QUALITY- AND DIFFICULTY-AWARE DATA SELECTION FOR MULTILINGUAL EXTRACTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Training models on translated data is widely used to improve multilingual capabilities, particularly for low-resource languages where data scarcity remains a long-standing challenge. However, translated data can contain noise, making effective data selection crucial. While data quality can help identify noisy instances, it does not indicate how informative an instance is for learning. We therefore propose selecting translated data by jointly considering data quality and example difficulty. We study this approach in extraction tasks, where predictions are made over words or phrases and must be precisely grounded in the input text. Such tasks are particularly valuable targets for translation, since fine-grained annotation is costly to produce in new languages. However, they also pose a particular challenge: label annotations may become invalid after translation even when sentence-level translation quality is high. We show that machine translation metrics fail to capture such errors and complement them with label transfer accuracy, assessed using an LLM-as-a-judge. We evaluate these selection strategies by fine-tuning an LLM on English extraction tasks and their translations into 20 languages. Our results show that selection based on quality alone generally outperforms random selection, while prioritizing high-quality instances within a target difficulty range further outperforms quality-based selection alone on 7 of 8 evaluation datasets. Finally, we show that fine-tuning on translated data remains beneficial even for LLMs with strong multilingual abilities, outperforming few-shot prompting on 6 of 8 datasets.

1 INTRODUCTION

In recent years, large language models (LLMs) have made remarkable progress in multilingual capabilities. Whereas earlier models were largely English-centric and often suffered from the "curse of multilinguality" (Conneau et al., 2020; Chang et al., 2024; Blevins et al., 2024), newer models have substantially narrowed the performance gap between English and other languages. Given this progress, it is unclear whether, or to what extent, training on translated English data - a long-standing approach to improving performance in low-resource languages - continues to benefit multilingual LLMs. In our study, we find that translated data remains useful, but its benefits depend critically on data selection. This raises the question of which translated data is worth keeping. We study this question in the context of extraction tasks, which require *locating specific words or phrases that convey target information*, such as personal disclosures (Dou et al., 2024), hateful expressions (Mathew et al., 2021), or hallucinated details (Vazquez et al., 2025). Unlike instance-level classification, extraction tasks provide interpretable, segment-level decisions that enable targeted interventions such as redaction or correction. Because annotation at this level of granularity is costly, replicating such annotations across target languages can be prohibitively expensive, making translation an attractive alternative. However, extraction tasks pose a particular challenge for translated data selection: even when a translation is faithful overall, the original annotations may no longer identify valid segments in the translated text (❶ in Figure 1). Segments may be aligned with incorrect translated text, reordered, split, or lost entirely. Consequently, sentence-level translation quality does not necessarily indicate whether a translated example provides useful supervision for extraction. These challenges motivate two research questions: for extraction tasks, how can we assess the quality of translated data to guide selection, and beyond data quality, how can we further improve data selection?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

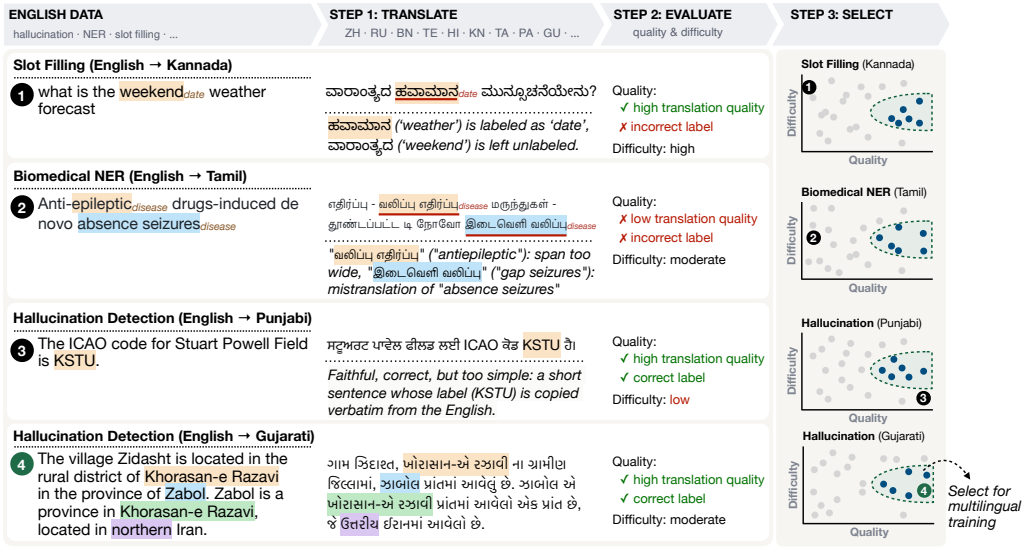


Figure 1: Overview of our quality- and difficulty-aware selection of translated training data. English data is translated into many languages, and each translated example is scored for quality (translation quality and label correctness) and difficulty. For each task–language pair, we retain high-quality translated examples within a target difficulty range (dashed green region), and combine the selected subsets across all pairs for multilingual fine-tuning. We show one translated example for each task–language pair and its position in the corresponding quality–difficulty space; matching colors (e.g., weekend) link each English label to its counterpart in the translation, and red underlines mark labeling errors. ❶ is discarded due to an incorrect label; ❷ is discarded because translation errors also corrupt its labels; ❸ has high quality but is too simple; and ❹ is selected.

To address the first question, we first examine whether existing reference-free machine translation metrics adequately capture the quality of translated data for extraction tasks. Prior work on general instruction tuning uses such metrics to filter translated data, discarding instances with low estimated translation quality (Zhang et al., 2024b). We find that these metrics alone are insufficient for extraction tasks. Specifically, we examine how well MT metrics correlate with downstream performance across 3 tasks and 12 target languages (§5). Translation quality correlates moderately to strongly with performance for low-resource languages, but only weakly to moderately for mid- to high-resource languages. For the latter, the overall translation quality is similar across translated datasets, while the main source of error instead lies in label transfer: annotated segments may not be correctly located in the translated text. We therefore introduce label transfer accuracy as a second dimension of data quality, using an LLM-as-a-judge to assess whether each annotated segment is correctly located in the translation. Label transfer accuracy correlates moderately to strongly with performance for most mid- to high-resource languages, complementing MT metrics, which are more informative for low-resource languages. Building on these findings, we select translated data using both dimensions, with XCOMET-XXL (Guerreiro et al., 2024) measuring translation quality and an LLM-as-a-judge measuring label transfer accuracy. We fine-tune multilingual instruction models on data translated into 20 languages, and find that selecting translated data based on the quality generally improves over random selection, although the gains vary across tasks.

Quality, however, is not the same as training usefulness: shorter and simpler sentences are both easier to translate and easier to transfer labels for (example ❸ in Figure 1), so quality-based selection may concentrate the training set on instances that the model already handles. We therefore hypothesize that selecting on both quality and difficulty provides greater training benefit than selecting on quality alone. To measure difficulty, we fine-tune a multilingual model on the English data and compute the loss of each translated instance under this model. Because the final model is trained on the English data together with selected translations in all target languages, this loss measures how much information a translated instance provides beyond what is already well covered by the English supervision. Difficulty alone is nonetheless ambiguous: a high loss may reflect a poor translation, an incorrect transferred label (example ❶ in Figure 1), or an instance that is rare or difficult to learn (Mindermann et al., 2022). The two signals are therefore complementary: weighting by quality sup-

presses instances whose difficulty comes from poor translation or incorrect label transfer, allowing difficulty to better reflect what the model has yet to learn. We score each translated instance as a product of its quality and a function of its difficulty that favors a target range, selecting instances that are both reliable and informative. Evaluating this selection strategy across the 8 downstream datasets, we find that it outperforms quality-based selection alone on 7 of them, often requiring less translated data (Figure 2).

Finally, we quantify the benefit of translate-train (i.e., training on translated versions of annotated English data) relative to few-shot prompting, comparing the two approaches using Gemma-3 models (Team et al., 2025) at different scales. For the 4B and 12B models, translate-train achieves higher performance than few-shot prompting on 6 of the 8 datasets, and the fine-tuned 12B model even surpasses the few-shot 27B model on 6 datasets (Figure 3). The two exceptions, where few-shot prompting performs better, are extractive QA and hate speech detection. For extractive QA, LLMs may already possess substantial task-relevant knowledge from pretraining, reducing the marginal benefit of additional translated supervision. For hate speech detection, hateful expressions are often highly language-specific and difficult to translate while preserving their intent, making translated supervision potentially less reliable. These findings show that even for LLMs with strong multilingual abilities, translate-train remains an effective strategy that complements few-shot prompting.

2 Problem Statement

In this section, we discuss the task formulation and the problem of selecting translated training data.

2.1 Background: Multilingual Training for Span-Level Extraction

Task formulation We frame extraction tasks within an instruction-tuning paradigm, where the model receives a task description, optional schema information, an input text, and generates a structured output containing labeled spans, as illustrated in Figure 10 in the Appendix.

Multilingual Training We adopt the translate-train approach, where the annotated English data are transferred to target languages via machine translation. To preserve the label spans after translation, we adopt the mark-then-translate approach, following Chen et al. (2025a). This approach inserts special markers around all labeled spans in the source sentence and then translates the modified sentence into the target languages. Specifically, we use the HTML tag to mark the labeled spans. The resulting multilingual data are combined with the original English data to form a unified training set. Formally, let D_{en} be the English dataset and D_{tr}^{l} be the translated dataset for each language $l \in L$, where L denotes the set of target languages. Let $D_{tr} = \cup_{l \in L} D_{tr}^{l}$ denote the set of all translated data. The model then is fine-tuned as follows: $\pi^* = \arg\min_{\pi} \mathcal{L}(\pi; D_{en} \cup D_{tr})$, where π denotes the model, $\mathcal{L}$ is the loss function.

2.2 Translated Data Selection Problem

As the translated training set D_{tr} grows with the number of target languages, including all translated instances becomes computationally expensive and may not improve performance. A common approach is to filter out low-quality instances, such as those with incorrectly transferred labels (Chen et al., 2023; Ebing & Glavaš, 2025) or unfaithful translations (Zhang et al., 2024b). Although effective, quality-based filtering focuses exclusively on whether an example is reliable, rather than on how much it contributes to learning. Yet even among correctly translated instances, some provide more learning signal than others. We therefore explore a complementary criterion: instance difficulty, which captures how informative a translated example is for learning. As we demonstrate, jointly considering quality and difficulty achieves strong performance more efficiently than quality-based selection alone across many multilingual settings.

3 Translated Data Selection Strategies

In this section, we describe several scoring functions for assessing translated training data and introduce the corresponding selection strategies.

Data Difficulty Translated training examples can differ in how much additional learning signal they provide even when they have similar quality. For instance, simple examples - such as extremely short phrases with no labels - are often easier to translate, resulting in high translation quality but limited learning value. We define the Data Difficulty as below:

$$\mathrm{d}(\bar{x}_i, \bar{y}_i) = \mathcal{L}(\pi_{En}(\bar{x}_i), \bar{y}_i) \tag{1}$$

where $(\bar{x}_i, \bar{y}_i)$ denotes a translated data point and its translated label, π_{En} is the model fine-tuned on the full English dataset, and $\mathcal{L}$ is the task-specific loss. Intuitively, if the model fine-tuned on English data can already solve a translated example, the example is considered *simple* and may provide limited additional learning signal. Conversely, examples that result in high loss are considered *difficult* and may be more informative for fine-tuning.

