# Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

Wissam Siblini*, Mathieu Ciancone**
Imene Kerboua***, Marion Schaeffer**

*Contributeur Individuel
wissam.siblini92@gmail.com,
**Wikit, France
{mathieu, marion}@wiki.ai
***INSA Lyon, LIRIS - Esker, France
imene.kerboua@liris.cnrs.fr

**Résumé.** Ce document présente les annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français" détaillant les données, les modèles et les résultats expérimentaux.

## Annexe 1 - Nouveaux jeux de données

### Syntec (Retrieval)

La convention collective de travail Syntec [1] comprend environ 90 articles. Ces derniers ont été extraits pour être utilisés comme documents dans un contexte de recherche d'information. Quatre annotateurs ont été divisés en deux groupes. Chaque groupe a reçu la moitié des articles et a répété le processus suivant : sélection d'un article au hasard et rédaction d'une question à son sujet. Chaque annotateur a rédigé 25 questions. Ainsi, cent questions ont été créées manuellement et associées aux articles contenant la réponse.

Un exemple de donnée est montré dans la Figure A. Le jeu de données Syntec peut également être utilisé pour la classification de textes, le clustering ou le topic modeling. En ce qui concerne la qualité, l'intégrité de chaque article a été vérifiée lors de la création manuelle des questions. Nous avons également vérifié manuellement que les questions ne pouvaient être répondues qu'à l'aide de l'article annoté comme pertinent.

### HAL (Clustering)

*Hyper Articles en Ligne* (HAL) est une archive ouverte française de documents scientifiques sur des domaines académiques variés. En analysant cette source, nous avons récupéré les détails de 85 000 publications en français. Nous avons extrait les identifiants, les titres et les domaines auto-annotés par les auteurs lors de la soumission sur HAL. Ce jeu de données peut être utilisé pour plusieurs tâches (e.g. topic modeling, la classification). Pour garantir que la qualité des données soit bonne, un nettoyage supplémentaire a été effectué :

1. https://www.syntec.fr/convention-collective/

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

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 ? |

FIG. A – *Exemple d'échantillon du jeu de données SyntecRetrieval.*

— Les doublons ont été éliminés.
— Les titres non pertinents (en raison d'erreurs d'indexation) ou les titres pas en français ont été supprimés après validation manuelle. Les échantillons avec des titres composés de deux mots ou moins ont aussi été supprimés (371 échantillons), amenant la moyenne à 13.4 (voir Figure C).
— Les échantillons appartenant aux classes avec moins de 500 exemples ont été supprimés, pour obtenir finalement un jeu de données avec 10 classes.
— Un sous-échantillonnage a été effectué sur les 2 classes contenant plus de 10 000 échantillons pour atténuer le déséquilibre (voir distribution des classes dans le Tableau A).

Deux extraits de données sont fournis dans la Figure B et des statistiques dans le Tableau A et la Figure C. La qualité des données est finalement évaluée en observant la capacité de 4 modèles de natures différentes à apprendre les tâches de classification et de topic modeling : *TF-IDF + SVM/LogReg*, le modèle *Camembert* (Martin et al., 2019), *GPT-4* avec une stratégie de type In-Context Learning (ICL), et LDA (Blei et al., 2003) pour le topic modeling. Les résultats sont indiqués dans le tableau B. L'ensemble de données est mis à disposition du public dans ses versions brute et nettoyée. Il est utilisé pour une tâche de clustering à partir du titre dans MTEB-FR (et le domaine sert de vérité terrain).

| **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 |

FIG. B – *Extraits du jeu de données HAL.*

| 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 (*Electro-magnétisme*) |
| instrum | 2 | N/A | Instrumentation [Physics] (*Instrumentation [Physique]*) |
| image | 1 | N/A | Image |

TAB. A – *Distribution des classes dans les données brutes HAL et dans le sous-ensemble final de MTEB-FR.*

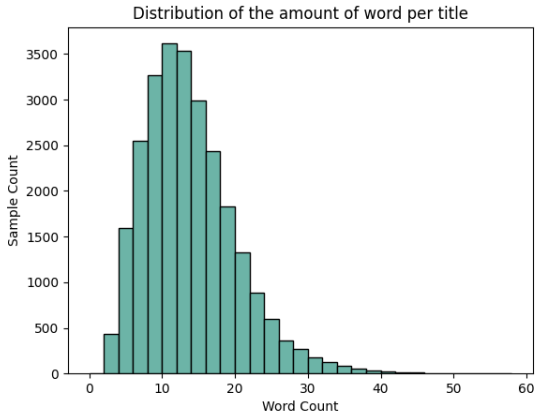


FIG. C – *Distribution du nombre de mots des titres des articles HAL dans le sous ensemble sélectionné pour MTEB-FR.*

| Tâche | Modèle | Score |
|---|---|---|
| **Classification (F1-score)** | TF-IDF + LogReg | $0.60\,(\pm 0.002)$ |
| | TF-IDF + SVM | $0.61\,(\pm 0.001)$ |
| | CamemBERT (fine-tuné)* | $0.6\,(\pm 0.008)$ |
| | GPT-4 (ICL)** | 0.30 |
| **Topic Modeling** | TF-IDF + LDA | 0.49 (Coherence) |
| | | -8.23 (Perplexity) |

