

SwissGov-RSD: A Human-annotated, Cross-lingual Benchmark for Token-level Recognition of Semantic Differences Between Related Documents

Michelle Wastl Jannis Vamvas Rico Sennrich

Department of Computational Linguistics, University of Zurich
{michelle.wastl, jannisnikos.vamvas, rico.sennrich}@uzh.ch

Abstract

Recognizing semantic differences across documents is crucial for text generation evaluation and content alignment, especially in cross-lingual settings. However, as a standalone task, it has received little attention. We address this by introducing *SwissGov-RSD*, the first naturalistic, document-level, cross-lingual dataset for semantic difference recognition. It encompasses a total of 224 multi-parallel documents in English–German, English–French, and English–Italian with token-level difference annotations by human annotators. We evaluate a variety of open-source and closed-source large language models as well as encoder models across different fine-tuning settings on this new benchmark. Our results show that current automatic approaches perform poorly compared to their performance on monolingual, sentence-level, and synthetic benchmarks, revealing a considerable gap for both LLMs and encoder models. We make our code and dataset publicly available¹.

1 Introduction

Recognizing semantic differences (RSD) across different versions of a text—whether monolingual or cross-lingual—is an underexplored challenge in natural language understanding. Beyond theoretical interest in tasks such as machine translation evaluation (Burchardt, 2013; Freitag et al., 2021; Rei et al., 2023) and interpretable semantic similarity (iSTS) (Agirre et al., 2016a,b), it has high practical relevance. For example, multilingual websites, especially in public and governmental contexts, are expected to convey equivalent content across languages. Yet in practice, discrepancies can go unnoticed, as illustrated in Figure 1 with

¹Code: <https://github.com/ZurichNLP/SwissGov-RSD>;
Data: <https://huggingface.co/datasets/ZurichNLP/SwissGov-RSD>. We release the SwissGov-RSD labels under a CC BY license, while the copyright of the text remains with the Swiss federal authorities: [copyright notice](#).

English

New economic operators in e-commerce

From mid-July 2021 on, any online retailer wishing to sell radio equipment or electrical appliances to customers in Switzerland must have an intermediary located in Switzerland. Also, online marketplaces or platforms will be required to cooperate with OFCOM in relation to market surveillance.

Please do not disturb

Internet and stores abroad ...

German

Neue Wirtschaftsakteurinnen im Bereich des E-Commerce

Ausländische Websites, über die Funkanlagen oder elektrische Geräte an Schweizer Kundinnen und Kunden verkauft werden, müssen ab Mitte Juli 2021 über einen Vermittler in der Schweiz verfügen. Ausserdem müssen die Online-Verkaufsplattformen im Rahmen der Marktüberwachung mit dem Bundesamt für Kommunikation (BAKOM) zusammenarbeiten.

Effizientere Notrufdienste

Ab dem 17. März 2022 wird die Standortidentifikation eines Notrufes über ein Smartphone genauer sein. So werden die Smartphones beim Aufbau des Notrufes automatisch ihre Position übermitteln. Jede und jeder von uns kann einen Unfall haben, Opfer eines Überfalls werden oder in eine andere schwierige Situation geraten. In solchen Fällen ist es unerlässlich, dass die Rettungsdienste so rasch wie möglich eingreifen können. Dazu müssen sie aber genau wissen, wohin sie fahren müssen.

Bitte nicht stören

Im Internet und in den Geschäften im Ausland ...

Figure 1: Excerpt from an English–German document pair with difference annotations from the SwissGov-RSD dataset. The differences range from small semantic shifts and explicitations to omitted paragraphs. The saturation of the red highlight indicates how large the semantic difference is, i.e., the paragraph marked in deep red contains information about emergency calls and is completely omitted in the English document.

excerpts from the multilingual Swiss government portal [admin.ch](https://www.admin.ch).

We collect document pairs with naturally occurring semantic differences from this portal across its language versions and release them with human-

Dataset	# Doc Pairs	# Total Tokens	Min. Doc Length	Max. Doc Length	Avg. Doc Length
SwissGov-RSD EN-DE	224	173,043	50	2,528	386
SwissGov-RSD EN-FR	224	195,930	52	2,837	437
SwissGov-RSD EN-IT	224	187,115	60	2,577	418

Table 1: General statistics for SwissGov-RSD. Length is measured in tokens separated by whitespaces.

annotated token-level difference spans in a new multi-parallel dataset: *SwissGov-RSD*. We then evaluate a range of systems, spanning unsupervised, few-shot and fine-tuned settings, on their ability to automatically detect these differences. Finally, we compare system performance on our dataset to results on the synthetically constructed iSTS-RSD benchmark (Vamvas and Sennrich, 2023) and explore the limitations of current systems.

Our contributions are the following:

- We construct and release SwissGov-RSD, the first human-annotated, document-level, cross-lingual dataset designed for token-level difference recognition for three language pairs: English–German, English–French, and English–Italian.
- We benchmark 6 LLMs and 12 encoder models using 3 different approaches (Figure 3) in combination with differing degrees of supervision on the naturalistic SwissGov-RSD and the synthetically constructed iSTS-RSD.
- We reveal a considerable gap between the performance on the naturalistic and synthetic datasets, showing that all systems perform substantially worse when evaluated on the naturalistic data with differences in Spearman correlation of up to 78.

Our results highlight the need for more specialized approaches to make semantic difference recognition applicable in real-world settings and point towards the limitations of synthetic benchmarking in general, where strong performance can mask a failure to generalize.

2 Recognition of Semantic Differences

Recognizing Semantic Differences (RSD) concerns identifying which parts of two texts differ in meaning. Rather than assigning a single similarity score to a text pair, RSD targets fine-grained, token-level differences. This framing is particularly relevant for comparing document versions, either across time or languages, where differences may range

from minor reformulations to omitted paragraphs, as illustrated in Figure 1.

To formalize the task, Vamvas and Sennrich (2023) propose treating RSD as a token-level regression problem, where each token is assigned a score indicating its semantic divergence from its aligned counterpart. Their benchmark repurposes human span-level similarity annotations from iSTS (Agirre et al., 2016a). They synthetically augment the iSTS dataset through extension with PAWS(-X) (Zhang et al., 2019; Yang et al., 2019), concatenation, reordering and machine translation to encompass cross-linguality and emulate document structures. In the remainder of this paper, we will refer to this benchmark as *iSTS-RSD*.

Initial evaluations on iSTS-RSD by Vamvas and Sennrich (2023) suggest that while longer texts increase the task difficulty, cross-lingual scenarios are the most challenging, highlighting the need for broader evaluation settings and more realistic data to evaluate on.

In this work, we address this need by assembling a human-annotated dataset inspired by realistic use cases and therefore enabling evaluation in settings that reflect real-world, cross-lingual applications.

3 SwissGov-RSD

Switzerland publishes its government websites at admin.ch in English and the country’s four official languages: German, French, Italian, and Romansh. These websites provide coherent texts that span a wide range of topics, making them a valuable source of natural multi-parallel data, which, either due to translation errors or unsynchronized content updates, contain semantic differences across language versions. We therefore collect multi-parallel documents from admin.ch and its subdomains in English, German, French, and Italian². The resulting dataset includes aligned document pairs in EN–DE, EN–FR, and EN–IT with descriptive statistics shown in Table 1.

²We exclude Romansh due to its limited availability, which would have substantially reduced dataset size.

Metric	EN-DE	EN-FR	EN-IT
Total # spans A1	348	148	276
Total # spans A2	139	192	382
Total # differences in tokens A1	2,543	1,167	4,451
Total # differences in tokens A2	1,699	1,453	3,420
# fuzzy matched span pairs ≥ 50	145	54	222
# fuzzy matched span pairs ≥ 75	52	36	85
# fuzzy matched span pairs ≥ 90	37	28	73
# exactly matching span pairs	27	23	64
Mean IoU (English)	23.98	24.06	32.10
Mean IoU (other language)	39.95	29.65	43.52
Mean F1 (English)	34.61	33.34	44.15
Mean F1 (other language)	51.14	40.04	55.60
Corr. fuzzy matched span pairs ≥ 50	76.47	90.21	88.12
Corr. fuzzy matched span pairs ≥ 75	78.36	98.05	93.71
Corr. fuzzy matched span pairs ≥ 90	79.50	97.66	92.89
Corr. exactly matched span pairs	73.27	91.09	82.76
Token-level Spearman correlation	56.14	43.61	65.09

Table 2: Statistics and agreement between annotators in 25 overlapping documents.

Scraping and Preprocessing We crawl [admin.ch](#) and its subdomains, which reflect different topics³, in English. If a given page includes links to corresponding versions in German, French, and Italian, we consider it for our dataset. We extract the HTML content for all four languages and apply the following filters: documents must contain at least three sentences of natural language text in the correct language; pages that are non-linguistic (e.g., link-only pages or templates) are discarded. After filtering, we retain 235 multi-parallel documents.

Labeling Scheme Following iSTS(-RSD) (Agirre et al., 2016b; Vamvas and Sennrich, 2023), we annotate semantic differences at the span level. Each annotation consists of at least one token in one of the texts, optionally aligned with a span in the other language. Spans need not be identical in length, position, or continuity across languages. For labeling, we adopt the inverse of the five-point scale previously used by Agirre et al. (2016a,b), due to our focus on semantic differences rather than explaining semantic similarity estimates, such that label 1 indicates minimal difference (minor unimportant detail), and 5 denotes complete semantic dissimilarity. Unannotated text is assumed to be semantically equivalent (label = 0). The scale is designed to support fine-grained cross-lingual comparison and accommodate both symmetric differences consisting of alignable token spans in both texts and asymmetric ones, where text from one

³See Appendix A.1 for an overview of the included subdomains.

side has no correspondence on the other side (e.g., omission/addition).

Annotation Process To annotate the collected data, we recruit two undergraduate/bachelor’s students in computational linguistics for each language pair. The annotators are either native speakers or highly proficient in the languages they are assigned to. All annotators receive detailed written guidelines outlining the annotation procedure and the semantic difference labeling scheme⁴.

The annotation is conducted in two phases: a trial phase and a main phase. During the trial phase, all annotators work on the same three text pairs for their assigned language pair. These annotations are evaluated both automatically (Table 6) and manually, and targeted feedback is provided based on the results. In the main phase, each annotator is assigned half the text pairs in a language pair with an additional 25 samples that overlap with the other annotator. The shared samples are used for the annotation validation described below.

An interactive annotation interface specifically developed for this task is used to streamline the process. Annotators are also instructed to flag any text pairs they consider faulty or unannotatable. Each annotator is compensated at a rate of approximately \$35 per hour and dedicated between 20 and 35 hours to complete the task.

