MTEB-French: Resources for French Sentence Embedding Evaluation and Analysis

Anonymous ACL submission

Abstract

Recently, numerous embedding models have been made available and widely used for various NLP tasks. The Massive Text Embedding Benchmark (MTEB) has primarily simplified the process of choosing a model that performs well for several tasks in English, but extensions to other languages remain challenging. This is why we expand MTEB to propose the first massive benchmark of sentence embeddings for French. We gather 15 existing datasets in an easy-to-use interface and create three new French datasets for a global evaluation of 8 task categories. We compare 51 carefully selected embedding models on a large scale, 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 exceptionally 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 word representations from contextual relationships in vast amounts of text. 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. Many models now exist with various architectures, monolingual or multilingual, pre-trained or fine-tuned (Naseem et al., 2021; Ding et al., 2023).

In this work, our primary objective is to introduce a large-scale embedding benchmark for French to enable the research community and industry to select the most relevant embedding methods based on one's specific needs, such as being open-source, versatile or targeted toward a particular task, having a small embedding dimension, the ability to process long texts or their performance. To achieve this goal, we undertake significant efforts in collecting datasets to conduct a broad comparison of models. We ensure that the datasets cover various tasks within a common, easy-to-use framework, and we create three new quality-checked datasets to enhance this collection. We select a diverse range of models, including prominent French and multilingual models deemed most efficient. The results of our study already enable the community to make informed model selections, whether for general purposes or specific tasks. Additionally, our implementation is open to the community and features a public leaderboard, allowing the results to evolve with new models or datasets. With this first large-scale comparison, we perform an in-depth analysis of the results, confirming well-known findings such as the correlation between performance and model/embedding dimensions and uncovering interesting nuances.

2 Related Work

Sentence Embeddings Sentence embeddings are required for many language tasks, such as Semantic Textual Similarity (STS) and knowledge retrieval. Many models have been proposed in the literature, leveraging pooling strategies (Devlin et al., 2019; Muennighoff, 2022) or similarity fine-tuning (Reimers and Gurevych, 2019) using a contrastive framework (Gao et al., 2021; Neelakantan et al., 2022; Ni et al., 2021; Wang et al., 2022; Zhang et al., 2023), leveraging prompts (Wang et al., 2023) or a two steps training process (Chen et al., 2024; Lee et al., 2024). Few French-language models have been proposed in the literature (Martin et al., 2019; Le et al., 2020). Most French models for sentence embeddings have been developed by the

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

open-source community[2], by fine-tuning models like *CamemBERT*(Martin et al., 2019) or *CroissantLLM*(Faysse et al., 2024).

Benchmarks Embedding models are generally compared on specific tasks, such as information retrieval, STS or reranking (Thakur et al., 2021; Agirre et al., 2016; Wang et al., 2021). Other works evaluate embedding models on multiple tasks (Wang et al., 2018; et al., 2022; Conneau and Kiela, 2018) or compare meta-embeddings (García-Ferrero et al., 2021). The most comprehensive benchmark to date is MTEB (Muennighoff et al., 2022). MTEB still has a critical limit: it mainly focuses on English. Some initiatives already extended this benchmark to other languages, such as Chinese (Xiao et al., 2024) 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 extends it to a new language.

3 MTEB for French

In this section, we describe 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 New Datasets

We identified 7 datasets relevant to French in the existing MTEB, which we assume are of good quality. We complemented these with 8 external relevant datasets proposed in the literature, such as BSARD (Louis and Spanakis, 2022) and Alloprof (Lefebvre-Brossard et al., 2023), which are proven to be good quality. We created 3 new ones presented in Table 1 and assessed their quality with various procedures and metrics. In addition to all performed checks, we run multiple models on these datasets and provide results to show that they are neither trivial nor impossible to solve (see Tables 10, 11, 12 and 13).

Therefore, as of today, our French MTEB runs on 18 datasets. Some datasets are framed differently according to the task category they are used with. For example, MasakhaNEWS dataset (Adelani et al., 2023) is used for both Classification (*MasakhaNEWSClassification*) and Clustering (*MasakhaNEWSClusteringS2S* and *MasakhaNEWSClusteringP2P*). Table 3 shows details of each task data used for running the benchmark.

This section describes the 3 new datasets we introduce, quality checks performed and an analysis of the semantic similarities between datasets.

3.1.1 Syntec (Retrieval)

The Syntec French collective bargaining agreement[3] comprises 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. The articles have been scraped for use as documents. Four annotators were divided into two groups. Each group was given half of the articles and asked to choose an article and write a question about it. Each annotator wrote 25 questions. Thus, a hundred questions have been manually created and paired with the articles containing the answer[4]. Examples of the dataset are available in the appendix Figure 5. This dataset could also be used for text classification, clustering or topic modeling. Regarding quality checks, every article's integrity has been reviewed while manually creating questions. We also manually checked that the questions could only be answered using the annotated article.

3.1.2 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[5]. We extracted IDs, titles and the author's choice among domain labels. The last 2 are provided by authors when submitting their papers to HAL. Since domain annotations are provided, the dataset can be used for many tasks, such as topic modeling or text classification. To ensure the dataset quality is suitable for a benchmark, further data cleaning has been performed:

- Duplicates are eliminated, retaining unique publications for each field.
- Irrelevant titles (due to API indexing mistakes) or titles in languages other than French have been manually removed.

[2]Models on the HuggingFace hub: *sentence-camebert*, *sentence_croissant_alpha_v0.3*, *Solon-embeddings-large-0.1*.

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

[4]The link to the publicly accessible dataset on Hugging Face will be added once the anonymization process is complete.

[5]The link to the publicly accessible dataset on Hugging Face will be added once the anonymization process is complete.

Dataset	Syntec	HAL	SummEvalFr
Samples	100 queries 90 documents	26233 samples 10 classes	100 texts 1100 human summaries 1600 machine summaries
Creation process	Scraping of Syntec collective bargaining agreement with articles as documents. Writing queries corresponding to articles.	Scraping of HAL articles with *id*, *title* and *domain*. Further cleaning with deduplication, language filtering and class subsampling.	Translation from English to French with Deepl of the SummEval dataset.
Annotation process	4 annotators divided into 2 groups. Each group was given half of the articles and asked to choose an article and ask a question about it. Each annotator wrote 25 questions.	Annotations provided by authors when submitting their paper. They choose the *domain* between existing academic fields.	Detailed annotation process provided in Fabbri et al. (2021).
Quality checks	Human verification of annotations.	Baseline models for classification and topic modeling.	Correlation between BLEU and ROUGE scores of the French and the original English datasets. LLM as-a-judge translation rating and human verification.

Table 1: New datasets details with the number of samples, the creation process, the annotation process and the quality checks. All datasets are test splits.

- Samples belonging to *domain* classes with less than 500 samples were removed, which leads us to keep only 10 classes.
- Subsampling was performed on 2 classes containing more than 10k samples each to lower the number of samples and mitigate the unbalance of the dataset.

More details about this process are provided in the appendix A.2 along with some extracts in Figure 6. We make the dataset publicly available in both their raw and clean versions. We use this dataset in a clustering setup to cluster publications by their title and use the domain as ground truth. To ensure the quality of this dataset, we run 3 baseline models for classification: *TF-IDF + SVM*, a fine-tuned *Camembert* (Martin et al., 2019) and *GPT-4* leveraging In-Context Learning (ICL). Furthermore, we run one baseline model for topic modeling: Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and report scores in the appendix A.2.

3.1.3 SummEvalFr (Summarization)

The original SummEval dataset (Fabbri et al., 2021) consists of 100 news articles from the CNN/DailyMail dataset. Each article has 11 human-written summaries and 16 machine-generated summaries annotated by 8 people with a score for coherence, consistency, fluency, and relevance. We translated it from English to French using DeepL API[6]. Since MTEB evaluation is based on the embedding similarity between machine-generated and human-generated summaries, we propose to compute the ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) metrics between machine and human summaries for both French and English version. In Table 2, we report the average of the scores as well as their correlations between the two languages. The correlation is high (above 0.7), showing that the word and n-gram overlap between human and machine summaries is highly preserved in the French version. One may argue that computing the metric on fully translated texts (human and machine summaries are both translated from English) may introduce biases and not assess the quality of the translations. For this purpose, we ensure the French human summaries are correctly translated from English. We use an LLM as-a-judge (Zheng et al.,

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

2023) where given the original human summary in English and its translation in French, the model rates the quality of the translation from 0 to 10, with 0 being of very bad quality and 10 being excellent. The prompt is available in Figure 8. Additionally, we manually check random translations with ratings between 9 and 10 to ensure the rating is relevant. We do the same for all translations with a score less than 9 and correct them (see the rating distribution in Table 6).

Dataset	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
SummEval	0.205	0.292	0.099	0.193
SummEvalFr	0.276	0.302	0.117	0.194
Correlation En-Fr	0.70	0.85	0.80	0.84

Table 2: Average ROUGE and BLUE scores computed between machine summaries and human summaries for the original English SummEval and its translation to French. The correlations of the individual scores between English and French are also reported.

3.1.4 Data for the Reranking task

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 found no French dataset that natively exhibits such a structure. Thus, to evaluate this task, we built data for the reranking task based on the *Syntec* and *Alloprof* (Lefebvre-Brossard et al., 2023) datasets. These already feature queries and labeled relevant documents. Irrelevant ones were added using the following process:

- To avoid bias, we use the BM25 algorithm (Robertson and Jones, 1976) (which is a deterministic method) to rank documents in terms of relevance regarding each query.
- The top 10 documents that are not labeled as relevant constitute the negative samples.

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

3.1.5 Similarity analysis

We investigate the proximity between the datasets' topics to give insights about the benchmark contents. The methodology introduced by Muennighoff et al. (2022), i.e. computing an average embedding of samples from each dataset, is used to build a dataset-similarity matrix (displayed in appendix Figure 3). The distances between averaged embedding vectors of each dataset (which range from 0.89 to 1 in Figure 3) remain hard to interpret into a dataset semantic proximity. Thus, we complement this by observing the dataset's clouds of embedding in a 2D plane using PCA in Figure 4.

Figures 4 and 3 seem to correlate, showing high similarity between two datasets when the same underlying data is used in different tasks. Dataset topics are pretty close, with some exceptions, such as the 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 help to understand the link between the results and the datasets' topics.

3.2 Models

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

- **Quantity:** The aim was to compare a substantial number of models (51 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 and some English models to asses their language-transferring 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 efficiently storing 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 are multilingual. Language is usually acquired during pre-training, but sometimes, models familiarize themselves 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 that claimed to be English (e.g. *all-MiniLM-L12-v2*[7]).

