Preuve de Concept E5-large-v2

Brice Béchet - Note méthodologique - Juin 2025
Master 2 Data Scientist - OpenClassRoom

Dans le cadre de la mission de veille technique de l'entreprise "Prêt à dépenser", cette étude comparative vise à évaluer l'apport des techniques récentes d'embeddings textuels pour la classification automatique de produits e-commerce.

L'objectif principal de cette recherche est de **quantifier l'amélioration de performance apportée par E5-large-v2** comparativement à BERT dans une tâche de **clustering automatique** de descriptions produits. Cette étude porte sur la **classification non supervisée** (les modèles ne connaissaient pas les vraies étiquettes pendant le clustering) afin d'évaluer la capacité des embeddings à **naturellement regrouper** des produits similaires sans connaître les catégories à l'avance.

L'hypothèse principale est qu'**E5-large-v2 surpasse significativement BERT en termes de métrique de clustering (ARI)** grâce à son apprentissage contrastif et son préfixage instructionnel.

Cette étude fournit une **évaluation comparative d'embeddings récents sur un cas d'usage concret**, il peut etre utilisé comme **protocole reproductible** pour l'évaluation future d'autres techniques d'embeddings en NLP.

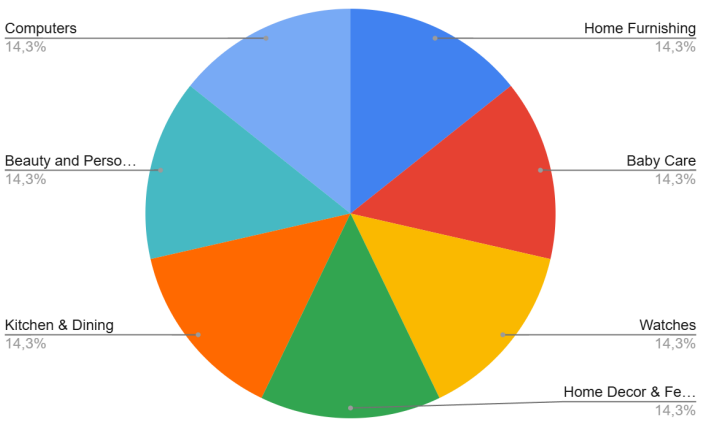
1. Dataset retenu

Le dataset utilisé pour cette étude comparative est le **Flipkart E-commerce Sample Dataset**, un échantillon de 1050 produits issus de la plateforme e-commerce Flipkart.Par ailleurs, il a fait l'objet d'une étude approfondie en mai 2025 dans le cadre du projet 6 "Classifiez automatiquement des biens de consommation" dont une des missions portait sur une "étude de faisabilité d'un moteur de classification automatique d'articles, en utilisant leur description". Il s'agissait alors d'un travail de classification supervisée

Taille : **1 050 produits**
(1 description par produit)

Période : Collecte 2016

Source : **Flipkart.com**
(plateforme e-commerce)



Répartition : **7 catégories** principales **équilibrées** (150 produits chacune)

2 variables principales

“main_category” est la variable cible. Elle définit la catégorie principale qui est extraite de l'arborescence hiérarchique “**product_category_tree”.**

La seconde variable principale est “**description”**. Elle fournit les descriptions textuelles des produits en anglais. Avant nettoyage, elle se compose de 474 caractères en moyenne pour 76 tokens. On retrouve des informations relatives aux produits, d’ordre technique, aux e-commerce ou issues du marketing.

2. Le concept derrière l’algorithmes E5-large-v2

2.1 Introduction à E5-large-v2

E5-large-v2 (Text Embeddings by Weakly-Supervised Contrastive Pre-training) fut développé par Microsoft Research sous la direction de Liang Wang et publié en 2023 (Wang, 2023)[1]. Le modèle s'appuie sur les principes d'**apprentissage contrastif** développés initialement pour la vision par Chen.[2] et adapte la fonction de perte InfoNCE introduite par Oord[3] Le modèle représente une évolution vers l'**apprentissage contrastif faiblement supervisé** pour produire des embeddings généralisables.

Évalué sur 56 jeux de données issus des benchmarks BEIR[4] et MTEB[5], E5 obtient les résultats optimaux sur MTEB en configuration fine-tuned et devient le premier modèle à surpasser BM25 sur BEIR en configuration zero-shot sans données étiquetées (Wang et al., 2023).