Annotation Validation We assess annotation quality using automatic agreement metrics computed on the 25 overlapping text pairs annotated by both annotators, A_1 and A_2 , for each language pair. We define a difference span as a sequence of tokens labeled with the same contiguous label (>0). For all such spans S , we compute the intersection-over-union (IoU)

$$\text{IoU} = \frac{|S_{A_1} \cap S_{A_2}|}{|S_{A_1} \cup S_{A_2}|},$$

where the intersection corresponds to all overlapping spans and the union to the total number of annotated difference spans in a document. We report the macro-average over all document IoU scores for each language. Additionally, we calculate F1 scores between the two annotators.

To evaluate consistency in label assignment for difference spans, we calculate the Spearman rank correlation between the annotators’ labels, where

⁴Appendix A.2 contains the full guidelines, including a screenshot of the interface and examples to illustrate the labeling scheme.

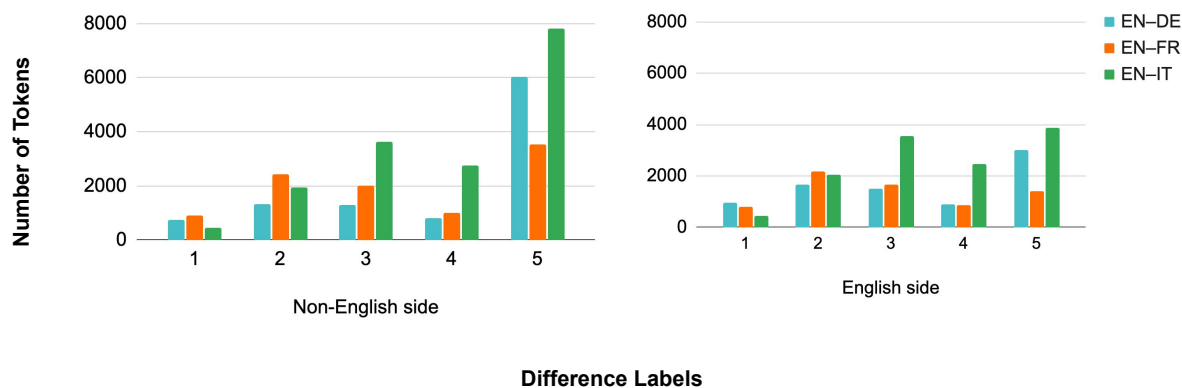


Figure 2: Label distribution of the final SwissGov-RSD dataset in tokens (separated by whitespaces). 0-labeled tokens are not included in this plot.

non-annotated tokens are excluded from the computation. Correlation is measured for both exact span matches and fuzzy matches, defined as spans with IoU scores larger than 50, 75, and 90 respectively. In addition, to facilitate comparison between annotator agreement and model performance, we report the Spearman correlation over all labels as defined in Subsection 5.4. Results are summarized in Table 2.

While the number of annotated spans varies across language pairs and annotators, IoU and F1 scores remain broadly comparable, suggesting that annotators recognize similar spans despite differing annotation strategies. Across language pairs, EN-DE shows the lowest label agreement (correlation 76.47 at ≥ 50 overlap) despite comparable overlap statistics with other language pairs, indicating diverging label choices over similar spans. EN-FR presents the inverse pattern: high label correlations (up to 98.05) alongside lower overlap statistics, reflecting stronger consensus on labels than on exact span boundaries. For EN-IT we observe the highest overall overlap agreement and high label correlation. The inter-annotator Spearman correlation ranges from 43.61 (EN-FR) to 65.09 (EN-IT)⁵. While these values may appear low, they are consistent with agreement reported in previous literature on cross-lingual, pairwise span-level annotation tasks such as error span annotation in translation evaluation (Kreutzer et al., 2020; Kocmi et al., 2024) and overlap statistics in span-level error detection (Lavie et al., 2025) or qual-

⁵After a preliminary quality assessment, we identified one below-average annotation run for EN-DE and repeated it with a new annotator. Replacing the annotations from the first run led to an increase in Spearman correlation from 9.4 to 56.14.

ity estimation for post-editing (Sarti et al., 2025), since these tasks are inherently difficult for human annotators as well. An analysis of the effect of annotation discrepancy on model evaluation is shown in Appendix A.4.

We merge the sets of annotated documents. Where documents have been annotated by both annotators, we manually inspect them. We include the annotations adhering more to the guidelines and the label distribution of the other annotators in the dataset. Furthermore, the documents flagged as faulty were also manually inspected and removed if necessary. After filtering 11 faulty documents, the final dataset comprises 224 multi-parallel documents, totaling 886 texts (Table 1).

Figure 2 shows the label distribution of the dataset. We observe that around 10–17% of tokens are labeled as semantically different—a considerably more skewed semantic equivalence-to-difference ratio than in the iSTS-RSD dataset with approximately 25–30%.

4 Automatic Detection Approaches

In order to assess how well state-of-the-art methods align with human judgment in recognizing semantic differences, we evaluate three different approaches: an unsupervised RSD algorithm and common specialization techniques, namely few-shot prompting and fine-tuning.

4.1 Unsupervised Baseline

As a baseline, we use DiffAlign, the best-performing RSD approach described by Vamvas and Sennrich (2023) (Figure 3a).

Given a pair of texts A and B , each is independently encoded into a sequence of contextualized

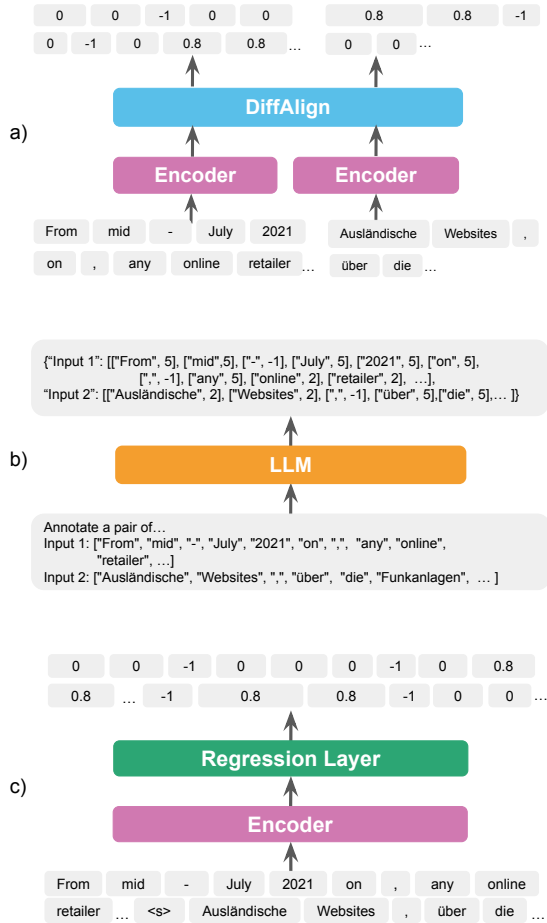


Figure 3: Architectures used in our experiments. An **unsupervised approach** (a), where each document is encoded separately and a difference alignment algorithm is used to predict difference scores for each token. We provide **LLMs** (b) with a natural language instruction and examples of the expected output. The two text segments to compare are provided in tokenized form. The LLM then autoregressively generates a JSON object with a score for each token. The **token regressor** (c) predicts a score for each encoded token in the sequence pair.

token embeddings $h(A) = [h(a_1), \dots, h(a_n)]$ and $h(B) = [h(b_1), \dots, h(b_m)]$. For this we use the final hidden states of a pretrained encoder model. For each token a_i in A , a soft alignment score is computed as the maximum cosine similarity between $h(a_i)$ and all token embeddings in B :

$$\text{diff}_{align}(a_i) = 1 - \max_{b_j \in B} \cos(h(a_i), h(b_j)).$$

The resulting score reflects the degree to which a token in one text is semantically aligned with the other. Tokens with low alignment confidence (i.e., high DiffAlign values) are interpreted as likely

semantic differences. The method is entirely unsupervised and requires no labeled training data.

4.2 Few-shot Prompting

We instruct an LLM to label every token of a sequence pair according to their semantic similarity on a scale from 0 to 5, equivalent to the labeling scheme employed by iSTS (Agirre et al., 2016a). The output is to be given in JSON format. We add three in-context examples that illustrate the task and the desired format (Figure 3b). The prompt is attached in Appendix E.

4.3 Fine-tuning

LLMs For fine-tuning LLMs, we use a supervised fine-tuning (SFT) objective that minimizes the cross-entropy between the model’s predictions and the gold response, given the few-shot prompt.

Encoder Models Here, we concatenate the two input sequences, separated by a special token, and encode them with the model. The final hidden states of the tokens are passed through a linear regression layer to predict a score for each token (Figure 3c). During training, we minimize the binary cross-entropy between the predicted scores and the human-annotated labels. The scores are on a continuous scale from 0 to 1, where 0 indicates perfect alignment and 1 the strongest semantic difference, which is consistent with the scale used by Vamvas and Sennrich (2023).

5 Experimental Setup

5.1 LLMs

We evaluate open-weight models of two sizes, Llama-3.1-8B-Instruct and its 405B equivalent (Grattafiori et al., 2024), as well as two closed-source commercial models, GPT-4o-mini and GPT-4o (OpenAI et al., 2024). Additionally, we evaluate two models that generate reasoning tokens before predicting the final response: the open-weight DeepSeek-R1 (Guo et al., 2025) and the closed-source commercial o3-mini (OpenAI et al., 2025), for which we set the reasoning effort to “low”.

For Llama, we generate responses using greedy decoding, while for GPT, we use the default settings of the API (i.e., sampling without temperature scaling) and enforce valid JSON output.

For fine-tuning Llama 8B, we train LoRA (Hu et al., 2022) as implemented in the PEFT library⁶

⁶<https://github.com/huggingface/peft>

with a rank of $r = 32$ and a scaling factor of $\alpha = 16$, and we use 4-bit quantization during training (Dettmers et al., 2023). We fine-tune the model with a batch size of 16 and a learning rate of 2×10^{-4} for 10 epochs, with early stopping on the validation set.

To fine-tune GPT-4o-mini, we use the default settings recommended by OpenAI (3 epochs, batch size 1, lr multiplier 1.8).

5.2 Encoders

For DiffAlign, we follow Vamvas and Sennrich (2023) and use XLM-R + SimCSE (Conneau et al., 2020; Gao et al., 2021; Wang et al., 2022), which has been shown to produce high-quality semantic representations for token-level alignment tasks, and LaBSE (Feng et al., 2022). Since the sequence length of XLM-R and LaBSE is restricted to 512, we also experiment with ModernBERT-large (Warner et al., 2025), its multilingual variants EuroBERT (210M, 610M, 2.1B) (Boizard et al., 2025), MmBERT (Marc et al., 2025), as well as bge-m3 (Chen et al., 2024), and Qwen3-Embedding (Zhang et al., 2025), which allow us to input a whole document without splitting, due to their native sequence length of 8,192 tokens.