- ***Model types:*** 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 contrastive 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 1, and all of their characteristics are summarized in appendix Table 7. 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 effectively. We provide information on the models' licenses as reported in the Hugging Face hub[8]. However, we encourage readers to conduct further research before utilizing a model.

3.3 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@10), Summarization and STS (Spearman correlation based on cosine similarity). BitextMining tasks are excluded from the average performance scores and therefore the figures, as this task evaluates 2 languages instead of one, and this benchmark focuses only on one language (French). We present the results for both *DiaBlaBitextMining* and *FloresBitextMining* in Table 12.

Using the overall benchmark results, our goal will be to answer the following research questions:
Q1: Is a model outstanding 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 to facilitate model choice for specific applications.
Q2: Are there any links between the model characteristics and performance?
In section 3.2, 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: Do monolingual models have multilingual capabilities?
We interrogate the ability of a model trained exclusively in one language to perform well in 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 3.1.5), subsequent analysis will be conducted regarding how they rank models. Additionally, complementary insights will be derived from examining cor-

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

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

relations of models relative 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 9, 10, 11, 12 and 13. Figure 1 shows the critical difference diagram of average score ranks.

As in MTEB (Muennighoff et al., 2022), no model claims state-of-the-art in all tasks even if the *text-embedding-3-large* model is in first place on average on all tasks (see Table 9). It ranks first for the classification and reranking tasks. For the clustering task, *text-embedding-ada-002* is the best model. The models *voyage-code-2*, *text-embedding-3-small* and *mistral-embed* share the top positions in the retrieval task ranking. For the pair classification task, *laser2* is ahead of its competitors. Finally, *sentence-camembert-large* leads on the STS task and *multilingual-e5-small* has the best results for summarization.

Figure 1 shows a global model comparison across all datasets. The models are arranged horizontally according to their performance, with the best models on the left. The black bars represent the statistical equivalence between the models' performances. The statistically equivalent top performers for this benchmark are OpenAI's models *text-embedding-3-large*, *text-embedding-3-small* and *text-embedding-ada-002*. Interestingly, many models do not show a significant performance gap between their base and large flavours. Some French models stand out among the multilingual models, such as *Solon-embeddings-large-0.1*, *sentence_croissant_alpha_v0.3* and *sentence-camembert-large*.

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:

- *Tuned for sentence similarity*: 0.727
- *Finetuned vs pretrained*: 0.544
- *Model number of parameters*: 0.49
- *Embedding dimension*: 0.452
- *Closed source*: 0.449
- *Max sequence length*: 0.336
- *Multilingual*: 0.103
- *English*: 0.025
- *English but tuned on other languages*: -0.025
- *French*: -0.134
- *Bilingual*: -0.135

Additionally, all cross-correlations between characteristics are reported in appendix Figure 10.

As expected, the score most strongly correlates with whether the evaluated models were trained on a sentence similarity task. Of course, this criterion is 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 1). However, some small models perform very well on the benchmark, such as the standard version of the multilingual universal sentence encoder or *Solon-embeddings-base-1.0*. Notably, the maximum sequence length, while an important criterion for generative tasks with LLMs, is less correlated with performance than the other dimensions. This can be explained by many datasets containing relatively small texts (see appendix Table 3 showing that 14 datasets have less than 50 tokens).

Regarding language, it is surprising that good performance is not particularly correlated with French models in particular. In reality, the other aspects of the models, such as being fine-tuned for similarity, prevail. Nevertheless, we can highlight the excellent performance of a few French models such as *sentence-camembert* and *sentence-croissant* and *Solon-embeddings*.

Lastly, we emphasize that closed-source models perform well on this benchmark (*text-embeddings*,

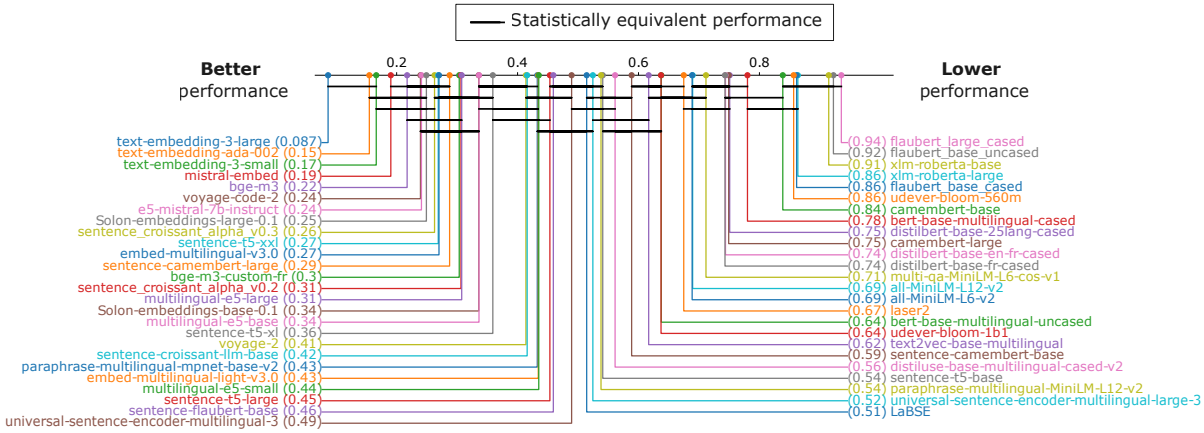

Figure 1: Critical difference diagram representing the significant rank gaps between models. The axis represents the normalized average rank of the models (lower is better). The black bars indicate that the difference in models' rank is not statistically significant, i.e. lower than the critical difference.

mistral-embed and *voyage*), but we lack information about their characteristics. As more open-source well-performing models get added in the future, we could expect this correlation to decrease. Note that the correlation between sequence length and performance could be dragged by closed-source models that have generally larger sequence lengths.

Q3: Do monolingual models have multilingual capabilities?

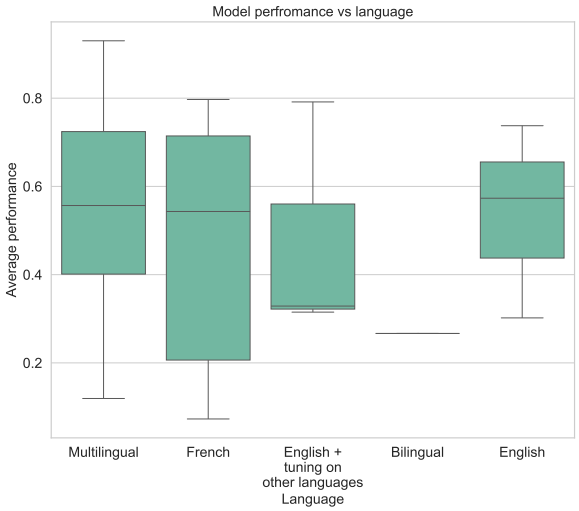

Figure 2: Model performance depending on the language of the data they have been trained on.

We also studied the capabilities of models on the French language when the language of the training data varies. It is surprising to note the absence of a clear correlation between the language the model is trained on and its performance on French, as shown by the large standard deviation in Figure 2. 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).

This is explained by the fact that a large part of the selected French models generate embeddings using a pooling strategy. Only a few are sentence transformer models, for which the pooled representation is part of the model and trained with it, leading to higher-quality embeddings. This is endorsed by the excellent results of *sentence-camembert-large*, a sentence transformer model trained on French corpus and confirms the recent findings in terms of model architecture (Gao et al., 2021).

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 cause 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 12. Except for two datasets (*MasakhaNEWSClusteringP2P*, *SummEvalFr*), the correlations, on average, are high. There is still enough diversity to make each dataset

interesting for the French MTEB benchmark. Two groups (*SyntecReranking*/ *SyntecRetrieval*, *MassiveScenarioClassification*/ *MTOPDomainClassification*/ *MassiveIntentClassification*) exhibit notably high correlations ($\sim$0.97). It is interesting to point out some sub-diagonal correlation blocks. The datasets being arranged by task indicate that models behave slightly 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 11. Strong correlations in behavior emerge among the variants of the same models (e.g. DistilBERT, sentence-croissant, sentence-t5, e5, etc.). Correlations are also generally observed among numerous models trained using the sentence transformers framework (Reimers and Gurevych, 2019), as well as proprietary models, e.g. from Cohere and OpenAI. Conversely, these models fine-tuned for sentence similarity, show minimal correlation with pre-trained models for which token-embedding pooling techniques are employed.

5 Conclusion and perspectives

In this work, we introduce a large-scale embedding benchmark for French to enable the research community and industry to select the most relevant embedding methods based on their specific needs. We undertake significant efforts in collecting 15 datasets and create 3 new quality-checked ones to enhance this collection. The whole French benchmark runs on 26 tasks. We select a diverse range of 51 models, including prominent French and multilingual models deemed most efficient to conduct a broad comparison. Our implementation is open to the community and features a public leaderboard, allowing the results to evolve with new models or datasets. After an in-depth analysis of the results, OpenAI models perform significantly better than the other models. However, other models should be considered for their performance on specific tasks, being open source or having a small embedding dimension.

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 path 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 and German), enabling a more comprehensive evaluation of multilingual model performance.

6 Limitations

Native French resources unavailability The availability of resources natively in French is an obvious limitation of our work. Regarding models, there are far fewer options than with more widespread languages such as English. Indeed, most of the existing French embedding models we found are trained using either older architectures or methods, unlike most recent multilingual models such as *NV-Embed-v1* (Lee et al., 2024) or *e5-mistral-7b-instruct* (Wang et al., 2023). Comparing models by family would be beneficial, particularly for evaluating French models against multilingual models on the same architecture using the same training technique. Resource limitations also apply to datasets. For example, the summarization task dataset is translated, which can be less relevant than a natively French dataset. We have also built datasets for reranking tasks using existing ones from retrieval task because we could not find any in French. This construction process introduces a bias as the model performance on both tasks may be correlated (see Figure 12). We preferred to propose datasets even if they could introduce biases rather than not address the task in the benchmark. Note that each task type can be considered individually. We hope additional resources will be developed in the French-speaking community to enrich our comparison.

Benchmark validity over time As with all benchmarks, their reliability over time can be discussed as the field evolves fast. The models selected for the analysis conducted in this paper are those available at this time, new outperforming models will be created and shall be evaluated. Our work extends MTEB and thus simplifies the addition of new datasets for evaluation and allows running new models. With this effort, we hope this will simplify the evaluation of new models proposed by the community to keep our work up to date.

