Évaluation de grands modèles de langue pour la classification de concepts et la génération de descriptions dans les études aréales

Xinyi Shen, Damien Nouvel, Peter Stockinger

Évaluation de grands modèles de langue pour la classification de concepts et la génération de descriptions dans les études aréales

Xinyi Shen[1] Damien Nouvel[2] Peter Stockinger
(1) LaCAS, 2 rue de Lille, 75007, Paris, F
(2) ERTIM, 2 rue de Lille, 75007, Paris, F
prenom.nom@inalco.fr

RÉSUMÉ

Récemment, les grands modèles de langue (LLM) ont connu un grand succès dans les tâches de TAL, mais demandent encore beaucoup de travaux pour bien identifier leurs capacités et leurs limitations. Dans ce travail, nous avons évalué l'efficacité des LLM dans le domaine scientifique, plus précisément dans les sciences humaines et les études aréales. À travers l'analyse de quatre modèles de langue (GPT, LLAMA3, , BERT, Mistral), pour deux tâches spécifiques, la classification et la génération de textes, nous avons découvert que GPT se distingue par sa capacité à produire des textes académiques cohérents mais souvent généraux. En matière de classification multi-label, GPT surpasse également LLAMA3 et Mistral, démontrant une meilleure adéquation avec les évaluations humaines et une plus grande cohérence dans les résultats.

ABSTRACT

EEvaluation of Large Language Models for Concept Classification and Description Generation in Areal Studies

Recently, large language models (LLMs) have seen great success in NLP tasks, but still require extensive work to properly identify their capabilities and limitations. In this study, we evaluated the effectiveness of LLMs in the scientific domain, specifically in the humanities and areal studies. Through the analysis of four language models (GPT, LLAMA3, BERT, Mistral), for two specific tasks, classification and text generation, we found that GPT stands out for its ability to produce coherent academic texts but often general. In multi-label classification, GPT also surpasses LLAMA3 and Mistral, demonstrating better alignment with human evaluations and greater consistency in results.

MOTS-CLÉS : Grand modèle de langue, Classification multi-label, génération automatique de texte.

KEYWORDS: Large Language Model, multi-label classification, automatic text generation.

1 Introduction

Ces dernières années, les grands modèles de langue (LLM) , basés sur des architectures de transformeurs (Vaswani *et al.*, 2017), ont profondément révolutionné le domaine du traitement automatique des langues (TAL). Ces modèles, tels que ChatGPT(Achiam *et al.*, 2023), ont démontré leur potentiel non seulement pour contribuer à répondre à des problèmes généraux tels que la recherche d'informations ou la génération de codes informatiques, mais aussi pour améliorer les procédures de veille, de classification ou de comparaison d'informations scientifiques.

Pour mieux comprendre les avantages et les faiblesses des LLM, un nombre croissant de chercheurs se consacrent à leur évaluation systématique. Les objectifs et les méthodes d'évaluation des divers LLM sont variés. Certains travaux d'évaluation se contentent d'un dispositif simple de questions-réponses (Kamalloo *et al.*, 2023; Sun *et al.*, 2023, 2024), d'autres travaux d'évaluation s'intéressent davantage à la génération des texte (Akkasi *et al.*, 2023; Karpinska *et al.*, 2021). Un troisième groupe d'évaluation se concentre sur la capacité des LLM à contribuer à la production de codes informatiques (Yuan *et al.*, 2023; Liu *et al.*, 2023). Enfin, pour terminer ce survol synthétique et non-exhaustif, citons encore un dernier groupe de travaux d'évaluation s'intéressant à la détection et à l'analyse de sentiments (He *et al.*, 2024; Buscemi & Proverbio, 2024). Ensemble, les différents projets et travaux d'évaluation permettent de mieux cerner les performances et les limites de ces technologies prometteuses que représentent les LLM.

Nous remarquons cependant que les évaluations actuelles des LLM reposent sur des jeux de données qui ont été construits pour des tâches spécifiques, alors que les fonds et corpus de données qui font partie des diverses archives numériques et/ou plateformes de valorisation ont des objectifs beaucoup plus larges, moins circonscrits et pouvant évoluer dans le temps et à travers des usages sociaux multiples. Lesdits fonds et corpus se distinguent également par le fait de réunir presque obligatoirement des données et des métadonnées représentant une certaine hétérogénéité - hétérogénéité de format, certes, mais aussi hétérogénéité en termes de qualité éditoriale, de précision sémantique, d'orientation pragmatique, etc.

Nous nous sommes fixés pour objectif de tester les avantages d'une utilisation plus systématique des LLM pour la recherche en sciences humaines et sociales. Pour ce faire, nous avons sélectionné un corpus de données et de métadonnées représentant une certaine hétérogénéité sémantique (en termes de contenu) et pragmatique (en termes d'utilisation), telle qu'on la rencontre dans les entrepôts, archives et plateformes de recherche (HAL et Nakala en France, Zenodo, OpenAire, etc.).

Le domaine scientifique choisi est celui des études aréales (*area studies* ou *regional studies*, en anglais). Il s'agit de recherches pluridisciplinaires consacrées à une aire géographique, socioculturelle ou historique. Une partie de ces recherches pluridisciplinaires se trouve décrite et documentée sur la plateforme *LaCAS - Open Archive in Language and Cultural Area Studies* [1], une plateforme initiée et portée par l'Institut National des Langues et Civilisations Orientales (Inalco), en partenariat avec l'Université Paris Cité, l'Institut National de l'Audiovisuel (INA) et la Région Île de France qui fonctionne avec le logiciel *Okapi* (*Open Knowledge Annotation and Publishing Interface*) (Beloued *et al.*, 2015, 2017).

Dans le cadre des diverses activités de création d'alignement et de documentation de concepts, d'indexation et d'annotation de ressources documentaires ou encore de publication de dossiers traitant une problématique particulière en études aréales, les LLM sont utilisés régulièrement pour automatiser nos différentes tâches.

