Estimating Machine Translation Difficulty

Lorenzo Proietti[1,*] **Stefano Perrella**[1,*] **Vilém Zouhar**[2,*]
Roberto Navigli[1,3] **Tom Kocmi**[4]
[1]Sapienza University of Rome [2]ETH Zurich [3]Babelscape [4]Cohere
{lproietti,perrella,navigli}@diag.uniroma1.it vzouhar@ethz.ch kocmi@cohere.com

Abstract

Machine translation quality has steadily improved over the years, achieving near-perfect translations in recent benchmarks. These high-quality outputs make it difficult to distinguish between state-of-the-art models and to identify areas for future improvement. In this context, automatically identifying texts where machine translation systems struggle holds promise for developing more discriminative evaluations and guiding future research.

In this work, we address this gap by formalizing the task of translation difficulty estimation, defining a text's difficulty based on the expected quality of its translations. We introduce a new metric to evaluate difficulty estimators and use it to assess both baselines and novel approaches. Finally, we demonstrate the practical utility of difficulty estimators by using them to construct more challenging benchmarks for machine translation. Our results show that dedicated models outperform both heuristic-based methods and LLM-as-a-judge approaches, with sentinel-src achieving the best performance. Thus, we release two improved models for difficulty estimation, sentinel-src-24 and sentinel-src-25, which can be used to scan large collections of texts and select those most likely to challenge contemporary machine translation systems.

1 Introduction

Not all data samples are equal: Machine learning models may struggle with some samples more than others. The ability to automatically assess sample difficulty is indispensable at various stages of model development. For example, during training, organizing samples from the easiest to the hardest, known as Curriculum Learning, improves both performance and training efficiency (Bengio et al., 2009; Wang et al., 2022; Soviany et al., 2022). Even during inference, computational costs can be reduced by early-exiting on easy examples (Teerapittayanon et al., 2016; Schwartz et al., 2020).

Evaluating models also benefits from estimates of sample difficulty, as too easy or too difficult benchmarks fail to effectively differentiate between models (Lalor et al., 2018; Rodriguez et al., 2021). This issue is particularly relevant in Machine Translation (MT), with recent state-of-the-art models obtaining near-perfect scores and performing close to the human level (Kocmi et al., 2024a; Proietti et al., 2025). With easy test sets, practitioners might struggle to differentiate between top-performing models and assess whether there is headroom for further model improvement. Additionally, while the MT Test Suites subtask of WMT (Kocmi et al., 2024a) targets specific complex translation phenomena, no systematic investigation of the broader concept of general translation difficulty has been carried out.

To address this gap, we explore the notion of sample difficulty in machine translation. First, we propose a definition of translation difficulty and formally introduce *translation difficulty estimation* as a novel task, where the source text's difficulty is automatically predicted. We then present *difficulty estimation correlation* (DEC), a measure designed to evaluate the performance of difficulty estimation methods. Finally, we test baselines and newly proposed approaches to difficulty estimation and validate their practical utility in the downstream task of creating a challenging benchmark, which involves automatically selecting subsets of challenging samples from a large corpus.

We find that approaches such as word rarity, syntactic complexity, or even LLM-as-a-Judge underperform dedicated solutions in capturing translation difficulty. Specifically, we show that sentinel-src – a model trained to predict the expected translation quality of a given text based solely on the source text itself (Perrella et al., 2024) – outper-

[*]Equal contribution.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 24256–24278
November 4-9, 2025

forms other methods at estimating translation difficulty. Therefore, we train two improved versions, called sentinel-src-24 and sentinel-src-25, and release them publicly.[1]

2 Related Work

Previous works can be divided into two categories depending on whether their focus is human or machine translation difficulty.

Human translation difficulty. The earliest works (Fang, 1959; Hale and Campbell, 2002) attempted to connect general text complexity to translation difficulty for humans. A more modern investigation by Mishra et al. (2013) framed human translation difficulty as the time needed to translate a sentence, and estimated it using data on translators' eye movements. They used text length, word polysemy degree, and syntactic complexity as predictors of translation difficulty. Vanroy et al. (2019) examined the correlation between error count, word translation entropy, and syntactic equivalence with translation duration, gaze, and other proxies for human translation difficulty. More recently, Lim et al. (2023, 2024) used word alignment distributions and decoder perplexity to predict human translation difficulty.

Machine translation difficulty. To implement a Curriculum Learning training schedule, Kocmi and Bojar (2017) estimated sample difficulty based on sentence length, word rarity, and the number of coordinating conjunctions in the text. Similarly, Platanios et al. (2019) used sentence length and rarity as proxies for difficulty. Beyond these linguistically-motivated criteria, Zhang et al. (2018) and Liu et al. (2020) predicted translation difficulty using the confidence and other intrinsics of the translation model in generating the text. Almeida (2017) treated difficulty estimation as a binary classification task, although they also used features from the target text, making it closer to quality estimation. Zhan et al. (2021b) used an artificial crowd-based approach that leverages automatic metrics and discovered that long segments, low-frequency words, and proper nouns are the most challenging to machine translate. Finally, Zhan et al. (2021a) estimated a text's difficulty using the embedding similarity between its tokens and those of its translations.

Closer to our work, Don-Yehiya et al. (2022) defined the PreQuEL task as predicting the quality of a given text's translation before the translation is generated. However, they adopted the evaluation of the WMT 2020 Quality Estimation Shared Task (Specia et al., 2020), which was designed for quality estimation rather than for assessing difficulty estimators. Furthermore, their test set included only two language directions, with all translations produced by the same MT model. Additionally, they did not explore the broader space of difficulty estimators or investigate their use in constructing challenging benchmarks.

In contrast, we define translation difficulty estimation as a distinct task with a dedicated evaluation metric. Moreover, we benchmark a wide range of difficulty estimation approaches using test sets that span 11 language directions, with 11 to 19 translations per segment across language pairs, produced by both MT models and human translators. As a result, our work constitutes the first extensive evaluation of translation difficulty estimators, establishing a new state of the art for the task.

3 The Difficulty Estimation Task

The difficulty of translating a given text can depend on multiple factors. A text may be challenging, for example, due to its length, syntactic complexity, idiomatic language, or the presence of rare or specialized vocabulary. Some aspects that affect translation difficulty may even depend on the translation direction, meaning that the same source text might be more difficult to translate into one language than into another. Moreover, translation difficulty might not be uniform across translators, as it can vary with the translator's cultural background and linguistic familiarity – in the case of human translators – or based on factors such as the number of parameters, training data, and model architecture – in the case of machine translation models.

Given these considerations, we avoid defining translation difficulty in absolute terms, as such a definition may not generalize well. Instead, we define difficulty relative to a given target language and to the accuracy of a particular translator, whether human or automatic. More specifically, given a text x, a model[2] m, and a target language l, we assign to x a difficulty score $d_{m,l}(x)$ equal to the quality

[1] Models: hf.co/collections/Prosho/translation-difficulty-estimators-6816665c008e1d22426eb6c4.
Code: github.com/zouharvi/translation-difficulty-estimation.

[2] For brevity, we use "model" to refer both to human translators and automatic models.

score assigned to a translation of x into language l produced by m. Lower scores indicate a lower translation quality and, therefore, greater difficulty associated with the source text.

As an example, suppose we have two texts, x_1 and x_2, and their respective translations t_1 and t_2 into language l, both produced by model m. A human rater evaluates these translations on a scale from 1 to 100, assigning a score of 60 to the first and 90 to the second. Then, $d_{m,l}(x_1) = 60$ and $d_{m,l}(x_2) = 90$. Since $d_{m,l}(x_1) < d_{m,l}(x_2)$, then x_1 is more difficult to translate into l than x_2, for model m. Importantly, the lower the score d, the higher the difficulty and vice versa.

Task Definition. Given a source text x, a model m, and a target language l, Difficulty Estimation is the task of automatically predicting $d_{m,l}(x)$. Different from Quality Estimation, difficulty estimation models do not have access to the translations whose quality is being estimated. Indeed, difficulty estimation can be seen as the task of estimating the *expected* quality of a given text's translation.

Evaluation. We evaluate difficulty estimation methods according to their ability to rank texts based on difficulty. Consider a collection of texts $\mathcal{X} = x_1, x_2, \ldots, x_N$, a collection of target languages $\mathcal{L} = l_1, l_2, \ldots, l_L$, and a collection of models translating into language l: $\mathcal{M}_l = m_1, m_2, \ldots, m_{M_l}$. Let us also define the vector of ground-truth difficulty scores for model m and language l as $D_{m,l} = d_{m,l}(x_1), d_{m,l}(x_2), \ldots, d_{m,l}(x_N)$ and the corresponding predictions of a difficulty estimation method as $\hat{D}_{m,l} = \hat{d}_{m,l}(x_1), \hat{d}_{m,l}(x_2), \ldots, \hat{d}_{m,l}(x_n)$. We measure the translation **Difficulty Estimation Correlation (DEC)** by averaging the Kendall's rank correlation coefficients τ_b across models and languages:

$$\text{DEC} = \frac{1}{|\mathcal{L}|} \sum_{l \in \mathcal{L}} \frac{1}{|\mathcal{M}_l|} \sum_{m \in \mathcal{M}_l} \tau_b\left(\hat{D}_{m,l}, D_{m,l}\right). \quad (1)$$

We refer the reader to Appendix A for further details on how the Kendall correlation coefficient τ_b is computed.

Contrasting DEC with standard MT meta-evaluation strategies. The evaluation approach used in Formula 1, termed "Group-by-System" by Deutsch et al. (2023), makes DEC fundamentally different from other meta-evaluation strategies used in MT evaluation and Quality Estimation, which typically rely on the Group-by-Item method instead (Deutsch et al., 2023). Group-by-Item calculates the correlation between assessments assigned to different translations of the **same source text** (i.e., the same evaluation item), and then averages these correlations across all source texts. The primary benefit of this method is that it mitigates spurious correlations between source text features (e.g., length) and translation quality judgments (Perrella et al., 2024).

However, we argue that these features are precisely what define a text's translation difficulty. Since our goal is to measure translation difficulty, we define DEC using Group-by-System. This method computes the correlation between human and metric assessments for translations of **different source texts** that were produced by the same MT system, and then averages the correlations across MT systems. By holding the MT system constant, this evaluation directly measures a metric's ability to identify which source texts were more challenging for that system to translate.

This distinction is crucial. Indeed, using Group-by-System to meta-evaluate standard MT or quality estimation metrics would favor those that predict source difficulty rather than purely translation quality. Conversely, using Group-by-Item to evaluate difficulty estimators would be inappropriate, as these estimators assign the same score to all translations of a given source.

