Extending the Massive Text Embedding Benchmark to French

Anonymous ACL submission

Abstract

In recent years, numerous embedding models have been made available and widely used for various NLP tasks. Choosing a model that performs well for several tasks in English has been largely simplified by the Massive Text Embedding Benchmark (MTEB), but extensions to other languages remain challenging. This is why we expand MTEB to propose the first massive benchmark of sentence embeddings for French. Not only we gather 22 existing datasets in an easy-to-use interface, but we also create three new French datasets for a global evaluation over 8 different tasks. We perform a large scale comparison with 46 carefully selected embedding models, conduct comprehensive statistical tests, and analyze the correlation between model performance and many of their characteristics. We find out that even if no model is the best on all tasks, large multilingual models pre-trained on sentence similarity perform particularly well. Our work comes with open-source code, new datasets and a public leaderboard[1].

1 Introduction

Embeddings are dense vector representations that capture the semantics of an input. The first emblematic example is Word2Vec, introduced by Mikolov et al. (2013). It consists of neural architectures trained to learn high-quality words representations from contextual relationships in huge amounts of texts. Other models were proposed since then, leveraging the transformer architecture (Vaswani et al., 2017) to produce both generic and contextualized word embeddings using self-attention. The most influential one is BERT (Devlin et al., 2019), a transformer-based encoder pre-trained on the Masked-Language Modeling and Next Sentence Prediction tasks to learn the semantics of a language. A multitude of models now exists with various architectures, monolingual or multilingual, pre-trained or fine-tuned (Naseem et al., 2021; Ding et al., 2023). Embeddings are essential for a large part of NLP applications, such as semantic search, topic modeling, classification, etc. Choosing the most suitable model for a specific application is not always obvious. Hence, initiatives such as MTEB (Muennighoff et al., 2022) enable these models to be compared across various datasets and tasks. However, most of the provided resources for the evaluations are in English.

We extend MTEB to produce the first massive benchmark on French embeddings. Our contributions are the following. We bring together 25 datasets (3 of which are new) in an easy to use toolbox. We make a large scale comparison of 46 models with diverse characteristics. We propose an in-depth performance analysis based on models characteristics, statistical tests and a meta analysis of all language-specific MTEB leaderboards.

We show that although no model is the best on all tasks, a set of methods seems to perform particularly well for most of the datasets of the benchmark (see Figure 2). We also find out a strong correlation between the performance and the size of a model, its training technique and the use of multilingual data. The code and all the resources used for this work are made publicly available as open-source.

2 Related Work

This section first describes the various family of models used in this benchmark. Previous works comparing embedding models are also discussed.

2.1 Sentence embeddings

Sentence embeddings are required for a large set of language tasks, such as Semantic Textual Similarity (STS) and knowledge retrieval. Before the standardization of the transformer architecture (Vaswani et al., 2017), the common architecture used to obtain sentence embeddings was Recurrent

[1] Access links will be available in the final version

Neural Networks (RNN) (Kiros et al., 2015). Many models have been proposed to build sentence representations leveraging this architecture, for example, *LASER* (Artetxe and Schwenk, 2018) which trains a BiLSTM encoder and a LSTM decoder to produce multilingual sentence embeddings.

Now that most language models leverage the transformer architecture, a large majority of embedding methods internally start by generating embeddings for tokens. An additional step is therefore required to compute sentence embeddings. A well-known strategy is pooling, which can be done either using the embedding of the CLS token (Devlin et al., 2019), or by mean-pooling which averages the tokens output vector representations. Alternatively, max-pooling is sometimes used and consists in taking the component-wise maximum across all token embeddings. Finally, for sequential and causal models, the common strategy is to use the last token representation (Muennighoff, 2022).

In addition to pooling, these embeddings can be learned by training a siamese network which outputs vector representations that can be compared using similarity measure (Reimers and Gurevych, 2019). For encoder-decoder based language models that do not come with a CLS token such as *T5* (Raffel et al., 2019), authors like Ni et al. (2021) propose to train a dual encoder paired with a contrastive learning framework (Gao et al., 2021; Neelakantan et al., 2022). Other works build multilingual models leveraging contrastive frameworks on sentence similarity such as *E5* (Wang et al., 2022).

More recently, Large Language Models (LLM) have been used to build sentence embeddings, also through fine-tuning on sentence similarity (Wang et al., 2023; Zhang et al., 2023). Some of these models are not openly accessible, namely OpenAI's *text-embedding-ada-002*[2], or Cohere[3] and Voyage models[4].

Regarding the French-specific literature, models based on the BERT architecture were proposed by the community, such as *camembert* (Martin et al., 2019), *flaubert* (Le et al., 2020), and very recently larger models such as *Vigogne*[5] and *CroissantLLM* (Faysse et al., 2024).

2.2 Benchmarks

Embedding models are generally compared on specific tasks. For example, BEIR (Thakur et al., 2021) focuses on Information Retrieval, while SemEval (Agirre et al., 2016) is used for STS. The Reranking task is also considered in some benchmarks (Wang et al., 2021). Others works evaluate embedding models on multiple tasks such as GLUE (Wang et al., 2018), BIG-bench (et al., 2022), SentEval (Conneau and Kiela, 2018) or the meta-embeddings comparison (García-Ferrero et al., 2021).

The most comprehensive benchmark to date is MTEB (Muennighoff et al., 2022). It compares 33 models across 8 embedding tasks on 58 datasets, allowing to take informed decision when selecting an embedding model. It results in no model is best on all tasks. Therefore, the different embedding models seem to exhibit task-dependent strengths and weaknesses. Performance is also strongly correlated with the model size, so users can define their trade-off between efficiency (speed, memory) and embedding quality. They can also have other criteria such as multilingualism. The MTEB benchmark mainly focuses on sentence embeddings. It attempts to select tasks that resemble downstream applications (such as retrieval, classification, clustering, etc.) in order to help the user project into such use-cases. However, it should be kept in mind that the datasets may not perfectly reflect the peculiarity of the downstream tasks.

MTEB still has an important limit: it does not easily allow to select a model for other languages than English or for multilingual applications. Some initiatives already evaluate embedding models for specific languages, such as Arabic (Elnagar et al., 2023) and German (Wehrli et al., 2024). Our work comes with the same ambition for French. It relies on the MTEB structure that provides a solid basis for analysis, and extend it to a new language.

3 MTEB for French

In this section, we describe the tasks, the datasets and the models that we propose for the French extension of MTEB. We also list the research questions we want to discuss with the results.

3.1 Tasks

Similarly to MTEB (Muennighoff et al., 2022), we evaluate the ability of models to produce relevant embedding vectors for 8 different tasks :
Bitext Mining evaluates a model's ability to pro-

[2] https://platform.openai.com/docs/guides/embeddings/embedding-models

[3] https://txt.cohere.com/introducing-embed-v3/

[4] https://docs.voyageai.com/embeddings/

[5] https://github.com/bofenghuang/vigogne

duce vector representations that preserve the semantics of a pair of sentences from different languages.
Classification ensures that an embedding model produces vectors that help classifiers correctly associate samples to their relevant classes.
Clustering ensures clusters of semantically close sentences can be built based on their embeddings.
Pair Classification assesses whether a model generates close vector representations for texts that carry the same information, and distant ones for texts that do not.
Retrieval implies that a model creates documents and queries embeddings for which the distance is correlated with matching relevance.
Reranking assesses whether a model produces embeddings that enable ordering documents according to their relevance regarding a given query.
Semantic Textual Similarity (STS) evaluates the ability to generate representations for which the distance between texts determine their relatedness.
Summarization evaluates the ability of a model to produce embedding vectors for both a text and its summary that are close if the summary is relevant.

More information about the tasks can be found in the MTEB paper (Muennighoff et al., 2022). Overall, most tasks have in common that they evaluate the relevance of embedding similarity in different contexts.

3.2 Datasets

We identified 8 datasets relevant for French in MTEB. We complemented these with 14 external relevant datasets and created 3 new ones. Therefore, as of today, our French MTEB runs on a total of 25 datasets (see Figure 1) spread over the 8 tasks mentioned above. This section briefly describes the 3 new datasets we introduce and goes along the analysis of the similarities between datasets.

3.2.1 New datasets

Syntec (Retrieval) The Syntec French collective bargaining agreement[6] is composed of around 90 articles. Despite its topic, the language used does not feature the specificity of the legal vocabulary, making the data suitable for benchmarking general purpose models. A hundred questions have been manually created, and paired with the articles containing the answer.

HAL (Clustering) Hyper Articles en Ligne (HAL) is a French open archive of scholarly documents from all academic fields. Scrapping this resource, we fetched 85,000 publications in French and extracted their *id*, *title* and *domain*. The publications can be clustered from their title and the domain can be used as ground truth.

SummEvalFr (Summarization) The original SummEval dataset (Fabbri et al., 2021) consists of 100 news articles from the CNN/DailyMail dataset. Each article comes with 10 human-written summaries and 16 machine-generated summaries annotated by 8 persons with a score for coherence, consistency, fluency, and relevance. We translated it from English to French using DeepL API[7].

Reranking datasets The reranking task, as evaluated in MTEB, requires datasets composed of a set of queries each associated with relevant and irrelevant documents. Despite our efforts, we did not find any French dataset that natively exhibits such structure. Thus, to evaluate this task, we built reranking datasets based on the *Syntec* and *Alloprof* (Lefebvre-Brossard et al., 2023) retrieval datasets. These already feature queries and labeled relevant documents. Irrelevant ones were added with the following process. The embedding model *all-MiniLM-L6-v2*, available on HugginFace[8], was used to compute the cosine similarity between each embedded query and all embedded documents. The 10 most similar documents not marked as relevant constitute the set of irrelevant documents.

We recognize that this process leads to a high correlation between the retrieval and reranking tasks. We still think it is important to make the latter available, with an open door to future improvement.

3.2.2 Similarity analysis

To give insights about the benchmark contents, we investigate the proximity between the datasets' topics. The methodology introduced by Muennighoff et al. (2022), i.e. computing a sampled average embedding of the datasets, is used to build a dataset-similarity matrix (displayed in appendix Figure 4). The distances between averaged embedding vectors of each dataset only brings little information (correlations range from 0.89 to 1 in Figure 4). So we complement this by observing the datasets clouds of embedding in a 2D plane using PCA in Figure 1.