Data contamination issues Bias may exist for models that use the training sets of the provided evaluation datasets for their training. It considerably improves their performance on the benchmark, favouring them over other models. This is particularly worrying for models that do not communicate about the datasets used during training, such as proprietary models. Generally speaking, it would be interesting to calculate the similarity between the datasets used to train the models and those used to test them to check that they are far enough apart to draw general conclusions.

Focus on sentence embeddings Finally, like the original version of MTEB, the comparison focuses mainly on sentence embeddings. Other tasks could be added to cover word embeddings and, therefore, more NLP tasks.

References

David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, Jesujoba Oluwadara Alabi, Atnafu Lambebo Tonja, Christine Mwase, Odunayo Ogundepo, Bonaventure F. P. Dossou, Akintunde Oladipo, Doreen Nixdorf, Chris C. Emezue, Sana Al-Azzawi, Blessing K. Sibanda, Davis David, Lolwethu Ndolela, Jonathan Mukiibi, Tunde Oluwaseyi Ajayi, Tatiana Moteu Ngoli, Brian Odhiambo, Abraham Toluwase Owodunni, Nnaemeka Obiefuna, Shamsuddeen Hassan Muhammad, Saheed Salahudeen Abdullahi, Mesay Gemeda Yigezu, Tajuddeen Rabiu Gwadabe, Idris Abdulmumin, Mahlet Taye Bame, Oluwabusayo Olufunke Awoyomi, Iyanuoluwa Shode, Tolulope Anu Adelani, Habiba Abdulganiy Kailani, Abdul-Hakeem Omotayo, Adetola Adeeko, Afolabi Abeeb, Anuoluwapo Aremu, Olanrewaju Samuel, Clemencia Siro, Wangari Kimotho, Onyekachi Raphael Ogbu, Chinedu E. Mbonu, Chiamaka Ijeoma Chukwuneke, Samuel Fanijo, Jessica Ojo, Oyinkansola F. Awosan, Tadesse Kebede Guge, Sakayo Toadoum Sari, Pamela Nyatsine, Freedmore Sidume, Oreen Yousuf, Mardiyyah Oduwole, Ussen Kimanuka, Kanda Patrick Tshinu, Thina Diko, Siyanda Nxakama, Abdulmejid Tuni Johar, Sinodos Gebre, Muhidin A. Mohamed, Shafie Abdi Mohamed, Fuad Mire Hassan, Moges Ahmed Mehamed, Evrard Ngabire, and Pontus Stenetorp. 2023. Masakhanews: News topic classification for african languages. In *International Joint Conference on Natural Language Processing*.

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.

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.

Rachel Bawden, Eric Bilinski, Thomas Lavergne, and Sophie Rosset. 2021. Diabla: A corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation*, 55:635–660.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation.

Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory. 2022. SemEval-2022 task 8: Multilingual news article similarity. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 1094–1106, Seattle, United States. Association for Computational Linguistics.

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.

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.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

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.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

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.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2024. Nv-embed: Improved techniques for training llms as generalist embedding models.

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.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2950–2962, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Antoine Louis and Gerasimos Spanakis. 2022. A statutory article retrieval dataset in French. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6789–6803, Dublin, Ireland. Association for Computational Linguistics.

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.

Julian McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In *Proceedings of the 7th ACM Conference on Recommender Systems*, RecSys '13, page 165–172, New York, NY, USA. Association for Computing Machinery.

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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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

Stephen E. Robertson and Karen Spärck Jones. 1976. Relevance weighting of search terms. *J. Am. Soc. Inf. Sci.*, 27:129–146.

Thomas Scialom, Paul-Alexis Dray, Sylvain Lamprier, Benjamin Piwowarski, and Jacopo Staiano. 2020. MLSUM: The multilingual summarization corpus. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8051–8067, Online. Association for Computational Linguistics.

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.

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.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packaged resources to advance general chinese embedding.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

A Supplementary materials for datasets

A.1 All datasets

Table 3 displays the size of each dataset along with the average number of tokens per sample and their references. The dataset's content was tokenized using *cl100k_base* encoding. For Retrieval, the two numbers refer to the queries and the documents. For Reranking, the three numbers refer to the queries, the pairs of queries with relevant documents and the pairs of queries with irrelevant ones, respectively. The pairs of queries and documents are obtained from the 90 documents extracted. For *SummEvalFr*, the three numbers refer to the texts, human and machine summaries, respectively.

Figure 3 represents the semantic similarity between each dataset. The methodology was as follows: 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).

We complement this analysis by observing the dataset's clouds of embedding in a 2D plane using PCA in Figure 4.

A.2 Created datasets

Syntec Figure 5 shows an extract from the Syntec dataset with a document and a query relative to this document.

HAL Figure 6 is an extract from the HAL dataset. Table 4 lists the distribution of classes (*domain* field) for the HAL dataset on *raw* subset and *mteb_eval* subset, which is used for MTEB evaluation. Labels descriptions can be found at this URL: https://api.archives-ouvertes.fr/ref/domain/?q=*:*&rows=393 or in Table 4. After pre-processing, *mteb_eval* covers titles from 10 domains as classes with less than 500 samples were removed. In the MTEB evaluation subset of the dataset, titles composed of 2 words or less have been removed (371 samples), resulting in an average word count of 13.4. Figure 7 shows the word count distribution per title. Furthermore, the dataset has been cleaned up by manually removing all non-French titles. Additionally, it can be observed in Table 4 that in the original *raw* dataset, the *shs* and *sdv* classes represent by far the majority of the dataset samples with respectively 58706 samples (73%) and 11049 samples (13%). In order to mitigate the class imbalance while preserving the majority of those classes, they have been randomly subsampled to 6701 and 4803 samples. Furthermore, baseline models have been trained and tested to assess the usability of this dataset in other tasks, such as classification and topic modeling. Table 5 shows the results obtained.

SummEvalFr Extracts of humans and machine summaries translated in French from SummEvalFr and the original ones in English from SummEval (Fabbri et al., 2021) are shown in Figure 9. As explained in section 3.1.3, we use a LLM to evaluate the quality of translations for human summaries, we provide the prompt used with *GPT-4* for this evaluation in Figure 8.

Table 6 shows the distribution of ratings given by the LLM. With the scale being 10, we manually verify random samples rated above 9. We verify all samples with ratings under 9 and those with no provided rating (N/A) due to the triggering of the OpenAI content management policy. The LLM suggests that 60 samples are not correctly translated. These were verified manually, and after checking, less than 10 samples only needed to be corrected.

B Supplementary materials for correlation analysis

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

Figure 10 represents cross-correlations between models' performances and their studied characteristics as a heatmap.

Figure 11 represents the Spearman correlations in terms of performance across models.

Figure 12 represents the Spearman correlations in terms of performance across datasets.

C Supplementary materials for models

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

For evaluating prompt-based models such as *intfloat/e5-mistral-instruct-7b*, we provide the prompts we used in Table 8.

D Evaluation results

This section presents the results obtained for each model on each task. To be relevant, we used the same metrics as in MTEB, which varies from one type of task to another:

Dataset x Task	Average # tokens	# samples	Reference	License
AmazonReviewsClassification	49.6	5000	McAuley and Leskovec (2013)	*N/A*
MasakhaNEWSClassification	1398.2	422	Adelani et al. (2023)	AFL-3.0
MassiveIntentClassification	11.4	2974	FitzGerald et al. (2023)	*N/A*
MassiveScenarioClassification	11.4	2974	FitzGerald et al. (2023)	*N/A*
MTOPDomainClassification	12.5	3193	Li et al. (2021)	*N/A*
MTOPIntentClassification	12.5	3193	Li et al. (2021)	*N/A*
AlloProfClusteringP2P	1021.8	2556	Lefebvre-Brossard et al. (2023)	MIT
AlloProfClusteringS2S	8.8	2556	Lefebvre-Brossard et al. (2023)	MIT
HALClusteringS2S	25.6	26233	*Introduced by our paper*	Apache-2.0
MasakhaNEWSClusteringP2P	1398.1	422	Adelani et al. (2023)	AFL-3.0
MasakhaNEWSClusteringS2S	21.7	422	Adelani et al. (2023)	AFL-3.0
MLSUMClusteringP2P	1062.1	15828	Scialom et al. (2020)	Other
MLSUMClusteringS2S	20.8	15828	Scialom et al. (2020)	Other
OpusparcusPC	9.7	1007	Creutz (2018)	CC-BY-NC-4.0
PawsX	34.9	2000	Yang et al. (2019)	Other
STSBenchmarkMultilingualSTS	18.4	1379	May (2021)	*N/A*
STS22	722.1	104	Chen et al. (2022)	*N/A*
SICKFr	15.1	4906	`https://huggingface.co/datasets/Lajavaness/SICK-fr`	Apache-2.0
DiaBLaBitextMining	12.02	5748	Bawden et al. (2021)	CC-BY-SA-4.0
FloresBitextMining	33.42	1012	Goyal et al. (2021)	CC-BY-SA-4.0
AlloprofReranking	48.3 - 1179.4 - 1196.4	2316 - 2975 - 22064	Lefebvre-Brossard et al. (2023)	MIT
SyntecReranking	19.2 - 402.2 - 467.2	100 - 100 - 917	*Introduced by our paper*	Apache-2.0
AlloprofRetrieval	48.31 - 1117.91	2316 - 2556	Lefebvre-Brossard et al. (2023)	MIT
BSARDRetrieval	144.03 - 24530.8	222 - 22600	Louis and Spanakis (2022)	CC-BY-NC-SA-4.0
SyntecRetrieval	19.22 - 295.65	100 - 90	*Introduced by our paper*	Apache-2.0
SummEvalFr	657.08 - 71.18 - 107.56	100 - 1100 - 1600	Created from Fabbri et al. (2021)	MIT

Table 3: Details of the data used for each task. The average number of tokens of texts is computed using the *cl100k_base* tokenizer. For Reranking, the three numbers refer to the queries, the pairs of queries with relevant documents and the pairs of queries with irrelevant ones, respectively. The pairs of queries and documents are obtained from the 90 dataset's documents. For Retrieval datasets, the two numbers refer to the queries and the documents, respectively. For *SummEvalFr*, the three numbers refer to the texts, human and machine summaries. References to all the datasets used are available.

