sequential fine-tuning results reveal a clear asymmetry. On legislative text, the sequential model (C8, 63.17 BLEU) outperforms the single-stage legislative model (C4, 61.96), suggesting that exposure to generic institutional text in the first fine-tuning stage provides useful horizontal knowledge, such as broader vocabulary coverage and diverse syntactic patterns, that benefits subsequent legislative specialisation. On generic text, however, the sequential model (C7, 43.46 BLEU) underperforms the single-stage generic model (C5, 44.95) by 1.49 BLEU. This trade-off is expected: the second-stage legislative fine-tuning partially overwrites generic-specific adaptations. Nevertheless, the sequential model retains 79% of the generic gain over the base model, making it a viable solution for institutions that handle both legislative and non-legislative translation.

The comparison with Claude Sonnet 4.6 is perhaps the most practically significant finding. Despite being a frontier proprietary model of considerably larger scale and broader training, Claude Sonnet is outperformed by the fine-tuned EuroLLM-22B by substantial margins: +8.06 BLEU on legislative text (C4 vs. C10) and +5.30 BLEU on generic text (C5 vs. C9). The sequential model extends the legislative margin further to +9.27 BLEU (C8 vs. C10). This demonstrates that targeted fine-tuning of a smaller open-weight model on institutional translation memories, data already available within the organisation, can decisively surpass a general-purpose commercial system. For the European Parliament, where leg-

islative texts routinely contain politically sensitive pre-decisional content, the ability to achieve superior translation quality without transmitting data to third-party services represents a decisive operational advantage in terms of both performance and data sovereignty.

## 6 Future Work

This study opens several lines of future research. First, the datasets produced for these experiments are available for testing [9] and will be integrated into the broader framework of the AI4TRAD project for the detection of bias and the quality assessment of LLM-based translation. The systematic analysis of translation biases, including gender bias, cultural bias, and legal-system bias, in the 24 official EU languages constitutes a natural extension of this work.

Second, human evaluation by specialised translators could complement the automatic metrics reported here. Expert assessment of legal equivalence, terminological consistency, and structural fidelity provides dimensions of quality that no automatic metric can fully capture. A planned human evaluation campaign should rate a subset of outputs from each experimental condition, providing both direct assessment scores and fine-grained error annotations.

Third, more sophisticated fine-tuning strategies merit investigation. Retrieval-augmented generation (RAG)(Lewis et al., 2020) that provides relevant legislative context at inference time represents a possible test. The interaction between fine-tuning and prompting strategies is similarly underexplored for legislative translation.

Fourth, the experimental framework can be extended to additional model architectures and sizes. Comparing EuroLLM-22B with smaller variants and with models not specifically designed for European languages (e.g., LLaMA-3(Grattafiori et al., 2024), Mistral, Tower+) could reveal the interaction between model design choices and domain adaptation effectiveness.

Finally, the deployment of fine-tuned models in operational institutional workflows raises questions about inference efficiency and quality assurance integration. Pilot deployments at the European Parliament's translation service, with feedback loops from professional translators, represent the natural next step toward practical impact.

[9] https://doi.org/10.5281/zenodo.20072253

## 7 Conclusion

This paper presents a systematic experimental investigation of the differential impact of domain-specific fine-tuning on legislative versus generic text translation using EuroLLM-22B for all 24 official languages of the EU. Our results confirm the central hypothesis: legislative fine-tuning brings a +12.30 BLEU improvement on legislative text, compared to +7.10 for generic fine-tuning on generic text, establishing a consistent legislative advantage of +5.20 BLEU, +3.23 chrF, −4.45 TER, and +0.49 COMET. This disproportionate impact holds despite the base model's already elevated performance on legislative text, underscoring the unique responsiveness of highly regulated language to domain adaptation.

Three additional findings emerge from the experimental design. First, cross-domain fine-tuning does not degrade out-of-domain performance; both domain-specific models improve translation quality on the other domain by approximately 3.5 BLEU points, providing evidence against catastrophic forgetting within institutional domain under LoRA adaptation. Second, sequential fine-tuning (generic then legislative) produces the highest overall legislative quality (63.17 BLEU, +13.51 over the base model) while retaining 79% of the generic fine-tuning gain, positioning it as the preferred single-model strategy for institutions handling mixed text types. Third, the fine-tuned open-weight EuroLLM-22B decisively outperforms Claude Sonnet 4.6, a frontier proprietary model, by +8.06 BLEU on legislative text and +5.30 on generic text, demonstrating that targeted domain adaptation of a smaller model using institutional translation memories can surpass general-purpose commercial systems.

Beyond its research contributions, the parallel datasets and experimental framework produced through this study constitute a resource for ongoing work on bias detection, quality assessment, and the responsible integration of LLM-based translation into institutional workflows. Data, configurations, and experimental code will be released with the camera-ready version of this paper to support reproducibility and further research on domain adaptation for institutional translation.

## References


[Akiba et al.2019] Akiba, Takuya, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. 2019. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631.