For our encoder fine-tuning experiments, we use ModernBERT-large (Warner et al., 2025). We train the models with a batch size of 32 for 4 epochs, using early stopping based on the validation loss and select the best model based on validation loss out of four learning rates: 1×10^{-5} , 3×10^{-5} , 5×10^{-5} , and 8×10^{-5} . The last learning rate performed best.

Details on the individual LLMs and encoder models are presented in Appendix G.

5.3 Fine-tuning Data

For fine-tuning, we use the train splits from iSTS-RSD (Vamvas and Sennrich, 2023), which are based on the human-annotated data from SemEval-2016 Task 2 (Agirre et al., 2016b), and 1500 human-validated paraphrases from PAWS (Zhang et al., 2019) to augment negative samples (texts without semantic differences). Label schemes of the latter are converted to the token-level. To train the models on longer inputs beyond single sentences, we apply the same data augmentations used by Vamvas and Sennrich (2023), concatenating inputs into synthetic ‘documents’ of up to 5 sentences, with up to 5 permutations of the sentence order between document pairs. To cover all lan-

guage pairs in SwissGov-RSD and enable a direct comparison, we extend iSTS-RSD to EN-IT⁷.

The iSTS-RSD data contains labeled examples only in English, limiting support to English-only fine-tuning. To support cross-lingual fine-tuning without requiring human annotations in the non-English language, we augment the machine-translated side of the training sets with label projection (Akbik et al., 2015; Chen et al., 2023). For each training example, we provide GPT-4o-mini with the English source sentence and its corresponding human-annotated token-level labels, prompting it to assign equivalent labels to the machine-translated version in German, French, or Italian. Prompt templates for each language pair are provided in Appendix D. This approach bypasses the need for explicit translation of span annotations and leverages the LLM’s multilingual and instruction-following capabilities. We refer to this approach as *direct LLM-based label projection* to distinguish it from multi-step or marker-based projection approaches (Chen et al., 2023; Le et al., 2024; Parekh et al., 2024). In Appendix F, we present the results of a manual evaluation of the label projections. For experiments using projected labels, we apply the same data sources and augmentation strategies, with the distinction that one side of each pair is in German, French, or Italian.

This approach allows us to invest comparable amounts of annotated data during fine-tuning across all models, while adjusting the number of augmentations based on computational resources: For fine-tuning the LLMs, which have more parameters and require more computational resources, we use 560 augmentations for training. For the encoder models, we generate 10,000 augmentations. For the multilingually fine-tuned encoder, we concatenate the EN-DE, EN-FR, and EN-IT training data with projected labels. Detailed statistics of the iSTS-RSD training and test sets are shown in Appendix B and an example of an augmented sample is presented in Appendix C.

5.4 Evaluation

We evaluate the above-described systems on the test split of iSTS-RSD as well as on the whole

⁷To do so, we use DeepL to machine translate the English iSTS dataset to Italian and re-align the sentence pairs to form an EN-IT dataset: <https://huggingface.co/datasets/ZurichNLP/rsd-ists-2016>. In addition, we machine-translate PAWS from English to Italian to also support augmentation with negative examples: <https://huggingface.co/datasets/ZurichNLP/paws-x-italian>.

Approach	iSTS-RSD			SwissGov-RSD		
	EN-DE	EN-FR	EN-IT	EN-DE	EN-FR	EN-IT
<i>DiffAlign (unsupervised)</i>						
XLM-R+SimCSE [†]	44.9	45.2	44.9	19.2	10.0	23.4
LaBSE [†]	40.6	42.3	45.3	20.6	8.2	25.6
bge-m3	47.1	45.8	47.5	14.8	9.9	13.6
gte-multilingual-base	31.2	30.2	30.5	10.7	5.9	12.8
ModernBERT-large	17.3	16.7	17.2	0.5	0.7	2.7
EuroBERT 210M	32.1	34.6	33.4	11.5	8.4	16.1
EuroBERT 610M	29.1	30.5	31.0	18.6	11.4	18.7
EuroBERT 2.1B	40.3	28.5	28.3	13.7	8.9	17.4
MmBERT-small	23.0	23.5	22.1	10.3	5.5	11.0
MmBERT-base	34.0	32.6	32.6	14.3	8.6	15.6
Qwen3-Embedding 4B	41.6	42.9	43.2	20.1	11.2	20.3
Qwen3-Embedding 8B	40.9	40.1	41.4	19.3	12.1	18.6
<i>LLMs with few-shot prompting</i>						
Llama-3.1 8B Instruct	2.8	3.7	2.6	-	-	-
Llama-3.1 405B Instruct	29.2	29.3	26.9	-0.8	-2.2	<u>7.4</u>
GPT-4o-mini	15.8	13.8	16.6	-	-	-
GPT-4o	43.0	40.5	42.5	<u>4.2</u>	<u>1.2</u>	5.5
DeepSeek-R1	38.2	40.1	34.6	-	-	-
o3-mini-low	<u>44.8</u>	<u>46.5</u>	<u>48.2</u>	-	-	-
<i>Fine-tuned LLMs</i>						
Llama-3.1 8B Instruct	66.7	67.3	66.9	-	-	-
GPT-4o-mini	81.6	79.9	78.2	<u>3.9</u>	<u>1.1</u>	<u>5.3</u>
<i>Fine-tuned encoder models</i>						
ModernBERT-large (EN)	55.3	55.4	53.8	5.7	1.3	4.0
ModernBERT-large (EN-DE*)	58.8	58.4	58.5	6.4	1.6	5.6
ModernBERT-large (EN-FR*)	58.4	60.7	60.1	5.4	<u>5.2</u>	8.7
ModernBERT-large (EN-IT*)	<u>59.2</u>	<u>61.2</u>	<u>63.0</u>	<u>7.5</u>	4.8	<u>12.4</u>
ModernBERT-large (multi*)	54.7	57.5	58.1	5.7	2.5	8.8

Table 3: Token-level Spearman correlations with gold labels for three language pairs, comparing results on the iSTS-RSD and the SwissGov-RSD dataset. (*) denotes encoders fine-tuned on data with projected labels, **bold** the best performance overall, and underline best performance within model category. (†) denotes that the input sequences for SwissGov had to be split due to length limitations of the model. Due to computational restraints, for LLM-based experiments, we only consider the best-performing systems during iSTS-RSD evaluation for SwissGov as well.

SwissGov-RSD dataset.

The test split of the iSTS-RSD benchmark includes 5 difficulty settings ranging from individual English sentence pairs (‘iSTS’) to a challenging cross-lingual setting where documents are sequences of 5 permuted sentences including 7 different language pairs: EN–DE, EN–FR, EN–IT, EN–ES, EN–ZH, EN–JA, and EN–KO. We restrict the evaluation to 100 samples per difficulty setting and language to reduce computational cost.

Some LLM-generated outputs were not in the expected format, either due to additional text outside the expected JSON format or a mismatch in the number of token-label pairs. In the former case, we apply heuristics to extract the JSON object; in the latter, we pad or truncate the output to align with the gold sequence length to enable evaluation. Furthermore, the LLM-generated output sometimes contains labels outside the expected range. These

were not altered and used as is for correlation computation.

To evaluate model performance, we use token-level Spearman correlations between all gold labels and predictions, where semantically equivalent tokens are labeled 0. Since Spearman correlations can be unstable if there are many ties, we also compute Kendall τ -b scores for the results in Table 3 and present them in Appendix I, showing no substantial differences from Spearman correlation-based evaluation.

6 Results and Discussion

Synthetically augmented datasets do not sufficiently reflect properties of a realistic scenario.

Table 3 presents a comparison of model performance on the EN–DE, EN–FR, EN–IT subsets

	English	German
Gold annotation	With the launch of Horizon Europe , the various types of multilateral initiatives were brought together under one umbrella	In Horizon Europe wurden die verschiedenen Arten von multilateralen Initiativen unter einem einheitlichen Dach zusammengefasst
DiffAlign MmBERT-base	With the launch of Horizon Europe , the various types of multilateral initiatives were brought together under one umbrella	In Horizon Europe wurden die verschiedenen Arten von multilateralen Initiativen unter einem einheitlichen Dach zusammengefasst
GPT-4o-mini	With the launch of Horizon Europe , the various types of multilateral initiatives were brought together under one umbrella	In Horizon Europe wurden die verschiedenen Arten von multilateralen Initiativen unter einem einheitlichen Dach zusammengefasst
Fine-tuned GPT-4o-mini	With the launch of Horizon Europe , the various types of multilateral initiatives were brought together under one umbrella	In Horizon Europe wurden die verschiedenen Arten von multilateralen Initiativen unter einem einheitlichen Dach zusammengefasst
Fine-tuned ModernBERT (multi)	With the launch of Horizon Europe , the various types of multilateral initiatives were brought together under one umbrella	In Horizon Europe wurden die verschiedenen Arten von multilateralen Initiativen unter einem einheitlichen Dach zusammengefasst

Figure 4: Excerpt of an EN–DE document pair with gold labels and predictions by one model from each of the system categories listed in Table 3.

of iSTS-RSD⁸ and the full SwissGov-RSD benchmark. Most notably, across all model types and language pairs, performance drops substantially when moving from iSTS-RSD to SwissGov-RSD. This confirms that models tuned or evaluated only on synthetically augmented data do not generalize well to naturalistic, real-world scenarios. The fine-tuned GPT-4o-mini model exemplifies this gap the most: Despite achieving the highest scores on iSTS-RSD, its performance drops to among the lowest on SwissGov-RSD, suggesting potential overfitting to the data augmentation patterns. Interestingly, while overall scores on SwissGov-RSD are lower, the unsupervised approaches show relatively strong robustness, outperforming all other few-shot and fine-tuned models. This may indicate that more specialized models suffer from out-of-domain effects, as SwissGov-RSD differs substantially from iSTS-RSD not only in terms of naturalness, but also in domain, length, and label distribution.

Encoders are competitive. Encoder-based models remain highly competitive, particularly in unsupervised settings (Table 3). Among long-context models, Qwen3-Embedding models achieve the strongest results, followed by EuroBERT 610M. While LaBSE and XLM-R + SimCSE achieve the highest scores in the unsupervised setting, their limited input length requires document segmentation prior to processing. This effectively re-

⁸A comprehensive overview of the results over the different augmentation categories and languages of iSTS-RSD are presented in Appendix H.

