EUROALPACA: Task-Preserving Localisation of Instruction Data for European Languages

Aleix Sant[1,2], **Jordi Luque**[1], **Carlos Escolano**[2]

[1]Scientific Research, Telefónica Innovación Digital
[2]Universitat Politècnica de Catalunya
Barcelona, Spain

{aleix.santsavall,jordi.luque}@telefonica.com
carlos.escolano@upc.edu

Abstract

Machine translation (MT) offers a scalable way to extend English instruction-tuning data to multiple languages, but it can distort task-critical constraints and required outputs, creating corrupted training examples and degrading models trained on such data. We introduce EUROALPACA, a task-preserving localisation pipeline and near-parallel resource covering 50 European languages and regional varieties, together with EUROPEAN-IFEVAL, a multilingual benchmark for verifiable instruction following. Depending on the example, our pipeline applies field-wise MT while preserving task-critical content or reconstructs a task-equivalent target-language instance, followed by validation of cross-field coherence and target-language consistency. Across LoRA experiments with four LLMs, training on directly translated data improves ROUGE-L and FBERT on the Aya Evaluation Suite, but reduces accuracy on EUROPEAN-IFEVAL by 29.8% relative to the unadapted baseline. In contrast, adaptation with EUROALPACA improves accuracy by 12.9% over the same baseline, reversing the degradation caused by direct MT, while also achieving the highest ROUGE-L and FBERT scores on Aya. These results show that preserving task semantics is essential for multilingual instruction tuning.

1 Introduction

Instruction tuning enables language models to follow diverse natural-language requests and generalise to unseen tasks (Wei et al., 2022; Ouyang et al., 2022). However, open instruction-tuning resources remain predominantly English-centric. These corpora encompass heterogeneous tasks, including question answering, summarisation, classification, reasoning, rewriting, constrained generation and code-related tasks. A common scalable strategy for extending them to new languages is to machine-translate English instruction datasets, including the textual fields of each instruction–input–output record (Upadhayay and Behzadan, 2023; Singh et al., 2024; Chen et al., 2024a; Shaham et al., 2024). Although scalable, this strategy conflates preserving sentence meaning with preserving the task encoded by the complete example (Roy et al., 2021).

The distinction matters when a task depends on linguistic form or on the role of a particular field. Previous work has shown that translated instruction data can corrupt language-dependent tasks such as grammatical-error correction, phonological constraints and examples containing code or language-specific knowledge (Chen et al., 2024b; Li et al., 2024). For example, translating an intentionally ungrammatical sentence may silently correct the error that the model is expected to identify. Similarly, constraints involving word length, rhyme, alliteration or acronym formation are not generally invariant under MT, because they depend on the linguistic form of the target-language expression. Translating classification labels, source code, URLs, paths or required output strings can likewise make an otherwise fluent example inconsistent. In such cases, we use the term *task-preserving localisation* to mean constructing a target-language instruction–input–output triple that preserves the underlying task, its explicit constraints, its difficulty and the answer relation, even when its surface form is not a literal translation.

In line with Skadins et al., 2014, we argue that task preservation is a central requirement for synthetic instruction data adaptation in new and low-resource languages. Fluency alone is insufficient. Translated supervision may appear natural while weakening a model's ability to satisfy verifiable instructions. We therefore formulate multilingual instruction-data generation as a structured decision process rather than uniform machine translation. Each example is treated as a structured instruction-following instance, and LLM-based judge modules

Proceedings of the 19th International Natural Language Generation Conference: Long Papers, pages 655–682
October 17-21, 2026 ©2026 Association for Computational Linguistics

determine whether it is suitable for direct translation, which fields must be preserved, and when task-level rewriting is required.

Starting from the 51,760 examples in Alpaca Cleaned[1], we build EUROALPACA. Our proposed data processing pipeline (i) annotates the task and domain, (ii) decides whether the sample can be machine-translated, (iii) judges which fields must remain unchanged, (iv) translates or rewrites the example for a target language ℓ, and (v) validates the complete target-language triple for coherence and language consistency. The resulting resource is near-parallel, i.e. examples remain aligned by source identifier and task intent, but language-dependent content may differ when literal equivalence would break the task. This makes task-preserving localisation a constrained NLG problem rather than a purely translation-based one, since some target-language content must be adapted or reconstructed to preserve the task.

We evaluate the usefulness of EUROALPACA, our localised version of the Alpaca Cleaned instruction dataset, by language-specific fine-tuning across four contemporary, state-of-the-art small-sized language models. By leveraging the Aya Evaluation Suite alongside our newly curated EUROPEAN-IFEVAL benchmark, we provide a comprehensive analysis of 50 European languages, which reveals an evaluation mismatch. Direct MT improves reference similarity but substantially reduces instruction-following accuracy. In contrast, models trained with EUROALPACA consistently improve both, supporting our assumption that multilingual instruction data should preserve task function rather than only target-language fluency and sentence meaning. Our contributions are:

- An LLM-guided pipeline for task-preserving multilingual instruction-data generation.
- EUROALPACA, a source-aligned, near-parallel resource covering English and 50 European languages and regional varieties.
- EUROPEAN-IFEVAL, a normalised multilingual instruction-following benchmark collection with explicit source provenance across the same language set.
- Controlled language-specific experiments across four instruction-tuned models.
- Matched comparisons with multilingual instruction datasets, including Bactrian-X, Okapi and MITS.

[1] https://github.com/gururise/AlpacaDataCleaned

2 Related Work

Synthetic instruction data. Self-Instruct and Alpaca showed that model-generated instructions can cheaply expand supervised instruction-tuning data (Wang et al., 2023; Taori et al., 2023). Alpaca Cleaned corrected flawed records from the original release. These resources follow a simple schema consisting of an *instruction*, which defines the task to be performed, an optional *input*, which provides additional context for the task, and an *output*, which contains the reference response expected from the model.

Multilingual instruction-tuning resources. Several efforts extend English instruction data to multiple languages through translation, templating, human curation and preference annotation. Okapi translates Alpaca-52K into 26 languages using ChatGPT and provides multilingual response-ranking data for RLHF (Lai et al., 2023). MITS translates Alpaca-52K-GPT4 and Dolly-15K into 132 languages with Google Cloud Translation (Upadhayay and Behzadan, 2023). Bactrian-X translates Alpaca- and Dolly-style prompts and pairs them with target-language responses generated by `gpt-3.5-turbo` (Li et al., 2023). The Aya Collection combines multilingual instruction–response data from multiple sources and construction strategies (Singh et al., 2024).

Instruction evaluation beyond English. IFEval evaluates compliance with objectively verifiable instructions rather than relying solely on reference-based answer similarity (Zhou et al., 2023). Multilingual extensions include M-IFEval, which covers French, Japanese and Spanish, and Marco-Bench-MIF, a localised benchmark covering 30 languages (Dussolle et al., 2025; Zeng et al., 2025), and XIF-Bench (Li et al., 2025b), introducing fine-grained multilingual constraint analysis. The Aya Evaluation Suite complements constraint-based evaluation with multilingual open-ended generation tasks and reference-based metrics (Singh et al., 2024). LLM-as-a-judge approaches provide an alternative for open-ended evaluation, although their judgements may exhibit evaluator and presentation biases (Liu et al., 2023; Zheng et al., 2023).

From direct translation to task localisation. Direct translation provides a scalable way to construct multilingual instruction data, but it can introduce translationese, create knowledge–language mismatches and corrupt language-dependent tasks

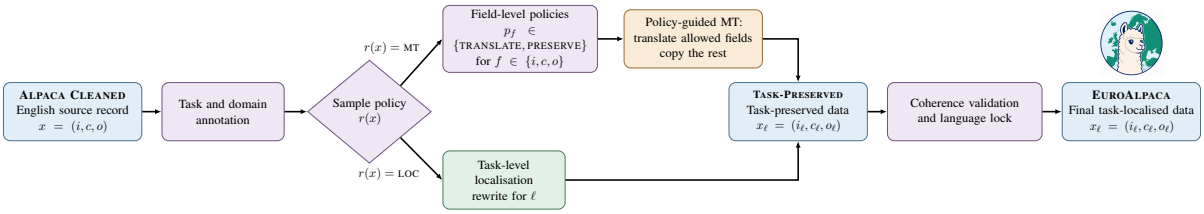


Figure 1: Task-preserving instruction-data generation pipeline. Blue nodes are data artefacts, purple are LLM stages, orange is policy-guided MT, and green is task-level rewriting. Gated routing directs samples to either per-field translation (TRANSLATE/PRESERVE) or task-level localisation, yielding TASK-PRESERVED which is validated to produce EUROALPACA for language ℓ.

such as grammatical error correction (Artetxe et al., 2020; Riley et al., 2020; Chen et al., 2024b; Hupkes and Bogoychev, 2025). Standard supervised fine-tuning (SFT) may further amplify translationese biases present in translated training data (Li et al., 2025a). Training-side methods such as X-CIT improve learning from parallel translated instructions but do not directly address whether translation preserves the validity of the underlying task (Wu et al., 2025). Other approaches move beyond direct translation by deriving instruction–response pairs from human-written target-language texts or by localising communicative intent and language- or culture-dependent content (Köksal et al., 2025; Vasselli et al., 2025; Maheshwary et al., 2025). Building on this shift towards localisation, our pipeline instead targets task-preserving localisation of *instruction–input–output* triples.

3 Task-Preserving Localisation

Let a source record be $x = (i, c, o)$, containing an instruction i, an optional input c and an output o. For a target language ℓ, the goal is to construct x_ℓ such that the task relation $T(i, c, o)$ remains valid. Literal semantic equivalence of every field is neither necessary nor sufficient.

We use a hierarchical routing process. First, a sample-level policy $r(x) \in \{\text{MT}, \text{LOC}\}$ determines whether the record is suitable for the field-wise MT route or instead requires task-level localisation. Only samples routed to MT undergo a second pass, which assigns each non-empty field $f \in \{i, c, o\}$ a policy $p_f \in \{\text{TRANSLATE}, \text{PRESERVE}\}$. Samples routed to LOC bypass field-wise MT and are reconstructed as task-equivalent target-language instances.

3.1 When Translation Breaks the Task

Table 2 shows the operational taxonomy used by the sample policy prediction. It distinguishes cases that require target-language rewriting from cases in which specific fields should remain unchanged.

Handling regime	Count	%
Direct translation	27,946	54.0
Field-preserving translation	12,726	24.6
Task-level localisation	11,088	21.4
Total	51,760	100.0

Table 1: Handling regimes for the 51,760 source examples. Direct translation translates all non-empty fields. Field-preserving translation retains at least one field unchanged. Task-level localisation reconstructs examples for which field-wise MT would invalidate the task.

Table 3 provides schematic illustrations of these examples, while Tables 11 and 12 in the Appendix present concrete examples of the different possible handling regimes. The appropriate handling depends on where task-critical information is encoded across the *instruction*, *input* and *output* fields.

3.2 Task-Preserving Data Generation Pipeline

Figure 1 illustrates the complete pipeline. We use an 8-bit quantised variant of `google/gemma-4-31B-it` (Google DeepMind, 2026) as the LLM-based decision module, with specialised prompts for semantic annotation, sample- and field-level policy predictions, task-preserving localisation (i.e., rewriting in the target language) and consistency validation with target-language enforcement. Complete inference settings and prompt templates for the different steps are reported in Appendix D. Prompt templates were iteratively refined through manual inspection of sampled outputs. LLM outputs are parsed as JSON and normalised against stage-specific formats defined in the prompts. Malformed outputs are handled conservatively. Sample policy prediction failures fall back to non-translatable decisions, while task-level localisation and consistency validation retry with an explicit repair prompt.

Failure type	Typical cases	Handling policy
Source-language form dependence	Grammar or spelling correction, punctuation judgements, tense/voice edits, alphabetic ordering	Generate an analogous target-language task and recompute the output.
Lexical or phonological constraints	Letter/character counts, word counts, rhymes, acrostics, mnemonics, puns, anagrams	Replace the constrained material with target-language content satisfying the same constraint.
Language- or culture-specific expressions	Idioms, proverbs, fixed expressions, English-only prompts, language-specific examples	Substitute a natural target-language analogue while preserving task intent and difficulty.
Protected literals	Code, URLs, paths, identifiers, formulas, quoted strings, names, titles, table columns	Preserve fields containing task-critical literal artefacts.
Fixed output vocabulary or schema	Classification labels, JSON keys, structured outputs, canonical options	Preserve fields defining or containing the required labels or output format.
Field-role dependence	Translation tasks, demonstrations, input–output examples, source/target language roles	Preserve the role of each field rather than translating all fields uniformly.

Table 2: Operational taxonomy of examples requiring non-uniform translation across fields.

Source task	Uniform MT failure	Task-preserving construction
Correct the grammar in an intentionally erroneous sentence.	Translation may silently remove the error, leaving nothing to correct.	Create a comparable target-language error and provide its correction.
Return a word that rhymes with a given word.	Rhyme depends on target-language phonology. Translated words may not rhyme.	Select a target-language cue and answer satisfying the same relation.
Write exactly ten words about a topic.	Word counts are not preserved under translation.	Generate target-language text satisfying the count.
Classify with exact labels.	Translating labels can break the required output vocabulary or automatic checker.	Translate the content but preserve the labels verbatim.
Interpret an idiom.	Literal translation may produce a nonsensical expression or a different idiom.	Replace it with a natural target-language idiom and explain that expression.
Translate a sentence from language A to language B.	Translating every field can invalidate the specified language pair.	Translate the instruction, but preserve the source and target texts.
Write a function in a specified programming language for a given operation.	Translating code can alter keywords, identifiers, literals or syntax.	Translate the natural-language instruction, but preserve the generated code exactly.

Table 3: Schematic examples where direct field-wise translation can break the target-language instruction task.

Semantic annotation. For each source example, we prompt Gemma-4 to assign one primary task and primary domain, as well as optional secondary tasks and domains, using predefined taxonomies (Appendix A.1). These annotations support corpus analysis and inform later sample- and field-level translation-policy predictions. In these later stages, they are provided as context to Gemma-4 and used by deterministic guardrails. The pipeline ensures that all predicted labels belong to the predefined taxonomies. For example, it replaces missing or invalid primary labels with *other* for tasks and *general* for domains. The resulting distributions reveal substantial imbalances across tasks and domains.

Sample- and field-level policy predictions. The sample-level gate first decides whether a record is suitable for field-wise MT or instead requires task-level localisation through rewriting. This gate combines a global include/exclude prediction from Gemma-4 with deterministic guardrails that can override the model's prediction and reject samples whose task depends on source-language form, such as spelling/grammar exercises, rhyme, mnemonics, idioms... For samples that pass the gate, Gemma-4 predicts a translation policy for each field: *instruction*, *input* and *output*. Deterministic field-level guardrails then adjust these predictions when necessary. Table 1 reports the resulting regimes.

Policy-guided MT and task localisation. For MT-routed samples, NLLB-3.3B (Costa-jussà et al., 2022) translates fields assigned TRANSLATE, while fields assigned PRESERVE are copied unchanged. Samples routed to localisation are instead reconstructed by Gemma-4 using a curated prompt, preserving the task type, difficulty, expected answer form and explicit constraints in the target language. These two routes then produce the intermediate TASK-PRESERVED artefact, which preserves task function. Consequently, some MT-routed samples may retain English content in preserved fields when translating it would invalidate the task.