Pour tester la qualité des résultats générés par les LLM dans le cadre de ces divers usages, nous nous sommes intéressés plus particulièrement aux deux tâches suivantes :la tâche de la génération de présentations structurées d'un concept du LaCAS et la tâche de la classification multi-lables des ressources documentaires et de termes eux-mêmes.

1. https://lacas.inalco.fr/

2 Aperçu de travaux connexes

2.1 Classification de ressources documentaires à l'aide de LLM

Dans le domaine de la classification, il est à noter que l'encodeur est généralement privilégié par rapport au décodeur dans les tâches impliquant des LLM. Ces modèles ont également été évalués pour diverses tâches circonscites. Peña *et al.* (2023) se concentrent sur la classification dans le domaine des affaires publiques en espagnol. Les résultats montrent qu'en combinant le classifieurs SVM, LLM (RoBERTa et GPT-2) peuvent améliorer l'exactitude de la classification. En ce qui concerne la classification dans le domaine de santé, Yang *et al.* (2023) a réalisé une évaluation de LLM sur la classification multi-classes en utilisant LLaMA, InstructGPT-3 et ChatGPT en testant diverses stratégies de spécification et de formulation des prompts. Cependant, les performances de ces LLM sur ce type de tâches restent généralement inférieures à celles obtenues avec les modèles de BERT.

Dans le domaine des articles scientifiques, certaines recherches, telles que celles de Xu *et al.* (2023) a utilisé des articles scientifiques pour des tâches de classification multi-étiquettes en utilisant des corpus en anglais. Il montrent que les LLM peuvent améliorer significativement la précision et l'efficacité de la classification, surtout dans le cas de manque des informations. Par ailleurs, Zhang *et al.* (2023) a affiné un modèle en utilisant des articles médicaux et informatiques pour réaliser des tâches de classification multi-étiquettes en montrant également que LLM peuvent augmenter la performance de classification en cas de manque des informations des étiquettes.

2.2 Génération des textes à l'aide de LLM

Dans le domaine de la génération de textes, il est pertinent de mentionner que c'est principalement la fonction de décodeur des LLM qui est utilisée. En ce qui concerne l'évaluation des LLM pour la génération de textes, plusieurs approches ont été adoptées. A titre d'exemple, mentionnons les approches qui s'intéressent à évaluer la qualité de descriptions ou de narrations générées et celles qui sont davantage concernées par la capacité d'un LLM à résumer des ressources documentaires. Par exemple, Tian *et al.* (2024) a évalué les compétences des LLM pour résumer des textes dans les domaines de la biomédecine et de la santé. Ils montrent que, même si ces modèles peuvent produire des résumés fluides et fidèles, incluant la gestion adéquate de termes professionnels complexes, ils présentent toujours des risques de lacunes informatives et d'inexactitudes. Gao *et al.* (2024) ont analysé la capacité des modèles GPT à produire des articles introductifs dans le style de Wikipédia à partir de différents prompts. Les résultats de cette recherche indiquent, lorsqu'ils sont guidés et qu'ils ont accès à des informations externes, peuvent rivaliser avec les humains dans la rédaction d'articles introductifs à caractère scientifique ; cependant, il est essentiel de mentionner que l'utilisation de LLM largement entraînés sur des données similaires, comme Wikipedia, nécessite une analyse critique de la contamination des données pour garantir l'intégrité des informations générées. Shen *et al.* (2023) ont exploré l'utilisation des LLM comme support à la rédaction scientifique. Ils ont découvert que, bien qu'il y ait de bonnes performances en compréhension de documents et en production de texte, il existe encore des limites concernant le développement de structures argumentatives et la capacité à créer des textes longs. Liang *et al.* (2023) ont évalué les performances des LLM en les soumettant à la critique d'articles scientifiques. Leur étude révèle que les résultats obtenus par ces modèles sont remarquablement proches de ceux des experts humains. Ces études illustrent les diverses applications et les défis rencontrés par les LLM dans la génération de textes scientifiques.

3 Classification de ressources documentaires

Une des principales tâches de la plateforme LaCAS est de lier les productions scientifiques aux concepts du thésaurus LaCAS qui couvre les grands domaines de connaissance abordés dans les études aréales. Cette tâche se présente sous forme d'une classification multi-label.

3.1 Jeux de données

Les annotations manuelles dans le TAL sont très coûteuses en raison du temps et de l'expertise. Impossible donc, pour une structure académique de procéder à l'indexation systématique de quelques 100.000 données scientifiques.

Actuellement, LaCAS contient 7000 concepts et 88000 dépôts textuels (complétés d'environ 52.000 données visuelles et 2500 données audiovisuelles et sonores), bien que les annotations ne soient pas complètes (la première version d'annotation humaine). D'une part, les données sont déséquilibrées : parmi les 7000 concepts, seulement 1043 sont documentés par plus de 50 ressources documentaires liées, 479 ont plus de 100 ressources documentaires liées, 281 ont plus de 150 documents liés, 133 ont plus de 200 documents liés. D'autre part, les annotations n'ont pas été suffisamment contrôlées, quelque 130.000 ressources documentaires multimédias ne sont pas toutes classées et les annotations existantes n'ont pas toujours été vérifiées par des experts.

Nous avons sélectionné aléatoirement 10 concepts du thésaurus parmi ceux qui disposent de plus de 200 ressources associées. Ensemble, les dix concepts sélectionnés sont représentés dans un corpus comprenant 1835 ressources documentaires. Les concepts sélectionnés sont : *Études indiennes* (*Indologie*), *Études japonaises*, *Études mexicaines*, *Études russes*, *Anthropologie*, *Linguistique*, *Architecture*, *Archéologie*, *Littératures du monde* et *Études cinématographiques*.

Pour évaluer les résultats de manière rigoureuse, nous avons divisé ces données en un ensemble d'entraînement et un ensemble de test, avec 367 ressources documentaires (soit 20 % du total) constituant les données de test. Pour éviter de dépasser le nombre de tokens limité pour LLM, nous n'avons pris en compte que le titre des articles. Nous avons ensuite ré-annoté manuellement les données de test (la deuxième version d'annotation humaine) pour garantir la qualité des annotations, étant donné que les annotations manuelles n'étaient pas toujours cohérentes et complètes.

