Les petits modèles sont bons : une étude empirique de classification dans un contexte zero-shot

Pierre Lepagnol, Thomas Gerald, Sahar Ghannay, Christophe Servan, Sophie Rosset

Les petits modèles sont bons : une étude empirique de classification dans un contexte zero-shot

Pierre Lepagnol[1,2], Thomas Gerald[1], Sahar Ghannay[1], Christophe Servan[1,3], Sophie Rosset[1]
[1]Université Paris-Saclay, CNRS, Laboratoire interdisciplinaire des sciences du numérique, 91405, Orsay, France
[2]SCIAM, 7508, Paris, France
[3]QWANT Research, 7 Rue spontini, 75116 Paris, France
{firstname.lastname}@lisn.upsaclay.fr

RÉSUMÉ

Ce travail s'inscrit dans le débat sur l'efficacité des grands modèles de langue par rapport aux petits pour la classification de texte par amorçage (prompting). Nous évaluons ici le potentiel des petits modèles de langue dans la classification de texte sans exemples, remettant en question la prédominance des grands modèles. À travers un ensemble diversifié de jeux de données, notre étude compare les petits et les grands modèles utilisant différentes architectures et données de pré-entraînement. Nos conclusions révèlent que les petits modèles peuvent générer efficacement des étiquettes et, dans certains contextes, rivaliser ou surpasser les performances de leurs homologues plus grands. Ce travail souligne l'idée que le modèle le plus grand n'est pas toujours le meilleur, suggérant que les petits modèles économes en ressources peuvent offrir des solutions viables pour des défis spécifiques de classification de données.

ABSTRACT

Here the title in English.

This study is part of the debate on the efficiency of large versus small language models for text classification by prompting. We assess the potential of small language models in zero-shot text classification, challenging the prevailing dominance of large models. Across a diverse set of datasets, our investigation benchmarks both small and large models using different architectures and pre-training data. Our findings reveal that small models can effectively generate labels and, in certain contexts, rival or surpass the performance of their larger counterparts. This research underscores the notion that bigger isn't always better, suggesting that resource-efficient small models may offer viable solutions for specific data classification challenges.

MOTS-CLÉS : Zero-shot, Prompting, amorçage, Modèle de langue, LLM, labelisation de données.

KEYWORDS: Zero-shot, Prompting, language modeling, LLMs, data labeling.

1 Introduction

Les grands modèles de langue (LLM) ont été largement favorisés par rapport aux modèles plus petits pour résoudre les tâches grâce aux méthodes par amorces (prompting) (Brown *et al.*, 2020; Hoffmann *et al.*, 2022; OpenAI, 2023; Chowdhery *et al.*, 2022) dans le contexte d'indisponibilité de données d'entraînement (zero-shot prompting).

Bien que les grand modèles soient très performants, leur utilisations présentent des difficultés - ils sont gourmands en ressources, coûteux à employer et leurs performances ne sont pas toujours garanties pour chaque tâche (Nityasya *et al.*, 2021).

De toujours plus grands modèles (Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022) ont été construits et des jeux de données toujours plus sophistiqués ont été nécessaires (Zhang *et al.*, 2023) pour atteindre de bonnes performances. Leurs performances a priori supérieures en ont fait un choix privilégié pour diverses tâches, même pour les tâches de classification de base.

Au fur et à mesure que le domaine du traitement du langage évolue, nous devons nous poser la question suivante : les grands modèles de langue sont-ils essentiels pour une classification efficace des données ?

Dans cet article, nous examinons dans quelle mesure les petits modèles peuvent rivaliser avec les grands modèles dans la création d'étiquettes. Sur différents jeux de données, nous voulons voir comment les petits modèles peuvent correctement étiqueter des textes sans exemple (zero-shot classification). Nous cherchons aussi à déterminer ce qui permet aux modèles d'obtenir de bons résultats sur les tâches de classification. Nous comparons le fonctionnement des petits et des grands modèles dans le cadre d'amorçage sans exemples (prompting zero-shot) sur différents jeux de données afin de déterminer s'il est possible d'obtenir de bons résultats avec moins de ressources.

Nous pensons que cette étude est le point de départ de la compréhension des capacités réelles des LLM lorsqu'ils sont utilisés pour des tâches de classification dans un contexte zero-shot.

Nos contributions principales sont :

1. Une évaluation d'un grand nombre de modèles de langue (de 77 millions à 70 milliards de paramètres) ajustés sur des jeux de données d'instructions, avec différentes architectures (encodeur-décodeur ou décodeur seul) et tailles sur de 15 jeux de données dans un contexte zero-shot.
2. Des preuves relativement solides de l'efficacité des petits modèles dans la classification zero-shot, où les performances des petits modèles sont comparables à celles de plus grands sur de nombreux jeux de données dans les tâches de classification.
3. Nos évaluations sont mises à la disposition de la communauté en open-source, présentant les méthodologies proposées, contribuant ainsi à l'intégrité et à la robustesse des études dans ce domaine. Le code est disponible en ligne dans le dépôt XXXX.

L'article est organisé comme suit : La section 2 présente une revue de la littérature sur les approches zero-shot. Dans la section 3, nous décrivons la méthodologie que nous suivons pour cette étude. La section 4 présente les résultats et analyses. Enfin, nous concluons dans la section 5 et discutons des limitations et travaux futurs.

2 Travaux connexes

Classification de texte sans exemples & amorçage (Prompting) Le prompting consiste à fournir un texte d'entrée (ou prompt) à un modèle de langue, qui génère ensuite un texte de sortie basé sur ce prompt. L'objectif de la classification de texte sans exemples est de catégoriser des textes avec des étiquettes sans entraînement préalable spécifique à cette tâche. Cette approche attire l'attention du monde industriel et de la communauté scientifique car elle vise à supprimer le besoin d'ajustement supplémentaire et, par conséquent, de données étiquetées additionnelles, qui sont souvent onéreuses à obtenir.

Pour que le système, ici le modèle de langue, obtienne de bonnes performances sur les classes non vues, il est nécessaire d'avoir des descriptions précises des classes non-vues, comme l'ont noté Xia *et al.* (2018) et Liu *et al.* (2019a). Fei *et al.* (2022) améliorent la classification zero-shot en segmentant les textes d'entrée et en exploitant les amorces (prompts) spécifiques aux classes. Meng *et al.* (2020) ont proposé une stratégie qui utilise des noms d'étiquettes combinés à un auto-apprentissage adapté à la classification zero-shot. De nombreuses méthodes nécessitent un jeu de données non étiquetées ou une base de connaissances pour extraire les étiquettes pertinentes et faciliter l'auto-apprentissage.