Approach	iSTS-RSD	SwissGov-RSD
<i>LLMs with few-shot prompting</i>		
Llama-3.1 8B Instruct	0.60%	–
Llama-3.1 405B Instruct	0.07%	7.00%
GPT-4o	0.03%	2.10%
<i>Fine-tuned LLMs</i>		
Llama-3.1 8B Instruct	0.10%	–
GPT-4o-mini	0.07%	7.40%

Table 4: Percentage of samples for which LLMs failed to produce the correct number of labels for the three investigated language pairs (see Table 13 for full iSTS-RSD). For SwissGov-RSD, each text of a pair is counted separately, hence, the model can fail twice for one sample.

duces the task to shorter units and avoids challenges associated with long-context modeling. The higher performance under segmentation suggests that long-context processing, even when supported, may weaken token-level cross-lingual alignment signals necessary for RSD. DiffAlign underperforms when using ModernBERT, likely due to its English-centricity. However, when fine-tuned, it surpasses LLMs in a naturalistic setting with few-shot prompting, all while being more efficient in terms of time and computational resources (see Appendix J for an inference time comparison).

LLMs are highly limited on this task. Table 3 shows that for iSTS-RSD, fine-tuned GPT-4o-mini performs best, while few-shot prompted LLMs fail to beat the unsupervised baseline, with reasoning bringing only little improvement. On SwissGov-RSD, however, LLMs show the largest

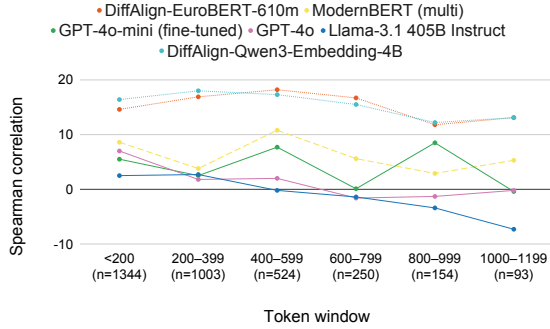


Figure 5: Average Spearman correlation coefficient at different positions in the documents, averaged across language pairs. n indicates the number of documents. Dotted lines represent DiffAlign approaches, dashed the fine-tuned regressor, and solid LLMs.

performance drops, especially fine-tuned GPT-4o-mini. Proprietary models generally outperform open-weight alternatives, and despite strong performance in synthetically augmented settings, LLMs are prone to output formatting errors, sequence length mismatches in particular, which worsen on SwissGov-RSD (Table 4), a problem encoder-based models avoid by design. Qualitatively, LLMs also tend to underlabel semantic differences, most noticeably toward document ends or in the second input document, as illustrated in Figure 4.

Long documents expose weaknesses. To better understand how models behave on extended inputs, we analyze the correlation between model predictions and gold labels by token position in Figure 5. While all systems show a slight downward trend as documents grow longer, encoder-based systems, specifically the unsupervised DiffAlign approach, appear more robust than LLMs. The plot suggests that LLaMA-3.1 405B is most sensitive to document length, often failing to produce complete output for inputs exceeding 1200 tokens. This suggests that unsupervised approaches with encoder models or in-domain fine-tuning may be promising strategies for cross-lingual RSD for long documents. For LLMs, however, future work may need to investigate strategies that preserve structured output quality for sequence labeling at extended input lengths.

7 Related Work

RSD is closely related to STS, paraphrase detection, and natural language inference (NLI). Apart from the above-mentioned (i-)STS(-RSD) and PAWS(-X) datasets (Agirre et al., 2016a,b;

Vamvas and Sennrich, 2023; Zhang et al., 2019; Yang et al., 2019), standard NLI datasets like SNLI, MNLI, and XNLI (Bowman et al., 2015; Williams et al., 2018; Conneau et al., 2018) also relate, but provide sentence-level annotations, whereas our task operates at token-level granularity.

RSD also intersects with evaluation of generated text, particularly in machine translation and hallucination detection. Datasets using the MQM framework (Lommel et al., 2024), such as MQM (Freitag et al., 2021), QT (Specia et al., 2017), and ACES (Amrhein et al., 2022; Moghe et al., 2025), annotate translation errors which can also include semantic differences. Hallucination detection benchmarks, such as HaDes (Liu et al., 2022), and multilingual Mushroom (Mickus et al., 2024; Vazquez et al., 2025) include semantic errors but are not parallel.

Our dataset also relates to comparable corpora, which contain semantically aligned but not fully parallel documents. Related corpora include SwissAdmin (Scherrer et al., 2014), the Bulletin Corpus (Volk et al., 2016), and 20min-XD (Wastl et al., 2025), which similarly leverage Swiss multilingual content but do not include token-level semantic annotations.

8 Conclusion

We release the first human-annotated, cross-lingual, document-level dataset consisting of naturally occurring semantic differences in multilingual government texts: SwissGov-RSD. Through a comprehensive evaluation of state-of-the-art systems, including LLMs, encoder models, and widely used specialization techniques on SwissGov-RSD and previous synthetically augmented benchmarks, we find a considerable performance gap between the synthetic and naturalistic evaluation settings. Our findings indicate that current systems struggle to generalize to real-world data, as data augmentation fails to capture the complexity of naturalistic settings. This emphasizes that document-level RSD across languages remains a complex challenge calling for new approaches that align better with human judgment and are applicable in real-world scenarios.

Limitations

Language and script diversity SwissGov-RSD is limited to high-resource, closely related languages (German, French, Italian) and Latin scripts.

Furthermore, the language pairs chosen for the dataset are English-centric. While this setting should be comparatively favorable, our results show that current models already struggle under these conditions. At the same time, the generalizability of our findings to typologically more distant languages and those using non-Latin scripts remains untested in a realistic setting. Due to the dataset’s multi-parallel nature, future work could extend the annotation to other language combinations, although we note the high annotation cost.

Annotation consistency Annotation quality may vary across annotators. Ideally, multiple annotators per document pair for all samples would be used to ensure high reliability. Due to resource constraints, we relied on a two-annotator setup with overlap-based quality checks.

Prompting simplicity Our prompting approach is intentionally simple and uniform across models to support comparability. While more elaborate prompts might improve performance marginally, the primary issues we observe suggest deeper limitations in LLM robustness that prompt tuning alone is unlikely to resolve.

Acknowledgements

This work was funded by the Swiss National Science Foundation (project InvestigaDiff; no. 10000503). We thank the reviewers for their constructive feedback and valuable suggestions, and are grateful to Cui Ding and Zach Hopton for proofreading.

References

- Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016a. [SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 497–511, San Diego, California. Association for Computational Linguistics.
- Eneko Agirre, Aitor Gonzalez-Agirre, Iñigo Lopez-Gazpio, Montse Maritxalar, German Rigau, and Larraitz Uribe. 2016b. [SemEval-2016 task 2: Interpretable semantic textual similarity](#). In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 512–524, San Diego, California. Association for Computational Linguistics.
- Alan Akbik, Laura Chiticariu, Marina Danilevsky, Yunyao Li, Shivakumar Vaithyanathan, and Huaiyu Zhu. 2015. [Generating high quality proposition Banks for multilingual semantic role labeling](#). In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 397–407, Beijing, China. Association for Computational Linguistics.
- Chantal Amrhein, Nikita Moghe, and Liane Guillou. 2022. [ACES: Translation accuracy challenge sets for evaluating machine translation metrics](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 479–513, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte Miguel Alves, Andre Martins, Ayoub Hammal, Caio Corro, Celine Hudelot, Emmanuel Malherbe, Etienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El Haddad, Manuel Faysse, Maxime Peyrard, Nuno M Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. [EuroBERT: Scaling multilingual encoders for european languages](#). In *Second Conference on Language Modeling*.
- Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. [A large annotated corpus for learning natural language inference](#). In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.
- Aljoscha Burchardt. 2013. [Multidimensional quality metrics: a flexible system for assessing translation quality](#). In *Proceedings of Translating and the Computer 35*, London, UK. Aslib.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Yang Chen, Chao Jiang, Alan Ritter, and Wei Xu. 2023. [Frustratingly easy label projection for cross-lingual transfer](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5775–5796, Toronto, Canada. Association for Computational Linguistics.
- 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, Online. Association for Computational Linguistics.

- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. [XNLI: Evaluating cross-lingual sentence representations](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. [Language-agnostic BERT sentence embedding](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.
- 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.
- Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. [SimCSE: Simple contrastive learning of sentence embeddings](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Naman Goyal, Jingfei Du, Myle Ott, Giri Anantharaman, and Alexis Conneau. 2021. [Larger-scale transformers for multilingual masked language modeling](#). In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 29–33, Online. Association for Computational Linguistics.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z F Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning](#). *Nature*, 645(8081):633–638.
- Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. [Lora: Low-rank adaptation of large language models](#). In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.
- Tom Kocmi, Vilém Zouhar, Eleftherios Avramidis, Roman Grundkiewicz, Marzena Karpinska, Maja Popović, Mrinmaya Sachan, and Mariya Shmatova. 2024. Error span annotation: A balanced approach for human evaluation of machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 1440–1453, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Julia Kreutzer, Nathaniel Berger, and Stefan Riezler. 2020. [Correct me if you can: Learning from error corrections and markings](#). In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 135–144, Lisboa, Portugal. European Association for Machine Translation.
- Alon Lavie, Greg Hanneman, Sweta Agrawal, Diptesh Kanojia, Chi-Kiu Lo, Vilém Zouhar, Frederic Blain, Chrysoula Zerva, Eleftherios Avramidis, Sourabh Deoghare, Archchana Sindhujan, Jiayi Wang, David Ifeoluwa Adelani, Brian Thompson, Tom Kocmi, Markus Freitag, and Daniel Deutsch. 2025. Findings of the WMT25 shared task on automated translation evaluation systems: Linguistic diversity is challenging and references still help. In *Proceedings of the Tenth Conference on Machine Translation*, pages 436–483, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Duong Minh Le, Yang Chen, Alan Ritter, and Wei Xu. 2024. [Constrained decoding for cross-lingual label projection](#). In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.
- Tianyu Liu, Yizhe Zhang, Chris Brockett, Yi Mao, Zhifang Sui, Weizhu Chen, and Bill Dolan. 2022. [A token-level reference-free hallucination detection benchmark for free-form text generation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6723–6737, Dublin, Ireland. Association for Computational Linguistics.
- Arle Lommel, Serge Gladkoff, Alan Melby, Sue Ellen Wright, Ingemar Strandvik, Katerina Gasova, Angelika Vaasa, Andy Benzo, Romina Marazzato Sparano, Monica Foresi, Johani Innis, Lifeng Han, and Goran Nenadic. 2024. [The multi-range theory of translation quality measurement: MQM scoring models and statistical quality control](#). In *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 2: Presentations)*, pages 75–94, Chicago, USA. Association for Machine Translation in the Americas.
- Marone Marc, Weller Orion, Fleshman William, Yang Eugene, Lawrie Dawn, and Benjamin Van Durme. 2025. MmBERT: A modern multilingual encoder with annealed language learning. *arXiv [cs.CL]*.

- Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, Vincent Segonne, Alessandro Raganato, and Marianna Apidianaki. 2024. [SemEval-2024 task 6: SHROOM, a shared-task on hallucinations and related observable overgeneration mistakes](#). In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.
- Nikita Moghe, Arnisa Fazla, Chantal Amrhein, Tom Kocmi, Mark Steedman, Alexandra Birch, Rico Sennrich, and Liane Guillou. 2025. [Machine translation meta evaluation through translation accuracy challenge sets](#). *Computational Linguistics*, 51(1):73–137.
- OpenAI, :, Ahmed El-Kishky, Alexander Wei, Andre Saraiva, Borys Minaiev, Daniel Selsam, David Dohan, Francis Song, Hunter Lightman, Ignasi Clavera, Jakub Pachocki, Jerry Tworek, Lorenz Kuhn, Lukasz Kaiser, Mark Chen, Max Schwarzer, Mostafa Rohaninejad, Nat McAleese, and 7 others. 2025. [Competitive programming with large reasoning models](#). *Preprint*, arXiv:2502.06807.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. [Gpt-4o system card](#). *Preprint*, arXiv:2410.21276.
- Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2024. [Contextual label projection for cross-lingual structured prediction](#). 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 5738–5757, Mexico City, Mexico. Association for Computational Linguistics.
- Ricardo Rei, Nuno M. Guerreiro, Marcos Treviso, Luisa Coheur, Alon Lavie, and André Martins. 2023. [The inside story: Towards better understanding of machine translation neural evaluation metrics](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1089–1105, Toronto, Canada. Association for Computational Linguistics.
- Gabriele Sarti, Vilém Zouhar, Malvina Nissim, and Arianna Bisazza. 2025. [Unsupervised word-level quality estimation for machine translation through the lens of annotators \(dis\)agreement](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 18320–18337.
- Yves Scherrer, Luka Nerima, Lorenza Russo, Maria Ivanova, and Eric Wehrli. 2014. [SwissAdmin: A multilingual tagged parallel corpus of press releases](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1832–1836, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Lucia Specia, Kim Harris, Frédéric Blain, Aljoscha Burchardt, Viviven Macketanz, Inguna Skadin, Matteo Negri, and Marco Turchi. 2017. [Translation quality and productivity: A study on rich morphology languages](#). In *Proceedings of Machine Translation Summit XVI: Research Track*, pages 55–71, Nagoya Japan.
- Jannis Vamvas and Rico Sennrich. 2023. [Towards unsupervised recognition of token-level semantic differences in related documents](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.
- 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. 2025. [SemEval-2025 task 3: MUSHROOM, the multilingual shared-task on hallucinations and related observable overgeneration mistakes](#). In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2472–2497, Vienna, Austria. Association for Computational Linguistics.
- Martin Volk, Chantal Amrhein, Noëmi Aepli, Mathias Müller, and Phillip Ströbel. 2016. [Building a parallel corpus on the world’s oldest banking magazine](#). In *KONVENS*. s.n.
- Yaushian Wang, Ashley Wu, and Graham Neubig. 2022. [English contrastive learning can learn universal cross-lingual sentence embeddings](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9122–9133, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. [Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.
- Michelle Wastl, Jannis Vamvas, Selena Calleri, and Rico Sennrich. 2025. [20min-XD: A comparable corpus of Swiss news articles](#). In *Proceedings of the 10th edition of the Swiss Text Analytics Conference*, pages 1–10, Winterthur, Switzerland. Association for Computational Linguistics.
- Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. [A broad-coverage challenge corpus for sentence understanding through inference](#). In *Proceedings of the 2018 Conference of the North American*

Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Yinfei Yang, Yuan Zhang, Chris Tar, and Jason Baldridge. 2019. [PAWS-X: A cross-lingual adversarial dataset for paraphrase identification](#). 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 3687–3692, Hong Kong, China. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, and 1 others. 2024. [mgte: Generalized long-context text representation and reranking models for multilingual text retrieval](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. [Qwen3 embedding: Advancing text embedding and reranking through foundation models](#). *arXiv [cs.CL]*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. [PAWS: Paraphrase adversaries from word scrambling](#). 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 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

A Details on SwissGov-RSD Dataset Collection Process

A.1 Domains

Domain	# Documents	Description
admin.ch	13	Main page of the Swiss federal government providing general information about its structure, political system, and key officeholders (Federal Council members).
bfe.admin.ch	86	Bundesamt für Energie (BFE): Official site of the Swiss Federal Office of Energy, which is the national center of expertise for energy supply and consumption.
bk.admin.ch	49	Bundeskanzlei (BK): Website of the Swiss Federal Chancellery, the staff organization of the federal government (Federal Council), providing details on the Federal Chancellor, the Chancellery's history, its organizational structure, and its various sections and functions.
seco.admin.ch	20	Staatssekretariat für Wirtschaft (SECO): The site of SECO (State Secretariat for Economic Affairs), serving as the federal government's center of expertise for economic and labor market policy.
uvek.admin.ch	21	Umwelt, Verkehr, Energie und Kommunikation (UVEK): Official site of the Federal Department of the Environment, Transport, Energy and Communications (DETEC), which functions as Switzerland's ministry for infrastructure and the environment.
sbfi.admin.ch	21	Staatssekretariat für Bildung, Forschung und Innovation (SBFI): The website of SERI (State Secretariat for Education, Research and Innovation), the federal agency specialized in national and international policies on education, research and innovation.
bakom.admin.ch	14	Bundesamt für Kommunikation (BAKOM): Official site of the Federal Office of Communications, the Swiss authority responsible for telecommunications, broadcasting (radio and television), and postal services regulation.
Total	224	

Table 5: Overview of crawled (sub-)domains and their document counts in the final SwissGov-RSD dataset.

A.2 Annotation Guidelines

Annotation Guidelines

Please read this document fully before starting the annotation process.

Semantic Difference Annotation

You will be shown a collection of parallel texts, one parallel text pair at a time. Each pair contains one text in English and one in your assigned language (either French, Italian, or German). Texts in a pair will generally convey the same information, although differences may be present. The task is to identify any such differences and assign a label depending on the degree of semantic difference on a scale from 1 to 5.

Examples

The scale is defined as follows. Italics show examples of sentence pairs with differences marked in yellow:

1: The two spans are mostly equivalent, but some unimportant details differ

- *Schweizerinnen und Schweizer verfügen über ausgedehnte politische Rechte*
vs.
The Swiss electorate enjoys extensive political rights
- *Les citoyens suisses disposent de droits politiques étendus . Ils peuvent notamment lancer et signer des initiatives et des référendums .*
vs.
The Swiss electorate enjoys extensive political rights . Registered voters are entitled to launch or sign initiatives and referendums .
- *[...] un mandato che comporta prevalentemente mansioni di rappresentanza.*
vs.
The office is largely ceremonial .

2: The two spans are roughly equivalent, but some important information differs/missing

- *Pour de plus amples informations sur les tâches de la Chancellerie fédérale*
vs.
Further information on the tasks of the Swiss Federal Chancellery
- *Ha conseguito la laurea in medicina umana all'Università di Zurigo (1987) [...]*
vs.

In 1987 he earned a degree in **medicine** from the University of Zurich .

- Das Medienzentrum bietet den Bundeshausjournalisten Arbeitsplätze und die notwendige Infrastruktur für die Berichterstattung aus „**Bundesbern**“ .

vs.

The centre also provides journalists with the office space and infrastructure they need to report on political news in **Bern**.

3: The two spans are not equivalent, but share some details.

- Il **est** marié et **père de deux enfants**.

vs.

Beat Jans is married and **has two daughters**.

- Von 1989-2000 **arbeitete sie** als selbständige Übersetzerin und **Lehrbeauftragte** einer Berufsschule.

vs.

Between 1989 and 2000 **she** worked as a freelance translator, and also **taught** at a vocational school.

- Il Consiglio federale utilizza la Residenza del Lohn principalmente per ricevere **capi di Stato e di Governo** .

vs.

The Federal Council uses the Lohn country residence primarily to receive **guests of state** .

- Dall'inizio del 2021 il settore Trasformazione digitale e governance delle TIC (TDT) **fa parte delle competenze della Cancelleria federale** [...]

vs.

As of the beginning of 2021, **the Federal Chancellery is responsible for** digital transformation and ICT steering.

4: The two spans are not equivalent, but are on the same topic

- **Besonderes Gewicht legt der Bundesrat** auf Pflege und Ausbau der Beziehungen zu den Nachbarstaaten und zur EU

vs.

This involves fostering and expanding relations with neighbouring states and the EU .

- [...] **un mandato che comporta prevalentemente mansioni di rappresentanza** .

vs.

The office is largely ceremonial .

- Letzte Änderung 08.02.2021
vs.
Dernière modification 12.04.2024


5: The two spans are completely dissimilar.

- Seit 2021 war er Regierungspräsident des Kantons Basel-Stadt.
vs.
He was the president of the Basel-Stadt cantonal council from 2021 until he was elected to serve on the Federal Council in late 2023.
- Il Dipartimento federale di giustizia e polizia (DFGP) si occupa di tematiche quali i diritti civili, la sicurezza interna, l'asilo e la migrazione .
vs.
The Federal Department of Justice and Police (FDJP) deals with issues such as asylum, migration and internal security .
- He was the president of the Basel-Stadt cantonal council from 2021 until he was elected to serve on the Federal Council in late 2023.
vs.
Il a ensuite été élu au Conseil d'État du canton de Bâle-Ville, qu'il a présidé de 2021 à 2023.

Further notes

- You will be making annotations at the span level.
- A span consists of at least one word in one of the texts that you select and annotate with a difference score.
- There is no upper limit on how many words you select for one difference pair on each side and they may not be the same across languages.
- There may be cases in which a whole span is present in one text but absent in the other. In that case, select the span on the present side and do not select anything in the text in which the span is absent (see above examples for difference score 5).
- Words that are selected for a text span do not have to appear in consecutive order (see above, second example for difference score 3).
- Text that is semantically equivalent should not be annotated.