3.2 Méthodologies et modèles

Nous avons évalué la performance de classification automatique de quatre différents LLM. Nous avons utilisé des expressions régulières pour obtenir les résultats de classification suite à la sortie des LLM. Chaque modèle possède des spécificités en termes de taille et de capacité. **BERT** (Devlin *et al.*, 2018) est le modèle de référence pour nos expérimentations. Pour nos expérimentations, nous avons utilisé la version bert-large-uncased qui comprend 24 couches et 336 millions de paramètres. Nous avons affiné BERT en ajoutant une couche de classification, en utilisant la première version de l'annotation. Le deuxième modèle est **gpt-4-0125-preview** qui est un LLM de question-réponse avancé basé sur l'architecture GPT-4 (Achiam *et al.*, 2023). Il est continuellement entraîné avec un mélange de *fine-tuning* supervisé et d'apprentissage par renforcement avec retour basé sur des commentaires humains (RLHF). La taille de contextes est 128, 000. Notre troisième modèle **Meta-**

Llama-3-8B (Touvron *et al.*, 2023) qui est développé par Meta AI, est un modèle ajusté par instruction et affiné pour des cas d'utilisation en dialogue. Il existe en deux tailles : 8 milliards et 70 milliards de paramètres. Nous avons adopté celle de 8 milliards de paramètres qui a une taille de contexte de 8 000 tokens. Le dernier modèle **Mistral-7B-Instruct-v0.2**(Jiang *et al.*, 2023) est également un modèle ajusté par instruction. Il a une taille de contexte de 32 000 tokens.

Pour comparer les résultats générés par les machines et les humains, nous avons adopté Alpha de Krippendorff pour mesurer d'accords.

3.3 Résultats

Nous avons mené trois séries d'expériences à zéro coup ("exemple", *shot* en anglais), à un coup et à deux coups. Les prompts utilisés pour ces expériences se trouvent également en annexe. Pour les expériences à un coup et à deux coups, nous avons utilisé des exemples de classes avec des F-mesures moyennement faibles par rapport aux annotations humaines. Pour le prompt à un coup, nous avons fourni un exemple pour le concept *Anthropologie*, pour le prompt à deux coups, nous avons fourni des exemples pour les concepts *Anthropologie* et *Études cinématographiques*.

Nous avons d'abord comparé les résultats en utilisant les F-mesures entre les résultats générés par les LLM et les annotations humaines. Ensuite, nous avons approfondi notre analyse en examinant l'indicateur *Alpha de Krippendorff* pour évaluer le degré de concordance. Les résultats de la F-mesure dans le tableau 3.3.1, nous mettons les matrices de confusion en annexe. Ceux de l'Alpha de Krippendorff sont présentés dans le tableau 3.3.2. Dans ces tableaux, H1 représente la première version d'annotation humaine, H2 représente la deuxième version d'annotation humaine, B représente le modèle BERT, G représente le modèle GPT, L représente le modèle LLAMA, M représente le modèle Mistral. Selon le modèle, le numéro représente le nombre de coups (0, 1 ou 2).

3.3.1 Résultats de F-mesure

	B	G0	G1	G2	L0	L1	L2	M0	M1	M2
Anthropologie	**0,48**	0,34	0,37	0,39	0,41	0,36	0,25	0,41	0,40	0,46
Architecture	**0,74**	0,68	0,68	0,65	0,61	0,67	0,60	0,67	0,67	0,63
Archéologie	**0,57**	0,47	0,48	0,52	0,35	0,41	0,38	0,35	0,35	0,35
Études cinématographiques	0,85	**0,91**	0,87	**0,91**	0,34	0,36	0,39	0,05	0,00	0,05
Linguistique	0,58	0,66	**0,67**	0,63	0,60	0,52	0,43	0,64	0,57	0,53
Littératures du monde	0,77	0,81	0,82	**0,83**	0,54	0,67	0,45	0,61	0,64	0,68
Études indiennes (Indologie)	0,72	0,76	**0,79**	**0,79**	0,31	0,45	0,47	0,40	0,47	0,42
Études japonaises	0,84	**0,89**	**0,89**	0,88	0,66	0,76	0,66	0,80	0,71	0,74
Études mexicaines	**0,87**	0,70	0,68	0,70	0,60	0,58	0,46	0,43	0,17	0,30
Études russes	0,76	**0,81**	**0,81**	0,78	0,65	0,69	0,67	0,73	0,70	0,64
Moyenne pondérée	**0,71**	0,70	**0,71**	**0,71**	0,51	0,55	0,47	0,52	0,48	0,49

TABLE 1 – Tableau des résultats de F-Mesure qui utilise la deuxième version des annotations humaine comme référence

Les F-mesures utilisent la deuxième version des annotations humaines comme référence (*gold standard* en anglais). Nous avons remarqué que les F-mesures les plus élevés sont systématiquement obtenus avec les résultats de BERT et de GPT pour zéro, un et deux coups. De plus, la moyenne pondérée de ces quatre résultats est presque identique.

Lorsque nous analysons les F-mesures de chaque modèle, nous constatons que pour les expériences avec les modèles GPT et Mistral, les F-mesures ne varient pas significativement entre zéro, un, ou deux coups. En revanche, pour les expériences avec les modèles LLAMA, par rapport au mode 0 coup, la F-mesure de moyenne pondérée augmente de 0,49 à 0,54. Cependant, le score de LLAMA en mode deux coups diminue.

En examinant de plus près les deux concepts *Anthropologie* et *Études cinématographiques* que nous avons utilisés comme exemples dans nos prompts, nous faisons trois obervation. Lorsque nous analysons des moyennes pondérées de chaque modèle, nous constatons que pour les expériences avec les modèles GPT et Mistral, les moyennes pondérées ne varient pas significativement entre zéro, un ou deux coups. En revanche, en comparant les modèles LLAMA au mode zéro coup, la moyenne pondérée du mode un coup augmente de 0,51 à 0,55. Cependant, le score de LLAMA en mode deux coups diminue. Enfin, nous remarquons que, pour toutes les expériences, lorsque nous ajoutons davantage d'exemples, les précisions augmentent, mais les rappels diminuent.