[Alves et al.2024] Alves, Duarte M, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

[Barman et al.2025] Barman, Amit, Atanu Mandal, and Sudip Kumar Naskar. 2025. From scratch to fine-tuned: A comparative study of transformer training strategies for legal machine translation. In *Proceedings of the 1st Workshop on NLP for Empowering Justice (JUST-NLP 2025)*. Association for Computational Linguistics.

[Bié et al.2020] Bié, Laurent, Aleix Cerdà-i Cucó, Hans Degroote, Amando Estela, Mercedes García-Martínez, Manuel Herranz, Alejandro Kohan, Maite Melero, Tony O'Dowd, Sinéad O'Gorman, Mārcis Pinnis, Roberts Rozis, Riccardo Superbo, and Artūrs Vasiļevskis. 2020. Neural translation for the European Union (NTEU) project. In Martins, André, Helena Moniz, Sara Fumega, Bruno Martins, Fernando Batista, Luisa Coheur, Carla Parra, Isabel Trancoso, Marco Turchi, Arianna Bisazza, Joss Moorkens, Ana Guerberof, Mary Nurminen, Lena Marg, and Mikel L. Forcada, editors, *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 477–478, Lisboa, Portugal, November. European Association for Machine Translation.

[Broder1997] Broder, Andrei Z. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.

[Chaplot2023] Chaplot, Devendra Singh. 2023. Albert q. jiang, alexandre sablayrolles, arthur mensch, chris bamford, devendra singh chaplot, diego de las casas, florian bressand, gianna lengyel, guillaume lample, lucile saulnier, lélio renard lavaud, marie-anne lachaux, pierre stock, teven le scao, thibaut lavril, thomas wang, timothée lacroix, william el sayed. *arXiv preprint arXiv:2310.06825*, 3.

[Dettmers et al.2023] Dettmers, Tim, Artidoro Pagnoni, Aman Srivastava, and Ari Holtzman. 2023. QLoRA: Efficient finetuning of quantized language models. In *Advances in Neural Information Processing Systems (NeurIPS 2023)*.

[Grattafiori et al.2024] Grattafiori, Aaron, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[Hu et al.2022] Hu, Edward J., 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. *arXiv preprint arXiv:2106.09685*.

[Huang and others2024] Huang, Jie et al. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, volume 1, pages 1416–1428, Bangkok, Thailand.

[Jaccard1912] Jaccard, Paul. 1912. The distribution of the flora in the alpine zone. 1. *New phytologist*, 11(2):37–50.

[Levenshtein and others1966] Levenshtein, Vladimir I et al. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

[Lewis et al.2020] Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

[Luo et al.2023] Luo, Yun, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2023. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. In *IEEE Transactions on Audio, Speech and Language Processing*.

[NLLB Team et al.2024] NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, 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, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2024. Scaling neural machine translation to 200 languages. *Nature*, 630(8018):841–846.

[Ramos et al.2026] Ramos, Miguel Moura, Duarte M. Alves, et al. 2026. EuroLLM-22B: Technical report. *arXiv preprint arXiv:2602.05879*.

[Rei et al.2025] Rei, Ricardo, Nuno M Guerreiro, José Pombal, João Alves, Pedro Teixeirinha, Amin Farajian, and André FT Martins. 2025. Tower+: Bridging generality and translation specialization in multilingual llms. *arXiv preprint arXiv:2506.17080*.

[Saunders2023] Saunders, Danielle. 2023. Domain adaptation and multi-domain adaptation for neural machine translation: A survey. *Journal of Artificial Intelligence Research*, 75:1–67.

[Seracini2021] Seracini, Francesca L. 2021. Phraseology in multilingual EU legislation: a corpus-based study of translated multi-word terms. *Perspectives*, 29(2).

[Shahnazaryan and others2024] Shahnazaryan, Lia et al. 2024. Defining boundaries: The impact of domain specification on cross-language and cross-domain transfer in machine translation. *arXiv preprint arXiv:2408.11926*.

[Touvron et al.2023] Touvron, Hugo, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

# A Appendix

## A.1 Data filtering Figures

**Table 8:** Legislative Corpus: Avg. Segment Lengths

| Split | Before | | After | |
|---|---|---|---|---|
| | chars | words | chars | words |
| Train | 186.2 | 28.6 | 186.2 | 28.6 |
| Eval | 185.5 | 28.5 | 189.4 | 29.1 |
| Test | 185.9 | 28.5 | 189.7 | 29.1 |
| **Wtd. Avg** | **186.1** | **28.6** | **186.6** | **28.7** |

**Table 9:** Non-Legislative Corpus: Avg. Segment Lengths

| Split | Before | | After | |
|---|---|---|---|---|
| | chars | words | chars | words |
| Train | 115.3 | 18.1 | 115.9 | 18.6 |
| Eval | 115.3 | 18.1 | 116.6 | 18.7 |
| Test | 115.0 | 18.0 | 116.1 | 18.6 |
| **Wtd. Avg** | **115.2** | **18.1** | **116.0** | **18.6** |

## A.2 Bayesian Optimization with TPE Optuna

### Optimization Objective

We aim to find the optimal hyperparameters with the Optuna framework (Akiba et al., 2019)

$$\theta = (\text{learning rate},\ \text{warmup ratio},\ \text{dropout}) \quad (1)$$

that minimize the validation loss:

$$\theta^* = \arg\min_{\theta \in \Theta} f(\theta) \quad (2)$$

where $f(\theta)$ denotes the validation loss after training. The goal is therefore to solve Eq. (2) over the parameter space defined in Eq. (1).