Consistency validation and target-language enforcement. In the final step, Gemma-4 examines the complete *instruction–input–output* triple for semantic contradictions, unclear wording, grammatical or orthographic errors, literal calques, mixed-language content and instruction–response mismatches. It enforces target-language coherence, applies recoverable edits and records post-edit metadata, while preserving non-target-language spans if a specific task requires it.

4 Generated Resources

4.1 Near-Parallel Instruction Data

We run the pipeline for 50 European languages and regional varieties, in addition to the annotated English source[2]. Each target record retains the source identifier, language tag, semantic labels, whole-sample and field policies, validation outcome and post-edit flags. The collection is *near-parallel*, with records aligned by source identifier and task intent while allowing lexical or structural differences when literal equivalence would invalidate the task.

The language set is European-centric but typologically diverse, as shown in Figure 2 and Appendix Table 13. It includes major national languages as well as regional and minority varieties from several language families and branches, including Romance, Germanic, Slavic, Celtic, Baltic, Uralic, Hellenic, Albanian, Semitic and Kartvelian, together with language isolates. It also spans the Latin, Cyrillic, Greek and Georgian scripts. We define resource tiers using clean monolingual token counts from MADLAD-400 (Kudugunta et al., 2023), with high denoting ≥ 5 billion tokens, medium denoting 100 million to < 5 billion, and low denoting < 100 million tokens or no separate clean monolingual partition in MADLAD-400. Combined with the near-parallel design of EUROALPACA, this diversity enables controlled multilingual instruction-following experiments across languages that differ in typology, script and resource availability, while keeping the underlying supervision aligned.

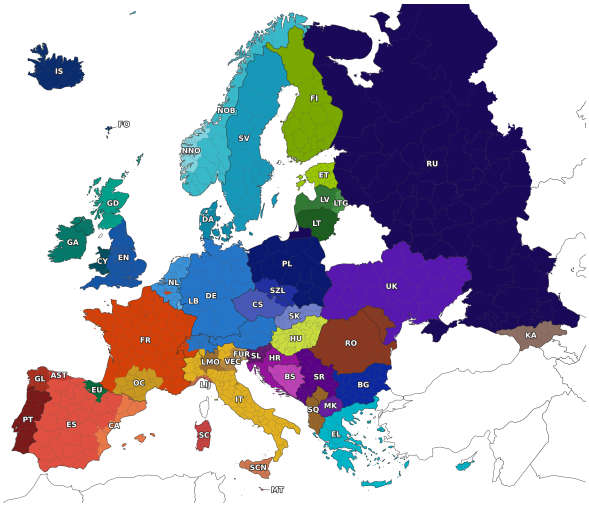


Figure 2: Approximate traditional geographic areas associated with the 50 European languages and regional varieties covered by EUROALPACA and EUROPEAN-IFEVAL. Boundaries are schematic and do not represent exclusive or current speaker distributions. The map is based on FUEN, 2013.

4.2 EUROPEAN-IFEVAL

EUROPEAN-IFEVAL is a separate evaluation resource, not an output of the EUROALPACA localisation pipeline[3]. It brings together existing IFEval-style resources for the same 50 target languages and converts them to a common evaluation format. For each language, we select one source using a fixed quality-oriented hierarchy. We first prioritise native-speaker-curated or professionally translated language-specific releases. When these are unavailable, we use Marco-Bench-MIF, which localises linguistic constraints and cultural references through automated processing followed by two rounds of human verification (Zeng et al., 2025). For the remaining languages, we use MultiIFEval, which translates and localises English IFEval prompts with an LLM while providing target-language Wikipedia content as contextual support.[4]

We convert the gathered datasets to a common record format containing the prompt, instruction-

[2] https://github.com/Telefonica-Scientific-Research/EuroAlpaca

[3] https://github.com/Telefonica-Scientific-Research/european-ifeval

[4] https://huggingface.co/datasets/danish-foundation-models/multi-ifeval

checker identifiers and their arguments, a sample identifier and an indicator of whether the response is required to be in the target language. A separate manifest records, for each language, the source, provenance category, missing source items and information about explicit response-language constraints. We harmonise checker identifiers and argument formats so that all datasets can be scored with a common evaluator. Response-language constraints are evaluated using GlotLID-M, selected for its coverage of low-resource languages and language varieties (Kargaran et al., 2023).

5 Experimental Setup

5.1 Models and Supervision Conditions

We evaluate four multilingual text-only instruction-tuned models in the 3–4B parameter range: Tiny-Aya-Global (Salamanca et al., 2026), Phi-4-Mini-instruct (Abouelenin et al., 2025), Ministral-3-3B-Instruct-2512 (Mistral AI, 2025) and Llama-3.2-3B-Instruct (Meta AI, 2024).

For each of the 200 model–language pairs (four models and 50 languages), we train three separate language-specific LoRA adapters (Hu et al., 2022), one for each supervision condition described below, yielding 600 fine-tuning runs in total. Each adapter is trained on 50,760 examples and validated on the same 1,000 held-out examples for that language. We also evaluate the unadapted checkpoint as a baseline. The three supervision conditions contrast: (i) **Direct MT**, translating all fields uniformly with NLLB-3.3B; (ii) **Task-Preserved**, an ablation using the intermediate dataset resulting from sample-level routing, field-level preservation policies and task-level localisation; and (iii) **EuroAlpaca**, the final dataset produced after validation.

Within each model–language pair, all training and evaluation settings are held fixed across supervision conditions, so performance differences primarily reflect the effect of the supervision data. We additionally fine-tune the same models on the Alpaca-derived subsets of Okapi, MITS and Bactrian-X, excluding non-Alpaca synthetic and preference data, and evaluate them using the same method. Each comparison is conducted on the languages shared by EuroAlpaca, the external resource and the corresponding benchmark.

5.2 Fine-Tuning

All runs use LoRA with rank 16, alpha 32, dropout 0.05, and adapters on query, key, value and output projections. We use AdamW, learning rate 10^{-5}, bf16 mixed precision, gradient clipping at 1.0, micro-batches of 2 and 32 gradient-accumulation steps for an effective batch size of 64. The random seed is 42. Early stopping selects the final adapter.

5.3 Evaluation and Aggregation

We use two complementary evaluation settings to assess different aspects of multilingual instruction-tuned models. The Aya Evaluation Suite (Singh et al., 2024) evaluates open-ended language generation using ROUGE-L for lexical overlap (Lin, 2004) and BERTScore for semantic similarity to reference responses (Zhang et al., 2020). We report the F1 variant of BERTScore, denoted FBERT throughout the paper.

In contrast, European-IFEval measures compliance with verifiable instructions and reports strict and loose accuracy scores. In this work, we report only strict accuracy. Together, they distinguish reference similarity from instruction following. European-IFEval covers all 50 languages, whereas Aya provides reference data for 36.

For each benchmark and metric, we macro-average scores across the available non-English languages within each model and then average the resulting means across the four models. Full per-language and per-model results are reported in Appendix C.

6 Results

6.1 Reference Similarity versus Instruction Following

As illustrated in Table 4, Direct MT improves ROUGE-L and FBERT on the Aya Evaluation Suite for all evaluated models, except for a small FBERT decrease for Llama, while substantially reducing European-IFEval accuracy for every model. These results show that greater reference similarity does not necessarily indicate better preservation of the underlying instruction-following task.

In contrast, EuroAlpaca achieves the highest scores across all models and metrics, surpassing Direct MT in the macro-average by 0.020 in Aya ROUGE-L, 0.010 in Aya FBERT and 0.132 in European-IFEval accuracy. As shown by the full language-level results in Appendix C, EuroAlpaca outperforms Direct MT in all 200 model–language comparisons. Figure 3 summarises changes relative to the off-the-shelf checkpoints.

Model	Training data	Aya ROUGE-L ↑ Mean	Aya ROUGE-L ↑ ΔOTS	Aya FBERT ↑ Mean	Aya FBERT ↑ ΔOTS	EUROPEAN-IFEVAL Acc. ↑ Mean	EUROPEAN-IFEVAL Acc. ↑ ΔOTS
Tiny-Aya-Global	Off-the-shelf	0.103	–	0.824	–	0.421	–
	Direct MT	0.228	+0.125	0.872	+0.049	0.295	−0.126
	Task-Preserved	0.232	+0.129	0.874	+0.050	0.446	+0.025
	EUROALPACA	**0.238**	**+0.135**	**0.877**	**+0.054**	**0.499**	**+0.078**
Phi-4-Mini-Instruct	Off-the-shelf	0.091	–	0.827	–	0.258	–
	Direct MT	0.198	+0.107	0.857	+0.031	0.192	−0.066
	Task-Preserved	0.200	+0.109	0.859	+0.033	0.262	+0.004
	EUROALPACA	**0.219**	**+0.128**	**0.867**	**+0.040**	**0.291**	**+0.033**
Ministral-3-3B-Instruct	Off-the-shelf	0.118	–	0.829	–	0.253	–
	Direct MT	0.191	+0.074	0.851	+0.022	0.187	−0.066
	Task-Preserved	0.193	+0.076	0.855	+0.026	0.277	+0.025
	EUROALPACA	**0.217**	**+0.100**	**0.867**	**+0.039**	**0.296**	**+0.044**
Llama-3.2-3B-Instruct	Off-the-shelf	0.186	–	0.860	–	0.305	–
	Direct MT	0.196	+0.010	0.858	−0.003	0.196	−0.109
	Task-Preserved	0.201	+0.015	0.860	+0.000	0.263	−0.042
	EUROALPACA	**0.218**	**+0.032**	**0.869**	**+0.009**	**0.310**	**+0.006**
Macro Mean (across models)	Off-the-shelf	0.124	–	0.835	–	0.309	–
	Direct MT	0.203	+0.079	0.860	+0.025	0.217	−0.092
	Task-Preserved	0.207	+0.083	0.862	+0.027	0.312	+0.003
	EUROALPACA	**0.223**	**+0.099**	**0.870**	**+0.035**	**0.349**	**+0.040**

Table 4: Language-averaged performance for each model, with macro-averages across the four models. Aya ROUGE-L and FBERT are averaged over the 36 languages with available reference data, whereas EUROPEAN-IFEVAL accuracy is averaged over all 50 evaluated languages. The ΔOTS columns report absolute score differences relative to the corresponding off-the-shelf checkpoint.

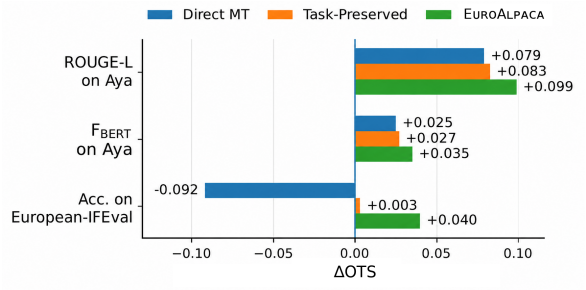


Figure 3: Macro-average ΔOTS values across the four evaluated models

6.2 Contributions of Task Preservation and Final Validation

The Task-Preserved condition isolates the sample-level gate, field-level preservation and task-level localisation before final validation. As shown in Table 4, these components improve ROUGE-L, FBERT and EUROPEAN-IFEVAL accuracy over Direct MT for all models. Notably, Task-Preserved recovers the instruction-following loss caused by Direct MT, slightly exceeds the off-the-shelf mean, and outperforms Direct MT in 199 of the 200 model–language comparisons.

Final coherence validation and target-language enforcement yield further gains across all three metrics, increasing Aya ROUGE-L by 0.016, Aya FBERT by 0.008 and EUROPEAN-IFEVAL accuracy by 0.037 at the macro-average level. EUROALPACA outperforms Task-Preserved in 180 of the 200 model–language comparisons, with one tie. These results suggest that task-aware processing prevents most translation-induced task corruption, while final validation addresses complementary errors.

6.3 Comparison with Existing Alpaca-Derived Resources

Table 5 reports results on the shared-language subset for each resource and benchmark, averaged across the four models. Because these subsets differ, each comparison is pairwise and should not be used to rank the external resources against one another. EUROALPACA outperforms Okapi and MITS on all reported metrics, except for a tie with Okapi on Aya FBERT. On EUROPEAN-IFEVAL, it improves accuracy by 0.020 over Okapi and by 0.070 over MITS. Compared with Bactrian-X, EUROALPACA achieves higher Aya scores but lower EUROPEAN-IFEVAL accuracy on the 20-language intersection (0.385 vs. 0.401).

Eval metric (Macro Avg)	n	Training data: Okapi	Training data: EUROALPACA	n	Training data: MITS	Training data: EUROALPACA	n	Training data: Bactrian-X	Training data: EUROALPACA
Aya ROUGE-L ↑	16	0.231	**0.239**	35	0.212	**0.223**	19	0.217	**0.232**
Aya FBERT ↑	16	0.879	**0.879**	35	0.864	**0.870**	19	0.871	**0.878**
EUROPEAN-IFEVAL Acc. ↑	17	0.377	**0.397**	38	0.289	**0.359**	20	**0.401**	0.385

Table 5: Pairwise comparison of LoRA fine-tuning runs on Alpaca-derived multilingual instruction resources and on our EUROALPACA. For each resource and benchmark, scores are macro-averaged over a fixed shared-language set within each model and then averaged across the four models. n denotes the number of languages shared by the resource and EUROALPACA that are also covered by the corresponding evaluation benchmark.

We also examine cross-language sample alignment in Table 6. Okapi and EUROALPACA obtain similar embedding-similarity scores (0.867 vs. 0.860), whereas Bactrian-X is less closely aligned (0.800). MITS does not expose identifiers that permit this analysis. We interpret embedding similarity as a measure of cross-language consistency rather than task validity or overall dataset quality.

Resource	Sample-level alignment	Embedding similarity
Okapi	Yes	0.867
EUROALPACA	Yes	0.860
Bactrian-X	Yes	0.800
MITS	No	–

Table 6: Sample-level alignment indicates whether corresponding records can be matched across languages. Embedding similarity is the mean cosine similarity between BGE-M3 representations of aligned instruction–input–output records over the common 11-language intersection of EUROALPACA, Okapi and Bactrian-X.

7 Discussion

The central finding is the divergence between reference- and constraint-based evaluation. Direct MT improves ROUGE-L and FBERT on Aya but causes an absolute decrease of 0.092 in EUROPEAN-IFEVAL accuracy. Thus, models trained on directly translated supervision can produce outputs that are closer to the references while becoming less reliable at following verifiable instructions. Multilingual instruction data should consequently be evaluated through both target-language response quality and verifiable instruction following.

The ablation results clarify how the pipeline addresses this problem. The Task-Preserved condition yields an accuracy gain of 0.095 over Direct MT, showing that sample-level routing, field preservation and task-level localisation prevent much of the translation-induced task corruption. Final validation adds a further gain of 0.037 and also improves Aya scores, indicating that cross-field consistency and target-language enforcement contribute beyond the initial generation policy. The two stages are therefore complementary.

Comparisons with external resources provide a broader view of the results. EUROALPACA outperforms Okapi and MITS on their respective shared-language subsets, whereas Bactrian-X achieves higher EUROPEAN-IFEVAL accuracy on its language set, possibly due to its direct target-language response generation approach. At the same time, EUROALPACA achieves cross-language embedding similarity close to Okapi's and higher than Bactrian-X's. These results show that near-parallel alignment is compatible with competitive downstream performance. However, differences in construction procedures prevent attributing the observed performance gaps solely to task-preserving localisation.