*CamemBERT est specialisé avec 5 passes, un pas d'apprentissage de $1e^{-4}$ (décroissant linéairement jusqu'à zéro) et une taille de batch de 64.
**En raison d'un budget limité, nous évaluons *GPT-4* avec une stratégie de In-Context Learning sur un sous ensemble du jeu de données (600 échantillons).

TAB. B – *Résultats de modèles sur HAL en classification et topic modeling.*

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

## SummEvalFr

L'ensemble de données original SummEval (Fabbri et al., 2021) se compose de 100 articles d'actualités issus des médias CNN et DailyMail. Chaque article comporte 11 résumés rédigés par des humains et 16 résumés générés par des algorithmes. Ces derniers sont annotés par 8 personnes avec un score de cohérence, de consistance, de fluidité et de pertinence. Une méthode d'embeddings peut ici être évaluée via la corrélation entre la similarité résumé humain - résumé machine et les scores annotés.

| **Type de résumé** | **Original (SummEval)** | **Traduit (SummEvalFr)** |
|---|---|---|
| Résumé humain | *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 milles.* |
| Résumé machine | *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 milles 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 milles. la côte nord du mexique est le nom russe pour "barbara".* |

FIG. D – *Extraits de traductions pour SummEvalFr.*

```
"""
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:
"""
```

FIG. E – *Prompt utilisé pour l'évaluation LLM-as-a-judge vis-à-vis de la qualité des traductions de SummEval vers SummEvalFr.*

Nous avons traduit ce jeu de données vers le français à l'aide de l'API DeepL [2]. Des extraits de traductions sont montrés en Figure D. Nous proposons ici de calculer les métriques ROUGE (Lin, 2004) et BLEU (Papineni et al., 2002) entre les résumés générés par la machine et par l'homme pour les versions française et anglaise. Dans le tableau C, nous reportons la moyenne des scores ainsi que les corrélations entre les deux langues. La corrélation est élevée (supérieure à 0,7), ce qui montre que le chevauchement des n-grammes entre les résumés humains et les résumés machines pour l'anglais est préservé dans la version française.

| **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 |

TAB. C – *Scores moyens ROUGE et BLUE calculés entre les résumés machine et les résumés humains pour le jeu SummEval original et sa traduction en français. Les corrélations des scores entre l'anglais et le français sont également reportées.*

| **Qualité** | **Note** | **Nombre d'échantillons** |
|---|---|---|
| Bonne qualité | 10.0 | 186 |
| | 9.5 | 661 |
| | 9.0 | 193 |
| Qualité insuffisante | 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 |

TAB. D – *Notes fournies par le LLM-as-a-judge concernant la qualité des traductions de SummEval vers SummEvalFr. (N/A) signifie que nous n'avons pas pu obtenir les notes car le texte ne respectait pas la politique de contenu d'OpenAI.*

En outre, nous nous assurons que les résumés humains français sont correctement traduits depuis l'anglais en utilisant la méthode de LLM-as-a-judge (Zheng et al., 2023) où, étant donné le résumé humain original en anglais et sa traduction en français, le modèle évalue la qualité avec une note de $0$ à $10$. Le prompt du LLM est disponible dans la figure E. Nous regardons manuellement toutes les notes inférieures à $9$ et les corrigeons (voir la distribution des notes dans le tableau D). De plus, nous vérifions aléatoirement des traductions avec des notes comprises entre $9$ et $10$ pour nous assurer que la note est pertinente.