### Bayesian Modeling (Inverse Density)

Instead of modeling the likelihood $p(y \mid \theta)$, the Tree-structured Parzen Estimator (TPE) models the inverse density:

$$p(\theta \mid y) \quad (3)$$

This inverse framing, shown in Eq. (3), allows TPE to decompose the search space into good and bad regions as described below.

### Splitting Observations

Let the observed losses be:

$$\mathcal{D} = \{(\theta_i, y_i)\}_{i=1}^{N} \quad (4)$$

A threshold is defined using quantile $\gamma$ over the data set in Eq. (4):

$$y^* = \text{quantile}_\gamma(\{y_1, y_2, \ldots, y_N\}) \quad (5)$$

Observations with loss below $y^*$ (Eq. (5)) are considered "good", while those above are considered "bad".

### Density Decomposition

Using the threshold from Eq. (5), the conditional density is split as:

$$p(\theta \mid y) = \begin{cases} l(\theta), & y < y^* \\ g(\theta), & y \geq y^* \end{cases} \quad (6)$$

where the two component densities are:

$$l(\theta) = p(\theta \mid y < y^*) \quad (7)$$

$$g(\theta) = p(\theta \mid y \geq y^*) \quad (8)$$

Both $l(\theta)$ (Eq. (7)) and $g(\theta)$ (Eq. (8)) are estimated using Parzen kernel density estimators over the good and bad observation sets defined by Eq. (6).

### Expected Improvement

The Expected Improvement (EI) objective used by TPE to score the candidate hyperparameters are proportional to the ratio of the densities in Eqs. (7) and (8):

$$\text{EI}(\theta) \propto \frac{l(\theta)}{g(\theta)} \quad (9)$$

Maximizing Eq. (9) favours regions that are dense under $l(\theta)$ and sparse under $g(\theta)$.

### Next Hyperparameter Selection

Based on the EI criterion in Eq. (9), the next hyperparameter configuration is selected as:

$$\theta_{\text{next}} = \arg\max_{\theta}\ \frac{l(\theta)}{g(\theta)} \quad (10)$$

### Optimization Loop

Starting from the data set in Eq. (4), the optimisation iterates by applying Eq. (10) at each step:

$$\{(\theta_i, y_i)\}_{i=1}^{N} \longrightarrow \{l(\theta),\, g(\theta)\} \longrightarrow \theta_{N+1} \quad (11)$$

until the evaluation budget is exhausted. Combining Eqs. (7), (8), and (10), the full selection rule is:

$$\boxed{\theta_{\text{next}} = \arg\max_{\theta} \frac{p(\theta \mid y < y^*)}{p(\theta \mid y \geq y^*)}} \quad (12)$$

### A.3 Optuna Results

Optuna results for the non-legislative and legislative corpora are displayed below. Both runs were conducted on 6 NVIDIA H200 GPUs, with one GPU assigned per parallel worker.

**Table 10:** Fixed parameters shared across all Optuna runs

| Parameter | Value |
|---|---|
| LoRA $r$ | 32 |
| LoRA $\alpha$ | 64 |
| Weight decay | 0.01 |
| LR scheduler | Cosine |
| Batch size | 8 |
| Gradient accumulation | 4 |
| Max steps | 350 |
| Eval steps | 50 |
| Train samples per language | 300 (6,900 train) |
| Val samples per language | 100 (2,300 val) |

#### Non-Legislative

The hyperparameter search for the non-legislative filtered data set was completed over 560 trials in total. The search space covered learning rate (log-uniform, $[10^{-5}, 2 \times 10^{-4}]$), warmup ratio (uniform, $[0.03, 0.15]$), and LoRA dropout (uniform, $[0.01, 0.1]$), while keeping the other parameters fixed for all trials (Table: 10) The best configuration was found at trial 120, achieving a validation loss of **1.0347**. The search converged well, with no improvement observed from trial 120 onward. The slightly higher validation loss compared to the unfiltered data set (1.0347 vs. 1.0234) is likely attributable to the reduced number of training samples after filtering (443k vs. 459k).

**Table 11:** Best hyperparameter configuration for Non-Legislative dataset

| Parameter | Value |
|---|---|
| Learning rate | $1.166 \times 10^{-4}$ |
| Warmup ratio | 0.0466 |
| LoRA dropout | 0.0971 |
| **Eval loss** | **1.0347** |

#### Legislative

The hyperparameter search for the legislative dataset was completed in 600 trials. The search space and fixed parameters were identical to the non-legislative run. The best configuration was found in trial 317, achieving a validation loss of **0.8187**. The search converged quickly: the loss dropped from 0.8720 in trial 0 to 0.8188 by trial 22, with only marginal gains in the fourth decimal place thereafter. Most trials beyond trial 100 were pruned at step 50. The TPE sampler converged to a learning rate around $2 \times 10^{-4}$, a warmup ratio in the range $[0.10, 0.12]$, and a dropout between $0.03$ and $0.07$.