4 Methods for Difficulty Estimation

In this section, we describe several difficulty estimation methods. We include both common and novel approaches. Specifically, we categorize difficulty estimators into four groups: heuristic-based, learned, LLM-as-a-Judge, and artificial crowd-based. We refer the reader to Appendix C for implementation details regarding all the considered models.

4.1 Heuristic-based estimators

We refer to estimators as heuristic-based if they rely on simple text features. This category includes estimators previously shown to correlate with other measures of difficulty (Mishra et al., 2013; Kocmi and Bojar, 2017; Araghi and Palangkaraya, 2024).

- **Text length** is the number of tokens in a text.
- **Word rarity** is the negative average of the frequencies (estimated from a reference corpus)

of the words in a text.

- **Syntactic complexity** is approximated as the height of the dependency tree associated with a text.

4.2 Learned estimators

Learned machine translation metrics are often trained to predict the quality of a translation given its source text and, optionally, a reference translation (Rei et al., 2020; Guerreiro et al., 2024b; Juraska et al., 2023, 2024). Similarly, neural models can be trained to predict the difficulty of a text. Previous research has explored training similar models for related purposes:

- **PreCOMET** is a suite of source-based regressors based on XLM-RoBERTa (Conneau et al., 2020) that predict the usefulness of a sample for evaluation (Zouhar et al., 2025b). Specifically, $\text{PreCOMET}_{\text{diversity}}$ prioritizes samples likely to elicit diverse machine translation outputs, while $\text{PreCOMET}_{\text{difficulty}}$ estimates difficulty as defined by item response theory (Santor and Ramsay, 1998).

- **sentinel-src** metric is a regression model based on XLM-RoBERTa. Perrella et al. (2024) trained sentinel-src to estimate translation quality from the source alone – i.e., without accessing the candidate translation – with the goal of learning spurious correlations between features of the source texts and translation quality scores.

4.3 LLM-as-a-Judge

LLM-as-a-Judge approaches have seen wide adoption across a range of applications (Zheng et al., 2023; Bavaresco et al., 2024). In this work, we investigate the effectiveness of the LLM-as-a-Judge paradigm for the task of difficulty estimation, using GPT-4o (OpenAI, 2024) and CommandA (Cohere Team, 2025). We prompt these models to determine the proficiency level required to translate a given text, optionally providing information about the target language, and return a scalar score between 0 and 120 indicating the difficulty level of the given text. See the prompts in Example 2.

4.4 Crowd-based Estimators

The methods discussed so far estimate translation difficulty based solely on the source text, and optionally, the target language. However, having defined translation difficulty as the expected quality of a model's translations (Section 3), we now introduce difficulty estimators that more closely mimic this definition.

Artificial Crowd. Artificial crowd-based methods first translate a source text and then use reference-less MT metrics to estimate the quality of the resulting translations.[3] Specifically, we translate the source texts from the test set using a diverse set of models to ensure variety in architecture and size: three instruction-tuned LLMs (Gemma-3-27B-IT, Qwen2.5-72B-IT, CommandA) and one standard encoder-decoder machine translation model (NLLB-moe-54B). For the evaluation step, we employ two state-of-the-art, reference-less MT metrics: XCOMET-QE-XXL (Guerreiro et al., 2024a) and MetricX-24-Hybrid-QE-XXL (Juraska et al., 2024), hereafter referred to as XCOMET and MetricX, respectively. The final difficulty score for each source text is the average quality score assigned to its translations by one of these metrics. This approach is inspired by the artificial crowd methods for efficient subset selection proposed by Zouhar et al. (2025b).

True Crowd. To establish a performance upper bound for Artificial Crowd estimators, we also define True Crowd estimators. Unlike Artificial Crowd, True Crowd estimators use XCOMET and MetricX to score the translations produced by the actual systems whose difficulty we aim to measure – namely, the translations that constitute the WMT24 test set used in our experiments.

Since they rely on the "ground-truth" translations, True Crowd estimators are effectively equivalent to quality estimators. Thus, they are not proper difficulty estimators; we employ them solely to report an upper bound on the performance of Artificial Crowd estimators.

5 Experiments

We benchmark the estimators using the difficulty estimation correlation measure (DEC, Formula 1).

5.1 Experimental Setup

We measure DEC on the test sets released at the WMT 2024 General MT and Metrics shared tasks (Kocmi et al., 2024a; Freitag et al., 2024). These test sets include a selection of source texts translated into multiple languages by automatic models

[3] Reference-less MT metrics estimate the quality of a translation by comparing it only to its source text, without requiring reference translations.

and human translators. Each translation is paired with quality annotations produced by human annotators following either the Error Span Annotation (ESA, Kocmi et al., 2024b) or the Multidimensional Quality Metrics (MQM, Lommel et al., 2014; Freitag et al., 2021) annotation protocols. Here we report results with the ESA annotation protocol. See Appendix Table 6 for results with the MQM protocol, and Appendix Tables 4 and 5 for data statistics.

We test all methods listed in Section 4. Additionally, we improve the top-performing learned estimator, sentinel-src, by expanding the training data used by Perrella et al. (2024) and training two new models, dubbed sentinel-src-24 and sentinel-src-25. The former is trained with data from previous WMT editions up to WMT 2023, while the latter also includes the WMT 24 test set.[4] See Appendix B for further details regarding the training pipeline and parameters of sentinel-src-24 and sentinel-src-25.

5.2 Results

We present the results in Table 1, with methods organized by category as described in Section 4. We also mark each method with the information it uses (i.e., true translations or target language), as detailed in the caption of Table 1, and include three distinct oracles to provide the reader with upper-bound performance values. The definition of oracles can be found in Appendix E.

Heuristic-based and Learned methods. These estimators base their predictions only on the input text. Consequently, the difficulty scores they assign to each text are the same across all target languages and models.[5] Within this group, all learned estimators outperform the heuristic-based ones. Furthermore, sentinel-src-24 achieves the highest difficulty estimation correlation overall, also higher than sentinel-src from Perrella et al. (2024), highlighting the effectiveness of our re-training.

LLM-as-a-Judge. LLM judges are optionally provided with the target language. For both models, the target language information improves performance. This is especially true for CommandA, where the target language information leads to a 0.032 points increase in correlation. However, the overall LLM judges' performance is poor, with scores even lower than the much simpler Text Length heuristics.

	Method	Trans.	Lang.	DEC
Oracle	Oracle	✓	✓	1.000
	Oracle	✗	✓	0.301
	Oracle	✗	✗	0.224
Heuristic	Text Length	✗	✗	0.121
	Syntactic Complexity	✗	✗	0.080
	Word Rarity	✗	✗	−0.040
Learned	sentinel-src-24	✗	✗	0.182
	sentinel-src	✗	✗	0.175
	PreCOMET Difficulty	✗	✗	0.153
	PreCOMET Diversity	✗	✗	0.142
LLM Judge	Command A	✗	✗	0.072
	Command A	✗	✓	0.104
	GPT-4o	✗	✗	0.077
	GPT-4o	✗	✓	0.080
Crowd Based	True (XCOMET)	✓	✓	0.221
	True (MetricX)	✓	✓	0.207
	Artificial (XCOMET)	✗	✓	0.177
	Artificial (MetricX)	✗	✓	0.166
	Random	✗	✗	0.003

Table 1: Difficulty Estimation Correlation (DEC) achieved by each method. We categorize the methods based on the type of information they have access to. Text-only estimators, such as the heuristic and learned ones, rely solely on the source text whose difficulty is being estimated. Instead, some methods also incorporate information of the target language (Lang.) or of the true translation included in the test set (Trans.).

Crowd-based estimators. As expected, the True Crowd methods, which utilize ground-truth translations and thus serve as an upper bound, yield the highest correlation. Since their performance depends solely on the reference-less metrics employed, this result also demonstrates that XCOMET outperforms MetricX on this task by a noticeable margin.

Instead, Artificial Crowd methods' performance is comparable to that of sentinel-src-24. However, Artificial Crowd approaches are considerably more resource-intensive than learned methods, as they require both the translation of source texts and a subsequent quality estimation step.

5.3 Discussion

Through our evaluation, we find that:

- Heuristic-based estimators, commonly used in previous works, are outperformed by most other methods.

[4] For this reason, sentinel-src-25 is not included in the results in Table 1.

[5] I.e., $\forall m_1, m_2 \in \mathcal{M}, l_1, l_2 \in \mathcal{L} : \hat{d}_{m_1,l_1}(x) = \hat{d}_{m_2,l_2}(x)$.

	IOL	GPT-4	Claude3.5	Tower70B
Human	0.137	0.137	0.127	0.109
Tower70B	0.176	0.158	0.151	
Claude3.5	0.178	0.221		
GPT-4	0.202			

Table 2: Average (across language directions) Kendall τ_b correlation matrix for human and four MT models.

- LLM-as-a-Judge approaches are also surpassed by most methods, including the much simpler Text Length heuristic.
- The performance of learned methods – i.e., models explicitly trained to predict text difficulty – is matched only by Artificial Crowd estimators, which are, however, considerably more expensive to operate.
- **sentinel-src-24 sets a state-of-the-art in difficulty estimation**, outperforming all other evaluated estimators.[6]

Based on these results, Section 6 examines the ability of the top estimators from each category – excluding True Crowd estimators, for the reasons discussed in Section 4.4 – on the downstream task of constructing difficult benchmarks.

Furthermore, even if Artificial Crowd estimators are included, using them for benchmark creation implicitly assumes that the generated test data will be used to evaluate models other than those involved in the difficulty estimation. In fact, because Artificial Crowd relies on an intermediate translation step, it would bias the resulting test set against the models used in its construction.

Full results with significance testing – including those restricted to the MQM-annotated portion of WMT24 – are provided in Appendix D.

5.4 Comparing Human and Machine Translation Difficulty

We now examine whether the texts that models find difficult to translate are also challenging for humans. To do this, we use the difficulty scores $d_{m,l}$ assigned to each source text in the WMT 24 test sets, varying m across human translators and MT models. We measure the Kendall's τ_b between the scores of all pairs of translators, averaging

[6]We intentionally exclude True Crowd estimators, which, as discussed in Section 4.4, are not genuine difficulty estimators and instead serve only as upper bounds for Artificial Crowd.

