

Examining Multilingual Embedding Models Cross-Lingually Through LLM-Generated Adversarial Examples

Andrianos Michail Simon Clematide Rico Sennrich

University of Zurich

{andrianos.michail, simon.clematide, sennrich}@cl.uzh.ch

Abstract

The evaluation of cross-lingual semantic search models is often limited to existing datasets from tasks such as information retrieval and semantic textual similarity. We introduce Cross-Lingual Semantic Discrimination (CLSD), a lightweight evaluation task that requires only parallel sentences and a Large Language Model (LLM) to generate adversarial distractors. CLSD measures an embedding model’s ability to rank the true parallel sentence above semantically misleading but lexically similar alternatives. As a case study, we construct CLSD datasets for German–French in the news domain. Our experiments show that models fine-tuned for retrieval tasks benefit from pivoting through English, whereas bitext mining models perform best in direct cross-lingual settings. A fine-grained similarity analysis further reveals that embedding models differ in their sensitivity to linguistic perturbations.¹

1 Introduction

When selecting a model for semantic search in a specific domain and language pair, standard embedding benchmarks such as (Multilingual) MTEB (Muennighoff et al., 2023; Enevoldsen et al., 2025) provide only partial coverage. We argue that evaluation should be carried out directly on the target language pair and domain of interest. However, the lack of suitable datasets makes it difficult to reach well-founded decisions about applications, such as choosing an embedding model or deciding whether to use English as a pivot language.

We propose a simple and efficient method for dataset generation and evaluation that simulates real-world challenges of cross-lingual semantic retrieval in large text collections. Our approach uses LLMs to create adversarial examples from parallel

sentences, designed to test the cross-lingual embedding capabilities of multilingual models. Concretely speaking, four distractor sentences are generated for each parallel target sentence. We require these sentences to be very similar to the original target sentence in terms of syntactic structure and surface word forms, but to be semantically dissimilar. Our proposed task, Cross-Lingual Semantic Discrimination (CLSD), involves identifying the correct target sentence from among four distractor sentences in the target language, given an original sentence in the source language.

To better understand how these distractor modifications influence semantic similarity both cross-lingually and monolingually, we perform a fine-grained analysis and correlate the semantic changes with linguistic properties at the level of part-of-speech tags.

Contributions

- We propose a new task that emulates cross-lingual semantic search within large text collections, requiring only a set of parallel sentences for dataset creation.
- We publish four such datasets in the news domain for the German–French language pair.
- We show that some embedding models perform better in direct cross-lingual retrieval, while others benefit from using English as a pivot language.
- We provide a linguistically informed analysis that offers insights into cross- and monolingual semantic representations of multilingual models.

2 Related Work

Cross-Lingual Semantic Search Cross-lingual semantic search is an umbrella term for tasks that

¹We release our code and datasets under AGPL-3.0: https://github.com/impresso/cross_lingual_semantic_discrimination

Original Sentence S	Distractor Sentence 1 D_1	Distractor Sentence 2 D_2
Die Linkspartei beschließt in Bonn ihr Programm zur Europawahl. <i>(The Left Party adopts its program for the European elections in Bonn.)</i>	Die Linkspartei beschließt in Bonn ihr Programm zur Bundestagswahl . <i>(The Left Party adopts its program for the Bundestag elections in Bonn.)</i>	Die Linkspartei verweigert in Bonn ihr Programm zur Europawahl. <i>(The Left Party refuses its program for the European elections in Bonn.)</i>
Là-bas, la perspective de la fin de la guerre reste toujours très éloignée. <i>(Over there, the prospect of the end of the war is still a long way off.)</i>	Là-bas, la fin de la guerre semble toujours très incertaine . <i>(Over there, the end of the war still seems very uncertain.)</i>	Là-bas, la perspective de la paix semble toujours très éloignée. <i>(Over there, the prospect of peace still seems a long way off.)</i>

Table 1: Examples of generated distractor sentences in German and French, with English translations provided in italics. Text in red indicates the modified segments.

involve searching for texts with similar or relevant semantics in another language. Examples of such cross-lingual tasks are information retrieval (CLIR) (Lawrie et al., 2023, 2024), question answering (CLQA) (Lewis et al., 2020), semantic text similarity (X-STS) (Cer et al., 2017), and bitext mining (Research; Zweigenbaum et al., 2017). As showcased in MMTEB (Enevoldsen et al., 2025), multilingual embedding models vary in their performance for different cross-lingual semantic search tasks and language pairs. When selecting models for a use case, we often rely on results from other domains or language pairs, as direct evaluation results are not always available.