**Table 12:** Best hyperparameter configuration for Legislative dataset

| Parameter | Value |
|---|---|
| Learning rate | $1.998 \times 10^{-4}$ |
| Warmup ratio | 0.1102 |
| LoRA dropout | 0.0453 |
| **Eval loss** | **0.8187** |

**Table 13:** Significant completed trials for the legislative dataset

| Trial | Learning rate | Warmup ratio | Dropout | Eval loss |
|---|---|---|---|---|
| 0 | $3.07 \times 10^{-5}$ | 0.1441 | 0.0759 | 0.8720 |
| 3 | — | — | — | 0.8194 |
| 6 | $6.01 \times 10^{-5}$ | 0.0487 | 0.0240 | 0.8466 |
| 18 | $1.93 \times 10^{-4}$ | 0.1216 | 0.0395 | 0.8301 |
| 22 | — | — | — | 0.8188 |
| 317 | $2.00 \times 10^{-4}$ | 0.1102 | 0.0453 | **0.8187** |

### A.4 Finetune Results

This section reports per-language results for each experimental condition across BLEU, chrF, TER, and COMET. Each table pairs the same model's performance on legislative and non-legislative test sets.

### A.5 Training Configuration

Table 19 consolidates the training setup for all three fine-tuning objectives. Sequential fine-tuning proceeds in two stages: Stage 1 trains a LoRA adapter on the filtered non-legislative corpus; Stage 2 loads that adapter and continues training on the legislative corpus at a reduced learning rate. Legislative and generic fine-tuning each train a fresh adapter in a single stage. All runs use PyTorch DDP (Distributed Data Parallel) on a single node with automatic checkpoint resumption via SLURM (Simple Linux Utility for Resource Management) script.

**Table 14:** Legislative fine-tuning: per-language results on legislative test set (left) and non-legislative test set (right). Seed = 42.

| | Test: Legislative | | | | | Test: Non-Legislative | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lang** | **BLEU** | **chrF** | **TER** | **COMET** | $N$ | **BLEU** | **chrF** | **TER** | **COMET** | $N$ |
| BG | 65.53 | 81.88 | 26.99 | 93.75 | 3,782 | 49.51 | 71.34 | 39.73 | 92.81 | 3,742 |
| CS | 60.61 | 78.29 | 31.88 | 94.28 | 3,765 | 33.90 | 60.11 | 57.25 | 92.53 | 3,747 |
| DA | 64.69 | 81.92 | 28.77 | 92.97 | 3,734 | 49.02 | 71.62 | 40.61 | 92.45 | 3,759 |
| DE | 54.79 | 77.24 | 37.58 | 89.73 | 3,777 | 35.99 | 63.71 | 55.16 | 88.82 | 3,736 |
| EL | 67.10 | 81.80 | 25.30 | 93.57 | 3,752 | 44.35 | 66.55 | 45.13 | 92.11 | 3,804 |
| ES | 69.57 | 83.44 | 23.31 | 90.58 | 3,777 | 52.77 | 72.69 | 37.25 | 90.17 | 3,748 |
| ET | 47.53 | 75.36 | 45.13 | 94.38 | 3,791 | 27.21 | 60.64 | 66.50 | 92.44 | 3,752 |
| FI | 47.06 | 75.63 | 46.79 | 94.37 | 3,805 | 27.23 | 62.03 | 66.52 | 93.11 | 3,755 |
| FR | 69.32 | 83.76 | 24.05 | 90.84 | 3,799 | 45.68 | 68.78 | 46.26 | 88.99 | 3,742 |
| GA | 61.06 | 80.05 | 32.00 | 87.70 | 3,781 | 39.96 | 65.72 | 51.11 | 85.12 | 3,784 |
| HR | 59.74 | 78.73 | 32.86 | 94.07 | 3,780 | 37.15 | 64.17 | 52.87 | 92.37 | 3,772 |
| HU | 49.81 | 75.15 | 43.44 | 92.94 | 3,765 | 29.84 | 60.30 | 62.46 | 91.28 | 3,759 |
| IT | 67.09 | 83.12 | 25.51 | 92.30 | 3,777 | 49.55 | 71.89 | 40.79 | 91.55 | 3,762 |
| LT | 51.67 | 76.38 | 41.49 | 93.38 | 3,759 | 33.24 | 63.00 | 58.19 | 92.17 | 3,778 |
| LV | 55.35 | 77.37 | 38.93 | 93.21 | 3,770 | 34.73 | 63.37 | 56.60 | 91.86 | 3,794 |
| MT | 69.53 | 88.36 | 20.97 | 80.82 | 3,761 | 47.30 | 75.95 | 38.77 | 76.33 | 3,757 |
| NL | 62.03 | 80.54 | 30.94 | 91.30 | 3,837 | 38.22 | 65.54 | 52.70 | 90.15 | 3,789 |
| PL | 58.39 | 77.53 | 34.24 | 93.54 | 3,779 | 34.02 | 61.94 | 57.97 | 92.03 | 3,776 |
| PT | 69.45 | 83.61 | 23.63 | 91.13 | 3,771 | 56.21 | 75.14 | 33.23 | 91.18 | 3,751 |
| RO | 70.66 | 84.60 | 23.16 | 93.91 | 3,784 | 48.75 | 71.00 | 42.59 | 92.75 | 3,750 |
| SK | 61.71 | 79.10 | 30.82 | 94.18 | 3,776 | 38.77 | 63.82 | 52.65 | 92.55 | 3,749 |
| SL | 62.03 | 79.67 | 32.23 | 93.81 | 3,808 | 35.56 | 63.03 | 57.08 | 91.71 | 3,773 |
| SV | 63.51 | 81.91 | 28.18 | 93.04 | 3,764 | 40.69 | 67.26 | 48.97 | 91.66 | 3,750 |
| **ALL** | **61.98** | **80.32** | **30.91** | **92.17** | **86,883** | **41.34** | **66.61** | **49.61** | **90.70** | **86,529** |