Figures 1 and 4 seem to correlate, showing high similarity between two datasets when the same underlying data is used in different tasks. Overall,

[6] https://www.syntec.fr/convention-collective/

[7] https://www.deepl.com

[8] https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

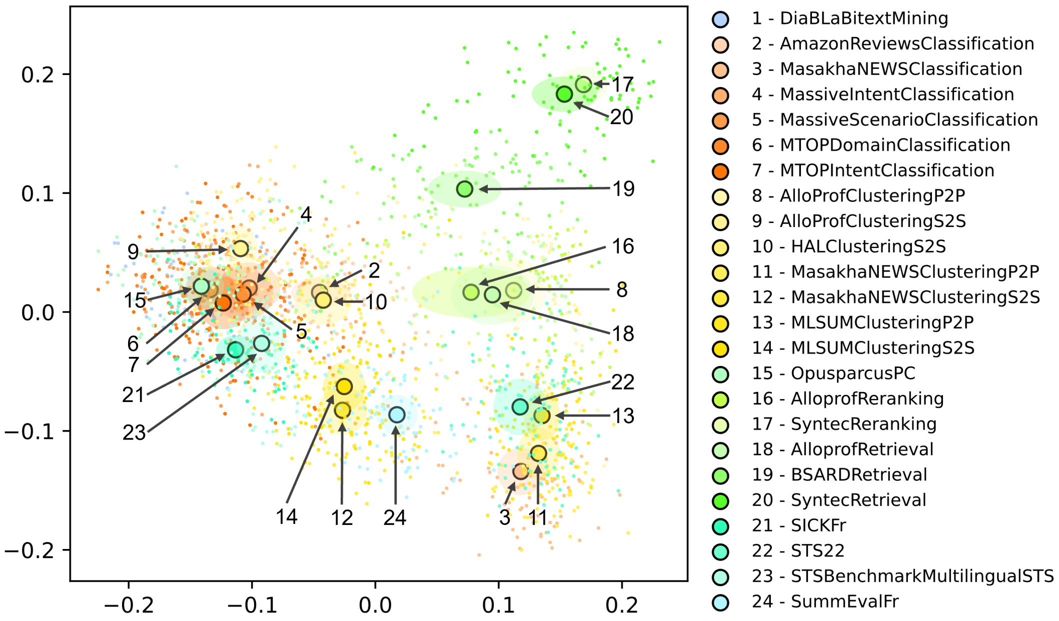

Figure 1: 2D projection of datasets. 90 random samples per dataset are embedded using the *multlingual e5 large* model (Wang et al., 2022). The embeddings are reduced to 2 dimensions using PCA. The centroid of each dataset is represented, along with the ellipse showing the standard deviation along each axis.

dataset topics are fairly close to each other, with some exceptions such as Syntec dataset. As more datasets are added to the benchmark, this analysis will help select new data that do not produce redundant results. It may also be a help understand the link between the results and the datasets' topics.

3.3 Models

For comparison on our benchmark, we selected various models to fulfill three objectives.

Quantity: The aim was to compare a substantial number of models (46 in total) to provide comprehensive results, facilitating the community in selecting effective French models.

Relevance: It was imperative to include top performers from the MTEB benchmark (Muennighoff et al., 2022). We mainly selected multilingual models but also some English ones with good cross-lingual abilities. Additionally, we integrated natively French transformer-based models such as *camembert* (Martin et al., 2019), *flaubert* (Le et al., 2020) and even the very recent *CroissantLLM* (Faysse et al., 2024).

Variety: Diverse model types were included to offer an insightful analysis across various model characteristics (dimension, training strategy, etc.).

In line with the third objective, we explicit below the studied characteristics of embedding models that will be discussed with the results.

Embedding dimension This critical element influences the expressiveness of the representation and, in practical applications, the underlying storage and compute costs. We selected models with embedding dimensions ranging from 384 to 4096.

Sequence length Being the number of tokens that a model can consider as input, the sequence length is important as it impacts the unit that can be encoded (sentence, paragraph, document). However, encoding overly long sequences requires the ability to efficiently store the relevant information into a single vector. Among the selected methods, this criterion varies from 128 tokens to 32768.

Model parameters Often correlated with the two first characteristics, parameter count is important for practical applications as it affects usability on resource-efficient machines. The selected models have a number of parameters ranging from 20 million (∼100Mb in float32) to 7 billion (∼28Gb).

Language This is a major feature of language models. Some are monolingual and others multilingual. Language is usually acquired during pre-training but sometimes model familiarize with new languages at tuning. For the benchmark, we selected French models, as well as bilingual or multilingual models. We also included a few ones claimed to be English (e.g. *all-MiniLM-L12-v2*[9]).

[9] https://huggingface.co/sentence-transformers/all-MiniLM-L12-v2

Model types As mentioned in the related works section, there are several strategies to generate text embeddings such as aggregating (e.g. with average pooling) token-level embeddings from raw pre-trained models, or adding an extra constrastive learning step on a sentence similarity task with, optionally, additional transformation layers. We included models of all types in our benchmark, summarizing the model type information under two relevant criteria: finetuned vs pretrained, and trained for sentence similarity or not.

The selected models are visible in Figure 2 and all of their characteristics are summarized in appendix Table 3. Overall, the selection includes the best models from the sentence transformers framework (Reimers and Gurevych, 2019), the most popular French NLP models (Le et al., 2020; Martin et al., 2019), their variants optimized for semantic similarity (Reimers and Gurevych, 2019), numerous multilingual models performing at the top on MTEB (e.g *E5* and *T5*), *Bloom* variants (Zhang et al., 2023), models based on very recent powerful LLMs (Wang et al., 2023; Faysse et al., 2024) and finally the proprietary models of OpenAI, Cohere and Voyage. Certain models were selected in multiple sizes to isolate the dimensionality effect more effectively. We provide information on the models' licenses as reported in the Hugging Face hub[10]. However, we encourage readers to conduct further research before utilizing a model.

3.4 Evaluation

For the sake of homogeneity, models are evaluated using the same metrics per task as in MTEB (Muennighoff et al., 2022): Classification (Accuracy), Bitext mining (F1 score), Pair classification (AP), Clustering (V measure), Reranking (MAP), Retrieval (NDCG@k), Summarization and STS (Spearman correlation based on cosine similarity).

Using the overall benchmark results, our goal will be to answer the following research questions:
Q1: Is there a model that outstands on all tasks?
As we are trying to find out whether one embedding model is statistically better than the others for French, the objective will also be to analyze the performance of the models by tasks in order to facilitate model choice for specific applications.
Q2: Are there any links between the model characteristics and performance?
In section 3.3, we undertook the substantial task of gathering the characteristics of all evaluated models. The goal here will be to analyze their impact on performance and draw conclusions about, for example, the relationship between embedding dimension and model ranking on the benchmark.
Q3.a: Do multilingual models perform similarly from one language to another?
We investigate the universal capabilities of multilingual models, i.e. whether they show similar performances among all available languages. To answer, we will collect the results for English, Chinese and Polish from the MTEB leaderboard[11] and compare them with our results for French.
Q3.b: Do monolingual models have multilingual capabilities?
We will interrogate the ability of model trained exclusively on one language to perform well on another language.
Q4: Are there any correlations between datasets with respect to model ranking?
To go further than the correlation analysis among datasets regarding their topics (see section 3.2.2), a subsequent analysis will be conducted regarding how they rank models. Additionally, complementary insights will be derived from examining correlations of models relatively to their strengths and weaknesses across different datasets.

4 Results and discussion

In this section, we present the results through the prism of our research questions.

Q1: Is there a model that outstands on all tasks?

Models performances for each task are presented in appendix Tables 4, 5, 6 and 7. The critical difference diagram of average score ranks is available in Figure 2.

As in MTEB (Muennighoff et al., 2022), no model claims the state-of-the-art in all tasks. The best performer varies depending on the task and dataset. For classification, *sentence-t5-xxl* and *text-embedding-ada-002* seem to give the best results. *text-embedding-ada-002* appears to be the best model for clustering and reranking even though other models are ranking first for specific datasets. It shares the first position for the retrieval task with *sentence-t5-xxl* and *voyage-code-2*. For the summarization task, *e5-mistral-7b-instruct* has the best

[10] https://huggingface.co/models

[11] https://huggingface.co/spaces/mteb/leaderboard

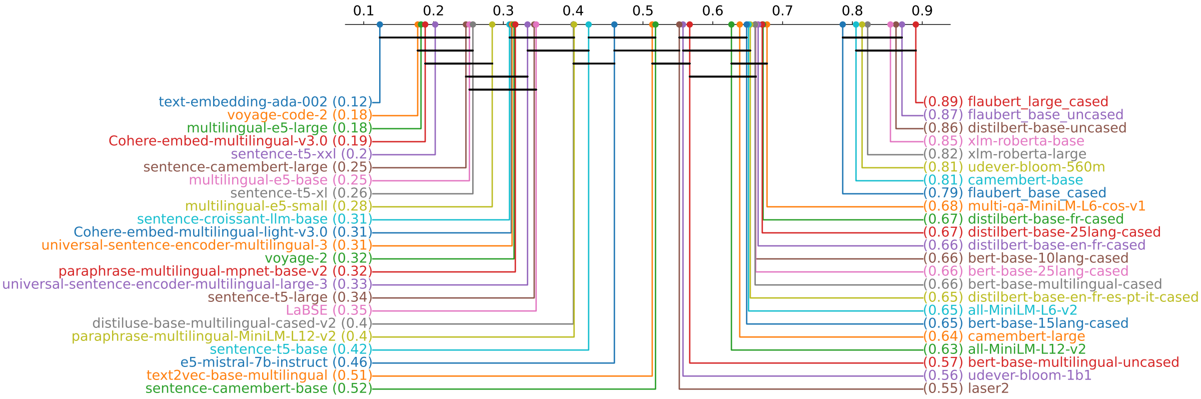

Figure 2: Critical difference diagram representing the significant rank gaps between models. The axis represents the normalized average rank of the models (lower is better). The black bars indicate that the difference of models' rank is not statistically significant, i.e. lower that the critical difference.

results. The only dataset available for pair classification, OpusparcusPC (Creutz, 2018), does not differentiate between models. *LaBSE* leads on the bitext mining task. Finally, *sentence-camembert-large* is, with Cohere models, ahead of its competitors for semantic textual similarity.