Pour utiliser des modèles de langues et les méthodes d'amorçage (prompting) Schick & Schütze (2021) propose d'exploiter des paires schéma-verbalisateur (pattern-verbalizer pairs) où le schéma représente le prompt et le verbalisateur représente un mot, un token, par classe qui sera associé sur la dite classe. Néanmoins, ils utilisent cette méthode non pas dans un cadre zero-shot mais dans un cadre de classification avec ajustement.

Plus récemment, Zhao *et al.* (2023b) ont proposé d'utiliser k-Nearest-Neighbor fondé sur la similarité entre plongement des mots du verbalisateur pour augmenter les performances de classification. Lu *et al.* (2023) ont proposé la sélection par la perplexité pour sélectionner les meilleurs prompts dans un contexte d'essai à zero-shot.

Alors que les travaux antérieurs se sont concentrés sur de nouvelles méthodes visant à rendre les modèles de langues plus performants en matière de classification sans exemples, nous souhaitons avoir un aperçu des caractéristiques des modèles et de leurs performances.

3 Dispositif expérimental

Bien que les auteurs de LLMs aient comparé leurs différentes tailles de modèles (Kaplan *et al.*, 2020; Hoffmann *et al.*, 2022), cette étude élargit cette analyse en comparant directement différentes architectures sur un ensemble étendu de jeux de données. Nous amorçons (promptons) divers modèles de langue en utilisant 4 fonctions de scoring différentes (voir Section 3) pour classifier les phrases. Nous évaluons la qualité de nos classifieurs par l'exactitude et le macro F1 score.

Tâches & Jeux de Données Nous examinons les performances des modèles sur 15 jeux de données, sélectionnés pour représenter divers défis en classification. Par exemple nous utilisons les jeux de données *AGNews*, avec ses 4 classes distinctes, et *BBCNews*, qui propose 5 classes. Pour la classification de sentiments, la plupart des jeux de données proposent un choix binaire, comme pour *ethos* (Mollas *et al.*, 2022) ou plus granulaire comme *sst-5* (Socher *et al.*, 2013) avec 5 classes.

La tâche de classification de spams inclut les jeux de données *youtube* (Alberto *et al.*, 2015) ou *sms* (Almeida & Hidalgo, 2012). La tâches de classification de relations inclut les jeux de données tels que *semeval* (Hendrickx *et al.*, 2010). La liste complète des jeux de donnée est en annexe B (Tableau 7).

Les jeux de données sélectionnés sont équilibrés en termes de classes. On considère un jeu de données comme déséquilibré si la classe majoritaire est au moins 2 fois plus grande que la classe minoritaire. Ainsi nous avons choisi le macro F1-score pour les jeux de données déséquilibrés et l'exactitude pour le reste.

Modèles Notre étude évalue un total de 72 modèles uniques. Nous sélectionnons à la fois des modèles encodeur-décodeur (comme T5 (Raffel *et al.*, 2020), mT0 (Muennighoff *et al.*, 2023) et Bart (Lewis *et al.*, 2020)) et des modèles uniquement décodeur causal (tels que Llama (Touvron *et al.*, 2023) et Falcon (Penedo *et al.*, 2023)). Nous optons pour différentes tailles pour les mêmes modèles, allant de quelques millions à des centaines de milliards de paramètres. Par exemple, le modèle Bart possède 255M ou 561M de paramètres, Falcon existe en version 7B ou 40B [1]. Ces modèles ont été choisis en fonction de leur prévalence dans la littérature, de leur efficacité rapportée sur des tâches similaires et du fait que des versions adaptées aux instructions étaient disponibles pour certains d'entre eux. L'ajustement sur les instructions fait référence à la stratégie de fine-tuning d'un modèle de langue sur des jeux de données d'instructions (Longpre *et al.*, 2023).

La liste complète des modèles est en annexe A (Tableaux 5 et 6).

Amorces (Prompts) Les prompts de nos expériences sont issues de la littérature et conçu pour être simples et répondu par un seul token par le modèle de langue. Les amorces sont soit des traductions de fonctions d'étiquetage issues du benchmark WRENCH (Zhang *et al.*, 2021), soit créées de zéro dans le même style. Elles sont adaptées à chaque tâche, *par exemple* les amorces pour le jeu de données *sms* sont formulées différemment de celles pour le jeu de données *bbcnews*. Ceci permet d'assurer la pertinence par rapport au domaine et maximiser la compréhension du modèle. La liste des amorces par jeux de données est en annexe C, tableau 9.

Ainsi pour *sms*, le couple amorce/verbalisateur est

Amorce (Prompt)

```
Is the following message spam? Answer by yes or no.\n"{TEXT}"
```

Verbalisateur Texte/Classe

```
{1:"yes", 0:"no"}
```

Pour *bbcnews*, le couple amorce/verbalisateur est :

Amorce (Prompt)

```
"{TEXT}" is about "
```

Verbalisateur Texte/Classe

```
{0: "tech", 1: "business", 2: "sport", 3: "entertainment", 4: "politics"}
```

1. Nous n'avons pas testé Falcon 180B, car il n'était pas disponible pendant nos expériences

Scoring Functions Dans la classification fondée sur les amorces (prompts), l'utilisation d'un verbalisateur associant des tokens aux étiquettes de classe est cruciale pour une classification précise. Comme suggéré par (Holtzman *et al.*, 2022), de nombreuses séquences valides peuvent représenter le même concept, ceci est appelé *compétition pour la forme de surface*. Par exemple, `"+"`, `"positif"`, `"Plus positif que l'opposé"` pourraient être utilisés pour représenter le même concept de positivité pour la tâche d'analyse des sentiments. Comme cette compétition existe, la manière dont les verbalisateurs sont conçus influence grandement l'efficacité des approche par prompting pour la classification. Zhao *et al.* (2023b) utilisent le k-Plus Proches Voisins pour la construction de verbalisateur et augmentent leurs verbalisateurs fondés sur la similarité des embeddings.

Dans cette étude nous utilisons plusieurs fonctions de scoring pour évaluer leur impact sur les performances de nos modèles, dont celles proposées par (Holtzman *et al.*, 2022).

Probability	$\arg\max_i \mathbb{P}(y_i \mid x')$
DCPMI	$\arg\max_i \frac{\mathbb{P}(y_i \mid x')}{\mathbb{P}(y_i \mid x_{\text{domain\_conditional}})}$
PMI	$\arg\max_i \frac{\mathbb{P}(y_i \mid x')}{\mathbb{P}(y_i \mid x_{\text{domain\_unconditional}})}$
Similarity	$\arg\max_{c_i \in C} \cos(e(t_0), e(y_i))$ [2]

Outils pour l'Analyse Statistique Les trois principaux outils statistiques utilisés sont détaillés ci-après :

Le Biweight Midcorrelation Coefficient est une alternative robuste au coefficient de corrélation de Pearson pour quantifier l'association entre deux échantillons. Il est conçu pour être moins sensible aux valeurs aberrantes que d'autres coefficients tels que celui de Pearson.