Distractor Generation PAWS-X (Yang et al., 2019; Zhang et al., 2019) generated adversarial examples by rule-based and backtranslation methods to measure the limitations of paraphrase identification models. More recently, XSim++ (Chen et al., 2023) used rule-based text augmentation to generate synthetic examples for tuning LASER models (replacements by antonyms, swapping entities). Within semantic search, InPars (Bonifacio et al., 2022) relies solely on in-domain synthetic data to fine-tune retrieval models that surpass strong baselines. Lastly, Li et al. (2025) used AMR (Banarescu et al., 2013) to craft hard negatives and examine the robustness of English embedding models.

Bitext Mining Bitext Mining is an established and most similar cross-lingual semantic search task to our proposed CLSD task. Bitext mining (Research; Zweigenbaum et al., 2017) challenges embedding models to identify an exact parallel sentence from a large pool—often consisting of thousands of unrelated or loosely related sentences—originating from large, homogeneous corpora.

3 Experiments

3.1 Cross-Lingual Semantic Discrimination

We propose a task that aims to evaluate multilingual embedding methods based on their capacity to find the closest cross-lingual semantic match amidst large bilingual text collections. The goal is to measure the model’s ability to identify a parallel sentence as the most similar to the original sentence in the source language. To accomplish this task, a dataset of parallel sentences is required that will be enriched with four distractor sentences in the target language, all of which are aimed to be semantically similar.

In this paper, we test the hypothesis that LLMs can generate difficult examples that look similar on the surface, but convey different meanings. We prompt GPT-4 (gpt-4-0613) (OpenAI, 2023) to monolingually generate four distractors for each original sentence (see Prompt 1). The distractors are requested to be structurally and lexically similar, whilst being semantically dissimilar. For each language, we manually examine 200 distractors of the original sentences and find that over 98% of the distractors meet these criteria. Example sentences are shown in Table 1.

CLSD Datasets We generate four datasets and validate samples from each, consisting of both language directions of German–French for the WMT19 (Barrault et al., 2019) and WMT21 (Akhbardeh et al., 2021) test sets. In the 18,928 generated distractors, we observe that the LLM used different modification strategies such as single polarity inversion, entity replacement or even multiple swaps with word-level Jaccard similarity between originals and distractors at ($\mu = 0.47$, $\sigma = 0.17$). Table 5 shows distractor samples.

CLSD Evaluation To evaluate a multilingual embedding model E , we assess its ability to discrimi-

Model (Hugging Face Name)	Alignment			WMT19		WMT21		Average
	Cross-lingual	Paraphrase	Retrieval	DE→FR	FR→DE	DE→FR	FR→DE	
Cross-Lingual CLSD Evaluation								
multilingual-e5-base	✓	✓	✓	91.51	88.46	86.34	81.97	87.07
multilingual-e5-large	✓	✓	✓	94.50	91.51	91.38	87.57	91.24
gte-multilingual-base	✓	✓	✓	90.22	89.68	90.48	91.60	90.50
paraphrase-multilingual-mpnet-base	✓	✓	✗	91.31	91.11	91.15	92.95	91.63
sentence-transformers/LaBSE	✓	✓	✗	95.18	94.30	94.06	94.18	94.43
English Pivot Translation by M2M 1.2B (Fan et al., 2021) CLSD Evaluation								
multilingual-e5-base	✓	✓	✓	90.50	89.55	89.81	93.39	90.81
multilingual-e5-large	✓	✓	✓	90.22	89.95	91.83	92.95	91.24
gte-multilingual-base	✓	✓	✓	89.68	89.07	90.26	92.95	90.49
paraphrase-multilingual-mpnet-base	✓	✓	✗	87.92	87.64	88.13	91.15	88.71
sentence-transformers/LaBSE	✓	✓	✗	90.84	90.16	91.83	94.40	91.81

Table 2: Precision@1 results on the Cross-Lingual Semantic Discrimination (CLSD) datasets derived from the WMT19 and WMT21 German–French test sets. For each dataset, distractors are added only in the target language, yielding one evaluation direction per set (DE→FR or FR→DE). The upper block reports direct cross-lingual evaluation, the lower block evaluation via English pivot translation. Columns under *Alignment* indicate whether the model was trained or fine-tuned with cross-lingual, paraphrase, or retrieval objectives. *Average* is the mean Precision@1 across all four evaluation sets.