8 Conclusion

We introduce EUROALPACA, an LLM-guided pipeline and near-parallel resource for task-preserving instruction-data localisation across 50 European languages and regional varieties. The pipeline combines sample- and field-level translation policy predictions, policy-guided MT, task-level rewriting and final target-language and cross-field coherence validation. Across experiments with four LLMs, direct MT improves reference similarity on the Aya Evaluation Suite but reduces EUROPEAN-IFEVAL accuracy by 29.8% relative to the unadapted baseline. In contrast, EUROALPACA improves accuracy by 12.9% over the same baseline, fully recovering the loss caused by direct MT, while also improving open-ended generation metrics and outperforming direct MT in all 200 model–language pairs on instruction following. These findings underscore the importance of task-preserving localisation and highlight the need for complementary reference- and constraint-based evaluation.

Limitations

This study covers only European languages and regional varieties, although the proposed approach could be applied to languages from other regions. The semantic labels, predicted policies, localised examples and validation edits are model-assisted and may therefore inherit model errors or biases. A key limitation of our pipeline is its reliance on Gemma-4 (`google/gemma-4-31B-it`) for translation policy prediction, rewriting and validation. Consequently, the quality of task-preserving localisations in low-resource varieties (such as Latgalian, Lombard or Friulian) is bounded by the LLM's pre-existing exposure to and knowledge of these languages. Although we manually inspected a small sample of outputs, a larger and more systematic error analysis across languages and processing stages would provide a stronger assessment of annotation, localisation and validation quality. We report results from a single training seed for each model-language–condition combination and therefore do not estimate optimisation variance. Comparisons with external datasets are restricted to matched model–language cells and should not be interpreted as comprehensive evaluations of those resources.

Data Availability Statement

At the time of submission, EUROALPACA and EUROPEAN-IFEVAL are not publicly available in order to preserve anonymity during the double-blind review process. Upon acceptance, we plan to release both datasets, the complete data-construction code and the intermediate annotations produced at each stage of the pipeline. These materials will include task and domain labels, sample- and field-level translation policies, examples produced through task-level localisation, validation decisions and post-edit metadata. We also plan to provide the prompt templates, processing configurations, provenance manifests, evaluation scripts and documentation required to reproduce the datasets and experiments.

AI-Generated Content Disclosure

We used a generative AI to assist in paraphrasing, improving clarity and grammar in parts of the manuscript and coding assistance. All generated content was reviewed and validated by the authors.

Acknowledgment

This work has received funding from the European Union's Horizon Europe research and innovation programme under the project ELOQUENCE (Grant Agreement No. 101135916). This work was supported by computational resources from the EuroHPC Joint Undertaking under the EuroHPC AI Factory grant EHPC-AIF-2026LS01-004. Carlos Escolano is funded by Grant LLM4SL (PID2024-157855OA-C33) MICIU/AEI/10.13039/501100011033 and by "ERDF/EU".

References

Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Jun-Kun Chen, Weizhu Chen, Yen-Chun Chen, Yi-ling Chen, Qi Dai, Xiyang Dai, Ruchao Fan, and 55 others. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *CoRR*, abs/2503.01743.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation Artifacts in Cross-lingual Transfer Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7674–7684. Association for Computational Linguistics.

Pinzhen Chen, Shaoxiong Ji, Nikolay Bogoychev, Andrey Kutuzov, Barry Haddow, and Kenneth Heafield. 2024a. Monolingual or Multilingual Instruction Tuning: Which Makes a Better Alpaca. In *Findings of the Association for Computational Linguistics: EACL 2024, St. Julian's, Malta, March 17-22, 2024*, volume EACL 2024 of *Findings of ACL*, pages 1347–1356. Association for Computational Linguistics.

Pinzhen Chen, Simon Yu, Zhicheng Guo, and Barry Haddow. 2024b. Is It Good Data for Multilingual Instruction Tuning or Just Bad Multilingual Evaluation for Large Language Models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 9706–9726. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Y. Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loïc Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, and 19 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *CoRR*, abs/2207.04672.

Antoine Dussolle, Andrea Cardeña Díaz, Shota Sato, and Peter Devine. 2025. M-IFEval: Multilingual Instruction-Following Evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, volume NAACL 2025 of *Findings of ACL*, pages 6161–6176. Association for Computational Linguistics.

FUEN. 2013. Map of Minorities & Regional and Minority Languages in Europe. Language Diversity project, funded with support from the European Commission.

Google DeepMind. 2026. Gemma 4 31B Instruct Model Card. Google AI for Developers and Hugging Face model documentation.

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.

Dieuwke Hupkes and Nikolay Bogoychev. 2025. MultiLoKo: a multilingual local knowledge benchmark for LLMs spanning 31 languages. *CoRR*, abs/2504.10356.

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. GlotLID: Language Identification for Low-Resource Languages. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 6155–6218. Association for Computational Linguistics.

Abdullatif Köksal, Marion Thaler, Ayyoob Imani, Ahmet Üstün, Anna Korhonen, and Hinrich Schütze. 2025. MURI: High-Quality Instruction Tuning Datasets for Low-Resource Languages via Reverse Instructions. *Trans. Assoc. Comput. Linguistics*, 13:1032–1055.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023 - System Demonstrations, Singapore, December 6-10, 2023*, pages 318–327. Association for Computational Linguistics.

Chong Li, Wen Yang, Jiajun Zhang, Jinliang Lu, Shaonan Wang, and Chengqing Zong. 2024. X-Instruction: Aligning Language Model in Low-resource Languages with Self-curated Cross-lingual Instructions. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 546–566. Association for Computational Linguistics.

Haonan Li, Fajri Koto, Minghao Wu, Alham Fikri Aji, and Timothy Baldwin. 2023. Bactrian-X : A Multilingual Replicable Instruction-Following Model with Low-Rank Adaptation. *CoRR*, abs/2305.15011.

Yafu Li, Ronghao Zhang, Zhilin Wang, Huajian Zhang, Leyang Cui, Yongjing Yin, Tong Xiao, and Yue Zhang. 2025a. Lost in Literalism: How Supervised Training Shapes Translationese in LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 12875–12894. Association for Computational Linguistics.

Zhenyu Li, Kehai Chen, Yunfei Long, Xuefeng Bai, Yaoyin Zhang, Xuchen Wei, Juntao Li, and Min Zhang. 2025b. XIFBench: Evaluating Large Language Models on Multilingual Instruction Following. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2511–2522. Association for Computational Linguistics.

Rishabh Maheshwary, Vikas Yadav, Hoang Nguyen, Khyati Mahajan, and Sathwik Tejaswi Madhusudhan. 2025. M2Lingual: Enhancing Multilingual, Multi-Turn Instruction Alignment in Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 9676–9713. Association for Computational Linguistics.

Meta AI. 2024. Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. Technical release and model documentation.

Mistral AI. 2025. Ministral 3 3B Instruct 2512 Model Card. Hugging Face model card.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Parker Riley, Isaac Caswell, Markus Freitag, and David Grangier. 2020. Translationese as a Language in "Multilingual" NMT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7737–7746. Association for Computational Linguistics.

Scott Roy, Cliff Brunk, Kyu-Young Kim, Justin Zhao, Markus Freitag, Mihir Kale, Gagan Bansal, Sidharth Mudgal, and Chris Varano. 2021. Using Machine Translation to Localize Task Oriented NLG Output. *CoRR*, abs/2107.04512.

Alejandro Salamanca, Diana Abagyan, Daniel D'souza, Ammar Khairi, David Mora, Saurabh Dash, Viraat Aryabumi, Sara Rajaee, Mehrnaz Mofakhami, Ananya Sahu, Thomas Euyang, Brittawnya Prince, Madeline Smith, Hangyu Lin, Acyr Locatelli, Sara Hooker, Tom Kocmi, Aidan N. Gomez, Ivan Zhang, and 7 others. 2026. Tiny Aya: Bridging Scale and Multilingual Depth. *CoRR*, abs/2603.11510.

Uri Shaham, Jonathan Herzig, Roee Aharoni, Idan Szpektor, Reut Tsarfaty, and Matan Eyal. 2024. Multilingual Instruction Tuning With Just a Pinch of Multilinguality. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 2304–2317. Association for Computational Linguistics.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Souza Moura, Dominik Krzeminski, Hakimeh Fadaei, Irem Ergün, Ifeoma Okoh, and 14 others. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 11521–11567. Association for Computational Linguistics.

Raivis Skadins, Marcis Pinnis, Andrejs Vasiljevs, Inguna Skadina, and Tomás Hudik. 2014. Application of machine translation in localization into low-resourced languages. In *Proceedings of the 17th Annual conference of the European Association for Machine Translation, EAMT 2014, Dubrovnik, Croatia, June 16-18, 2014*, pages 209–216. European Association for Machine Translation.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpaca: A Strong, Replicable Instruction-Following Model. Stanford Center for Research on Foundation Models blog. Accessed 2026-06-08.

Bibek Upadhayay and Vahid Behzadan. 2023. TaCo: Enhancing Cross-Lingual Transfer for Low-Resource Languages in LLMs through Translation-Assisted Chain-of-Thought Processes. *CoRR*, abs/2311.10797.

Justin Vasselli, Eunike Andriani Kardinata, Yusuke Sakai, and Taro Watanabe. 2025. Multilingual Dialogue Generation and Localization with Dialogue Act Scripting. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, EMNLP 2025, Suzhou, China, November 4-9, 2025*, pages 32896–32911. Association for Computational Linguistics.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13484–13508. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Linjuan Wu, Haoran Wei, Baosong Yang, and Weiming Lu. 2025. From English to Second Language Mastery: Enhancing LLMs with Cross-Lingual Continued Instruction Tuning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 23006–23023. Association for Computational Linguistics.

Bo Zeng, Chenyang Lyu, Sinuo Liu, Mingyan Zeng, Minghao Wu, Xuanfan Ni, Tianqi Shi, Yu Zhao, Yefeng Liu, Chenyu Zhu, Ruizhe Li, Jiahui Geng, Qing Li, Yu Tong, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. Marco-Bench-MIF: On Multilingual Instruction-Following Capability of Large Language. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*

(Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 24058–24072. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-Following Evaluation for Large Language Models. *CoRR*, abs/2311.07911.

A Dataset and Policy Statistics

A.1 Semantic Annotation of EUROALPACA

TASK TAXONOMY	
Creative writing	Question answering
Reasoning	Information extraction
Classification	Paraphrasing
Procedural instructions	Coding
Summarisation	Style transfer
Evaluation	Translation
Other	

Table 7: Task-label taxonomy used for semantic annotation of the source examples.

DOMAIN TAXONOMY	
General	Creative
Technology	Business
Science	Humanities
Mathematics	Education
Programming	Social science
Medicine	Law

Table 8: Domain-label taxonomy used for semantic annotation of the source examples.

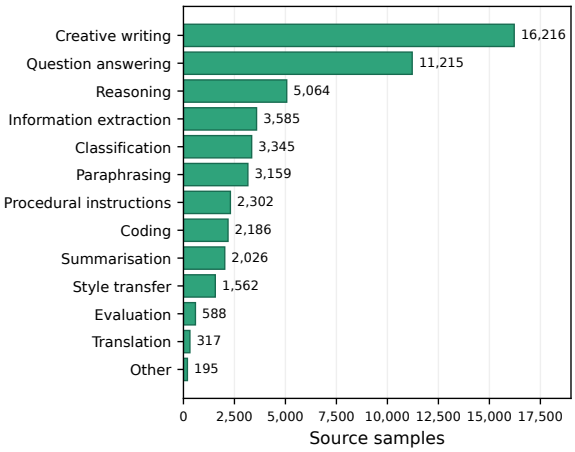


Figure 4: Primary task-label distribution across the 51,760 English source examples, as assigned by the LLM using the predefined task taxonomy.

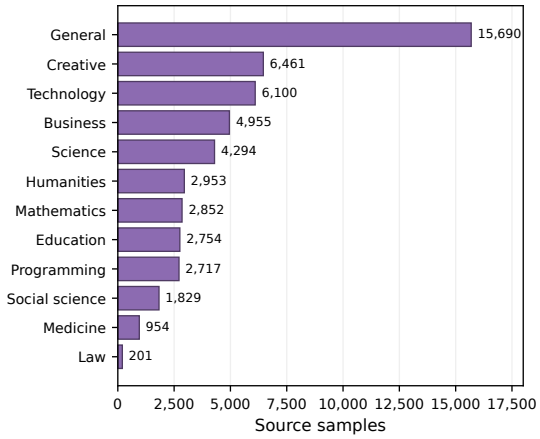


Figure 5: Primary domain-label distribution across the 51,760 English source examples, as assigned by the LLM using the predefined domain taxonomy.

Task-level localisation reason	Count	%
English grammar or spelling	4,694	42.3
Letter or character constraint	2,369	21.4
Language-specific task	1,165	10.5
Word/line-count constraint	1,053	9.5
Rhyme or sound constraint	644	5.8
Idiom or proverb	411	3.7
Wordplay/pun/anagram/palindrome	387	3.5
Other non-literal case	220	2.0
Mnemonic or initial letters	85	0.8
Source output violates instruction	60	0.5
Total	11,088	100.0

Table 9: Reasons for routing 11,088 examples to task-level localisation.

Field	Translate	Preserve
Instruction	40,174	498
Input	7,328	33,344
Output	30,179	10,493

Table 10: Field policies for the 40,672 examples that pass the whole-sample translation gate. Empty inputs are counted as preserved.

A.2 Examples Across Task-Preserving Handling Regimes

ID	Regime	Source example	Direct field-wise translation	Task-preserving target example
139	Task-equivalent rewriting in the target-language: Source-language form	**English** **Instruction:** Edit the following sentence to correct the grammar and syntax errors. **Input:** You was late to the party. **Output:** You were late to the party.	**French** **Instruction:** Modifie la phrase suivante pour corriger les erreurs de grammaire et de syntaxe. **Input:** Tu étais en retard à la fête. **Output:** Tu étais en retard à la fête.	**French** **Instruction:** Modifie la phrase suivante pour corriger les erreurs de grammaire et de syntaxe. **Input:** Tu était en retard à la fête. **Output:** Tu étais en retard à la fête.
	Failure and handling:	Direct translation can turn the erroneous source sentence into a grammatical target-language sentence, making the input and output identical. The localised sample must introduce a comparable target-language error and provide its correction.		
1480	Task-equivalent rewriting in the target-language: Rhyme / phonology	**English** **Instruction:** Generate a list of words that rhyme with "cat". **Output:** Bat, rat, hat, mat, pat, ...	**German** **Instruction:** Erstelle eine Liste von Wörtern, die sich auf "Katze" reimen. **Output:** Fledermaus, Ratte, Hut, Matte, Klaps, ...	**German** **Instruction:** Erstelle eine Liste von Wörtern, die sich auf "Haus" reimen. **Output:** Maus, Klaus, Applaus, hinaus.
	Failure and handling:	Rhyme is tied to target-language phonology. The cue word and answer set must be selected in the target language rather than translated word by word.		
218	Task-equivalent rewriting in the target-language: Structural constraint	**English** **Instruction:** Generate a haiku poem. **Output:** Fallen crimson leaves / Sighing softly in the breeze / Nature's tranquil peace.	**Catalan** **Instruction:** Genera un poema haiku. **Output:** Fulles carmesines caigudes / sospirant suaument a la brisa / pau tranquil.la de la natura.	**Catalan** **Instruction:** Genera un haiku. **Output:** Fulles al vent / la tarda calla lenta / pau de tardor.
	Failure and handling:	Line, syllable and poetic-form constraints are not preserved by literal translation. The poem must be regenerated so that the target text satisfies the intended form.		
115	Task-equivalent rewriting in the target-language: Idiom	**English** **Instruction:** What is the meaning of the following idiom? **Input:** It's raining cats and dogs. **Output:** It means that it is raining very heavily.	**Danish** **Instruction:** Hvad betyder følgende idiom? **Input:** Det regner katte og hunde. **Output:** Det betyder, at det regner meget kraftigt.	**Danish** **Instruction:** Hvad betyder følgende udtryk? **Input:** Det står ned i stænger. **Output:** Det betyder, at det regner meget kraftigt.
	Failure and handling:	The literal translation of the English idiom is not a natural Danish expression. The source idiom should be replaced by a target-language idiom with the same function.		
744	Task-equivalent rewriting in the target-language: Mnemonic / initials	**English** **Instruction:** Create a mnemonic to remember the eight planets. **Output:** "My Very Eager Mother Just Served Us Nachos"...	**Italian** **Instruction:** Crea una frase mnemonica per ricordare gli otto pianeti. **Output:** "Mia Madre Molto Entusiasta Ha Appena Servito Nachos"...	**Italian** **Instruction:** Crea una frase mnemonica per ricordare gli otto pianeti. **Output:** "Mia Vecchia Terra Mi Guarda Sempre Un Nuvolone", per Mercurio, Venere, Terra, Marte, Giove, Saturno, Urano e Nettuno.
	Failure and handling:	The English mnemonic encodes initial-letter constraints. The target sample must create a new mnemonic compatible with the target-language planet names.		
2575	Task-equivalent rewriting in the target-language: Letter-count constraint	**English** **Instruction:** Find a five-letter word that begins with 'd' and ends with 's'. **Output:** Dress	**Spanish** **Instruction:** Encuentra una palabra de cinco letras que empiece por 'd' y termine en 's'. **Output:** Vestido	**Spanish** **Instruction:** Encuentra una palabra de cinco letras que empiece por 'd' y termine en 's'. **Output:** Dosis
	Failure and handling:	The English answer satisfies the letter constraint, but its direct translation does not. The localised sample must choose a target-language answer that satisfies the same formal constraint.		