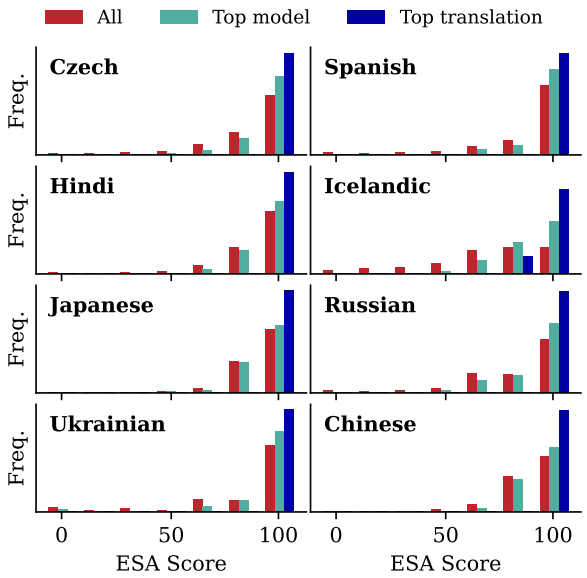

Figure 1: Distribution of human scores assigned to the translations of the texts included in the WMT 2024 test set (Kocmi et al., 2024a). We report the scores of all models (ALL), the scores of the top-performing model for each language (Top model), and the scores of the best translation for each input text (Top translation). The chosen bin width is 15 ESA points.

them across all language directions. For consistency, we restrict the analysis to the models and human translators for which we have annotated translations for all language directions, namely, one human translator and the following four models: Unbabel-Tower70B (Alves et al., 2024), IOL-Research (Zhang, 2024), Claude-3.5, and GPT-4 (OpenAI, 2024).

The results in Table 2 show that the correlations with the human translator are consistently lower (ranging 0.109 to 0.137) than those between machine translation models (ranging 0.151 to 0.221). This suggests that human translators may perceive translation difficulty differently from automatic models. Notably, the highest agreement is observed between GPT-4 and Claude-3.5, which might be due to both models being general-purpose LLMs, unlike Unbabel-Tower70B and IOL-Research, which were explicitly trained for machine translation.

6 Creating Difficult Benchmarks

In this section, we use the top-performing difficulty estimators to create difficult machine translation benchmarks. First, we show that the test set employed at the WMT 2024 General MT shared task (Kocmi et al., 2023) is too easy for current MT

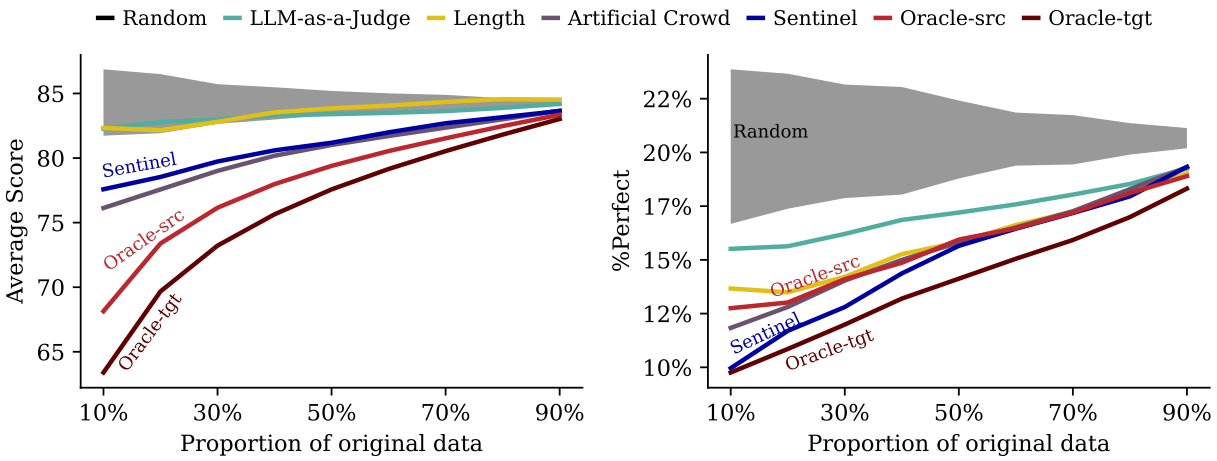

Figure 2: Average score and proportion of "perfect" source texts when creating a difficult test set. Lower values are better. Averaged across all language pairs from WMT24 on which the subset selection is simulated. Random selection shows 99% confidence t-test interval from 10 runs.

models. Then, we define the task of selecting a difficult subset of samples from a given dataset, and evaluate our estimators.

WMT24 is easy. In Figure 1, we report the distribution of scores assigned by human annotators to the translations of the WMT 2024 General MT shared task (Kocmi et al., 2024a). Notably, the best model for each language almost always attains 90 to 100 ESA points, and so does at least one system for each input text. This is particularly concerning for the English-to-Spanish translation direction, where top systems made barely any errors. These findings highlight the need to create more difficult benchmarks. To do this, we use difficulty estimators to sample difficult texts from among a larger collection.

6.1 Setup

Given a large set of source texts $\mathcal{X}$, we aim to extract a subset $\mathcal{X}' \subseteq \mathcal{X}$ of maximum difficulty of size $|\mathcal{X}'| = B$. In practice, we select B texts with the highest difficulty as determined by the top-performing difficulty estimators. As discussed in Section 5.2, we exclude True Crowd estimators.

We again rely on the WMT 24 test set, which we use as $\mathcal{X}$ (see Section 5 for further details). Specifically, we focus on its English source texts, which were translated into Chinese, Czech, Hindi, Icelandic, Japanese, Russian, Spanish, and Ukrainian, as well as its Czech sources translated into Ukrainian. Accordingly, since we select B *source* texts from $\mathcal{X}$, any subset $\mathcal{X}' \subseteq \mathcal{X}$ necessarily contains only English and Czech source texts, while target languages are considered solely for evaluation purposes.

Task definition. We assign a single difficulty score $\hat{d}(x)$ to each sample $x \in \mathcal{X}$. For source text-only difficulty estimators, such as heuristics and learned methods, this is straightforward, as they rely only on the given source text. Instead, for Artificial Crowd methods, we assign to each text x the average quality score of its translations, estimated using XCOMET, averaging across both the MT models employed and the target languages. Finally, we construct $\mathcal{X}'$ by selecting the B most difficult source texts.

Evaluation. One goal of constructing a difficult benchmark is to identify samples where contemporary models still struggle, in order to expose their shortcomings and guide improvements in future iterations. Therefore, we evaluate the usefulness of difficulty estimators based on the drop in the average human score obtained by the models' translations on the test set. As additional information, we also report the proportion of "perfect" outputs (i.e., those that received a full score of 100/100 ESA points from human annotators) that remain in the subsampled test set. The exact formulas for these measures are provided in Appendix C.

6.2 Results

We extract several $\mathcal{X}' \subset \mathcal{X}$ by varying the size of the subsample, and report the curves of the Average Score and %Perfect measures in Figure 2. First, we wish to highlight that the oracles serve as a performance upper bound only in terms of Average

	Source		Diversity		Unique
	length	errors	embd	chrF	outputs
Random	0.00	0.00	0.00	0.00	0.00
LLM-as-a-Judge	−0.61	0.26	0.19	0.23	−0.60
Length	−1.00	0.25	0.31	0.24	−0.52
Artificial Crowd	−0.63	0.04	−0.11	−0.17	−0.46
Sentinel	−0.66	0.12	−0.01	−0.09	−0.36
Oracle-src	−0.22	−0.16	−0.47	−0.49	−0.28
Oracle-tgt	−0.22	−0.16	−0.47	−0.49	−0.28

Table 3: Pearson correlations between difficulty estimators and variables of interest (source length, number of errors per source word, output diversity, and proportion of unique outputs). All estimators assign lower values to more difficult source texts. Therefore, negative correlation indicates a positive correlation between difficulty and the variable of interest. See Appendix Figure 7 for detailed visualization and Appendix C for implementation details.

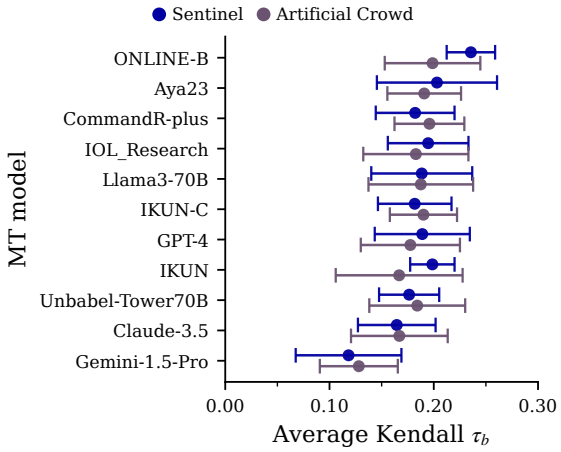

Figure 3: Average Kendall τ_b between difficulty estimators and the human judgments assigned to the MT models in the WMT24 test set. For each MT model and difficulty estimator, circles mark the average correlation across en→X language pairs and bars report ±1 standard deviation.

Score, and not in terms of %Perfect, because they are designed to select the sources with the lowest average score, rather than the lowest %Perfect. In this respect, oracle-src selects the sources with the lowest average difficulty score across models and target languages; instead, oracle-tgt selects a different source text for each target language, averaging difficulty scores only across MT models. As we can see from Figure 2, text length heuristics and LLM-as-a-Judge-based methods show very close performance to random subset selection, especially in terms of Average Score. Instead, sentinel-src-24 and Artificial Crowd perform closer to the oracles, even surpassing oracle-src in terms of %Perfect. In Appendix F, we report detailed quantitative results for the scenario where we subsample 25% of the test set, including per-domain performance breakdowns.

Additionally, returning to the original complaint of existing test sets being too easy (Figure 1), in Appendix Figure 6 we show how the score distribution changes when selecting difficult texts.

6.3 Potential Pitfalls of Selecting by Difficulty

Selecting samples by anything other than random sampling may harbor unexpected dangers. For example, the texts selected by a difficulty estimator might be grammatically incorrect or poorly formed. Here, we investigate potential pitfalls one might encounter when using difficulty as a subsampling criterion. Specifically, we focus on:

- **Source length**: Longer texts are more difficult to translate compared to shorter ones. We are interested in quantifying the extent to which difficulty estimators rely on text length.
- **Source errors**: Translating incomprehensible source texts is naturally difficult. Nevertheless, it might be undesirable to create test sets containing many garbled sources.
- **Output diversity**: When creating a benchmark, source texts that lead to more diverse outputs are more desirable, as they help distinguish between models. See Appendix C for implementation details.