Distractor Generation Prompt

Can you provide me with four tricky sentences (numbered) that look structurally and lexically similar but don't have the same meaning. The sentences should be within similar topics and share commonalities with the original. Answer in {German/French}!
{Original Sentence in German/French}

Prompt 1: Distractor generation prompt for GPT-4

nate between a source sentence S , its parallel target sentence T , and a set of distractors D . For a given source sentence S and its corresponding translation T in the target language, along with a set of four distractor sentences D , the model must produce an embedding of S that is more similar to the embedding of T than to that of any distractor $d_i \in D$.

To measure performance of embedding models, we use Precision@1 (P@1), defined as the proportion of samples in which the model ranks the true translation as more similar to the source sentence than the four distractors. For example, in Table 2, the column WMT19 DE→FR reports the P@1 of identifying the original French target sentence as more similar than its French distractors.

Models Evaluated We examine the following five embedding models available through Sentence-Transformers (Reimers and Gurevych, 2019).

multilingual-e5-base/-large (M-E5-b/l) Wang et al. (2024); Wang et al. (2024) transform XLM-RoBERTa (Conneau et al., 2020) into a bi-encoder through weakly supervised contrastive pre-training with multilingual data and a second step of supervised fine-tuning on retrieval datasets.

gte-multilingual-base (M-GTE) Zhang et al. (2024) pre-train a multilingual long-context encoder (8192) and derive a bi-encoder by a process similar to multilingual-e5 that also includes contrastive training against hard negatives.

paraphrase-multilingual-mpnet-base-v2 (M-MP) Reimers and Gurevych (2020) use knowledge distillation through cosine loss training on parallel sentences between an English paraphrase trained teacher model (paraphrase-mpnet-base-v2) and a multilingual student model (XLM-RoBERTa (Conneau et al., 2020)) to teach the student to represent sentences closely across multiple languages.

LaBSE Language-agnostic BERT Sentence Embedding (Feng et al., 2022) is trained with translation ranking loss and negative samples.

3.2 CLSD Results

The results of the models on the CLSD datasets are shown in Table 2. Regarding the direct cross-lingual evaluation, the Bitext Mining specialist LaBSE is the best model in this task with an average score of 94.43 (P@1), demonstrating its