Table 11: Examples requiring task-level localisation. The fourth column illustrates the failure cases, while the final column shows a task-preserving target-language construction.

ID	Regime	Source example	Direct field-wise translation	Task-preserving target example
66	Direct translation	**English** **Instruction:** Find the main idea of the following passage. **Input:** Climate change is an increasingly contentious global issue... **Output:** The main idea is that climate change is severely impacting the environment...	**Finnish** **Instruction:** Etsi seuraavan tekstin pääajatus. **Input:** Ilmastonmuutos on yhä kiistanalaisempi maailmanlaajuinen kysymys... **Output:** Tekstin pääajatus on, että ilmastonmuutos vaikuttaa vakavasti ympäristöön...	**Finnish** **Instruction:** Etsi seuraavan tekstin pääajatus. **Input:** Ilmastonmuutos on yhä kiistanalaisempi maailmanlaajuinen kysymys... **Output:** Tekstin pääajatus on, että ilmastonmuutos vaikuttaa vakavasti ympäristöön...
	Failure and handling:	This is a direct-translation case. All non-empty fields can be translated without changing the task relation. The direct and task-preserving target examples are therefore identical.		
192	Direct translation	**English** **Instruction:** Compose an informative sentence explaining the given term. **Input:** Green bonds **Output:** Green bonds are debt instruments issued to raise funds for environmentally friendly projects.	**Romanian** **Instruction:** Scrie o propoziție informativă care explică termenul dat. **Input:** Obligațiuni verzi **Output:** Obligațiunile verzi sunt instrumente de datorie emise pentru a strânge fonduri pentru proiecte ecologice.	**Romanian** **Instruction:** Scrie o propoziție informativă care explică termenul dat. **Input:** Obligațiuni verzi **Output:** Obligațiunile verzi sunt instrumente de datorie emise pentru a strânge fonduri pentru proiecte ecologice.
	Failure and handling:	This is a direct-translation case. The instruction, input, and output express ordinary semantic content, so translating all non-empty fields preserves the task relation.		
98	Field-preserving	**English** **Instruction:** Classify the following statement as true or false. **Input:** The Supreme Court is the highest court in the US. **Output:** True	**Spanish** **Instruction:** Clasifica la siguiente afirmación como verdadera o falsa. **Input:** El Tribunal Supremo es el tribunal más alto de EE. UU. **Output:** Es cierto.	**Spanish** **Instruction:** Classify the following statement as true or false. **Input:** El Tribunal Supremo es el tribunal más alto de EE. UU. **Output:** True
	Failure and handling:	The task defines a fixed label set. Translating the labels into natural Spanish changes the required answer space, so the instruction label set and the output label must be preserved verbatim.		
2092	Field-preserving	**English** **Instruction:** Construct a valid SQL statement. **Input:** Retrieve the names and ages of all the students who live in Brisbane. **Output:** `SELECT name, age FROM students WHERE city = 'Brisbane';`	**Albanian** **Instruction:** Ndërto një deklaratë të vlefshme SQL. **Input:** Merr emrat dhe moshat e të gjithë studentëve që jetojnë në Brisbane. **Output:** `ZGJIDH emri, mosha NGA studentet KU qyteti = 'Brisbane';`	**Albanian** **Instruction:** Ndërto një deklaratë të vlefshme SQL. **Input:** Retrieve the names and ages of all the students who live in Brisbane. **Output:** `SELECT name, age FROM students WHERE city = 'Brisbane';`
	Failure and handling:	The output is executable SQL. Translating SQL keywords, column names, table names or string literals can make the query invalid or change its semantics, so the SQL statement is preserved.		
38024	Field-preserving	**English** **Instruction:** Translate English text to French. **Input:** The sun was shining brightly in the sky. **Output:** Le soleil brillait fort dans le ciel.	**Welsh** **Instruction:** Cyfieithwch destun Saesneg i'r Ffrangeg. **Input:** Roedd yr haul yn disgleirio'n llachar yn yr awyr. **Output:** Roedd yr haul yn disgleirio'n gryf yn yr awyr.	**Welsh** **Instruction:** Cyfieithwch destun Saesneg i'r Ffrangeg. **Input:** The sun was shining brightly in the sky. **Output:** Le soleil brillait fort dans le ciel.
	Failure and handling:	The source and target languages are part of the task definition. Translating every field changes the English–French translation pair into Welsh text, so the source sentence and fixed-language output must be preserved.		
3798	Field-preserving	**English** **Instruction:** Complete the following sentence by filling in the <mask>. **Input:** I wanted to throw a party but the <mask> became an obstacle. **Output:** I wanted to throw a party but the pandemic became an obstacle.	**Icelandic** **Instruction:** Ljúktu við eftirfarandi setningu með því að fylla inn í <gríma>. **Input:** Ég vildi halda veislu en <gríma> varð hindrun. **Output:** Ég vildi halda veislu en faraldurinn varð hindrun.	**Icelandic** **Instruction:** Ljúktu við eftirfarandi setningu með því að fylla inn í <mask>. **Input:** I wanted to throw a party but the <mask> became an obstacle. **Output:** I wanted to throw a party but the pandemic became an obstacle.
	Failure and handling:	The mask token is part of the task format. Translating it changes the literal token that must be filled, so the token and masked sentence are preserved.		
32293	Field-preserving	**English** **Instruction:** Translate the following sentence into Latin. **Input:** He is happy. **Output:** Is laetus est.	**Czech** **Instruction:** Přelož následující větu do latiny. **Input:** Je šťastný. **Output:** Je šťastný.	**Czech** **Instruction:** Přelož následující větu do latiny. **Input:** Je šťastný. **Output:** Is laetus est.
	Failure and handling:	The instruction and source sentence can be localised, but the output is the requested Latin translation. Translating the output into the target language removes the answer, so the output field is preserved.		

Table 12: Examples of direct and field-preserving translation. Direct field-wise translation preserves the task in the first two examples, but later examples require preserving task-critical fields such as labels, code, literal tokens and fixed-language outputs.

B Language Coverage and Analysis Tiers

EUROALPACA comprises an annotated English source collection and near-parallel versions in 50 European languages and regional varieties. Table 13 lists the target languages alphabetically, together with the ISO tags used in the dataset and experiments and their assigned resource tiers.

The high-, medium- and low-resource categories are internal analysis strata reflecting the relative availability of digital resources considered during dataset construction. They should not be interpreted as measures of language vitality, speaker population or sociolinguistic status.

ISO tag	Language	Resource tier
sq	Albanian	Medium
ast	Asturian	Low
eu	Basque	Medium
bs	Bosnian	Medium
bg	Bulgarian	High
ca	Catalan	High
hr	Croatian	Medium
cs	Czech	High
da	Danish	High
nl	Dutch	High
et	Estonian	Medium
fo	Faroese	Low
fi	Finnish	High
fr	French	High
fur	Friulian	Low
gl	Galician	Medium
ka	Georgian	Medium
de	German	High
el	Greek	High
hu	Hungarian	High
is	Icelandic	Medium
ga	Irish	Medium
it	Italian	High
ltg	Latgalian	Low
lv	Latvian	Medium
lij	Ligurian	Low
lt	Lithuanian	High
lmo	Lombard	Low
lb	Luxembourgish	Low
mk	Macedonian	Medium
mt	Maltese	Medium
nob	Norwegian Bokmål	High
nno	Norwegian Nynorsk	Low
oc	Occitan	Low
pl	Polish	High
pt	Portuguese	High
ro	Romanian	High
ru	Russian	High
sc	Sardinian	Low
gd	Scottish Gaelic	Low
sr	Serbian	Medium
scn	Sicilian	Low
szl	Silesian	Low
sk	Slovak	High
sl	Slovenian	Medium
es	Spanish	High
sv	Swedish	High
uk	Ukrainian	High
vec	Venetian	Low
cy	Welsh	Medium

Table 13: Language coverage and MADLAD-400-based resource tiers for EUROALPACA. Tiers are defined using clean monolingual token counts: high ($\geq$ 5 billion), medium (100 million to $<$ 5 billion) and low ($<$ 100 million or no separate clean partition). The classification contains 21 high-, 15 medium- and 14 low-resource languages and is used only for aggregate analysis.

B.1 EUROPEAN-IFEVAL Provenance

Tier	Languages	Hugging Face Dataset ID
Native-speaker or professional translator (human-curated)	ca	projecte-aina/IFEval_ca
	da	danish-foundation-models/ifeval-da
	de	jzhang86/de_ifeval
	el	ilsp/ifeval_greek
	es	BSC-LT/IFEval_es
	et	tartuNLP/ifeval_et
	eu	HiTZ/ifeval_eu
	fi	LumiOpen/ifeval_mt (Finnish)
	gl	HiTZ/ifeval_gl
	sv	LumiOpen/ifeval_mt (Swedish)
Localised and verified	cs, fr, hu, it, nl, pl, pt, ro, ru, uk	AIDC-AI/Marco-Bench-MIF
Machine-translated with localisation	ast, bg, bs, cy, fo, fur, ga, gd, hr, is, ka, lb, lij, lmo, lt, ltg, lv, mk, mt, nno, nob, oc, sc, scn, sk, sl, sq, sr, szl, vec	danish-foundation-models/multi-ifeval

Table 14: Language-level provenance summary for EUROPEAN-IFEVAL data. The released manifest stores the exact URL, source tier, missing-sample count and source-language answer-constraint count for each language.

C Complete Language-Level Results

C.1 EUROPEAN-IFEVAL: Accuracy after Language-Specific Fine-Tuning

Model	Condition	ast	bg	bs	ca	cs	cy	da	de	el	es	et	eu	fi	fo	fr	fur	ga	gd	gl	hr
Tiny-Aya-Global	Off-the-shelf	.392	.531	.442	.506	.460	.454	.549	.468	.497	.508	.468	.410	.475	.314	.506	.292	.416	.263	.495	.473
	Direct MT	.184	.359	.346	.362	.316	.338	.377	.322	.298	.370	.299	.255	.359	.220	.322	.267	.311	.216	.394	.370
	Task-Preserved	.405	.548	.518	.514	.494	.485	.508	.482	.494	.527	.436	.388	.540	.281	.512	.372	.433	.290	.543	.540
	EUROALPACA	**.512**	**.601**	**.554**	**.560**	**.506**	**.563**	**.575**	**.525**	**.571**	**.599**	**.508**	**.416**	**.562**	**.346**	**.543**	**.412**	**.525**	**.382**	**.584**	**.592**
Phi-4-Mini-Instruct	Off-the-shelf	.312	.282	.294	.316	.277	.202	.277	.333	**.290**	.427	.153	**.231**	.325	.164	.357	.216	.135	.122	.329	.252
	Direct MT	.115	.181	.224	.229	.227	.158	.220	.255	.185	.266	.153	.118	.255	.126	.216	.174	.153	.115	.227	.214
	Task-Preserved	.252	.279	.285	.320	.261	.200	.316	.375	.238	.409	.163	.189	.323	.182	.388	**.237**	**.179**	.153	.322	.269
	EUROALPACA	**.354**	**.307**	**.323**	**.357**	**.322**	**.225**	**.346**	**.433**	.266	**.508**	**.181**	.196	**.338**	**.197**	**.431**	.231	.174	**.177**	**.335**	**.326**
Ministral-3-3B-Instruct	Off-the-shelf	.258	.309	**.317**	.275	**.305**	.214	.336	.298	.248	.333	.251	.205	.316	.164	.274	.193	.176	.139	.264	.300
	Direct MT	.117	.191	.195	.231	.181	.170	.203	.214	.211	.253	.165	.168	.209	.172	.166	.197	.172	.158	.255	.198
	Task-Preserved	.270	.284	.283	**.344**	.292	.208	.329	.299	.270	.401	.222	**.227**	.311	.266	.331	.218	.219	.177	**.342**	.282
	EUROALPACA	**.348**	**.324**	.279	.338	.283	**.225**	**.348**	**.322**	**.320**	**.427**	**.255**	.216	**.329**	**.293**	**.348**	**.281**	**.221**	**.189**	.336	**.324**
Llama-3.2-3B-Instruct	Off-the-shelf	.386	.317	**.382**	.299	**.355**	.168	**.392**	**.427**	.257	**.523**	.200	**.222**	**.318**	.222	**.445**	.218	.183	.160	**.409**	**.359**
	Direct MT	.103	.189	.207	.238	.203	.218	.227	.240	.185	.255	.174	.150	.233	.155	.213	.202	.174	.128	.288	.198
	Task-Preserved	.254	.294	.273	.285	.294	.221	.298	.338	.214	.362	.172	.200	.279	.203	.364	**.256**	.189	.174	.338	.277
	EUROALPACA	**.421**	**.324**	.314	**.373**	.349	**.258**	.362	**.427**	**.272**	.445	**.226**	.220	.311	**.226**	.440	.244	**.233**	**.204**	.399	.345

Table 15: EUROPEAN-IFEVAL accuracy after language-specific fine-tuning (languages 1–20 of 50).