Figure 2 shows a global model comparison across all datasets. The statistically equivalent top performers for this benchmark, from highest to lowest empirical performance, are *text-embedding-ada-002*, *voyage-code-2*, *multilingual-e5-large*, *sentence-t5-xxl*, *sentence-camembert-large*, and *multilingual-e5-base*. Interestingly, many models do not show significant performance gap between their small and large flavours.

Q2: Are there any links between model characteristics and performance

The Spearman correlations between the average rank of the models and their characteristics are the following: *Finetuned vs pretrained* (0.511), *Model number of parameters* (0.292), *Max sequence length* (-0.161), *Embedding dimension* (0.108), *Tuned for sentence similarity* (0.745), *Bilingual* (-0.13), *English* (-0.017), *English but tuned on other languages* (-0.051), *French* (-0.259), *Multilingual* (0.168), *Closed source* (0.359). Additionally, all cross-correlations between characteristics are reported in appendix Figure 5.

As expected, the score most strongly correlates with whether the evaluated models were trained on a sentence similarity task. Of course, this criterion is itself connected to the more general *Finetuned* one. The only top-performing models solely pre-trained are from the *E5* family, where the pre-training is in fact contrastive and optimized for similarity. Conversely, models pre-trained on token-level tasks, and generating embeddings via pooling, appear less well-suited for the benchmark tasks.

Furthermore, we observe a performance correlation with the embedding dimension and the model's number of parameters, which are often correlated themselves. This appears very clearly on the relative ranking of *E5* and *T5* models (see Figure 2). However, some small models, such as the standard version of multilingual universal sentence encoder, perform very well on the benchmark. It is noteworthy that the maximum sequence length, while an important criterion for generative tasks with LLMs, does not correlate with performance. This can be explained both by the fact that a lot of datasets contain relatively small texts (see appendix Table 2 showing that 14 datasets have less than 50 tokens) and that even if the model can manage long sequences, it is still challenging to summarize long contents in single global embeddings.

Regarding language, it is surprising that good performance is associated with the multilingual models we selected rather than French ones. In reality, we can observe that our multilingual selection is more often fine-tuned for similarity (these two criteria have a positive correlation), which could explain the result. Nevertheless, we highlight the excellent performance of *sentence-camembert-large* and *sentence-croissant-llm-base*, two French models trained using the sentence BERT methodology (Reimers and Gurevych, 2019).

Lastly, we emphasize that closed-source models perform very well on this benchmark, but we lack information about their characteristics. As more open-source well-performing models get added in the future, we can expect this correlation to decrease.

Q3.a: Do multilingual models perform similarly from one language to another?

We extract the overall average performance, for each of the four languages (FR, EN, ZH and PL), of a selected number of multilingual models. We normalize these models ranks over the total number of models evaluated for each selected language, and report the results in Table 1.

The normalized ranks seem to be distributed differently from one language to another. We suppose that it is due to both a higher volume of models is some languages like Chinese or English, and to the fact that multilingual models have unbalanced abilities across languages. For better comparison, we also compute the selected models relative rank.

We notice that *e5-mistral-7b* outperforms other models in English but shows poor performances in other languages (in a low range of absolute ranking). This might be due to its training strategy, *i.e.* the model was fine-tuned on *mistral-7b*, which is an English model, on synthetic multilingual data.

Furthermore, we see that for Chinese, monolingual models performance transcends that of multilingual ones (where the best of our selection is only ranked in the middle). This might be due to the richness and the specificity of the language (tokens, encoding, etc.).

Overall, *multilingual-e5-large* shows the best performance across at least the 4 languages that are evaluated in Table 1. Without being the best performer in each language, it ranks at the top among other multilingual models.

Based on the available data, we conclude that multilingual models are not always the best performers in other languages than the high-resource ones such as English or French. Also, multilinguality and good performance in one language does not guarantee good performance in others languages.

Model	Normalized rank in % (rel. rank)			
	FR	EN	ZH	PL
text-embedding-ada-002	2 (1)	22 (4)	66 (6)	-
sentence-t5-xxl	4 (2)	28 (6)	-	-
multilingual-e5-large	9 (3)	19 (3)	52 (2)	22 (1)
Cohere-multilingual-v3	11 (4)	5 (2)	-	-
multilingual-e5-base	13 (5)	29 (7)	60 (3)	44 (2)
sentence-t5-xl	17 (6)	38 (10)	-	-
multilingual-e5-small	20 (7)	38 (9)	62 (4)	50 (3)
paraph-multi-mpnet-base-v2	22 (8)	-	78 (8)	72 (4)
Cohere-multilingual-light-v3	24 (9)	25 (5)	-	-
sentence-t5-large	33 (10)	42 (11)	-	-
LaBSE	39 (11)	58 (14)	-	83 (5)
sentence-t5-base	43 (12)	50 (13)	-	-
laser2	50 (13)	62 (15)	-	-
e5-mistral-7b-instruct	52 (14)	2 (1)	46 (1)	-
udever-bloom-1b1	59 (15)	34 (8)	64 (5)	-
udever-bloom-560m	91 (16)	48 (12)	68 (7)	-

Table 1: Ranking of models according to their average performance over all tasks in French (FR), English (EN), Chinese (ZH) and Polish (PL). The first value corresponds to a normalized rank computed by dividing the rank of the model compared to its competitors on a language, by the total number of evaluated models for this language. This total number is 46, 141, 60 and 18 respectively for French, English, Chinese and Polish. The normalized rank can be interpreted as a quantile, *e.g.* *text-embedding-ada-002* is in the top 2 percent quantile for French. The value in parenthesis is the relative rank between the models in the table only.

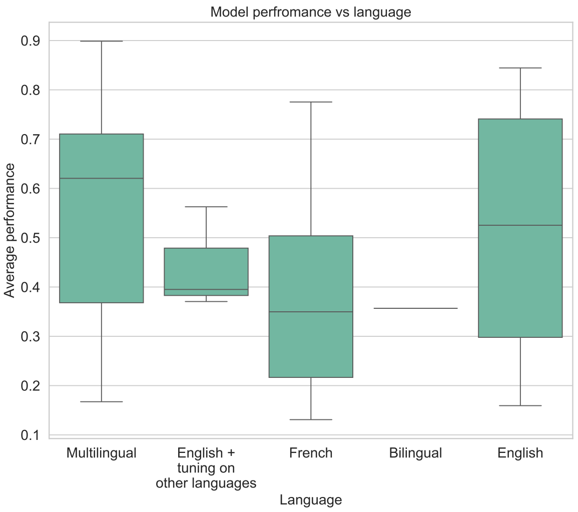

Figure 3: Model performance depending on the language of the data they have been trained on.

Q3.b: Do monolingual models have multilingual capabilities?

It is surprising to note the absence of clear correlation between the language the model is trained on and its performance on French, as shown by the large standard deviation in Figure 3. Furthermore, monolingual models trained exclusively on English such as *voyage-code-2* show very good results on French datasets compared to models trained exclusively on French such as *flaubert* derivatives and *distilbert-base-fr-cased* (see Table D.1).

A large part of the selected French models generate embeddings using a pooling strategy. We hypothesize that this lowers the results in comparison with sentence transformer models where the pooled representation is part of the model and trained with

it. This hypothesis is endorsed by the excellent results of *sentence-camembert-large*, a sentence transformer model trained on French corpus.
Finally, it should be noted that a significant portion of the French data used to train the selected French models actually comes from English datasets that have been machine translated (May, 2021). Despite the tremendous progress of machine translation, it is well known that the generated data may be unrepresentative of the language used by native speakers and causes a reduced final performance (Barbosa et al., 2021).

Q4: Are there any correlations between datasets with respect to model ranking?

The datasets correlation w.r.t model ranking are presented in appendix Figure 7. Except for three datasets (BSARDRetrieval, AlloProfClusteringS2S, SummEvalFr), the correlations, on average, are high but there is still enough diversity to make each dataset interesting for the French MTEB benchmark. Three pairs (*SynteсReranking/Retrieval*, *MassiveScenarioClassification/MTOPDomainClassification* and the two variants of *FloresBitextMining*) exhibit notably high correlations (∼0.97). As we enrich the benchmark with novel datasets in the future, we may consider only retaining a single element of each pair. It is interesting to point out the sub-diagonal correlation triangles. The datasets being arranged by task, from Classification to Bibtext Mining, this indicates that models behave more similarly within the same task than between two different tasks. This underscores the importance of having multiple tasks in the benchmark to select general-purpose models. For readers interested in specific tasks, it is more relevant to examine task-specific rankings rather than the overall one. The complementary results of model correlations w.r.t to strengths and weaknesses on datasets are displayed in appendix Figure 6. Strong correlations in behavior emerge among several multilingual variants of BERT and DistilBERT (trained under the same protocol (Abdaoui et al., 2020)). Correlations are also observed among numerous models trained using the sentence transformers framework (Reimers and Gurevych, 2019), models from the *E5* family (Wang et al., 2022), as well as proprietary models, e.g. from Cohere and OpenAI. Conversely, these models show minimal correlation with pre-trained models for which token-embedding pooling techniques were employed.

5 Conclusion and perspectives

Given the potential variability in model behavior across languages (see Table 1), it is important to contribute by introducing more language-specific MTEB variants. Our work focuses on this objective, and proposes the French one. We gather and introduce novel data for a total of 25 datasets across 8 tasks. We perform a large scale comparison with 46 models and carry a deep analysis around several research questions. The results offer interesting insights to help selecting embeddings optimized for French. It also highlights the importance of statistical tests, often overlooked in leaderboards, and demonstrate that certain empirical dominances may not be statistically significant. Furthermore, through the analysis of performance vs model characteristics, it provides readers with hints to make new proposals that may outperform existing ones, on one or multiple target languages. This involves exploring the latest model architectures (Jiang et al., 2023; Touvron et al., 2023), proposing various dimension variants, training for sentence or document similarity, harnessing multilinguality through appropriate dataset selection, and more.