Pour les modèles GPT en mode zéro coup, la F-mesure d' *Anthropologie* est la plus basse (0,34), tandis que celle d' *Études cinématographiques* est la plus élevée (0,91). La matrice de confusion en annexe montre que *Anthropologie* en mode zéro coup présente beaucoup de confusion en relation avec *Linguistique* et *Études indiennes (Indologie)*. En mode un et deux coups de GPT, nous observons que la F-mesure pour *Anthropologie* augmente, tandis que les résultats pour *Linguistique* et *Études indiennes (Indologie)* diminuent. la table de confusion en annexe montre que la précision (vrais positifs) pour *Anthropologie* ne varie pas beaucoup (36 en mode zéro coup, 36 en mode un coup, et 35 en mode deux coups). En mode deux coups, bien que la précision pour *Études cinématographiques* reste élevée, son score baisse par rapport au mode zéro coup.

Pour les modèles LLAMA en mode zéro coup, les F-mesures pour *Anthropologie* et *Études cinématographiques* sont respectivement de 0,41 et 0,34, ce qui n'est pas très satisfaisant. En mode zéro coup, pour le concept *Anthropologie*, il y a plus de FN et FP sur le concept *Études indiennes (Indologie)* (35) que de VP (29). En mode un coup, la F-mesure d' *Anthropologie* diminue, bien que la précision augmente. En mode deux coups, la F-mesure et la précision pour *Anthropologie* diminuent, tandis que la F-mesure et la précision pour *Études cinématographiques* augmentent. En modes un coup et deux coups, les FN et FP sur *Études indiennes (Indologie)* diminuent, mais les VP diminuent également.

Pour Mistral en mode zéro coup, la F-mesure pour *Anthropologie* est de 0,41, tandis que celle pour *Études cinématographiques* n'est que de 0,05. En mode un coup, la précision pour *Anthropologie* augmente légèrement, mais le rappel diminue, entraînant une légère baisse de la F-mesure. En mode deux coups, la précision pour *Anthropologie* augmente, le rappel diminue, mais la F-mesure augmente. En revanche, la précision pour *Études cinématographiques* reste faible et la F-mesure demeure à 0,05. La matrice de confusion pour ce modèle montre des tendances similaires à celles observées avec le modèle LLAMA.

3.3.2 Résultats d'Alpha de Krippendorff

Le meilleur résultat d'Alpha de Krippendorff a été observé entre GPT en mode un coup (G1) et deux coups (G2, 0,84), suivi de GPT en mode zéro coup (G0) et un coup (0,83). La première version d'annotation humaine (H1) et BERT (B) a obtenu un score de 0,76. GPT en mode zéro coup et deux coups (0,75). Tous ces scores sont supérieurs ou égales à 0,75. À titre de comparaison, le score des annotations entre les humains (H1 et H2) est de seulement 0,63, tandis que la deuxième version

	H1	H2	B	G0	G1	G2	L0	L1	L2	M0	M1	M2
H1	1,00	0,63	**0,76**	0,50	0,54	0,58	0,33	0,37	0,29	0,30	0,29	0,32
H2		1,00	0,61	0,46	0,50	0,54	0,33	0,35	0,28	0,32	0,27	0,28
B			1,00	0,48	0,53	0,57	0,34	0,38	0,28	0,31	0,31	0,32
G0				1,00	**0,83**	**0,75**	0,36	0,31	0,32	0,36	0,41	0,44
G1					1,00	**0,84**	0,37	0,34	0,34	0,38	0,42	0,44
G2						1,00	0,37	0,36	0,35	0,39	0,41	0,47
L0							1,00	0,41	0,36	0,40	0,35	0,34
L1								1,00	0,45	0,33	0,33	0,31
L2									1,00	0,32	0,33	0,31
M0										1,00	0,51	0,46
M1											1,00	0,53
M2												1,00

TABLE 2 – Tableau des valeurs de Krippendorff's Alpha

d'annotation (H2) et BERT a obtenu un score de 0,61.

Les scores compris entre 0,5 et 0,6 incluent les résultats de comparaison entre GPT et les annotations humaines, de comparaison entre GPT et BERT, ainsi que les comparaisons avec Mistral. Les comparaisons entre les modèles, c'est-à-dire les scores entre GPT et Mistral, et entre Mistral et LLAMA, sont moins élevées. Les comparaisons entre annotations humaines et Mistral ou LLAMA n'ont pas non plus donné de bons résultats.

En fournissant des exemples spécifiques pour les concepts les moins bien traités par les prompts en zéro coup permet d'améliorer les performances progressivement, davantage en accord avec les annotations humaines. Si nous comparons les résultats de la deuxième version d'annotation avec GPT, le score d'Alpha de Krippendorff pour GPT en mode zéro coup est de 0,46, en mode un coup est de 0,50, et en mode deux coups est de 0,54. Cependant, les résultats de LLAMA en mode deux coups sont surprenants : si l'on compare avec la deuxième version d'annotation, LLAMA en mode deux coups obtient un score de 0,28, inférieur à celui de LLAMA en mode zéro coup (0,33) et en mode un coup (0,35).

Nous avons également pu remarquer que les résultats de GPT sont très cohérents entre eux. Les scores entre les trois modes de GPT sont parmi les meilleurs, suggérant une forte cohérence intra-modèle.

En conclusion, les modèles GPT ont atteint des performances équivalentes à celles du modèle BERT en classification, ce qui illustre l'exceptionnelle capacité de GPT, étant donné que BERT est déjà considéré comme très performant dans ce domaine. L'ajout de coups supplémentaires tend à améliorer la précision des modèles et diminue légèrement les erreurs de confusion pour certains concepts, mais peut aussi entraîner une diminution du rappel. Les performances varient selon les modèles de langue et les classes spécifiques, indiquant que l'efficacité des modèles de langue peut être optimisée en fonction des types de données et du nombre de coups utilisés. Les modèles GPT montrent une cohérence notable, tandis que LLAMA et Mistral ont des performances moins stables.