Analyse de Covariance - ANCOVA combine les techniques d'ANOVA et de régression pour évaluer si les moyennes d'une variable dépendante sont égales à travers les niveaux d'une variable indépendante catégorielle tout en contrôlant statistiquement pour les effets d'autres variables continues (covariables).

Test de Kruskal-Wallis est une méthode non paramétrique pour tester si des échantillons proviennent de la même distribution. Nous l'avons utilisé comme une méthode non paramétrique, qui ne suppose pas une distribution normale des résidus, contrairement à l'analyse de variance standard à un facteur.

4 Résultats

Sur les jeux de données mentionnés précédemment, nous comparons la performance des modèles de langue et nous étudions : 1) la relation entre les performances des modèles et leurs tailles (le nombre de paramètres), 2) les performances et leurs architectures, et 3) les performances et si le modèle a été fine-tuné sur des jeux de données d'instructions. Ensuite, pour les deux types d'architectures (encodeur-décodeur et décodeur seul), nous étudions l'impact de l'ajustement sur des instructions.

Le tableau 1 présente les scores de l'état de l'art pour chaque jeu de données [3].

3. Les jeux de données *agnews*, *imdb*, *yelp*, *trec* sont inclus dans l'entraiement du modèle `mT0`. Nous ne considérons donc pas ses scores sur ces jeux de données.

dataset	SOTA	Classe Maj.	Meilleur Score	Modèle	# Paramètres
agnews	0.625	0.266	**0.734**	MBZUAI/LaMini-GPT-124M	163.0 M
bbcnews	NaN	0.236	0.869	bigscience/mt0-large	1.2 B
cdr	NaN	0.676	0.717	bigscience/bloomz-3b	3.6 B
chemprot	0.172	0.049	**0.192**	bigscience/bloomz-3b	3.6 B
ethos	0.667	0.566	0.597	bigscience/bloomz-1b1	1.5 B
financial_phrasebank	0.528	0.254	**0.744**	MBZUAI/LaMini-GPT-774M	838.4 M
imdb	0.718	0.500	**0.933**	MBZUAI/LaMini-Flan-T5-783M	783.2 M
semeval	0.435	0.054	0.270	bigscience/mt0-xxl	12.9 B
sms	0.340	0.464	**0.699**	mosaicml/mpt-7b	6.6 B
spouse	0.630	0.479	0.521	gpt2	163.0 M
sst-2	0.710	0.501	**0.956**	bigscience/bloomz-3b	3.6 B
sst-5	0.598	0.286	0.485	tiiuae/falcon-40b-instruct	41.8 B
trec	NaN	0.072	0.324	mosaicml/mpt-7b-instruct	6.6 B
yelp	0.888	0.522	**0.977**	MBZUAI/LaMini-Flan-T5-783M	783.2 M
youtube	0.468	0.528	**0.716**	tiiuae/falcon-40b	41.8 B

TABLE 1 – Tableau illustrant les mesures de performance pour différents jeux de données : Les colonnes sont (1) le nom de l'ensemble de données, (2) les scores de l'état de l'art (SOTA), (3) les scores obtenus en prédisant systématiquement la classe majoritaire, (4) les scores les plus élevés (surlignés en rouge lorsqu'ils sont meilleurs), (5) les modèles ayant ces meilleurs scores, et (6) le nombre de paramètres pour chaque modèle.
Notez la présence d'entrées `NaN`, signifiant des jeux de données pour lesquels les références SOTA n'ont pas été établies ou trouvées.

4.1 La taille du modèle n'a pas vraiment d'importance

La Figure 1 présente la relation entre le nombre de paramètres et la performance en termes de scores Acc/F1 à travers divers jeux de données. Nous calculons le Biweight Midcorrelation Coefficient et les p-valeurs associées pour chaque jeu de données. Ces résultats sont détaillés dans le tableau 2.

dataset	correlation coef	pvalue
agnews	-0.1418	**0.0536**
bbcnews	0.0489	0.4877
cdr	0.2541	**0.0002**
chemprot	0.1318	**0.0531**
ethos	-0.1519	**0.0256**
financial_phrasebank	0.0419	0.5406
imdb	-0.2862	**0.0001**
semeval	-0.0506	0.4595
sms	-0.1209	0.0763
spouse	-0.0254	0.7106
sst-2	0.0755	0.2693
sst-5	0.0061	0.9293
trec	-0.1085	0.1403
yelp	-0.0620	0.4008
youtube	-0.0014	0.9836

TABLE 2 – Biweight Midcorrelation Coefficients et p-valeurs mesurant la relation entre le score et la taille du modèle (log du nombre de paramètres) selon les datasets

D'après notre analyse, 10 des 15 jeux de données présentent des p-valeurs supérieures à 0,05, ce

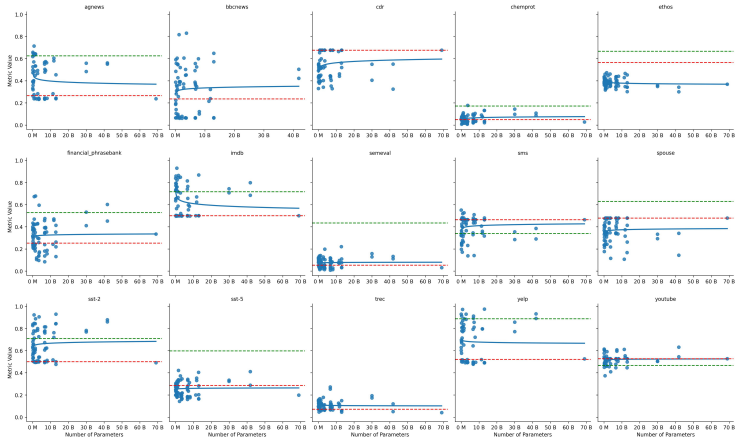

FIGURE 1 – Comparaison de la performance de différentes tailles de modèles à travers les jeux de données. Les ligne en bleu indiquent la tendance générale, les lignes pointillées en rouge indiquent pour les résultats des prédictions de la classe majoritaire, et la ligne verte indique les résultats état de l'art pour les méthodes de prompting zero-shot.

qui suggère qu'il n'y a pas de lien significatif entre les scores de performance et la taille du modèle. Cependant, trois jeux de données présentent des p-valeurs inférieures à 0,05, ce qui indique une corrélation notable. Parmi ceux-ci, la corrélation est positive pour le jeu de données *cdr* mais négative pour *ethos* et *imdb*. Deux jeux de données, à savoir *agnews* et *chemprot*, présentent des p-valeurs proches du seuil de 0,05, ce qui rend leur corrélation peu concluante.