This work opens several doors for future improvements. By examining dataset diversity in terms of topics and model ranking, we observe that the benchmark would benefit from additional datasets that introduce higher diversity. Beyond classification, many tasks focus on semantic similarity, explaining the strong performance of models trained for similarity. Exploring novel tasks in the generative spectrum or evaluating token embeddings (contextualized or not) on tasks like Named Entity Recognition could be an interesting paths for future exploration. There are also opportunities for improvements on the model side. With numerous existing models that could be added to the leaderboard and many new proposals awaiting. For instance, we can already see the promising capabilities of early variants of recent models (Faysse et al., 2024) and expect that future proposals will come to compete strongly with Closed Source models. Ultimately, we hope to see the emergence of other language-specific MTEB variants (e.g. for high-resource languages like Spanish, German), enabling a more comprehensive evaluation of multilingual model performance.

References

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. 2020. Load what you need: Smaller versions of mutililingual bert. In *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123.

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. 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.

Mikel Artetxe and Holger Schwenk. 2018. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Arthur Barbosa, Máverick Ferreira, Rafael Ferreira Mello, Rafael Dueire Lins, and Dragan Gasevic. 2021. The impact of automatic text translation on classification of online discussions for social and cognitive presences. In *LAK21: 11th International Learning Analytics and Knowledge Conference*, LAK21, page 77–87, New York, NY, USA. Association for Computing Machinery.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations. *ArXiv*, abs/1803.05449.

Mathias Creutz. 2018. Open subtitles paraphrase corpus for six languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *North American Chapter of the Association for Computational Linguistics*.

Ning Ding, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2023. Sentence and document representation learning. In *Representation Learning for Natural Language Processing*, pages 81–125. Springer Nature Singapore Singapore.

Ashraf Elnagar, Sane Yagi, Youssef Mansour, Leena Lulu, and Shehdeh Fareh. 2023. A benchmark for evaluating arabic contextualized word embedding models. *Information Processing & Management*, 60(5):103452.

Aarohi Srivastava et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *ArXiv*, abs/2206.04615.

Alexander R Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. Summeval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Manuel Faysse, Patrick Fernandes, Nuno M. Guerreiro, António Loison, Duarte M. Alves, Caio Corro, Nicolas Boizard, João Alves, Ricardo Rei, Pedro H. Martins, Antoni Bigata Casademunt, François Yvon, André F. T. Martins, Gautier Viaud, Céline Hudelot, and Pierre Colombo. 2024. Croissantllm: A truly bilingual french-english language model.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Iker García-Ferrero, Rodrigo Agerri, and German Rigau. 2021. Benchmarking meta-embeddings: What works and what does not. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3957–3972, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. *Advances in neural information processing systems*, 28.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: Unsupervised language model pre-training for french.

Antoine Lefebvre-Brossard, Stephane Gazaille, and Michel C. Desmarais. 2023. Alloprof: a new french question-answer education dataset and its use in an information retrieval case study.

Louis Martin, Benjamin Muller, Pedro Ortiz Suarez, Yoann Dupont, Laurent Romary, Eric Villemonte de la Clergerie, Djamé Seddah, and Benoît Sagot. 2019. Camembert: a tasty french language model. In *Annual Meeting of the Association for Computational Linguistics*.

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Tomas Mikolov, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations*.

Niklas Muennighoff. 2022. Sgpt: Gpt sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Usman Naseem, Imran Razzak, Shah Khalid Khan, and Mukesh Prasad. 2021. A comprehensive survey on word representation models: From classical to state-of-the-art word representation language models. *Transactions on Asian and Low-Resource Language Information Processing*, 20(5):1–35.

Arvind Neelakantan, Tao Xu, Raul Puri, Alec Radford, Jesse Michael Han, Jerry Tworek, Qiming Yuan, Nikolas Tezak, Jong Wook Kim, Chris Hallacy, et al. 2022. Text and code embeddings by contrastive pre-training. *arXiv preprint arXiv:2201.10005*.

Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B. Hall, Daniel Cer, and Yinfei Yang. 2021. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text -to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *CoRR*, abs/2104.08663.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP@EMNLP*.

Kexin Wang, Nils Reimers, and Iryna Gurevych. 2021. TSDAE: Using transformer-based sequential denoising auto-encoderfor unsupervised sentence embedding learning. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 671–688, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Silvan Wehrli, Bert Arnrich, and Christopher Irrgang. 2024. German text embedding clustering benchmark.

Xin Zhang, Zehan Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, Meishan Zhang, and Min Zhang. 2023. Language models are universal embedders. *ArXiv*, abs/2310.08232.

A Datasets similarity and number of tokens

Figure 4 represents the semantic similarity between each dataset. The methodology was as follow : 90 random samples per dataset are embedded using the *multilingual-e5-large* model. The embeddings of each dataset's samples are averaged. The similarity between each dataset is then calculated using cosine similarity as in (Muennighoff et al., 2022).

Table 2 displays the average number of tokens per sample for each dataset. The dataset's content were tokenized using *cl100k_base* encoding. For Retrieval datasets, the two numbers refer to the queries and the documents. For Reranking, the three numbers refer to the queries, the relevant documents and the irrelevant ones. For *SummEvalFr*, the three numbers refer to the texts, human summaries and machine summaries.

B Correlations

This section presents various correlations computed based on the model results on the proposed benchmark.

Figure 5 represents cross-correlations between models' performances and their studied characteristics as a heatmap.

Figure 6 represents the spearman correlations in terms of performance across models

Figure 7 represents the spearman correlations in terms of performance across datasets

C Model characteristics

We present in this section the model characteristics we collected for the 46 evaluated models.

D Evaluation results

We present in this section the results obtained for each model on each task. In order to be relevant, we used the same metrics as in MTEB, which varies from one type of task to another :

- Bitext Mining : F1 score
- Classification : Accuracy
- Clustering : V measure
- Pair Classification : Average Precision (AP)
- Reranking : Mean Average Precision (MAP)
- Retrieval : Normalized Discounted Cumulative Gain at k (NDCG@k)
- STS : Spearman correlation based on cosine similarity
- Summarization : Spearman correlation based on cosine similarity

D.1 Average performance per task type

Table 4 presents the average performance of each model on each task type.

D.2 Evaluation results per task

Tables 5, 6 and 7 present the models' performance on each task type. Table 5 presents the performance on classification and pair classification tasks. Table 6 presents the performance on bitext mining, reranking and retrieval tasks. Table 7 presents the performance on summarization and clustering tasks.

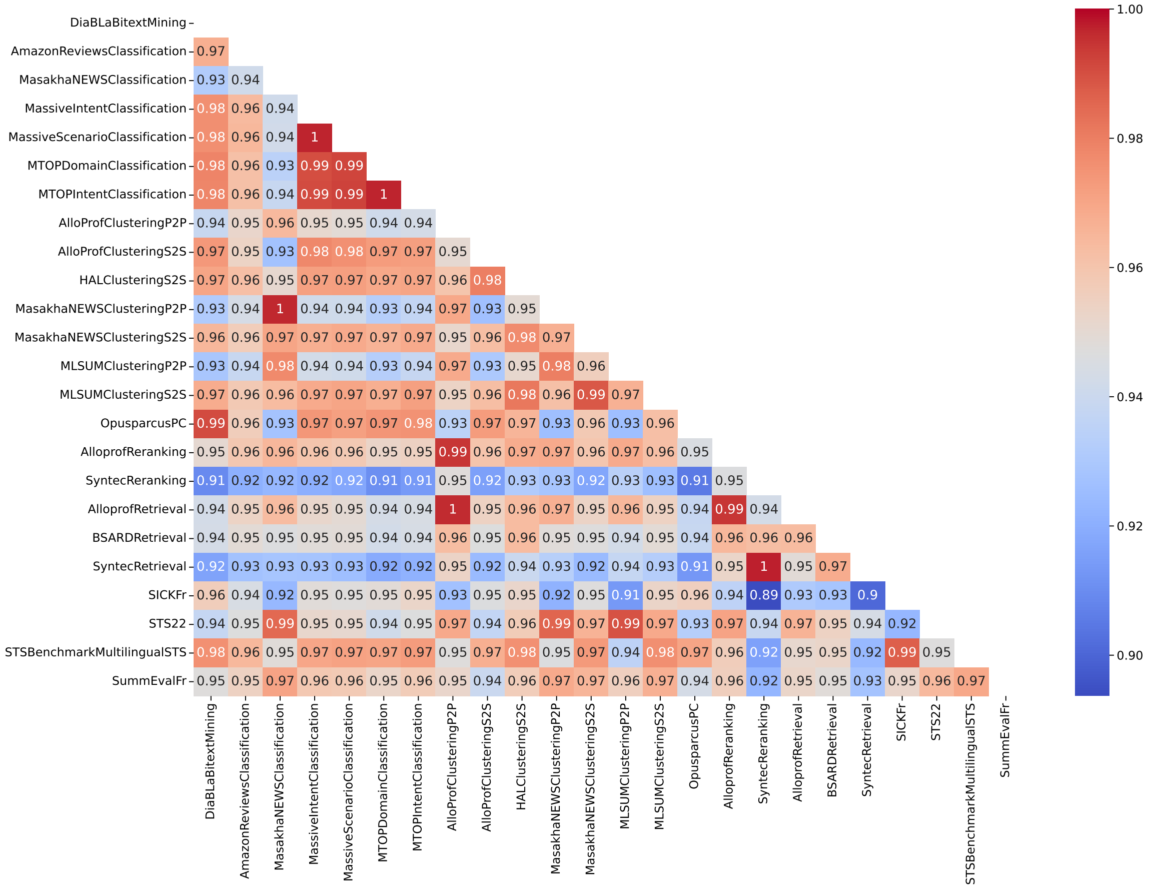

Figure 4: Cosine similarity between datasets. 90 random samples per dataset are embedded using the *multilingual-e5-large* model. The embeddings of each dataset samples are averaged. The similarity between each dataset is then calculated using cosine similarity as in (Muennighoff et al., 2022).