4 Génération de présentations structurées de concept

La deuxième tâche centrale pour enrichir le contenu de LaCAS est celle de produire des présentations structurées des concepts. Pour réaliser cette tâche, nous avons donc sollicité un groupe de chercheurs pour évaluer la qualité des présentations textuelles produits par Scholar AI qui est propulsé par GPT4. Pour classer et analyser les retours critiques des chercheurs, l'outil psychométrique appellé *échelle de Likert* (à 5 points) a été utilisé.

4.1 Prompts

L'objectif de cette expérience était de vérifier si le ChatGPT (GPT-4) pouvait générer une description d'un concept de LaCAS avec une crédibilité élevée. En d'autres mots, nous avons cherché à comprendre, si ChatGPT (GPT-4) est en mesure de saisir les détails d'un domaine de connaissance et s'il est capable de localiser et d'exploiter des références scientifiques pertinentes pour générer une présentation. Les prompts utilisés dans le cadre de cette expérience sont en annexe.

4.2 Jeux de données

Nous avons sélectionné 102 concepts qui font partie du thésaurus LaCAS et ont demandé à Scholar AI de générer une description pour chaque concept. Les concepts sont divisés en deux ensembles : concepts généraux et concepts spécifiques. Le modèle de description prévoit trois paragraphes : le premier est réservé à la définition d'un concept ; le deuxième est réservé à la production d'une petite liste d'exemples illustrant le contenu du premier paragraphe ; le troisième comprend la liste des références scientifiques ainsi que les liens permettant de les consulter en ligne. Les descriptions sont limitées à 250 mots. Pour chaque chercheur, un jeu de cinq descriptions de concepts correspondant à son domaine de recherche spécifique a été préparé. Parmi les 102 descriptions automatiquement générées, 41 descriptions ont été évaluées, avec un total de 48 réponses. Afin de réduire tout biais éventuel dans le processus d'évaluation, le questionnaire adressé aux chercheurs précisait que l'une des cinq descriptions générées était le travail d'un spécialiste humain.

4.3 Méthode d'évaluation par les humains

Les chercheurs ont été invité à évaluer des descriptions générées par Scholar AI selon les six critères suivants (Howcroft *et al.*, 2020) :

1. *Pertinence du contenu* Ce critère évalue dans quelle mesure le contenu de la description est adapté au contexte spécifié. Il détermine si la description est adaptée pour pouvoir servir comme une introduction générale fiable à la compréhension d'un concept.
2. *Pertinence lexicale* Ce critère évalue si le choix du registre lexico-terminologique est approprié dans le contexte donné. Ce critère examine si les mots utilisés sont adaptés à une introduction générale du terme, assurant que le texte est accessible et compréhensible pour le public ciblé.
3. *Pertinence discursive* Ce critère évalue si la description est développée de manière structurée, logique et cohérente.
4. *Ressemblance humaine* Si la description est générée par la machine, ce critère mesure si la description produite par la machine est semblable à celle qu'un humain aurait écrite.
5. *Précision / généralité* Ce critère évalue si la description fournie par la machine est trop générale pour le domaine spécifié. Il examine si la description reste pertinente et spécifique lorsqu'elle traite de sujets particuliers. Ce critère interroge notamment si la quantité de détails spécifiques au domaine est adéquate et si la description évite d'être trop vague ou interchangeable.
6. *Neutralité* Ce critère évalue si la description se présente sans biais scientifique, politique, éthique ou autres.

Chaque critère est mesuré sur une échelle de Likert à 5 points où 1 est le score le plus bas et 5 est le score le plus haut.

Critère / Catégorie	Tous les concepts	concepts spécifiques	concepts généraux
Pertinence du contenu	3.67	3.68	3.65
Pertinence lexicale	3.71	3.86	3.58
Pertinence discursive	4.02	4.13	3.92
Ressemblance humaine	3.91	4.00	3.85
Précision / généralité	3.37	3.41	3.35
Neutralité	3.98	3.78	4.15
Tous	3.78	3.81	3.75

TABLE 3 – Résultat d'évaluation de génération des textes

4.4 Résultats

Les résultats sont présentés dans le Tableau 4.3. Pour chaque critère, nous avons calculé les moyennes des scores pour l'ensemble des concepts, ainsi que séparément pour les concepts spécifiques et les concepts généraux. Ensuite, nous avons déterminé la moyenne des scores globaux pour tous les concepts, ainsi que pour les concepts spécifiques et généraux, pour chaque critère. Nous avons également calculé les écarts-types pour chaque critique et pour chaque catégorie de concepts ((i) Tous les Concepts ; (ii) Concepts spécifiques ; (iii)Concepts généraux).

En examinant les résultats, nous avons constaté que le critère de Pertinence (Construction discursive) obtient le meilleur score avec une moyenne de 4,02, suivi de près par la Ressemblance humaine avec un score de 3,91. En revanche, le critère de Précision/Généralité affiche le score le plus bas à 3,37, suivi par la Pertinence (Contenu) avec un score de 3,67. Les variations des écarts-types se situent entre 0,73 et 1,2, ce qui est considéré comme raisonnablement stable.

Il est intéressant de noter que, à l'exception du critère de Neutralité, les scores des concepts spécifiques sont plus élevés que ceux des concepts généraux. Cela pourrait indiquer qu'il est plus facile pour GPT-4 de définir des concepts très précis. De plus, la performance élevée en Pertinence (Construction discursive) peut être attribuée aux prompts détaillant clairement comment structurer les trois paragraphes. La ressemblance humaine, avec un score élevé, montre également la capacité de GPT4 à produire des textes qui imitent bien les écrits humains.

Nous avons également reçu plusieurs commentaires des chercheurs. Ces retours sont majoritairement des critiques négatives, pointant des problèmes tels que : (1) Des descriptions trop générales (définitions ou exemples). (2) Des textes peu courants. (3) Des définitions ou exemples insuffisants, ainsi que des références bibliographiques obsolètes ou non pertinentes. Une structure de paragraphe parfois étrange. (4) Un usage de vocabulaire non professionnel. (5) Une traduction de l'anglais au français trop évidente. (6) Des erreurs dans les descriptions.