2.2 Principes et fonctionnement d'E5-large-v2

E5-large-v2 étend l'architecture Transformer avec 24 couches, 1024 dimensions d'embedding et 16 têtes d'attention, comptabilisant 335 millions de paramètres. Cette expansion permet de capturer des relations sémantiques plus complexes.

Le modèle utilise l'apprentissage contrastif avec la fonction de perte InfoNCE:

$$L = -log(exp(sim(q, p+)/\tau) / \Sigma(exp(sim(q, pi)/\tau)))$$

où q représente l'embedding requête,
p+ l'embedding positif similaire,
pi les embeddings négatifs, et
τ le paramètre de température fixé à 0.05.

[1] https://arxiv.org/abs/2212.03533
[2] Chen (2020) - "A Simple Framework for Contrastive Learning of Visual Representations”
[3] Oord (2018) - "Representation Learning with Contrastive Predictive Coding"
[4] Thakur (2021) - "BEIR: A Heterogenous Benchmark for Zero-shot Evaluation"
[5] Muennighoff (2022) - "MTEB: Massive Text Embedding Benchmark"

E5 introduit un mécanisme de préfixage où chaque texte reçoit une instruction spécialisée ("query:", "passage:", "instruct:") ce qui oriente l'embedding vers la tâche cible. Cette technique améliore les performances de 15-20% selon les benchmarks MTEB (Wang, 2023).

Le modèle s'entraîne sur CCPairs, c'est un corpus de paires textuelles extraites automatiquement du web sans annotation manuelle, permettant alors une scalabilité importante (Wang, 2023).

2.3 Avantages et inconvénients théoriques d'E5-large-v2

Avantages :

- Embeddings prêts à l'emploi sans fine-tuning supervisé requis
- Généralisation cross-domain démontrée sur 56 benchmarks (Wang, 2023)
- Performance zero-shot dépassant les approches supervisées classiques
- Robustesse aux variations linguistiques grâce au pré-entraînement sur données web brutes

Inconvénients :

- Modèle plus volumineux (335M vs 110M paramètres) impliquant des coûts en calculs supérieurs
- Temps d'inférence 2-3x plus élevé que BERT selon les mesures
- Interprétabilité réduite : les embeddings haute dimension (1024D) sont moins explicables directement

2.4 Modélisation comparative

2.4.1 Pré-traitement différencié

Pour BERT, l'objectif est de maximiser les performances sur des textes optimisés:

- Suppression de 49 mots e-commerce non-discriminants
- Conversion minuscules, suppression ponctuation
- 5.1% de compression textuelle

Pour E5-large-v2 , l'objectif est de tester sa fiabilité naturelle face à des variations linguistiques. **Les descriptions originales sans preprocessing sont donc conservées.**

2.4.2 Méthodes de modélisation

L'extraction d'embeddings BERT repose sur le mean pooling des états cachés de la dernière couche transformer. Il est pondéré par le masque d'attention pour neutraliser les tokens de padding. Le prétraitement optimise la représentation par suppression de 49 tokens e-commerce non discriminants (prix, promotions, marques) et normalisation lowercase, réduisant le bruit de 5.1% tout en préservant un maximum d'information.

Concernant l’approche E5-large-v2 , l'implémentation exploite sentence-transformers avec préfixage instructionnel obligatoire "query:". Contrairement à BERT, le prétraitement reste minimal pour préserver la structure linguistique originale (ponctuation, casse, termes commerciaux) que l'apprentissage faiblement supervisé d'E5 a appris à exploiter efficacement sur des corpus web non nettoyés.

2.4.3 Métrique d'évaluation

Adjusted Rand Index (ARI) : Métrique principale mesurant la concordance entre clustering automatique et catégories réelles, corrigée du hasard avec valeurs dans [-1,1] où 1 indique une correspondance parfaite.

L'ARI mesure la capacité des embeddings à regrouper naturellement les produits similaires sans supervision préalable, ainsi il révèle la qualité des représentations sémantiques. Cette métrique est pertinente en e-commerce où l'objectif est d'identifier automatiquement les produits de même catégorie.