**Table 15:** Non-legislative fine-tuning: per-language results on non-legislative test set (left) and legislative test set (right). Seed = 42.

| | Test: Non-Legislative | | | | | Test: Legislative | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| **Lang** | **BLEU** | **chrF** | **TER** | **COMET** | $N$ | **BLEU** | **chrF** | **TER** | **COMET** | $N$ |
| BG | 54.91 | 74.87 | 35.49 | 93.78 | 3,742 | 58.52 | 77.06 | 33.16 | 92.68 | 3,782 |
| CS | 37.17 | 62.24 | 54.60 | 93.35 | 3,747 | 51.89 | 71.78 | 40.81 | 93.08 | 3,765 |
| DA | 52.34 | 73.90 | 38.18 | 93.11 | 3,759 | 56.85 | 76.54 | 36.21 | 91.74 | 3,734 |
| DE | 39.13 | 65.88 | 52.38 | 89.52 | 3,736 | 47.10 | 71.58 | 44.85 | 88.13 | 3,777 |
| EL | 48.20 | 69.09 | 42.01 | 92.86 | 3,804 | 59.05 | 76.58 | 32.02 | 92.50 | 3,752 |
| ES | 56.33 | 74.90 | 34.53 | 90.92 | 3,748 | 62.20 | 78.86 | 29.31 | 89.11 | 3,777 |
| ET | 30.20 | 62.54 | 62.56 | 93.04 | 3,752 | 38.68 | 68.09 | 54.37 | 93.13 | 3,791 |
| FI | 30.55 | 63.53 | 62.90 | 93.68 | 3,755 | 35.99 | 67.14 | 57.50 | 92.89 | 3,805 |
| FR | 48.91 | 70.70 | 43.09 | 89.69 | 3,742 | 59.86 | 78.49 | 30.88 | 89.57 | 3,799 |
| GA | 45.11 | 69.11 | 46.56 | 86.77 | 3,784 | 51.46 | 73.57 | 40.63 | 86.07 | 3,781 |
| HR | 42.83 | 68.12 | 47.49 | 93.49 | 3,772 | 49.42 | 71.54 | 42.20 | 92.84 | 3,780 |
| HU | 32.71 | 62.17 | 59.49 | 92.02 | 3,759 | 38.97 | 67.47 | 53.51 | 91.34 | 3,765 |
| IT | 53.19 | 74.15 | 37.55 | 92.32 | 3,762 | 59.74 | 78.77 | 31.66 | 91.23 | 3,777 |
| LT | 37.82 | 66.05 | 54.04 | 93.25 | 3,778 | 43.25 | 70.78 | 49.77 | 92.40 | 3,759 |
| LV | 38.18 | 65.61 | 53.60 | 92.96 | 3,794 | 47.73 | 72.02 | 46.68 | 92.37 | 3,759 |
| MT | 53.97 | 79.26 | 34.01 | 77.79 | 3,757 | 53.89 | 81.33 | 33.07 | 77.78 | 3,761 |
| NL | 41.20 | 67.30 | 49.62 | 90.84 | 3,789 | 53.49 | 74.67 | 38.75 | 89.99 | 3,837 |
| PL | 38.41 | 64.36 | 53.18 | 92.88 | 3,776 | 49.94 | 71.46 | 41.63 | 92.31 | 3,779 |
| PT | 59.87 | 77.29 | 30.60 | 91.83 | 3,751 | 62.21 | 79.16 | 29.58 | 89.86 | 3,771 |
| RO | 51.64 | 72.81 | 39.19 | 93.43 | 3,750 | 62.60 | 79.57 | 29.27 | 92.97 | 3,784 |
| SK | 42.73 | 66.52 | 48.87 | 93.49 | 3,749 | 53.68 | 73.29 | 38.44 | 93.12 | 3,776 |
| SL | 39.34 | 65.68 | 53.08 | 92.57 | 3,773 | 51.59 | 72.50 | 42.37 | 92.47 | 3,808 |
| SV | 44.11 | 69.50 | 45.81 | 92.51 | 3,750 | 53.80 | 74.65 | 36.70 | 91.38 | 3,764 |
| **ALL** | **45.13** | **69.04** | **46.09** | **91.57** | **86,529** | **53.12** | **74.32** | **38.82** | **90.82** | **86,883** |