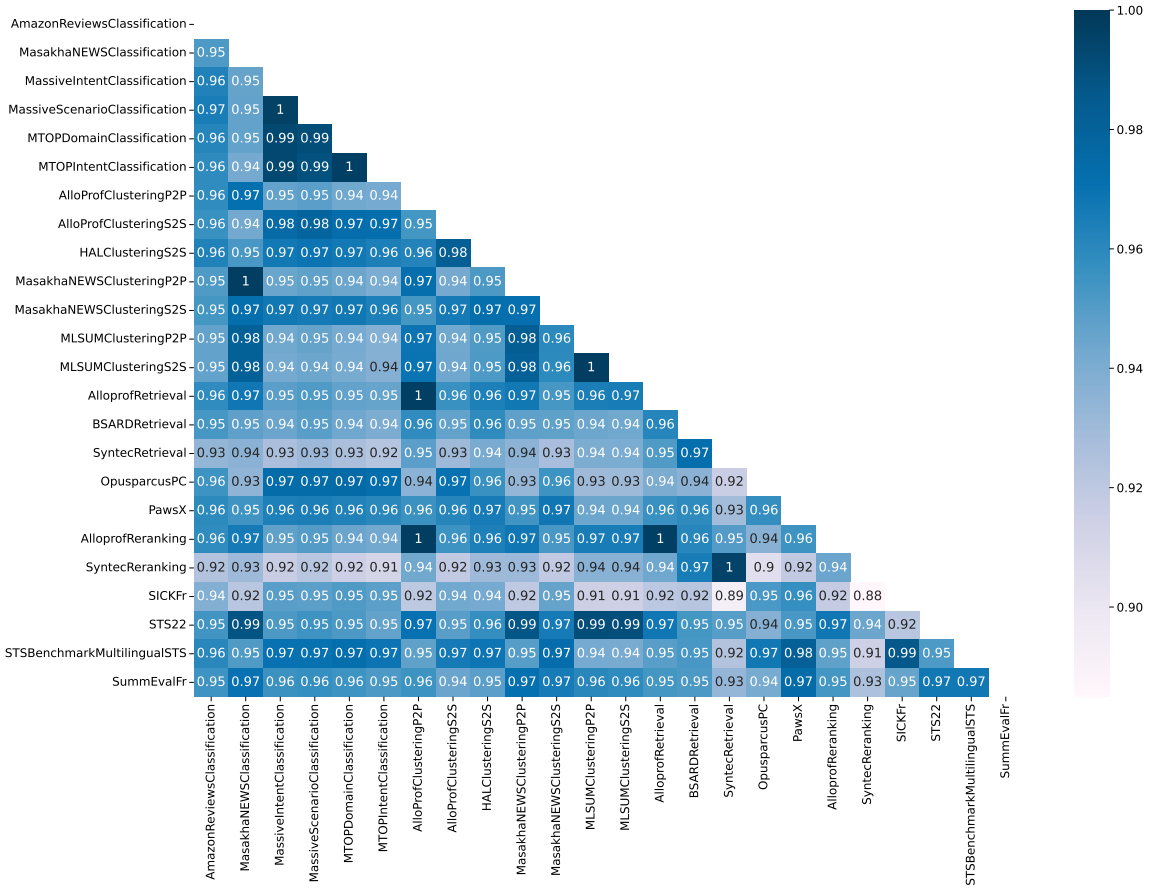

Figure 3: Cosine similarity between tasks' data. Ninety random samples per task's data are embedded using the *multilingual-e5-small* model. The embeddings of each task's data sample are averaged. The similarity between each dataset is then calculated using cosine similarity as in (Muennighoff et al., 2022).

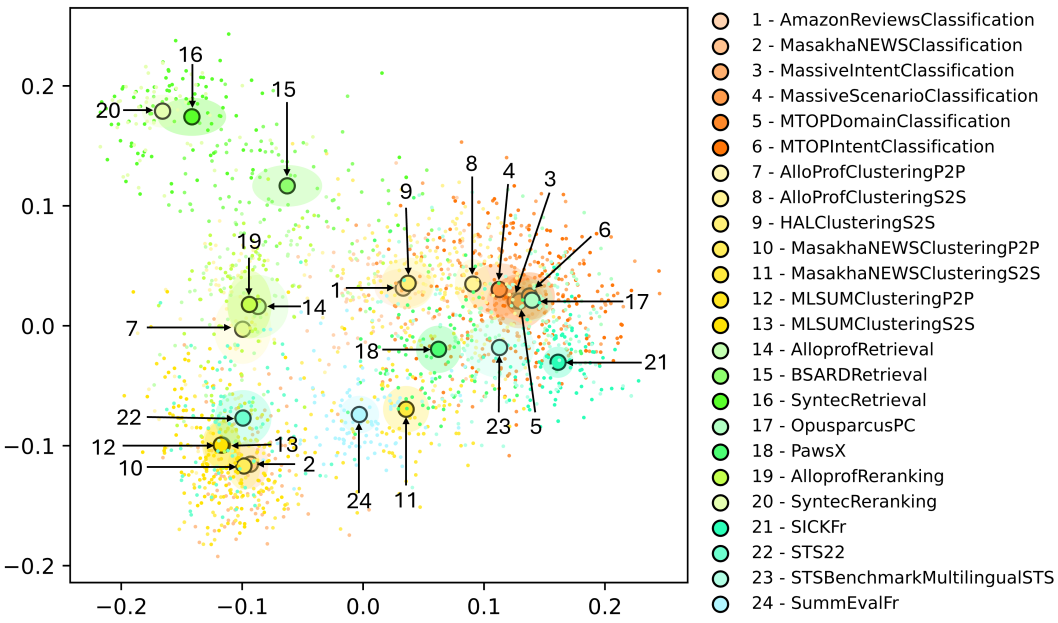

Figure 4: 2D projection of tasks' data. 90 random samples per task's data are embedded using *multlingual-e5-small* model (Wang et al., 2022). The embeddings are reduced to 2 dimensions using PCA. The centroid of each task's data is represented, along with the ellipse showing the standard deviation along each axis.

Label	# *raw*	# *mteb_eval*	Description
shs	58706	6701	Human and social sciences (*Sciences humaines et sociales*)
sdv	11049	4803	Life science [Biology] (*Sciences du vivant [Biologie]*)
spi	3601	3451	Engineering science (*Sciences de l'ingénieur [Physics]*)
info	3446	3263	Computer Science (*Informatique*)
sde	2830	2754	Environment science (*Sciences de l'environnement*)
phys	2003	1926	Physics (*Physique*)
sdu	1177	1158	Planet and Universe [Physics] (*Planète et Univers [Physique]*)
math	862	824	Mathematics (*Mathématiques*)
chim	764	734	Chemistry (*Chimie*)
scco	652	619	Cognitive sciences (*Sciences cognitives*)
qfin	183	N/A	Economy and quantitative finance (*Économie et finance quantitative*
stat	52	N/A	Statistics (*Statistiques*)
other	18	N/A	Other (*Autre*)
stic	14	N/A	N/A
nlin	12	N/A	Non-linear Science [Physics] (*Science non linéaire [Physique]*)
electromag	3	N/A	Electro-magnetism (*Electromagnétisme*)
instrum	2	N/A	Instrumentation [Physics] (*Instrumentation [Physique]*)
image	1	N/A	Image

Table 4: Distribution of classes in HAL the *raw* and *mteb_eval* subsets of the dataset.

Task type	Model	Score
Classification (F1-score)	TF-IDF + LR	0.60 ($\pm$ 0.002)
	TF-IDF + SVC	0.61 ($\pm$ 0.001)
	CamemBERT (fine-tuned)*	0.6 ($\pm$ 0.008)
	GPT-4 (ICL)**	0.30
Topic Modeling	TF-IDF + LDA	0.49 (Coherence)
		-8.23 (Perplexity)

Table 5: Baselines results for HAL on a classification task and topic modeling.
* CamemBERT was finetuned for 5 epochs with learning rate of $1e^{-4}$ (+ lr scheduler) and a batch size of 64.
** Due to limited budget, we evaluate *GPT-4* ICL capabilities on a limited subset of our dataset (600 first samples from the test set that is generated using the same seed as for other experiments).

- 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

Document

id	article-14
url	https://www.syntec.fr/convention-collective/resiliation-du-contrat-de-travail/#article-14
title	Article 14 : Préavis pendant la période d'essai
section	Résiliation du contrat de travail
content	Modification Avenant n° 7 du 5/07/1991 Au cours de cette période, les deux parties peuvent se séparer avec un préavis d'une journée de travail pendant le premier mois. Après le premier mois, le temps de préavis réciproque sera d'une semaine par mois complet passé dans l'entreprise. Après le premier mois, le temps de préavis réciproque sera d'une semaine par mois passé dans l'entreprise. Le préavis donne droit au salarié de s'absenter pour la recherche d'un emploi dans les conditions fixées à l'article 16. Le salarié sera payé au prorata du temps passé pendant la période d'essai.

Query

article	article-14
question	Quel est le préavis en période d'essai ?

Figure 5: Extracts of Syntec dataset.

hal_id	Domain	Title
hal-02899209	shs	La transformation digitale du management des ressources humaines et de ses enjeux pour les entreprises
tel-03993881	math	Sur l'approximation numérique de quelques problèmes en mécanique des fluides

Figure 6: Extracts of HAL dataset.

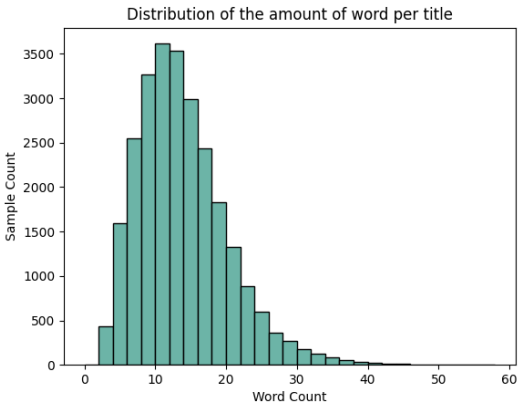

Figure 7: Distribution of the word count per title in HAL dataset, *mteb_eval* subset.

```
"""
You will be given a couple of texts in
    English and their translation in French.

Your task is to provide a 'rating' score on
    how well the system translated the
    English text into French.

Give your answer as a float on a scale of 0
    to 10, where 0 means that the
    system_translation is bad and does not
    represent what is being said in the
    original English text, and 10 means that
     the translation is good and represents
    the original English text.

No need to mind the quality of the text as
    original English text may be of bad
    quality.

Provide your feedback as follows:

Feedback:::
Total rating: (your rating, as a float
between 0 and 10)

Now here are the English and French texts.

Original text in English: {english_text}
Translation in French: {french_translation}

Feedback:::
Total rating:
"""
```

Figure 8: Prompt used for LLM as-judge evaluation of SummEval dataset translation.