Data quality For data quality, we consider two criteria: (1) translation quality and (2) label transfer accuracy. Translation quality refers to how faithfully the target sentence conveys the meaning of the source sentence. To measure translation quality, we use XCOMET-XXL (Guerreiro et al., 2024), a reference-free machine translation metric. Label transfer accuracy measures the percentage of label spans that are correctly projected to the translated sentence and have precise boundaries. Specifically, we prompt an LLM to judge whether the target label span is a correct translation of the source label span and whether the target span incorrectly includes any surrounding words (details of the prompt are shown in Figure 18 in the Appendix). We use Gemma-3-27B-IT (Team et al., 2025) as the judge for this metric. To assess its reliability, we evaluate the correlation between the metric and downstream performance in §5, and observe a moderate-to-strong positive correlation in most settings. Further analysis is provided in §5 and Appendix D.

Quality-Difficulty Quality and difficulty can sometimes conflict. For instance, a high-quality example can be too simple, while a high-difficulty example may be noisy, as illustrated in ❶ in Figure 1. Additionally, extremely difficult examples may represent outliers (Mindermann et al., 2022). Therefore, we hypothesize that fine-tuning on high-quality translations that are not too extreme in terms of difficulty may yield better performance. To capture this intuition, we define a deterministic scoring function S that combines translation quality with a difficulty-based weighting term:

$$S(\bar{x}_i, \bar{y}_i) = q(\bar{x}_i, \bar{y}_i) \cdot \mathcal{N}(d(\bar{x}_i, \bar{y}_i); \tau, \sigma) \tag{2}$$

where q and d denote the quality (XCOMET-XXL) and difficulty scores (Eq. 1), respectively. In practice, we use the percentile ranks of q and d before computing S. $\mathcal{N}$ denotes a Gaussian weighting function centered at a target difficulty percentile τ with standard deviation σ. Importantly, the scoring function is fully deterministic and does not involve stochastic sampling. The Gaussian term is used only to smoothly prioritize examples whose difficulty is near the target percentile while softly reducing the weight of examples farther from the target difficulty. We set $\tau = 0.4$ and $\sigma = 0.7$ (details in Appendix B.4). We additionally compare the proposed Gaussian weighting formulation against alternative difficulty integration strategies in Appendix B.5.

Selection Strategies We introduce the selection strategies explored in this study. For all strategies, we first prioritize all translated data that has perfect label transfer accuracy. Then, the remaining data are ranked according to translation quality, difficulty, or a combination of both. This yields three strategies: **Quality**, **Difficulty**, and **Quality-Difficulty**, respectively.

4 The Effectiveness of Quality and Difficulty in Translated Data Selection

In this section, we verify which strategy is the most effective for selecting the translated data.

4.1 Setup

Data We sample 10,000 training examples from 9 English datasets for instruction fine-tuning and evaluate on 8 multilingual test sets spanning diverse tasks and languages. Dataset details are summarized in Table 2. For MultiCoNER-v2 (Fetahu et al., 2023), we evaluate on 2,000 randomly sampled examples per language, which enables efficient evaluation while providing reliable performance estimates. Based on the non-English languages represented in the test sets, we select 20

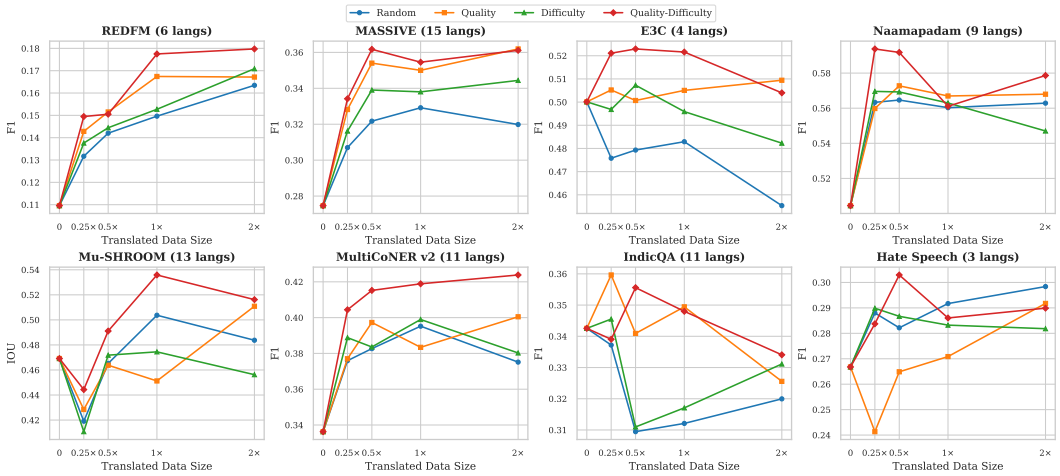


Figure 2: Performance of Gemma-3-4B-IT across benchmarks under different translated data selection strategies. The x-axis shows translated data size as a multiple ($\times$) of the English training data size, with the English data size held constant. For each benchmark, results are averaged across all non-English languages. Selecting data using both quality metrics and difficulty (**quality-difficulty**) enables the model to reach competitive peak performance earlier on most benchmarks.

target languages, including Arabic, Chinese, German, French, Italian, Spanish, Portuguese, Persian, Czech, Russian, Bulgarian, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu. We translate the English training data into all target languages and fine-tune Gemma-3-4B-IT using QLoRA (Dettmers et al., 2023) on the full English training set together with a subset of the translated data. Additional details, including the evaluation languages, the per-task breakdown of the English training data, and hyperparameters, are provided in Appendix B.4.

Setting & Baselines We use translations generated by Gemma-3-27B-IT with the mark-then-translate label projection method (Chen et al., 2023). Translated examples are ranked according to a selection strategy, and a subset is added to the training data. We vary the amount of translated data using translated-to-English ratios of 0.25, 0.5, 1, and 2, where the ratio is defined relative to the 10,000 English training examples (e.g., a ratio of 0.5 corresponds to 5,000 translated examples). Selection is performed independently for each task-language pair, ensuring proportional allocation of translated examples. In addition to **Quality**, **Difficulty**, and **Quality-Difficulty** (Eq. 2), we include **Random**, which ranks translated examples randomly, as a baseline.

4.2 Results

Figure 2 shows the performance of Gemma-3-4B-IT fine-tuned under different selection strategies and ratios of translated to English data across multilingual test datasets. Overall, Quality-Difficulty achieves the best peak performance on 7 of the 8 datasets and remains competitive on IndicQA. Compared to Quality-only selection, Quality-Difficulty is also more data-efficient: on all datasets except IndicQA, it surpasses the best score of Quality-only selection with the same or a smaller amount of translated data. Moreover, at the $1\times$ configuration, where the translated and English data are of equal size, Quality-Difficulty outperforms Quality-only in 50 of the 72 task–language pairs, a significant margin under a binomial sign test ($p < 0.01$), suggesting that the improvement is consistent across tasks and languages rather than driven by a small number of settings. We also conduct experiments comparing the two selection methods with Llama-3.2-3B-Instruct and observe a similar trend: based on the performance curves, Quality-Difficulty outperforms Quality-only on 5 of the 8 datasets and performs comparably on the remaining three. Details of the significance analysis and the full Llama results are provided in Appendix B.5. To isolate the contribution of the difficulty criterion, we also evaluate Difficulty-only selection. It performs comparably to Quality-only selection but peaks lower than Quality-Difficulty on every dataset, suggesting that quality and difficulty are complementary criteria that yield stronger results when combined. Finally, Random selection gen-

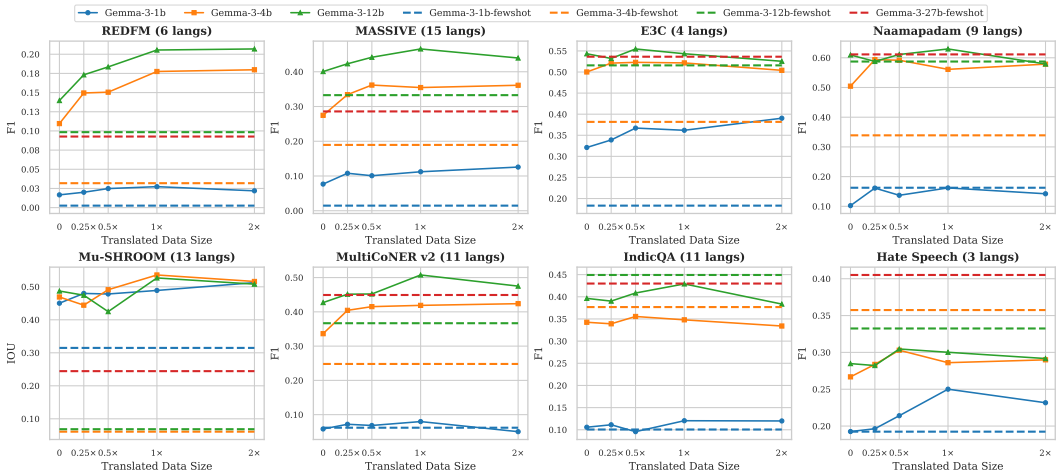


Figure 3: Few-shot prompting (dashed lines) and cross-lingual fine-tuning performance of different models from the Gemma family across benchmarks. The x-axis shows the translated data size as a multiple (×) of the English training data size, with the English data size held constant. For each benchmark, results are averaged across all non-English languages. Cross-lingual fine-tuning shows better performance over few-shot prompting for most benchmarks except IndicQA and Hate speech extraction for larger size models (4B and 12B).

erally lags behind other strategies, with the exception of Hate Speech. In this dataset, it outperforms Quality-only selection at every data size, and Quality-Difficulty is better only at the 0.5× ratio, although this is the best result on the dataset. A possible explanation is that the English Hate Speech datasets are derived from social media and contain noisy, informal, and often ungrammatical text, which makes translation quality harder to assess reliably with automatic metrics.

4.3 Analyses with different model sizes

We adopt the same experimental setup as the previous section and investigate the effect of model size. Specifically, we focus on Gemma-3 model family of the following sizes 1B, 4B, 12B. We use the Quality-Difficulty selection strategy to select top translated examples as it regularly achieves the best performance across datasets. The results for this analysis are shown in Figure 3.

The optimal ratio of the translated to English data Figure 3 illustrates the performance of different Gemma-3 models across datasets. For the 4B and 12B models, competitive performance is typically reached at translated-to-English ratios of 0.5 or 1.0. Beyond these points, performance often increases slowly or degrades, indicating that only a fraction of the translated data is needed for effective instruction fine-tuning.

Tasks that benefit from translate-train We compare performance of few-shot prompting (3 English demonstrations) with the original Gemma-3-IT models against the same models after fine-tuning on translated data. For Gemma-3-27B-IT, we report only few-shot performance, as a reference for prompting a larger model of the same family. As shown in Figure 3, even for instruction-tuned LLMs with strong multilingual abilities such as the Gemma-3 family, fine-tuning on translated data remains beneficial: for the 4B and 12B models, it outperforms few-shot prompting on most datasets, and the fine-tuned 12B model surpasses the few-shot 27B model on 6 of the 8 datasets. The exceptions are IndicQA and Hate Speech, where fine-tuned 4B and 12B models remain below their few-shot counterparts. For IndicQA, a likely reason is that question answering is a common task on which recent LLMs are already extensively trained, leaving little room for improvement from additional fine-tuning. For Hate Speech extraction, fine-tuning on translated data benefits only the smaller model (1B) but decreases performance for larger models. This may be due to cultural specificity: many hate speech expressions cannot be translated literally or may lose their intensity after translation.