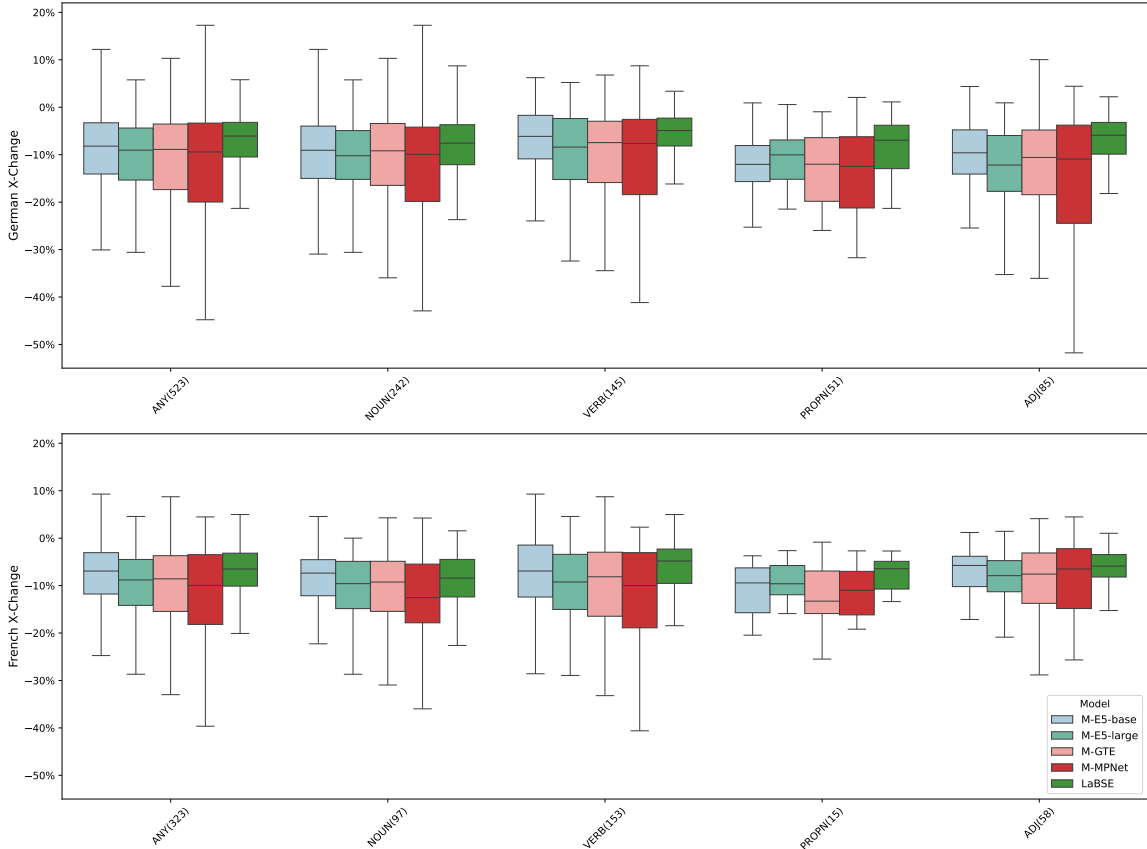


Figure 2: Change in cross-lingual cosine similarity between original and distractor sentence pairs with exactly one token swapped, grouped by the part of speech of the differing token.

consistent ability to rank the true translation T higher than any distractor $d_i \in D$. Another pattern that emerges is that models (M-E5-b/l, M-GTE) that were also trained on English-centric retrieval datasets, such as the MS MARCO (Nguyen et al., 2016), perform worse within our stricter task.

Pivot Evaluation Setup For comparison, we also evaluate models in an English-pivot setting. For each CLSD dataset, we translate both the source sentence and its candidate set (true translation plus distractors) into English using the M2M 1.2B model (Fan et al., 2021). The evaluation is then performed monolingually in English. For example, in the WMT19-DE→FR dataset, the German source sentence and all five French candidates are translated into English, and the model must rank the true English translation above its distractors. This yields four additional evaluation setups corresponding to the four CLSD datasets.

In the pivot evaluation, both LaBSE and M-MP show a performance decrease of 2.6/2.8 points compared to direct cross-lingual evaluation. In contrast, models with weaker performance in direct cross-

lingual show only minor improvements, except the M-E5-b model, which gains 3.7 points.

We conduct a qualitative analysis of the 279 samples where LaBSE succeeds in the direct cross-lingual evaluation but fails in the pivot-through-English evaluation. Our analysis reveals that, in most cases, language-specific terms are translated into broader and common English words, reducing the distinction between the original texts and the distractors. Less frequently, we observe omitted details in the translations, hallucinations, and, rarely, poor quality translations.

3.3 Fine-grained similarity analysis

How do minor surface changes affect the semantic similarity between sentence pairs? Can linguistic analysis provide a more precise characterisation of the changes? To address these questions, we pick original-distractor pairs (T, d_i) in which the distractor d_i differs from T by only a single word. This subset includes 523 German pairs (5.5%) and 324 French pairs (3.4%). Using Stanza (v 1.8) (Qi et al., 2020) and further validation by a linguist annotator, we assign part-of-speech tags to the differ-

ing token. We measure the cross-lingual semantic as the difference between $\cos(E(S), E(T))$ and $\cos(E(S), E(d_i))$. To improve comparability in our analysis across different models, we normalize² this score by dividing with $\|-\|_{\text{DE} \leftrightarrow \text{FR}}$ for the corresponding model. Formally, we measure:

$$\frac{(1 - \cos(E(S), E(T))) - (1 - \cos(E(S), E(d)))}{\|-\|_{\text{DE} \leftrightarrow \text{FR}}}$$

Figure 2 presents model results, showing that the M-MP model experiences the largest overall shifts in cosine similarity when single-word modifications are made. The column ‘‘ANY’’ reflects changes across all items, facilitating within-model comparisons to determine which linguistically characterized subgroups exhibit more pronounced changes. Notably, LaBSE, the best performing model within CLSD, represents single-word changes with relatively uniform similarity shifts across linguistic categories, unlike the other models. Additionally, we observe that replacing proper nouns consistently decreases semantic similarity more than replacing other parts of speech across all languages and models.