Summary type	Original (SummEval)	Translated (SummEvalFr)
Human summary	*The whale, Varvara, swam a round trip from Russia to Mexico, nearly 14,000 miles. The previous record was set by a humpback whale that migrated more than 10,000 miles.*	*La baleine, Varvara, a parcouru à la nage un trajet aller-retour entre la Russie et le Mexique, soit près de 14 000 milles. Le précédent record avait été établi par une baleine à bosse qui avait migré sur plus de 10 000 miles.*
Machine summary	*north pacific gray whale has earned a spot in the record for the longest migration of a mammal ever recorded . the whale , named varvara , swam nearly 14,000 miles from the guinness worlds records . the record was set by a whale whale whale that swam a mere 10,190-mile round trip . the north coast of mexico is russian for "barbara".*	*la baleine grise du pacifique nord a obtenu une place dans le record de la plus longue migration d'un mammifère jamais enregistrée. la baleine, nommée varvara, a nagé près de 14 000 miles depuis les records du monde guinness. le record a été établi par une baleine baleine qui a nagé un voyage aller-retour de seulement 10 190 miles. la côte nord du mexique est le nom russe pour "barbara".*

Figure 9: Extracts of SummEvalFr dataset.

Quality	Rating	# samples
Good quality	10.0	186
	9.5	661
	9.0	193
Not good enough	8.5	16
	8.0	5
	7.5	7
	7.0	3
	6.0	3
	5.0	2
	4.0	1
	3.0	1
	2.0	3
	N/A	19

Table 6: Ratings provided by the LLM judge for the quality of human summaries translations of SummEvalFr from English to French.

- Summarization: Spearman correlation based on cosine similarity

D.1 Average performance per task type

Table 9 presents the average performance of each model on each task type.

D.2 Evaluation results per task

Tables 10, 11 12 and 13 present the models' performance on each task type. Table 10 presents the performance on classification and pair classification tasks. Table 11 presents the reranking and retrieval performance. Table 12 presents the performance on bitext mining, semantic textual similarity and summarization. Table 13 presents the performance on the clustering tasks.

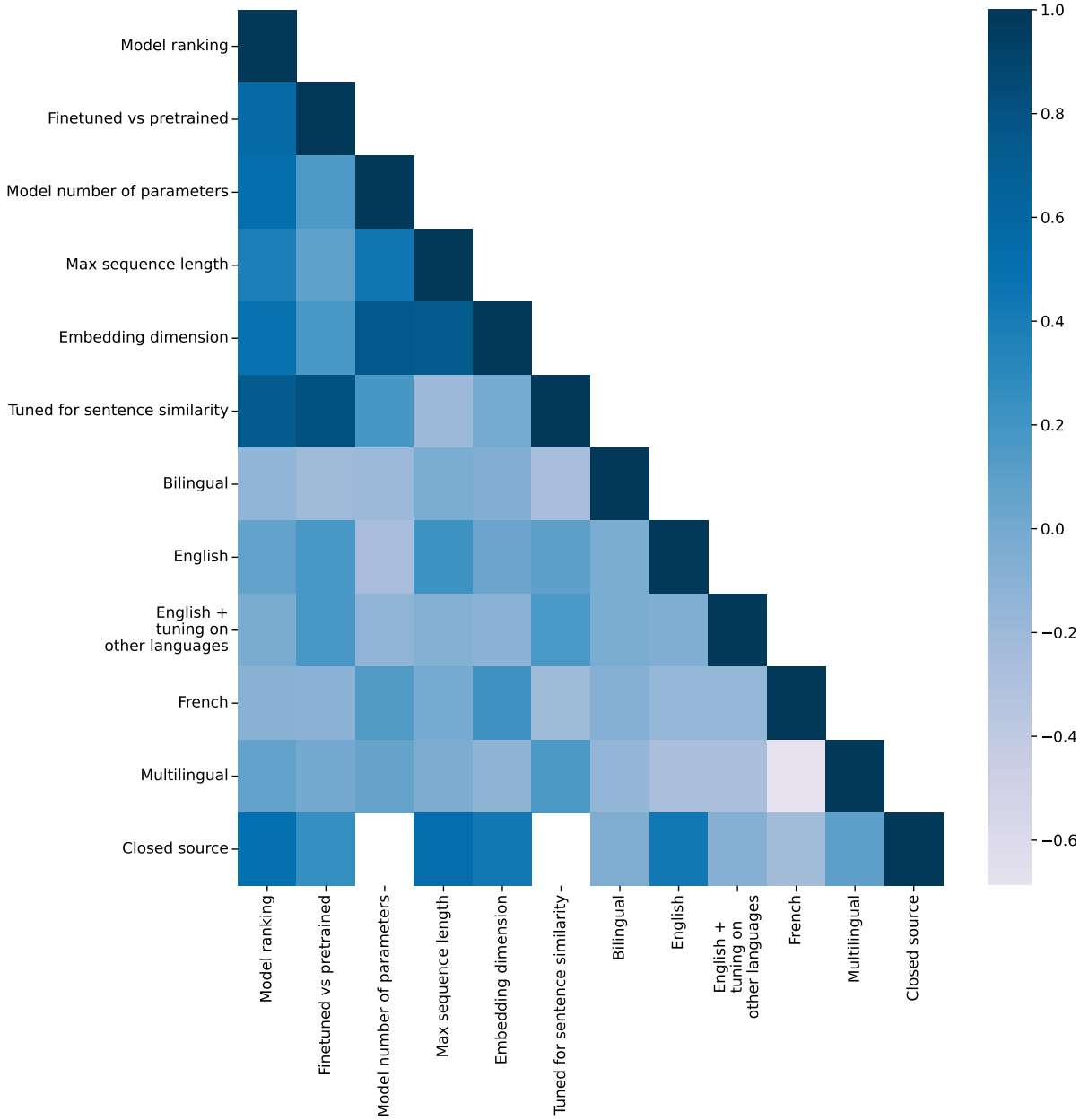

Figure 10: Heatmap representing cross-correlations between models' characteristics and models' performances.

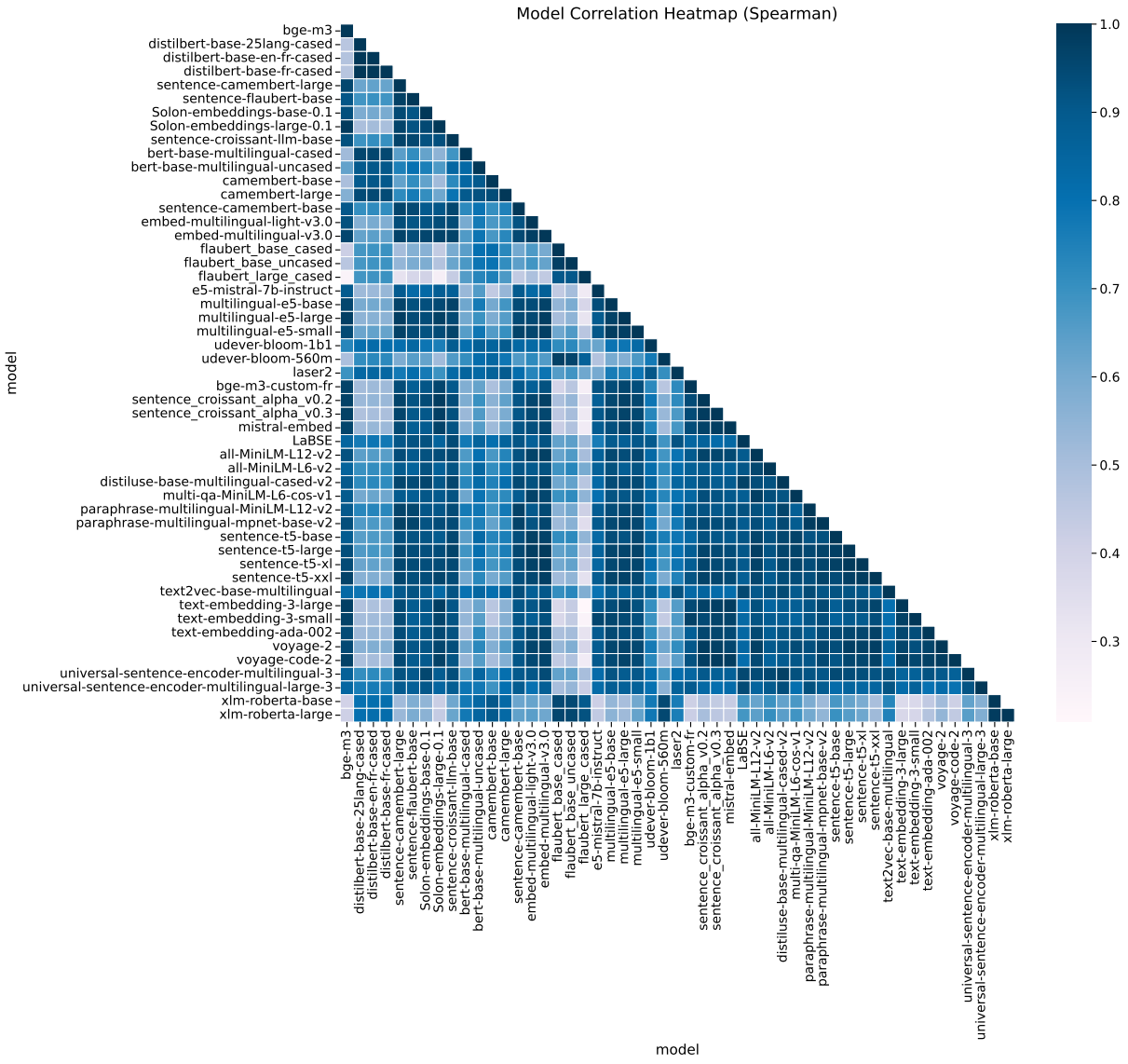

Figure 11: Heatmap representing the Spearman correlations in terms of performance across models.