Bien que ces commentaires ne puissent pas tous être catégorisés selon nos critères, le problème le plus souvent soulevé concerne la Précision/Généralité, reflétant les points 1, 2 et 3. De plus, des problèmes ont été relevés dans les catégories de Pertinence (Contenu), notamment les erreurs dans les descriptions, Pertinence (Choix lexical) avec des critiques sur le vocabulaire et la traduction, et Pertinence (Construction discursive) en raison de structures de paragraphe inappropriées.

De plus, lors de l'expérience de classification, nous avons constaté que la classe *Anthropologie* pose un problème pour les modèles GPT, LLAMA et Mistral. Nous souhaitions également évaluer la performance de ces modèles sur des sous-domaines spécifiques de l'Anthropologie. Pour ce faire, nous avons sélectionné des concepts pertinents tels que : *Anthropologie cognitive*, *Anthropologie culturelle*, *Anthropologie sociale* et *Anthropologie - Madagascar*, pour un total de six réponses. Les résultats sont présentés dans le 4.4. Nous avons observé que ces concepts produisent des performances

	PC	PL	PD	RH	P/G	N
Anthropologie cognitive	2	3	2	4	3	2
Anthropologie culturelle	2	3	3	3	3	3
Anthropologie culturelle	3	3	4	3	3	5
Anthropologie sociale	3	3	3	3	3	3
Anthropologie sociale	3	2	3	3	3	4
Anthropologie – Madagascar	3	2	3	3	3	3
Total	2,67	2,67	3	3,17	3	3,17

TABLE 4 – Résultat d'évaluation de générations des textes sur les concepts *Anthropologie* (PC : Pertinence du Contenu, PL : Pertinence Lexicale, PD Pertinence Discursive, RH : Ressemblance Humaine, P/G : Précision / généralité, N : Neutralité)

inférieures par rapport aux résultats précédents.

5 Conclusion

Cet article porte sur l'évaluation des contributions éventuelles des LLM pour la recherche en sciences humaines et, plus précisément, dans les études aréales. Deux types d'évaluations ont été réalisés : la classification multi-label des articles et la génération de textes pour présenter des concepts qui représentent un domaine ou une partie d'un domaine de connaissance.

Dans l'expérience de classification, nous avons constaté que le modèle GPT est plus aligné sur les jugements humains comparativement à LLAMA et Mistral. De plus, GPT offre des réponses plus stables, montrant une cohérence interne supérieure aux deux autres modèles. Nous avons également remarqué que l'ajout d'exemples dans le prompt, en particulier ceux provenant de domaines moins performants, peut améliorer la précision des réponses sans nécessairement affecter les performances globales. Notre prochaine étape consiste à intégrer des expériences utilisant des modèles fine-tunés pour une évaluation plus approfondie.

Nous avons constaté que le modèle GPT affiche de bonnes performances dans ces deux tâches. Dans l'expérience de génération de textes, nous avons observé que GPT est capable de produire des textes académiques de qualité tout à fait satisfaisante et convaincante. Avec un prompt bien détaillé, GPT peut structurer les paragraphes de manière cohérente. Cependant, les textes générés tendent à être trop généraux, surtout lorsque les concepts fournis sont peu précis. Le manque de données supplémentaires pour cette expérience est regrettable, mais nous prévoyons de continuer à enrichir notre base de données. Nous allons également continuer à conduire des nouvelles expériences pour évaluer la génération de textes scientifiques, notamment dans les études aréales.

Dans l'ensemble, ces travaux montrent qu'il est possible d'évaluer des modèles de langue pour les tâches considérées, mais qu'elle requiert de mettre en place une méthodologie solide et coûteuse, fondée sur des annotations humaines de qualité apportées par des experts. Les résultats quantitatifs montrent une légère supériorité de BERT et GPT pour la classification et la domination des modèles GPT pour la génération de textes. Les évaluations qualitatives des experts en sciences humaines indiquent que, pour ces tâches relativement bien prises en charge par ces modèles, leur fiabilité semble être au rendez-vous, même si certains concepts en sciences humaines, comme l'*Anthropologie*, sont moins bien générés par le modèle GPT et demandent des vérifications humaines avant d'être publiées.

Remerciements

Nous souhaitons remercier chaleureusement les personnes suivantes pour avoir répondu à notre questionnaire et enrichi notre étude de leurs précieuses perspectives : Arnaud Arslangul, Joelle Dalegre, Georges Kostakiotis, Bénédicte Parvaz, Frosa Pejoska, Miharitiana Rakotonirina, Rola Skaf, Assen Slim, Sibylle Pouillaude, Mathieu Valette, et Julien Vercueil. Leur participation a été essentielle à la réussite de cette recherche.

Références

ACHIAM J., ADLER S., AGARWAL S., AHMAD L., AKKAYA I., ALEMAN F. L., ALMEIDA D., ALTENSCHMIDT J., ALTMAN S., ANADKAT S. *et al.* (2023). Gpt-4 technical report. *arXiv preprint arXiv :2303.08774*.

AKKASI A., FRASER K. C. & KOMEILI M. (2023). Reference-free summarization evaluation with large language models. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, p. 193–201.

BELOUED A., LALANDE S. & STOCKINGER P. (2015). Modélisation et formalisation rdfs/owl d'une ontologie de description audiovisuelle. *Les Cahiers du numérique*, (3), 39–70.

BELOUED A., LALANDE S. & STOCKINGER P. (2017). Studio campus aar. une plateforme sémantique pour l'analyse et la publication de corpus audiovisuels. *Intelligence collective et archives numériques*, p. 85–108.

BUSCEMI A. & PROVERBIO D. (2024). Chatgpt vs gemini vs llama on multilingual sentiment analysis. *arXiv preprint arXiv :2402.01715*.