Mono- and Cross-Lingual Change Correlations

How language-agnostic are the base models when dealing with distractors? Does it matter whether the source sentence is presented in the source or target language? We measure the monolingual similarity change of the original to the distractor by substituting $E(S)$ with $E(T)$ in the calculation. The corresponding monolingual semantic change plots can be found in Figure 3 in the Appendix. By correlating the results of the cross- and monolingual change for all one-word change cases (i.e. ‘‘ANY’’): The correlation between monolingual and cross-lingual changes is strong for M-MP with 0.93 closely followed by M-GTE with 0.85 and LaBSE with 0.82. The M-E5-b/l shows lower correlation with 0.79 and 0.71 respectively. However, specific nuances depend on the language, model, or type of replaced word. For example, M-GTE shows a low correlation of 0.64 for proper nouns between cross- and monolingual evaluations in French, compared to 0.77 for German.

3.4 Successful Distractors Analysis

We conduct an analysis that examines the distribution of successful distractors $d_i \in D$ with respect

²The average cosine similarity difference between parallel and unrelated cross-lingual sentence pairs in the original WMT19/21 datasets is the normalization factor in Figure 2.

Lev Sim.	$ D_{\text{bin}} $	M-E5-b	M-E5-l	M-GTE	M-MP	LaBSE
French						
0.9–1.0	1120	25.9%	20.1%	26.6%	30.4%	32.3%
0.8–0.9	2138	27.1%	23.9%	31.4%	24.2%	28.2%
0.7–0.8	2951	31.6%	36.5%	26.2%	28.5%	25.0%
0.6–0.7	2394	12.6%	16.4%	12.7%	13.0%	12.1%
0.3–0.6	861	2.8%	3.1%	3.1%	3.9%	2.4%
German						
0.9–1.0	1523	28.4%	28.7%	30.6%	28.9%	40.4%
0.8–0.9	2534	32.9%	32.5%	33.2%	35.1%	32.4%
0.7–0.8	2914	25.4%	23.6%	22.7%	22.2%	19.9%
0.6–0.7	1890	10.3%	11.8%	10.5%	10.3%	5.9%
0.3–0.6	690	3.0%	3.4%	3.1%	3.6%	1.5%

Table 3: Successful distractors for each model, grouped by Levenshtein similarity to the original sentence. Values show the fraction of successful distractors (%). $|D_{\text{bin}}|$ gives the total number of distractors in each bin.

to their Levenshtein similarity to the true translation T . Specifically, we binned the distractors per model according to their similarity score and report the relative percentage per bin. Full results are available in Table 3.

Across all models, the majority of successful distractors fall into high-similarity bins (especially 0.8–0.99), confirming that minimal changes to the true translation T can suffice to mislead the models. However, LaBSE stands out with a relatively lower proportion of successful distractors in lower similarity bins, particularly for German. This suggests a greater robustness to more substantial perturbations, potentially explaining its superior performance on our task.

4 Conclusion

This paper advances cross-lingual semantic search by identifying gaps in current evaluation practices and introducing adversarial examples for testing multilingual embeddings. In a German–French case study, we contribute four new adversarial cross-lingual evaluation datasets in the news domain. Our comparative experiments highlight that direct cross-lingual and pivot-through-English evaluation each have advantages depending on the embedding model. Fine-grained analyses further demonstrate how generated distractors reveal differences in the cross-lingual capabilities of multilingual models and show how small perturbations can affect their semantic representations. Future work should extend this approach to additional language pairs and explore more refined adversarial generation methods to improve the evaluation of semantic search models.

Limitations

Our pipeline enables cross-lingual evaluation of semantic search for any language pair with available parallel texts. However, we evaluate only one language pair and rely on a single LLM for distractor generation, whereas simpler methods might suffice. Automatic prediction of part-of-speech tags and named entities in our fine-grained analysis can introduce noise, although manual validation reduces this risk. However, the capabilities of current models in the language of interest should be critically assessed. Moreover, the evaluation covers only a few models due to space constraints; nevertheless, smaller models could also yield valuable insights. Future work should address these limitations by broadening the range of languages, models, and adversarial generation techniques, ultimately supporting large-scale evaluation across many languages, including low-resource ones.

Acknowledgments

This research is conducted under the project *Impresso – Media Monitoring of the Past II. Beyond Borders: Connecting Historical Newspapers and Radio*. Impresso is a research project funded by the Swiss National Science Foundation (SNSF 213585) and the Luxembourg National Research Fund (17498891). We would also like to thank our colleagues Michelle Wastl and Juri Opitz for their support with annotations and further discussions.