2. `https://www.deepl.com`

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

# Annexe 2 - Analyses et références des jeux de données de MTEB-FR

Le tableau E indique la taille de chaque ensemble de données, le nombre moyen de tokens des échantillons, et les références.

| Données et tâches | Nombre moyen de tokens | Nombre d'échantillons | Référence | Licence |
|---|---|---|---|---|
| AmazonReviewsClassification | 49.6 | 5000 | McAuley et 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 | 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 et 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 |

TAB. E – *Détails des données utilisées dans MTEB-FR. Le nombre moyen de tokens est calculé à l'aide du tokenizer* cl100k_base*. Pour la recherche d'information, les deux nombres font référence aux requêtes et aux documents. Pour le reranking, les trois nombres font référence aux requêtes, aux paires de requêtes avec des documents pertinents et aux paires de requêtes avec des documents non pertinents. Pour* SummEvalFr*, les trois nombres font référence aux textes, aux résumés humains et aux résumés machine.*

Nous montrons également la proximité entre les domaines des jeux de données en suivant la méthodologie de l'article original de MTEB (Muennighoff et al., 2022). Nous utilisons la méthode d'*embeddings* multlingual-e5-small et calculons les représentations d'un sous-échantillon de chaque jeu de données. Cela nous permet de construire une matrice de similarité (Figure G). Les distances entre les vecteurs moyens de chaque jeu de données (qui vont de $0,89$ à $1$ dans la Figure G) étant difficiles à interpréter, nous observons les nuages de points dans un plan grâce à l'Analyse en Composantes Principales (Figure F).

Les Figures F et G ont des résultats consistants. Il y a une forte similarité entre les ensembles qui ont les mêmes données sous-jacentes. On peut aussi voir quelques clusters de sujets. Syntec et BSARD, du domaine juridique, sont légèrement plus éloignés du reste. À l'avenir, à mesure que de nouveaux jeux de données seront ajoutés au benchmark, ce type d'analyse aidera à sélectionner ceux qui produisent des résultats moins redondants.

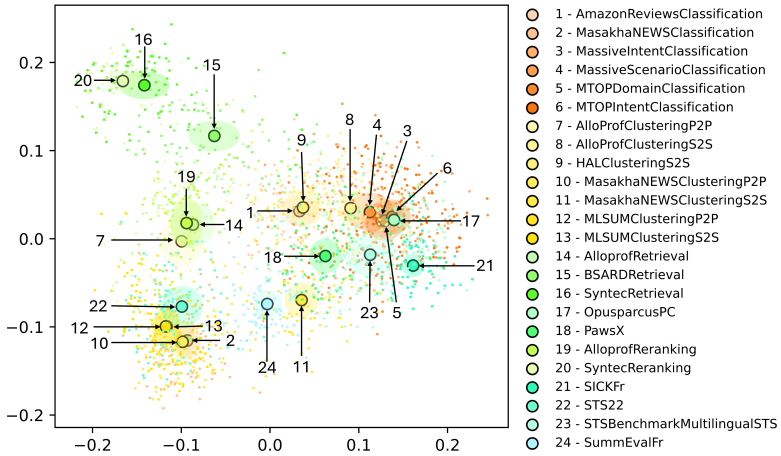


FIG. F – *Projection 2D (ACP) des données de MTEB-FR. Pour chaque jeu, 90 échantillons aléatoires sont vectorisés à l'aide du modèle* multlingual-e5-small *(Wang et al., 2022) et nous représentons les centroïdes ainsi que les nuages de points.*