We present the correlation between difficulty estimators' predictions and these variables of interest in Table 3 and Appendix Figure 7. We wish to remind the reader that we defined difficulty using translation quality, meaning that lower estimators' scores indicate higher difficulty. Therefore, a negative correlation with difficulty scores should be interpreted as a strong correlation with the concept of difficulty, as estimated by our models.

As expected, all estimators show a strong negative correlation with source length, indicating that they are all biased toward selecting longer outputs. In contrast, this does not seem to be the case for source errors, suggesting that our difficulty estimators do not prioritize texts containing many errors. Finally, our results suggest that sentinel-src-24 and Artificial Crowd select source texts that lead to more diverse outputs.

Bias toward MT models. A practical concern when creating a test set using difficulty as a selection criterion is the potential bias toward certain MT models. An estimator's difficulty scores may select the texts that some MT models find more challenging than others, thereby biasing the resulting benchmark. To investigate this, we measure the alignment between our two best-performing difficulty estimators – i.e., sentinel-src-24 and Artificial Crowd (XCOMET) – and the concept of translation difficulty, as it was defined in Section 3, for each MT model in isolation. Specifically, for each model and each en→X translation direction, we measure the Kendall's τ_b correlation between the difficulty scores from the estimators and the human quality judgments assigned to that model's translations. Then, we average these correlations across language pairs, reporting the mean and standard deviation in order to capture the variability across translation directions. To ensure that the averages are robust, this analysis only includes MT models that translated the WMT24 source texts into at least five target languages. The results are presented in Figure 3.

Overall, the mean Kendall τ_b correlations fall within a narrow range and have comparable standard deviations. This suggests that the notion of difficulty captured by both estimators aligns relatively uniformly with MT models' perceived difficulty. However, we note two exceptions: ONLINE-B exhibits a higher than average correlation with sentinel-src-24, whereas Gemini-1.5-Pro shows a lower average correlation with both estimators. While this is a preliminary investigation, this finding suggests that a benchmark created using these estimators could be disproportionately difficult for ONLINE-B and easier for Gemini-1.5-Pro. We leave for future work a deeper investigation into whether this discrepancy stems from a bias in the difficulty estimations or simply from variance, especially given that the correlations were averaged over a limited number of translation directions.

6.4 Qualitative Analysis

We manually inspected 200 source texts, half of which were deemed easy and half difficult by sentinel-src-25, and we separated them into 10 length-based buckets. In general, difficulty levels assigned by sentinel-src-25 align well with human perception of difficulty. Indeed, we find that difficult segments often contain complex constructions (Example 1.1), consist of incomplete sentences, such as headlines (Example 1.2), or include indirect speech (Example 1.3). On the other hand, the segments classified as easy by sentinel-src-25 are typically single words, have simple sentence structures, or are concatenations of short, simple sentences (Example 1.4 to Example 1.6).

1 (difficult): City get a nice easy draw at home.

2 (difficult): Alex Bregman Predicted To Betray Astros, Sign With Shocking Blue Jays

3 (difficult): Some folks really do deserve a badge of honour for their pedantry (C8). Veronica Coyne of Springfield claims that "when bemoaning the loss of the express lane at Woolies "12 items or less," a friend told me she'd never used it on principle as it should have been "12 items or fewer.""

4 (easy): Washington

5 (easy): Developing the next generation of hybrid vehicles in Europe

6 (easy): We cannot allow this to happen. This legislation is enormously unpopular. It is exactly what the American people do not want. It must not be passed by Congress.

Example 1: The most difficult and easiest English source texts from the WMT24 dataset, as selected by sentinel-src.

7 Conclusion

In this work, we formally define the task of translation difficulty estimation and introduce the Difficulty Estimation Correlation (DEC), a dedicated measure for evaluating the performance of difficulty estimators. We conduct a comprehensive evaluation of existing and newly proposed estimators, finding that models explicitly trained for the task significantly outperform traditional, heuristic-based methods and LLM-as-a-judge approaches.

Our analysis identifies sentinel-src-24 as the current state-of-the-art in translation difficulty estimation. We further validate the performance of difficulty estimators in the downstream task of creating difficult benchmarks, demonstrating that they successfully identify samples where modern MT models underperform. In this downstream task too, sentinel-src-24 remains the top-performing method. Building on these findings, we develop sentinel-src-25 by incorporating additional data into the training pipeline of sentinel-src-24, and release both models publicly. Finally, we conduct a qualitative analysis of sentinel-src-25's predictions, offering intuitive insights into the types of texts it deems difficult.

Limitations

The concept of translation difficulty. This work is based on the assumption that we can proxy the difficulty of a given text using the quality of the translations it produces. While we acknowledge that translation difficulty should ideally be an intrinsic property of the source – independently of any specific translation model – this working assumption serves our purposes, particularly for the downstream task of creating challenging machine translation benchmarks. Indeed, our research objective is to identify texts that are difficult for contemporary MT models to translate, rather than to explore the abstract, model-independent notion of translation difficulty.

Impact of the target language on translation difficulty. As discussed in Section 3, the difficulty of translating a given text may depend on the target language, as corroborated theoretically by Bugliarello et al. (2020). We acknowledge that this aspect is only mentioned briefly herein and that we do not provide an investigation into this phenomenon. Nonetheless, our experiments support this hypothesis: the performance of the LLM-as-a-Judge improves when the model is given information about the target language. We therefore encourage future research to explore the influence of the target language on translation difficulty more thoroughly and to investigate how this information might be incorporated into other difficulty estimation methods effectively.

Using the WMT 2024 test set to analyze difficulty estimators. In Section 6.3, we investigated potential concerns of subsampling large data sets using our difficulty estimators. However, to do this, we used the WMT 2024 test set, which has a limitation. The sources contained in this test set were vetted by humans, making the distribution of the phenomena we investigate artificial. To mitigate this issue, we conduct the same analysis using a larger batch of data and report results in Appendix Figure 5.

Ethics Statement

We foresee no ethical issues with our work.

Acknowledgements

The authors gratefully acknowledge the support of the PNRR MUR project PE0000013-FAIR, and the CREATIVE project (CRoss-modal understanding and gEnerATIon of Visual and tExtual content), which is funded by the MUR Progetti di Rilevante Interesse Nazionale programme (PRIN 2020). The authors acknowledge the CINECA award IsCb9_mtmit under the ISCRA initiative for the availability of high-performance computing resources.

This work was carried out while Lorenzo Proietti was enrolled in the Italian National Doctorate on Artificial Intelligence run by Sapienza University of Rome.

References

Ana Sofia Vieira de Jesus Almeida. 2017. Difficulty estimation of machine translation. MSc Thesis.

Duarte Miguel Alves, José Pombal, Nuno M Guerreiro, Pedro Henrique Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, Pierre Colombo, José G. C. de Souza, and Andre Martins. 2024. Tower: An open multilingual large language model for translation-related tasks. In *First Conference on Language Modeling*.

Sahar Araghi and Alfons Palangkaraya. 2024. The link between translation difficulty and the quality of machine translation: A literature review and empirical investigation. *Language Resources and Evaluation*, 58(4):1093–1114.

Marta Bañón, Pinzhen Chen, Barry Haddow, Kenneth Heafield, Hieu Hoang, Miquel Esplà-Gomis, Mikel L. Forcada, Amir Kamran, Faheem Kirefu, Philipp Koehn, Sergio Ortiz Rojas, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Elsa Sarrías, Marek Strelec, Brian Thompson, William Waites, Dion Wiggins, and Jaume Zaragoza. 2020. ParaCrawl: Web-scale acquisition of parallel corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4555–4567. Association for Computational Linguistics.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K Surikuchi, Ece Takmaz, and Alberto Testoni. 2024. LLMs instead of human judges? a

large scale empirical study across 20 NLP evaluation tasks. *Preprint*, arXiv:2406.18403.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Emanuele Bugliarello, Sabrina J. Mielke, Antonios Anastasopoulos, Ryan Cotterell, and Naoaki Okazaki. 2020. It's easier to translate out of English than into it: Measuring neural translation difficulty by cross-mutual information. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1640–1649. Association for Computational Linguistics.

Cohere Team. 2025. Command a: An enterprise-ready large language model. *Preprint*, arXiv:2504.00698.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451. Association for Computational Linguistics.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2021. A statistical analysis of summarization evaluation metrics using resampling methods. *Transactions of the Association for Computational Linguistics*, 9:1132–1146.

Daniel Deutsch, George Foster, and Markus Freitag. 2023. Ties matter: Meta-evaluating modern metrics with pairwise accuracy and tie calibration. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12914–12929. Association for Computational Linguistics.

Shachar Don-Yehiya, Leshem Choshen, and Omri Abend. 2022. PreQuEL: Quality estimation of machine translation outputs in advance. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11170–11183. Association for Computational Linguistics.

Achilles Fang. 1959. *Some Reflections on the Difficulty of Translation*, pages 111–134. Harvard University Press, Cambridge, MA and London, England.

Marina Fomicheva, Shuo Sun, Erick Fonseca, Chrysoula Zerva, Frédéric Blain, Vishrav Chaudhary, Francisco Guzmán, Nina Lopatina, Lucia Specia, and André F. T. Martins. 2022. MLQE-PE: A multilingual quality estimation and post-editing dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4963–4974. European Language Resources Association.

Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. Association for Computational Linguistics.

Markus Freitag, Nitika Mathur, Chi-kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Tom Kocmi, Frederic Blain, Daniel Deutsch, Craig Stewart, Chrysoula Zerva, Sheila Castilho, Alon Lavie, and George Foster. 2023. Results of WMT23 metrics shared task: Metrics might be guilty but references are not innocent. In *Proceedings of the Eighth Conference on Machine Translation*, pages 578–628. Association for Computational Linguistics.

Gemma, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41. Association for Computational Linguistics.

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

Nuno M Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André FT Martins. 2024b. xcomet: Transparent machine translation evaluation through fine-grained error detection. *Transactions of the Association for Computational Linguistics*, 12:979–995.

Sandra Hale and Stuart Campbell. 2002. The interaction between text difficulty and translation accuracy. *Babel*, 48:14–33.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. Association for Computational Linguistics.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. 2023. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In *Proceedings of the Eighth Conference on Machine Translation*, pages 756–767. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, and 3 others. 2024a. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1–46. Association for Computational Linguistics.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Philipp Koehn, Benjamin Marie, Christof Monz, Makoto Morishita, Kenton Murray, Masaaki Nagata, Toshiaki Nakazawa, Martin Popel, and 3 others. 2023. Findings of the 2023 conference on machine translation (WMT23): LLMs are here but not quite there yet. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1–42. Association for Computational Linguistics.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45. Association for Computational Linguistics.