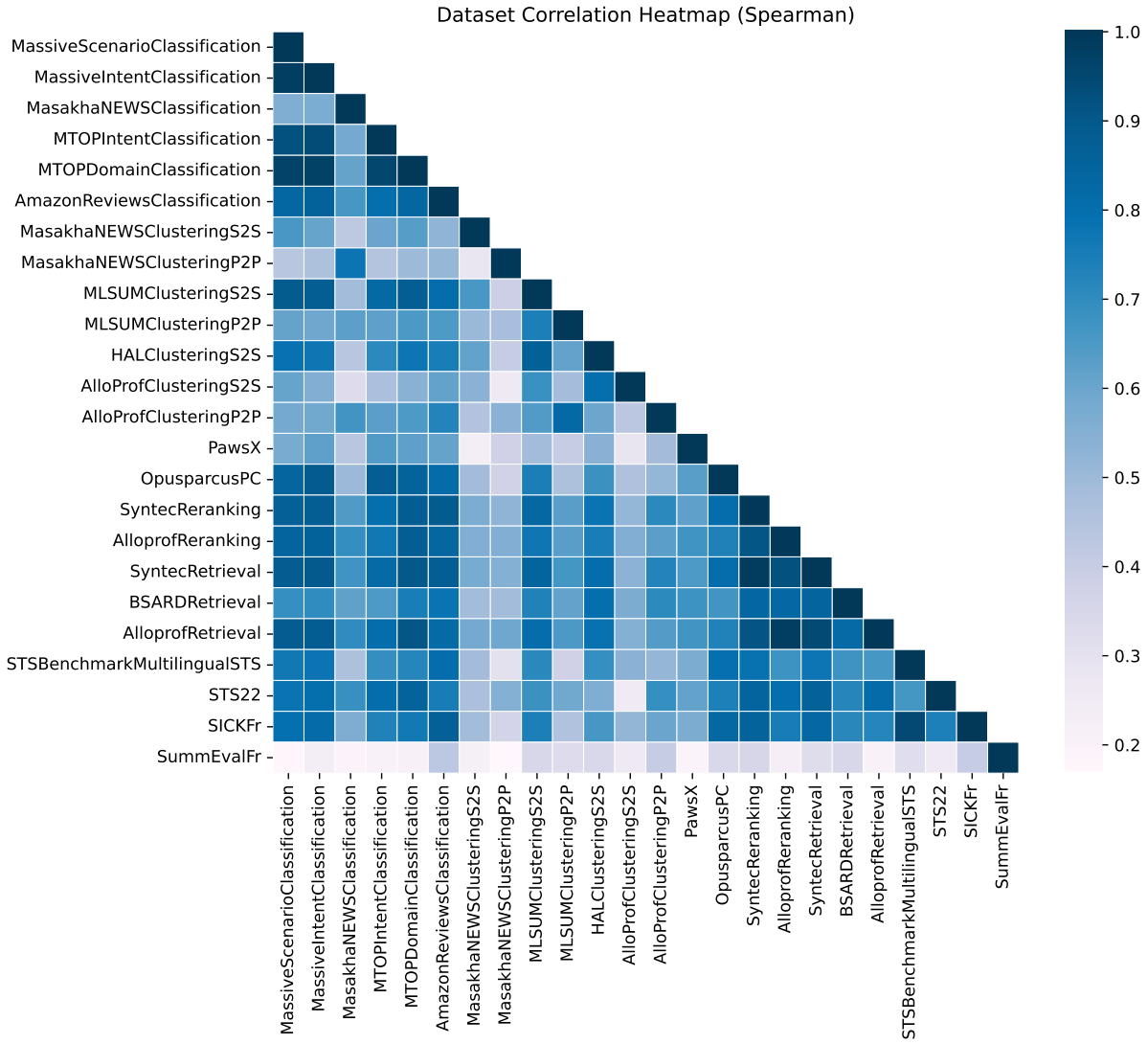

Figure 12: Heatmap representing the correlation regarding model performance across tasks.

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
sentence-flaubert-base	Yes	french	1,37e+08	0.55	512	768	Apache-2.0	Yes
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
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-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
text-embedding-3-small	N/A	multilingual	N/A	N/A	8191	1536	Closed source	N/A
text-embedding-3-large	N/A	multilingual	N/A	N/A	8191	3072	Closed source	N/A
mistral-embed	N/A	multilingual	N/A	N/A	16384	1024	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
Solon-embeddings-large-0.1	Yes	french	5.60e+08	2.239561728	512.0	1024.0	MIT	Yes
Solon-embeddings-base-0.1	Yes	french	2.78e+08	1.112174592	512.0	768.0	MIT	Yes
sentence-croissant-alpha-v0.3	Yes	french	1.28e+09	5.11954944	1024.0	2048.0	MIT	Yes
sentence-croissant-alpha-v0.2	Yes	french	1.28e+09	5.11954944	1024.0	2048.0	MIT	Yes
bge-m3	Yes	multilingual	5.68e+08	2.271019008	8192.0	1024.0	MIT	Yes
bge-m3-custom-fr	Yes	multilingual	5.68e+08	2.271019008	8192.0	1024.0	MIT	Yes

Table 7: 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.

Task type	Prompt
Classification	"Classify the following task: "
Clustering	"Identify the topic or theme based on the text: "
Retrieval	"Retrieve semantically similar text: "
Reranking	"Re-rank the following text: "
Pair Classification	"Classify the following pair of text: "
STS	"Determine the similarity between the following text: "
Summarization	"Summarize the following text: "
Bitext Mining	"Translate the following text: "

Table 8: Prompts used for the evaluation of *e5-mistral-7b-instruct*.

	Average	BitextMining	Classification	Clustering	PairClassification	Reranking	Retrieval	STS	Summarization
bge-m3	0.68	0.95	0.69	0.43	0.77	0.81	0.65	0.81	0.31
distilbert-base-25lang-cased	0.43	0.65	0.46	0.37	0.69	0.34	0.10	0.53	0.31
distilbert-base-en-fr-cased	0.43	0.65	0.46	0.38	0.69	0.34	0.10	0.54	0.31
distilbert-base-fr-cased	0.41	0.45	0.46	0.38	0.69	0.34	0.10	0.54	0.31
sentence-camembert-large	0.65	0.90	0.66	0.43	0.77	0.72	0.56	0.82	0.31
sentence-flaubert-base	0.59	0.80	0.61	0.41	0.76	0.65	0.43	0.79	0.31
Solon-embeddings-base-0.1	0.64	0.95	0.67	0.43	0.76	0.78	0.41	0.78	0.31
Solon-embeddings-large-0.1	0.67	0.96	0.69	0.42	0.77	0.79	0.63	0.80	0.30
sentence-croissant-llm-base	0.62	0.91	0.65	0.43	0.77	0.68	0.52	0.76	0.29
bert-base-multilingual-cased	0.44	0.75	0.46	0.34	0.70	0.38	0.10	0.50	0.29
bert-base-multilingual-uncased	0.49	0.76	0.48	0.41	0.70	0.46	0.19	0.56	0.31
camembert-base	0.35	0.18	0.42	0.34	0.68	0.31	0.02	0.57	0.30
camembert-large	0.37	0.26	0.49	0.36	0.65	0.34	0.07	0.59	0.17
sentence-camembert-base	0.57	0.72	0.57	0.36	0.74	0.66	0.43	**0.78**	0.29
embed-multilingual-light-v3.0	0.63	0.89	0.61	0.39	0.74	0.76	0.55	0.78	0.31
embed-multilingual-v3.0	0.66	0.94	0.67	0.41	0.77	0.79	0.54	0.81	0.31
flaubert _base _cased	0.34	0.23	0.25	0.27	0.67	0.36	0.08	0.52	0.31
flaubert _base _uncased	0.31	0.12	0.23	0.22	0.68	0.40	0.09	0.43	0.29
flaubert _large _cased	0.27	0.11	0.25	0.25	0.65	0.30	0.01	0.33	0.29
e5-mistral-7b-instruct	0.68	0.95	0.64	0.50	0.76	0.82	0.64	0.79	0.31
multilingual-e5-base	0.65	0.95	0.65	0.43	0.75	0.75	0.56	0.78	0.31
multilingual-e5-large	0.66	0.95	0.66	0.40	0.76	0.76	0.59	0.81	0.31
multilingual-e5-small	0.63	0.94	0.60	0.39	0.75	0.73	0.52	0.78	**0.32**
udever-bloom-1b1	0.47	0.52	0.55	0.35	0.74	0.43	0.28	0.62	0.29
udever-bloom-560m	0.36	0.32	0.30	0.29	0.71	0.39	0.11	0.51	0.24
laser2	0.52	0.95	0.58	0.30	**0.82**	0.44	0.13	0.67	0.31
bge-m3-custom-fr	0.66	0.94	0.67	0.40	0.77	0.79	0.59	0.80	0.30
sentence _croissant _alpha _v0.2	0.66	0.92	0.66	0.44	0.80	0.77	0.61	0.74	0.30
sentence _croissant _alpha _v0.3	0.67	0.92	0.66	0.46	0.79	0.78	0.65	0.77	0.31
mistral-embed	0.68	0.92	0.69	0.46	0.78	0.80	**0.68**	0.80	0.31
LaBSE	0.59	**0.96**	0.65	0.39	0.74	0.61	0.33	0.74	0.30
all-MiniLM-L12-v2	0.51	0.48	0.52	0.34	0.72	0.68	0.43	0.67	0.27
all-MiniLM-L6-v2	0.50	0.40	0.52	0.35	0.71	0.65	0.38	0.68	0.28
distiluse-base-multilingual-cased-v2	0.60	0.94	0.64	0.39	0.72	0.69	0.40	0.75	0.28
multi-qa-MiniLM-L6-cos-v1	0.49	0.38	0.51	0.33	0.72	0.64	0.39	0.67	0.28
paraphrase-multilingual-MiniLM-L12-v2	0.60	0.93	0.60	0.39	0.74	0.68	0.44	0.75	0.29
paraphrase-multilingual-mpnet-base-v2	0.63	0.94	0.63	0.40	0.76	0.74	0.50	0.78	0.30
sentence-t5-base	0.59	0.83	0.58	0.41	0.72	0.70	0.45	0.75	0.30
sentence-t5-large	0.62	0.90	0.62	0.42	0.76	0.73	0.51	0.75	0.30
sentence-t5-xl	0.65	0.91	0.65	0.43	0.78	0.76	0.55	0.77	0.32
sentence-t5-xxl	0.67	0.94	0.67	0.44	0.79	0.78	0.60	0.78	0.30
text2vec-base-multilingual	0.57	0.92	0.56	0.34	0.79	0.59	0.32	0.78	0.29
text-embedding-3-large	**0.71**	**0.96**	**0.74**	0.48	0.80	**0.86**	0.73	0.81	0.30
text-embedding-3-small	0.69	0.95	0.70	0.49	0.77	0.81	**0.68**	0.79	0.30
text-embedding-ada-002	0.69	0.95	0.69	**0.51**	0.77	0.82	0.67	0.78	0.30
voyage-code-2	0.67	0.86	0.67	0.47	0.77	0.81	**0.68**	0.78	0.28
universal-sentence-encoder-multilingual-3	0.60	0.94	0.64	0.43	0.72	0.68	0.35	0.75	0.28
universal-sentence-encoder-multilingual-large-3	0.59	0.95	0.66	0.37	0.74	0.67	0.33	0.74	0.28
xlm-roberta-base	0.36	0.48	0.31	0.28	0.68	0.30	0.01	0.51	0.29
xlm-roberta-large	0.35	0.35	0.31	0.29	0.69	0.35	0.03	0.49	0.29

Table 9: Average performance of models per task category.