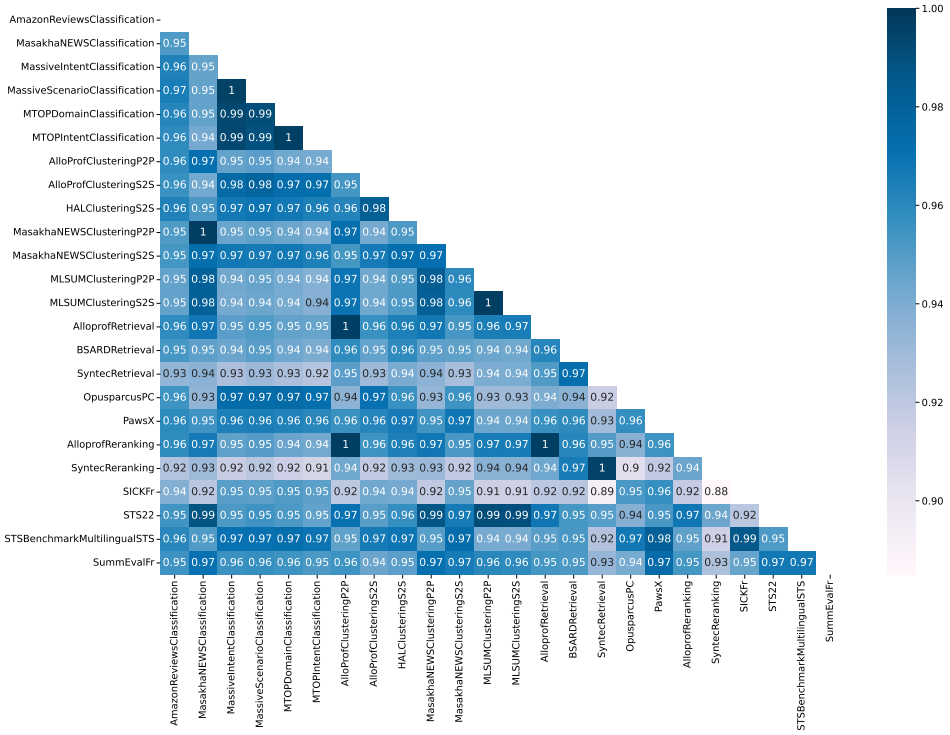


FIG. G – *Similarités du cosinus entre des embeddings moyens d'échantillons des différents jeux de données de MTEB-FR.*

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

# Annexe 3 - Matériel supplémentaire concernant l'analyse des corrélations

Cette section présente différentes corrélations calculées à partir des résultats des modèles sur le benchmark MTEB-FR. La Figure H montre, sous forme de heatmap, les corrélations croisées entre les performances des modèles et les caractéristiques étudiées. La Figure I représente les corrélations de Spearman entre les modèles en termes de performances sur les différents jeux de données.

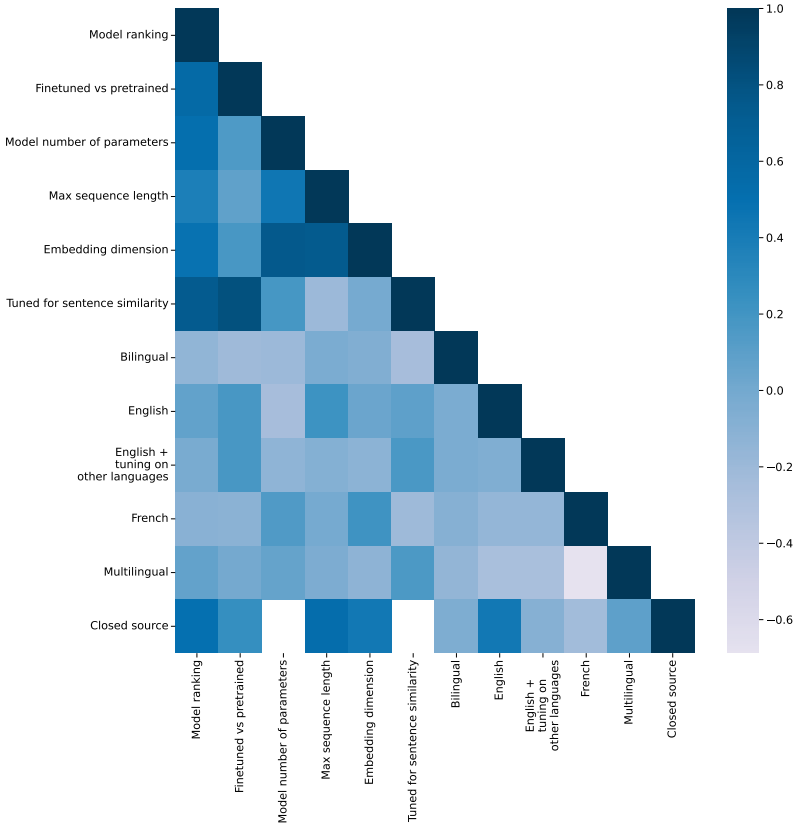


FIG. H – *Heatmap représentant les corrélations croisées entre caractéristiques des modèles et performances.*