Model	Condition	hu	is	it	ka	lb	lij	lmo	lt	ltg	lv	mk	mt	nl	nno	nob	oc	pl	pt	ro	ru
Tiny-Aya-Global	Off-the-shelf	.442	.288	.457	.189	.235	.365	.332	.485	.265	.479	.377	.498	.530	.559	.576	.368	.418	.436	.433	.409
	Direct MT	.262	.227	.355	.202	.233	.179	.246	.294	.183	.284	.278	.359	.327	.319	.389	.284	.255	.338	.290	.272
	Task-Preserved	.427	.271	.567	.233	.332	.338	.384	.523	.256	.416	.429	.494	.549	.475	.521	.443	**.429**	.540	.460	**.510**
	EUROALPACA	**.484**	**.328**	**.584**	**.254**	**.382**	**.405**	**.435**	**.567**	**.357**	**.487**	**.525**	**.574**	**.553**	**.567**	**.605**	**.513**	.407	**.571**	**.499**	.505
Phi-4-Mini-Instruct	Off-the-shelf	.238	.193	.336	.206	.191	.202	.216	**.239**	.155	.197	**.314**	.200	.298	.296	.338	.275	.251	.355	.268	**.329**
	Direct MT	.177	.164	.268	.177	.174	.145	.176	.176	.143	.166	.203	.155	.227	.216	.286	.183	.231	.244	.185	.185
	Task-Preserved	.211	.195	.375	**.231**	.187	.181	**.246**	.231	.172	.200	.266	**.229**	.335	.296	.376	.254	.281	.372	.242	.292
	EUROALPACA	**.253**	**.225**	**.401**	.221	**.235**	**.239**	.235	.227	**.197**	**.244**	.291	.223	**.368**	**.344**	**.422**	**.309**	**.292**	**.429**	**.272**	**.329**
Ministral-3-3B-Instruct	Off-the-shelf	.262	.218	.292	.231	.185	.198	.197	.277	.155	**.252**	.284	.172	.261	.303	**.323**	.235	.248	.281	.246	.307
	Direct MT	.159	.179	.213	.198	.141	.145	.189	.181	.168	.149	.216	.174	.203	.198	.218	.206	.185	.209	.192	.179
	Task-Preserved	.261	.250	.366	.284	.189	.254	.258	.279	.191	.223	.318	.252	**.331**	.313	.321	.282	**.292**	**.364**	.290	.336
	EUROALPACA	**.277**	**.265**	**.386**	**.286**	**.223**	**.269**	**.311**	**.284**	**.208**	.239	**.328**	**.256**	.320	**.361**	**.323**	**.338**	.275	**.364**	**.301**	**.362**
Llama-3.2-3B-Instruct	Off-the-shelf	.290	.221	.449	.177	.229	**.269**	.250	**.219**	.158	.195	.316	**.221**	.394	.342	**.397**	**.357**	.301	.442	**.364**	**.384**
	Direct MT	.179	.172	.262	.151	.212	.174	.191	.162	.134	.162	.190	.172	.244	.212	.284	.210	.187	.251	.203	.205
	Task-Preserved	.250	.202	.390	.191	.258	.248	.231	.202	.172	.216	.282	.216	.360	.288	.317	.281	.261	.372	.290	.303
	EUROALPACA	**.307**	**.256**	**.455**	**.225**	**.263**	.263	**.298**	.210	**.197**	**.218**	**.330**	**.221**	**.434**	**.366**	.389	.342	**.316**	**.451**	.323	.351

Table 16: EUROPEAN-IFEVAL accuracy after language-specific fine-tuning (languages 21–40 of 50).

Model	Condition	sc	scn	sk	sl	sq	sr	sv	szl	uk	vec
Tiny-Aya-Global	Off-the-shelf	.366	.326	.487	.447	.321	.433	.468	.365	.440	.344
	Direct MT	.244	.275	.399	.313	.240	.280	.383	.210	.272	.261
	Task-Preserved	.355	.368	.561	.523	.330	.471	.567	.338	.470	.420
	EUROALPACA	**.412**	**.441**	**.569**	**.552**	**.372**	**.594**	**.625**	**.393**	**.484**	**.466**
Phi-4-Mini-Instruct	Off-the-shelf	.231	.208	.258	.246	.219	**.295**	.338	**.242**	.272	.218
	Direct MT	.143	.210	.195	.168	.191	.222	.268	.162	.181	.168
	Task-Preserved	.248	.246	.279	.237	**.246**	.289	.349	.214	.264	**.240**
	EUROALPACA	**.250**	**.252**	**.344**	**.254**	.237	.291	**.394**	.233	**.279**	.237
Ministral-3-3B-Instruct	Off-the-shelf	.187	.233	**.303**	**.275**	.204	.284	**.342**	.202	.299	.202
	Direct MT	.160	.166	.197	.164	.187	.215	.209	.166	.155	.183
	Task-Preserved	.227	**.279**	.261	.252	.242	.293	.307	.227	.272	.281
	EUROALPACA	**.254**	.267	.296	**.275**	**.265**	**.307**	.314	**.239**	**.305**	**.305**
Llama-3.2-3B-Instruct	Off-the-shelf	.275	.286	**.380**	**.332**	**.254**	**.297**	**.434**	.227	**.298**	.244
	Direct MT	.195	.172	.198	.160	.172	.193	.250	.141	.161	.200
	Task-Preserved	.275	.254	.242	.227	.206	.253	.344	.212	.237	.269
	EUROALPACA	**.303**	**.303**	.309	.269	.242	.295	.375	**.250**	.283	**.284**

Table 17: EUROPEAN-IFEVAL accuracy after language-specific fine-tuning (languages 41–50 of 50).

C.2 Aya Evaluation Suite: ROUGE-L after Language-Specific Fine-Tuning

Model	Condition	bg	ca	cs	cy	da	de	el	es	et	eu	fi	fr	ga	gd	gl	hu	is	it
Tiny-Aya-Global	Off-the-shelf	.041	.099	.134	.116	.134	.106	.053	.095	.116	.129	.118	.116	.119	.083	.113	.131	.087	.109
	Direct MT	**.123**	.271	.248	.235	.272	**.258**	.158	.294	.222	.176	.208	.279	.236	.198	.289	.269	.203	.275
	Task-Preserved	.114	.283	**.265**	.254	**.284**	**.258**	.157	.299	**.232**	.184	**.219**	**.290**	.248	.188	**.299**	.278	.214	.283
	EUROALPACA	.122	**.285**	.255	**.259**	.280	**.258**	**.160**	**.312**	**.232**	**.188**	.212	.285	**.251**	**.216**	.294	**.287**	**.220**	**.288**
Phi-4-Mini-Instruct	Off-the-shelf	.073	.123	.106	.072	.103	.078	.087	.158	.045	.052	.078	.129	.073	.066	.091	.106	.057	.130
	Direct MT	**.126**	.259	.217	.160	.261	.258	.130	.294	.147	.127	.184	.267	.167	.155	.276	.235	.173	.245
	Task-Preserved	.109	.251	.218	.166	.259	.255	.138	.287	.144	.130	.178	.273	.170	.165	.276	.234	.176	.255
	EUROALPACA	.117	**.276**	**.241**	**.200**	**.267**	**.267**	**.157**	**.300**	**.172**	**.148**	**.200**	**.282**	**.195**	**.203**	**.293**	**.261**	**.200**	**.267**
Ministral-3-3B-Instruct	Off-the-shelf	.056	.163	.138	.105	.156	.117	.066	.158	.100	.063	.115	.161	.098	.092	.166	.135	.129	.140
	Direct MT	.095	.255	.228	.166	.245	.218	.140	.278	.172	.127	.183	.252	.148	.149	.261	.210	.181	.252
	Task-Preserved	.099	.256	.221	.165	.241	.213	.132	.279	.174	.136	.181	.256	.163	.149	.265	.199	.192	.258
	EUROALPACA	**.100**	**.272**	**.254**	**.214**	**.272**	**.246**	**.153**	**.287**	**.199**	**.154**	**.203**	**.286**	**.197**	**.186**	**.281**	**.239**	**.221**	**.273**
Llama-3.2-3B-Instruct	Off-the-shelf	.090	.222	.220	.138	.234	.205	.133	.254	.163	.125	.163	.247	.132	.159	.229	.225	.194	.243
	Direct MT	.078	.254	.219	.172	.243	.230	.121	.285	.166	.141	.170	.266	.166	.170	.271	.234	.185	.261
	Task-Preserved	.085	.267	.217	.175	.236	.250	**.149**	.296	.175	.131	.178	**.273**	.166	.167	.274	.238	.191	.287
	EUROALPACA	**.112**	**.283**	**.241**	**.203**	**.271**	**.254**	.133	**.302**	**.193**	**.164**	**.183**	.267	**.188**	**.215**	**.291**	**.264**	**.202**	**.288**

Table 18: ROUGE-L on Aya Evaluation Suite after language-specific fine-tuning (languages 1–18 of 36).

Model	Condition	lb	lt	lv	mk	mt	nl	nno	nob	pl	pt	ro	ru	sk	sl	sq	sr	sv	uk
Tiny-Aya-Global	Off-the-shelf	.047	.089	.083	.047	.120	.134	.162	.171	.106	.105	.123	.047	.119	.118	.128	.033	.099	.085
	Direct MT	.205	.206	.215	.125	.231	.279	.238	.284	.228	.305	.263	.131	.247	.243	.220	.130	**.284**	**.157**
	Task-Preserved	.184	.212	.221	.120	.231	.278	.240	.282	**.242**	.309	**.282**	.127	.250	.249	.231	.134	.274	.140
	EUROALPACA	**.221**	**.213**	**.223**	**.134**	**.260**	**.286**	**.245**	**.291**	.238	**.311**	.273	**.147**	**.264**	**.255**	**.251**	**.135**	.281	.148
Phi-4-Mini-Instruct	Off-the-shelf	.059	.088	.089	.083	.047	.089	.067	.102	.114	.143	.121	.091	.079	.063	.116	.072	.126	.100
	Direct MT	.190	.163	.166	.124	.140	.262	.204	.259	.216	.286	.227	.117	.203	.189	.204	.122	.251	.139
	Task-Preserved	.192	.170	.176	.129	.151	.253	.206	.259	.222	.292	.235	.119	.193	.195	.219	.120	.253	.136
	EUROALPACA	**.227**	**.195**	**.184**	**.137**	**.169**	**.276**	**.236**	**.283**	**.227**	**.310**	**.250**	**.128**	**.216**	**.211**	**.233**	**.144**	**.271**	**.141**
Ministral-3-3B-Instruct	Off-the-shelf	.078	.090	.103	.071	.066	.156	.147	.158	.122	.172	.149	.071	.130	.116	.132	.091	.171	.052
	Direct MT	.164	.168	.165	.102	.157	.236	.216	.253	.203	.273	.230	.095	.214	.175	.211	.100	.244	.113
	Task-Preserved	.155	.174	.164	.115	.162	.243	.213	.244	.208	.279	.222	.093	.207	.195	.208	.103	.259	**.133**
	EUROALPACA	**.209**	**.191**	**.195**	**.126**	**.202**	**.263**	**.239**	**.263**	**.224**	**.298**	**.257**	**.113**	**.227**	**.212**	**.239**	**.117**	**.276**	.132
Llama-3.2-3B-Instruct	Off-the-shelf	.165	.176	**.184**	.098	.141	.257	.203	.253	.211	.274	.250	.093	.179	.188	.178	.079	**.275**	.110
	Direct MT	.196	.157	.165	.111	.163	.270	.220	.256	.204	.286	.226	.083	.202	.188	.215	.098	.256	.124
	Task-Preserved	.221	.181	.174	.090	.174	.262	.221	.251	.206	.285	.241	**.107**	.190	.185	.225	.090	.258	.123
	EUROALPACA	**.227**	**.202**	.182	**.120**	**.199**	**.281**	**.233**	**.267**	**.222**	**.303**	**.264**	.086	**.210**	**.210**	**.257**	**.121**	**.275**	**.127**

Table 19: ROUGE-L on Aya Evaluation Suite after language-specific fine-tuning (languages 19–36 of 36).

C.3 Aya Evaluation Suite: FBERT after Language-Specific Fine-Tuning

Model	Condition	bg	ca	cs	cy	da	de	el	es	et	eu	fi	fr	ga	gd	gl	hu	is	it
Tiny-Aya-Global	Off-the-shelf	.822	.822	.840	.825	.834	.826	.826	.819	.832	.821	.827	.829	.821	.791	.829	.831	.798	.828
	Direct MT	.881	.876	.878	.860	.883	.884	.881	.885	.873	.862	.873	.884	.854	.828	**.885**	.877	.845	.883
	Task-Preserved	.884	.881	**.882**	.865	.886	.884	.883	.887	.874	.866	.877	**.887**	.856	.824	**.885**	.880	.848	.886
	EUROALPACA	**.886**	**.883**	**.882**	**.867**	**.887**	**.886**	**.887**	**.890**	**.877**	**.869**	**.878**	**.887**	**.857**	**.835**	**.885**	**.885**	**.855**	**.887**
Phi-4-Mini-Instruct	Off-the-shelf	.839	.842	.838	.797	.842	.841	.835	.858	.796	.800	.830	.853	.789	.772	.833	.831	.794	.848
	Direct MT	.870	.870	.864	.819	.876	.881	.853	.884	.840	.828	.859	.880	.820	.805	.876	.863	.833	.875
	Task-Preserved	.872	.866	.868	.824	.878	.882	.862	.884	.835	.829	.862	.882	.820	.814	.874	.862	.840	.877
	EUROALPACA	**.876**	**.878**	**.874**	**.840**	**.884**	**.885**	**.867**	**.888**	**.850**	**.839**	**.872**	**.885**	**.832**	**.827**	**.879**	**.872**	**.845**	**.884**
Ministral-3-3B-Instruct	Off-the-shelf	.841	.843	.839	.803	.843	.837	.829	.846	.821	.806	.837	.847	.790	.774	.845	.824	.815	.842
	Direct MT	.863	.862	.863	.816	.869	.867	.857	.873	.840	.829	.849	.873	.799	.797	.870	.850	.828	.872
	Task-Preserved	.865	.868	.864	.815	.869	.870	.864	.874	.845	.838	.856	.880	.813	.800	.875	.845	.839	.877
	EUROALPACA	**.878**	**.876**	**.876**	**.842**	**.882**	**.882**	**.879**	**.885**	**.859**	**.849**	**.872**	**.886**	**.831**	**.822**	**.878**	**.861**	**.852**	**.883**
Llama-3.2-3B-Instruct	Off-the-shelf	.864	.868	.870	.826	.875	.871	.864	.878	.856	.850	.861	.878	.827	.823	.866	.867	.846	.879
	Direct MT	.855	.869	.861	.829	.876	.874	.861	.883	.845	.837	.854	.879	.820	.820	.876	.858	.836	.878
	Task-Preserved	.859	.874	.863	.830	.873	.881	.864	.885	.850	.838	.861	**.883**	.822	.816	.878	.863	.839	.885
	EUROALPACA	**.875**	**.881**	**.873**	**.848**	**.880**	**.884**	**.877**	**.890**	**.861**	**.851**	**.869**	.882	**.833**	**.839**	**.882**	**.876**	**.849**	**.888**

Table 20: FBERT on Aya Evaluation Suite after language-specific fine-tuning (languages 1–18 of 36).