	MassiveScenario	MassiveIntent	MasakhaNEWS	MTOPIntent	MTOPDomain	AmazonReviews	PawsX	OpusparcusPC
	Classification						PairClassification	
bge-m3	0.73	0.67	0.77	0.62	0.89	0.45	0.60	0.93
distilbert-base-25lang-cased	0.44	0.35	0.68	0.35	0.62	0.29	0.51	0.86
distilbert-base-en-fr-cased	0.44	0.35	0.68	0.35	0.62	0.29	0.51	0.86
distilbert-base-fr-cased	0.44	0.35	0.68	0.35	0.62	0.29	0.51	0.86
sentence-camembert-large	0.70	0.64	0.74	0.61	0.87	0.38	0.61	0.94
sentence-flaubert-base	0.63	0.59	0.71	0.53	0.79	0.40	0.58	0.93
Solon-embeddings-base-0.1	0.70	0.65	0.75	0.62	0.87	0.41	0.59	0.93
Solon-embeddings-large-0.1	0.71	0.67	0.76	0.69	0.89	0.42	0.60	0.94
sentence-croissant-llm-base	0.65	0.59	0.79	0.63	0.86	0.35	0.63	0.91
bert-base-multilingual-cased	0.44	0.37	0.64	0.38	0.64	0.29	0.53	0.87
bert-base-multilingual-uncased	0.44	0.38	0.76	0.39	0.64	0.29	0.53	0.87
camembert-base	0.39	0.31	0.66	0.29	0.58	0.30	0.52	0.83
sentence-camembert-base	0.61	0.52	0.70	0.43	0.77	0.36	0.57	0.92
sentence-camembert-large	0.69	0.63	0.81	0.59	0.86	0.38	0.60	0.95
embed-multilingual-light-v3.0	0.59	0.56	0.83	0.50	0.81	0.39	0.57	0.91
embed-multilingual-v3.0	0.67	0.63	0.83	0.61	0.86	0.42	0.61	0.94
flaubert_base_cased	0.11	0.07	0.71	0.09	0.26	0.25	0.52	0.82
flaubert_base_uncased	0.11	0.06	0.63	0.09	0.28	0.24	0.53	0.82
flaubert_large_cased	0.23	0.16	0.56	0.10	0.24	0.22	0.54	0.75
e5-mistral-7b-instruct	0.70	0.60	0.75	0.53	0.82	0.44	0.60	0.92
multilingual-e5-base	0.66	0.61	0.80	0.56	0.85	0.41	0.57	0.93
multilingual-e5-large	0.68	0.64	0.79	0.59	0.86	0.42	0.59	0.94
multilingual-e5-small	0.61	0.56	0.78	0.46	0.81	0.40	0.56	0.93
udever-bloom-1b1	0.50	0.43	0.81	0.51	0.69	0.35	0.62	0.86
udever-bloom-560m	0.22	0.15	0.68	0.16	0.35	0.27	0.60	0.82
laser2	0.59	0.53	0.66	0.57	0.76	0.34	0.70	0.94
bge-m3-custom-fr	0.75	0.67	0.70	0.61	0.90	0.42	0.61	0.93
sentence_croissant_alpha_v0.2	0.70	0.64	0.76	0.61	0.89	0.38	0.67	0.93
sentence_croissant_alpha_v0.3	0.70	0.65	0.76	0.59	0.88	0.36	0.65	0.93
mistral-embed	0.70	0.63	0.81	0.66	0.90	0.42	0.62	0.93
LaBSE	0.65	0.60	0.77	0.62	0.84	0.39	0.55	0.94
all-MiniLM-L12-v2	0.54	0.45	0.72	0.39	0.76	0.28	0.56	0.87
all-MiniLM-L6-v2	0.51	0.43	0.74	0.40	0.75	0.27	0.55	0.87
distiluse-base-multilingual-cased-v2	0.67	0.60	0.77	0.56	0.85	0.36	0.51	0.92
multi-qa-MiniLM-L6-cos-v1	0.50	0.43	0.76	0.37	0.73	0.27	0.57	0.88
paraphrase-multilingual-MiniLM-L12-v2	0.65	0.58	0.76	0.48	0.78	0.37	0.57	0.92
paraphrase-multilingual-mpnet-base-v2	0.68	0.62	0.78	0.52	0.80	0.40	0.58	0.93
sentence-t5-base	0.60	0.51	0.81	0.44	0.75	0.37	0.55	0.89
sentence-t5-large	0.64	0.57	0.80	0.48	0.80	0.41	0.60	0.91
sentence-t5-xl	0.66	0.61	0.80	0.54	0.85	0.44	0.63	0.92
sentence-t5-xxl	0.69	0.66	0.79	0.58	0.86	0.46	0.64	0.94
text2vec-base-multilingual	0.58	0.52	0.74	0.45	0.72	0.34	0.66	0.92
text-embedding-3-large	0.76	0.71	0.82	0.74	0.93	0.46	0.65	0.96
text-embedding-3-small	0.73	0.68	0.76	0.68	0.91	0.43	0.61	0.94
text-embedding-ada-002	0.71	0.65	0.82	0.64	0.89	0.44	0.60	0.94
voyage-code-2	0.70	0.63	0.82	0.59	0.88	0.42	0.61	0.93
universal-sentence-encoder-multilingual-3	0.70	0.61	0.82	0.54	0.85	0.34	0.52	0.91
universal-sentence-encoder-multilingual-large-3	0.73	0.66	0.72	0.64	0.88	0.35	0.54	0.93
xlm-roberta-base	0.23	0.14	0.60	0.19	0.44	0.27	0.51	0.85
xlm-roberta-large	0.24	0.16	0.66	0.15	0.37	0.27	0.53	0.84

Table 10: Performance of each model for Classification and Pair Classification.

	SyntecReranking	AlloprofReranking	SyntecRetrieval	BSARDRetrieval	AlloprofRetrieval
	Reranking		Retrieval		
bge-m3	0.88	0.74	0.85	0.60	0.49
distilbert-base-25lang-cased	0.39	0.29	0.18	0.11	0.01
distilbert-base-en-fr-cased	0.39	0.29	0.18	0.11	0.01
distilbert-base-fr-cased	0.39	0.29	0.18	0.11	0.01
sentence-camembert-large	0.82	0.63	0.79	0.56	0.33
sentence-flaubert-base	0.81	0.48	0.69	0.42	0.18
Solon-embeddings-base-0.1	0.85	0.71	0.81	0.00	0.41
Solon-embeddings-large-0.1	0.87	0.72	0.85	0.58	0.47
sentence-croissant-llm-base	0.78	0.57	0.74	0.52	0.30
bert-base-multilingual-cased	0.43	0.32	0.19	0.10	0.02
bert-base-multilingual-uncased	0.59	0.33	0.35	0.16	0.06
camembert-base	0.36	0.26	0.06	0.00	0.00
camembert-large	0.36	0.33	0.18	0.01	0.02
sentence-camembert-base	0.74	0.58	0.69	0.39	0.22
embed-multilingual-light-v3.0	0.82	0.70	0.77	0.52	0.35
embed-multilingual-v3.0	0.84	0.74	0.79	0.44	0.38
flaubert_base_cased	0.43	0.29	0.21	0.02	0.02
flaubert_base_uncased	0.49	0.30	0.22	0.03	0.02
flaubert_large_cased	0.32	0.29	0.02	0.00	0.01
e5-mistral-7b-instruct	0.90	0.74	0.83	0.64	0.45
multilingual-e5-base	0.83	0.67	0.80	0.53	0.36
multilingual-e5-large	0.83	0.69	0.81	0.59	0.38
multilingual-e5-small	0.82	0.65	0.76	0.52	0.27
udever-bloom-1b1	0.48	0.39	0.41	0.32	0.12
udever-bloom-560m	0.47	0.31	0.24	0.06	0.02
laser2	0.49	0.39	0.29	0.08	0.03
bge-m3-custom-fr	0.85	0.74	0.79	0.53	0.45
sentence_croissant_alpha_v0.2	0.82	0.72	0.79	0.60	0.45
sentence_croissant_alpha_v0.3	0.82	0.74	0.80	0.66	0.49
mistral-embed	0.81	0.78	0.79	0.68	0.57
LaBSE	0.68	0.55	0.55	0.23	0.20
all-MiniLM-L12-v2	0.69	0.67	0.61	0.34	0.33
all-MiniLM-L6-v2	0.67	0.63	0.60	0.27	0.28
distiluse-base-multilingual-cased-v2	0.75	0.62	0.65	0.29	0.27
multi-qa-MiniLM-L6-cos-v1	0.65	0.63	0.58	0.30	0.30
paraphrase-multilingual-MiniLM-L12-v2	0.73	0.62	0.66	0.38	0.27
paraphrase-multilingual-mpnet-base-v2	0.81	0.67	0.76	0.43	0.31
sentence-t5-base	0.76	0.63	0.67	0.40	0.28
sentence-t5-large	0.78	0.68	0.71	0.47	0.35
sentence-t5-xl	0.81	0.71	0.74	0.50	0.40
sentence-t5-xxl	0.82	0.75	0.79	0.56	0.46
text2vec-base-multilingual	0.63	0.56	0.50	0.26	0.19
text-embedding-3-large	0.92	0.80	0.87	0.73	0.60
text-embedding-3-small	0.89	0.74	0.87	0.66	0.52
text-embedding-ada-002	0.89	0.76	0.86	0.64	0.52
voyage-code-2	0.87	0.76	0.83	0.68	0.53
universal-sentence-encoder-multilingual-3	0.74	0.62	0.70	0.00	0.35
universal-sentence-encoder-multilingual-large-3	0.69	0.64	0.64	0.00	0.34
xlm-roberta-base	0.32	0.28	0.03	0.00	0.00
xlm-roberta-large	0.39	0.31	0.07	0.01	0.01

Table 11: Performance of each model for Retrieval and Reranking.