References

- Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. [Findings of the 2021 conference on machine translation \(WMT21\)](#). In *Proceedings of the Sixth Conference on Machine Translation*, pages 1–88, Online. Association for Computational Linguistics.
- Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. [Abstract Meaning Representation for sembanking](#). In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.
- Loïc Barrault, Ondřej Bojar, Marta R. Costa-jussa, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. [Findings of the 2019 conference on machine translation \(WMT19\)](#). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Florence, Italy. Association for Computational Linguistics.
- Luiz Bonifacio, Vitor Jeronimo, Hugo Queiroz Abonizio, Israel Campiotti, Marzieh Fadaee, Roberto Lotufo, and Rodrigo Nogueira. 2022. [mMARCO: A multilingual version of the MS MARCO passage ranking dataset](#). *Preprint*, arXiv:2108.13897.
- Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. [SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation](#). In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.
- Mingda Chen, Kevin Heffernan, Onur Çelebi, Alexandre Mourachko, and Holger Schwenk. 2023. [xSIM++: An improved proxy to bitext mining performance for low-resource languages](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 101–109, 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.
- Kenneth Enevoldsen, Isaac Chung, Imene Kerboua, Márton Kardos, Ashwin Mathur, David Stap, Jay Gala, Wissam Siblini, Dominik Krzemiński, Genta Indra Winata, et al. 2025. [MMTEB: Massive multilingual text embedding benchmark](#). In *The Thirteenth International Conference on Learning Representations*.
- Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Michael Auli, and Armand Joulin. 2021. [Beyond english-centric multilingual machine translation](#). *Journal of Machine Learning Research*, 22(107):1–48.

- Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Ariavazhagan, 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.
- Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2023. [Overview of the TREC 2022 NeuCLIR track](#). *Preprint*, arXiv:2304.12367.
- Dawn Lawrie, Sean MacAvaney, James Mayfield, Paul McNamee, Douglas W. Oard, Luca Soldaini, and Eugene Yang. 2024. [Overview of the trec 2023 neuclir track](#). *Preprint*, arXiv:2404.08071.
- Patrick Lewis, Barlas Oguz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. [MLQA: Evaluating cross-lingual extractive question answering](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7315–7330, Online. Association for Computational Linguistics.
- Hongji Li, Andrianos Michail, Reto Gubelmann, Simon Clematide, and Juri Opitz. 2025. [Sentence smith: Controllable edits for evaluating text embeddings](#). *EMNLP 2025*.
- Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. [MTEB: Massive text embedding benchmark](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.
- Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. [MS MARCO: A human generated machine reading Comprehension dataset](#). *CoRR*, abs/1611.09268.
- OpenAI. 2023. [GPT-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A Python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.
- 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, Hong Kong, China. Association for Computational Linguistics.
- Nils Reimers and Iryna Gurevych. 2020. [Making monolingual sentence embeddings multilingual using knowledge distillation](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.
- Facebook Research. [Tatoeba multilingual bitext mining test set](#).
- Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2024. [Text embeddings by weakly-supervised contrastive pre-training](#).
- Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. [Multilingual e5 text embeddings: A technical report](#). *arXiv e-prints*, arXiv:2402.05672.
- 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, Meishan Zhang, Wenjie Li, and Min Zhang. 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, Miami, Florida, US. Association for Computational Linguistics.
- 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.
- Pierre Zweigenbaum, Serge Sharoff, and Reinhard Rapp. 2017. [Overview of the second BUCC shared task: Spotting parallel sentences in comparable corpora](#). In *Proceedings of the 10th Workshop on Building and Using Comparable Corpora*, pages 60–67, Vancouver, Canada. Association for Computational Linguistics.