Dataset	**Average number of tokens**
AmazonReviewsClassification	48.97
MasakhaNEWSClassification	1386.24
MassiveIntentClassification	11.41
MassiveScenarioClassification	11.41
MTOPDomainClassification	12.41
MTOPIntentClassification	12.41
AlloProfClusteringP2P	1021.79
AlloProfClusteringS2S	8.81
HALClusteringS2S	24.1
MasakhaNEWSClusteringP2P	1398.1
MasakhaNEWSClusteringS2S	21.44
MLSUMClusteringP2P	1082.76
MLSUMClusteringS2S	20.8
OpusparcusPC	9.19
STSBenchmarkMultilingualSTS	20.01
STS22	722.15
SICKFr	15.15
DiaBLaBitextMining	12.02
FloresBitextMining	33.42
SyntecReranking	19.22 - 392.19 - 1318.42
AlloprofReranking	48.83 - 1500.58 - 7547.37
AlloprofReranking	48.31 - 1117.91
BSARDRetrieval	144.03 - 24530.8
SyntecRetrieval	19.22 - 295.65
SummEvalFr	657.08 - 71.18 - 107.56

Table 2: Average number of tokens of texts, using the *cl100k_base* tokenizer, in the datasets from the Massive Text Embedding Benchmark for French. For Retrieval datasets, the two numbers refer to the queries and the documents. For Reranking, the three numbers refer to the queries, the relevant documents and the irrelevant ones. For *SummEvalFr*, the three numbers refer to the texts, human summaries and machine summaries.

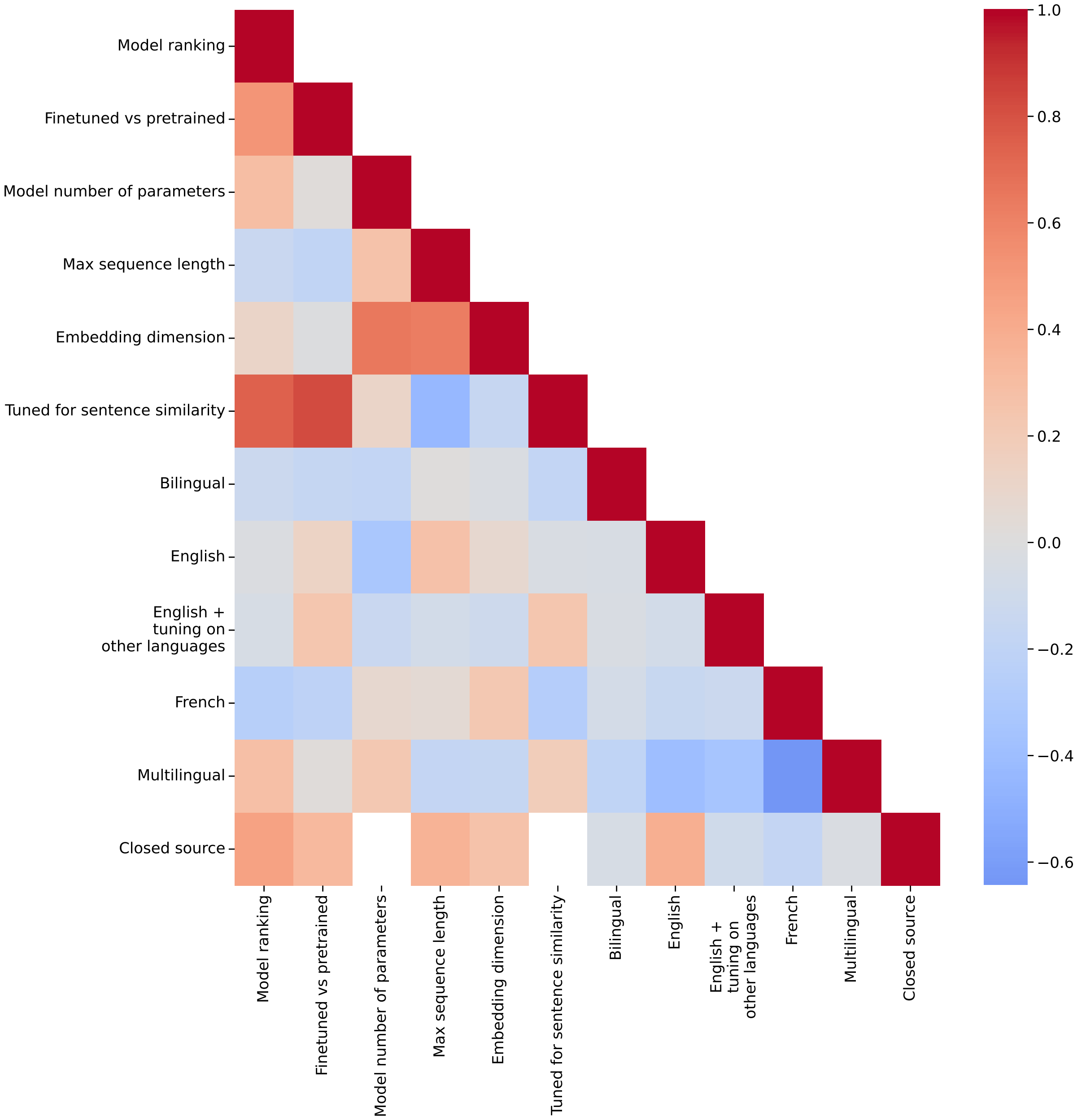

Figure 5: Heatmap representing cross-correlations between models' characteristics and models' performances.

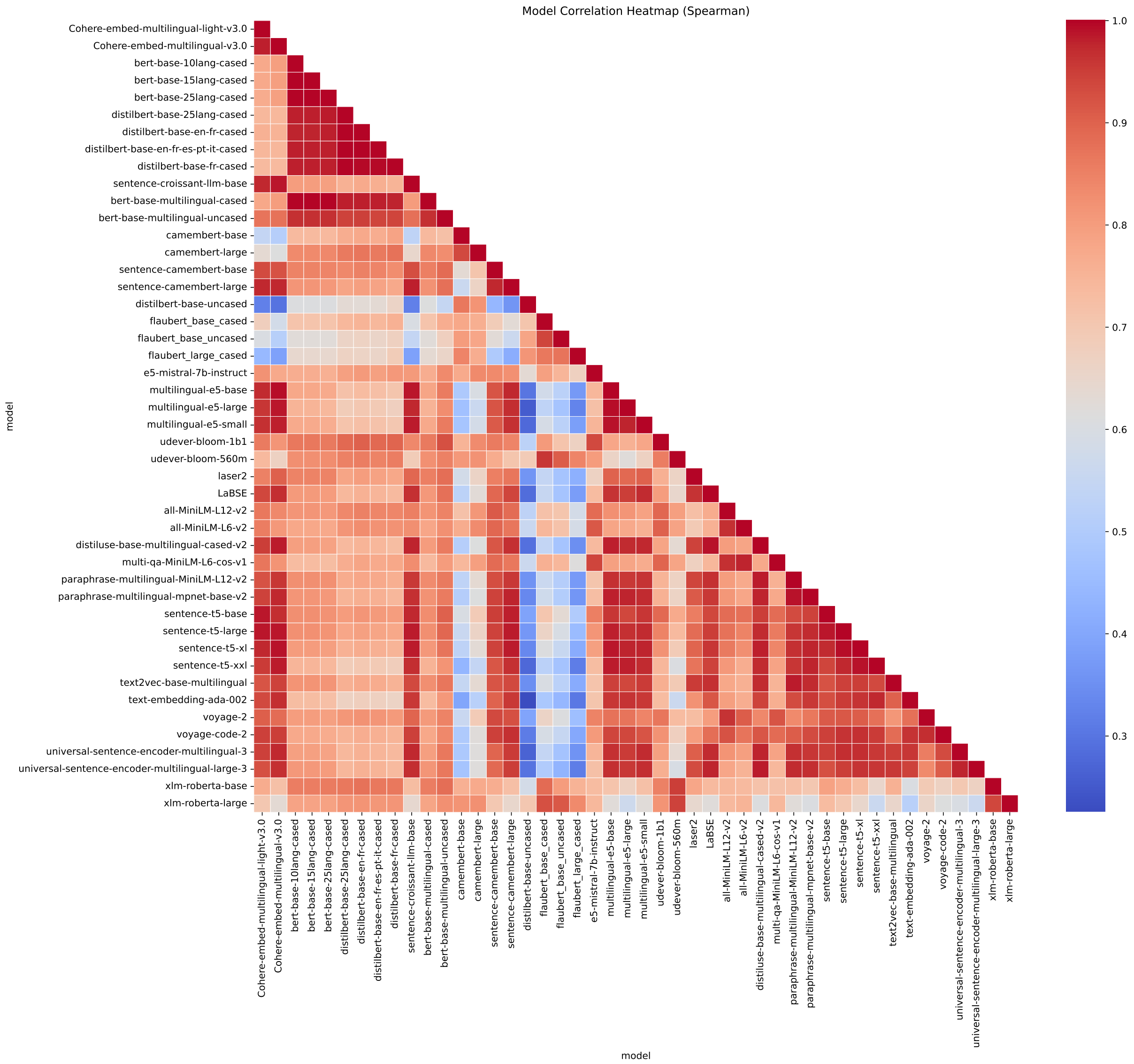

Figure 6: Heatmap representing the Spearman correlations in terms of performance across models.