En conclusion, alors que de nombreux jeux de données ne montrent pas de relation directe entre des tailles de modèle plus grandes et une amélioration des performances, des jeux de données comme *cdr*, *ethos*, et *imdb* le font. De plus, la variance du coefficient de corrélation entre les jeux de données suggère que la taille du modèle n'est pas le seul facteur déterminant des performances.

4.2 Impact de l'architecture sur les performances

La figure 2 illustre les variations de performances entre les architectures encodeur-décodeur et décodeur seul. En utilisant l'ANCOVA, nous mesurons l'impact du choix de l'architecture sur les scores de performance, tout en contrôlant l'effet de la taille du modèle. Les résultats sont présentés dans le tableau 3.

D'une part, 7 des 15 jeux de données, nommément *agnews*, *bbcnews*, *chemprot*, *semeval*, *sms*, *spouse* et *youtube*, présentent des p-valeurs inférieures à 0,05, ce qui suggère que l'architecture a un impact significatif. En revanche, les jeux de données tels que *cdr*, *ethos* et *financial_phrasebank* ne sont pas affectés par le choix de l'architecture. Le jeux de données *imdb* présente un impact non concluant. En conclusion, bien que la taille du modèle ne soit pas un facteur dominant, le choix de l'architecture a un impact significatif sur les performances dans ces jeux de données spécifiques.

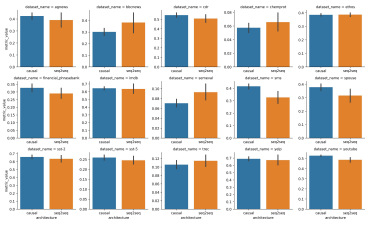

FIGURE 2 – Variation des performances entre les différentes architectures.

Dataset	statistique	p-valeur	Variances Egales
agnews	4.0676	**0.0452**	True
bbcnews	7.0640	**0.0085**	False
cdr	0.2519	0.6163	True
chemprot	4.4883	**0.0353**	True
ethos	0.3945	0.5306	False
financial_phrasebank	1.4592	0.2284	False
imdb	3.6687	0.0570	True
semeval	8.2301	**0.0045**	True
sms	11.9951	**0.0006**	False
spouse	4.7794	**0.0299**	True
sst-2	0.2501	0.6175	True
sst-5	0.7852	0.3766	True
trec	0.3382	0.5616	False
yelp	0.7103	0.4004	True
youtube	18.0011	**0.0000**	False

FIGURE 3 – ANCOVA indiquant l'impact de l'architecture sur les scores selon le dataset

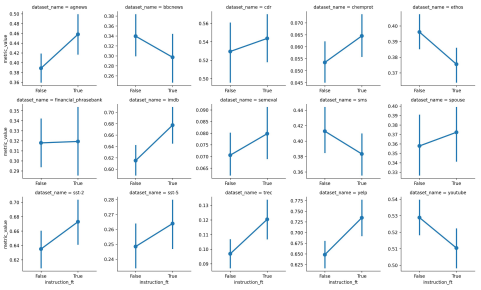

FIGURE 4 – Comparaison des performances des modèles avec ou sans instruction sur l'ensemble des données.

dataset	statistique	p-valeur	Variances Egales
agnews	10.5411	**0.0014**	True
bbcnews	1.9492	0.1642	True
cdr	0.1635	0.6864	True
chemprot	2.3152	0.1296	True
ethos	5.8015	**0.0169**	True
financial_phrasebank	0.0001	0.9917	False
imdb	13.6945	**0.0003**	True
semeval	1.4016	0.2378	False
sms	2.6667	0.1039	True
spouse	0.3379	0.5617	True
sst-2	3.0055	0.0844	False
sst-5	1.8271	0.1779	True
trec	8.3534	**0.0043**	False
yelp	12.5571	**0.0005**	True
youtube	5.8369	**0.0165**	True

FIGURE 5 – ANCOVA indiquant l'impact de l'instruction fine-tunning sur les scores selon le dataset

4.3 Impact de l'Instruction Fine-tuning sur les performances

De la même manière que pour l'architecture, nous avons quantifié l'impact de l'ajustement sur des instructions sur les performances tout en contrôlant le nombre de paramètres. Nous avons utilisé l'ANCOVA pour tester si les moyennes de nos scores ACC/F1 sont égales pour toutes les catégories de la variable `instruction_ft`, tout en contrôlant statistiquement l'effet du nombre de paramètres. Les résultats sont présentés dans le tableau 5.

La figure 4 montre l'impact de l'ajustement sur des instructions sur les scores de performance selon les datasets. L'axe des y de chaque graphique affiche le score de performance (Acc/F1). L'axe des x a deux valeurs : `False` et `True`, indiquant si le modèle a été ajusté sur des instructions ou non.

Pour de nombreux jeux de données, le fine-tuning sur les instructions améliore les performances par rapport à l'absence de fine-tuning (*agnews*, *ethos*, *imdb*, *trec*, *yelp*, et *youtube*) comme le suggère les p-valeurs significatives de l'ANCOVA. Une diminution des perfomances semble survenir lorsque les modèles sont ajustés sur des instructions pour certains jeux de données. Pour *bbcnews*, *youtube* et

sms, l'ANCOVA nous indique que cette diminution n'est pas significative, en revanche, pour *ethos*, elle est significative.

Pour les autres jeux de données, bien qu'il puisse y avoir des différences visuelles dans les performances avec et sans fine-tuning sur les instructions, ces différences ne sont pas statistiquement significatives d'après les p-valeurs. Par conséquent, bien que le fine-tuning sur les instructions ait le potentiel d'améliorer les performances des modèles sur de nombreux jeux de données, son impact peut varier en fonction des jeux de données spécifiques.

4.4 Relation entre la taille du modèle et les performances par architecture

Le tableau 3 présente les coefficients de corrélation moyenne pondérée entre les tailles de modèle (log du nombre de paramètres) et les scores de performance pour les deux types architectures étudiées. On peut y retrouver une corrélation légère mais significative pour les modèles de décodeurs, mais largement non-significative pour les modèles de encodeur-décodeur. Cela suggère que le décodeur seul pourrait être plus sensible au nombre de paramètres ; un trop grand nombre de paramètres pourrait nuire aux performances.

dataset	correlation coef	pvalue
causal	-0.0435	**0.0299**
seq2seq	0.0065	0.8728

TABLE 3 – Biweight Midcorrelation Coefficients et p-valeurs mesurant la relation entre le score et la taille du modèle (log du nombre de paramètres) selon les architectures

4.5 Impact de l'ajustement sur des instructions et des performances par architecture

La figure 6 compare visuellement l'impact de l'ajustement sur des instructions et les scores de performance pour les deux architectures. En ordonnée est le score de performance (Acc/F1), en abscisse une variable à deux modalité indiquant si le modèle a été ajusté sur des instructions.