Tom Kocmi and Ondřej Bojar. 2017. Curriculum learning and minibatch bucketing in neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 379–386. INCOMA Ltd.

Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024b. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453. Association for Computational Linguistics.

John P. Lalor, Hao Wu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Understanding deep learning performance through an examination of test set difficulty: A psychometric case study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4711–4716. Association for Computational Linguistics.

Zheng Wei Lim, Trevor Cohn, Charles Kemp, and Ekaterina Vylomova. 2023. Predicting human translation difficulty using automatic word alignment. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11590–11601. Association for Computational Linguistics.

Zheng Wei Lim, Ekaterina Vylomova, Charles Kemp, and Trevor Cohn. 2024. Predicting human translation difficulty with neural machine translation. *Transactions of the Association for Computational Linguistics*, 12:1479–1496.

Xuebo Liu, Houtim Lai, Derek F. Wong, and Lidia S. Chao. 2020. Norm-based curriculum learning for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 427–436. Association for Computational Linguistics.

Arle Lommel, Hans Uszkoreit, and Aljoscha Burchardt. 2014. Multidimensional quality metrics (MQM): A framework for declaring and describing translation quality metrics. *Tradumàtica*, 0(12):0455–463.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR.

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997. Association for Computational Linguistics.

Abhijit Mishra, Pushpak Bhattacharyya, and Michael Carl. 2013. Automatically predicting sentence translation difficulty. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 346–351. Association for Computational Linguistics.

Jinjie Ni, Yifan Song, Deepanway Ghosal, Bo Li, David Junhao Zhang, Xiang Yue, Fuzhao Xue, Zian Zheng, Kaichen Zhang, Mahir Shah, Kabir Jain, Yang You, and Michael Shieh. 2024a. MixEval-X: Any-to-Any evaluations from real-world data mixtures. *Preprint*, arXiv:2410.13754.

Jinjie Ni, Fuzhao Xue, Xiang Yue, Yuntian Deng, Mahir Shah, Kabir Jain, Graham Neubig, and Yang You. 2024b. MixEval: Deriving wisdom of the crowd from LLM benchmark mixtures. *Preprint*, arXiv:2406.06565.

NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, and 20 others. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

OpenAI. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774.

Stefano Perrella, Lorenzo Proietti, Alessandro Scirè, Edoardo Barba, and Roberto Navigli. 2024. Guardians of the machine translation meta-evaluation: Sentinel metrics fall in! In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16216–16244. Association for Computational Linguistics.

Emmanouil Antonios Platanios, Otilia Stretcu, Graham Neubig, Barnabas Poczos, and Tom Mitchell. 2019. Competence-based curriculum learning for neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1162–1172. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, and Roberto Navigli. 2025. Has machine translation evaluation achieved human parity? the human reference and the limits of progress. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 790–813. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Parker Riley, Daniel Deutsch, George Foster, Viresh Ratnakar, Ali Dabirmoghaddam, and Markus Freitag. 2024. Finding replicable human evaluations via stable ranking probability. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4908–4919. Association for Computational Linguistics.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. 2021. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *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)*, pages 4486–4503. Association for Computational Linguistics.

Jie Ruan, Xiao Pu, Mingqi Gao, Xiaojun Wan, and Yuesheng Zhu. 2024. Better than random: Reliable NLG human evaluation with constrained active sampling. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18915–18923.

Darcy A Santor and James O Ramsay. 1998. Progress in the technology of measurement: Applications of item response models. *Psychological assessment*, 10(4):345.

Roy Schwartz, Gabriel Stanovsky, Swabha Swayamdipta, Jesse Dodge, and Noah A. Smith. 2020. The right tool for the job: Matching model and instance complexities. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6640–6651. Association for Computational Linguistics.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *Int. J. Comput. Vision*, 130(6):1526–1565.

Lucia Specia, Zhenhao Li, Juan Pino, Vishrav Chaudhary, Francisco Guzmán, Graham Neubig, Nadir Durrani, Yonatan Belinkov, Philipp Koehn, Hassan Sajjad, Paul Michel, and Xian Li. 2020. Findings of the WMT 2020 shared task on machine translation robustness. In *Proceedings of the Fifth Conference on Machine Translation*, pages 76–91. Association for Computational Linguistics.

Robyn Speer. 2022. rspeer/wordfreq: V3.0.

Surat Teerapittayanon, Bradley McDanel, and H.T. Kung. 2016. Branchynet: Fast inference via early exiting from deep neural networks. In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pages 2464–2469.

Bram Vanroy, Orphee de clercq, and Lieve Macken. 2019. Correlating process and product data to get an insight into translation difficulty. *Perspectives*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2022. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Guillaume Wenzek, Vishrav Chaudhary, Angela Fan, Sahir Gomez, Naman Goyal, Somya Jain, Douwe Kiela, Tristan Thrush, and Francisco Guzmán. 2021. Findings of the WMT 2021 shared task on large-scale multilingual machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 89–99. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021a. Difficulty-aware machine translation evaluation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 26–32. Association for Computational Linguistics.

Runzhe Zhan, Xuebo Liu, Derek F. Wong, and Lidia S. Chao. 2021b. Variance-aware machine translation test sets. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Wenbo Zhang. 2024. IOL research machine translation systems for WMT24 general machine translation shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 147–154. Association for Computational Linguistics.

Xuan Zhang, Gaurav Kumar, Huda Khayrallah, Kenton Murray, Jeremy Gwinnup, Marianna J Martindale, Paul McNamee, Kevin Duh, and Marine Carpuat. 2018. An empirical exploration of curriculum learning for neural machine translation. *Preprint*, arXiv:1811.00739.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Vilém Zouhar, Tom Kocmi, and Mrinmaya Sachan. 2025a. AI-assisted human evaluation of machine translation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4936–4950. Association for Computational Linguistics.

Vilém Zouhar, Peng Cui, and Mrinmaya Sachan. 2025b. How to select datapoints for efficient human evaluation of NLG models? *Preprint*, arXiv:2501.18251.

	EN→DE	EN→ES	JA→ZH
#Source texts	486	622	559
#Translators	19	15	15

Table 4: Statistics of the test set released at the WMT 2024 Metrics Shared Task (Freitag et al., 2024). "#Source texts" indicates the number of source texts in the test set, and "#Translators" indicates the number of available translations for each source text.

A Kendall τ_b

Kendall's τ variant b is defined as:

$$\tau_b = \frac{C - D}{\sqrt{(C + D + T_g)(C + D + T_h)}}. \quad (2)$$

Here C and D are the numbers of *concordant* and *discordant* pairs when comparing gold scores r_i with hypothesis scores $\hat{r}_i$: a pair (i, j) is concordant if $(r_i - r_j)(\hat{r}_i - \hat{r}_j) > 0$ and discordant if $(r_i - r_j)(\hat{r}_i - \hat{r}_j) < 0$. T_h counts pairs tied *only* in the hypothesis ($\hat{r}_i = \hat{r}_j$ and $r_i \neq r_j$), and T_g counts pairs tied *only* in the gold ($r_i = r_j$ and $\hat{r}_i \neq \hat{r}_j$); pairs tied in both rankings are ignored.

We use Kendall's τ_b rather than Pearson correlation because τ_b evaluates *relative order* (ranks) and is therefore more robust to scale differences and outliers. This is important in our setting: as visible in Figure 1, most WMT24 segment-level scores lie in a narrow band (e.g., $[90, 100]$), which magnifies the effect of outliers on Pearson, whereas τ_b depends only on pairwise rankings. This behavior is also discussed by Mathur et al. (2020).

B Training sentinel-src-24 and sentinel-src-25

Our new learned difficulty estimation models, sentinel-src-24 and sentinel-src-25, follow the same architecture and training pipeline used for the sentinel-src model introduced by Perrella et al. (2024). Both models are based on XLM-RoBERTa large as the backbone encoder, followed by a multi-layer feedforward network on top of the `[CLS]` token. They are trained to minimize the Mean Squared Error (MSE) between predicted and human scalar scores.

We adopt the same two-stage training approach as the sentinel-src model. In the first stage, the model is trained on Direct Assessment (DA, Graham et al., 2013) data. In the second stage, it is fine-tuned on MQM annotations. The key differences between our models and sentinel-src lie in the training data used at each stage.

- **Stage 1: DA training.** For sentinel-src-24, we extend the DA training data used by Perrella et al. (2024) by including annotations from WMT 21 (Wenzek et al., 2021), as well as DA+SQM annotations from WMT 22 (Kocmi et al., 2022) and WMT 23 (Kocmi et al., 2023). sentinel-src-25 further includes the ESA annotations from WMT 24. For both model versions, we also incorporate MLQE-PE data (Fomicheva et al., 2022) in the training set for this stage.

- **Stage 2: MQM fine-tuning.** In this phase, we expand the MQM training set by adding MQM annotations from WMT 23 (Freitag et al., 2023). Unlike the sentinel-src training pipeline, we do not average multiple scores per translation. Instead, we include all available annotations as individual training instances, preserving variability across raters. This applies to WMT 20 and WMT 22 MQM datasets, which include three human scores per translation (Freitag et al., 2021; Riley et al., 2024). Similarly to the first training stage on DA, in the case of the sentinel-src-25 model, we also include MQM annotations from WMT 24.

Following the approach of Perrella et al. (2024), we treat each pair consisting of a source text segment and its associated human score as an independent training instance. Since human scores are assigned to individual translations, multiple annotations may exist for the same source text. We do not combine these scores in any way but include them all in the training data for both DA and MQM stages. Training hyperparameters match those used by Perrella et al. (2024) for sentinel-src. All models are trained using a single NVIDIA GeForce RTX 4090 GPU. The estimated training time is approximately three GPU hours for the first (DA) stage and one GPU hour for the second (MQM) fine-tuning stage. These estimates apply to both sentinel-src-24 and sentinel-src-25.

C Implementation Details

- For the word rarity heuristic, we compute word frequencies using the `wordfreq` Python library (Speer, 2022).

- For the syntactic complexity heuristic and text length, we obtain dependency trees and

	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH	EN→CS	CS→UK
#Source texts	634	634	634	634	634	634	634	634	1954
#Translators	14	11	11	13	14	11	13	16	12

Table 5: Statistics of the test set released at the WMT 2024 General Machine Translation Shared Task (Kocmi et al., 2024a). "#Source texts" indicates the number of source texts in the test set, and "#Translators" indicates the number of available translations for each source text.

corresponding tokens using `spaCy`. Specifically, we use language-specific pipelines: i) `en_core_web_sm` for English, ii) `ja_core_news_sm` for Japanese, and iii) `spacy_udpipe` for Czech.

