**Note méthodologique**

## **Dataset retenu**

Le dataset utilisé pour ce projet est celui du projet 6, il s’agit de la description des produits d’un site d’e-commerce. Il comprend 7 catégories de produit, reparties dans des classes égales. Il est donc possible d’utiliser l’accuracy comme métriques.

En se basant sur le nombre de catégories, le score de base d’un modèle faisant des prédictions au hasard sera de 1/7, soit 0.143

## **Modèle envisagé**

Recemment, les LLMs ont demontrés de très fortes capacités de compréhension de texte, en particulier grâce au très large nombre de texte utilisé pour l’entrainement de ces modèles.

Nous allons essayer d’appliquer Mistral7B sur ces données pour créer une classification des catégories de notre site d’e-commerce.

Afin de comparer les performances de notre modèle, nous allons comparer l’accuracy et le temps nécessaire pour effectuer les prédictions avec un modèle BERT pour les embeddings et un réseau de neurones profond entrainé sur ces embeddings pour prédire les catégories.

## **Références bibliographiques**

* [Zero-shot Topical Text Classification with LLMs - an Experimental Study](https://aclanthology.org/2023.findings-emnlp.647/) (Gretz et al., Findings 2023)
* [Pushing the limit of LLMs for text classification](https://dl.acm.org/doi/abs/10.1145/3701716.3715528) (Zhang et al., 2025)
* [LLMs classification performance is overclaimed](https://arxiv.org/abs/2406.16203) (Xu et al., 2024)

## **Explication de votre démarche de test du nouvel algorithme (votre preuve de concept)**

Pour savoir si Mistral7B est performant pour cette tâche, nous allons comparer les résultats à plusieurs modèles utilisés pour le projet 6. Comme nous n’avions fait qu’une étude de faisabilité, il faudra aussi ajouter un classificateur pour catégorifier les embeddings obtenues avec BERT. Ensuite nous comparerons les résultats de ces modèles avec Mistral7B avec et sans pré-entrainement des données en utilisant des métriques comme l’accuracy et le temps utilisé pour la prédiction des données.

Le but est de savoir si le gain de performance vaut le temps de prédiction potentiellement plus long lorsque l’on utilise un LLM pour classifier du texte.

Dans nos tests, l’accuracy du modèle de base (BERT) est de 0.88 pour un temps total de moins d’une minute pour effectuer les embeddings, entrainer le réseau de neurones, puis effectuer les prédictions.

Notre classificateur basé sur Mistral 7B n’a quant à lui qu’une accuracy de 0.73, bien plus pas que le modèle de référence aussi bien en zero-shot learning qu’en few-shot learning. Par ailleurs, le temps nécessaire pour effectuer les prédictions sur le jeu de test est de plus de 10 minutes.

Le modèle a donc une accuracy plus faible et un temps de prédiction plus long. Il n’est donc pas efficace d’utiliser un LLM dans notre cas.

En utilisant des LLMs plus récents, nous pourrions augmenter l’accuracy. Par ailleurs, il est possible de traiter plus de données en passant par l’API des modèles, qui nécessite un coût par token.