Une ANCOVA est réalisée pour quantifier l'impact de l'ajustement sur des instructions sur chaque architecture (encodeur-décodeur/décodeur-seul) tout en contrôlant l'effet de la taille du modèle. Le tableau 4 présente les statistiques et les p-valeurs.

Pour l'architecture décodeur-seul, il n'y a pas d'impact significatif de l'ajustement sur des instructions sur les scores. La p-valeur est ici de 0,6693, bien supérieure à 0,05. Pour l'architecture encodeur-décodeurs, il y a un impact significatif de l'ajustement sur des instructions sur les scores. La p-valeur pour l'architecture encodeur-décodeur est surlignée en rouge et s'élève à 0,0086, soit moins de 0,05.

La différence de résultats entre les deux architectures suggère que l'impact de l'ajustement sur des instructions pourrait dépendre de l'architecture. Tant l'analyse graphique que l'ANCOVA montrent un effet de l'ajustement sur des instructions sur l'architecture encodeur-décodeur.

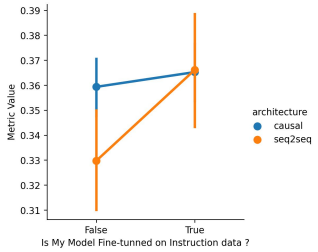

FIGURE 6 – Comparaison des performances entre les modèles ajustés par instruction et les modèles non ajustés, selon l'architecture

dataset	statistique	p-valeur	Variances Egales
causal	0.1825	0.6693	True
seq2seq	6.9406	**0.0086**	False

TABLE 4 – "ANCOVA indiquant l'impact de `instruction_ft` sur Acc/F1 selon les architectures

5 Conclusion & Perspectives

Ce travail avait pour but de déterminer si l'utilisation de grands modèles était nécessaire pour aborder les tâches de classification en utilisant des techniques de prompting.

La performance des modèles de langue varie en fonction de multiples facteurs, y compris la taille du modèle, les choix architecturaux et les stratégies de fine-tuning. Si un modèle plus grand entraîne une amélioration des performances, ce n'est pas un facteur clef, le choix d'architecture du modèle a un impact plus notable sur les résultats obtenus sur nos jeux de données. L'impact de l'ajustement sur des instructions est également évident, mais son efficacité dépend de l'architecture. Une étude complète d'autres architectures émergentes, telles que l'architecture RWKV (Peng *et al.*, 2023) ou les modèles fondés sur les states spaces models, pourrait apporter des nuances et des détails à cette analyse. L'impact varié de l'ajustement sur des instructions à travers les jeux de données suggère le besoin de techniques de fine-tuning plus avancées comme l'incorporation de recherche d'informations pour assurer de meilleures performances de classification lors de l'armocage.

6 Limitations

Dans cette étude, nous avons limité notre évaluation à des prompts simples, non optimisés pour obtenir les meilleurs réponses et nous n'avons pas étudié la variablité des résulats pour différents prompts (test d'un seul prompt pour pour chaque jeu de données). De plus, nous avons concentré notre étude sur les modèles encodeur-décodeur et décodeurs-seul sans les comparer avec des modèles encodeurs-seul. Nous n'avons pas étudié pas la sensibilité des performances à certains facteurs externes tels que le temps de pré-entraînement, la qualité des données de pré-entraînement ou les biais potentiels dans les jeux de données. Ces facteurs externes pourraient influencer les résultats ou le caractère universel des conclusions. Le choix et les hypothèses des outils statistiques pourraient influencer les résultats. Cette étude n'inclut pas les modèles publiés très récemment. Ainsi le comportement de modèles très récents comme les modèles conversationnels entraînés avec du RLHF/DPO pourrait exhiber des différences dans nos conclusions.

Références

ALBERTO T. C., LOCHTER J. V. & ALMEIDA T. A. (2015). TubeSpam : Comment Spam Filtering on YouTube. *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)*, p. 138–143. DOI : 10.1109/ICMLA.2015.37.

ALMEIDA T. & HIDALGO J. (2012). SMS Spam Collection. UCI Machine Learning Repository. DOI : https ://doi.org/10.24432/C5CC84.

BIDERMAN S., SCHOELKOPF H., ANTHONY Q., BRADLEY H., O'BRIEN K., HALLAHAN E., KHAN M. A., PUROHIT S., PRASHANTH U. S., RAFF E., SKOWRON A., SUTAWIKA L. & VAN DER WAL O. (2023). *Pythia : A Suite for Analyzing Large Language Models Across Training and Scaling*. Rapport interne. arXiv :2304.01373 [cs] type : article, DOI : 10.48550/arXiv.2304.01373.

BIGSCIENCE WORKSHOP (2022). BLOOM (revision 4ab0472). DOI : 10.57967/hf/0003.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CHO H., KIM Y. & LEE S.-G. (2023). *CELDA : Leveraging Black-box Language Model as Enhanced Classifier without Labels*. Rapport interne. arXiv :2306.02693 [cs] type : article.

CHOWDHERY A., NARANG S., DEVLIN J., BOSMA M., MISHRA G., ROBERTS A., BARHAM P., CHUNG H. W., SUTTON C., GEHRMANN S., SCHUH P., SHI K., TSVYASHCHENKO S., MAYNEZ J., RAO A., BARNES P., TAY Y., SHAZEER N., PRABHAKARAN V., REIF E., DU N., HUTCHINSON B., POPE R., BRADBURY J., AUSTIN J., ISARD M., GUR-ARI G., YIN P., DUKE T., LEVSKAYA A., GHEMAWAT S., DEV S., MICHALEWSKI H., GARCIA X., MISRA V., ROBINSON K., FEDUS L., ZHOU D., IPPOLITO D., LUAN D., LIM H., ZOPH B., SPIRIDONOV A., SEPASSI R., DOHAN D., AGRAWAL S., OMERNICK M., DAI A. M., PILLAI T. S., PELLAT M., LEWKOWYCZ A., MOREIRA E., CHILD R., POLOZOV O., LEE K., ZHOU Z., WANG X., SAETA B., DIAZ M., FIRAT O., CATASTA M., WEI J., MEIER-HELLSTERN K., ECK D., DEAN J., PETROV S. & FIEDEL N. (2022). *PaLM : Scaling Language Modeling with Pathways*. Rapport interne. arXiv :2204.02311 [cs] type : article.