- For the output diversity assessment (Section 6.3), we compute multilingual sentence embeddings from the source texts using sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2 (Reimers and Gurevych, 2019). Specifically, for each pair of translations, we measure the inner product between their multilingual sentence embeddings and their chrF score computed one against the other.

For the artificial crowd, we use the following:

- **NLLB-moe-54B**: sparsely-gated mixture-of-experts encoder-decoder translation model (NLLB et al., 2022).
- **Gemma-3-27B-IT**: multimodal instruction-tuned LLM from the Gemma family (Gemma et al., 2025).
- **Qwen2.5-72B-IT**: largest instruction-tuned LLM from the Qwen2.5 family (Qwen et al., 2025).
- **CommandA**: 111B-parameter LLM for real-world enterprise use cases (Cohere Team, 2025).

For evaluation of Section 6 we use average model score and %Perfect. For average model score, given $\mathcal{X}' \subseteq \mathcal{X}$ and a set of models $\mathcal{M}_l$, we report for each subset $\mathcal{X}' \subseteq \mathcal{X}$ with $|\mathcal{X}'| = B$:

$$\text{AvgScore} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} d_{m,l}(x)\,, \quad (3)$$

which is the average human score on the subset. For proportion of perfect translations, we use:

$$\%\text{Perfect} = \frac{1}{B \cdot |\mathcal{M}_l|} \sum_{\substack{x \in \mathcal{X}' \\ m \in \mathcal{M}_l}} \mathbb{1}[d_{m,l}(x) = 100\%] \quad (4)$$

D Complete Results

Table 6 presents the difficulty estimation correlation scores of all considered methods when the ground truth is based on MQM annotations, rather than ESA.

Instead, Tables 7 and 8 present the per-language breakdown of all methods' difficulty estimation correlation scores on the ESA-annotated and MQM-annotated WMT24 test data, respectively. These tables also include ranks derived from statistical significance analysis. Specifically, we used the PERM-BOTH hypothesis test, introduced by Deutsch et al. (2021).

E Oracles

Oracle methods adopt the true human judgments used to derive difficulty scores, as detailed in Section 3. We consider three oracles that differ in the type of information they have access to:

- **Oracle (source text + target language + target translation)** assigns to each source text x the true $d_{m,l}(x)$, for each m and l.
- **Oracle (source text + target language)** estimates the difficulty of x by averaging the true $d_{m,l}(x)$ across all models ($\forall m \in \mathcal{M}_l$), meaning that its estimates do not vary across models, but only across target languages.
- **Oracle (source text only)** averages the true $d_{m,l}(x)$ across both models and target languages, assigning the same score to each source text regardless of target language or translator.

F Creating Difficult Benchmarks – Quantitative Results

To quantitatively evaluate the effectiveness of our difficulty estimators for constructing challenging benchmarks, we simulate a 25% budget scenario. That is, for each method, we select the 25% most difficult source texts from the WMT 24 test sets and

	Method	System	Lang	DEC
Oracle	Oracle	✓	✓	1.000
	Oracle (source text + target lang)	✗	✓	0.430
	Oracle (source text only)	✗	✗	0.404
Heuristic	Text Length	✗	✗	0.222
	Syntactic Complexity	✗	✗	0.170
	Word Rarity	✗	✗	−0.052
Learned	sentinel-src-24	✗	✗	0.246
	sentinel-src	✗	✗	0.235
	PreCOMET Difficulty	✗	✗	0.169
	PreCOMET Diversity	✗	✗	0.167
LLM Judge	Command A (source text only)	✗	✗	0.114
	Command A (source text + target lang)	✗	✓	0.120
	GPT-4o (source text only)	✗	✗	0.090
	GPT-4o (source text + target lang)	✗	✓	0.090
Crowd Based	True (XCOMET-QE-XXL)	✓	✓	**0.278**
	True (MetricX-24-Hybrid-QE-XXL)	✓	✓	0.248
	Artificial (XCOMET-QE-XXL)	✗	✓	0.207
	Artificial (MetricX-24-Hybrid-QE-XXL)	✗	✓	0.185
	Random	✓	✓	0.002

Table 6: Difficulty Estimation Correlation (DEC) achieved by each method on the MQM-annotated WMT24. We categorize the methods based on the type of information they have access to. Text-only estimators, such as the heuristic and learned ones, rely solely on the source text whose difficulty is being estimated. Instead, some methods also incorporate information on the target language of translation, while others further leverage knowledge of the specific translator who produced the translations in the test set.

assess the resulting subset using human annotations of translation quality.

Table 9 and Table 10 report the results of this evaluation for the ESA and MQM human annotation protocols, respectively, averaged across all language directions in the corresponding test sets. We consider two quantitative indicators: (1) AvgScore, the average human score assigned to the selected subset (lower indicates higher difficulty), and (2) %Perfect, the proportion of model outputs in the selected subset that receive a perfect human score (lower is also better).

Results confirm the strong performance of our dedicated difficulty estimation model, sentinel-src-24, which achieves substantially lower AvgScore and %Perfect values than random selection. It also achieves the best results among all automatic methods that rely solely on the source text. In particular, in Table 9, it is outperformed in AvgScore only by Artificial Crowd (XCOMET), a more computationally intensive approach that requires translating each source text with multiple large models and evaluating those translations using an XXL MT metric. Furthermore, Artificial Crowd methods can produce difficulty scores conditioned on the target language, unlike sentinel-src-24, which relies exclusively on the source text. On the other hand, sentinel-src-24 obtains the best %Perfect score in Table 9. In Table 10, sentinel-src-24 outperforms all automatic methods in both AvgScore and %Perfect, including Artificial Crowd.

As for the other automatic methods, the Text Length heuristic consistently outperforms LLM-as-a-Judge (based on Command A), despite the latter requiring significantly more computational resources. Notably, in both Table 9 and Table 10, Command A only marginally improves over random selection, reinforcing the limitations of LLM-as-a-Judge methods already observed in Table 1.

Tables 11 and 12 provide a fine-grained breakdown of results across the WMT 24 domains (News, Social, Literary, and Speech) for the ESA and MQM test sets, respectively. These results show that the overall patterns hold consistently across domains. While absolute performance varies, sentinel-src-24 and Artificial Crowd achieve the strongest results in nearly all domain-specific evaluations.

This analysis supports the practical utility of difficulty estimation for controlled test set construction and confirms that learned estimators such as sentinel-src-24 offer effective and reliable means for identifying source segments where MT systems

are more likely to struggle.

G Related work for Benchmark Creation

We extend the related work in Section 2 by discussion on previous attempts to automatically create challenging subsets.

Maia Polo et al. (2024, tinyBenchmarks) and Rodriguez et al. (2021) make heavy use of Item Response Theory (Santor and Ramsay, 1998), which is a set of statistical models for educational testing of human subjects. However, this is not applicable to machine translation, where the quality of the output is represented as a continuous score. Other works (Ni et al., 2024b,a; Ruan et al., 2024; Zouhar et al., 2025b) attempt to be more broadly applicable to natural language generation tasks, though their optimization goals are usually efficient testing (i.e. obtaining the same model ranking with fewer evaluated examples) rather than creating difficult testsets.

For machine translation specifically, Zhan et al. (2021a) use proxy of machine translation difficulty to inform better evaluation. Again, Zouhar et al. (2025a) automatically remove examples that are too easy from the evaluation set, corresponding to our True Crowd with quality estimation.

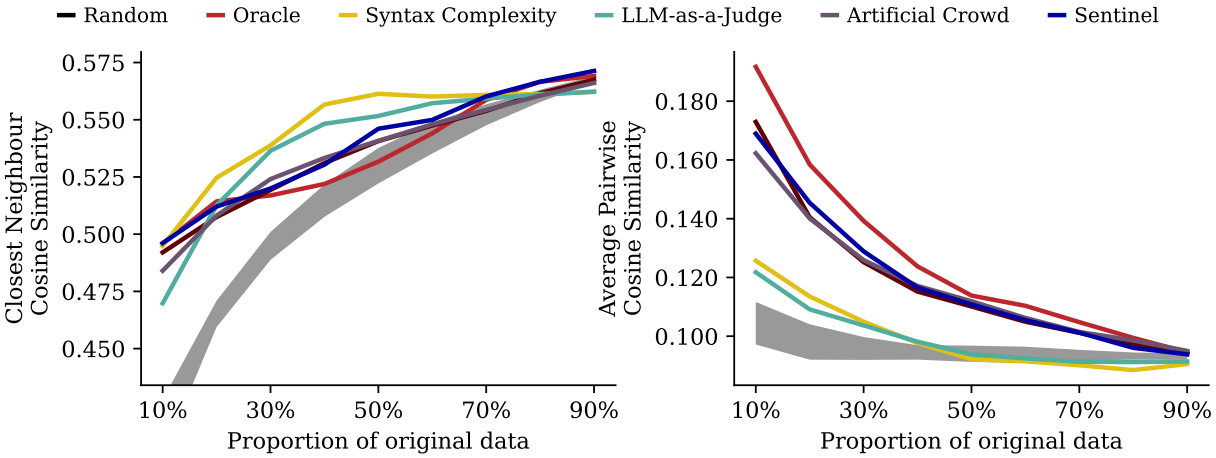

Figure 4: Average similarity between two closest (left) and any two (right) source texts in $\mathcal{X}$ based on embeddings and cosine similarity. The left curves go up because the vector space saturates and nearest neighbours become closer. Random selection shows 99% confidence t-test interval from 10 runs.