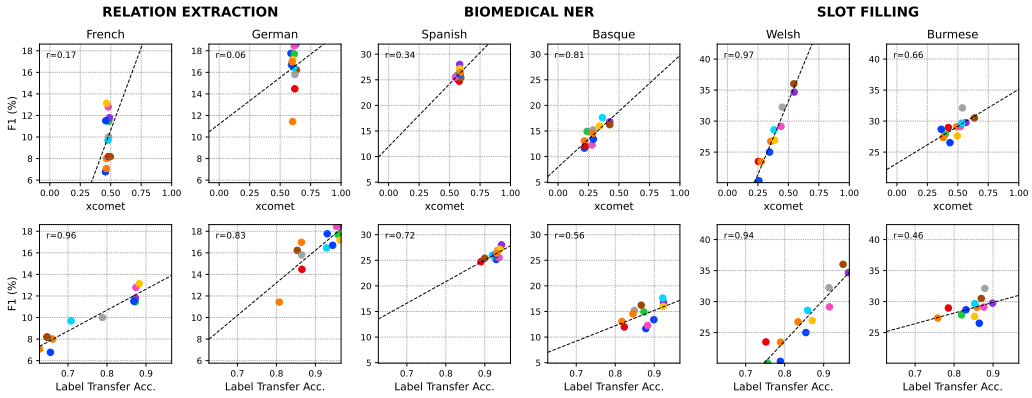


Figure 4: Pearson r correlations between model performance and XCOMET - translation quality (**Top**) or label transfer accuracy (**Bottom**). Each color indicates a different pipeline used to translate English training data. After translation, models are fine-tuned on the translated data; performance is shown on the y-axis. Translation quality has low or moderate correlation with performance in French, German (Relation Extraction) and Spanish (Biomedical NER), whereas label transfer accuracy shows a substantially stronger correlation in these settings. This difference arises because, in these settings, XCOMET scores are similar across translation datasets, while label transfer accuracy varies substantially. Correlations for additional task-language pairs are shown in Table 1.

5 How do Quality Metrics correlate with Downstream Performance?

In the previous section, we adopted a practical design combining a reference-free MT metric (XCOMET-XXL) with label transfer accuracy to assess translated data quality. Here, we revisit the broader question of how translated training data quality should be assessed. Specifically, we extend the analysis to additional MT metrics and investigate whether these metrics alone are sufficient or whether label-aware signals provide complementary information. For this purpose, we evaluate how each metric correlates with downstream performance in a controlled single-task, single-language setting. This setup minimizes confounding factors from multilingual multitask training, such as task and language interference. For each task-language pair, we construct multiple equal-sized translated datasets from the same English training data, thereby controlling cross-lingual difficulty and isolating data quality. We then combine each translated dataset with the original English data to fine-tune a separate model. We compute dataset-level scores by averaging instance-level quality scores and analyze how strongly different metrics correlate with downstream model performance.

5.1 Setup

Tasks and Datasets We study three downstream tasks: Relation Extraction, Biomedical NER and Slot Filling, spanning 12 target languages. Regarding the dataset, we use RED$^{\text{FM}}$ (Huguet Cabot et al., 2023) for Relation Extraction, E3C (Magnini et al., 2021) for Biomedical NER, and MASSIVE (FitzGerald et al., 2023) for Slot Filling. We translate the English training dataset of RED$^{\text{FM}}$, E3C, and MASSIVE, respectively, to construct the training data in target languages. For evaluation, we use the original test split for each target language in each dataset. Due to computational constraints, we select 5 low-resource languages from MASSIVE to experiment on. Details about the selected language per dataset are shown in Table 1. We use the F1 score as the evaluation metric for all tasks.

Experimental setup We construct 12 translated datasets for each task-language pair by leveraging different MT models and label projection methods (details in the Appendix C). As the correlation analysis requires training multiple models, we fine-tune Gemma-3-270M-IT for all the task-language pairs for computational efficiency.

Data quality metrics We study the following quality metrics: (1) translation quality, including MetricX-24 (Juraska et al., 2024), XCOMET-XXL (Guerreiro et al., 2024), and GEMBA-ESA

(Zouhar et al., 2024); (2) label transfer accuracy, which measures the percentage of label span being correctly transferred to the target language (§3). We use Gemma-3-27B-IT as the judge for both GEMBA-ESA and the label transfer accuracy.

5.2 RESULTS

Table 1 demonstrates the Pearson correlation coefficients (r) between data quality metrics and downstream performance.

Table 1: Pearson correlation (r) between data quality metrics and downstream performance (LT Acc.: Label transfer accuracy), computed over 12 data points per task-language pair. "*": statistical significance ($p < 0.05$). In Relation Extraction and Biomedical NER, translation quality metrics (MetricX, XCOMET, GEMBA-ESA) show low or moderate correlations in several languages, where translated datasets have similarly high translation quality but large differences in label transfer accuracy (see Fig 4).

Tasks	Benchmark	Data Quality Metric			
		MetricX	XCOMET	GEMBA-ESA	LT Acc.
Relation Extraction	RED^{FM}				
	Arabic	0.11	0.09	0.10	0.57
	German	0.11	0.06	0.12	0.83*
	Spanish	0.31	0.35	0.16	0.68*
	French	0.12	0.17	0.14	0.96*
	Italian	0.02	-0.04	-0.05	0.70*
	Chinese	0.19	0.07	0.02	0.33
	Overall	0.14	0.12	0.08	0.75
Biomedical NER	E3C				
	Spanish	0.48	0.34	0.54	0.72*
	French	0.22	0.04	0.36	0.98*
	Italian	0.29	0.09	0.44	0.74*
	Basque	0.81*	0.81*	0.82*	0.56
	Overall	0.50	0.38	0.57	0.82
Slot Filling	MASSIVE				
	Armenian	0.84*	0.87*	0.81*	0.79*
	Khmer	0.70*	0.70*	0.66*	0.61*
	Burmese	0.62*	0.66*	0.59*	0.46
	Kannada	0.62*	0.61*	0.56	0.75*
	Welsh	0.97*	0.97*	0.97*	0.94*
	Overall	0.81	0.83	0.79	0.76

MT metrics fail when label span errors dominate. In the Relation Extraction and Biomedical NER tasks, MT metrics generally show weak or moderate correlations with downstream performance, whereas label transfer accuracy correlates strongly in most languages. This occurs because the overall translation quality is similarly high across translated datasets, leaving limited variation for MT metrics to distinguish, while label transfer errors vary substantially, as seen in the French column of Figure 4. Manual analysis of the French examples reveals a common error where function words, such as articles, are incorrectly included at the beginning of entity spans: "le 4 décembre 2019", and "l'Université Pompeu Fabra". These types of error directly reduce downstream task performance, even when the translation itself is correct. In these examples, the main source of error is not incorrect semantic alignment, but mismatches between annotation conventions (include the functional words in the label spans or not). Consequently, MT metrics fail to capture these projection errors, whereas label transfer accuracy, which checks the projected label directly, provides a more reliable signal of dataset quality.

MT metrics succeed when translation errors dominate. Unlike cases dominated by annotation mismatches, MT metrics correlate strongly with downstream performance when translation quality is the main bottleneck. In Slot Filling, MT metrics correlate significantly with downstream performance for all languages (with the single exception of GEMBA-ESA on Kannada) and slightly outperform label transfer accuracy overall (Table 1). A similar pattern holds for Basque in Biomedical NER, where all three MT metrics reach correlations above 0.8. These languages typically receive lower-quality machine translation than high-resource languages such as French, Italian, and Spanish. When translation quality is poor, errors affect not only the label spans, which may be mistranslated, omitted, or difficult to align with the source, but also the surrounding context, degrading the training signal as a whole. MT metrics capture this sentence-level degradation, whereas label transfer accuracy, which checks only the projected labels, can miss it: for Basque and Burmese, MT metrics correlate strongly while label transfer accuracy does not reach significance.

6 RELATED WORK

Translation-based Data Augmentation for Extraction Tasks Cross-lingual transfer for extraction tasks often relies on machine translation with label projection, which maps source spans to translated text. Prior work has explored several label projection strategies, including word-alignment-based projection (Fei et al., 2020; Ebing & Glavaš, 2025), approaches that rely on MT systems to jointly translate the sentence and project labels (Lee et al., 2018; Hu et al., 2020; Chen et al., 2023;

Table 2: Datasets used in the instruction fine-tuning experiments.

Dataset	Domain	Task	Languages	Usage
NCBIDisease (Doğan et al., 2014)	Biomedical	NER	English	Train
BC5CDR (Li et al., 2016)	Biomedical	NER	English	Train
MultiSpanQA (Li et al., 2022)	Wiki	QA	English	Train
$\text{MASSIVE}_{\text{En}}$ (FitzGerald et al., 2023)	Spoken	Slot filling	English	Train
$\text{RED}^{\text{FM}}{}_{\text{En}}$ (Huguet Cabot et al., 2023)	Wiki	Relation extraction	English	Train
Few-NERD (Ding et al., 2021)	Wiki	Fine-grained NER	English	Train
TBO (Zampieri et al., 2023)	Social	Targeted Offensive Detection	English	Train
HateXplain (Mathew et al., 2021)	Social	Hate speech explanation	English	Train
PsiloQA (Rykov et al., 2025)	Wiki	Hallucination detection	English	Train
E3C (Magnini et al., 2021)	Biomedical	NER	4 non-En	Eval.
Naamapadam (Mhaske et al., 2023)	News	NER	9 Indic	Eval.
IndicQA (Doddapaneni et al., 2023)	Wiki	QA	11 Indic	Eval.
MASSIVE (FitzGerald et al., 2023)	Spoken	Slot filling	15 low-resource	Eval.
RED^{FM} (Huguet Cabot et al., 2023)	Wiki	Relation extraction	6 non-En	Eval.
MultiCoNER-v2 (Fetahu et al., 2023)	Wiki	Fine-grained NER	11 non-En	Eval.
Mu-SHROOM (Vazquez et al., 2025)	Wiki	Hallucination detection	13 non-En	Eval.
Hate Speech				
ToxiCN (Bai et al., 2025)	Social	Toxic span extraction	Chinese	Eval.
KOLD (Jeong et al., 2022)	Social	Offensive span extraction	Korean	Eval.
ViHOS (Hoang et al., 2023)	Social	Hateful/offensive span extraction	Vietnamese	Eval.

Le et al., 2024), dedicated projection models trained for span projection (García-Ferrero et al., 2023), and more recently prompting LLMs to perform span mapping (Parekh et al., 2024; Zuo et al., 2025).

Selection of Translated Training Data In extraction tasks, a common practice is to discard translated instances with failed label projections (García-Ferrero et al., 2023; Parekh et al., 2024), sometimes combined with rule-based heuristics (Chen et al., 2023; Le et al., 2024; Ebing & Glavaš, 2025). Beyond selection, Zuo et al. (2025) refine translated sentences and projected spans through rephrasing to improve translation and label quality. In this work, we assess translated data quality in terms of translation quality, measured with reference-free MT metrics, and label transfer accuracy, judged by LLMs. While reference-free MT metrics are widely used to filter noisy parallel corpora (Peter et al., 2023), their effectiveness for translated data with span-level labels remains unclear.