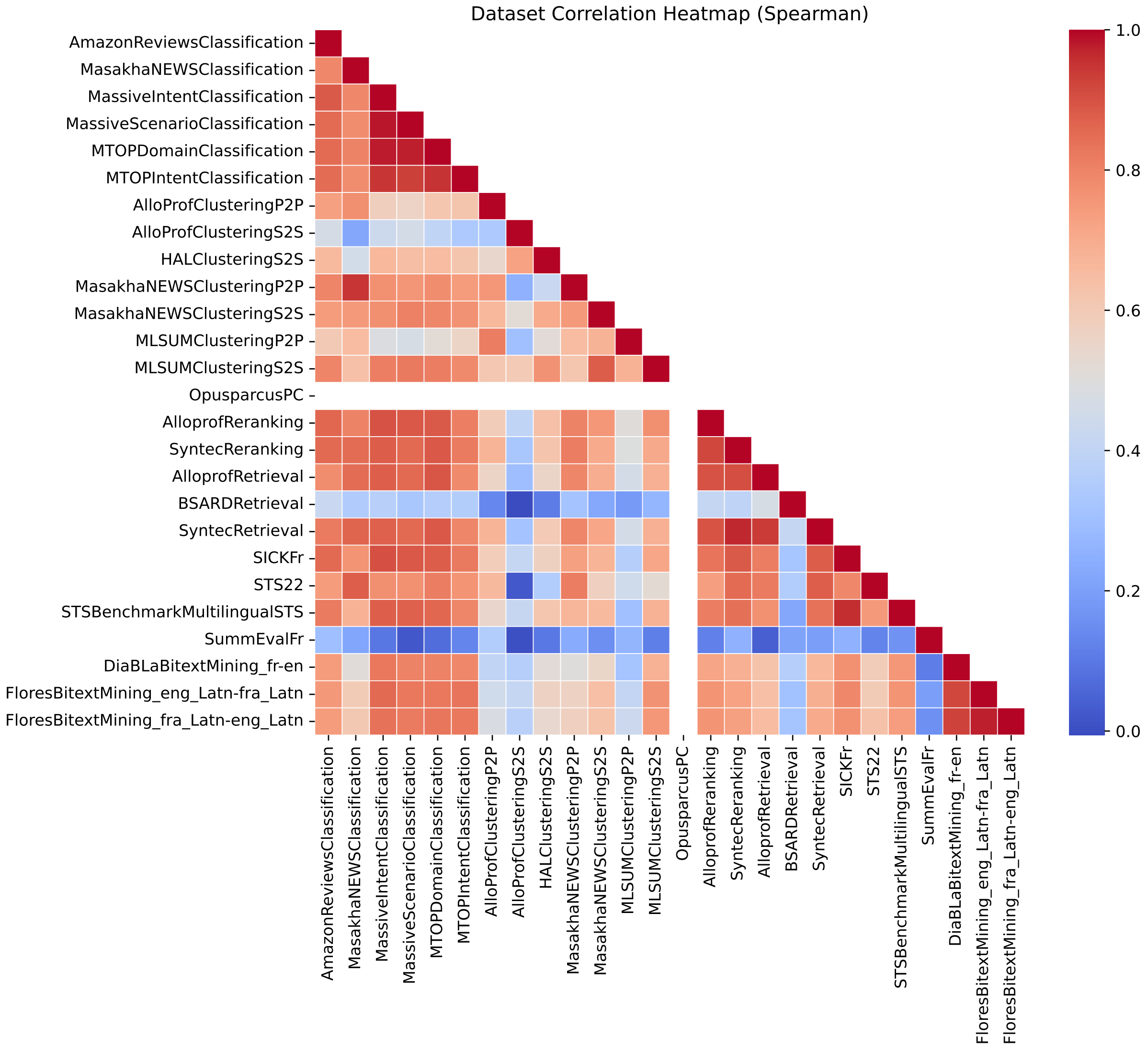

Figure 7: Heatmap representing the correlation in terms of model performance across datasets.

Model	Finetuned	Language	# params	Size (Gb)	Seq. Len.	Emb. dim.	License	Sentence sim
bert-base-multilingual-cased	No	multilingual	1,78e+08	0.71	512	768	Apache-2.0	No
bert-base-multilingual-uncased	No	multilingual	1,67e+08	0.67	512	768	Apache-2.0	No
camembert-base	No	french	1,11e+08	0.44	514	768	MIT	No
camembert-large	No	french	3,37e+08	1.35	514	1024	MIT	No
sentence-camembert-base	Yes	french	1,11e+08	0.44	128	768	Apache-2.0	Yes
sentence-camembert-large	Yes	french	3,37e+08	1.35	514	1024	Apache-2.0	Yes
distilbert-base-uncased	No	english	6,64e+07	0.27	512	768	Apache-2.0	No
embed-multilingual-light-v3.0	N/A	multilingual	N/A	N/A	512	384	Closed source	N/A
embed-multilingual-v3.0	N/A	multilingual	N/A	N/A	512	1024	Closed source	N/A
flaubert-base-cased	No	french	1,38e+08	0.55	512	768	MIT	No
flaubert-base-uncased	No	french	1,37e+08	0.55	512	768	MIT	No
flaubert-large-cased	No	french	3,73e+08	1.49	512	1024	MIT	No
bert-base-10lang-cased	No	multilingual	1,38e+08	0.55	512	768	Apache-2.0	No
bert-base-15lang-cased	No	multilingual	1,41e+08	0.56	512	768	Apache-2.0	No
bert-base-25lang-cased	No	multilingual	1,51e+08	0.61	512	768	Apache-2.0	No
distilbert-base-25lang-cased	No	multilingual	1,08e+08	0.43	512	768	Apache-2.0	No
distilbert-base-en-fr-cased	No	bilingual	6,86e+07	0.27	512	768	Apache-2.0	No
distilbert-base-en-fr-es-pt-it-cased	No	multilingual	7,61e+07	0.3	512	768	Apache-2.0	No
distilbert-base-fr-cased	No	french	6,17e+07	0.25	512	768	Apache-2.0	No
multilingual-e5-base	No	multilingual	2,78e+08	1.11	512	768	MIT	Yes
multilingual-e5-large	No	multilingual	5,60e+08	2.24	512	1024	MIT	Yes
multilingual-e5-small	No	multilingual	1,18e+08	0.47	512	384	MIT	Yes
e5-mistral-7b-instruct	Yes	english-plus	7,11e+09	28.44	32768	4096	MIT	Yes
udever-bloom-1b1	Yes	multilingual	1,07e+09	4.26	2048	1536	bloom-rail-1.0	Yes
udever-bloom-560m	Yes	multilingual	5,59e+08	2.24	2048	1024	bloom-rail-1.0	Yes
laser2	Yes	multilingual	4,46e+07	0.18	N/A	1024	BSD License	Yes
all-MiniLM-L12-v2	Yes	english-plus	3,34e+07	0.13	128	384	Apache-2.0	Yes
all-MiniLM-L6-v2	Yes	english-plus	2,27e+07	0.09	256	384	Apache-2.0	Yes
distiluse-base-multilingual-cased-v2	Yes	multilingual	1,35e+08	0.54	128	512	Apache-2.0	Yes
LaBSE	Yes	multilingual	4,72e+08	1.89	256	768	Apache-2.0	Yes
multi-qa-MiniLM-L6-cos-v1	Yes	english	2,27e+07	0.09	512	384	N/A	Yes
paraphrase-multilingual-MiniLM-L12-v2	Yes	multilingual	1,18e+08	0.47	128	384	Apache-2.0	Yes
sentence-t5-base	Yes	multilingual	1,10e+08	0.44	256	768	Apache-2.0	Yes
sentence-t5-large	Yes	multilingual	3,36e+08	1.34	256	768	Apache-2.0	Yes
sentence-t5-xl	Yes	multilingual	1,24e+09	4.97	256	768	Apache-2.0	Yes
sentence-t5-xxl	Yes	multilingual	4,87e+09	19.46	256	768	Apache-2.0	Yes
text2vec-base-multilingual	Yes	multilingual	1,18e+08	0.47	256	384	Apache-2.0	Yes
text-embedding-ada-002	N/A	multilingual	N/A	N/A	8191	1536	Closed source	N/A
universal-sentence-encoder-multilingual-3	Yes	multilingual	6,89e+07	0.28	N/A	512	Apache-2.0	Yes
universal-sentence-encoder-multilingual-large-3	Yes	multilingual	8,52e+07	0.34	N/A	512	Apache-2.0	Yes
xlm-roberta-base	No	multilingual	2,78e+08	1.11	514	768	MIT	No
xlm-roberta-large	No	multilingual	5,60e+08	2.24	514	1024	MIT	No
sentence-croissant-llm-base	Yes	french	1,28e+09	5.12	256	2048	MIT	Yes
paraphrase-multilingual-mpnet-base-v2	No	multilingual	2,78e+08	1.11	128	768	Apache-2.0	Yes
voyage-2	N/A	english	N/A	N/A	4000	1024	Closed source	N/A
voyage-code-2	N/A	english	N/A	N/A	16000	1536	Closed source	N/A

Table 3: Models included in the benchmark with their main characteristics. The size in Gb is estimated using the number of parameters counted as float32 numbers. *Sentence sim* refers to the fact that the model was trained on a task that favors semantic similarity.