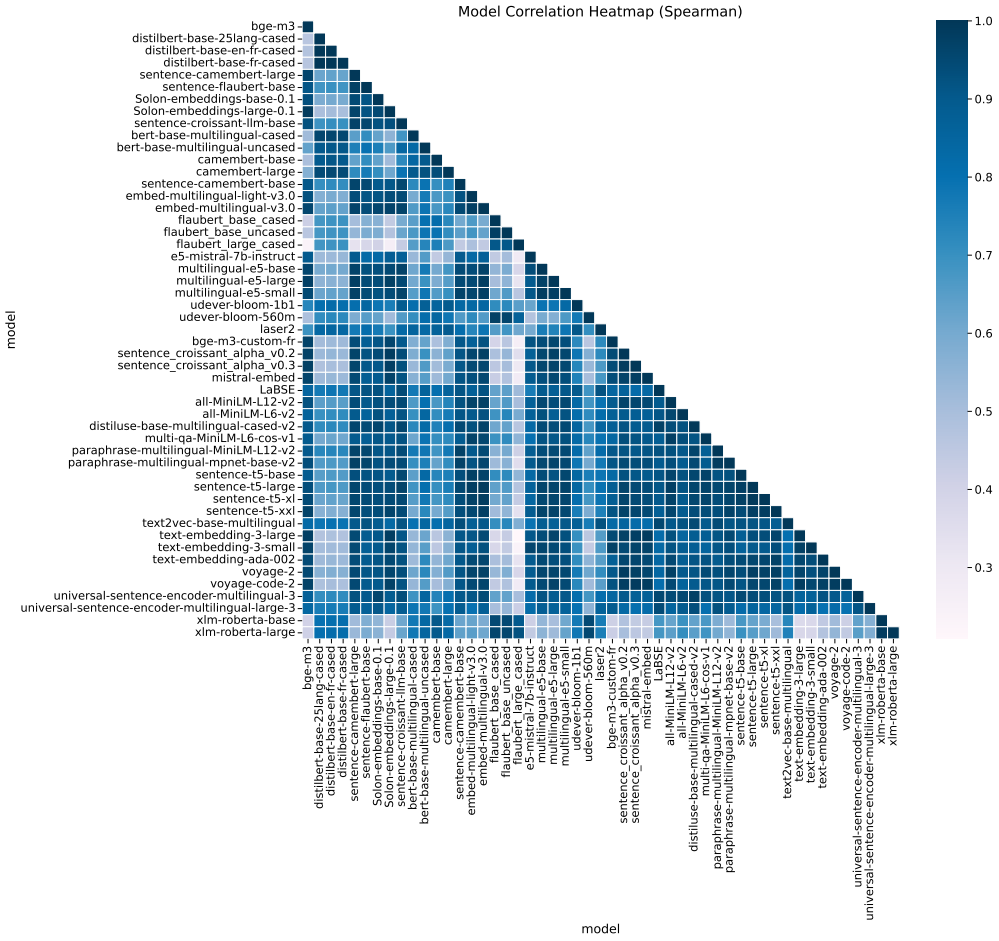


FIG. I – *Heatmap représentant les corrélations de Spearman entre modèles en termes de performances sur les différents jeux de données.*

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

# Annexe 4 - Informations supplémentaires concernant les modèles

Nous présentons dans cette section les caractéristiques des 51 modèles que nous avons selectionnés dans notre experience. Les détails sont donnés dans le Tableau F.