CHUNG H. W., HOU L., LONGPRE S., ZOPH B., TAY Y., FEDUS W., LI Y., WANG X., DEHGHANI M., BRAHMA S., WEBSON A., GU S. S., DAI Z., SUZGUN M., CHEN X., CHOWDHERY A., CASTRO-ROS A., PELLAT M., ROBINSON K., VALTER D., NARANG S., MISHRA G., YU A., ZHAO V., HUANG Y., DAI A., YU H., PETROV S., CHI E. H., DEAN J., DEVLIN J., ROBERTS A., ZHOU D., LE Q. V. & WEI J. (2022). *Scaling Instruction-Finetuned Language Models*. Rapport interne. arXiv :2210.11416 [cs] type : article.

CLARKE C., HENG Y., KANG Y., FLAUTNER K., TANG L. & MARS J. (2023). Label Agnostic Pre-training for Zero-shot Text Classification. In *Findings of the Association for Computational Linguistics : ACL 2023*, p. 1009–1021, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-acl.64.

DAVIS A. P., GRONDIN C. J., JOHNSON R. J., SCIAKY D., KING B. L., MCMORRAN R., WIEGERS J., WIEGERS T. C. & MATTINGLY C. J. (2016). The comparative toxicogenomics database : update 2017. *Nucleic Acids Res*, **45**(D1), D972–D978.

DEY N., GOSAL G., ZHIMING, CHEN, KHACHANE H., MARSHALL W., PATHRIA R., TOM M. & HESTNESS J. (2023). *Cerebras-GPT : Open Compute-Optimal Language Models Trained on the Cerebras Wafer-Scale Cluster*. Rapport interne. arXiv :2304.03208 [cs] type : article.

FEI Y., MENG Z., NIE P., WATTENHOFER R. & SACHAN M. (2022). Beyond prompting : Making Pre-trained Language Models Better Zero-shot Learners by Clustering Representations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, p. 8560–8579, Abu Dhabi, United Arab Emirates : Association for Computational Linguistics. DOI : 10.18653/v1/2022.emnlp-main.587.

HENDRICKX I., KIM S. N., KOZAREVA Z., NAKOV P., Ó SÉAGHDHA D., PADÓ S., PENNACCHIOTTI M., ROMANO L. & SZPAKOWICZ S. (2010). SemEval-2010 task 8 : Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, p. 33–38, Uppsala, Sweden : Association for Computational Linguistics.

HOFFMANN J., BORGEAUD S., MENSCH A., BUCHATSKAYA E., CAI T., RUTHERFORD E., CASAS D. D. L., HENDRICKS L. A., WELBL J., CLARK A., HENNIGAN T., NOLAND E., MILLICAN K., DRIESSCHE G. V. D., DAMOC B., GUY A., OSINDERO S., SIMONYAN K., ELSEN E., RAE J. W., VINYALS O. & SIFRE L. (2022). *Training Compute-Optimal Large Language Models*. Rapport interne. arXiv :2203.15556 [cs] type : article.

HOLTZMAN A., WEST P., SHWARTZ V., CHOI Y. & ZETTLEMOYER L. (2022). *Surface Form Competition : Why the Highest Probability Answer Isn't Always Right*. Rapport interne. arXiv :2104.08315 [cs] type : article, DOI : 10.48550/arXiv.2104.08315.

HSIEH C.-Y., LI C.-L., YEH C.-K., NAKHOST H., FUJII Y., RATNER A., KRISHNA R., LEE C.-Y. & PFISTER T. (2023). *Distilling Step-by-Step ! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes*. Rapport interne. arXiv :2305.02301 [cs] type : article.

KAPLAN J., MCCANDLISH S., HENIGHAN T., BROWN T. B., CHESS B., CHILD R., GRAY S., RADFORD A., WU J. & AMODEI D. (2020). *Scaling Laws for Neural Language Models*. Rapport interne. arXiv :2001.08361 [cs, stat] type : article, DOI : 10.48550/arXiv.2001.08361.

KRALLINGER M., RABAL O., AKHONDI S., PÉREZ M., SANTAMARÍA J., RODRÍGUEZ G. P., TSATSARONIS G., INTXAURRONDO A., LÓPEZ J. A. B., NANDAL U., BUEL E. V., CHANDRASEKHAR A., RODENBURG M., LÆGREID A., DOORNENBAL M. A., OYARZÁBAL J., LOURENÇO A. & VALENCIA A. (2017). Overview of the BioCreative VI chemical-protein interaction Track.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A., LEVY O., STOYANOV V. & ZETTLEMOYER L. (2020). BART : Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 7871–7880, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.703.

LIU H., ZHANG X., FAN L., FU X., LI Q., WU X.-M. & LAM A. Y. (2019a). Reconstructing Capsule Networks for Zero-shot Intent Classification. 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)*, p. 4799–4809, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1486.

LIU P., YUAN W., FU J., JIANG Z., HAYASHI H. & NEUBIG G. (2021). *Pre-train, Prompt, and Predict : A Systematic Survey of Prompting Methods in Natural Language Processing*. Rapport interne. arXiv :2107.13586 [cs] type : article.

LIU T. & LOW B. K. H. (2023). *Goat : Fine-tuned LLaMA Outperforms GPT-4 on Arithmetic Tasks*. Rapport interne. arXiv :2305.14201 [cs] type : article.

LIU Y., OTT M., GOYAL N., DU J., JOSHI M., CHEN D., LEVY O., LEWIS M., ZETTLEMOYER L. & STOYANOV V. (2019b). Roberta : A robustly optimized bert pretraining approach.

LONGPRE S., HOU L., VU T., WEBSON A., CHUNG H. W., TAY Y., ZHOU D., LE Q. V., ZOPH B., WEI J. & ROBERTS A. (2023). The flan collection : Designing data and methods for effective instruction tuning.

LOSHCHILOV I. & HUTTER F. (2019). *Decoupled Weight Decay Regularization*. Rapport interne. arXiv :1711.05101 [cs, math] version : 3 type : article.

LU J., ZHU D., HAN W., ZHAO R., MAC NAMEE B. & TAN F. (2023). What makes pre-trained language models better zero-shot learners ? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 2288–2303, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.128.

LUDAN J. M., MENG Y., NGUYEN T., SHAH S., LYU Q., APIDIANAKI M. & CALLISON-BURCH C. (2023). *Explanation-based Finetuning Makes Models More Robust to Spurious Cues*. Rapport interne. arXiv :2305.04990 [cs] version : 2 type : article.