	Model	**Overall Average**	**Bitext Mining**	**Clustering**	**Pair Classification**	**Summarization**	**Reranking**	**Classification**	**STS**	**Retrieval**
1	text-embedding-ada-002	**0.7**	0.95	**0.47**	1	0.3	**0.9**	**0.69**	0.78	0.46
2	sentence-t5-xxl	0.66	0.94	0.4	1	0.3	0.77	0.67	0.78	**0.43**
3	voyage-code-2	0.66	0.86	0.45	1	0.28	0.79	0.67	0.78	0.45
4	multilingual-e5-large	0.65	0.95	0.38	1	0.31	0.72	0.66	0.8	0.4
5	Cohere-embed-multilingual-v3	0.65	0.94	0.39	1	0.31	0.68	0.67	0.81	0.39
6	multilingual-e5-base	0.65	0.95	0.39	1	0.31	0.72	0.65	0.78	0.38
7	sentence-camembert-large	0.65	0.89	0.4	1	0.31	0.73	0.66	**0.82**	0.37
8	sentence-t5-xl	0.65	0.91	0.4	1	**0.32**	0.73	0.65	0.77	0.38
9	multilingual-e5-small	0.64	0.94	0.39	1	**0.32**	0.71	0.6	0.78	0.34
10	paraphrase-multilingual-mpnet-base-v2	0.63	0.94	0.39	1	0.29	0.69	0.63	0.78	0.34
11	Cohere-embed-multilingual-light-v3	0.63	0.89	0.38	1	0.31	0.7	0.61	0.78	0.36
12	sentence-croissant-llm-base	0.63	0.91	0.39	1	0.29	0.68	0.65	0.76	0.34
13	universal-sentence-encoder-multilingual-3	0.63	0.94	0.4	1	0.28	0.65	0.64	0.75	0.34
14	universal-sentence-encoder-multilingual-large-3	0.63	0.95	0.38	1	0.29	0.66	0.67	0.75	0.32
15	sentence-t5-large	0.62	0.9	0.39	1	0.3	0.69	0.62	0.75	0.35
16	voyage-2	0.62	0.76	0.41	1	0.31	0.73	0.59	0.72	0.4
17	distiluse-base-multi-cased-v2	0.61	0.94	0.36	1	0.28	0.63	0.63	0.75	0.3
18	LaBSE	0.61	**0.96**	0.36	1	0.3	0.61	0.65	0.74	0.23
19	paraphrase-multilingual-MiniLM-L12-v2	0.61	0.92	0.37	1	0.29	0.62	0.6	0.75	0.3
20	sentence-t5-base	0.6	0.83	0.38	1	0.3	0.64	0.58	0.74	0.3
21	text2vec-base-multilingual	0.59	0.92	0.31	1	0.29	0.61	0.56	0.78	0.22
22	sentence-camembert-base	0.58	0.72	0.32	1	0.29	0.64	0.57	0.78	0.29
23	laser2	0.54	0.95	0.26	1	**0.32**	0.46	0.57	0.67	0.09
24	e5-mistral-7b-instruct	0.52	0.37	0.39	1	**0.32**	0.62	0.58	0.65	0.23
25	bert-base-multi-uncased	0.51	0.76	0.35	1	0.31	0.53	0.48	0.57	0.11
26	all-MiniLM-L12-v2	0.51	0.48	0.3	1	0.27	0.57	0.52	0.66	0.3
27	udever-bloom-1b1	0.5	0.52	0.35	1	0.29	0.51	0.55	0.62	0.16
28	all-MiniLM-L6-v2	0.49	0.41	0.32	1	0.28	0.46	0.52	0.68	0.29
29	multi-qa-MiniLM-L6-cos-v1	0.49	0.38	0.29	1	0.28	0.53	0.51	0.67	0.29
30	bert-base-15lang-cased	0.48	0.75	0.33	1	0.29	0.45	0.46	0.5	0.05
31	bert-base-10lang-cased	0.48	0.75	0.33	1	0.29	0.45	0.46	0.5	0.05
32	bert-base-multi-cased	0.48	0.75	0.33	1	0.29	0.45	0.46	0.5	0.05
33	bert-base-25lang-cased	0.48	0.75	0.33	1	0.29	0.45	0.46	0.5	0.05
34	distilbert-base-en-fr-cased	0.47	0.65	0.34	1	0.31	0.42	0.45	0.54	0.06
35	distilbert-base-en-fr-es-pt-it-cased	0.47	0.65	0.34	1	0.31	0.42	0.45	0.53	0.06
36	distilbert-base-25lang-cased	0.47	0.65	0.34	1	0.31	0.42	0.45	0.53	0.06
37	distilbert-base-fr-cased	0.45	0.45	0.34	1	0.31	0.42	0.45	0.54	0.06
38	camembert-large	0.43	0.26	0.36	1	0.28	0.42	0.49	0.59	0.05
39	xlm-roberta-base	0.4	0.48	0.25	1	0.29	0.35	0.31	0.51	0
40	xlm-roberta-large	0.39	0.35	0.25	1	0.29	0.39	0.31	0.49	0.02
41	camembert-base	0.39	0.19	0.29	1	0.3	0.33	0.42	0.57	0.02
42	udever-bloom-560m	0.39	0.32	0.25	1	0.24	0.4	0.3	0.51	0.07
43	flaubert_base_cased	0.38	0.23	0.23	1	0.31	0.45	0.25	0.52	0.06
44	flaubert_base_uncased	0.35	0.12	0.18	1	0.29	0.46	0.23	0.43	0.06
45	distilbert-base-uncased	0.33	0.04	0.23	1	0.31	0.35	0.32	0.39	0.02
46	flaubert_large_cased	0.32	0.11	0.21	1	0.29	0.35	0.25	0.33	0.01

Table 4: Average performance of models per task type.

Model	**OpusparcusPC**	**AmazonReviewsClassification**	**MasakhaNEWSClassification**	**MassiveIntentClassification**	**MassiveScenarioClassification**	**MTOPDomainClassification**	**MTOPIntentClassification**
	Pair Classif.			Classification			
Cohere-embed-multilingual-light-v3.0	**1.00**	0.39	**0.83**	0.56	0.59	0.81	0.50
Cohere-embed-multilingual-v3.0	**1.00**	0.42	**0.83**	0.63	0.67	0.86	0.61
LaBSE	**1.00**	0.39	0.77	0.60	0.65	0.84	0.62
all-MiniLM-L12-v2	**1.00**	0.28	0.72	0.45	0.54	0.76	0.39
all-MiniLM-L6-v2	**1.00**	0.27	0.74	0.43	0.51	0.75	0.40
bert-base-10lang-cased	**1.00**	0.29	0.64	0.37	0.44	0.64	0.38
bert-base-15lang-cased	**1.00**	0.29	0.64	0.37	0.44	0.64	0.38
bert-base-25lang-cased	**1.00**	0.29	0.64	0.37	0.44	0.64	0.38
bert-base-multilingual-cased	**1.00**	0.29	0.64	0.37	0.44	0.64	0.38
bert-base-multilingual-uncased	**1.00**	0.29	0.76	0.38	0.44	0.64	0.39
camembert-base	**1.00**	0.30	0.66	0.31	0.39	0.58	0.29
camembert-large	**1.00**	0.31	0.71	0.36	0.46	0.68	0.42
distilbert-base-25lang-cased	**1.00**	0.29	0.67	0.35	0.44	0.62	0.35
distilbert-base-en-fr-cased	**1.00**	0.29	0.68	0.35	0.44	0.62	0.35
distilbert-base-en-fr-es-pt-it-cased	**1.00**	0.29	0.68	0.35	0.44	0.62	0.35
distilbert-base-fr-cased	**1.00**	0.29	0.68	0.35	0.44	0.62	0.35
distilbert-base-uncased	**1.00**	0.25	0.55	0.21	0.28	0.44	0.21
distiluse-base-multilingual-cased-v2	**1.00**	0.36	0.77	0.60	0.67	0.85	0.56
e5-mistral-7b-instruct	**1.00**	0.37	0.81	0.46	0.54	0.75	0.54
flaubert-base-cased	**1.00**	0.25	0.71	0.07	0.11	0.26	0.09
flaubert-base-uncased	**1.00**	0.24	0.63	0.06	0.11	0.28	0.09
flaubert-large-cased	**1.00**	0.22	0.56	0.16	0.23	0.24	0.10
laser2	**1.00**	0.34	0.66	0.53	0.59	0.76	0.57
multi-qa-MiniLM-L6-cos-v1	**1.00**	0.27	0.76	0.43	0.50	0.73	0.37
multilingual-e5-base	**1.00**	0.41	0.80	0.61	0.66	0.85	0.56
multilingual-e5-large	**1.00**	0.42	0.79	0.64	0.68	0.86	0.59
multilingual-e5-small	**1.00**	0.40	0.78	0.56	0.61	0.81	0.46
paraphrase-multilingual-MiniLM-L12-v2	**1.00**	0.37	0.76	0.58	0.65	0.78	0.48
paraphrase-multilingual-mpnet-base-v2	**1.00**	0.40	0.78	0.62	0.68	0.80	0.52
sentence-camembert-base	**1.00**	0.36	0.70	0.52	0.61	0.77	0.43
sentence-camembert-large	**1.00**	0.38	0.81	0.63	0.69	0.86	0.59
sentence-croissant-llm-base	**1.00**	0.35	0.79	0.59	0.65	0.86	0.63
sentence-t5-base	**1.00**	0.37	0.81	0.51	0.60	0.75	0.44
sentence-t5-large	**1.00**	0.41	0.80	0.57	0.64	0.80	0.48
sentence-t5-xl	**1.00**	0.44	0.80	0.61	0.66	0.85	0.54
sentence-t5-xxl	**1.00**	**0.46**	0.79	**0.66**	0.69	0.86	0.58
text-embedding-ada-002	**1.00**	0.44	0.82	0.65	0.71	**0.89**	**0.64**
text2vec-base-multilingual	**1.00**	0.34	0.74	0.52	0.58	0.72	0.45
udever-bloom-1b1	**1.00**	0.35	0.81	0.43	0.50	0.69	0.51
udever-bloom-560m	**1.00**	0.27	0.68	0.15	0.22	0.35	0.16
universal-sentence-encoder-multilingual-3	**1.00**	0.34	0.82	0.61	0.70	0.85	0.54
universal-sentence-encoder-multilingual-large-3	**1.00**	0.35	0.76	**0.66**	**0.73**	0.88	**0.64**
voyage-2	**1.00**	0.37	0.78	0.54	0.62	0.80	0.46
voyage-code-2	**1.00**	0.42	0.82	0.63	0.70	0.88	0.59
xlm-roberta-base	**1.00**	0.27	0.61	0.14	0.23	0.44	0.19
xlm-roberta-large	**1.00**	0.27	0.66	0.16	0.24	0.37	0.15

Table 5: Results obtained for each model on each dataset for the Classification and Pair Classification tasks.