**Table 16:** Sequential fine-tuning (gen. $\rightarrow$ leg.): per-language results on legislative (left) and non-legislative (right) test sets. Seed = 42.

| | Test: Legislative | | | | Test: Non-Legislative | | | |
|---|---|---|---|---|---|---|---|---|
| **Lang** | **BLEU** | **chrF** | **TER** | **COMET** | **BLEU** | **chrF** | **TER** | **COMET** |
| BG | 66.67 | 82.52 | 26.14 | 93.83 | 52.70 | 73.47 | 37.40 | 93.36 |
| CS | 62.00 | 79.20 | 30.53 | 94.42 | 35.85 | 61.64 | 55.29 | 93.00 |
| DA | 66.33 | 82.90 | 27.46 | 93.11 | 51.21 | 73.26 | 38.64 | 92.80 |
| DE | 55.48 | 77.70 | 36.78 | 89.79 | 37.70 | 64.76 | 53.66 | 89.04 |
| EL | 68.12 | 82.40 | 24.46 | 93.63 | 46.31 | 68.08 | 43.36 | 92.43 |
| ES | 70.59 | 84.00 | 22.53 | 90.67 | 54.93 | 73.99 | 35.72 | 90.56 |
| ET | 48.53 | 76.19 | 44.16 | 94.50 | 28.86 | 61.86 | 64.42 | 92.54 |
| FI | 48.58 | 76.53 | 44.57 | 94.44 | 28.45 | 62.89 | 65.07 | 93.24 |
| FR | 70.13 | 84.19 | 23.41 | 90.97 | 47.83 | 70.14 | 44.65 | 89.36 |
| GA | 63.31 | 81.50 | 30.10 | 88.20 | 43.69 | 68.53 | 48.01 | 86.29 |
| HR | 61.83 | 80.10 | 30.56 | 94.34 | 41.06 | 67.14 | 49.01 | 92.95 |
| HU | 50.80 | 75.82 | 42.46 | 92.93 | 31.09 | 61.32 | 61.43 | 91.47 |
| IT | 67.87 | 83.60 | 24.98 | 92.41 | 51.64 | 73.30 | 39.07 | 91.88 |
| LT | 52.54 | 76.90 | 40.59 | 93.46 | 35.89 | 64.85 | 55.44 | 92.66 |
| LV | 56.08 | 77.98 | 37.90 | 93.35 | 36.25 | 64.44 | 54.62 | 92.18 |
| MT | 71.60 | 89.34 | 19.51 | 81.46 | 50.61 | 77.99 | 36.39 | 77.38 |
| NL | 62.87 | 81.18 | 30.05 | 91.43 | 40.27 | 67.00 | 50.38 | 90.59 |
| PL | 59.47 | 78.12 | 33.38 | 93.62 | 36.03 | 63.56 | 55.94 | 92.57 |
| PT | 70.84 | 84.43 | 22.63 | 91.31 | 58.28 | 76.48 | 31.83 | 91.59 |
| RO | 71.95 | 85.37 | 22.17 | 94.04 | 50.21 | 72.05 | 41.39 | 93.04 |
| SK | 62.87 | 79.85 | 29.78 | 94.29 | 40.45 | 65.28 | 50.92 | 93.01 |
| SL | 63.51 | 80.66 | 30.70 | 93.95 | 38.00 | 64.91 | 54.61 | 92.16 |
| SV | 64.59 | 82.59 | 27.04 | 93.16 | 42.59 | 68.79 | 47.21 | 92.14 |
| **ALL** | **63.21** | **81.08** | **29.79** | **92.32** | **43.53** | **68.17** | **47.61** | **91.14** |

**Table 17:** EuroLLM-22B-Instruct base model (no fine-tuning): per-language results on legislative (left), non-legislative (centre), and Flores+ (right) test sets.