DEVLIN J., CHANG M.-W., LEE K. & TOUTANOVA K. (2018). Bert : Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv :1810.04805*.

GAO F., JIANG H., YANG R., ZENG Q., LU J., BLUM M., LIU D., SHE T., JIANG Y. & LI I. (2024). Evaluating large language models on wikipedia-style survey generation.

HE L., OMRANIAN S., MCROY S. & ZHENG K. (2024). Using large language models for sentiment analysis of health-related social media data : empirical evaluation and practical tips. *medRxiv*, p. 2024–03.

HOWCROFT D. M., BELZ A., CLINCIU M., GKATZIA D., HASAN S. A., MAHAMOOD S., MILLE S., VAN MILTENBURG E., SANTHANAM S. & RIESER V. (2020). Twenty years of confusion in human evaluation : Nlg needs evaluation sheets and standardised definitions. In *13th International Conference on Natural Language Generation 2020*, p. 169–182 : Association for Computational Linguistics.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S., CASAS D. D. L., BRESSAND F., LENGYEL G., LAMPLE G., SAULNIER L. *et al.* (2023). Mistral 7b. *arXiv preprint arXiv :2310.06825*.

KAMALLOO E., DZIRI N., CLARKE C. L. & RAFIEI D. (2023). Evaluating open-domain question answering in the era of large language models. *arXiv preprint arXiv :2305.06984*.

KARPINSKA M., AKOURY N. & IYYER M. (2021). The perils of using mechanical turk to evaluate open-ended text generation. *arXiv preprint arXiv :2109.06835*.

LIANG W., ZHANG Y., CAO H., WANG B., DING D., YANG X., VODRAHALLI K., HE S., SMITH D., YIN Y. *et al.* (2023). Can large language models provide useful feedback on research papers ? a large-scale empirical analysis. *arXiv preprint arXiv :2310.01783*.

LIU J., XIA C. S., WANG Y. & ZHANG L. (2023). Is your code generated by chatgpt really correct. *Rigorous evaluation of large language models for code generation. CoRR, abs/2305.01210*.

PEÑA A., MORALES A., FIERREZ J., SERNA I., ORTEGA-GARCIA J., PUENTE I., CORDOVA J. & CORDOVA G. (2023). Leveraging large language models for topic classification in the domain of public affairs. In *International Conference on Document Analysis and Recognition*, p. 20–33 : Springer.

SHEN Z., AUGUST T., SIANGLIULUE P., LO K., BRAGG J., HAMMERBACHER J., DOWNEY D., CHANG J. C. & SONTAG D. (2023). Beyond summarization : Designing ai support for real-world expository writing tasks. *arXiv preprint arXiv :2304.02623*.

SUN J., SHAIB C. & WALLACE B. C. (2023). Evaluating the zero-shot robustness of instruction-tuned language models. *arXiv preprint arXiv :2306.11270*.

SUN L., HAN Y., ZHAO Z., MA D., SHEN Z., CHEN B., CHEN L. & YU K. (2024). Scieval : A multi-level large language model evaluation benchmark for scientific research. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, p. 19053–19061.

TIAN S., JIN Q., YEGANOVA L., LAI P.-T., ZHU Q., CHEN X., YANG Y., CHEN Q., KIM W., COMEAU D. C. *et al.* (2024). Opportunities and challenges for chatgpt and large language models in biomedicine and health. *Briefings in Bioinformatics*, **25**(1), bbad493.

TOUVRON H., LAVRIL T., IZACARD G., MARTINET X., LACHAUX M.-A., LACROIX T., ROZIÈRE B., GOYAL N., HAMBRO E., AZHAR F. *et al.* (2023). Llama : Open and efficient foundation language models. *arXiv preprint arXiv :2302.13971*.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł. & POLOSUKHIN I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

XU R., YU Y., HO J. & YANG C. (2023). Weakly-supervised scientific document classification via retrieval-augmented multi-stage training. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, p. 2501–2505.

YANG K., JI S., ZHANG T., XIE Q., KUANG Z. & ANANIADOU S. (2023). Towards interpretable mental health analysis with large language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

YUAN Z., LIU J., ZI Q., LIU M., PENG X. & LOU Y. (2023). Evaluating instruction-tuned large language models on code comprehension and generation. *arXiv preprint arXiv :2308.01240*.

ZHANG Y., JIN B., CHEN X., SHEN Y., ZHANG Y., MENG Y. & HAN J. (2023). Weakly supervised multi-label classification of full-text scientific papers. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, p. 3458–3469.

A Prompts

A.1 Prompts de classification

Prompt 0 coup : Classifiez un texte à une ou plusieurs catégories correspondantes.

Les catégories sont 'Études indiennes (Indologie)', 'Études japonaises', 'Études mexicaines' , 'Études russes', 'Anthropologie', 'Linguistique' , 'Architecture', 'Archéologie', 'Littératures du monde', 'Etudes cinématographiques'

Si le texte correspond uniquement à une catégorie, répondez directement le nom de cette catégorie

Si le texte correspond à plusieurs catégories, répondez directement avec les noms des catégories séparés par des ','

Le texte est

Prompt 1 coup : Classifiez un texte à une ou plusieurs catégories correspondantes.

Les catégories sont 'Études indiennes (Indologie)', 'Études japonaises', 'Études mexicaines' , 'Études russes', 'Anthropologie', 'Linguistique' , 'Architecture', 'Archéologie', 'Littératures du monde', 'Etudes cinématographiques'

Si le texte correspond uniquement à une catégorie, répondez directement le nom de cette catégorie

Si le texte correspond à plusieurs catégories, répondez directement avec les noms des catégories séparés par des ','

Par exemple : Centre de recherche berbère (CRB - équipe du LACNAD, EA 4092)Le Centre de recherche berbère (CRB) est un pôle de recherche en linguistique descriptive, historique et appliquée berbère, en littérature berbère et anthropologie culturelle du monde berbère, s'appuyant sur la seule et la plus ancienne structure universitaire d'enseignement complète du berbère existant en France et en Europe (du premier au troisième cycle). Fondé en 1992 à l'Inalco en tant qu'équipe d'accueil (EA 3577), le CRB intègre en 2006 le laboratoire Langues et Cultures du Nord de l'Afrique et Diasporas (LACNAD, EA 4092). Ses activités s'organisent autour de trois axes : linguistique berbère, littérature berbère, culture et société berbères.