Instruction-Based Formulations for Extraction Tasks Several studies have extended instruction tuning to extraction tasks by reformulating them as instruction-following problems and fine-tuning LLMs accordingly (Wang et al., 2023; Sainz et al., 2024; Mirza et al., 2024). Some work further applies this approach to multilingual settings, particularly English-Chinese scenarios (Xiao et al., 2023; Gui et al., 2024; Zuo et al., 2025). More recently, Chen et al. (2025a) extended instruction tuning to extraction tasks across 36 languages, including many low-resource ones. However, prior work mainly focuses on traditional extraction tasks such as NER and event extraction, with less attention to emerging tasks. Our work builds on these formulations to study translated data selection for multilingual extraction across diverse tasks. More related work is discussed in Appendix A.

7 Conclusion

In this work, we study how to select translated data for multilingual extraction tasks. We show that reference-free MT metrics often miss translation-induced label errors, while label transfer accuracy provides a complementary quality signal. Prioritizing translated data by quality, considering both translation quality and label transfer accuracy, generally improves performance over random selection, and additionally accounting for example difficulty leads to further gains. We also find that, for the Gemma-3-IT 4B and 12B models, using translated data at $0.5\times$ or $1\times$ the size of the English data achieves competitive or even superior performance compared to training with larger translated datasets. Finally, even for LLMs with strong multilingual abilities, fine-tuning on translated data remains useful, outperforming few-shot prompting on 6 of 8 datasets.

AI USE STATEMENT

We used generative AI tools to assist with editing and polishing the manuscript for clarity and readability, and to support software coding during the research process. All AI-assisted code was reviewed, tested, and verified by the authors. We did not use generative AI to formulate the research methodology, hypotheses, mathematical claims, proofs, or interpret experimental results. We reviewed all AI-assisted content and take full responsibility for the final content of the paper, including all claims, code, and artifacts.

REPRODUCIBILITY STATEMENT

To support reproducibility, we provide the details of the datasets used in our experiments in Table 3, while hyperparameter configurations and experimental settings are described in the Appendix B and C. Upon publication, we will release the source code together with the processed data or the scripts needed to reproduce it.

REFERENCES

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1998–2022, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.130. URL https://aclanthology.org/2022.emnlp-main.130/.

Wasi Ahmad, Jianfeng Chi, Tu Le, Thomas Norton, Yuan Tian, and Kai-Wei Chang. Intent classification and slot filling for privacy policies. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 4402–4417, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.340. URL https://aclanthology.org/2021.acl-long.340/.

Fan Bai, Alan Ritter, and Wei Xu. Pre-train or annotate? domain adaptation with a constrained budget. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 5002–5015, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.409. URL https://aclanthology.org/2021.emnlp-main.409/.

Zewen Bai, Liang Yang, Shengdi Yin, Junyu Lu, Jingjie Zeng, Haohao Zhu, Yuanyuan Sun, and Hongfei Lin. STATE ToxiCN: A benchmark for span-level target-aware toxicity extraction in Chinese hate speech detection. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 10206–10219, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.532. URL https://aclanthology.org/2025.findings-acl.532/.

Terra Blevins, Tomasz Limisiewicz, Suchin Gururangan, Margaret Li, Hila Gonen, Noah A. Smith, and Luke Zettlemoyer. Breaking the curse of multilinguality with cross-lingual expert language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10822–10837, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.604. URL https://aclanthology.org/2024.emnlp-main.604/.

Alexander Bukharin, Shiyang Li, Zhengyang Wang, Jingfeng Yang, Bing Yin, Xian Li, Chao Zhang, Tuo Zhao, and Haoming Jiang. Data diversity matters for robust instruction tuning. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 3411–3425, Miami, Florida, USA, November 2024.

Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.195. URL https://aclanthology.org/2024.findings-emnlp.195/.

Yihan Cao, Yanbin Kang, Chi Wang, and Lichao Sun. Instruction mining: Instruction data selection for tuning large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=wF6k0aWjAu.

Tommaso Caselli and Piek Vossen. The event StoryLine corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pp. 77–86, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-2711. URL https://www.aclweb.org/anthology/W17-2711.

Tyler A. Chang, Catherine Arnett, Zhuowen Tu, and Benjamin K. Bergen. When is multilinguality a curse? language modeling for 250 high- and low-resource languages. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4074–4096, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.236. URL https://aclanthology.org/2024.emnlp-main.236/.

Lichang Chen, Shiyang Li, Jun Yan, Hai Wang, Kalpa Gunaratna, Vikas Yadav, Zheng Tang, Vijay Srinivasan, Tianyi Zhou, Heng Huang, and Hongxia Jin. Alpagasus: Training a better alpaca with fewer data. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=FdVXgSJhvz.

Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. Frustratingly easy label projection for cross-lingual transfer. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5775–5796, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.357. URL https://aclanthology.org/2023.findings-acl.357/.

Yang Chen, Vedaant Shah, and Alan Ritter. Translation and fusion improves cross-lingual information extraction. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7744–7764, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.382. URL https://aclanthology.org/2025.acl-long.382/.

Yicheng Chen, Yining Li, Kai Hu, Ma Zerun, HaochenYe HaochenYe, and Kai Chen. MIG: Automatic data selection for instruction tuning by maximizing information gain in semantic space. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9902–9915, Vienna, Austria, July 2025b. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.515. URL https://aclanthology.org/2025.findings-acl.515/.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.747. URL https://aclanthology.org/2020.acl-main.747/.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 10088–10115. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/1feb87871436031bdc0f2beaa62a049b-Paper-Conference.pdf.

Ning Ding, Guangwei Xu, Yulin Chen, Xiaobin Wang, Xu Han, Pengjun Xie, Haitao Zheng, and Zhiyuan Liu. Few-NERD: A few-shot named entity recognition dataset. In Chengqing

Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3198–3213, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.248. URL https://aclanthology.org/2021.acl-long.248/.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12402–12426, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.693. URL https://aclanthology.org/2023.acl-long.693/.

Simona Emilova Doneva, Tilia Ellendorff, Beate Sick, Jean-Philippe Goldman, Amelia Elaine Cannon, Gerold Schneider, and Benjamin Victor Ineichen. NeuroTrialNER: An annotated corpus for neurological diseases and therapies in clinical trial registries. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18868–18890, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.1050. URL https://aclanthology.org/2024.emnlp-main.1050/.

Yao Dou, Isadora Krsek, Tarek Naous, Anubha Kabra, Sauvik Das, Alan Ritter, and Wei Xu. Reducing privacy risks in online self-disclosures with language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13732–13754, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.741. URL https://aclanthology.org/2024.acl-long.741/.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: A resource for disease name recognition and concept normalization. *Journal of Biomedical Informatics*, 47: 1–10, 2014. ISSN 1532-0464. doi: https://doi.org/10.1016/j.jbi.2013.12.006. URL https://www.sciencedirect.com/science/article/pii/S1532046413001974.

Qianlong Du, Chengqing Zong, and Jiajun Zhang. Mods: Model-oriented data selection for instruction tuning. *ArXiv*, abs/2311.15653, 2023. URL https://api.semanticscholar.org/CorpusID:265457248.

Benedikt Ebing and Goran Glavaš. The devil is in the word alignment details: On translation-based cross-lingual transfer for token classification tasks. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 18111–18128, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.931. URL https://aclanthology.org/2025.findings-acl.931/.

Hao Fei, Meishan Zhang, and Donghong Ji. Cross-lingual semantic role labeling with high-quality translated training corpus. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7014–7026, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.627. URL https://aclanthology.org/2020.acl-main.627/.

Besnik Fetahu, Zhiyu Chen, Sudipta Kar, Oleg Rokhlenko, and Shervin Malmasi. MultiCoNER v2: a large multilingual dataset for fine-grained and noisy named entity recognition. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 2027–2051, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.134. URL https://aclanthology.org/2023.findings-emnlp.134/.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In

Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4277–4302, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.235. URL https://aclanthology.org/2023.acl-long.235/.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. T-projection: High quality annotation projection for sequence labeling tasks. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15203–15217, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.1015. URL https://aclanthology.org/2023.findings-emnlp.1015/.

Arushi Goel, Basura Fernando, Frank Keller, and Hakan Bilen. Who are you referring to? coreference resolution in image narrations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15247–15258, 2023.

Pierpaolo Goffredo, Mariana Chaves, Serena Villata, and Elena Cabrio. Argument-based detection and classification of fallacies in political debates. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 11101–11112, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.684. URL https://aclanthology.org/2023.emnlp-main.684/.

Cristina González, Nicolás Ayobi, Isabela Hernández, José Hernández, Jordi Pont-Tuset, and Pablo Arbeláez. Panoptic narrative grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1364–1373, 2021.

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

Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995, 2024. doi: 10.1162/tacl_a_00683. URL https://aclanthology.org/2024.tacl-1.54/.

Honghao Gui, Lin Yuan, Hongbin Ye, Ningyu Zhang, Mengshu Sun, Lei Liang, and Huajun Chen. IEPile: Unearthing large scale schema-conditioned information extraction corpus. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 127–146, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-short.13. URL https://aclanthology.org/2024.acl-short.13/.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *AAAI Conference on Artificial Intelligence*, 2023. URL https://api.semanticscholar.org/CorpusID:260887222.

Hassan Hamad, Abhinav Kumar Thakur, Nijil Kolleri, Sujith Pulikodan, and Keith Chugg. FIRE: A dataset for financial relation extraction. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3628–3642, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.230. URL https://aclanthology.org/2024.findings-naacl.230/.

David Heineman, Yao Dou, and Wei Xu. Thresh: A unified, customizable and deployable platform for fine-grained text evaluation. In Yansong Feng and Els Lefever (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 336–345, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-demo.30. URL https://aclanthology.org/2023.emnlp-demo.30/.

Phu Gia Hoang, Canh Duc Luu, Khanh Quoc Tran, Kiet Van Nguyen, and Ngan Luu-Thuy Nguyen. ViHOS: Hate speech spans detection for Vietnamese. In Andreas Vlachos and Isabelle Augenstein (eds.), *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 652–669, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.eacl-main.47. URL https://aclanthology.org/2023.eacl-main.47/.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4411–4421. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/hu20b.html.

Pere-Lluís Huguet Cabot, Simone Tedeschi, Axel-Cyrille Ngonga Ngomo, and Roberto Navigli. RED^{fm}: a filtered and multilingual relation extraction dataset. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4326–4343, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.237. URL https://aclanthology.org/2023.acl-long.237/.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. KOLD: Korean offensive language dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10818–10833, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.744. URL https://aclanthology.org/2022.emnlp-main.744/.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. The state and fate of linguistic diversity and inclusion in the NLP world. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.560. URL https://aclanthology.org/2020.acl-main.560/.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 492–504, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.wmt-1.35.

Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. Named entity recognition in Indian court judgments. In Nikolaos Aletras, Ilias Chalkidis, Leslie Barrett, Cătălina Goanță, and Daniel Preoțiuc-Pietro (eds.), *Proceedings of the Natural Legal Language Processing Workshop 2022*, pp. 184–193, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.nllp-1.15. URL https://aclanthology.org/2022.nllp-1.15/.

Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. Constrained decoding for cross-lingual label projection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=DayPQKXaQk.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. Semi-supervised training data generation for multilingual question answering. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL https://aclanthology.org/L18-1437/.

Haonan Li, Martin Tomko, Maria Vasardani, and Timothy Baldwin. MultiSpanQA: A dataset for multi-span question answering. In Marine Carpuat, Marie-Catherine de Marneffe, and

Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 1250–1260, Seattle, United States, July 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.naacl-main.90. URL https://aclanthology.org/2022.naacl-main.90/.