Model	Condition	lb	lt	lv	mk	mt	nl	nno	nob	pl	pt	ro	ru	sk	sl	sq	sr	sv	uk
Tiny-Aya-Global	Off-the-shelf	.771	.814	.810	.817	.818	.835	.849	.856	.826	.824	.832	.828	.835	.824	.818	.815	.818	.841
	Direct MT	.826	.873	.871	.875	.852	.883	.870	.888	.876	.887	.879	.883	.877	.879	.855	.876	.884	.877
	Task-Preserved	.820	.873	.873	.874	.852	.885	.872	.889	.881	.888	**.883**	.885	.878	.881	.857	.876	.885	.882
	EUROALPACA	**.835**	**.875**	**.874**	**.878**	**.858**	**.887**	**.874**	**.890**	**.882**	**.889**	.882	**.889**	**.885**	**.884**	**.868**	**.881**	**.887**	**.884**
Phi-4-Mini-Instruct	Off-the-shelf	.783	.824	.817	.837	.781	.838	.818	.841	.842	.855	.836	.857	.828	.818	.825	.830	.843	.847
	Direct MT	.819	.846	.843	.867	.819	.878	.858	.877	.866	.883	.861	.883	.859	.857	.846	.864	.872	.870
	Task-Preserved	.821	.850	.848	.865	.819	.877	.859	.876	.870	.883	.866	.885	.860	.855	.853	.864	.875	.873
	EUROALPACA	**.832**	**.857**	**.853**	**.872**	**.830**	**.883**	**.867**	**.886**	**.876**	**.888**	**.872**	**.887**	**.867**	**.865**	**.858**	**.873**	**.882**	**.878**
Ministral-3-3B-Instruct	Off-the-shelf	.790	.815	.816	.837	.784	.845	.837	.846	.839	.848	.839	.846	.832	.829	.822	.838	.847	.834
	Direct MT	.809	.841	.839	.854	.819	.864	.855	.872	.865	.876	.859	.869	.857	.847	.844	.862	.867	.857
	Task-Preserved	.808	.847	.839	.857	.820	.868	.860	.873	.872	.879	.860	.877	.860	.854	.844	.858	.869	.864
	EUROALPACA	**.826**	**.862**	**.856**	**.872**	**.839**	**.879**	**.868**	**.880**	**.875**	**.884**	**.873**	**.884**	**.868**	**.870**	**.860**	**.872**	**.882**	**.875**
Llama-3.2-3B-Instruct	Off-the-shelf	.815	.859	**.859**	.857	.823	.879	.865	.878	.872	.879	.874	.878	.860	.860	.845	.852	.878	.870
	Direct MT	.823	.849	.849	.858	.826	.878	.864	.875	.866	.883	.864	.870	.857	.857	.847	.861	.871	.864
	Task-Preserved	.830	.856	.852	.865	.830	.879	.864	.874	.868	.884	.870	.875	.855	.856	.854	.862	.869	.868
	EUROALPACA	**.837**	**.867**	.858	**.872**	**.840**	**.884**	**.869**	**.880**	**.877**	**.889**	**.881**	**.881**	**.866**	**.869**	**.867**	**.869**	**.879**	**.878**

Table 21: FBERT on Aya Evaluation Suite after language-specific fine-tuning (languages 19–36 of 36).

C.4 Per-Language Primary Comparison

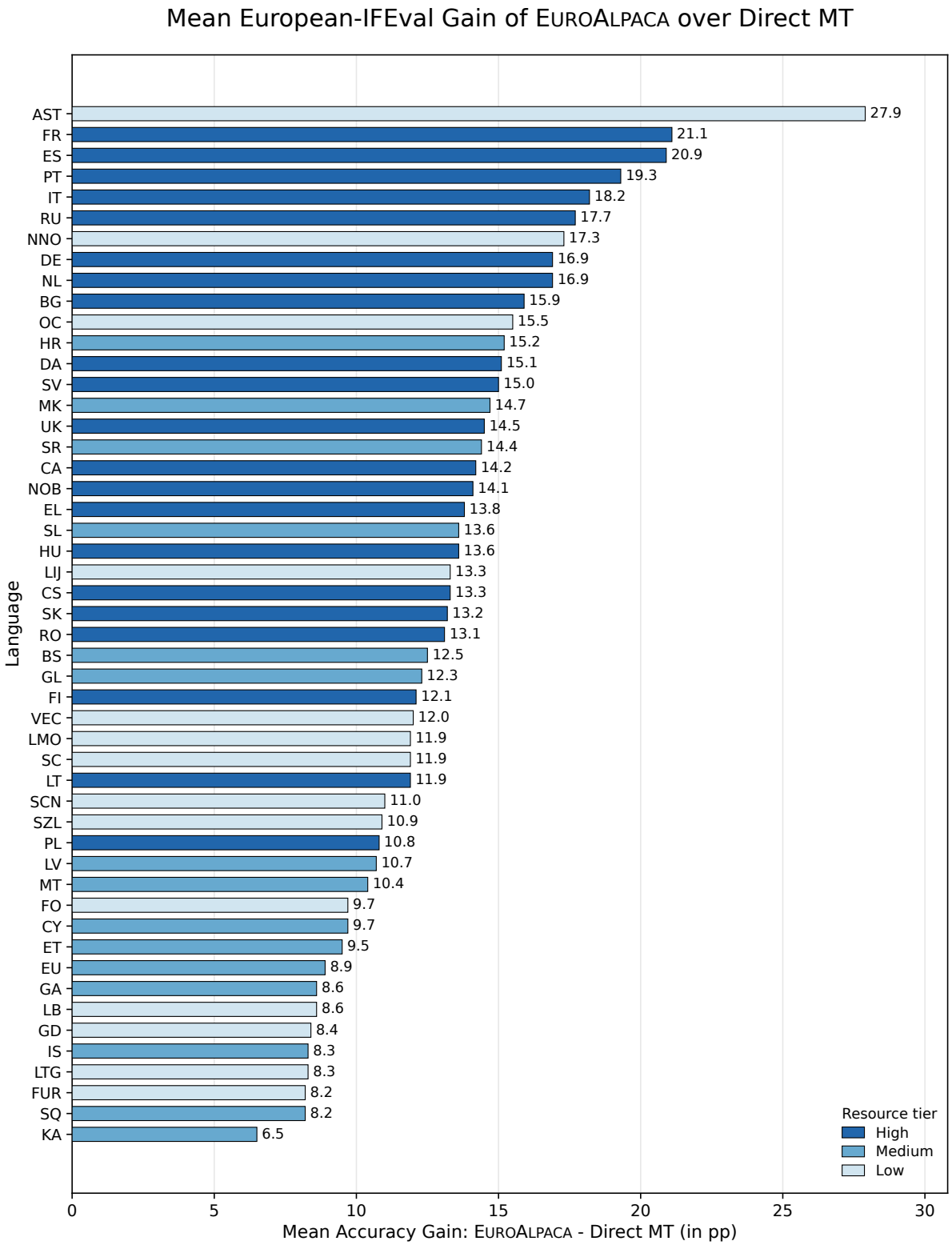


Figure 6: Mean EUROPEAN-IFEVAL accuracy gain of EUROALPACA over direct MT for each language, averaged across the four evaluated models. Languages are ordered by the magnitude of the gain and coloured by their internal resource tier. All 50 language-level gains are positive.

C.5 Pairwise Comparisons with Existing Alpaca-Derived Resources on EUROPEAN-IFEVAL Accuracy

C.5.1 Okapi

Model	Training resource	ca	da	de	es	eu	fr	hr	hu	it	nl	pt	ro	ru	sk	sr	sv	uk
Tiny-Aya-Global	Okapi	**.573**	.566	**.529**	.573	.410	.481	.582	.468	.492	.508	.490	.449	.471	**.605**	.579	.555	.460
	EUROALPACA	.560	**.575**	.525	**.599**	**.416**	**.543**	**.592**	**.484**	**.584**	**.553**	**.571**	**.499**	**.505**	.569	**.594**	**.625**	**.484**
Phi-4-Mini-Instruct	Okapi	.348	**.357**	**.442**	.449	**.207**	.423	.305	**.266**	.399	**.384**	.360	**.275**	**.342**	.323	**.318**	**.399**	**.336**
	EUROALPACA	**.357**	.346	.433	**.508**	.196	**.431**	**.326**	.253	**.401**	.368	**.429**	.272	.329	**.344**	.291	.394	.279
Ministral-3-3B-Instruct	Okapi	**.351**	**.372**	**.372**	.392	.213	.338	**.330**	.257	.368	.312	.331	.294	.322	.288	**.330**	**.314**	.301
	EUROALPACA	.338	.348	.322	**.427**	**.216**	**.348**	.324	**.277**	**.386**	**.320**	**.364**	**.301**	**.362**	**.296**	.307	**.314**	**.305**
Llama-3.2-3B-Instruct	Okapi	.320	.311	.403	.375	.216	.364	.315	.246	.390	.368	.377	.294	.296	.298	.276	.360	.218
	EUROALPACA	**.373**	**.362**	**.427**	**.445**	**.220**	**.440**	**.345**	**.307**	**.455**	**.434**	**.451**	**.323**	**.351**	**.309**	**.295**	**.375**	**.283**

Table 22: Pairwise comparison of Okapi and EUROALPACA after language-specific fine-tuning on EUROPEAN-IFEVAL over their shared-language subset.

C.5.2 MITS

Model	Training resource	bg	bs	ca	cs	cy	da	de	el	es	et	eu	fi	fr	ga	gd	gl	hr	hu	is
Tiny-Aya-Global	MITS	.489	.468	.453	.384	.473	.477	.429	.455	.458	.433	.360	.512	.442	.435	.267	.479	.517	.364	.269
	EUROALPACA	**.601**	**.554**	**.560**	**.506**	**.563**	**.575**	**.525**	**.571**	**.599**	**.508**	**.416**	**.562**	**.543**	**.525**	**.382**	**.584**	**.592**	**.484**	**.328**
Phi-4-Mini-Instruct	MITS	.296	.312	.296	.294	.162	.314	.316	.237	.373	.159	.187	.305	.359	.166	.135	.307	.281	.227	.189
	EUROALPACA	**.307**	**.323**	**.357**	**.322**	**.225**	**.346**	**.433**	**.266**	**.508**	**.181**	**.196**	**.338**	**.431**	**.174**	**.177**	**.335**	**.326**	**.253**	**.225**
Ministral-3-3B-Instruct	MITS	.244	.228	.264	.196	.158	.264	.244	.244	.335	.207	.172	.272	.253	.174	**.189**	.274	.235	.200	.223
	EUROALPACA	**.324**	**.279**	**.338**	**.283**	**.225**	**.348**	**.322**	**.320**	**.427**	**.255**	**.216**	**.329**	**.348**	**.221**	**.189**	**.336**	**.324**	**.277**	**.265**
Llama-3.2-3B-Instruct	MITS	.231	.289	.274	.251	.185	.303	.281	.235	.346	.190	.198	.285	.364	.160	.170	.351	.254	.226	.202
	EUROALPACA	**.324**	**.314**	**.373**	**.349**	**.258**	**.362**	**.427**	**.272**	**.445**	**.226**	**.220**	**.311**	**.440**	**.233**	**.204**	**.399**	**.345**	**.307**	**.256**

Table 23: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on EUROPEAN-IFEVAL accuracy over their shared-language subset (languages 1–19 of 38).

Model	Training resource	it	ka	lb	lt	lv	mk	mt	nl	nob	pl	pt	ro	ru	sk	sl	sq	sr	sv	uk
Tiny-Aya-Global	MITS	.482	.235	.273	.429	.408	.375	.460	.444	.552	.298	.449	.373	.342	.506	.473	.275	.381	.545	.355
	EUROALPACA	**.584**	**.254**	**.382**	**.567**	**.487**	**.525**	**.574**	**.553**	**.605**	**.407**	**.571**	**.499**	**.505**	**.569**	**.552**	**.372**	**.594**	**.625**	**.484**
Phi-4-Mini-Instruct	MITS	.370	.193	.191	.216	.176	.289	.191	.314	.366	.264	.333	.211	.266	.273	**.260**	.208	.249	.364	.261
	EUROALPACA	**.401**	**.221**	**.235**	**.227**	**.244**	**.291**	**.223**	**.368**	**.422**	**.292**	**.429**	**.272**	**.329**	**.344**	.254	**.237**	**.291**	**.394**	**.279**
Ministral-3-3B-Instruct	MITS	.279	.218	.219	.206	.221	.234	.233	.268	.261	.183	.296	.200	.213	.233	.216	.195	.211	.292	.203
	EUROALPACA	**.386**	**.286**	**.223**	**.284**	**.239**	**.328**	**.256**	**.320**	**.323**	**.275**	**.364**	**.301**	**.362**	**.296**	**.275**	**.265**	**.307**	**.314**	**.305**
Llama-3.2-3B-Instruct	MITS	.383	.189	.191	.185	.181	.238	.197	.336	.340	.226	.344	.266	.231	.239	.239	.179	.218	.320	.207
	EUROALPACA	**.455**	**.225**	**.263**	**.210**	**.218**	**.330**	**.221**	**.434**	**.389**	**.316**	**.451**	**.323**	**.351**	**.309**	**.269**	**.242**	**.295**	**.375**	**.283**

Table 24: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on EUROPEAN-IFEVAL accuracy over their shared-language subset (languages 20–38 of 38).

C.5.3 Bactrian-X

Model	Training resource	cs	de	es	et	fi	fr	gl	hr	it	lt	lv	mk	nl	pl	pt	ro	ru	sl	sv	uk
Tiny-Aya-Global	Bactrian-X	.466	.460	**.612**	.490	**.577**	**.586**	**.608**	.576	**.614**	.557	**.510**	.506	**.577**	**.481**	**.608**	.429	**.571**	**.584**	**.640**	**.545**
	EUROALPACA	**.506**	**.525**	.599	**.508**	.562	.543	.584	**.592**	.584	**.567**	.487	**.525**	.553	.407	.571	**.499**	.505	.552	.625	.484
Phi-4-Mini-Instruct	Bactrian-X	**.329**	**.447**	.468	.170	**.366**	**.473**	**.410**	**.328**	**.458**	**.242**	.225	**.364**	**.423**	**.312**	**.471**	**.290**	**.377**	**.254**	**.416**	**.336**
	EUROALPACA	.322	.433	**.508**	**.181**	.338	.431	.335	.326	.401	.227	**.244**	.291	.368	.292	.429	.272	.329	**.254**	.394	.279
Ministral-3-3B-Instruct	Bactrian-X	**.336**	**.364**	.392	.253	**.335**	**.386**	**.377**	**.336**	**.396**	.271	**.254**	**.354**	**.392**	**.301**	**.410**	**.311**	**.409**	**.303**	**.370**	**.351**
	EUROALPACA	.283	.322	**.427**	**.255**	.329	.348	.336	.324	.386	**.284**	.239	.328	.320	.275	.364	.301	.362	.275	.314	.305
Llama-3.2-3B-Instruct	Bactrian-X	.340	**.451**	.423	.200	**.322**	.433	**.438**	.340	**.457**	**.223**	.198	**.343**	.425	**.320**	**.460**	**.331**	**.366**	**.303**	**.405**	.248
	EUROALPACA	**.349**	.427	**.445**	**.226**	.311	**.440**	.399	**.345**	.455	.210	**.218**	.330	**.434**	.316	.451	.323	.351	.269	.375	**.283**

Table 25: Pairwise comparison of Bactrian-X and EUROALPACA after language-specific fine-tuning on EUROPEAN-IFEVAL accuracy over their shared-language subset.