	Flores_fr-en	Flores_en-fr	DiaBla_fr-en	STSBenchmarkMultilingual	STS22	SICKFr	SummEvalFr
	BitextMining			STS			Summarization
bge-m3	1.00	1.00	0.85	0.82	0.82	0.78	0.31
distilbert-base-25lang-cased	0.92	0.91	0.11	0.57	0.41	0.62	0.31
distilbert-base-en-fr-cased	0.92	0.91	0.11	0.57	0.42	0.62	0.31
distilbert-base-fr-cased	0.63	0.65	0.06	0.57	0.43	0.62	0.31
sentence-camembert-large	0.99	1.00	0.70	0.86	0.82	0.78	0.31
sentence-flaubert-base	0.96	0.97	0.47	0.86	0.74	0.78	0.31
Solon-embeddings-base-0.1	1.00	1.00	0.85	0.79	0.81	0.75	0.31
Solon-embeddings-large-0.1	1.00	1.00	0.87	0.80	0.83	0.77	0.30
sentence-croissant-llm-base	1.00	1.00	0.74	0.79	0.79	0.70	0.29
bert-base-multilingual-cased	0.97	0.98	0.30	0.52	0.39	0.59	0.29
bert-base-multilingual-uncased	0.95	0.98	0.36	0.55	0.56	0.58	0.31
camembert-base	0.26	0.25	0.04	0.55	0.61	0.54	0.30
sentence-camembert-base	0.90	0.90	0.36	0.82	0.78	0.74	0.29
sentence-camembert-large	0.99	1.00	0.68	0.86	0.82	0.78	0.31
embed-multilingual-light-v3.0	1.00	1.00	0.66	0.76	0.83	0.76	0.31
embed-multilingual-v3.0	1.00	1.00	0.83	0.82	0.83	0.79	0.31
flaubert_base_cased	0.31	0.36	0.02	0.37	0.65	0.54	0.31
flaubert_base_uncased	0.25	0.08	0.03	0.33	0.55	0.42	0.29
flaubert_large_cased	0.15	0.17	0.01	0.16	0.49	0.35	0.29
e5-mistral-7b-instruct	1.00	1.00	0.85	0.83	0.76	0.79	0.31
multilingual-e5-base	1.00	1.00	0.85	0.81	0.78	0.76	0.31
multilingual-e5-large	1.00	1.00	0.85	0.83	0.80	0.79	0.31
multilingual-e5-small	1.00	1.00	0.82	0.79	0.80	0.76	0.32
udever-bloom-1b1	0.75	0.78	0.03	0.50	0.77	0.60	0.29
udever-bloom-560m	0.50	0.37	0.08	0.37	0.61	0.55	0.24
laser2	1.00	1.00	0.86	0.70	0.65	0.65	0.31
bge-m3-custom-fr	1.00	1.00	0.83	0.81	0.82	0.76	0.30
sentence_croissant_alpha_v0.2	1.00	1.00	0.75	0.73	0.79	0.69	0.30
sentence_croissant_alpha_v0.3	1.00	1.00	0.77	0.78	0.81	0.72	0.31
mistral-embed	1.00	1.00	0.75	0.80	0.83	0.76	0.31
LaBSE	1.00	1.00	0.88	0.75	0.78	0.70	0.30
all-MiniLM-L12-v2	0.71	0.62	0.10	0.67	0.70	0.63	0.27
all-MiniLM-L6-v2	0.62	0.56	0.03	0.65	0.77	0.62	0.28
distiluse-base-multilingual-cased-v2	1.00	1.00	0.83	0.77	0.76	0.72	0.28
multi-qa-MiniLM-L6-cos-v1	0.55	0.50	0.09	0.64	0.75	0.62	0.28
paraphrase-multilingual-MiniLM-L12-v2	1.00	1.00	0.78	0.80	0.71	0.75	0.29
paraphrase-multilingual-mpnet-base-v2	1.00	1.00	0.81	0.85	0.74	0.76	0.30
sentence-t5-base	0.97	0.96	0.55	0.74	0.78	0.72	0.30
sentence-t5-large	0.99	0.99	0.71	0.78	0.75	0.73	0.30
sentence-t5-xl	0.99	0.99	0.76	0.79	0.77	0.75	0.32
sentence-t5-xxl	1.00	1.00	0.83	0.81	0.77	0.77	0.30
text2vec-base-multilingual	0.99	0.99	0.78	0.83	0.74	0.77	0.29
text-embedding-3-large	1.00	1.00	0.88	0.83	0.82	0.79	0.30
text-embedding-3-small	1.00	1.00	0.86	0.81	0.81	0.76	0.30
text-embedding-ada-002	0.99	0.99	0.86	0.78	0.81	0.76	0.30
voyage-code-2	1.00	0.99	0.60	0.79	0.80	0.74	0.28
universal-sentence-encoder-multilingual-3	1.00	1.00	0.82	0.75	0.78	0.71	0.28
universal-sentence-encoder-multilingual-large-3	1.00	1.00	0.84	0.78	0.71	0.74	0.28
xlm-roberta-base	0.70	0.53	0.21	0.46	0.57	0.49	0.29
xlm-roberta-large	0.65	0.26	0.13	0.42	0.55	0.50	0.29

Table 12: Performance of each model for Bitext Mining, Semantic Textual Similarity (STS) and Summarization.

	MasakhaNEWSS2S	MasakhaNEWSP2P	MLSUMS2S	MLSUMP2P	HALS2S	AlloProfS2S	AlloProfP2P
	Clustering						
bge-m3	0.42	0.45	0.44	0.43	0.31	0.37	0.59
distilbert-base-25lang-cased	0.33	0.32	0.31	0.41	0.24	0.43	0.57
distilbert-base-en-fr-cased	0.34	0.34	0.31	0.41	0.25	0.42	0.57
distilbert-base-fr-cased	0.35	0.34	0.31	0.41	0.24	0.43	0.57
sentence-camembert-large	0.37	0.44	0.43	0.43	0.32	0.40	0.62
sentence-flaubert-base	0.30	0.49	0.41	0.41	0.32	0.40	0.57
Solon-embeddings-base-0.1	0.36	0.50	0.42	0.43	0.30	0.37	0.61
Solon-embeddings-large-0.1	0.31	0.46	0.43	0.43	0.32	0.37	0.63
sentence-croissant-llm-base	0.41	0.54	0.34	0.43	0.29	0.33	0.64
bert-base-multilingual-cased	0.24	0.24	0.32	0.41	0.25	0.43	0.51
bert-base-multilingual-uncased	0.42	0.50	0.31	0.43	0.26	0.35	0.61
camembert-base	0.27	0.44	0.27	0.41	0.16	0.29	0.54
camembert-large	0.33	0.42	0.35	0.44	0.03	0.34	0.59
sentence-camembert-base	0.31	0.36	0.27	0.36	0.25	0.39	0.59
embed-multilingual-light-v3.0	0.29	0.57	0.33	0.43	0.20	0.31	0.62
embed-multilingual-v3.0	0.32	0.53	0.35	0.45	0.24	0.36	0.64
flaubert_base_cased	0.21	0.42	0.17	0.39	0.04	0.14	0.53
flaubert_base_uncased	0.23	0.28	0.15	0.33	0.02	0.13	0.43
flaubert_large_cased	0.25	0.26	0.19	0.38	0.07	0.22	0.41
e5-mistral-7b-instruct	0.65	0.38	0.44	0.45	0.37	0.58	0.64
multilingual-e5-base	0.51	0.48	0.39	0.43	0.28	0.33	0.62
multilingual-e5-large	0.31	0.41	0.38	0.44	0.28	0.32	0.63
multilingual-e5-small	0.39	0.40	0.38	0.43	0.21	0.33	0.61
udever-bloom-1b1	0.27	0.40	0.30	0.44	0.16	0.27	0.62
udever-bloom-560m	0.21	0.38	0.25	0.36	0.08	0.22	0.54
laser2	0.30	0.32	0.27	0.35	0.12	0.26	0.48
bge-m3-custom-fr	0.42	0.29	0.42	0.42	0.31	0.39	0.58
sentence_croissant_alpha_v0.2	0.32	0.56	0.44	0.45	0.33	0.38	0.62
sentence_croissant_alpha_v0.3	0.38	0.58	0.44	0.44	0.35	0.41	0.60
mistral-embed	0.40	0.48	0.43	0.45	0.35	0.49	0.62
LaBSE	0.38	0.46	0.35	0.42	0.25	0.32	0.55
all-MiniLM-L12-v2	0.32	0.43	0.29	0.34	0.25	0.32	0.46
all-MiniLM-L6-v2	0.41	0.35	0.28	0.37	0.23	0.32	0.52
distiluse-base-multilingual-cased-v2	0.33	0.54	0.35	0.40	0.22	0.35	0.56
multi-qa-MiniLM-L6-cos-v1	0.27	0.54	0.26	0.35	0.14	0.26	0.49
paraphrase-multilingual-MiniLM-L12-v2	0.34	0.37	0.37	0.40	0.30	0.42	0.56
paraphrase-multilingual-mpnet-base-v2	0.31	0.42	0.38	0.41	0.31	0.45	0.54
sentence-t5-base	0.36	0.62	0.30	0.41	0.22	0.36	0.58
sentence-t5-large	0.31	0.59	0.32	0.42	0.25	0.40	0.62
sentence-t5-xl	0.32	0.63	0.34	0.42	0.27	0.41	0.60
sentence-t5-xxl	0.38	0.61	0.35	0.42	0.30	0.44	0.61
text2vec-base-multilingual	0.33	0.39	0.30	0.36	0.21	0.33	0.49
text-embedding-3-large	0.40	0.53	0.46	0.46	0.37	0.54	0.62
text-embedding-3-small	0.55	0.45	0.46	0.46	0.36	0.51	0.61
text-embedding-ada-002	0.49	0.68	0.42	0.45	0.35	0.54	0.65
voyage-code-2	0.35	0.57	0.41	0.45	0.35	0.51	0.62
universal-sentence-encoder-multilingual-3	0.40	0.61	0.36	0.44	0.24	0.38	0.57
universal-sentence-encoder-multilingual-large-3	0.40	0.24	0.38	0.41	0.23	0.38	0.54
xlm-roberta-base	0.24	0.29	0.24	0.40	0.09	0.20	0.52
xlm-roberta-large	0.22	0.34	0.19	0.43	0.06	0.21	0.57

Table 13: Performance of each model for Clustering.