Jiangnan Li, Thuy-Trang Vu, Christian Herold, Amirhossein Tebbifakhr, Shahram Khadivi, and Gholamreza Haffari. CONGRAD: Conflicting gradient filtering for multilingual preference alignment. In Vera Demberg, Kentaro Inui, and Lluís Marquez (eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6371–6387, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.299. URL https://aclanthology.org/2026.eacl-long.299/.

Jiao Li, Yueping Sun, Robin J. Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J. Mattingly, Thomas C. Wiegers, and Zhiyong Lu. Biocreative v cdr task corpus: a resource for chemical disease relation extraction. *Database*, 2016:baw068, 05 2016. ISSN 1758-0463. doi: 10.1093/database/baw068. URL https://doi.org/10.1093/database/baw068.

Manling Li, Alireza Zareian, Qi Zeng, Spencer Whitehead, Di Lu, Heng Ji, and Shih-Fu Chang. Cross-media structured common space for multimedia event extraction. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2557–2568, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.230. URL https://aclanthology.org/2020.acl-main.230/.

Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. From quantity to quality: Boosting LLM performance with self-guided data selection for instruction tuning. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7602–7635, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.421. URL https://aclanthology.org/2024.naacl-long.421/.

Swee Kiat Lim, Aldrian Obaja Muis, Wei Lu, and Chen Hui Ong. MalwareTextDB: A database for annotated malware articles. In Regina Barzilay and Min-Yen Kan (eds.), *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1557–1567, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1143. URL https://aclanthology.org/P17-1143/.

Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. A token-level reference-free hallucination detection benchmark for free-form text generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6723–6737, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.464. URL https://aclanthology.org/2022.acl-long.464/.

Wei Liu, Weihao Zeng, Keqing He, Yong Jiang, and Junxian He. What makes good data for alignment? a comprehensive study of automatic data selection in instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=BTKAeLqLMw.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks. *ArXiv*, abs/1511.06343, 2015. URL https://api.semanticscholar.org/CorpusID:5324823.

Teresa Lynn, Malik H. Altakrori, Samar M. Magdy, Rocktim Jyoti Das, Chenyang Lyu, Mohamed Nasr, Younes Samih, Kirill Chirkunov, Alham Fikri Aji, Preslav Nakov, Shantanu Godbole, Salim Roukos, Radu Florian, and Nizar Habash. From multiple-choice to extractive QA: A case study for English and Arabic. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend

Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 2456–2477, Abu Dhabi, UAE, January 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.coling-main.168/.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. The e3c project: European clinical case corpus. In *Annual Conference of the Spanish Society for Natural Language Processing*, 2021. URL https://api.semanticscholar.org/CorpusID:248700142.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 14867–14875, 2021.

Arnav Mhaske, Harshit Kedia, Sumanth Doddapaneni, Mitesh M. Khapra, Pratyush Kumar, Rudra Murthy, and Anoop Kunchukuttan. Naamapadam: A large-scale named entity annotated data for Indic languages. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10441–10456, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.582. URL https://aclanthology.org/2023.acl-long.582/.

Sören Mindermann, Jan M Brauner, Muhammed T Razzak, Mrinank Sharma, Andreas Kirsch, Winnie Xu, Benedikt Höltgen, Aidan N Gomez, Adrien Morisot, Sebastian Farquhar, et al. Prioritized training on points that are learnable, worth learning, and not yet learnt. In *International Conference on Machine Learning*, pp. 15630–15649. PMLR, 2022.

Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. ILLUMINER: Instruction-tuned large language models as few-shot intent classifier and slot filler. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 8639–8651, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.758/.

Sungrim Moon, Serguei Pakhomov, Nathan Liu, James O Ryan, and Genevieve B Melton. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. *Journal of the American Medical Informatics Association*, 21(2):299–307, 2014.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain J Marshall, Ani Nenkova, and Byron C Wallace. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2018, pp. 197, 2018.

Jinlong Pang, Jiaheng Wei, Ankit Shah, Zhaowei Zhu, Yaxuan Wang, Chen Qian, Yang Liu, Yujia Bao, and Wei Wei. Improving data efficiency via curating LLM-driven rating systems. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=DKkQtRMowq.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. Contextual label projection for cross-lingual structured prediction. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5738–5757, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-long.321. URL https://aclanthology.org/2024.naacl-long.321/.

John Pavlopoulos, Léo Laugier, Alexandros Xenos, Jeffrey Sorensen, and Ion Androutsopoulos. From the detection of toxic spans in online discussions to the analysis of toxic-to-civil transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL 2022).*, Dublin, Ireland, May 2022. Association for Computational Linguistics.

Jan-Thorsten Peter, David Vilar, Daniel Deutsch, Mara Finkelstein, Juraj Juraska, and Markus Freitag. There's no data like better data: Using QE metrics for MT data filtering. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 561–577, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.50. URL https://aclanthology.org/2023.wmt-1.50/.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pp. 2641–2649, 2015.

Hanoona Rasheed, Muhammad Maaz, Sahal Shaji, Abdelrahman Shaker, Salman Khan, Hisham Cholakkal, Rao M Anwer, Eric Xing, Ming-Hsuan Yang, and Fahad S Khan. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13009–13018, 2024.

Elisei Rykov, Kseniia Petrushina, Maksim Savkin, Valerii Olisov, Artem Vazhentsev, Kseniia Titova, Alexander Panchenko, Vasily Konovalov, and Julia Belikova. When models lie, we learn: Multilingual span-level hallucination detection with PsiloQA. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11663–11682, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.626. URL https://aclanthology.org/2025.findings-emnlp.626/.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Y3wpuxd7u9.

Taneeya Satyapanich, Francis Ferraro, and Tim Finin. Casie: Extracting cybersecurity event information from text. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 8749–8757, 2020.

Megha Sundriyal, Atharva Kulkarni, Vaibhav Pulastya, Md. Shad Akhtar, and Tanmoy Chakraborty. Empowering the fact-checkers! automatic identification of claim spans on Twitter. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 7701–7715, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.525. URL https://aclanthology.org/2022.emnlp-main.525/.

Fiona Anting Tan, Hansi Hettiarachchi, Ali Hürriyetoğlu, Nelleke Oostdijk, Tommaso Caselli, Tadashi Nomoto, Onur Uca, Farhana Ferdousi Liza, and See-Kiong Ng. RECESS: Resource for extracting cause, effect, and signal spans. In Jong C. Park, Yuki Arase, Baotian Hu, Wei Lu, Derry Wijaya, Ayu Purwarianti, and Adila Alfa Krisnadhi (eds.), *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 66–82, Nusa Dua, Bali, November 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.ijcnlp-main.6. URL https://aclanthology.org/2023.ijcnlp-main.6/.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. FinEntity: Entity-level sentiment classification for financial texts. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15465–15471, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.956. URL https://aclanthology.org/2023.emnlp-main.956/.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Raul Vazquez, Timothee Mickus, Elaine Zosa, Teemu Vahtola, Jörg Tiedemann, Aman Sinha, Vincent Segonne, Fernando Sanchez Vega, Alessandro Raganato, Jindřich Libovický, Jussi Karlgren, Shaoxiong Ji, Jindřich Helcl, Liane Guillou, Ona De Gibert, Jaione Bengoetxea, Joseph Attieh, and Marianna Apidianaki. SemEval-2025 task 3: Mu-SHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes. In Sara Rosenthal, Aiala Rosá, Debanjan Ghosh, and Marcos Zampieri (eds.), *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 2472–2497, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-273-2. URL https://aclanthology.org/2025.semeval-1.322/.

Xiao Wang, Weikang Zhou, Can Zu, Han Xia, Tianze Chen, Yuansen Zhang, Rui Zheng, Junjie Ye, Qi Zhang, Tao Gui, et al. Instructuie: Multi-task instruction tuning for unified information extraction. *arXiv preprint arXiv:2304.08085*, 2023.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 926–941, Abu Dhabi, United Arab Emirates, December 2022a. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.60. URL https://aclanthology.org/2022.emnlp-main.60/.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Anjana Arunkumar, Arjun Ashok, Arut Selvan Dhanasekaran, Atharva Naik, David Stap, et al. Super-naturalinstructions:generalization via declarative instructions on 1600+ tasks. In *EMNLP*, 2022b.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A Smith, Mari Ostendorf, and Hannaneh Hajishirzi. Fine-grained human feedback gives better rewards for language model training. *Advances in Neural Information Processing Systems*, 36:59008–59033, 2023.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=PG5fV50maR.

Xinglin Xiao, Yijie Wang, Nan Xu, Yuqi Wang, Hanxuan Yang, Minzheng Wang, Yin Luo, Lei Wang, Wenji Mao, and Daniel Zeng. Yayi-uie: A chat-enhanced instruction tuning framework for universal information extraction. *arXiv preprint arXiv:2312.15548*, 2023.

Xiaoyu Yang, Stephen Obadinma, Huasha Zhao, Qiong Zhang, Stan Matwin, and Xiaodan Zhu. SemEval-2020 task 5: Counterfactual recognition. In Aurelie Herbelot, Xiaodan Zhu, Alexis Palmer, Nathan Schneider, Jonathan May, and Ekaterina Shutova (eds.), *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 322–335, Barcelona (online), December 2020. International Committee for Computational Linguistics. doi: 10.18653/v1/2020.semeval-1.40. URL https://aclanthology.org/2020.semeval-1.40/.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. DocRED: A large-scale document-level relation extraction dataset. In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 764–777, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1074. URL https://aclanthology.org/P19-1074/.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. Target-based offensive language identification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.66. URL https://aclanthology.org/2023.acl-short.66/.

Jia Zhang, Chen-Xi Zhang, Yao Liu, Yi-Xuan Jin, Xiao-Wen Yang, Bo Zheng, Yi Liu, and Lan-Zhe Guo. D3: diversity, difficulty, and dependability-aware data selection for sample-efficient llm instruction tuning. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, IJCAI '25, 2025. ISBN 978-1-956792-06-5. doi: 10.24963/ijcai.2025/928. URL https://doi.org/10.24963/ijcai.2025/928.

Qi Zhang, Zhijia Chen, Huitong Pan, Cornelia Caragea, Longin Jan Latecki, and Eduard Dragut. SciER: An entity and relation extraction dataset for datasets, methods, and tasks in scientific documents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 13083–13100, Miami, Florida, USA, November 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.726. URL https://aclanthology.org/2024.emnlp-main.726/.

Yuanchi Zhang, Yile Wang, Zijun Liu, Shuo Wang, Xiaolong Wang, Peng Li, Maosong Sun, and Yang Liu. Enhancing multilingual capabilities of large language models through self-distillation from resource-rich languages. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11189–11204, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.603. URL https://aclanthology.org/2024.acl-long.603/.

Chunguang Zhao, Yilun Liu, Pufan Zeng, Yuanchang Luo, Shimin Tao, Minggui He, Weibin Meng, Song Xu, Ziang Chen, Chen Liu, et al. A method for improving multilingual quality and diversity of instruction fine-tuning datasets. *arXiv preprint arXiv:2509.15549*, 2025.

Yang Zhong, Jingfeng Yang, Wei Xu, and Diyi Yang. WIKIBIAS: Detecting multi-span subjective biases in language. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 1799–1814, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.findings-emnlp.155.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, LILI YU, Susan Zhang, Gargi Ghosh, Mike Lewis, Luke Zettlemoyer, and Omer Levy. LIMA: Less is more for alignment. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=KBMOKmX2he.