| Modèle | Spécialisé | Langue | # params | Taille (Gb) | Long. seq. | Dim. emb. | Licence | Sim phrases |
|---|---|---|---|---|---|---|---|---|
| bert-base-multilingue-cased | Non | multilingue | 1,78e+08 | 0.71 | 512 | 768 | Apache-2.0 | Non |
| bert-base-multilingue-uncased | Non | multilingue | 1,67e+08 | 0.67 | 512 | 768 | Apache-2.0 | Non |
| camembert-base | Non | français | 1,11e+08 | 0.44 | 514 | 768 | MIT | Non |
| camembert-large | Non | français | 3,37e+08 | 1.35 | 514 | 1024 | MIT | Non |
| sentence-camembert-base | Oui | français | 1,11e+08 | 0.44 | 128 | 768 | Apache-2.0 | Oui |
| sentence-camembert-large | Oui | français | 3,37e+08 | 1.35 | 514 | 1024 | Apache-2.0 | Oui |
| sentence-flaubert-base | Oui | français | 1,37e+08 | 0.55 | 512 | 768 | Apache-2.0 | Oui |
| embed-multilingue-light-v3.0 | N/A | multilingue | N/A | N/A | 512 | 384 | Closed source | N/A |
| embed-multilingue-v3.0 | N/A | multilingue | N/A | N/A | 512 | 1024 | Closed source | N/A |
| flaubert-base-cased | Non | français | 1,38e+08 | 0.55 | 512 | 768 | MIT | Non |
| flaubert-base-uncased | Non | français | 1,37e+08 | 0.55 | 512 | 768 | MIT | Non |
| flaubert-large-cased | Non | français | 3,73e+08 | 1.49 | 512 | 1024 | MIT | Non |
| distilbert-base-25lang-cased | Non | multilingue | 1,08e+08 | 0.43 | 512 | 768 | Apache-2.0 | Non |
| distilbert-base-en-fr-cased | Non | bilingue | 6,86e+07 | 0.27 | 512 | 768 | Apache-2.0 | Non |
| distilbert-base-fr-cased | Non | français | 6,17e+07 | 0.25 | 512 | 768 | Apache-2.0 | Non |
| multilingue-e5-base | Non | multilingue | 2,78e+08 | 1.11 | 512 | 768 | MIT | Oui |
| multilingue-e5-large | Non | multilingue | 5,60e+08 | 2.24 | 512 | 1024 | MIT | Oui |
| multilingue-e5-small | Non | multilingue | 1,18e+08 | 0.47 | 512 | 384 | MIT | Oui |
| e5-mistral-7b-instruct | Oui | anglais-plus | 7,11e+09 | 28.44 | 32768 | 4096 | MIT | Oui |
| udever-bloom-1b1 | Oui | multilingue | 1,07e+09 | 4.26 | 2048 | 1536 | bloom-rail-1.0 | Oui |
| udever-bloom-560m | Oui | multilingue | 5,59e+08 | 2.24 | 2048 | 1024 | bloom-rail-1.0 | Oui |
| laser2 | Oui | multilingue | 4,46e+07 | 0.18 | N/A | 1024 | BSD License | Oui |
| all-MiniLM-L12-v2 | Oui | anglais-plus | 3,34e+07 | 0.13 | 128 | 384 | Apache-2.0 | Oui |
| all-MiniLM-L6-v2 | Oui | anglais-plus | 2,27e+07 | 0.09 | 256 | 384 | Apache-2.0 | Oui |
| distiluse-base-multilingue-cased-v2 | Oui | multilingue | 1,35e+08 | 0.54 | 128 | 512 | Apache-2.0 | Oui |
| LaBSE | Oui | multilingue | 4,72e+08 | 1.89 | 256 | 768 | Apache-2.0 | Oui |
| multi-qa-MiniLM-L6-cos-v1 | Oui | anglais | 2,27e+07 | 0.09 | 512 | 384 | N/A | Oui |
| paraphrase-multilingue-MiniLM-L12-v2 | Oui | multilingue | 1,18e+08 | 0.47 | 128 | 384 | Apache-2.0 | Oui |
| sentence-t5-base | Oui | multilingue | 1,10e+08 | 0.44 | 256 | 768 | Apache-2.0 | Oui |
| sentence-t5-large | Oui | multilingue | 3,36e+08 | 1.34 | 256 | 768 | Apache-2.0 | Oui |
| sentence-t5-xl | Oui | multilingue | 1,24e+09 | 4.97 | 256 | 768 | Apache-2.0 | Oui |
| sentence-t5-xxl | Oui | multilingue | 4,87e+09 | 19.46 | 256 | 768 | Apache-2.0 | Oui |
| text2vec-base-multilingue | Oui | multilingue | 1,18e+08 | 0.47 | 256 | 384 | Apache-2.0 | Oui |
| text-embedding-ada-002 | N/A | multilingue | N/A | N/A | 8191 | 1536 | Closed source | N/A |
| text-embedding-3-small | N/A | multilingue | N/A | N/A | 8191 | 1536 | Closed source | N/A |
| text-embedding-3-large | N/A | multilingue | N/A | N/A | 8191 | 3072 | Closed source | N/A |
| mistral-embed | N/A | multilingue | N/A | N/A | 16384 | 1024 | Closed source | N/A |
| universal-sentence-encoder-multilingue-3 | Oui | multilingue | 6,89e+07 | 0.28 | N/A | 512 | Apache-2.0 | Oui |
| universal-sentence-encoder-multilingue-large-3 | Oui | multilingue | 8,52e+07 | 0.34 | N/A | 512 | Apache-2.0 | Oui |
| xlm-roberta-base | Non | multilingue | 2,78e+08 | 1.11 | 514 | 768 | MIT | Non |
| xlm-roberta-large | Non | multilingue | 5,60e+08 | 2.24 | 514 | 1024 | MIT | Non |
| sentence-croissant-llm-base | Oui | français | 1,28e+09 | 5.12 | 256 | 2048 | MIT | Oui |
| paraphrase-multilingue-mpnet-base-v2 | Non | multilingue | 2,78e+08 | 1.11 | 128 | 768 | Apache-2.0 | Oui |
| voyage-2 | N/A | anglais | N/A | N/A | 4000 | 1024 | Closed source | N/A |
| voyage-code-2 | N/A | anglais | N/A | N/A | 16000 | 1536 | Closed source | N/A |
| Solon-embeddings-large-0.1 | Oui | français | 5.60e+08 | 2.239561728 | 512.0 | 1024.0 | MIT | Oui |
| Solon-embeddings-base-0.1 | Oui | français | 2.78e+08 | 1.112174592 | 512.0 | 768.0 | MIT | Oui |
| sentence-croissant-alpha-v0.3 | Oui | français | 1.28e+09 | 5.11954944 | 1024.0 | 2048.0 | MIT | Oui |
| sentence-croissant-alpha-v0.2 | Oui | français | 1.28e+09 | 5.11954944 | 1024.0 | 2048.0 | MIT | Oui |
| bge-m3 | Oui | multilingue | 5.68e+08 | 2.271019008 | 8192.0 | 1024.0 | MIT | Oui |
| bge-m3-custom-fr | Oui | multilingue | 5.68e+08 | 2.271019008 | 8192.0 | 1024.0 | MIT | Oui |