MAAS A. L., DALY R. E., PHAM P. T., HUANG D., NG A. Y. & POTTS C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics : Human Language Technologies - Volume 1*, HLT '11, p. 142–150, USA : Association for Computational Linguistics.

MALO P., SINHA A., KORHONEN P., WALLENIUS J. & TAKALA P. (2014). Good debt or bad debt : Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, **65**.

MENG Y., ZHANG Y., HUANG J., XIONG C., JI H., ZHANG C. & HAN J. (2020). Text Classification Using Label Names Only : A Language Model Self-Training Approach. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, p. 9006–9017, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.emnlp-main.724.

MOLLAS I., CHRYSOPOULOU Z., KARLOS S. & TSOUMAKAS G. (2022). ETHOS : an Online Hate Speech Detection Dataset. *Complex & Intelligent Systems*, **8**(6), 4663–4678. arXiv :2006.08328 [cs, stat], DOI : 10.1007/s40747-021-00608-2.

MOSQUERA A. (2022). Tackling Data Drift with Adversarial Validation : An Application for German Text Complexity Estimation. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, p. 39–44, Potsdam, Germany : Association for Computational Linguistics.

MUENNIGHOFF N., WANG T., SUTAWIKA L., ROBERTS A., BIDERMAN S., SCAO T. L., BARI M. S., SHEN S., YONG Z.-X., SCHOELKOPF H., TANG X., RADEV D., AJI A. F., ALMUBARAK K., ALBANIE S., ALYAFEAI Z., WEBSON A., RAFF E. & RAFFEL C. (2023). *Crosslingual Generalization through Multitask Finetuning*. Rapport interne. arXiv :2211.01786 [cs] type : article, DOI : 10.48550/arXiv.2211.01786.

NITYASYA M. N., WIBOWO H. A., PRASOJO R. E. & AJI A. F. (2021). *Costs to Consider in Adopting NLP for Your Business*. Rapport interne. arXiv :2012.08958 [cs] type : article.

OPENAI (2023). *GPT-4 Technical Report*. Rapport interne. arXiv :2303.08774 [cs] type : article, DOI : 10.48550/arXiv.2303.08774.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C. L., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A., SCHULMAN J., HILTON J., KELTON F., MILLER L., SIMENS M., ASKELL A., WELINDER P., CHRISTIANO P., LEIKE J. & LOWE R. (2022). *Training language models to follow instructions with human feedback*. Rapport interne. arXiv :2203.02155 [cs] type : article.

PENEDO G., MALARTIC Q., HESSLOW D., COJOCARU R., CAPPELLI A., ALOBEIDLI H., PANNIER B., ALMAZROUEI E. & LAUNAY J. (2023). *The RefinedWeb Dataset for Falcon LLM : Outperforming Curated Corpora with Web Data, and Web Data Only*. Rapport interne. arXiv :2306.01116 [cs] type : article, DOI : 10.48550/arXiv.2306.01116.

PENG B., ALCAIDE E., ANTHONY Q., ALBALAK A., ARCADINHO S., CAO H., CHENG X., CHUNG M., GRELLA M., GV K. K., HE X., HOU H., KAZIENKO P., KOCON J., KONG J., KOPTYRA B., LAU H., MANTRI K. S. I., MOM F., SAITO A., TANG X., WANG B., WIND J. S., WOZNIAK S., ZHANG R., ZHANG Z., ZHAO Q., ZHOU P., ZHU J. & ZHU R.-J. (2023). Rwkv : Reinventing rnns for the transformer era.

PILÁN I. & VOLODINA E. (2018). Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, p. 49–58, Santa Fe, New-Mexico : Association for Computational Linguistics.

QIN C., ZHANG A., ZHANG Z., CHEN J., YASUNAGA M. & YANG D. (2023). *Is ChatGPT a General-Purpose Natural Language Processing Task Solver ?* Rapport interne. arXiv :2302.06476 [cs] type : article.

RADFORD A., WU J., CHILD R., LUAN D., AMODEI D. & SUTSKEVER I. (2019). Language models are unsupervised multitask learners.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S., MATENA M., ZHOU Y., LI W. & LIU P. J. (2020). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer*. Rapport interne. arXiv :1910.10683 [cs, stat] type : article.

SANH V., WEBSON A., RAFFEL C., BACH S. H., SUTAWIKA L., ALYAFEAI Z., CHAFFIN A., STIEGLER A., SCAO T. L., RAJA A., DEY M., BARI M. S., XU C., THAKKER U., SHARMA S. S., SZCZECHLA E., KIM T., CHHABLANI G., NAYAK N., DATTA D., CHANG J., JIANG M. T.-J., WANG H., MANICA M., SHEN S., YONG Z. X., PANDEY H., BAWDEN R., WANG T., NEERAJ T., ROZEN J., SHARMA A., SANTILLI A., FEVRY T., FRIES J. A., TEEHAN R., BERS T., BIDERMAN S., GAO L., WOLF T. & RUSH A. M. (2022). *Multitask Prompted Training Enables Zero-Shot Task Generalization*. Rapport interne. arXiv :2110.08207 [cs] type : article.

SCHICK T. & SCHÜTZE H. (2021). *Exploiting Cloze Questions for Few Shot Text Classification and Natural Language Inference*. Rapport interne. arXiv :2001.07676 [cs] type : article.

SMITH R., FRIES J. A., HANCOCK B. & BACH S. H. (2022). *Language Models in the Loop : Incorporating Prompting into Weak Supervision*. Rapport interne. arXiv :2205.02318 [cs] type : article.

SOCHER R., PERELYGIN A., WU J., CHUANG J., MANNING C. D., NG A. & POTTS C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, p. 1631–1642, Seattle, Washington, USA : Association for Computational Linguistics.

SUN Y., DONG L., HUANG S., MA S., XIA Y., XUE J., WANG J. & WEI F. (2023). Retentive network : A successor to transformer for large language models.

TAY Y., DEHGHANI M., TRAN V. Q., GARCIA X., WEI J., WANG X., CHUNG H. W., SHAKERI S., BAHRI D., SCHUSTER T., ZHENG H. S., ZHOU D., HOULSBY N. & METZLER D. (2023). *UL2 : Unifying Language Learning Paradigms*. Rapport interne. arXiv :2205.05131 [cs] type : article.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F., RODRIGUEZ A., JOULIN A., GRAVE E. & LAMPLE G. (2023). Llama : Open and efficient foundation language models.

WANG T., ROBERTS A., HESSLOW D., SCAO T. L., CHUNG H. W., BELTAGY I., LAUNAY J. & RAFFEL C. (2022). *What Language Model Architecture and Pretraining Objective Work Best for Zero-Shot Generalization ?* Rapport interne. arXiv :2204.05832 [cs, stat] type : article.