Haotian Zhou, Tingkai Liu, Qianli Ma, Yufeng Zhang, Jianbo Yuan, Pengfei Liu, Yang You, and Hongxia Yang. DavIR: Data selection via implicit reward for large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9220–9237, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.452. URL https://aclanthology.org/2025.acl-long.452/.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. Ai-assisted human evaluation of machine translation. *arXiv preprint arXiv:2406.12419*, 2024.

Yuxin Zuo, Wenxuan Jiang, Wenxuan Liu, Zixuan Li, Long Bai, Hanbin Wang, Yutao Zeng, Xiaolong Jin, Jiafeng Guo, and Xueqi Cheng. KnowCoder-X: Boosting multilingual information extraction via code. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14486–14509, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.748. URL https://aclanthology.org/2025.findings-acl.748/.

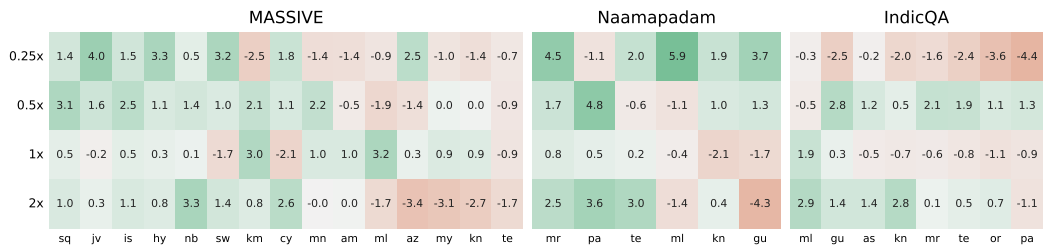


Figure 5: Performance gap in F1 (%) between two selection strategies, which are Quality-Difficulty and Quality, in low-resource languages across three datasets. The x-axis represents languages, while the y-axis denotes the translated data size as a multiple ($\times$) of the English training data size, with the English data size held constant. Green cells indicate superior performance by Quality-Difficulty. Quality-Difficulty's improvements are most pronounced in extreme low-data regimes ($0.25\times$ and $0.5\times$) across the MASSIVE and Naamapadam datasets, while showing less consistent improvements on IndicQA.

A Additional Related Work

Our work is related to prior literature on data selection for instruction-formatted training data. Most prior work focuses on English instruction-formatted data and general instruction-following settings. Zhou et al. (2023) demonstrates that training on a small set of carefully curated instruction data can yield strong instruction-following performance. Subsequent work has proposed automatic data selection frameworks that prioritize training examples according to desirable characteristics. One line of research studies how to estimate data quality and prioritize high-quality instruction data (Chen et al., 2024; Cao et al., 2024). Other work incorporates complementary dimensions such as diversity (Du et al., 2023; Liu et al., 2024; Bukharin et al., 2024; Pang et al., 2025; Chen et al., 2025b), or difficulty (Li et al., 2024; Zhang et al., 2025). Prior work has explored loss-based data selection strategies that prioritize informative or learnable training instances (Loshchilov & Hutter, 2015; Mindermann et al., 2022; Zhou et al., 2025). Another line of work leverages gradient information to select examples that are most relevant to a small set of representative instances (Xia et al., 2024). Our work also studies the data selection problem, focusing on two characteristics of training data: quality and difficulty. However, we consider a different setting: selecting translated data for multilingual training on span-extraction tasks, which requires different approaches for assessing these characteristics. In contrast to prior instruction-tuning settings, assessing the quality of translated span-extraction data requires verifying both translation fidelity and the correctness of cross-lingual span alignment. Moreover, unlike prior work that studies the intrinsic difficulty of instruction data, we study cross-lingual difficulty in multilingual training, focusing on the additional learning signal provided by translated examples beyond the corresponding English data.

There have been efforts to study the instruction data selection in the multilingual setting. Zhao et al. (2025) selects the multilingual data based on the quality and diversity dimensions. Different from our work, Zhao et al. (2025) assumes access to curated multilingual data to train a scoring model to assess the data quality, whereas our setting assumes only English supervision during training and relies on translated data without target-language annotations. Finally, Li et al. (2026) study multilingual data selection in the context of preference alignment, proposing a gradient-based method to mitigate negative interference during multilingual training. Unlike our work, their setting focuses on preference optimization rather than supervised fine-tuning.

B Additional Details and Results for Data Selection Experiments

In this section, we provide more details about the experiments conducted in §4.

B.1 TARGET LANGUAGES IN THE EXPERIMENTS

We translate the English datasets to the following target languages: Arabic, Chinese, German, French, Italian, Spanish, Portuguese, Persian, Czech, Russian, Bulgarian, Bengali, Gujarati, Hindi, Kannada, Malayalam, Marathi, Punjabi, Tamil, and Telugu. The target languages are chosen based on the language coverage of our test sets.

B.2 DATASET

Table 3 provides details about the license and languages for each test dataset. For the English training data, we sample 1,500 examples each from the MultiSpanQA and PsiloQA datasets, and 1,000 examples from each of the remaining 7 English datasets, yielding a total of 10,000 training examples.

B.3 INSTRUCTION TUNING PROMPTS

Details about the instruction tuning prompts are illustrated in Figures 11, 12, 13, 14, 15, 16, 17.

B.4 EXPERIMENTAL SETUP

Fine-tuning We fine-tune the instruction-tuned variants of the Gemma-3 models at three parameter scales: 1B, 4B, and 12B. Fine-tuning is performed using the Unsloth library. Specifically, we use the Unsloth-optimized checkpoints of the models[1]. Parameter-efficient fine-tuning is conducted using QLoRA applied to all linear layers. The LoRA configuration uses rank $r = 32$ and scaling factor $\alpha = 32$. Training is performed with a learning rate of 2e-4, number of training epochs of 2, batch size of 8, and gradient accumulation of 2 steps. We employ a linear learning-rate scheduler with a warm-up ratio of 0.1. We use 1 A40 GPU 48GB to fine-tune the models. Finally, performing 1,000 updates for 1B, 4B, and 12B models takes 1.1, 3.1, 5.4 GPU hours, respectively. All strategies except Random are trained once per configuration, while Random results are averaged over three independently sampled datasets to account for sampling variance. We assess the robustness of the single-run comparisons with a paired bootstrap significance analysis in §B.5.

Choosing hyperparameters for the Quality-Difficulty function The Quality-Difficulty function defined in Eq. 2 contains two hyperparameters, τ and σ. We first fix $\sigma = 0.7$ and search over $\tau \in \{0.2, 0.4, 0.6, 0.8\}$. After determining the best value of τ, we then search over $\sigma \in \{0.3, 0.5, 0.7\}$. We use the English validation set for selecting the optimal configuration, which follows the standard cross-lingual setting in prior work (Chen et al., 2023), where only source-language data is available during the training time. The English validation set contains 1,000 examples sampled from the nine English datasets listed in Table 2. The optimal setting is $\tau = 0.4$ and $\sigma = 0.7$, which is used in the main experiments. As a post hoc analysis, we additionally evaluated all hyperparameter configurations on the multilingual test sets in Table 2. Although hyperparameters were selected solely using the English validation set, the chosen configuration achieved the best overall performance on 5 of the 8 multilingual test sets. This observation suggests that English validation performance can serve as a useful signal for hyperparameter selection in our setting.

Inference We use vLLM library for efficient inference. The inference is performed with the temperature of 1.0, top-k of 64, top-p of 0.95, and min-p of 0.

B.5 ADDITIONAL RESULTS AND DISCUSSIONS

Statistical Significance of Quality-Difficulty Improvements We conducted paired bootstrap resampling (N=2000) for each task–language pair using the 1x configuration (72 evaluations in total). Quality-Difficulty outperforms Quality-only in 50/72 cases (69.4% win rate), and a binomial sign test shows that this directional advantage is statistically significant ($p < 0.01$), suggesting that the observed trend is unlikely to be explained by random variation alone. Furthermore, the 95% confidence intervals of the performance differences (Quality-Difficulty vs. Quality-only, computed from

[1] unsloth/gemma-3-1b-it-unsloth-bnb-4bit, unsloth/gemma-3-4b-it-unsloth-bnb-4bit, unsloth/gemma-3-12b-it-unsloth-bnb-4bit

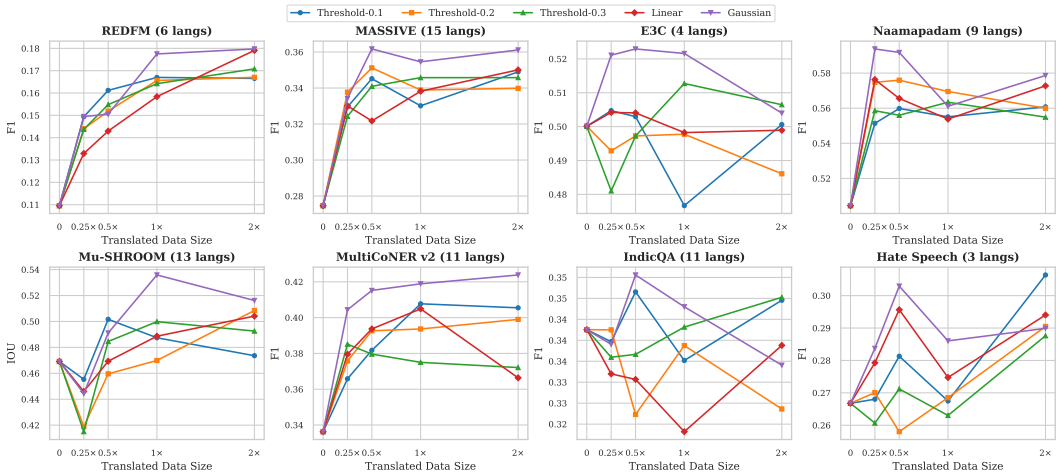


Figure 6: Performance of Gemma-3-4B-IT across benchmarks under different difficulty integration strategies. The x-axis shows the translated data size as a multiple ($\times$) of the English training data size, with the English data size held constant. For each benchmark, results are averaged across all non-English languages. Threshold-0.1/0.2/0.3 denote the Hard-Threshold baselines with $\sigma = 0.1$, 0.2, and 0.3, respectively. Gaussian weighting regularly enables the model to reach competitive peak performance earlier on most benchmarks.

the same bootstrap samples) are strictly positive in 22 cases while only 3 cases are strictly negative. Overall, these results are broadly consistent with a directional advantage for Quality-Difficulty over Quality-only.

Comparison on low-resource languages We investigate whether the Quality-Difficulty selection method provides benefits in low-resource scenarios. To this end, we compare the performance gap between Quality-Difficulty and the Quality-only baseline across all low-resource languages in the multilingual test sets (Table 3). We classify the language as lower, mid or higher resource according to the taxonomy classes introduced by Joshi et al. (2020), where the lower-resource category includes languages from classes 0, 1, and 2.

Figure 5 presents the F1 performance gap between Quality-Difficulty and Quality across the MASSIVE, Naamapadam, and IndicQA datasets under varying translated data sizes. We observe that the benefits of Quality-Difficulty are most pronounced in extreme low-translated-data regimes (0.25$\times$ and 0.5$\times$) on MASSIVE and Naamapadam, suggesting that incorporating difficulty information is particularly beneficial when translated supervision is scarce. In contrast, improvements on IndicQA are less consistent.

Gaussian Weighting vs. Alternative Difficulty Integration Strategies In Eq. 2, we employ a **Gaussian weighting** (*Gaussian*) to incorporate difficulty into the scoring function S. In this section, we investigate alternative formulations for integrating difficulty into the scoring process.