C.6 Pairwise Comparisons with Existing Alpaca-Derived Resources on Aya Evaluation Suite with ROUGE-L

C.6.1 Okapi

Model	Training resource	ca	da	de	es	eu	fr	hu	it	nl	pt	ro	ru	sk	sr	sv	uk
Tiny-Aya-Global	Okapi	.276	**.288**	**.268**	.311	.179	**.289**	.282	**.305**	.280	**.319**	**.279**	.110	.252	.102	**.284**	.137
	EUROALPACA	**.285**	.280	.258	**.312**	**.188**	.285	**.287**	.288	**.286**	.311	.273	**.147**	**.264**	**.135**	.281	**.148**
Phi-4-Mini-Instruct	Okapi	.266	**.275**	.259	**.313**	**.158**	**.290**	**.262**	**.283**	**.287**	**.312**	.238	.107	.211	.116	**.281**	.119
	EUROALPACA	**.276**	.267	**.267**	.300	.148	.282	.261	.267	.276	.310	**.250**	**.128**	**.216**	**.144**	.271	**.141**
Ministral-3-3B-Instruct	Okapi	.266	.249	.231	.283	.142	.272	**.242**	.271	.256	.286	.230	.093	.212	.087	.253	.106
	EUROALPACA	**.272**	**.272**	**.246**	**.287**	**.154**	**.286**	.239	**.273**	**.263**	**.298**	**.257**	**.113**	**.227**	**.117**	**.276**	**.132**
Llama-3.2-3B-Instruct	Okapi	.273	.262	.251	.285	**.165**	**.280**	.252	.285	.275	.281	.256	**.099**	.201	.074	.263	.121
	EUROALPACA	**.283**	**.271**	**.254**	**.302**	.164	.267	**.264**	**.288**	**.281**	**.303**	**.264**	.086	**.210**	**.121**	**.275**	**.127**

Table 26: Pairwise comparison of Okapi and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with ROUGE-L over their shared-language subset.

C.6.2 MITS

Model	Training resource	bg	ca	cs	cy	da	de	el	es	et	eu	fi	fr	ga	gd	gl	hu	is	it
Tiny-Aya-Global	MITS	**.123**	.267	.246	.240	.274	.255	**.160**	.295	.220	.174	.207	.274	.237	.192	.287	.273	.207	.280
	EUROALPACA	.122	**.285**	**.255**	**.259**	**.280**	**.258**	**.160**	**.312**	**.232**	**.188**	**.212**	**.285**	**.251**	**.216**	**.294**	**.287**	**.220**	**.288**
Phi-4-Mini-Instruct	MITS	**.125**	.259	.225	.182	.264	.249	.145	**.304**	.153	.138	.187	.274	.194	.178	.284	.236	.181	.252
	EUROALPACA	.117	**.276**	**.241**	**.200**	**.267**	**.267**	**.157**	.300	**.172**	**.148**	**.200**	**.282**	**.195**	**.203**	**.293**	**.261**	**.200**	**.267**
Ministral-3-3B-Instruct	MITS	**.115**	.261	.232	.196	.243	.233	.138	.285	.191	.147	.187	.266	.192	.185	.265	.223	.214	.253
	EUROALPACA	.100	**.272**	**.254**	**.214**	**.272**	**.246**	**.153**	**.287**	**.199**	**.154**	**.203**	**.286**	**.197**	**.186**	**.281**	**.239**	**.221**	**.273**
Llama-3.2-3B-Instruct	MITS	**.115**	.270	**.245**	.196	.267	.234	**.139**	.291	.183	.145	**.184**	**.282**	**.204**	.198	.276	.251	.196	.282
	EUROALPACA	.112	**.283**	.241	**.203**	**.271**	**.254**	.133	**.302**	**.193**	**.164**	.183	.267	.188	**.215**	**.291**	**.264**	**.202**	**.288**

Table 27: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with ROUGE-L over their shared-language subset (languages 1–18 of 35).

Model	Training resource	lb	lt	lv	mk	mt	nl	nob	pl	pt	ro	ru	sk	sl	sq	sr	sv	uk
Tiny-Aya-Global	MITS	.204	.203	.211	**.142**	.142	.268	.268	.229	.289	.256	.141	.248	.239	.224	.120	.270	**.157**
	EUROALPACA	**.221**	**.213**	**.223**	.134	**.260**	**.286**	**.291**	**.238**	**.311**	**.273**	**.147**	**.264**	**.255**	**.251**	**.135**	**.281**	.148
Phi-4-Mini-Instruct	MITS	.214	.172	.183	.132	**.176**	.259	.271	.217	.298	.233	.127	.208	.210	.223	.122	.261	**.152**
	EUROALPACA	**.227**	**.195**	**.184**	**.137**	.169	**.276**	**.283**	**.227**	**.310**	**.250**	**.128**	**.216**	**.211**	**.233**	**.144**	**.271**	.141
Ministral-3-3B-Instruct	MITS	.207	.183	.183	**.128**	.183	.245	.251	.209	.278	.242	**.121**	.203	.194	.231	.113	.259	**.142**
	EUROALPACA	**.209**	**.191**	**.195**	.126	**.202**	**.263**	**.263**	**.224**	**.298**	**.257**	.113	**.227**	**.212**	**.239**	**.117**	**.276**	.132
Llama-3.2-3B-Instruct	MITS	.222	**.203**	.181	**.129**	.187	.263	.259	.214	.292	.240	**.106**	.205	.209	.217	.114	.265	**.131**
	EUROALPACA	**.227**	.202	**.182**	.120	**.199**	**.281**	**.267**	**.222**	**.303**	**.264**	.086	**.210**	**.210**	**.257**	**.121**	**.275**	.127

Table 28: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with ROUGE-L over their shared-language subset (languages 19–35 of 35).

C.6.3 Bactrian-X

Model	Training resource	cs	de	es	et	fi	fr	gl	it	lt	lv	mk	nl	pl	pt	ro	ru	sl	sv	uk
Tiny-Aya-Global	Bactrian-X	.127	.113	.304	.218	**.213**	.283	.273	**.292**	**.222**	.218	.116	.275	.228	.299	.122	.117	.242	**.290**	**.158**
	EUROALPACA	**.255**	**.258**	**.312**	**.232**	.212	**.285**	**.294**	.288	.213	**.223**	**.134**	**.286**	**.238**	**.311**	**.273**	**.147**	**.255**	.281	.148
Phi-4-Mini-Instruct	Bactrian-X	.236	.250	**.303**	**.176**	.198	.281	.248	.263	.184	.180	.116	.271	.217	.306	.243	.118	.196	**.276**	**.147**
	EUROALPACA	**.241**	**.267**	.300	.172	**.200**	**.282**	**.293**	**.267**	**.195**	**.184**	**.137**	**.276**	**.227**	**.310**	**.250**	**.128**	**.211**	.271	.141
Ministral-3-3B-Instruct	Bactrian-X	.239	.236	**.289**	.179	.190	.262	.238	.265	.178	.177	.124	.254	.218	.276	.243	.106	.196	.258	.130
	EUROALPACA	**.254**	**.246**	.287	**.199**	**.203**	**.286**	**.281**	**.273**	**.191**	**.195**	**.126**	**.263**	**.224**	**.298**	**.257**	**.113**	**.212**	**.276**	**.132**
Llama-3.2-3B-Instruct	Bactrian-X	.224	.241	.295	.191	.173	.257	.255	.273	.180	.180	.109	.276	.211	.289	.255	**.111**	.200	.258	**.127**
	EUROALPACA	**.241**	**.254**	**.302**	**.193**	**.183**	**.267**	**.291**	**.288**	**.202**	**.182**	**.120**	**.281**	**.222**	**.303**	**.264**	.086	**.210**	**.275**	**.127**

Table 29: Pairwise comparison of Bactrian-X and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with ROUGE-L over their shared-language subset.

C.7 Pairwise Comparisons with Existing Alpaca-Derived Resources on Aya Evaluation Suite with FBERT

C.7.1 Okapi

Model	Training resource	ca	da	de	es	eu	fr	hu	it	nl	pt	ro	ru	sk	sr	sv	uk
Tiny-Aya-Global	Okapi	.881	**.892**	**.888**	**.891**	.866	**.888**	**.885**	**.893**	**.887**	**.892**	**.886**	.888	**.885**	.875	**.888**	**.884**
	EUROALPACA	**.883**	.887	.886	.890	**.869**	.887	**.885**	.887	**.887**	.889	.882	**.889**	**.885**	**.881**	.887	**.884**
Phi-4-Mini-Instruct	Okapi	.877	**.886**	**.887**	**.892**	**.844**	**.888**	**.877**	**.886**	**.887**	**.891**	**.874**	**.888**	**.870**	.865	**.885**	**.878**
	EUROALPACA	**.878**	.884	.885	.888	.839	.885	.872	.884	.883	.888	.872	.887	.867	**.873**	.882	**.878**
Ministral-3-3B-Instruct	Okapi	**.877**	.877	.877	**.885**	.843	.882	**.864**	**.884**	.878	.883	.872	.881	.866	.866	.877	.871
	EUROALPACA	.876	**.882**	**.882**	**.885**	**.849**	**.886**	.861	.883	**.879**	**.884**	**.873**	**.884**	**.868**	**.872**	**.882**	**.875**
Llama-3.2-3B-Instruct	Okapi	.878	**.881**	**.884**	.888	.847	**.885**	.871	.887	**.885**	.886	.875	**.883**	.862	.856	**.883**	.875
	EUROALPACA	**.881**	.880	**.884**	**.890**	**.851**	.882	**.876**	**.888**	.884	**.889**	**.881**	.881	**.866**	**.869**	.879	**.878**

Table 30: Pairwise comparison of Okapi and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with FBERT over their shared-language subset.

C.7.2 MITS

Model	Training resource	bg	ca	cs	cy	da	de	el	es	et	eu	fi	fr	ga	gd	gl	hu	is	it
Tiny-Aya-Global	MITS	.879	.878	.877	.862	.883	.883	.879	.886	.869	.862	.873	.883	.853	.824	.880	.879	.847	.885
	EUROALPACA	**.886**	**.883**	**.882**	**.867**	**.887**	**.886**	**.887**	**.890**	**.877**	**.869**	**.878**	**.887**	**.857**	**.835**	**.885**	**.885**	**.855**	**.887**
Phi-4-Mini-Instruct	MITS	.869	.870	.867	.836	.880	.881	.863	.887	.836	.832	.863	.882	.830	.820	.877	.864	.836	.878
	EUROALPACA	**.876**	**.878**	**.874**	**.840**	**.884**	**.885**	**.867**	**.888**	**.850**	**.839**	**.872**	**.885**	**.832**	**.827**	**.879**	**.872**	**.845**	**.884**
Ministral-3-3B-Instruct	MITS	.865	.871	.869	.836	.874	.874	.871	.882	.852	.838	.862	.880	.822	.815	.872	.855	.839	.876
	EUROALPACA	**.878**	**.876**	**.876**	**.842**	**.882**	**.882**	**.879**	**.885**	**.859**	**.849**	**.872**	**.886**	**.831**	**.822**	**.878**	**.861**	**.852**	**.883**
Llama-3.2-3B-Instruct	MITS	.870	.875	.869	.843	.878	.877	.867	.886	.855	.837	.858	**.883**	.832	.831	.876	.868	.840	.882
	EUROALPACA	**.875**	**.881**	**.873**	**.848**	**.880**	**.884**	**.877**	**.890**	**.861**	**.851**	**.869**	.882	**.833**	**.839**	**.882**	**.876**	**.849**	**.888**

Table 31: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with FBERT over their shared-language subset (languages 1–18 of 35).

Model	Training resource	lb	lt	lv	mk	mt	nl	nob	pl	pt	ro	ru	sk	sl	sq	sr	sv	uk
Tiny-Aya-Global	MITS	.825	.869	.870	.873	.854	.882	.885	.876	.886	.877	.883	.878	.877	.858	.874	.882	.878
	EUROALPACA	**.835**	**.875**	**.874**	**.878**	**.858**	**.887**	**.890**	**.882**	**.889**	**.882**	**.889**	**.885**	**.884**	**.868**	**.881**	**.887**	**.884**
Phi-4-Mini-Instruct	MITS	.825	.848	.849	.864	.828	.878	.879	.869	.884	.864	.882	.861	.861	.853	.866	.876	.873
	EUROALPACA	**.832**	**.857**	**.853**	**.872**	**.830**	**.883**	**.886**	**.876**	**.888**	**.872**	**.887**	**.867**	**.865**	**.858**	**.873**	**.882**	**.878**
Ministral-3-3B-Instruct	MITS	.825	.857	.849	.862	.830	.872	.874	.869	.878	.868	.876	.861	.859	.854	.868	.873	.868
	EUROALPACA	**.826**	**.862**	**.856**	**.872**	**.839**	**.879**	**.880**	**.875**	**.884**	**.873**	**.884**	**.868**	**.870**	**.860**	**.872**	**.882**	**.875**
Llama-3.2-3B-Instruct	MITS	.831	.857	.853	.861	.834	.878	.877	.865	.887	.869	.878	.860	.863	.851	.864	.877	.872
	EUROALPACA	**.837**	**.867**	**.858**	**.872**	**.840**	**.884**	**.880**	**.877**	**.889**	**.881**	**.881**	**.866**	**.869**	**.867**	**.869**	**.879**	**.878**

Table 32: Pairwise comparison of MITS and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with FBERT over their shared-language subset (languages 19–35 of 35).

C.7.3 Bactrian-X

Model	Training resource	cs	de	es	et	fi	fr	gl	it	lt	lv	mk	nl	pl	pt	ro	ru	sl	sv	uk
Tiny-Aya-Global	Bactrian-X	.838	.831	.888	.871	**.878**	.885	.880	.886	.871	.871	.870	.883	.876	.887	.832	.886	.877	.885	.880
	EUROALPACA	**.882**	**.886**	**.890**	**.877**	**.878**	**.887**	**.885**	**.887**	**.875**	**.874**	**.878**	**.887**	**.882**	**.889**	**.882**	**.889**	**.884**	**.887**	**.884**
Phi-4-Mini-Instruct	Bactrian-X	**.875**	.883	**.888**	.849	.870	.884	.872	.878	.854	.850	.867	.880	.872	.887	.869	.886	.855	.879	.876
	EUROALPACA	.874	**.885**	**.888**	**.850**	**.872**	**.885**	**.879**	**.884**	**.857**	**.853**	**.872**	**.883**	**.876**	**.888**	**.872**	**.887**	**.865**	**.882**	**.878**
Ministral-3-3B-Instruct	Bactrian-X	.872	.878	**.885**	.849	.866	.879	.869	.880	.855	.851	.865	.876	.872	.881	.870	.876	.856	.870	.869
	EUROALPACA	**.876**	**.882**	.885	**.859**	**.872**	**.886**	**.878**	**.883**	**.862**	**.856**	**.872**	**.879**	**.875**	**.884**	**.873**	**.884**	**.870**	**.882**	**.875**
Llama-3.2-3B-Instruct	Bactrian-X	.867	.881	.887	.854	.861	.880	.876	.881	.852	.851	.866	.880	.869	.887	.874	.877	.862	**.880**	.868
	EUROALPACA	**.873**	**.884**	**.890**	**.861**	**.869**	**.882**	**.882**	**.888**	**.867**	**.858**	**.872**	**.884**	**.877**	**.889**	**.881**	**.881**	**.869**	.879	**.878**

Table 33: Pairwise comparison of Bactrian-X and EUROALPACA after language-specific fine-tuning on Aya Evaluation Suite with FBERT over their shared-language subset.

D Reproducibility Details

D.1 Checkpoint Identifiers

The four evaluated checkpoints are:

- `CohereLabs/tiny-aya-global`
- `microsoft/Phi-4-mini-instruct`
- `mistralai/Ministral-3-3B-Instruct-2512`
- `meta-llama/Llama-3.2-3B-Instruct`

D.2 Fine-Tuning Hyperparameters

All language-specific runs use the same optimisation and adapter configuration, reported in Table 34.