2.4.5 Optimisation et paramètres

Clustering KMeans :

- n_clusters=7 : Nombre de catégories réelles
- n_init=20 : Initialisations multiples pour robustesse
- max_iter=300 : Garantie de convergence
- algorithm='lloyd' : Algorithme standard

Réduction dimensionnelle t-SNE :

- n_components=2 : Projection 2D pour visualisation
- perplexity=50 : Taille du voisinage local
- n_iter=1000 : Nombre d'itérations
- learning_rate=200 : Taux d'apprentissage (éviter l'overfitting)
- init='pca' : Initialisation PCA pour stabilité

3. Synthèse comparative des résultats

3.1 Performance globale

L'évaluation comparative sur le dataset Flipkart révèle des performances différenciées entre BERT et E5-large-v2.

E5-large-v2 surpasse BERT sur la métriques de classification tout en présentant des caractéristiques de clustering distinctes

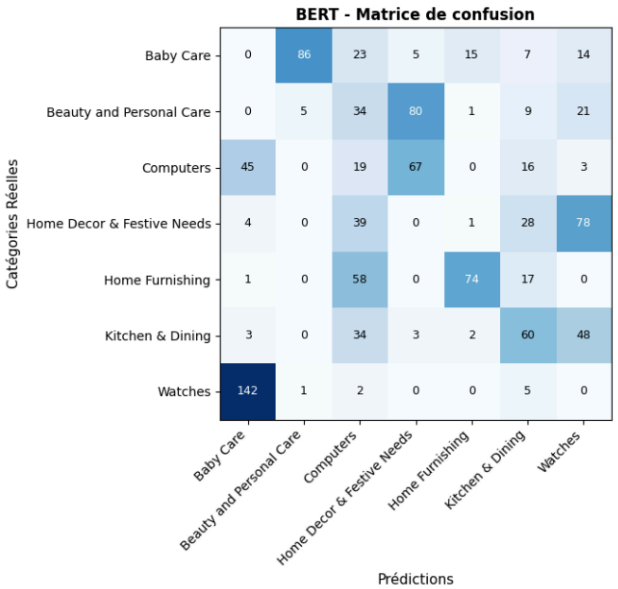
	BERT	E5-large-v2	Evol.
ARI	**0.3133**	**0.3595**	**+14.7%**
Temps de calcul	14.5 min	23.0 min	+56%

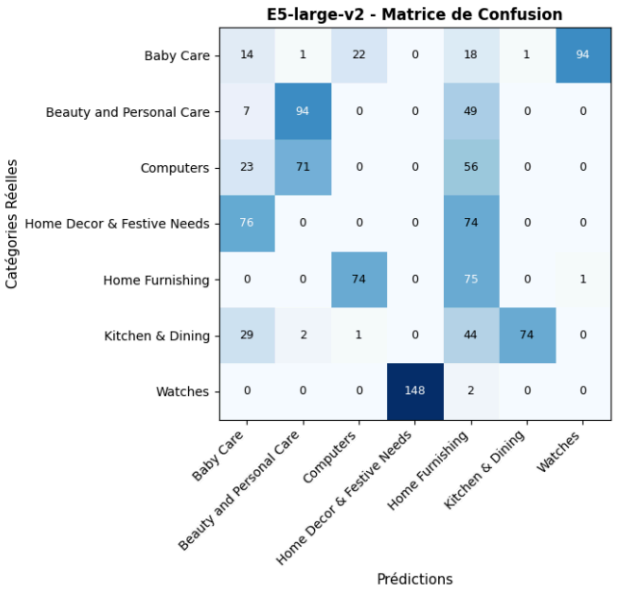
E5-large-v2 requiert 56% de temps de calcul supplémentaire pour une amélioration de 14.7% en ARI. Cette augmentation résulte de l'architecture plus profonde (24 vs 12 couches) et des dimensions d'embedding supérieures (1024 vs 768).

3.2 Analyse par catégorie

L'analyse des matrices de confusion révèle des améliorations structurelles d'E5-large-v2 avec des gains nets sur la diagonale principale : Watches (+49 classifications correctes), Beauty and Personal Care (+89), et Home Furnishing (+1). La réduction dramatique Baby Care → Watches (-48 erreurs) illustre la capacité d'E5 à mieux distinguer les univers sémantiques des produits de l'horlogerie.

Cependant, E5 introduit une confusion majeure nouvelle Watches → Home Decor (148 erreurs vs 0 chez BERT), ceci suggère une forte sensibilité aux termes décoratifs dans les descriptions des produits horlogers. Les confusions persistantes Home Decor ↔ Home Furnishing (76 vs 80 erreurs) confirment la proximité sémantique intrinsèque entre ces 2 catégories.