The first baseline replaces the Gaussian weighting with a hard-threshold mechanism that filters out translated instances whose difficulty scores fall outside a window of width 2σ centered at the target difficulty τ. We refer to this approach as **Hard-Threshold** (*Threshold*), which is defined as:

$$S_{\text{hard}}(\bar{x}_i, \bar{y}_i) = q(\bar{x}_i, \bar{y}_i) \times \mathbf{1}_{|d(\bar{x}_i, \bar{y}_i) - \tau| \leq \sigma} \tag{3}$$

where $\mathbf{1}$ denotes the indicator function. In practice, we use the percentile ranks of q and d. We set τ as 0.4, the same as the value used in our main experiment for the Gaussian weighting approach. For σ, we experiment with the values of 0.1, 0.2, or 0.3.

The second baseline replaces the Gaussian-based difficulty weighting with a linear weighting scheme that directly favors more difficult instances. We refer to this approach as **Linear Weighting** (*Linear*), which is defined as:

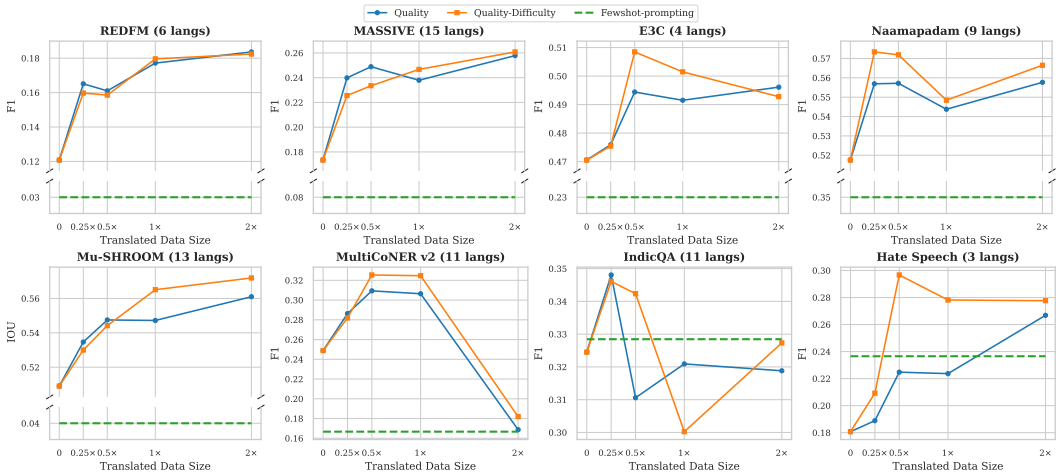


Figure 7: Few-shot prompting and cross-lingual fine-tuning performance of Llama-3.2-3B-Instruct under two data selection approaches across benchmarks. The x-axis shows the translated data size as a multiple (×) of the English training data size, with the English data size held constant. For each benchmark, results are averaged across all non-English languages. Adding translated data improves over English-only fine-tuning and few-shot prompting on most benchmarks, and Quality-Difficulty selection generally matches or outperforms Quality-only selection.

$$S_{\text{linear}}(\bar{x}_i, \bar{y}_i) = q(\bar{x}_i, \bar{y}_i) \times d(\bar{x}_i, \bar{y}_i) \tag{4}$$

We follow the same setup as the experiments in §4 to compare the three weighting functions for data selection. Figure 6 shows that Gaussian weighting enables the model to achieve competitive peak performance across most benchmarks.

Experiments with a model from another family We follow the setup of the data selection experiments in §4 but changing the fine-tuning model to Llama-3.2-3B-Instruct [2] (Grattafiori et al., 2024). Figure 7 shows that fine-tuning on translated data remains helpful: with an appropriate amount of translated data, both fine-tuned models outperform the few-shot prompting baseline on most datasets. Between the two translated data selection methods, Quality-Difficulty reaches better performance than Quality on 5 datasets, while being comparable on the remaining three.

C Details about the Correlation Study

In this section, we provide more details about the experiments conducted in §5.

C.1 Translated Data Construction

To create different translated versions of each English dataset on a target language, we leverage the following LLMs to do the translation: Gemma-3 (4B, 12B, 27B) and Qwen-3 (8B, 14B, 30B-A3B-Instruct-2507). We also employ two label projection methods per model: *Mark-then-translate* (Chen et al., 2023) and *CLaP* (Parekh et al., 2024). This process yields 12 translated datasets for each task-language pair.

C.2 Experimental Setup

For E3C and MASSIVE, we fine-tune Gemma-3-270M-IT with a learning rate of 1e-5, an effective batch size of 16, and a maximum of 500 training steps for E3C and 5,000 for MASSIVE. For

[2] https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct

Table 3: Datasets used in the Instruction Fine-tuning experiments in §4.

Dataset	License	Eval.	Languages (language codes)
Few-NERD (Ding et al., 2021)	CC BY-SA 4.0		English (en)
NCBIDisease (Doğan et al., 2014)	Unknown		English (en)
BC5CDR (Li et al., 2016)	Unknown		English (en)
MultiSpanQA (Li et al., 2022)	CC BY-SA 4.0		English (en)
TBO (Zampieri et al., 2023)	Creative Commons		English (en)
HateXplain (Mathew et al., 2021)	MIT License		English (en)
PsiloQA (Rykov et al., 2025)	CC BY 4.0		English (en)
$\text{RED}^{\text{FM}}_{\text{en}}$ (Huguet Cabot et al., 2023)	CC BY-SA 4.0		English (en)
$\text{MASSIVE}_{\text{en}}$ (FitzGerald et al., 2023)	CC BY 4.0		English (en)
RED^{FM} (Huguet Cabot et al., 2023)	CC BY-SA 4.0	✓	Arabic (ar), Chinese (zh), German (de), French (fr), Italian (it), and Spanish (es)
MASSIVE (FitzGerald et al., 2023)	CC BY 4.0	✓	Amharic (am), Azerbaijani (az), Welsh (cy), Armenian (hy), Icelandic (is), Javanese (jv), Khmer (km), Kannada (kn), Malayalam (ml), Mongolian (mn), Burmese (my), Norwegian (nb), Albanian (sq), Swahili (sw), Telugu (te)
E3C (Magnini et al., 2021)	CC-BY-4.0	✓	Spanish (es), Basque (eu), French (fr), Italian (it)
MultiCoNER-v2 (Fetahu et al., 2023)	CC BY 4.0	✓	Bengali (bn), Chinese (zh), Persian (fa), French (fr), German (de), Hindi (hi), Italian (it), Portuguese (pt), Spanish (es), Swedish (sv), Ukrainian (uk)
Naamapadam (Mhaske et al., 2023)	CC0 1.0 Universal	✓	Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Punjabi (pa), Tamil (ta), Telugu (te)
IndicQA (Doddapaneni et al., 2023)	CC BY 4.0	✓	Assamese (as), Bengali (bn), Gujarati (gu), Hindi (hi), Kannada (kn), Malayalam (ml), Marathi (mr), Odia (or), Punjabi (pa), Tamil (ta), Telugu (te)
Mu-SHROOM (Vazquez et al., 2025)	CC BY 4.0	✓	Arabic (ar), Basque (eu), Catalan (ca), Chinese (zh), Czech (cs), Persian (fa), Finnish (fi), French (fr), German (de), Hindi (hi), Italian (it), Spanish (es), Swedish (sv)
ToxiCN (Bai et al., 2025)	CC BY-NC 4.0	✓	Chinese (zh)
KOLD (Jeong et al., 2022)	Unknown	✓	Korean (ko)
ViHOS (Hoang et al., 2023)	MIT license	✓	Vietnamese (vi)

RED^{FM}, we fine-tune Gemma-3-270M-IT with a learning rate of 3e-5, a maximum of 2000 steps, an effective batch size of 16. All models are fine-tuned on a single A40 GPU with 48GB of memory. The GPU hours for training a model in E3C, MASSIVE and RED^{FM} are 0.1, 1.4, and 1.2, respectively. For each task-language pair, we fine-tune separate models on each of the 12 translated datasets. For each translated dataset, we fine-tune the model 3 times using different random seeds and consider the average as the model's final performance.

D LLM-AS-A-JUDGE STUDY

Prompt The LLM-as-a-judge prompt is illustrated in Fig 18

Human Agreement We consider three language pairs: English-Chinese, English-Vietnamese, and English-Korean. For each pair, we sample 132 English-translation instances from the span-level extraction tasks listed in Table 4. These samples contain 664 projected label spans in total. Three annotators who are native speakers of Chinese, Vietnamese, and Korean evaluate whether each label span is transferred correctly. We then compute Cohen's kappa between the annotations produced by Gemma-3-27B-it and the human annotations. The resulting coefficients for Chinese, Vietnamese, and Korean are 0.54, 0.54, and 0.56, respectively, indicating a moderate level of agreement.

Annotation Platform We developed an annotation platform based on Thresh (Heineman et al., 2023). A screenshot of the platform is shown in Figure 8, and the annotation guidelines are presented in Figure 9. The annotators were university-level students who are native speakers of the annotated

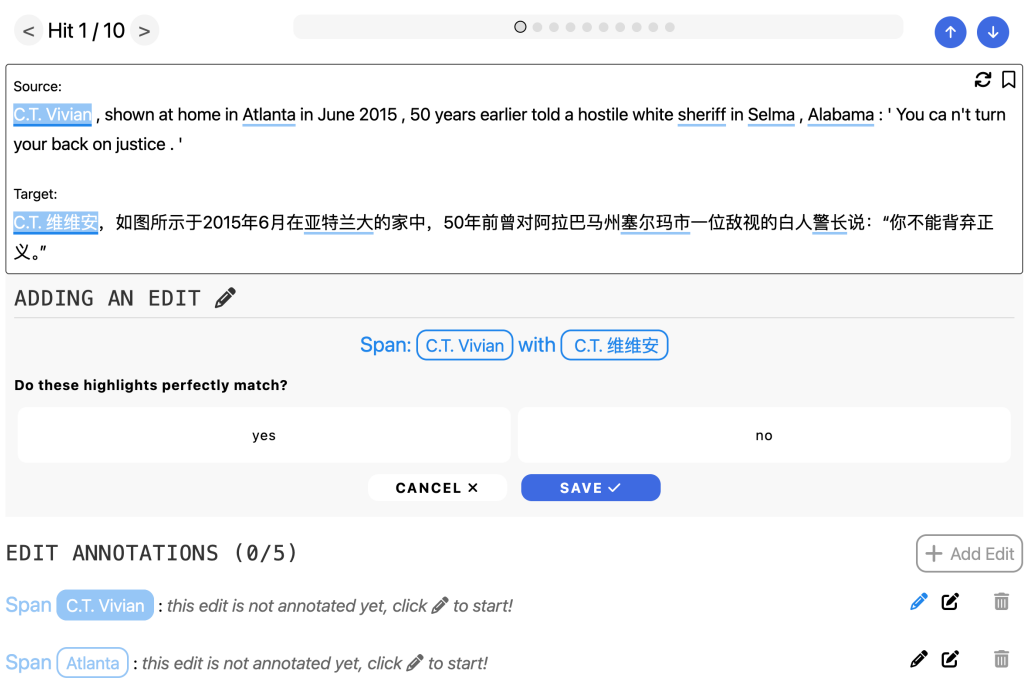


Figure 8: A screenshot of our annotation interface.

```
Input: You will be given an English source text and its translation
     into another language (target text). There are some source-
    target text span pairs highlighted.
(Warning: Some data may contain content that may be offensive or
    upsetting.)

Task: Your task is to verify each pair of span whether it is a
    perfect match. You should check if the boundary of the target
    span is correct:
- Do the source and target spans appear in the same corresponding
    sentence, and in the same corresponding part within the sentence
    ?
- Does the target span include any extra words?
- Does the target span miss any words?
```

Figure 9: Annotation guidelines

languages. The annotators were informed that the collected annotations would be used for research purposes - evaluating the quality of the translated data.