| | Test: Legislative | | | | Test: Non-Legislative | | | | Test: Flores+ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Lang** | **BLEU** | **chrF** | **TER** | **COMET** | **BLEU** | **chrF** | **TER** | **COMET** | **BLEU** | **chrF** | **TER** | **COMET** |
| BG | 53.83 | 74.73 | 36.99 | 92.26 | 46.56 | 69.60 | 41.96 | 92.78 | 42.04 | 67.87 | 46.82 | 91.75 |
| CS | 49.87 | 70.64 | 41.72 | 92.96 | 31.66 | 58.27 | 58.72 | 92.44 | 33.56 | 59.95 | 54.66 | 92.24 |
| DA | 52.41 | 74.21 | 39.28 | 91.60 | 45.40 | 69.38 | 43.35 | 92.24 | 48.21 | 70.65 | 39.30 | 91.78 |
| DE | 46.10 | 71.29 | 45.15 | 88.25 | 34.16 | 62.19 | 57.48 | 88.57 | 42.35 | 67.76 | 45.15 | 88.98 |
| EL | 54.42 | 73.98 | 35.41 | 92.08 | 40.79 | 63.80 | 48.51 | 91.71 | 29.17 | 55.38 | 56.54 | 90.14 |
| ES | 58.18 | 76.98 | 31.68 | 88.95 | 49.47 | 70.46 | 40.42 | 89.86 | 30.38 | 57.65 | 55.25 | 87.40 |
| ET | 35.81 | 67.05 | 56.67 | 92.87 | 24.61 | 58.61 | 68.56 | 92.02 | 28.86 | 61.59 | 59.18 | 92.10 |
| FI | 33.95 | 66.21 | 59.75 | 92.88 | 25.18 | 60.00 | 67.53 | 93.01 | 26.73 | 60.81 | 61.86 | 93.03 |
| FR | 55.63 | 76.91 | 34.41 | 89.31 | 39.24 | 65.62 | 51.20 | 88.61 | 53.78 | 73.58 | 35.75 | 89.20 |
| GA | 43.87 | 68.72 | 47.76 | 85.01 | 35.44 | 62.65 | 54.06 | 84.86 | 30.79 | 58.33 | 58.92 | 81.32 |
| HR | 44.81 | 68.63 | 46.43 | 92.44 | 33.58 | 61.88 | 55.84 | 92.18 | 29.18 | 58.75 | 59.69 | 90.91 |
| HU | 35.47 | 65.62 | 57.22 | 91.00 | 26.58 | 57.90 | 64.85 | 90.94 | 26.62 | 58.53 | 62.28 | 90.33 |
| IT | 55.68 | 76.98 | 34.12 | 91.00 | 46.28 | 69.86 | 43.69 | 91.17 | 33.04 | 61.44 | 53.84 | 89.45 |
| LT | 40.05 | 68.84 | 51.98 | 91.95 | 30.15 | 60.88 | 60.70 | 91.92 | 28.03 | 59.78 | 61.31 | 91.01 |
| LV | 45.09 | 70.85 | 48.72 | 92.03 | 32.25 | 61.94 | 58.90 | 91.78 | 32.32 | 61.82 | 56.80 | 91.15 |
| MT | 52.20 | 79.07 | 35.87 | 76.72 | 44.04 | 74.01 | 42.06 | 75.40 | 41.02 | 70.39 | 43.93 | 72.56 |
| NL | 51.35 | 73.85 | 39.55 | 90.01 | 35.82 | 63.92 | 54.45 | 90.07 | 28.80 | 59.61 | 59.37 | 88.74 |
| PL | 48.61 | 70.94 | 43.00 | 92.25 | 32.35 | 60.95 | 58.86 | 92.00 | 24.14 | 54.01 | 66.59 | 90.53 |
| PT | 54.29 | 74.94 | 34.32 | 89.23 | 44.74 | 68.27 | 42.20 | 90.05 | 51.89 | 72.70 | 35.36 | 90.46 |
| RO | 60.09 | 78.33 | 31.20 | 92.82 | 45.82 | 69.16 | 43.48 | 92.61 | 43.41 | 66.69 | 44.86 | 91.48 |
| SK | 51.31 | 71.99 | 39.92 | 92.90 | 37.24 | 62.79 | 53.36 | 92.55 | 35.92 | 61.89 | 53.14 | 91.39 |
| SL | 49.99 | 71.81 | 43.05 | 92.30 | 33.74 | 61.62 | 58.49 | 91.44 | 32.29 | 59.66 | 57.68 | 90.71 |
| SV | 52.38 | 74.42 | 37.10 | 91.37 | 38.73 | 65.85 | 50.13 | 91.52 | 48.08 | 70.70 | 38.26 | 91.92 |
| **ALL** | **49.66** | **72.57** | **41.41** | **90.53** | **37.85** | **64.42** | **52.25** | **90.42** | **36.33** | **63.04** | **51.98** | **89.50** |

**Table 18:** Claude Sonnet 4.6 (zero-shot): per-language results on legislative (left) and non-legislative (right) test sets.