Annotation Process

Access has been provided to the following Github repository: 

Clone the branch of the repository with your name and the assigned language (German: "de", French: "fr", Italian: "it") to your computer and switch to the branch :

```
### for Windows users only ###
# 1. Open Command Prompt as Administrator
# 2. Enable long paths
fsutil behavior set SymlinkEvaluation L2L:1 R2L:1 R2R:1 L2R:1
#####

git clone --branch yourname_assigned-language
https://github.com/[REDACTED]-single-branch
```

Check whether you are on your assigned branch:

```
cd admin-ch
git config --system core.longpaths true #for Windows users only
git branch --show-current
```

The output should be:

```
yourname_language
```

Afterwards, create a new virtual environment (Python 3.11 recommended) and install the requirements:

```
pip install -r reqs.txt
```

Trial round

Once the requirements are installed, you can run the following to start the annotation interface:

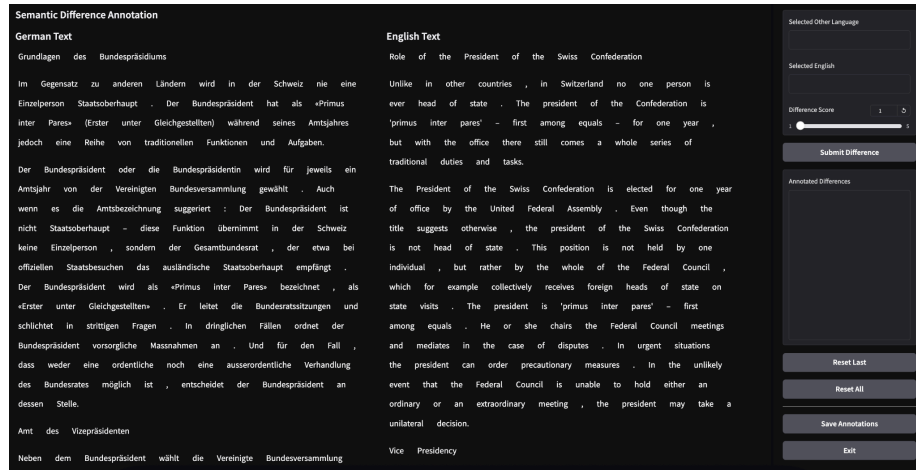
```
python scripts/collect_annotations_trial.py
```

If everything runs correctly, you should see the following message:

```
* Running on local URL: [REDACTED]
* Running on public URL: [REDACTED]
```

This share link expires in 72 hours. For free permanent hosting and GPU upgrades, run `gradio deploy` from the terminal in the working directory to deploy to Hugging Face Spaces (<https://huggingface.co/spaces>)

Open the local URL (marked in yellow) in your browser (Google Chrome recommended). You should arrive at an interface showing an English text and its counterpart in your assigned language:



In most cases, the texts will follow a similar structure, resulting in a natural alignment of paragraphs. Each paragraph, and if possible each sentence, should be compared with its counterpart in the other language, and any semantically different text spans—according to the previously defined scale—should be annotated.

However, exceptions may occur and the structure between texts within a pair might differ. Hence, it is important to first read both of the texts fully.

Clicking on a word will select it for the current difference pair and it will be shown in the boxes on the top right. Clicking the word again will deselect it. Select the words within a span from left to right. If a span crosses a paragraph boundary, split it and annotate each part as a separate span within its respective paragraph.

Once you have selected all words for a difference pair, add the label with the slider. Afterwards, push the “Submit Difference” button to save the difference and move on to selecting the next difference pair.

In case a difference pair was erroneously submitted, it can be removed by pushing “Reset Last”. If all the annotations for a text pair should be removed, “Reset All” can be used.

Once the last difference of a text pair has been submitted, the “Save Annotations” button can be pushed. This saves all annotations to a file in the background. Wait a few seconds until the launching message (shown above) is renewed in your terminal. Once renewed, refresh

the demo in your browser by reloading the page in your browser. This brings you to the next text pair.

Note that the annotations for a text pair are only saved once the “Save Annotations” button has been pressed, if you close the window before doing so the annotations for the current text pair will be lost.

If you want to take a break, finish annotating the current text pair, press “Save Annotations”, refresh the page and press the “Exit” button. This will finish the process in your console.

All annotations are stored in files located within the “annotations(_trial)” folder. These files may be used to review previous annotations. Although it is technically possible to modify the annotations manually within these files, such changes should be made only when absolutely necessary.

In the trial round, you will only annotate 3 text pairs. Once you are finished with the trial round, commit and push your “annotation_trial” folder to your branch of the repo.

```
git branch --show-current # check whether you are committing to
the correct branch
git add .
git commit -m "trial round annotations" .
git push
```

Please inform me once you finish annotating the documents in the trial round and how much time you have spent on annotating them. I will take a look at your annotations and give you feedback.

Additionally, please also use the trial round to inform me about any inconsistencies in the annotation interface that you encounter and could possibly impair your annotations or the annotation process.

Once you have received your feedback and confirmation from my side, you can continue to annotate the rest in the main annotation phase.

Main annotation phase

The main annotation phase follows the same principles as the described in trial round. But instead of running the trial round script, please run

```
python scripts/collect_annotations.py.
```

Since there are 100+ documents to annotate, push annotations to your branch after completing 5–10 text pairs even before finishing all annotations to avoid information loss.

```
git branch --show-current # check whether you are committing to
the correct branch
git add .
```

```
git commit -m "new/final batch of annotations" .  
git push
```

Inform me once you finished annotating all files and pushed the final version of your annotations-folder to the repo.

Using Automated Tools

The use of translation software such as Google Translate or DeepL to translate individual words or snippets that may not be fully understood, as well as the use of search engines to look up definitions, is generally permitted. However, large language models (e.g., ChatGPT, Claude) must not be used to automatically suggest semantic difference annotations, nor should any tools designed to automatically detect semantic differences or similarities be employed for this task.

Handling of Irregular Texts

If a text pair appears irregular—for example, if there are some structural errors noticeable or if there is no discernible overlap in meaning between the two texts—try your best to perform the annotations nonetheless, but please also copy the filenames into the designated “Potentially Faulty” Google Doc under your name and language, along with a brief comment explaining why the pair is considered irregular.

Further questions

If you have any further questions, feel free to reach out to me via email:



A.3 Evaluation

Metric	EN-DE	EN-FR	EN-IT
Total # differences annotator 1	89	87	99
Total # differences annotator 2	63	115	105
Total # differences in tokens annotator 1	430	1,082	772
Total # differences in tokens annotator 2	321	821	756
# fuzzy matched span pairs ≥ 50	31	33	43
# fuzzy matched span pairs ≥ 75	21	23	34
# fuzzy matched span pairs ≥ 90	17	15	30
Exactly matching spans	16	10	26
Corr. fuzzy matched span pairs ≥ 50	59.97	59.12	27.43
Corr. fuzzy matched span pairs ≥ 75	78.66	63.24	31.05
Corr. fuzzy matched span pairs ≥ 90	74.60	52.99	41.45
Corr. exactly matching spans	71.22	43.00	52.64
Mean IoU EN	40.75	34.80	42.42
Mean IoU OTHER	32.39	34.17	40.90
Mean F1 EN	55.81	51.61	59.02
Mean F1 OTHER	46.89	49.73	56.97

Table 6: Results from the automatic evaluation of human annotations in the trial phase.

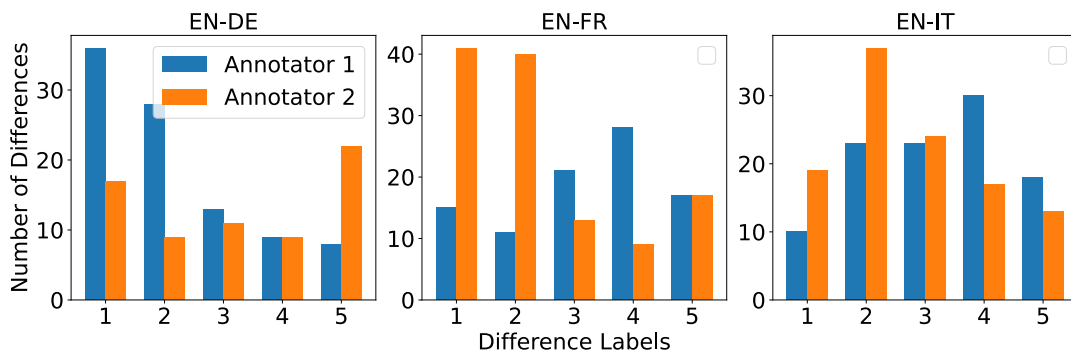


Figure 6: Label distribution for all languages by each annotator in the trial phase.

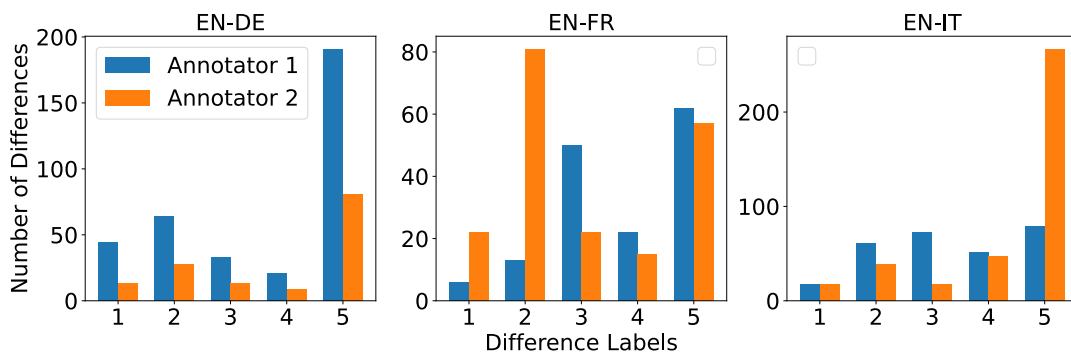


Figure 7: Label distribution for all languages by each annotator in the main phase.

A.4 Impact of Annotator Discrepancies on RSD Evaluation

Approach	EN-DE		EN-FR		EN-IT	
	ann. 1	ann. 2	ann. 1	ann. 2	ann. 1	ann. 2
<i>DiffAlign (unsupervised)</i>						
DiffAlign XLM-R SimCSE (Spearman)	0.211	0.183	0.015	0.172	0.215	0.261
DiffAlign XLM-R SimCSE (Kendall)	0.169	0.148	0.012	0.138	0.171	0.209
<i>LLMs with few-shot prompting</i>						
GPT-4o (Spearman)	0.053	0.026	0.040	-0.007	0.086	0.031
GPT-4o (Kendall)	0.049	0.025	0.039	-0.007	0.080	0.030
<i>Fine-tuned encoder models</i>						
ModernBERT (multi) (Spearman)	0.049	0.060	0.079	-0.015	0.150	0.028
ModernBERT (multi) (Kendall)	0.040	0.049	0.064	-0.012	0.119	0.023

Table 7: Spearman and Kendall correlations on SwissGov-RSD subsets by the individual annotators. Note that the results for each annotator are for the most part based on different document pairs.