	DiaBLaBitextMining:fr-en	**FloresBitextMining:fr-en**	**FloresBitextMining:en-fr**	**AlloprofRetrieval**	**BSARDRetrieval**	**SyntecRetrieval**	**AlloprofReranking**	**SyntecReranking**
Model	Bitext Mining			Retrieval			Reranking	
Cohere-embed-multilingual-light-v3.0	0.66	**1.00**	**1.00**	0.35	0.00	0.73	0.52	0.88
Cohere-embed-multilingual-v3.0	0.83	**1.00**	**1.00**	0.38	0.01	0.78	0.51	0.86
LaBSE	**0.88**	**1.00**	**1.00**	0.20	0.01	0.50	0.50	0.73
all-MiniLM-L12-v2	0.10	0.71	0.62	0.33	0.00	0.58	0.46	0.68
all-MiniLM-L6-v2	0.03	0.62	0.56	0.28	0.00	0.57	0.32	0.60
bert-base-10lang-cased	0.30	0.97	0.98	0.02	0.00	0.14	0.36	0.53
bert-base-15lang-cased	0.30	0.97	0.98	0.02	0.00	0.14	0.36	0.53
bert-base-25lang-cased	0.30	0.97	0.98	0.02	0.00	0.14	0.36	0.53
bert-base-multilingual-cased	0.30	0.97	0.98	0.02	0.00	0.14	0.36	0.53
bert-base-multilingual-uncased	0.36	0.95	0.98	0.06	0.00	0.27	0.39	0.66
camembert-base	0.04	0.26	0.25	0.00	0.00	0.05	0.24	0.41
camembert-large	0.06	0.40	0.32	0.02	0.00	0.13	0.33	0.51
distilbert-base-25lang-cased	0.11	0.92	0.91	0.01	0.00	0.16	0.32	0.52
distilbert-base-en-fr-cased	0.11	0.92	0.91	0.01	0.00	0.16	0.32	0.52
distilbert-base-en-fr-es-pt-it-cased	0.11	0.92	0.91	0.01	0.00	0.16	0.32	0.52
distilbert-base-fr-cased	0.06	0.63	0.65	0.01	0.00	0.15	0.32	0.52
distilbert-base-uncased	0.01	0.05	0.07	0.00	0.00	0.07	0.24	0.46
distiluse-base-multilingual-cased-v2	0.83	**1.00**	**1.00**	0.27	0.00	0.62	0.52	0.75
e5-mistral-7b-instruct	0.01	0.48	0.63	0.16	0.00	0.52	0.47	0.77
flaubert-base-cased	0.02	0.31	0.36	0.02	0.00	0.17	0.35	0.56
flaubert-base-uncased	0.03	0.25	0.08	0.02	0.00	0.17	0.35	0.57
flaubert-large-cased	0.01	0.15	0.17	0.01	0.00	0.01	0.26	0.43
laser2	0.86	**1.00**	**1.00**	0.03	0.01	0.24	0.35	0.56
multi-qa-MiniLM-L6-cos-v1	0.09	0.55	0.50	0.30	0.00	0.57	0.40	0.65
multilingual-e5-base	0.85	**1.00**	**1.00**	0.36	0.00	0.78	0.58	0.85
multilingual-e5-large	0.85	**1.00**	**1.00**	0.38	0.01	0.80	0.57	0.87
multilingual-e5-small	0.82	**1.00**	**1.00**	0.27	0.00	0.74	0.56	0.87
paraphrase-multilingual-MiniLM-L12-v2	0.78	**1.00**	**1.00**	0.27	0.00	0.63	0.49	0.75
paraphrase-multilingual-mpnet-base-v2	0.81	**1.00**	**1.00**	0.31	0.00	0.72	0.54	0.83
sentence-camembert-base	0.36	0.90	0.90	0.22	0.00	0.65	0.49	0.80
sentence-camembert-large	0.68	0.99	**1.00**	0.32	0.00	0.79	0.58	0.88
sentence-croissant-llm-base	0.74	**1.00**	**1.00**	0.30	0.00	0.73	0.53	0.83
sentence-t5-base	0.55	0.97	0.96	0.28	0.01	0.63	0.50	0.78
sentence-t5-large	0.71	0.99	0.99	0.35	0.00	0.70	0.58	0.80
sentence-t5-xl	0.76	0.99	0.99	0.40	0.01	0.72	0.63	0.83
sentence-t5-xxl	0.83	**1.00**	**1.00**	0.46	**0.06**	0.77	0.68	0.85
text-embedding-ada-002	0.86	0.99	0.99	0.52	0.02	**0.85**	nan	**0.90**
text2vec-base-multilingual	0.78	0.99	0.99	0.19	0.00	0.46	0.51	0.70
udever-bloom-1b1	0.03	0.75	0.78	0.12	0.00	0.36	0.39	0.63
udever-bloom-560m	0.08	0.50	0.37	0.02	0.00	0.19	0.29	0.51
universal-sentence-encoder-multilingual-3	0.82	**1.00**	**1.00**	0.35	0.00	0.68	0.56	0.74
universal-sentence-encoder-multilingual-large-3	0.84	**1.00**	**1.00**	0.34	0.00	0.61	0.55	0.77
voyage-2	0.32	0.99	0.98	0.45	0.01	0.73	0.64	0.83
voyage-code-2	0.60	**1.00**	0.99	**0.53**	0.02	0.81	**0.71**	0.87
xlm-roberta-base	0.21	0.70	0.53	0.00	0.00	0.00	0.26	0.44
xlm-roberta-large	0.13	0.65	0.26	0.01	0.00	0.06	0.29	0.49

Table 6: Results obtained for each model on each dataset for the Bitext Mining, Reranking and Retrieval tasks.

Model	AlloProfClusteringP2P	AlloProfClusteringS2S	HALClusteringS2S	MasakhaNEWSClusteringP2P	MasakhaNEWSClusteringS2S	MLSUMClusteringP2P	MLSUMClusteringS2S	SummEvalFr
	Clustering							Summ.
Cohere-embed-multilingual-light-v3.0	0.62	0.31	0.17	**0.57**	0.21	0.43	0.33	0.31
Cohere-embed-multilingual-v3.0	0.64	0.36	0.20	0.49	0.23	0.45	0.35	0.31
LaBSE	0.55	0.32	0.21	0.43	0.28	0.42	0.35	0.30
all-MiniLM-L12-v2	0.46	0.32	0.20	0.27	0.21	0.34	0.29	0.27
all-MiniLM-L6-v2	0.52	0.32	0.19	0.33	0.22	0.37	0.28	0.28
bert-base-10lang-cased	0.53	0.43	0.20	0.23	0.19	0.41	0.32	0.29
bert-base-15lang-cased	0.53	0.43	0.20	0.23	0.20	0.41	0.32	0.29
bert-base-25lang-cased	0.53	0.43	0.20	0.23	0.20	0.41	0.32	0.29
bert-base-multilingual-cased	0.51	0.43	0.21	0.23	0.20	0.41	0.32	0.29
bert-base-multilingual-uncased	0.61	0.35	0.21	0.34	0.22	0.43	0.31	0.31
camembert-base	0.54	0.29	0.14	0.25	0.13	0.41	0.27	0.30
camembert-large	0.59	0.34	0.18	0.31	0.26	0.44	0.35	0.28
distilbert-base-25lang-cased	0.57	0.43	0.19	0.25	0.22	0.41	0.31	0.31
distilbert-base-en-fr-cased	0.57	0.42	0.20	0.25	0.24	0.41	0.31	0.31
distilbert-base-en-fr-es-pt-it-cased	0.57	0.43	0.20	0.27	0.22	0.41	0.31	0.31
distilbert-base-fr-cased	0.57	0.43	0.20	0.26	0.22	0.41	0.31	0.31
distilbert-base-uncased	0.37	0.26	0.12	0.15	0.12	0.32	0.24	0.31
distiluse-base-multilingual-cased-v2	0.56	0.35	0.18	0.41	0.25	0.40	0.35	0.28
e5-mistral-7b-instruct	0.61	0.28	0.20	0.52	0.36	**0.46**	0.32	**0.32**
flaubert-base-cased	0.53	0.14	0.04	0.27	0.04	0.39	0.17	0.31
flaubert-base-uncased	0.43	0.13	0.02	0.14	0.04	0.33	0.15	0.29
flaubert-large-cased	0.41	0.22	0.05	0.16	0.05	0.38	0.19	0.29
laser2	0.48	0.26	0.12	0.22	0.11	0.35	0.27	**0.32**
multi-qa-MiniLM-L6-cos-v1	0.49	0.26	0.12	0.35	0.18	0.35	0.26	0.28
multilingual-e5-base	0.62	0.33	0.22	0.41	0.29	0.43	0.39	0.31
multilingual-e5-large	0.63	0.32	0.22	0.38	0.29	0.44	0.38	0.31
multilingual-e5-small	0.61	0.33	0.19	0.44	0.34	0.43	0.38	**0.32**
paraphrase-multilingual-MiniLM-L12-v2	0.56	0.42	0.23	0.33	0.27	0.40	0.37	0.29
paraphrase-multilingual-mpnet-base-v2	0.54	0.45	0.24	0.40	0.29	0.41	0.38	0.29
sentence-camembert-base	0.59	0.39	0.20	0.27	0.16	0.36	0.27	0.29
sentence-camembert-large	0.63	0.42	0.24	0.47	0.27	0.42	0.32	0.31
sentence-croissant-llm-base	0.64	0.33	0.23	0.43	0.33	0.43	0.34	0.29
sentence-t5-base	0.58	0.36	0.18	0.54	0.28	0.41	0.30	0.30
sentence-t5-large	0.62	0.40	0.19	0.55	0.26	0.42	0.32	0.30
sentence-t5-xl	0.60	0.41	0.20	0.54	0.30	0.42	0.34	**0.32**
sentence-t5-xxl	0.61	0.44	0.21	0.53	0.25	0.42	0.35	0.30
text-embedding-ada-002	**0.65**	**0.54**	0.26	0.52	**0.46**	0.45	**0.42**	0.30
text2vec-base-multilingual	0.49	0.33	0.16	0.32	0.19	0.36	0.30	0.29
udever-bloom-1b1	0.62	0.27	0.14	0.45	0.21	0.44	0.30	0.29
udever-bloom-560m	0.54	0.22	0.08	0.20	0.08	0.36	0.25	0.24
universal-sentence-encoder-multilingual-3	0.57	0.38	0.19	0.51	0.38	0.44	0.36	0.28
universal-sentence-encoder-multilingual-large-3	0.54	0.38	0.19	0.42	0.37	0.41	0.38	0.29
voyage-2	0.58	0.42	0.25	0.41	0.35	0.45	0.39	0.31
voyage-code-2	0.62	0.51	**0.27**	0.48	0.39	0.45	0.41	0.28
xlm-roberta-base	0.52	0.20	0.09	0.23	0.07	0.40	0.24	0.29
xlm-roberta-large	0.57	0.21	0.06	0.27	0.06	0.43	0.19	0.29

Table 7: Results obtained for each model on each dataset for the Summarization and Clustering tasks.