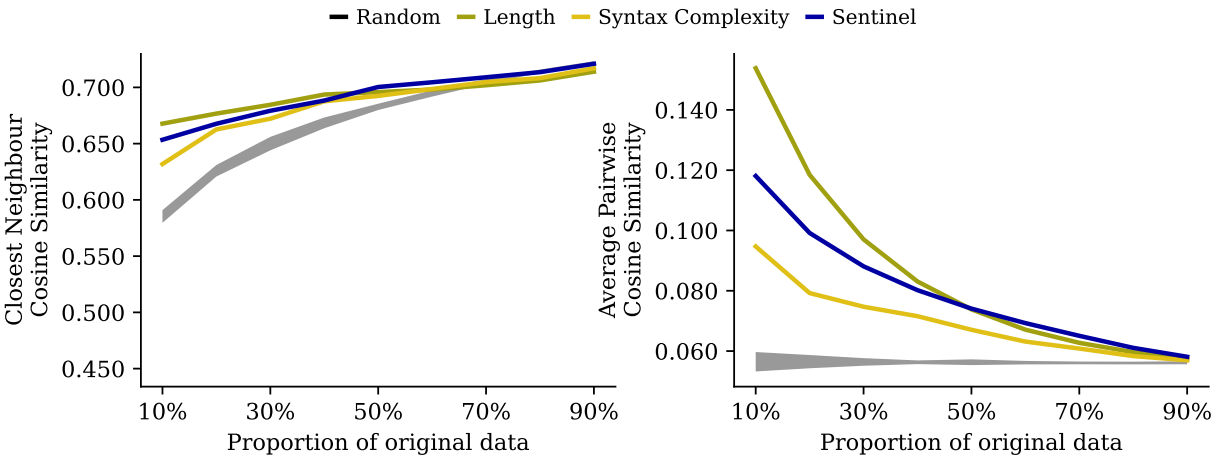

Figure 5: In contrast to Figure 4, we have collected contemporary news articles (40k segments through crawling, Bañón et al., 2020) to evaluate how our difficulty sampling would perform in a real world. Average similarity between two closest (left) and any two (right) source texts based on embeddings and cosine similarity in $\mathcal{X}$ on raw 40k English segments (not WMT24). The left curves go up because the vector space saturates and nearest neighbours become closer. Random selection shows 99% confidence t-test interval from 10 runs. LLM-as-a-Judge and Artificial crowd were not included due to compute costs. Oracle is not present due to the absence of model outputs and human scores.

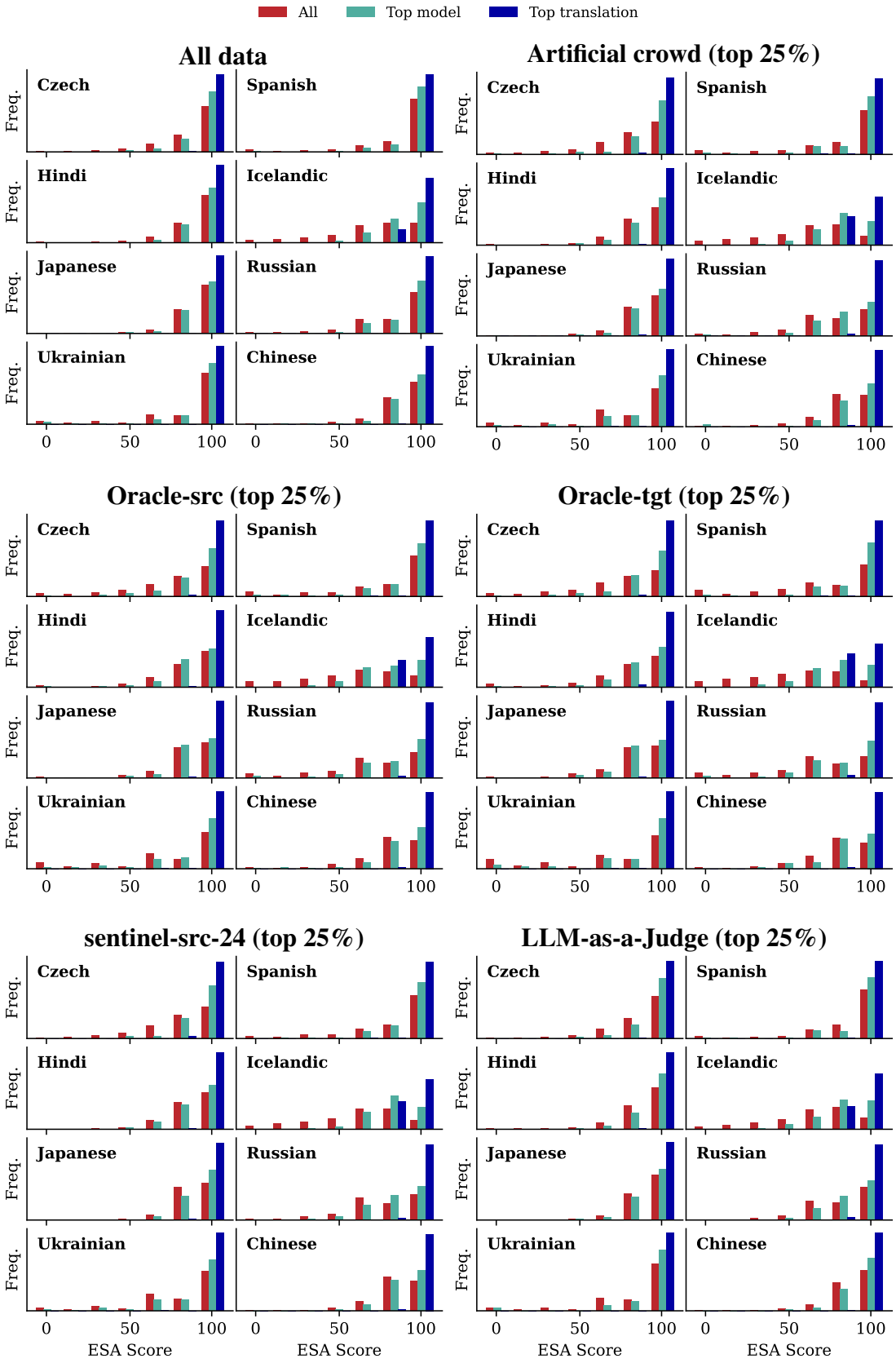

Figure 6: Distribution of human scores for machine translation models in WMT 2024 (Kocmi et al., 2024a) of all models, top model in each language, and top model for each input segment. Subset selection methods select top 25% most difficult segments. Extends Figure 1.

	Average		CS→UK	EN→CS	EN→ES	EN→HI	EN→IS	EN→JA	EN→RU	EN→UK	EN→ZH
	Rank	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC	DEC
Oracle	1	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Oracle (source text + target lang)	2	0.301	0.298	0.303	0.280	0.271	0.381	0.252	0.320	0.305	0.302
Oracle (source text only)	3	0.224	0.298	0.246	0.201	0.176	0.238	0.172	0.253	0.213	0.218
True Crowd (XCOMET-QE-XXL)	3	0.221	0.203	0.271	0.194	0.213	0.273	0.173	0.238	0.195	0.233
True Crowd (MetricX-24-Hybrid-QE-XXL)	4	0.207	0.211	0.256	0.184	0.212	0.221	0.176	0.229	0.175	0.203
sentinel-src-24	5	0.182	0.167	0.216	0.169	0.173	0.220	0.142	0.204	0.150	0.197
Artificial Crowd (XCOMET-QE-XXL)	6	0.177	0.175	0.192	0.146	0.179	0.240	0.128	0.194	0.160	0.183
sentinel-src	6	0.175	0.164	0.205	0.159	0.171	0.223	0.118	0.201	0.141	0.190
Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	7	0.166	0.181	0.174	0.121	0.162	0.247	0.128	0.180	0.136	0.162
PreCOMET Difficulty	8	0.153	0.137	0.193	0.131	0.139	0.188	0.120	0.166	0.131	0.170
PreCOMET Diversity	9	0.142	0.059	0.167	0.134	0.129	0.213	0.120	0.159	0.130	0.165
Text Length	10	0.121	0.024	0.133	0.129	0.143	0.206	0.078	0.142	0.100	0.132
LLM-as-a-Judge (Command A, tgt-based)	11	0.104	0.077	0.100	0.098	0.120	0.190	0.068	0.117	0.072	0.097
Syntactic Complexity	12	0.080	0.018	0.078	0.072	0.112	0.181	0.035	0.090	0.050	0.079
LLM-as-a-Judge (GPT-4o, tgt-based)	12	0.080	0.061	0.067	0.072	0.116	0.179	0.035	0.079	0.037	0.071
LLM-as-a-Judge (GPT-4o, src-based)	13	0.077	0.038	0.066	0.072	0.111	0.188	0.029	0.083	0.036	0.071
LLM-as-a-Judge (Command A, src-based)	14	0.072	0.045	0.063	0.062	0.103	0.169	0.026	0.079	0.029	0.072
Random	15	0.003	-0.001	0.004	0.004	-0.008	0.010	0.008	0.005	0.008	0.000
Word Rarity	16	-0.040	0.016	-0.034	-0.044	-0.065	-0.093	-0.032	-0.043	-0.022	-0.043

Table 7: Difficulty Estimation Correlation (DEC) achieved by each method, per language, on the ESA-annotated WMT24. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2024), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

	Average		EN→DE	EN→ES	JA→ZH
	Rank	DEC	DEC	DEC	DEC
Oracle	1	1.000	1.000	1.000	1.000
Oracle (source text + target lang)	2	0.430	0.505	0.280	0.503
Oracle (source text only)	3	0.404	0.488	0.221	0.503
True Crowd (XCOMET-QE-XXL)	4	0.278	0.309	0.208	0.315
True Crowd (MetricX-24-Hybrid-QE-XXL)	5	0.248	0.268	0.192	0.284
sentinel-src-24	5	0.246	0.278	0.168	0.291
sentinel-src	6	0.235	0.273	0.165	0.268
Text Length	7	0.222	0.262	0.147	0.256
Artificial Crowd (XCOMET-QE-XXL)	8	0.207	0.243	0.159	0.220
Artificial Crowd (MetricX-24-Hybrid-QE-XXL)	9	0.185	0.209	0.145	0.201
Syntactic Complexity	10	0.170	0.158	0.073	0.278
PreCOMET Difficulty	10	0.169	0.219	0.129	0.159
PreCOMET Diversity	10	0.167	0.241	0.143	0.117
LLM-as-a-Judge (Command A, tgt-based)	11	0.120	0.122	0.088	0.150
LLM-as-a-Judge (Command A, src-based)	11	0.114	0.117	0.060	0.165
LLM-as-a-Judge (GPT-4o, tgt-based)	12	0.090	0.096	0.064	0.110
LLM-as-a-Judge (GPT-4o, src-based)	12	0.090	0.111	0.049	0.109
Random	13	0.002	0.003	0.004	0.000
Word Rarity	14	-0.052	-0.114	-0.043	0.001

Table 8: Difficulty Estimation Correlation (DEC) achieved by each method, per language, on the MQM-annotated WMT24. Ranks represent clusters of statistical significance and are computed following Freitag et al. (2024), which leverage the PERM-BOTH hypothesis test introduced by Deutsch et al. (2021).