A Appendix

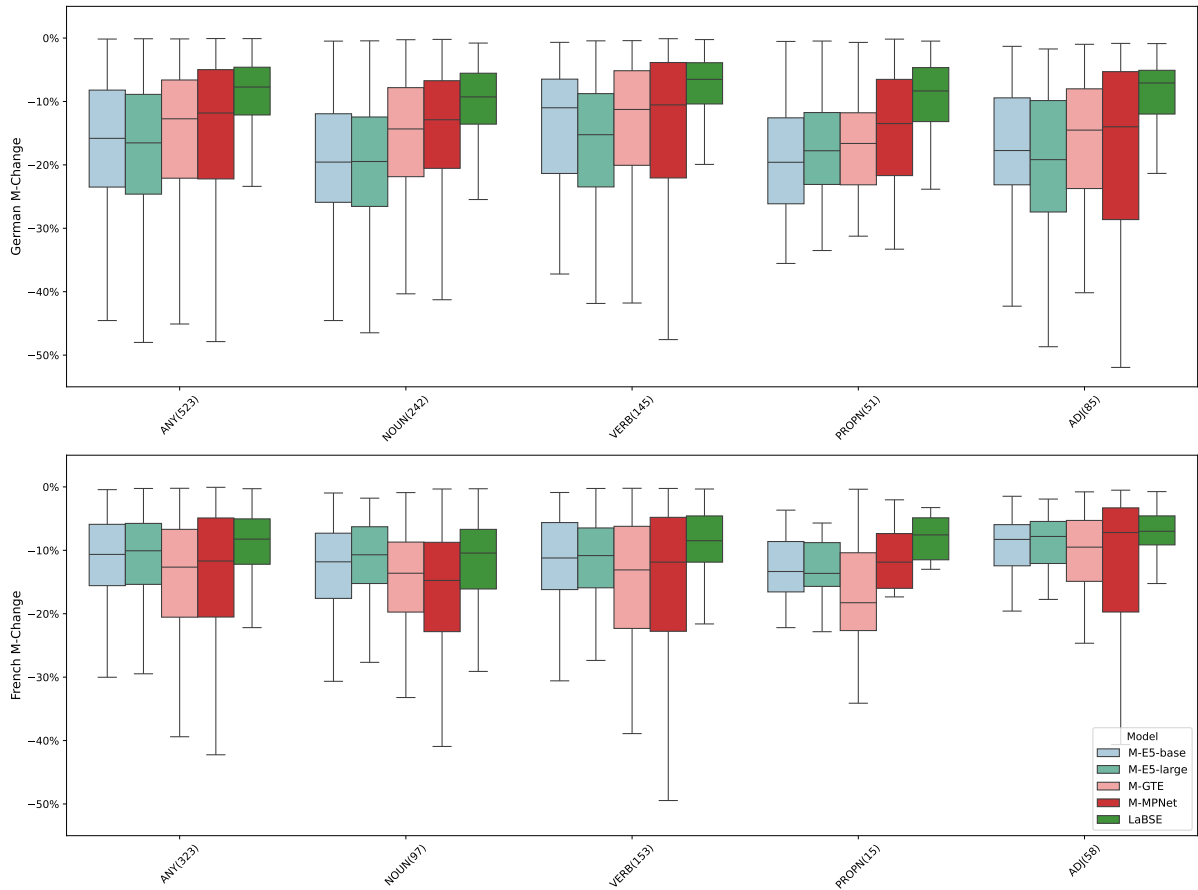


Figure 3: Monolingual cosine similarity change in original-distractor sentence pairs with exactly one token swapped, grouped by the part of speech of the differing token.

Model Parameter	Value
Temperature	1.0
Top-P	1.0

Table 4: Parameters used for GPT-4 via chat completion for the distractor generation. Undefined parameters are the default.