Input Prompt

````
Given a sentence, identify and extract all named entities of the
    following types:
{SCHEMA INFORMATION} // SKIP...

The output should follow the following JSON format:
```
[
   {{
      "text": "ENTITY_TEXT",
      "label": "ENTITY_TYPE"
   }}
]
```
Output an empty list if there is no entities in the provided text.
    Do not include any additional text outside the JSON.

### Input text to analyze:
In February 2008 , Church 's Chicken entered the UK market under
    the `` Texas Chicken `` name , claiming to have signed up 50
    former Dixy Chicken franchisees .
````

Model Output

```
[
   {
      "text": "Church 's Chicken",
      "label": "organization-company"
   },
   {
      "text": "Texas Chicken",
      "label": "organization-company"
   },
   {
      "text": "Dixy Chicken",
      "label": "organization-company"
   }
]
```

Figure 10: Example of the input prompt and the corresponding output of the NER task.

````
Given a sentence, identify and extract all named entities of the
    following types:
{LABEL_SCHEMA}

The output should follow the following JSON format:
```
[
   {{
      "text": "ENTITY_TEXT",
      "label": "ENTITY_TYPE"
   }}
]
```
Output an empty list if there is no entities in the provided text.
    Do not include any additional text outside the JSON.
````

Figure 11: Instruction prompt for the NER task

```
Identify and extract all relations triplets consisting of two
    entities and their relation from the input text. Each triplet
    consists of one subject entity, one object entity and their
    relationship.

Below is the list of relation types to consider:
{RELATION_SCHEMA}

Below is the list of entity types to consider:
{ENTITY_SCHEMA}

The output should be in JSON format as follows:
```
[
   {{
      "subject": {{
         "text": "<text>",
         "label": "<entity_type>"
      }},
      "relation": "<relation_type>",
      "object": {{
         "text": "<text>",
         "label": "<entity_type>"
      }}
   }}
]
```
If no relationship triplets are found, output an empty JSON array:
    []. Do not include any additional text outside the JSON.
```

Figure 12: Instruction prompt for the Relation Extraction task

```
Given a text, identify and extract information that fills specific
    slots defined for the task.

The slot types are:
{LABEL_SCHEMA}

The output should follow the following JSON format:
```
[
   {{
      "text": "SLOT_TEXT",
      "label": "SLOT_TYPE"
   }}
]
```
Output an empty list if there is no slot in the provided text. Do
    not include any additional text outside the JSON.
```

Figure 13: Instruction prompt for the Slot Filling task

```
Extract all spans of text that express hate speech or offensive
    content from the given sentence.

The output should follow the following JSON format:
```
[
   "SPAN_TEXT",
]
```
Output an empty list if there is no hate speech or offensive
    content in the provided text. Do not include any additional text
     outside the JSON.
```

Figure 14: Instruction prompt for the Hate Speech Detection task

```
Given a sentence or short passage, identify all offensive spans (
    arguments), their corresponding targets (if any), and whether
    each argument is harmful to its target.

Note: The relationships between TARGETs and ARGUMENTs can be one-to
    -many, many-to-one, or many-to-many. Each triple should reflect
    a meaningful link between a specific argument and its relevant
    target, along with a judgment of its harmfulness.

The output should follow the following JSON format:
```
[
   {{
      "argument": "ARGUMENT_TEXT",
      "target": "TARGET_TEXT",
      "harmfulness": "HARMFULNESS"
   }}
]
```
Output an empty list if there is no offensive content in the
    provided text. Do not include any additional text outside the
    JSON.
```

Figure 15: Instruction prompt for the Target-based Offensive Language Identification task

```
Answer the question based on the provided context. The answer must
    be extracted from the context. Return a list of one
    discontiguous span that answers the question.

The output should follow the following JSON format:
```
[
   "SPAN_TEXT",
]
```
If the question cannot be answered based on the context, return an
    empty list. Do not include any additional text outside the JSON.
```

Figure 16: Instruction prompt for the Extractive QA task

```
Given a context containing factual information, a question based on
    the context, and an answer generated by a large language model
    (LLM), identify and extract all hallucinated text spans in the
    answer, if any.

The output should follow the following JSON format:
```
[
   "SPAN_TEXT",
]
```
If the answer contains no hallucinations, output an empty list. Do
    not include any additional text outside the JSON.
```

Figure 17: Instruction prompt for the Hallucination Detection task

Table 4: Collection of English Datasets with span-level annotations

Category	Dataset	Domain	Task	License
AI safety	WIKIBIAS (Zhong et al., 2021)	Wikipedia	Subjective Biases Detection	CC BY-SA 3.0, GFDL
	TOXICSPANS (Pavlopoulos et al., 2022)	Social	Toxic spans detection	CC0-1.0 license
	TBO (Zampieri et al., 2023)	Social	Target-based Offensive language identification	CC
	Reddit Self Disclosure (Dou et al., 2024)	Social	Self-Disclosure Detection	CC BY-NC-SA 4.0
	HateXplain (Mathew et al., 2021)	Social	Hate Speech Detection	MIT
	QA-Feeback (Wu et al., 2023)	Open	QA with fine-grained feedback	Apache-2.0
	HADES (Liu et al., 2022)	Wikipedia	Hallucination Detection	MIT
Multimodality	M-HalDetect (Gunjal et al., 2023)	Open	Multimodal Hallucination Detection	CC BY-NC 4.0
	CIN (Goel et al., 2023)	Open	Coreference resolution in image narrations	research purpose
	M2E2 (Li et al., 2020)	News	Multimedia Event Extraction	Unknown
	GRAD-f (Rasheed et al., 2024)	Open	Grounded Conversation Generation	Apache 2.0
	PNG (González et al., 2021)		Panoptic Narrative Grounding	Unknown
	Flickr30k Entities (Plummer et al., 2015)		Grounded Image Captioning	Unknown
Causality/ Temporality	EventStoryLine (Caselli & Vossen, 2017)	News	Event Causal, Temporal RE	CC BY 3.0
	MAVEN-ERE (Wang et al., 2022a)	Wikipedia	Event coreference, temporal, causal, subevent RE	GPL-3.0
	RECESS (Tan et al., 2023)	News	Cause-Effect-Signal Span Detection	CC0-1.0
	SemEval-2020-Task-5 (Yang et al., 2020)	News	Detecting Antecedent and Consequent	CC BY 4.0
Expert Domain	SciER (Zhang et al., 2024a)	Science	NER, RE	GPL-3.0
	FinEntity (Tang et al., 2023)	Financial	Entity-level Sentiment Classification	ODC-By 1.0
	FIRE (Hamad et al., 2024)	Financial	NER, RE	CC-BY-4.0
	MalwareTextDB (Lim et al., 2017)	Cybersecurity	NER	Unknown
	CASIE (Satyapanich et al., 2020)	Cybersecurity	Event Extraction	Unknown
	LegalNER (Kalamkar et al., 2022)	Legal	NER	MIT
	PUBMEDM&M (Bai et al., 2021)	Biomedical	NER, RE	MIT
	CHEMSYN (Bai et al., 2021)	Biomedical	NER, RE	MIT
	NeuroTrialNER (Doneva et al., 2024)	Biomedical	NER	MIT
	ClinicalIE (Agrawal et al., 2022; Moon et al., 2014)			MIT
	- Coreference	Biomedical	Coreference Resolution	-
	- Medication_Status	Biomedical	Medication Status Extraction	-
	- Medication_Attr	Biomedical	Medication Attribute Extraction	-
	EBM-NLP (Nye et al., 2018)	Biomedical	Biomedical Evidence Extraction	Unknown
General domain	Few-NERD (Ding et al., 2021)	Wikipedia	Fine-grained NER	CC BY-SA 4.0
	PolicyIE (Ahmad et al., 2021)	Web	Slot Filling for Privacy Policies	MIT
	MultiSpanQA (Li et al., 2022)	Wikipedia	Multi-Span QA	CC BY-SA 4.0
	BELEBELE-EQA (Lynn et al., 2025)	Wikipedia	Extractive QA	CC BY-SA 4.0
	DocRED (Yao et al., 2019)	Wikipedia	Document-Level RE	MIT
	CURT (Sundriyal et al., 2022)	Social	Claim Span Identification	Unknown
	ElecDeb60to20 (Goffredo et al., 2023)	Politic	Fallacy Detection	Unknown
	Super-Natural Instruction (Wang et al., 2022b)			Apache 2.0
	- Task329	Wikipedia	Coreference resolution	-
	- Task033		Coreference resolution	-
	- Task249		Coreference resolution	-

```
# TASK
Your task is to analyze whether a "Projected Span" in a target language sentence correctly
      and precisely represents a "Source Label" from an English sentence. You are acting
     as a linguistic auditor looking for boundary precision.

## INSTRUCTIONS:
Analyze the relationship between the Source Label and the Projected Span by following
     these steps:

STEP 1: Analysis
1. Exact Translation: Identify the exact translation of the <source language> span in the
     translated text.
2. Comparison: Compare the exact translation to the <target language> projected span.
3. Character/Word Audit: List any extra characters (punctuation, symbols), extra words (
     articles, adjectives), or parenthetical additions.
4. Morphological Check: Is the exact translation fused into a larger word due to target
     language grammar (e.g., suffixes, case markers)?

STEP 2: Categorization
Classify the Projected Span into exactly one of the following categories:
* EXACT_MATCH: The projected span is the same as the exact translation.
* MORPHOLOGICAL_MATCH: The projected span is inside a single word with mandatory
     grammatical affixes/suffixes.
* BILINGUAL_DUPLICATION: The projected span includes both the target translation AND the
     original English term (e.g., "coche (car)"), while the exact translation does not.
* SUPERSET_SYNTACTIC: The projected span includes extra leading or trailing functional
     words (articles like "the/a", prepositions like "of/to").
* SUPERSET_SEMANTIC: The projected span includes extra modifiers or descriptive words (
     adjectives like "red", "big").
* BOUNDARY_NOISE: The projected span includes adjacent punctuation (commas, quotes,
     periods) or formatting marks.
* INCOMPLETE: Missing part of the required translation or cuts off a word.
* MISMATCH: The projected span is semantically unrelated or the wrong word entirely.

*Priority Note:* If multiple errors exist, prioritize word-level errors (SEMANTIC/
     SYNTACTIC/BILINGUAL) over BOUNDARY_NOISE.

## EXAMPLES:

{INCONTEXT-EXAMPLES}

## INPUT DATA:
* <source language> text:
<source text>

* <source language> span: <source span>

* <target language> translation:
<translated text>

* <target language> projected span: <target span>

## OUTPUT FORMAT:
<source language> span: <text>
Analysis: <Find the exact translation, compare it to the projected span, analyze the
     differences (if any), and categorize it into one of the categories above>
Decision: <CATEGORY_NAME>

(Do not add any summary or closing remarks.)
```

Figure 18: LLM-as-a-judge prompt