| | Test: Legislative | | | | Test: Non-Legislative | | | |
|---|---|---|---|---|---|---|---|---|
| **Lang** | **BLEU** | **chrF** | **TER** | **COMET** | **BLEU** | **chrF** | **TER** | **COMET** |
| BG | 58.50 | 77.81 | 32.36 | 92.96 | 46.56 | 69.60 | 41.96 | 92.78 |
| CS | 53.88 | 73.87 | 37.62 | 93.52 | 31.66 | 58.27 | 58.72 | 92.44 |
| DA | 58.29 | 78.35 | 34.53 | 92.32 | 45.40 | 69.38 | 43.35 | 92.24 |
| DE | 50.18 | 74.52 | 41.17 | 89.08 | 34.16 | 62.19 | 57.48 | 88.57 |
| EL | 58.12 | 77.04 | 30.81 | 92.66 | 40.79 | 63.80 | 48.51 | 91.71 |
| ES | 62.42 | 79.57 | 28.28 | 89.47 | 49.47 | 70.46 | 40.42 | 89.86 |
| ET | 40.92 | 70.95 | 51.28 | 93.67 | 24.61 | 58.61 | 68.56 | 92.02 |
| FI | 41.47 | 72.43 | 51.22 | 93.96 | 25.18 | 60.00 | 67.53 | 93.01 |
| FR | 58.55 | 79.06 | 31.60 | 89.94 | 39.24 | 65.62 | 51.20 | 88.61 |
| GA | 50.83 | 74.25 | 39.64 | 86.40 | 35.44 | 62.65 | 54.06 | 84.86 |
| HR | 52.01 | 73.97 | 38.96 | 93.49 | 33.58 | 61.88 | 55.84 | 92.18 |
| HU | 43.13 | 71.16 | 48.95 | 92.16 | 26.58 | 57.90 | 64.85 | 90.94 |
| IT | 58.63 | 79.02 | 31.87 | 91.36 | 46.28 | 69.86 | 43.69 | 91.17 |
| LT | 43.03 | 71.22 | 48.60 | 92.56 | 30.15 | 60.88 | 60.70 | 91.92 |
| LV | 47.88 | 73.09 | 46.40 | 92.64 | 32.25 | 61.94 | 58.90 | 91.78 |
| MT | 52.78 | 81.38 | 33.01 | 77.65 | 44.04 | 74.01 | 42.06 | 75.40 |
| NL | 55.40 | 76.93 | 36.24 | 90.56 | 35.82 | 63.92 | 54.45 | 90.07 |
| PL | 52.01 | 73.74 | 39.01 | 92.73 | 32.35 | 60.95 | 58.86 | 92.00 |
| PT | 60.95 | 79.11 | 29.02 | 90.19 | 44.74 | 68.27 | 42.20 | 90.05 |
| RO | 62.35 | 80.14 | 28.77 | 93.17 | 45.82 | 69.16 | 43.48 | 92.61 |
| SK | 52.82 | 73.46 | 38.24 | 93.25 | 37.24 | 62.79 | 53.36 | 92.55 |
| SL | 53.09 | 74.18 | 39.97 | 92.97 | 33.74 | 61.62 | 58.49 | 91.44 |
| SV | 55.80 | 77.13 | 34.34 | 91.97 | 38.73 | 65.85 | 50.13 | 91.52 |
| **ALL** | **53.90** | **75.85** | **37.14** | **91.25** | **37.85** | **64.42** | **52.25** | **90.42** |

**Table 19:** Training configuration for all fine-tuning objectives. Sequential fine-tuning uses two stages: legislative and generic fine-tuning, each use a single stage with identical settings (only the input data differs).

| **Parameter** | **Sequential Stage 1** | **Sequential Stage 2** | **Leg./Gen. single-stage** |
|---|---|---|---|
| *Model and adapter* | | | |
| Base model | EuroLLM-22B-Instruct | | |
| LoRA targets | `q, k, v, o, gate, up, down_proj` | | |
| LoRA $r$ / $\alpha$ | 32 / 64 | | |
| Adapter init | Gaussian ($A: \mathcal{N}(0,1)$, $B: \mathbf{0}$) | | Kaiming uniform |
| rsLoRA | ✓ | | × |
| Precision / Attention | `bfloat16` / FlashAttention-2 | | |
| Grad. checkpointing | ✓ | | |
| *Training hyperparameters (from Optuna)* | | | |
| Learning rate | $1.166 \times 10^{-4}$ | $6.66 \times 10^{-5}$ ($\eta_1/3$) | $1.998 \times 10^{-4}$ |
| Warmup ratio | 0.0466 | 0.1102 | 0.1102 |
| LoRA dropout | 0.0971 | 0.0453 | 0.0453 |
| Optimizer / schedule | AdamW / Cosine | | |
| Max sequence length | 512 tokens | | |
| Max epochs | 5 | | |
| Max gradient norm | 1.0 | | |
| Eval frequency | every 200 steps | | |
| Checkpoint selection | min. validation loss | | |
| Early stopping | 3 evals (thr. $10^{-3}$) | | 2 evals |
| Seed | 42 | | |
| *Batch size and hardware* | | | |
| Batch / GPU | 8 | | 8 |
| Grad. accumulation | 12 steps | | 2 steps |
| GPUs | 4×RTX PRO 6000 Blackwell (98 GiB) | | 6×H200 NVL (141 GiB) |
| Effective batch | $8 \times 12 \times 4 = 384$ | | $8 \times 2 \times 6 = 96$ |
| Parallelism | PyTorch DDP, single-node, `torchrun` | | |

### A.5.1 Non-legislative Fine-Tuning

Generic fine-tuning follows the same setup as legislative fine-tuning described above. The model architecture, hyperparameters, and hardware configuration remain identical to those shown in Table 19. The only difference is that the training and evaluation data are drawn from the fil-

tered non-legal corpus instead of the legal corpus. This setup enables a direct comparison between domain-specific and general-domain adaptation under consistent conditions.