WANG Y., YU Z., ZENG Z., YANG L., WANG C., CHEN H., JIANG C., XIE R., WANG J., XIE X., YE W., ZHANG S. & ZHANG Y. (2023). *PandaLM : An Automatic Evaluation Benchmark for LLM Instruction Tuning Optimization*. Rapport interne. arXiv :2306.05087 [cs] type : article.

WEI J., BOSMA M., ZHAO V., GUU K., YU A. W., LESTER B., DU N., DAI A. M. & LE Q. V. (2022). Finetuned Language Models are Zero-Shot Learners.

WU M., WAHEED A., ZHANG C., ABDUL-MAGEED M. & AJI A. F. (2023). Lamini-lm : A diverse herd of distilled models from large-scale instructions. *CoRR*, **abs/2304.14402**.

XIA C., ZHANG C., YAN X., CHANG Y. & YU P. S. (2018). *Zero-shot User Intent Detection via Capsule Neural Networks*. Rapport interne. arXiv :1809.00385 [cs] type : article, DOI : 10.48550/arXiv.1809.00385.

YEH H.-S., LAVERGNE T. & ZWEIGENBAUM P. (2022). Decorate the Examples : A Simple Method of Prompt Design for Biomedical Relation Extraction. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, p. 3780–3787, Marseille, France : European Language Resources Association.

ZHANG J., YU Y., LI Y., WANG Y., YANG Y., YANG M. & RATNER A. J. (2021). WRENCH : A Comprehensive Benchmark for Weak Supervision. *ArXiv*.

ZHANG S., DONG L., LI X., ZHANG S., SUN X., WANG S., LI J., HU R., ZHANG T., WU F. & WANG G. (2023). *Instruction Tuning for Large Language Models : A Survey*. Rapport interne. arXiv :2308.10792 [cs] type : article.

ZHAO W. X., ZHOU K., LI J., TANG T., WANG X., HOU Y., MIN Y., ZHANG B., ZHANG J., DONG Z., DU Y., YANG C., CHEN Y., CHEN Z., JIANG J., REN R., LI Y., TANG X., LIU Z., LIU P., NIE J.-Y. & WEN J.-R. (2023a). *A Survey of Large Language Models*. Rapport interne. arXiv :2303.18223 [cs] type : article.

ZHAO X., OUYANG S., YU Z., WU M. & LI L. (2023b). Pre-trained language models can be fully zero-shot learners. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 15590–15606, Toronto, Canada : Association for Computational Linguistics. DOI : 10.18653/v1/2023.acl-long.869.

A Models

Model	Number of Parameters	Instruction-Tunned
bigscience/bloom (**?**)	560M, 1B1, 1B7, 3B, 7B1	No
bigscience/bloomz (Muennighoff *et al.*, 2023)	560M, 1B1, 1B7, 3B, 7B1	Yes
tiiuae/falcon	7B, 40B	Yes/No
tiiuae/falcon-rw	7B, 40B	No
MBZUAI/LaMini-Cerebras (Wu *et al.*, 2023)	111M, 256M, 590M, 1.3B	Yes
MBZUAI/LaMini-GPT (Wu *et al.*, 2023)	124M, 774M, 1.5B	Yes
mosaicml/mpt	7B 30b	Yes/No
databricks/dolly-v2	3b, 7B, 12b	Yes
EleutherAI/pythia (Biderman *et al.*, 2023)	70M, 160M, 410M, 1B, 1.4B, 2.8, 6.9B, 12B	No
openlm-research/open_llama	3B 7B 13B	No
openlm-research/open_llama_v2	3B 7B	No
pankajmathur/orca_dolly	3B	Yes
pankajmathur/orca_alpaca	3B	Yes
pankajmathur/orca_mini	7B, 3B, 13B	Yes
pankajmathur/orca_mini_v2	7B, 13B	Yes
pankajmathur/orca_mini_v3	7B, 13B	Yes

TABLE 5 – Decoder Only Models

Model	Number of Parameters	Instruction-Tunned
MBZUAI/LaMini-Flan-T5 (Wu *et al.*, 2023)	77M, 248M, 783M	Yes
T5 vanilla (Raffel *et al.*, 2020)	77M, 248M, 770M, 3B, 11B	No
bigscience/mt0 (Muennighoff *et al.*, 2023)	300M, 582, 1.2B, 3.8B, 13B	Yes
Bart (Lewis *et al.*, 2020)	255M, 561M	No

TABLE 6 – Encoder-Decoder Only Models

B Datasets

Datasets	Tasks	#Classes	#Test Examples	Balance ratios
AGNews	Topic Classification	4	12000	0.897
BBCNews	Topic Classification	5	2000	0.742
CDR bio	Relation Classification	2	4673	0.478
Chemprot	Chemical Relation Classification	10	1607	0.004
ETHOS	Sentiment Classification	2	998	0.766
financial_phrasebank	Topic Classification	3	2264	0.218
IMDB	Sentiment Classification	2	2500	1.000
SemEval	Relation Classification	9	600	0.042
SMS	Spam Classification	2	500	0.155
Spouse	Relation Classification	2	2701	0.088
SST2	Sentiment Classification	2	1821	0.997
SST5	Sentiment Classification	5	2210	0.441
TREC	Question Classification	6	500	0.065
Yelp	Sentiment Classification	2	3800	0.915
Youtube	Spam Classification	2	250	0.894

TABLE 7 – Descriptions des jeux de données

C Prompts

TABLE 8 – Prompt used

dataset	prompts
sms	`Is the following message spam? Answer by yes or no.\n"TEXT"`
youtube	`Is the following comment spam? Answer by yes or no.\n"TEXT"`
spouse	`Context: "TEXT"\n\nAre ENTITY2 and ENTITY1 married? Answer by yes or`
cdr	`Context: "TEXT"\n\nDoes ENTITY1 induce ENTITY2 ? Answer by yes or no.`
chemprot	`Context: "TEXT"\n\nWhat is the relation between ENTITY1 and ENTITY2 ?`
semeval	`Context: "TEXT"\n\nWhat is the relation between ENTITY1 and ENTITY2 ?`
sst-2	`"TEXT" has a tone that is`
sst-5	`"TEXT" has a tone that is`
yelp	`"TEXT" has a tone that is`
imdb	`"TEXT" has a tone that is`
ethos	`"TEXT" has a tone that is`
financial_phrasebank	`"TEXT" has a tone that is`
trec	`"TEXT" is about`
agnews	`"TEXT" is about`
bbcnews	`"TEXT" is about`

TABLE 9 – Description des amorces et verbalizers