TAB. F – *Modèles inclus dans le benchmark avec leurs principales caractéristiques. La taille en Go est estimée à l'aide du nombre de paramètres stockés en float32. La colonne "*Sim phrases*" fait référence au fait que le modèle ait été entraîné sur une tâche qui rend interprétable la similarité entre phrases. "anglais-plus" indique que le modèle est anglais mais qu'il a pu voir d'autres langues pendant sa spécialisation.*

Nous fournissons les prompts utilisés pour évaluer *intfloat/e5-mistral-instruct-7b* (modèle basé sur un LLM et du prompt-engineering) dans le Tableau G.

| Tâche | 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 : " |

TAB. G – *Structure des prompts utilisés pour obtenir des embeddings à partir du modèle* e5-mistral-7b-instruct*. Ces formats suivent les recommandations de la librairie SentenceTransformers.*

# Annexe 5 - Résultats détaillés de l'expérience

Cette section présente les résultats obtenus avec chaque modèle sur chaque tâche. Nous avons utilisé les mêmes métriques que dans MTEB, c'est-à-dire :

- Bitext Mining : Score F1
- Classification : Accuracy
- Clustering : Mesure V
- Pair Classification : Précision moyenne (Average Precision)
- Reranking : Mean average precision (MAP)
- Retrieval : Normalized Discounted Cumulative Gain at k (NDCG@k)
- STS et Summarization : Corrélation de Spearman entre (i) la similarité du cosinus entre deux embeddings et (ii) le score de vérité terrain pour les deux phrases correspondantes.

Le Tableau H présente les performances moyennes de chaque modèle sur chaque tâche. Les Tableaux I, J, K et L présentent les détails des performances des modèles sur chaque jeu de données. Le tableau I présente les performances sur les tâches de classification et de classification de paires. Le tableau J présente les performances de reranking et de recherche d'information. Le tableau K présente les performances de bibtex mining, de similarité textuelle sémantique et du résumé. Enfin, le tableau L présente les performances sur les tâches de clustering.

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

| | 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 |

TAB. H – *Performance moyenne des modèles par tâche.*

| | 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 |

TAB. I – *Performance de chaque modèle sur les jeux de données de classification de classification de paires.*

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

| | 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 |

TAB. J – *Performance de chaque modèle sur les jeux de données de recherche d'information et de 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 |

TAB. K – *Performance de chaque modèle sur les jeux de données de bitext mining, de similarité textuelle sémantique (STS) et de résumé.*

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"

| | 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 |

TAB. L – *Performance de chaque modèle sur les jeux de données de clustering.*

# Références

FitzGerald, J., C. Hench, C. Peris, S. Mackie, K. Rottmann, A. Sanchez, A. Nash, L. Urbach, V. Kakarala, R. Singh, S. Ranganath, L. Crist, M. Britan, W. Leeuwis, G. Tur, et P. Natarajan (2023). MASSIVE : A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In A. Rogers, J. Boyd-Graber, et N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Toronto, Canada, pp. 4277–4302. Association for Computational Linguistics, doi:10.18653/v1/2023.acl-long.235.

Adelani, D. I., M. Masiak, I. A. Azime, J. O. Alabi, A. L. Tonja, C. Mwase, O. Ogundepo, B. F. P. Dossou, A. Oladipo, D. Nixdorf, C. C. Emezue, S. Al-Azzawi, B. K. Sibanda, D. David, L. Ndolela, J. Mukiibi, T. O. Ajayi, T. M. Ngoli, B. Odhiambo, A. T. Owodunni, N. Obiefuna, S. H. Muhammad, S. S. Abdullahi, M. G. Yigezu, T. R. Gwadabe, I. Abdulmumin, M. T. Bame, O. O. Awoyomi, I. Shode, T. A. Adelani, H. A. Kailani, A.-H. Omotayo, A. Adeeko, A. Abeeb, A. Aremu, O. Samuel, C. Siro, W. Kimotho, O. R. Ogbu, C. E. Mbonu, C. I. Chukwuneke, S. Fanijo, J. Ojo, O. F. Awosan, T. K. Guge, S. T. Sari, P. Nyatsine, F. Sidume, O. Yousuf, M. Oduwole, U. Kimanuka, K. P. Tshinu, T. Diko, S. Nxakama, A. T. Johar, S. Gebre, M. A. Mohamed, S. A. Mohamed, F. M. Hassan, M. A. Mehamed, E. Ngabire, et P. Stenetorp (2023). Masakhanews : News topic classification for african languages. In *International Joint Conference on Natural Language Processing*.