Distractor	InDistr Jaccard Sim.
<i>Original: PARIS (Reuters) - Die New Yorker Börse schloss am Freitag mit 0,68 % im Minus und beendete die Woche mit einem allgemeinen Rückgang, da sie von enttäuschenden Ergebnissen, einem Wiederanstieg der Coronavirus-Infektionen und geopolitischen Unsicherheiten beeinträchtigt wurde.</i>	/
Adv1: LONDON (Reuters) - Die Londoner Börse schloss am Dienstag mit 0,68 % im Plus und startete die Woche mit allgemeinem Wachstum , da sie von erfreulichen Ergebnissen, einem kontinuierlichen Rückgang der Coronavirus-Infektionen und geopolitischen Stabilitäten angetrieben wurde.	0.434
Adv2: BERLIN (Reuters) - Die Berliner Börse schloss am Donnerstag mit 0,68 % im Minus und beendete die Woche mit einem insgesamt flachen Verlauf, das durch gemischte Ergebnisse, eine Stagnation der Coronavirus-Infektionen und geopolitische Spannungen beeinflusst wurde.	0.300
Adv3: TOKYO (Reuters) - Die Tokioter Börse schloss am Mittwoch mit 0,68 % im Plus und begann den Monat mit allgemeiner Erholung , da sie von positiven Ergebnissen, einem eutlichen Rückgang der Coronavirus-Infektionen und geopolitischer Sicherheit profitierte .	0.359
Adv4: MADRID (Reuters) - Die Madriдер Börse schloss am Montag mit 0,68 % im Minus und startete die Woche mit allgemeinem Rückgang , da sie von enttäuschenden Ergebnissen, einem Wiederanstieg der Coronavirus-Infektionen und geopolitischen Unsicherheiten beeinträchtigt wurde.	0.438
<i>Original: Der Nasdaq verzeichnete die schlechteste Woche der letzten vier.</i>	/
Adv1: Der Nasdaq hat die beste Woche der letzten vier verzeichnet .	0.630
Adv2: Der Nasdaq verzeichnete die aktivste Woche der letzten vier.	0.623
Adv3: Der Nasdaq hat die ruhigste Woche der letzten vier verzeichnet .	0.630
Adv4: Der Nasdaq verzeichnete die turbulenteste Woche der letzten vier.	0.623
<i>Original: Die Beamten werden im Haushalt für 2021 jedoch nicht vergessen werden.</i>	/
Adv1: Die Beamten werden im Haushalt für 2021 jedoch nicht besprochen werden.	0.818
Adv2: Die Beamten werden im Haushalt für 2021 jedoch nicht entlassen werden.	0.818
Adv3: Die Beamten werden im Haushalt für 2021 jedoch nicht befördert werden.	0.818
Adv4: Die Beamten werden im Haushalt für 2021 jedoch nicht belastet werden.	0.818

Table 5: Original-Distractor sets with their Intra-Distractor Jaccard Similarity. Red font indicates modified text.

Original	Distractor	Tricked Model
WMT19- DE->FR-CLSD		
Kipping au congrès de die Linke sur l'Europe : l'Europe est depuis longtemps un continent d'immigration.	Kipping au congrès de die Linke sur l'Asie : l'Asie est depuis longtemps un continent de diversité.	MPNet-M
Par exemple, nous allons à présent recevoir un quartier général pour les munitions de l'UE, à Bruxelles, ce que les Britanniques ont jusqu'ici toujours empêché.	Par exemple, nous prévoyons de lancer un bureau principal pour les ressources de l'UE, à Bruxelles, ce que les Français ont toujours refusé jusqu'ici.	MPNet-M
L'appel de Macron: l'Europe est davantage qu'un "projet"	L'appel de Macron: l'Europe est plus qu'une "idée".	M-E5-b
Trois commissions ont accepté ma candidature, l'une d'elle était de justesse contre.	Trois comités ont rejeté ma candidature, l'un d'eux était presque en faveur.	M-E5-b
WMT19- FR->DE-CLSD		
Würden sie dies tun, gäbe es niemanden in Europa, der sich dagegen wehren würde.	Würden sie dies tun, gäbe es niemanden in Europa, der sich dafür aussprechen würde.	MPNet-M M-E5-b
Um dem etwas entgegenzusetzen, ist die EU-Kommission auf Kooperation mit den Internetriesen angewiesen.	Um das zu fördern, ist die EU-Kommission auf Zusammenarbeit mit den Internetgiganten angewiesen.	MPNet-M M-E5-b
Und in der Sozialpolitik sehe ich bei ihr ebenfalls keinen Rechtsschwenk.	Und bei ihr sehe ich in der Sozialpolitik ebenfalls keinen Linksschwenk.	MPNet-M
Bohnenpfopf, Hähnchen mit Pommes oder Kalbfleisch mit Soße und Kartoffeln: In der Mensa-Eingangshalle der Universität Karlsruhe drängeln sich um die Mittagszeit die jungen Leute.	Pizzaschnitten, Hähnchen mit Reis oder Rinderbraten mit Gemüse und Kartoffeln: Um die Abendstunde füllt sich der Festsaal der Universität Stuttgart mit hungrigen Studenten.	MPNet-M
Nein. Zurückweisungen gibt es ja auch in anderen EU-Ländern.	Nein, Zuwanderungen gibt es ja auch in anderen EU-Ländern.	M-E5-b
Und es ist zugleich die Emanzipation der Linken von ihrem Ex-Vorsitzenden Oskar Lafontaine.	Und es ist zugleich die Spaltung der Linken wegen ihrem Ex-Vorsitzenden Oskar Lafontaine.	M-E5-b

Table 6: Randomly selected examples of successful distractors. The red font indicates modified text.