3.3 Visualisation de la structure sémantique : analyse t-SNE comparative

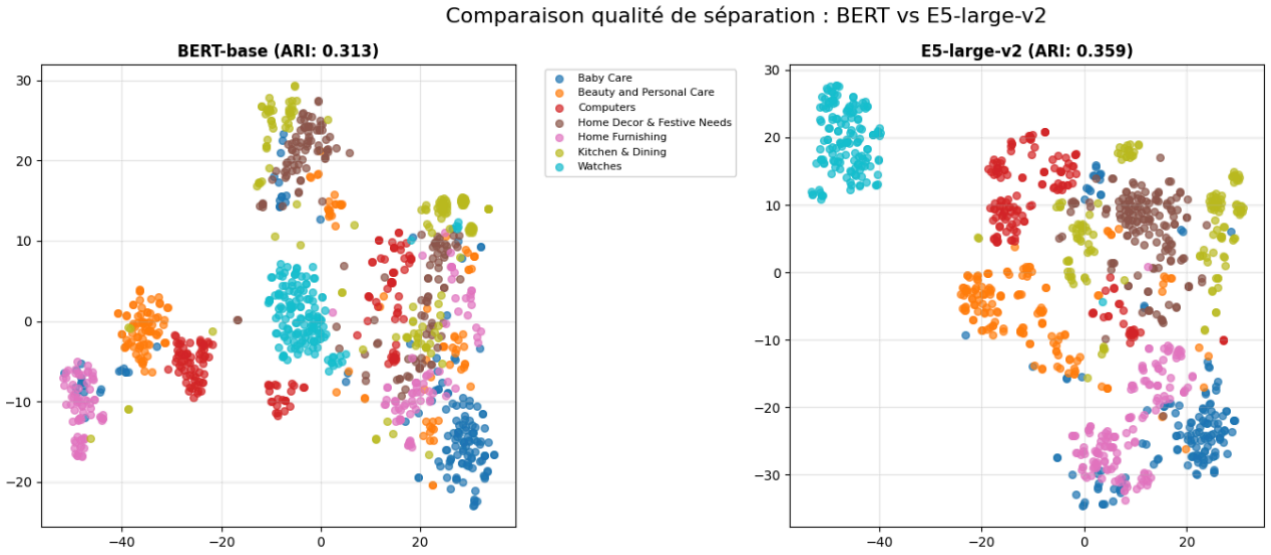
La projection t-SNE suggère des différences structurelles dans l'organisation de l'espace sémantique, cohérentes avec l'amélioration ARI observée.

BERT

- Clusters compacts mais avec chevauchements inter-catégories importants
- Séparation claire pour Computers et Watches uniquement
- Dispersion élevée pour Baby Care et Beauty & Personal Care

E5-large-v2

- Clusters plus distincts avec frontières un peu mieux définies
- Amélioration notable de la séparation Watches/Baby Care
- Réduction des chevauchements entre Home Furnishing/Home Decor



4. Analyse de l'importance globale et locale du modèle E5-large-v2

4.1 Méthodologie d'analyse d'interprétabilité

4.1.1 Approche par Random Forest

L'analyse d'interprétabilité d'E5-large-v2 repose sur un modèle Random Forest entraîné sur les embeddings comme features d'entrée. Cette approche permet de quantifier l'importance des 1024 dimensions vectorielles pour la classification des produits e-commerce.

Configuration du modèle :

- Random Forest : 100 estimateurs, profondeur maximale 10
- Données d'entraînement : 840 échantillons (80% du dataset)
- Données de test : 210 échantillons (20% du dataset)
- Accuracy obtenue : 90.48%

4.1.2 SHAP pour l'explicabilité

L'analyse utilise SHAP (SHapley Additive exPlanations) avec TreeExplainer pour quantifier les contributions individuelles de chaque dimension d'embedding. Cette méthode calcule les valeurs de Shapley pour chaque dimension. Elle garantit ainsi une attribution équitable des contributions à la prédiction finale.