Zheng, L., W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, et I. Stoica (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.

Lefebvre-Brossard, A., S. Gazaille, et M. C. Desmarais (2023). Alloprof : a new french question-answer education dataset and its use in an information retrieval case study. doi:10.48550/ARXIV.2302.07738.

Wang, L., N. Yang, X. Huang, B. Jiao, L. Yang, D. Jiang, R. Majumder, et F. Wei (2022). Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv :2212.03533*.

Louis, A. et G. Spanakis (2022). A statutory article retrieval dataset in French. In S. Muresan, P. Nakov, et A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, Dublin, Ireland, pp. 6789–6803. Association for Computational Linguistics, doi:10.18653/v1/2022.acl-long.468.

Chen, X., A. Zeynali, C. Camargo, F. Flöck, D. Gaffney, P. Grabowicz, S. Hale, D. Jurgens, et M. Samory (2022). SemEval-2022 task 8 : Multilingual news article similarity. In G. Emerson, N. Schluter, G. Stanovsky, R. Kumar, A. Palmer, N. Schneider, S. Singh, et S. Ratan (Eds.), *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, Seattle, United States, pp. 1094–1106. Association for Computational Linguistics, doi:10.18653/v1/2022.semeval-1.155.

Muennighoff, N., N. Tazi, L. Magne, et N. Reimers (2022). Mteb : Massive text embedding benchmark. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Fabbri, A. R., W. Kryściński, B. McCann, C. Xiong, R. Socher, et D. Radev (2021). Summeval : Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics 9*, 391–409.

Annexes de l'article "MTEB-FR: une expérience à large échelle pour l'apprentissage de représentation en français"


Li, H., A. Arora, S. Chen, A. Gupta, S. Gupta, et Y. Mehdad (2021). MTOP : A comprehensive multilingual task-oriented semantic parsing benchmark. In P. Merlo, J. Tiedemann, et R. Tsarfaty (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, Online, pp. 2950–2962. Association for Computational Linguistics, doi:10.18653/v1/2021.eacl-main.257.

May, P. (2021). Machine translated multilingual sts benchmark dataset.

Goyal, N., C. Gao, V. Chaudhary, P.-J. Chen, G. Wenzek, D. Ju, S. Krishnan, M. Ranzato, F. Guzmán, et A. Fan (2021). The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Bawden, R., E. Bilinski, T. Lavergne, et S. Rosset (2021). Diabla : A corpus of bilingual spontaneous written dialogues for machine translation. *Language Resources and Evaluation 55*, 635–660, doi:10.1007/s10579-020-09514-4.

Scialom, T., P.-A. Dray, S. Lamprier, B. Piwowarski, et J. Staiano (2020). MLSUM : The multilingual summarization corpus. In B. Webber, T. Cohn, Y. He, et Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online, pp. 8051–8067. Association for Computational Linguistics, doi:10.18653/v1/2020.emnlp-main.647.

Yang, Y., Y. Zhang, C. Tar, et J. Baldridge (2019). PAWS-X : A cross-lingual adversarial dataset for paraphrase identification. In K. Inui, J. Jiang, V. Ng, et X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, Hong Kong, China, pp. 3687–3692. Association for Computational Linguistics, doi:10.18653/v1/D19-1382.

Martin, L., B. Muller, P. O. Suarez, Y. Dupont, L. Romary, E. V. de la Clergerie, D. Seddah, et B. Sagot (2019). Camembert : a tasty french language model. In *Annual Meeting of the Association for Computational Linguistics*.

Creutz, M. (2018). Open subtitles paraphrase corpus for six languages. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, et T. Tokunaga (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

McAuley, J. et J. 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, New York, NY, USA, pp. 165–172. Association for Computing Machinery, doi:10.1145/2507157.2507163.

Lin, C.-Y. (2004). ROUGE : A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, pp. 74–81. Association for Computational Linguistics.

Blei, D. M., A. Y. Ng, et M. I. Jordan (2003). Latent dirichlet allocation. *Journal of machine Learning research 3*(Jan), 993–1022.

Papineni, K., S. Roukos, T. Ward, et W.-J. Zhu (2002). Bleu : a method for automatic evaluation of machine translation. In P. Isabelle, E. Charniak, et D. Lin (Eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Phi-

ladelphia, Pennsylvania, USA, pp. 311–318. Association for Computational Linguistics, doi:10.3115/1073083.1073135.

## Summary

Appendices of the article "MTEB-FR: a large-scale experiment for representation learning in French" detailing the data, the models and the experimental results.