The largest absolute difference of the English–French results for the two sets is 0.157 (DiffAlign XLM-R SimCSE (Spearman); 0.015 vs 0.172). We consider both results to be low values, indicating that our main conclusion—systems performing poorly on the document-level RSD task—holds for both annotators.

To make the sources of disagreement between annotators more tangible, in Figure 8 we show excerpts from the overlapping documents with annotations by both annotators. The examples show how differences can arise due to subjective annotation strategies, e.g.: for EN–DE, annotator 1 tends to label semantic differences more minimalistically, while annotator 2 labels the whole phrase in which the difference occurs.

	English	German
Annotator 1	The Federal Chancellery is responsible for legislation regulating the procedures of the government and the Federal Administration . This includes laws on the consultation procedure, publications, government and administrative organisation and on Parliament. The Federal Chancellery drafts and enforces laws in these areas.	Die Bundeskanzlei ist zuständig für Gesetze, die den Betrieb von Regierung und Verwaltung regeln. Dazu gehören das Publikationsrecht, das Regierungs- und Verwaltungsorganisationsrecht, das Vernehmlassungsrecht und das Parlamentsrecht. Die Bundeskanzlei bereitet in diesen Bereichen die Gesetze vor und vollzieht sie.
Annotator 2	The Federal Chancellery is responsible for legislation regulating the procedures of the government and the Federal Administration . This includes laws on the consultation procedure, publications, government and administrative organisation and on Parliament. The Federal Chancellery drafts and enforces laws in these areas.	Die Bundeskanzlei ist zuständig für Gesetze, die den Betrieb von Regierung und Verwaltung regeln . Dazu gehören das Publikationsrecht, das Regierungs- und Verwaltungsorganisationsrecht, das Vernehmlassungsrecht und das Parlamentsrecht. Die Bundeskanzlei bereitet in diesen Bereichen die Gesetze vor und vollzieht sie.
Annotator 1	Double digit percentage growth is anticipated in the global market for fuel cells.	Global liegt das Wachstum im Markt für Brennstoffzellen bezogen auf die installierte Leistung im zweistelligen Prozentbereich.
Annotator 2	Double digit percentage growth is anticipated in the global market for fuel cells.	Global liegt das Wachstum im Markt für Brennstoffzellen bezogen auf die installierte Leistung im zweistelligen Prozentbereich .

Figure 8: Different annotation strategies: same difference, but one annotator labels the whole phrase while the other only annotates specific words.

B Description of RSD-iSTS Datasets

Dataset	Human-annotated sentence pairs	Augmentations	Avg. tokens per input
<i>Training set for encoder models</i>			
Train (EN)	2800	10000	67.1
Validation (EN)	200	200	62.9
Train (EN-DE*)	2800	10000	65.2
Validation (EN-DE*)	200	200	67.8
Train (EN-FR*)	2800	10000	76.4
Validation (EN-FR*)	200	200	80.3
Train (EN-IT*)	2800	10000	70.8
Validation (EN-IT*)	200	200	74.5
Train (multi*)	2800	30000	70.8
Validation (multi*)	200	200	74.2
<i>Training set for LLMs</i>			
Train (EN)	2800	560	67.1
Validation (EN)	200	200	62.9
<i>Test set</i>			
iSTS	100	100	18.6
+ Negatives	100	100	30.4
+ Documents	500	100	156.6
+ Permuted	500	100	156.6
+ Cross-lingual			
EN-DE	500 [†]	100	154.8
EN-ES	500 [†]	100	164.5
EN-FR	500 [†]	100	172.6
EN-JA	500 [†]	100	111.1
EN-KO	500 [†]	100	119.2
EN-ZH	500 [†]	100	106.0
EN-IT	500 [†]	100	163.7

Table 8: Statistics of the iSTS-RSD-based (Vamvas and Sennrich, 2023) datasets used for fine-tuning and evaluation. Augmentations are random concatenations of up to 5 human-annotated sentences and do not introduce any new information. [†]The cross-lingual test sets only have annotations on the English side of the sentence pairs.

C Example of an Augmented Input

Couple sailing in a small sailboat . The birds are swimming in the water . Thai protesters launch Bangkok ‘ shutdown ’

Thai opposition protesters begin Bangkok shutdown Two ducks are standing by the water . Two people sailing a small white sail boat .

Figure 9: Example of an automatically created augmentation based on three individual, randomly selected sentence pairs. The order of sentences in the two documents has been permuted to make difference recognition more challenging. The highlights visualize the token-level gold labels.

D English–German Example for Label Projection Prompts

Annotate a pair of input sentences. The goal is to label every word in each sentence regarding its semantic similarity to the words in the other sentence. To help you with that task, you will be given the English translation and the labels for the English translation. Your task is to project the labels of the English translation onto the original sentence.

Example 1

```
sentence1_en: "Nevada : 2 dead , 2 hurt in middle school shooting"  
sentence2_en: "2 dead , 2 injured in middle school shooting in Nevada"  
labels1_en: [5, -1, 5, 5, -1, 5, 5, 5, 5, 5, 5]  
labels2_en: [5, 5, -1, 5, 5, 5, 5, 5, 5, 5, 5]
```

```
sentence1: "2 dead , 2 injured in middle school shooting in Nevada"  
sentence2: "Nevada : 2 Tote , 2 Verletzte bei Schießerei an Mittelschule in Nevada"
```

```
Output: {"text_a": "Nevada : 2 dead , 2 hurt in middle school shooting", "text_b": "2 Tote , 2  
Verletzte bei Schießerei an Mittelschule in Nevada", "labels_a": [5, -1, 5, 5, -1, 5, 5, 5, 5, 5, 5],  
"labels_b": [5, 5, -1, 5, 5, 5, 5, 5, 5, 5, 5]}
```

Example 2

```
sentence1_en: "sheep standing in a field ."  
sentence2_en: "A sheep grazing in a field ."  
labels1_en: [4, 2, 5, 5, 5, -1]  
labels2_en: [1, 4, 2, 5, 5, -1]
```

```
text_a: "sheep standing in a field ."  
text_b: "Ein Schaf grast auf einem Feld ."
```

```
Output: {"text_a": "Sheep standing in a field .", "text_b": "Ein Schaf grast auf einem Feld .",  
"labels_a": [4, 2, 5, 5, 5, -1], "labels_b": [1, 4, 2, 5, 5, -1]}
```

Example 3

```
sentence1_en: "A black dog standing in front of yellow flowers ."  
sentence2_en: "A black dog standing in a field ."  
labels1_en: [5, 5, 5, 5, 2, 2, 2, 2, 2, -1]  
labels2_en: [5, 5, 5, 5, 2, 2, 2, -1]
```

```
text_a: "A black dog standing in front of yellow flowers ."  
text_b: "Ein schwarzer Hund steht auf einem Feld ."
```

```
Output: {"text_a": "A black dog standing in front of yellow flowers .", "text_b": "Ein schwarzer Hund  
steht auf einem Feld .", "labels_a": [5, 5, 5, 5, 2, 2, 2, 2, 2, -1], "labels_b": [5, 5, 5, 5, 2, 2,  
2, -1]}
```

Score scale

- **5**: Complete equivalence
- **3-4**: Very similar or closely similar in terms of semantics.
- **1-2**: Slightly similar or somewhat similar.
- **0**: No relation (there is no word in the other sentence that is even slightly similar in terms of semantics).
- **-1**: Punctuation

Other guidelines

- Make sure to output the correct JSON format and to preserve the provided sentence tokenization.

Output only the raw JSON, without any additional text.

E Few-shot Prompt

Annotate a pair of input sentences. The goal is to label every word in each sentence regarding its semantic similarity to the words in the other sentence.

Example 1:

Input sentence 1: ["Iran", "hopes", "nuclear", "talks", "will", "yield", "", "roadmap", "']

Input sentence 2: ["Iran", "Nuclear", "Talks", "in", "Geneva", "Spur", "High", "Hopes"]

Output: {"sentence1": [{"Iran": 5}, {"hopes": 4}, {"nuclear": 5}, {"talks": 5}, {"will": 3}, {"yield": 3}, {"": -1}, {"roadmap": 2}, {"": -1}], "sentence2": [{"Iran": 5}, {"Nuclear": 5}, {"Talks": 5}, {"in": 0}, {"Geneva": 0}, {"Spur": 3}, {"High": 2}, {"Hopes": 2}]}

Example 2:

Input sentence 1: ["Books", "To", "Help", "Kids", "Talk", "About", "Boston", "Marathon", "News"]

Input sentence 2: ["Report", "of", "2", "explosions", "at", "finish", "line", "of", "Boston", "Marathon"]

Output: {"sentence1": [{"Books": 1}, {"To": 0}, {"Help": 0}, {"Kids": 0}, {"Talk": 0}, {"About": 4}, {"Boston": 4}, {"Marathon": 4}, {"News": 4}], "sentence2": [{"Report": 1}, {"of": 0}, {"2": 0}, {"explosions": 0}, {"at": 0}, {"finish": 0}, {"line": 0}, {"of": 4}, {"Boston": 4}, {"Marathon": 4}]}

Example 3:

Input sentence 1: ["Chinese", "shares", "close", "lower", "Wednesday"]

Input sentence 2: ["Chinese", "shares", "close", "higher", "Friday"]

Output: {"sentence1": [{"Chinese": 5}, {"shares": 5}, {"close": 5}, {"lower": 0}, {"Wednesday": 3}], "sentence2": [{"Chinese": 5}, {"shares": 5}, {"close": 5}, {"higher": 0}, {"Friday": 3}]}

Score scale

- **5**: Complete equivalence
- **3-4**: Very similar or closely similar in terms of semantics.
- **1-2**: Slightly similar or somewhat similar.
- **0**: No relation (there is no word in the other sentence that is even slightly similar in terms of semantics).
- **-1**: Punctuation

Other guidelines

- Make sure to output the correct JSON format and to preserve the provided sentence tokenization. Output only the raw JSON, without any additional text.

Input to Annotate

Sentence 1: {{ sentence1 }}

Sentence 2: {{ sentence2 }}

Respond with the JSON object.

F Evaluation of projected labels

Language Pair	Edited Samples	Total Samples	Edited Labels	Total Labels	% Edited Samples	% Edited Labels
EN-DE	17	50	58	3379	34	1.72
EN-IT	17	50	81	3361	34	2.41
EN-FR	23	50	137	3916	46	3.50

Table 9: Number of edited labels across language pairs.