Method	**AvgScore**	**%Perfect**
Random	84.4	21.0%
Oracle (source text only)	74.9	13.3%
Oracle (source text + target lang)	71.6	11.4%
Text Length	82.7	14.1%
sentinel-src-24	79.1	12.1%
Artificial Crowd (XCOMET-QE-XXL)	78.3	13.3%
Command A (source text + target lang)	83.0	16.1%

Table 9: Comparison of methods for selecting the most difficult 25% of samples from the ESA test set, evaluated using (1) the average human score on the selected subset and (2) the proportion of model outputs in the selected subset that achieve a perfect human score. Results are calculated per language pair and then averaged. The entire test set has an average score (AvgScore) of 84.4 and a percentage of perfect outputs (%Perfect) of 20.7%.

Method	**AvgScore**	**%Perfect**
Random	-2.5	58.8%
Oracle (source text only)	-6.6	32.7%
Oracle (source text + target lang)	-6.8	30.5%
Text Length	-4.5	43.6%
sentinel-src-24	-5.1	39.6%
Artificial Crowd (XCOMET-QE-XXL)	-4.4	43.8%
Command A (source text + target lang)	-3.1	51.1%

Table 10: Comparison of methods for selecting the most difficult 25% of samples from the MQM test set, evaluated using (1) the average human score on the selected subset and (2) the proportion of model outputs in the selected subset that achieve a perfect human score. Results are calculated per language pair and then averaged. The entire test set has an average score (AvgScore) of -2.5 and a percentage of perfect outputs (%Perfect) of 57.7%.

Method	AvgScore				%Perfect			
	News	Social	Literary	Speech	News	Social	Literary	Speech
Random	86.5	84.7	84.7	80.3	19.3%	22.6%	19.7%	12.3%
Oracle (source text only)	82.5	75.8	76.0	71.1	14.1%	16.9%	11.0%	7.0%
Oracle (source text + target lang)	79.6	71.3	72.9	68.3	11.8%	13.7%	8.0%	4.7%
Text Length	84.6	83.1	78.4	82.0	15.0%	15.5%	9.7%	11.8%
sentinel-src-24	84.1	80.2	78.7	77.5	14.4%	15.1%	10.0%	8.6%
Artificial Crowd (XCOMET-QE-XXL)	84.6	79.6	77.6	75.6	15.3%	16.6%	11.9%	8.1%
Command A (source text + target lang)	84.9	82.1	79.7	78.8	15.5%	17.4%	10.6%	10.2%

Table 11: Fine-grained evaluation of the most difficult 25% of test set samples from the ESA test set, selected independently for each domain (News, Social, Literary, Speech) and averaged across the language pairs. Results are shown for AvgScore (average human score) and %Perfect (proportion of model outputs with a perfect human score).

Method	AvgScore				%Perfect			
	News	Social	Literary	Speech	News	Social	Literary	Speech
Random	−1.4	−1.4	−3.5	−5.5	64.6%	68.9%	56.5%	37.5%
Oracle (source text only)	−4.5	−3.1	−5.9	−10.5	37.0%	45.6%	40.9%	24.6%
Oracle (source text + target lang)	−4.7	−3.4	−5.9	−11.0	33.6%	41.8%	40.7%	22.2%
Text Length	−3.3	−2.0	−5.2	−6.0	47.4%	58.5%	46.4%	37.2%
sentinel-src-24	−2.7	−2.2	−4.9	−7.1	48.7%	56.5%	48.5%	30.2%
Artificial Crowd (XCOMET-QE-XXL)	−2.5	−2.0	−3.5	−7.6	51.0%	57.8%	50.4%	30.4%
Command A (source text + target lang)	−2.3	−1.6	−4.1	−4.4	50.0%	64.2%	51.2%	40.5%

Table 12: Fine-grained evaluation of the most difficult 25% of test set samples from the MQM test set, selected independently for each domain (News, Social, Literary, Speech) and averaged across the language pairs. Results are shown for AvgScore (average human score) and %Perfect (proportion of model outputs with a perfect human score).

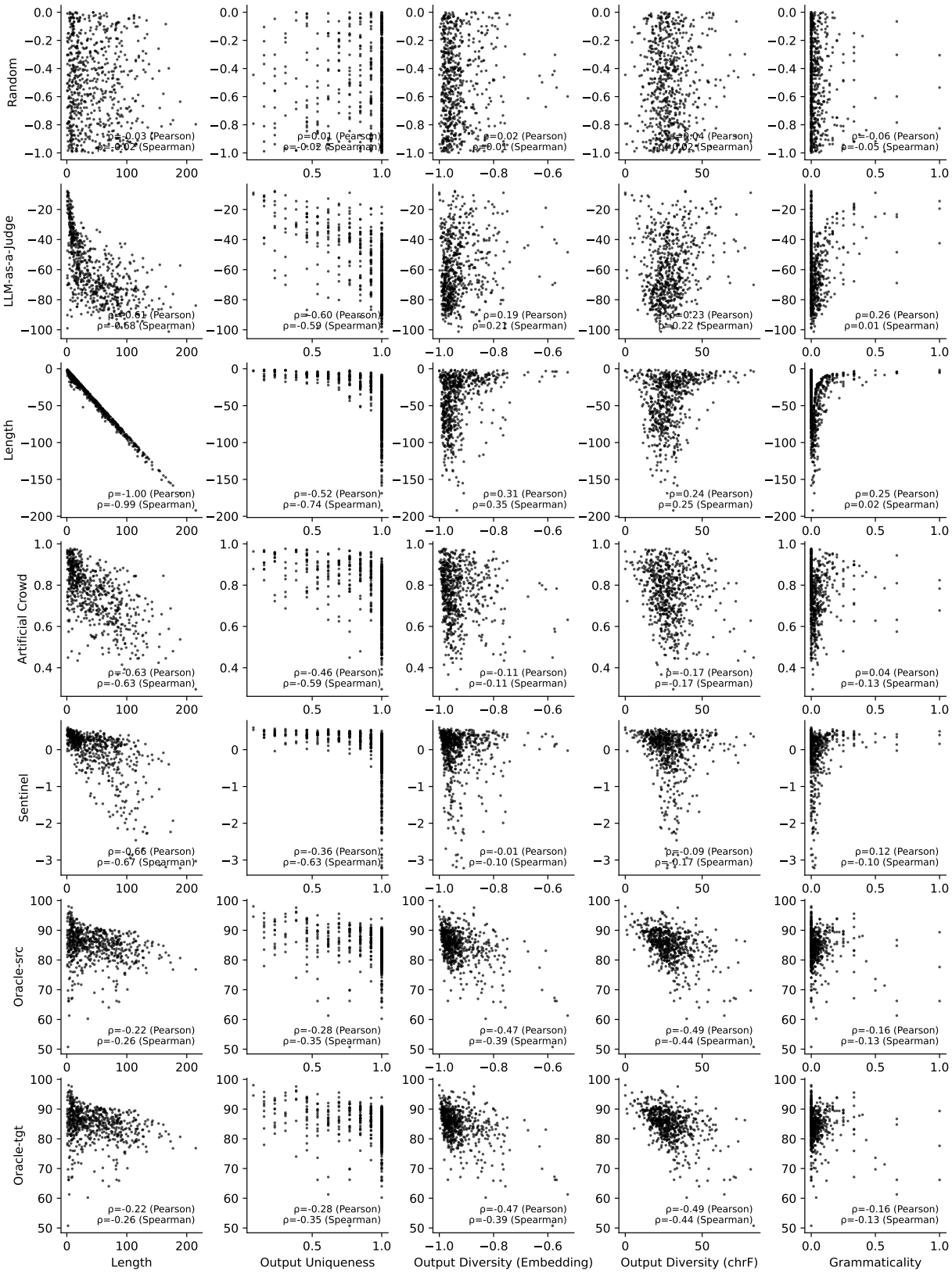

Figure 7: Relationship between selectors and variables of interest (source length, source number of errors per word, output diversity measured by pairwise embeddings inner product and chrF, and proportion of unique outputs). For all methods, lower values indicate more difficult source texts, so negative correlation implies stronger positive connection between difficulty and the target variable. See Table 3 for an aggregated perspective.

Prompt for LLM-as-a-judge (source text only):

You are given a source text. Your goal is to determine the approximate proficiency level required to translate this text, based on a detailed analysis of its complexity. The final result should be reported as a single numeric score on a scale of 0 to 120, where higher numbers correspond to a higher difficulty (i.e., more advanced language proficiency requirements). You should also relate this numeric score to commonly recognized proficiency levels (e.g., A1, A2, B1, B2, C1, C2). Here is the expected mapping: 0-20 for A1 (Beginner); 21-40 for A2 (Elementary); 41-60 for B1 (Intermediate); 61-80 for B2 (Upper Intermediate); 81-100 for C1 (Advanced); 101-120 for C2 (Mastery).

Instructions: First, examine the text to identify features that affect reading difficulty, including complexity of vocabulary, grammar, semantic density, and any specialized knowledge required. Then, provide a brief explanation of your reasoning for each major factor. Consider whether the text includes domain-specific terminology, cultural references, idiomatic expressions, or advanced grammatical constructions. Finally, assign a numeric score from 0 to 120 and map that score to one of the CEFR levels. Conclude with a final statement that clearly states your numeric score and the corresponding proficiency level surrounded by triple square brackets, for example [[[86, C1 (Advanced)]]]

Analyze following text:
{src}

Prompt for LLM-as-a-judge (source text + target language):

You are given a source text. Your goal is to determine the approximate proficiency level required to translate this text into {target_language}, based on a detailed analysis of its complexity. The final result should be reported as a single numeric score on a scale of 0 to 120, where higher numbers correspond to a higher difficulty (i.e., more advanced language proficiency requirements). You should also relate this numeric score to commonly recognized proficiency levels (e.g., A1, A2, B1, B2, C1, C2). Here is the expected mapping: 0-20 for A1 (Beginner); 21-40 for A2 (Elementary); 41-60 for B1 (Intermediate); 61-80 for B2 (Upper Intermediate); 81-100 for C1 (Advanced); 101-120 for C2 (Mastery).

Instructions: First, examine the text to identify features affecting the translation into {target_language}, which affect reading difficulty, including complexity of vocabulary, grammar, semantic density, and any specialized knowledge required. Then, provide a brief explanation of your reasoning for each major factor. Consider whether the text includes domain-specific terminology, cultural references, idiomatic expressions, or advanced grammatical constructions. Finally, assign a numeric score from 0 to 120 and map that score to one of the CEFR levels. Conclude with a final statement that clearly states your numeric score and the corresponding proficiency level surrounded by triple square brackets, for example [[[86, C1 (Advanced)]]].

Analyze following text:
{src}

Example 2: Prompts used to estimate the difficulty of a given text using LLM-as-a-judge (Section 4.3).