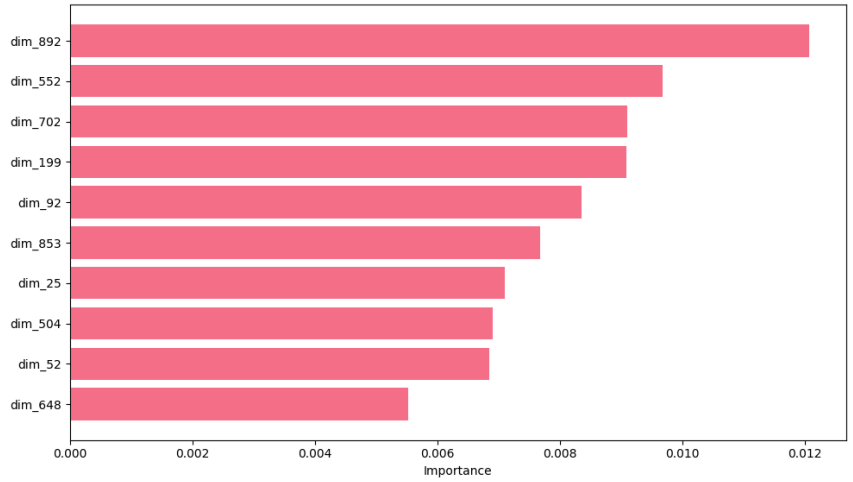
4.2 Importance globale des dimensions

4.2.1 Feature Importance Random Forest

L'analyse Random Forest révèle une concentration de l'information discriminante sur un nombre restreint de dimensions :

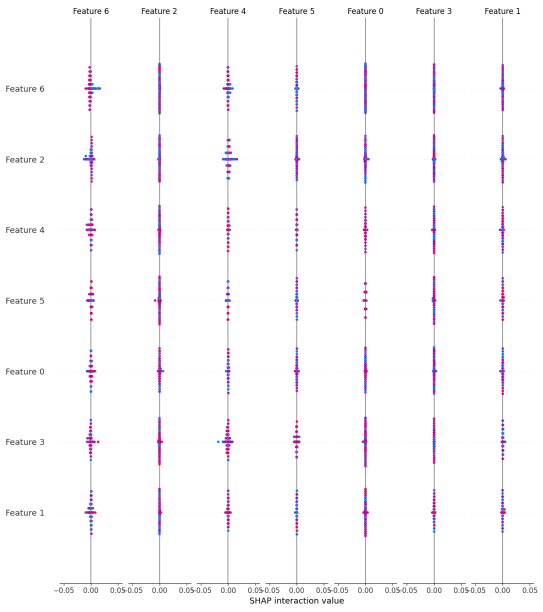
Les dimensions comme dim_892, dim_552, et dim_702 ont une forte influence sur les prédictions, car elles apparaissent dans le top 10 des features les plus importantes selon Random Forest.

Top 10 dimensions les plus importantes



E5-large-v2 concentre l'information sémantique discriminante sur environ 15-20 dimensions critiques parmi les 1024 disponibles.

4.2.2 Analyse SHAP globale



Ce graphique montre l'importance globale des dimensions d'embeddings E5-large-v2 dans la classification des produits e-commerce.

Les points représentent des valeurs SHAP pour chaque dimension d'embedding, et chaque colonne représente une classe de produit. Les points roses indiquent les valeurs élevées de la dimension, tandis que les points bleus montrent les valeurs faibles.

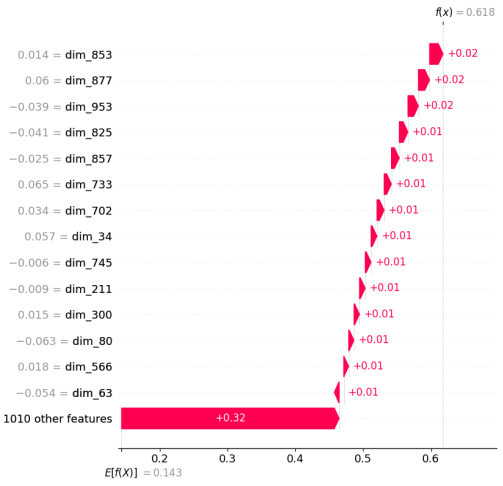
Les dimensions comme dim_6, dim_2, et dim_4 ont une forte influence sur les prédictions, car les points sont bien dispersés sur l'axe horizontal, indiquant qu'elles contribuent fortement à discriminer les catégories de produits.

Les points proches de 0 sur l'axe horizontal signifient que la dimension n'a pas beaucoup d'influence sur la classification pour certaines instances.

Chaque colonne (classe de produit) présente une signature particulière , elle montre qu'E5-large-v2 capture les nuances sémantiques entre les catégories e-commerce.