Parameter	Value
Method	LoRA
Rank / alpha / dropout	16 / 32 / 0.05
Target modules	q, k, v, and o projections
Maximum sequence length	1,300 tokens by default
Optimiser	AdamW
Learning rate	1×10^{-5}
Micro-batch size	2
Gradient-accumulation steps	32
Effective batch size	64
Gradient clipping	1.0
Precision	bf16
Random seed	42
Checkpoint selection	Validation loss with early stopping

Table 34: Fine-tuning configuration shared by all language-specific runs. Based on token-length analysis, we use a 1,600-token limit for `ga`, `gd`, `is`, `mk`, `mt` and `sq` to reduce truncation in the upper tail of the sequence-length distribution.

D.2.1 Prompt Formatting and Language-Specific Alpaca Templates

For fine-tuning, each example was serialised using the model-specific chat template provided by each model's tokeniser, ensuring compatibility with the format expected by each model. The user turn was constructed with an Alpaca-style prompt in the same language as the instruction.

We translated the original English Alpaca template into each target language using ChatGPT. In every version, we modified the original instruction, *Write a response that appropriately completes the request*, by adding an explicit response-language requirement. For English, this yielded *Write a response in English that appropriately completes the request*. Equivalent specifications, such as *auf Deutsch*, *en català* and *in italiano*, were added for the other languages. This language requirement is not present in the original Alpaca template. Each example was therefore paired with a template written in the same language as the instruction and explicitly requesting a response in that language.

Following the original format, we used two template variants. Records with a non-empty input included separate `Instruction`, `Input` and `Response` blocks, whereas records without an input omitted the `Input` block.

English template

```
Below is an instruction that describes a task, paired with an
input that provides further context. Write a response in
English that appropriately completes the request.

### Instruction:
{instruction text}

### Input:
{optional input text}

### Response:
{expected output}
```

Examples in other languages:

German template

```
Nachfolgend finden Sie eine Anweisung, die eine Aufgabe
beschreibt. Schreiben Sie eine Antwort auf Deutsch, die die
Anfrage angemessen erfüllt.

### Anweisung:
{instruction text}

### Eingabe:
{optional input text}

### Antwort:
{expected output}
```

Catalan template

```
A continuació es mostra una instrucció que descriu una tasca,
juntament amb una entrada que proporciona més context.
Escriviu una resposta en català que completi adequadament la
sol·licitud.

### Instrucció:
{instruction text}

### Entrada:
{optional input text}

### Resposta:
{expected output}
```

D.3 Gemma-4 Decision-Module Configuration

Component	Configuration
Decision model	Gemma-4-31B-IT, `Q8_K_XL` GGUF
Draft model	Gemma-4-E2B-IT, `Q4_K_XL` GGUF
Decoding	Temperature $= 0$, top-$p = 0.95$, min-$p = 0.01$
Penalties	None explicitly configured
Reasoning	Disabled
Speculative decoding	5–16 draft tokens
Prompt formatting	Gemma-4 Jinja template
Context / parallelism	8,912 tokens / 2 server slots
KV cache / attention	`q8_0`; unified KV cache; Flash Attention
Batch / physical batch	1,024 / 512
GPU offloading	Up to 99 main- and draft-model layers

Table 35: Inference and decoding configuration of the Gemma-4 decision module.

D.4 Gemma-4 Prompt Templates

This section reports the system and user prompt templates used by the Gemma-4 decision module for semantic annotation, sample-level routing, field-level policy prediction, task-level localisation and consistency validation. Placeholders enclosed in braces are populated at runtime.

D.4.1 Semantic Annotation

System prompt

```
You are a careful dataset annotator. Output JSON only.
```

User prompt template

```
You are labelling BOTH the TASK TYPE(S) and the DOMAIN(S) of
the request.

You must pick:
- one PRIMARY task type (exactly one)
- zero or more SECONDARY task types (may be empty)
- one PRIMARY domain (exactly one)
- zero or more SECONDARY domains (may be empty)

Allowed task types (use EXACT strings):
question_answering, reasoning, summarisation, translation,
paraphrasing, classification, information_extraction,
creative_writing, procedural_instructions, coding,
style_transfer, evaluation, other

Allowed domain types (use EXACT strings):
general, programming, technology, science, mathematics,
medicine, law, business, education, humanities,
social_science, creative, other

Rules (tasks):
- primary_task must be exactly one of the allowed task types.
- secondary_tasks must be a JSON array of allowed task types.
- secondary_tasks must NOT repeat primary_task.
- If uncertain, use "other".

Rules (domains):
- primary_domain must be exactly one of the allowed domain
types.
- secondary_domains must be a JSON array of allowed domain
types.
- secondary_domains must NOT repeat primary_domain.
- Use "general" when the topic is everyday or non-specialised.
- Use "other" only if none fit.

Return ONLY valid JSON with exactly these keys:
{
  "primary_task": "<one_task_type>",
  "secondary_tasks": ["<task_type>", "..."],
  "primary_domain": "<one_domain_type>",
  "secondary_domains": ["<domain_type>", "..."]
}

No extra text. No markdown. No code fences.
---
Below is an instruction that describes a task, paired with an
input that provides further context. Write a response that
appropriately completes the request.

### Instruction:
{instruction}

### Input:
{input}

### Response:
```

D.4.2 Sample-Level MT Suitability Gate

System prompt

```
You are a strict data curation assistant. Return JSON only.
```

User prompt template

```
You are deciding the SAMPLE-LEVEL translation policy for a
dataset row.
This first pass is a global gate: decide whether the task can
be preserved through field-wise machine translation.
A later field-level pass will decide which individual fields
to translate or preserve.
Source language: {source_lang_name}
Target languages: {target_desc_text}
Primary task: {primary_task}
Primary domain: {primary_domain}
Secondary tasks: {secondary_tasks}
Secondary domains: {secondary_domains}

Field values:
instruction: {instruction}
input: {input}
output: {output}

Rules:
- Keep task semantics intact.
- Decide if this sample should be translated to other
languages at all.
- Set include_sample=false when the task depends on
source-language form rather than meaning.
- Set include_sample=false for language-form tasks (spelling,
grammar, tense/voice edits, typo correction).
- Set include_sample=false for token-level/wordplay tasks
(letter counts, rhymes/rhyming/rhymes with, homophones,
anagrams).
- Set include_sample=false for idioms, proverbs, puns,
mnemonics, first-letter tasks, and exact letter/character
constraints.
- Set include_sample=false when the output already violates a
strict word/line/count constraint.
- For classification, discard figurative-language detection
tasks if MT may alter label logic.
- For coding, discard mislabelled
language-conjugation/grammar-only tasks.
- For creative writing, discard strict form/sound constraints
(haiku/limerick/acrostic/rhyme/wordplay).
- For evaluation, discard English-form correction/judgement
tasks (grammar/spelling/punctuation/usage).
- For information extraction, discard
grammar/syntax/token-form extraction tasks.
- For other, discard language-form
editing/spelling/punctuation/phonetic/accent tasks.
- For paraphrasing, discard
grammar/tense/voice/person/word-level constrained rewrites.
- Set include_sample=false for English-constrained
generation/rewrite tasks (e.g., 'in formal/simple/modern
English', English-sentence-only constraints).
- Set include_sample=false for English alphabet/pangram
constraints.
- Set include_sample=false for language-specific tasks
(English-only constraints).
- Keep include_sample=true for normal QA, reasoning, coding,
summarisation, and creative generation tasks.
- Keep include_sample=true for paraphrasing focused on
clarity/conciseness/flow (not language-form mechanics).
- Translation tasks can still be include_sample=true; field
policy will protect fixed-language fields.
- If uncertain whether the sample involves a language-form
constraint or ordinary meaning-preserving translation, choose
include_sample=false.
- If include_sample=true, reason_code must be
directly_translatable and evidence must be empty.
- If include_sample=false, choose exactly one reason_code
from the allowed list and provide short evidence from the
sample.

Allowed reason_code values:
- directly_translatable
- english_grammar_or_spelling
- rhyme_or_sound_constraint
- letter_or_character_constraint
- word_count_or_line_count_constraint
- idiom_or_proverb
```

```
- wordplay_pun_anagram_palindrome
- mnemonic_or_initial_letters
- language_specific_task
- output_already_violates_instruction
- other_not_directly_translatable

Return ONLY valid JSON with exactly this schema:
{
  "sample": {
    "include_sample": true,
    "reason_code": "directly_translatable",
    "evidence": ""
  }
}
No markdown. No extra keys.
```

D.4.3 Field-Level Translation Policy

System prompt

```
You are a strict data curation assistant. Return JSON only.
```

User prompt template

```
You are deciding the FIELD-LEVEL translation policy for a
dataset row.
This sample has passed the global gate and can be preserved
through field-wise machine translation.
This second pass decides which individual fields should be
translated or preserved.
Make decisions in one pass for ALL provided target languages,
using a default plus language-specific exceptions.
Source language: {source_lang_name}
Target languages: {target_desc_text}
Primary task: {primary_task}
Primary domain: {primary_domain}
Secondary tasks: {secondary_tasks}
Secondary domains: {secondary_domains}

Field values:
instruction: {instruction}
input: {input}
output: {output}

Rules:
- Keep task semantics intact.
- Set should_translate=false only when translation would break
task intent.
- If a field is empty or missing, should_translate=false.
- Do not set false just because task is
rewriting/paraphrasing.
- For language-specific tasks, sample-level policy already
filtered unsuitable samples.
- For classification: preserve input for language-ID/code/
formula detection tasks.
- For classification with constrained labels, preserve output
labels.
- For coding: keep code and exact literal-string
transformations unchanged.
- If output is already in a fixed language required by the
task, usually do not translate it.
- Proper names/titles may need to remain unchanged depending
on task context.
- Example: if task is translate English into French and the
output is
French text, do not translate that output.
- Example: if input contains a film title or named entity
that should remain original, avoid translating it.
- IMPORTANT: reason is only needed when decision is false.
- If decision is true, omit reason or leave it empty.
- Reasons must be very brief (max 8 words).
- If unsure, set should_translate=true.

Return ONLY valid JSON with exactly this schema:
{
  "instruction": {
    "all_languages": {"should_translate": true},
    "per_language": {}
  },
  "input": {
    "all_languages": {"should_translate": true},
    "per_language": {}
  },
  "output": {
    "all_languages": {"should_translate": true},
    "per_language": {}
  }
}
- Use language tags exactly from target list.
- all_languages is the default for every target language.
- per_language contains only exceptions.
No markdown. No extra keys.
```

D.4.4 Task-Level Localisation

System prompt

```
You localise instruction-following dataset samples.
Return one valid JSON object only.
Do not use markdown code fences.
```

User prompt template

```
You are creating a localised counterpart of one English
Alpaca-style sample that cannot be directly translated.

Target language: {target_lang_name} ({target_lang})

Original sample metadata:
primary_task: {primary_task}
primary_domain: {primary_domain}
secondary_tasks: {secondary_tasks}
secondary_domains: {secondary_domains}

Original English sample:
instruction: {instruction}
input: {input}
output: {output}

Localisation rules:
- Write the final instruction, input, and output in
{target_lang_name}.
- Make the localised sample as semantically close to the
original as the target language allows.
- Treat the result as a near-parallel sample: preserve the
same meaning, intent, task type, difficulty, constraints, and
expected answer style whenever possible.
- The first priority is that the sample makes sense in the
target language and that the task logic is correct.
- The second priority is semantic closeness to the original
so the pair can reasonably be treated as parallel data.
- Do not do literal machine translation when it would break
the task logic or sound unnatural in the target language.
- Adapt grammar, orthography, syntax, vocabulary, morphology,
punctuation, and idioms to the target language.
- For spelling, grammar, tense, voice, punctuation, typos,
word-count, rhyme, or wordplay tasks, create a
target-language equivalent with target-language examples and
a correct target-language answer.
- If the original sample uses an English-specific sentence,
word, alphabet, proverb, idiom, or error, replace it with a
natural equivalent in the target language.
- When replacing English-specific material, keep the
replacement as close as possible in topic, tone, length,
complexity, and answer structure.
- Preserve all functional information from the
instruction/input fields, including separate lines, criteria,
labels, options, constraints, and examples.
- Keep code, URLs, numbers, formal symbols, and named
entities unchanged unless they are part of the
language-specific exercise.
- Keep empty input empty when the localised task does not
need an input field.
- The localised output must correctly answer the localised
instruction/input, not the English source.
- Do not mention that the sample was translated or localised.
- Do not add explanations outside the requested output field.

Return ONLY valid JSON with exactly this schema:
{
  "instruction": "...",
  "input": "...",
  "output": "..."
}

No markdown. No extra keys. Do not wrap JSON in code fences.
Escape internal double quotes and newline characters
correctly.
```

D.4.5 Consistency Validation

System prompt

```
You validate multilingual instruction-following samples with
strict language-specific coherence checks. Return JSON only.
```

User prompt template

```
Task:
Validate one sample in language: {language_name}.

Input sample:
primary_task: {primary_task}
instruction: {instruction}
input: {input}
output: {output}

Coherence criteria (language-specific):
- Judge coherence from the actual text in {language_name},
not from the abstract task idea.
- Set is_coherent=false if wording in {language_name} makes
the instruction/input/output semantically wrong,
contradictory, or unclear.
- Set is_coherent=false if there is severe
grammar/orthography/syntax degradation that changes or blocks
meaning in {language_name}.
- Set is_coherent=false if mistranslation/literal calque
causes wrong meaning in {language_name}.
- Set is_coherent=false if language mixing or wrong-language
text breaks task understanding in {language_name}.
- If errors are minor and meaning remains clear and logically
aligned, is_coherent can be true.

Language lock (mandatory):
- Final corrected instruction/input/output must be in
{language_name}.
- Exception: keep spans in another language only when the
task explicitly requires them (for example, translation tasks
or quoted source text that must remain in the source
language).
- Keep named entities, code, URLs, numbers, and formal
symbols unchanged unless clearly malformed.
- Outside those required spans, avoid language mixing and
normalise wording to {language_name}.

Task and constraint preservation:
- Treat explicit instruction constraints as part of the task
meaning, including format, number of items, casing,
punctuation restrictions, required/forbidden words, exact
phrases, and requested response language.
- Set is_coherent=false if the instruction/input/output is
internally inconsistent with an explicit instruction
constraint, or if the output clearly violates one.

Return ONLY JSON with this schema:
{
  "is_coherent": true,
  "issues": "",
  "changes": {}
}

Rules:
- The issues field must contain a very short reason only when
not coherent.
- If coherent, issues must be an empty string.
- Always correct grammar, orthography, punctuation, and
syntax when needed.
- Put edits only in changes. The changes object must include
ONLY modified fields among instruction/input/output as
optional keys.
- If nothing needs editing, return changes as {}.
- If is_coherent=false, changes must repair the sample so it
becomes coherent.
- If is_coherent=true, changes can still contain language
fixes.
- Keep text in the dataset language, except required
cross-language spans.
- Preserve original intent, constraints, factual content, and
task type whenever possible.
- Do not add new requirements not implied by the sample.
- Keep URLs, code, numbers, and named entities unless clearly
malformed.
- Keep empty input empty when appropriate.
- Do not use markdown fences. Do not return multiple JSON
objects. Do not explain.
```