We randomly select 50 samples per language pair from the label-projected training data and report the statistics on post-edits that we perform in Table 9. The collected statistics show that while a moderate number of full text samples (34–46%) contain at least one edit, the proportion of edited labels at the token level remains very low across all language pairs (1.72–3.50%), suggesting that the LLM-projected labels are largely reliable.

G Description of Open-weight Models

Name	Param.	Vocab.	License	Citation	URL
XLM-R+SimCSE	277M	250k	MIT	Conneau et al. (2020) ; Vamvas and Sennrich (2023)	↗
ModernBERT-large	396M	50k	Apache 2.0	Warner et al. (2025)	↗
EuroBERT 210M	210M	128k	Apache 2.0	Boizard et al. (2025)	↗
EuroBERT 610M	610M	128k	Apache 2.0	Boizard et al. (2025)	↗
EuroBERT 2.1B	2.1B	128k	Apache 2.0	Boizard et al. (2025)	↗
MmBERT-small	140M	256k	MIT	Marc et al. (2025)	↗
MmBERT-base	307M	256k	MIT	Marc et al. (2025)	↗
LaBSE	500M	500k	Apache 2.0	Feng et al. (2022)	↗
bge-m3	560M	250k	MIT	Chen et al. (2024)	↗
gte-multilingual-base	300M	129k	Apache 2.0	Zhang et al. (2024)	↗
XLM-R-XL	3.48B	250k	MIT	Goyal et al. (2021)	↗
Llama-3.1 8B Instruct	8.03B	128k	Custom license	Grattafiori et al. (2024)	↗
Llama-3.1 405B Instruct	406B	128k	Custom license	Grattafiori et al. (2024)	↗
DeepSeek-R1	685B	129k	MIT	Guo et al. (2025)	↗

Table 10: Number of parameters, vocabulary size and licensing information of the open-weight models used in this paper.

H Comprehensive iSTS-RSD Result Overview

H.1 Results for All Augmentation Categories

Approach	iSTS-RSD	+Negatives 50% paraphrases	+Documents 5 sentences	+Permuted 5 inversions	+Cross-lingual 7 language pairs
<i>DiffAlign (unsupervised)</i>					
XLM-R+SimCSE	64.4	62.8	<u>56.6</u>	<u>54.3</u>	36.3
LaBSE	62.4	64.3	49.8	49.7	35.3
bge-m3	<u>67.4</u>	<u>64.4</u>	49.5	47.9	<u>42.5</u>
gte-multilingual-base	<u>60.1</u>	<u>61.1</u>	44.8	38.5	<u>23.6</u>
ModernBERT-large	52.1	47.2	49.7	48.1	12.9
EuroBERT 210M	38.7	47.1	52.3	51.7	27.8
EuroBERT 610M	30.7	38.3	41.8	41.0	24.7
EuroBERT 2.1B	30.0	39.5	41.2	40.3	20.8
mmBERT-small	54.0	49.1	46.0	46.0	17.1
mmBERT-base	62.3	55.4	51.3	50.5	25.5
Qwen3-Embedding 4B	59.7	59.7	48.6	46.8	38.4
Qwen3-Embedding 8B	56.1	53.2	48.1	45.3	37.1
<i>LLMs with few-shot prompting</i>					
Llama-3.1 8B Instruct	44.1	38.2	12.5	12.5	3.3
Llama-3.1 405B Instruct	60.6	63.3	57.8	55.2	22.8
GPT-4o-mini	54.6	60.5	38.0	26.2	12.4
GPT-4o	<u>61.1</u>	<u>64.9</u>	<u>60.7</u>	<u>61.2</u>	34.8
DeepSeek-R1	<u>57.3</u>	<u>62.7</u>	56.8	53.9	30.6
o3-mini-low	59.6	64.7	58.0	58.5	<u>36.6</u>
<i>Fine-tuned LLMs</i>					
Llama-3.1 8B Instruct	78.0	87.6	80.9	81.0	49.2
GPT-4o-mini	<u>85.9</u>	<u>92.2</u>	<u>88.7</u>	<u>87.9</u>	<u>62.8</u>
<i>Fine-tuned encoder models</i>					
ModernBERT-large (EN)	86.9	<u>84.0</u>	<u>81.0</u>	<u>81.3</u>	51.8
ModernBERT-large (EN-DE*)	81.5	69.4	65.0	65.2	49.3
ModernBERT-large (EN-FR*)	81.9	70.9	69.2	69.6	<u>52.2</u>
ModernBERT-large (EN-IT*)	82.6	72.3	68.7	69.0	48.3
ModernBERT-large (multi*)	79.2	70.0	67.2	67.4	45.4

Table 11: Comparison of different models and approaches on iSTS-RSD (Vamvas and Sennrich, 2023). The table reports token-level Spearman correlations between predicted and gold labels. The variations described in the column headers are cumulative: the rightmost column refers to a cross-lingual test set of permuted documents containing negative examples. The cross-lingual results are averages from the results in Table 12.

H.2 Results For All iSTS-RSD Language Pairs

Approach	EN-DE	EN-FR	EN-IT	EN-ES	EN-JA	EN-KO	EN-ZH
<i>DiffAlign (unsupervised)</i>							
XLM-R+SimCSE	44.9	45.2	44.9	<u>47.2</u>	15.9	33.1	23.1
LaBSE	40.6	42.3	45.3	44.3	20.0	27.8	26.7
bge-m3	<u>47.1</u>	<u>45.8</u>	<u>47.5</u>	46.3	<u>32.4</u>	<u>42.0</u>	<u>36.0</u>
gte-multilingual-base	31.2	30.2	30.5	31.5	14.9	18.7	7.9
ModernBERT-large	17.3	16.7	17.2	16.4	4.2	12.4	6.1
EuroBERT 210M	32.1	34.6	33.4	38.1	14.7	24.0	17.4
EuroBERT 610M	29.1	30.5	31.0	31.4	17.1	17.2	16.9
EuroBERT 2.1B	40.3	28.5	28.3	30.2	11.8	9.3	12.4
mmBERT-small	23.0	23.5	22.1	26.7	5.8	9.6	9.3
mmBERT-base	34.0	32.6	32.6	35.6	11.8	18.0	13.6
Qwen3-Embedding 4B	41.6	42.9	43.2	42.8	29.1	37.5	31.6
Qwen3-Embedding 8B	40.9	40.1	41.4	40.8	30.3	34.2	32.2
<i>LLMs with few-shot prompting</i>							
Llama-3.1 8B Instruct	2.8	3.7	2.6	7.7	-2.1	7.0	1.7
Llama-3.1 405B Instruct	29.2	29.3	26.9	29.3	18.2	18.4	8.0
GPT-4o-mini	15.8	13.8	16.6	14.8	2.9	13.9	9.0
GPT-4o	43.0	40.5	42.5	39.8	<u>25.0</u>	<u>26.2</u>	26.6
DeepSeek-R1	38.2	40.1	34.6	39.8	<u>19.3</u>	19.9	22.0
o3-mini-low	<u>44.8</u>	<u>46.5</u>	<u>48.2</u>	<u>48.5</u>	21.9	19.2	<u>27.1</u>
<i>Fine-tuned LLMs</i>							
Llama-3.1 8B Instruct	66.7	67.3	66.9	68.4	19.0	29.5	26.4
GPT-4o-mini	81.6	79.9	78.2	82.0	<u>35.4</u>	<u>39.6</u>	<u>41.5</u>
<i>Fine-tuned encoder models</i>							
ModernBERT-large (EN)	55.3	55.4	53.8	55.4	45.9	48.8	48.2
ModernBERT-large (EN-DE*)	58.8	58.4	58.5	58.4	32.1	43.2	37.9
ModernBERT-large (EN-FR*)	58.4	60.7	60.1	<u>61.2</u>	35.8	45.9	40.7
ModernBERT-large (EN-IT*)	<u>59.2</u>	<u>61.2</u>	<u>63.0</u>	60.7	27.1	37.4	34.0
ModernBERT-large (multi*)	54.7	57.5	58.1	57.5	25.3	35.4	30.3

Table 12: Token-level Spearman correlations between predicted and gold labels for all individual language pairs covered by the iSTS-RSD. (*) denotes encoders fine-tuned on data with projected labels, **bold** the best performance overall, and underline best performance within model category.

H.3 LLMs Fail to Produce Labels

Approach	iSTS-RSD
<i>LLMs with few-shot prompting</i>	
Llama-3.1 8B Instruct	10.5%
Llama-3.1 405B Instruct	2.0%
GPT-4o	0.6%
<i>Fine-tuned LLMs</i>	
Llama-3.1 8B Instruct	0.6%
GPT-4o-mini	1.3%

Table 13: Percentage of samples for which LLMs failed to produce the correct number of labels for all iSTS-RSD predictions.

I Kendall τ -b Results

Approach	EN-DE	EN-FR	EN-IT
<i>DiffAlign (unsupervised)</i>			
DiffAlign XLM-R+SimCSE	15.5	8.1	18.7
DiffAlign ModernBERT	1.1	0.6	2.1
<i>LLMs with few-shot prompting</i>			
Llama-3.1 405B Instruct	-0.8	-2.2	7.1
GPT-4o	4.0	1.2	5.1
GPT-4o-mini	3.8	1.1	5.1
<i>Fine-tuned encoder models</i>			
ModernBERT-large (EN)	3.8	1.1	3.2
ModernBERT-large (EN-DE*)	5.2	1.3	4.5
ModernBERT-large (EN-FR*)	4.3	4.2	7.0
ModernBERT-large (EN-IT*)	6.1	3.9	10.0
ModernBERT-large (multi*)	4.6	2.0	7.1

Table 14: Kendall τ -b scores on the SwissGov-RSD task (equivalent to the Spearman correlation in Table 3). The Kendall scores are generally lower than Spearman scores, but model ranking stays consistent.

J Inference Time Comparison

Model	Parameters	Short sequence pair (s)	Long sequence pair (s)
ModernBERT-large (DiffAlign)	396M	0.082	0.251
ModernBERT-large (fine-tuned)	396M	0.072	0.267
Llama-3.1 8B Instr. (few-shot)	8.03B	6.400	341.900

Table 15: Comparison of inference time when processing either a short sentence pair or a long sequence pair with different approaches. The table reports average time of 100 inferences for encoder models and 10 inferences for LLMs in seconds measured on a single (encoder) or eight (LLM) NVIDIA GeForce RTX 4090. The short sequence pair has a total token count (separated by whitespaces) of 21, while the long sentence pair counts 962 tokens.