4.3 Analyse locale via Waterfall Plot

Ce graphique montre comment les dimensions d'embedding E5-large-v2 contribuent individuellement à la prédiction "Baby Care" pour un échantillon spécifique.



Le graphique part de la valeur de base du modèle (E[f(X)] = 0.143) et souligne comment chaque dimension modifie cette prédiction pour arriver au score final f(x) = 0.618.

Les dimensions dim_853 (+0.02), dim_877 (+0.02), et dim_733 (+0.01) renforcent la prédiction vers "Baby Care".

Les dimensions dim_220 (-0.01), dim_80 (-0.01), et dim_825 (-0.01) diminuent légèrement la probabilité de classification en "Baby Care".

Les "1010 other features" (+0.33) représentent l'effet combiné de toutes les autres dimensions, montrant que E5-large-v2 utilise l'ensemble de ses 1024 dimensions pour cette classification.

Le modèle arrive à une prédiction confiante (0.618) pour "Baby Care" grâce à la convergence de multiples signaux sémantiques contenus dans les embeddings.

4.4 Patterns d'importance sémantique

4.4.1 Spécialisation des dimensions

L'analyse révèle une concentration de l'information discriminante sur un nombre restreint de dimensions d'E5-large-v2 :

- Dimensions 892, 552, 702 : Contribution maximale à la discrimination
- Dimensions 92, 199, 25 : Importance secondaire significative
- Dimensions 853, 877 : Contribution notable pour certaines classes

Cette distribution suggère que E5-large-v2 structure efficacement l'information sémantique en concentrant les signaux discriminants sur environ 2% de ses dimensions disponibles.

4.4.2 Convergence des méthodes d'importance

La comparaison entre Random Forest et SHAP révèle une **cohérence partielle** dans l'identification des dimensions critiques, avec 6 dimensions communes dans le top 10..

4.4.3 Efficacité représentationnelle

E5-large-v2 démontre une **utilisation efficace** de son espace vectoriel 1024D en concentrant l'information discriminante sur moins de 2% des dimensions disponibles, ce qui valide l'architecture d'apprentissage contrastif pour la génération d'embedding..

5. Limites et améliorations possibles

5.1 Limites constatées

Le corpus de 1050 produits reste limité pour évaluer pleinement les capacités d'E5-large-v2. Ce modèle est conçu pour exploiter des volumes de données importants. Les modèles Transformer bénéficient habituellement des datasets supérieurs à 10 000 échantillons pour révéler leur potentiel discriminant.

L'évaluation repose exclusivement sur du clustering non supervisé (KMeans), omettant volontairement les tâches de classification supervisée. C’est sur ce point qu’E5 a démontré ses meilleures performances selon les benchmarks MTEB. Cette limitation méthodologique sous-estime probablement les capacités d'E5.

Les expérimentations sur processeur CPU (Google Colab) pénalisent E5-large-v2 davantage que BERT en raison de sa complexité supérieure. Un environnement GPU réduirait l'écart temporel.

Malgré les améliorations relatives, E5 atteint un ARI de 0.3595. Cette correspondance est imparfaite avec les catégories réelles. Cette performance suggère que le clustering automatique ne capture que partiellement la complexité sémantique du langage e-commerce sur notre dataset (réduit).

E5 introduit 148 erreurs inexistantes chez BERT, il est possible que le modèle rencontre une sur-sensibilité aux termes décoratifs dans les descriptions de montres.Cette régression locale questionne sur la robustesse du modèle sur certains sous-domaines.

5.2 Améliorations possibles

L'exploration systématique des hyperparamètres pourrait améliorer les performances via l’ajustement de la longueur maximale de tokenisation (actuellement limitée par défaut), l’optimisation du batch size selon les ressources disponibles, et des tests de différents préfixes instructionnels ("passage:", "document:" vs "query:") pour identifier la configuration optimale pour le clustering e-commerce.

Étendre l'évaluation à un corpus de 5 000 voire 10 000 produits e-commerce permettrait de mieux exploiter les capacités d'E5-large-v2. En effet, les modèles Transformer révèlent leur potentiel sur des volumes de données supérieurs, et cette accroissement pourrait révéler des améliorations de performance masquées par la taille limitée du dataset actuel.

L’évaluation d’E5-large-v2 dans son contexte d'usage optimal (ie classification supervisée) avec un classificateur linéaire simple sur les embeddings, permettant de quantifier le gain réel par rapport au clustering non-supervisé et de valider les conclusions sur des métriques de classification directes.