Réponse : Anthropologie

Le texte est

Prompt deux coups : Classifiez un texte à une ou plusieurs catégories correspondantes.

Les catégories sont 'Études indiennes (Indologie)', 'Études japonaises', 'Études mexicaines' , 'Études russes', 'Anthropologie', 'Linguistique' , 'Architecture', 'Archéologie', 'Littératures du monde', 'Etudes cinématographiques'

Si le texte correspond uniquement à une catégorie, répondez directement le nom de cette catégorie

Si le texte correspond à plusieurs catégories, répondez directement avec les noms des catégories séparés par des ','

Par exemple :

Interactions, transferts, ruptures artistiques et culturels (InTRu ; EA 6301)Au-delà de ces convergences méthodologiques, ils ont en commun un certain nombre d'objets thématiques. L'un de ces pôles thématiques concerne la construction des cultures visuelles contemporaines, entre peinture, photographie, cinéma, graphisme, design et image imprimée. Un autre de ces pôles concerne l'histoire de l'architecture et de l'urbanisme, la représentation des espaces habités, les liens entre l'action publique et les enjeux politiques et sociaux. Entre ces deux pôles circule enfin un intérêt commun pour les enjeux critiques et les stratégies d'émancipation qui permettent de comprendre et de déjouer

les formes de la domination spatiale, culturelle et visuelle. | L'InTRu est une équipe de recherches de l'université de Tours (Équipe d'accueil n° 6301). L'acronyme « InTRu » renvoie aux termes « Interactions, Transferts, Ruptures artistiques et culturelles ».Le laboratoire InTRu réunit des chercheurs et des chercheuses issues de l'histoire de l'art et de l'architecture, de la littérature, la philosophie, l'esthétique de la bande dessinée, l'histoire de la photographie, du cinéma, du design.

Réponse : Etudes cinématographiques

Par exemple : Centre de recherche berbère (CRB - équipe du LACNAD, EA 4092)Le Centre de recherche berbère (CRB) est un pôle de recherche en linguistique descriptive, historique et appliquée berbère, en littérature berbère et anthropologie culturelle du monde berbère, s'appuyant sur la seule et la plus ancienne structure universitaire d'enseignement complète du berbère existant en France et en Europe (du premier au troisième cycle). Fondé en 1992 à l'Inalco en tant qu'équipe d'accueil (EA 3577), le CRB intègre en 2006 le laboratoire Langues et Cultures du Nord de l'Afrique et Diasporas (LACNAD, EA 4092). Ses activités s'organisent autour de trois axes : linguistique berbère, littérature berbère, culture et société berbères.

Réponse : Anthropologie

Le texte est

A.2 Prompt de génération des textes

Je vais vous demander de me fournir une présentation de termes ou de concepts que je vous soumettrai par la suite, un par un, en les mettant entre crochets (exemple : [Terme]). Voici les instructions pour rédiger la présentation.

« Rédaction de la présentation d'un [Terme] en trois paragraphes.

1) Organisation globale de la présentation en cinq paragraphes :

1.1) Le premier paragraphe est réservé à une définition et/ou une description synthétique de [Terme].

1.2) Le deuxième paragraphe est réservé à deux ou trois exemples saillants qui illustrent la définition ou la description du [Terme] fournie dans le premier paragraphe.

1.3) Le troisième paragraphe est réservé à l'énumération de cinq à dix références bibliographiques scientifiques.

2) Rédaction du premier paragraphe réservé à une définition/description du [Terme]

2.1) Ce premier paragraphe est précédé de l'intitulé : Résumé :

2.2) La définition/description doit se baser sur des références scientifiques (bibliographiques) clairement identifiées.

2.3) La référence ou les références bibliographiques doivent être rédigées comme suit : (Auteur(s). Année. Titre court ; Auteur(s). Année. Titre court ; ...).

2.4) Les références doivent prioritairement provenir de fonds scientifiques internationaux reconnus tels que Cairn, arXiv, Hal, JSTOR, CiteSeer, Isidore, OpenAire, Zenodo, OCLC, BNF, ORCID, ...

2.5) Utilisez exclusivement des références bibliographiques vérifiables et des liens qui fonctionnent réellement !

3) Sélection et présentation des exemples dans le deuxième paragraphe de la présentation

3.1) Ce deuxième paragraphe est précédé de l'intitulé : Exemples :

3.2) Les exemples cités dans le deuxième paragraphe de la présentation doivent provenir des références scientifiques énumérées dans le troisième paragraphe.

3.2) Pour chaque exemple, il faut indiquer la référence (bibliographique) où on peut le trouver.

3.3) La référence bibliographique doit être rédigée comme suit : (Auteur(s). Année. Titre court).

3.4) Les références doivent prioritairement provenir de fonds scientifiques internationaux reconnus tels que Cairn, arXiv, Hal, JSTOR, CiteSeer, Isidore, OpenAire, Zenodo, OCLC, BNF, ORCID, ...

3.5) Utilisez exclusivement des références bibliographiques vérifiables et des liens qui fonctionnent réellement !

4) Sélection et présentation des références dans le troisième paragraphe :

4.1) Ce paragraphe est précédé de l'intitulé : Références scientifiques :

4.2) La référence bibliographique doit être construite comme suit : Auteur(s). Année. Intitulé complet. Revue ou Ouvrage (si applicable). Éditeur.

4.3) Merci de supprimer les liens vers votre base documentaire et de fournir uniquement les informations bibliographiques décrites en 5.2 permettant d'identifier correctement une référence.

4.4) La bibliographie doit être rédigée par ordre alphabétique en prenant en compte le premier auteur composant une référence.

4.5) Merci de faire attention à un certain équilibre géographique et linguistique dans le choix des références (essayez de ne pas seulement fournir des références « occidentales »).

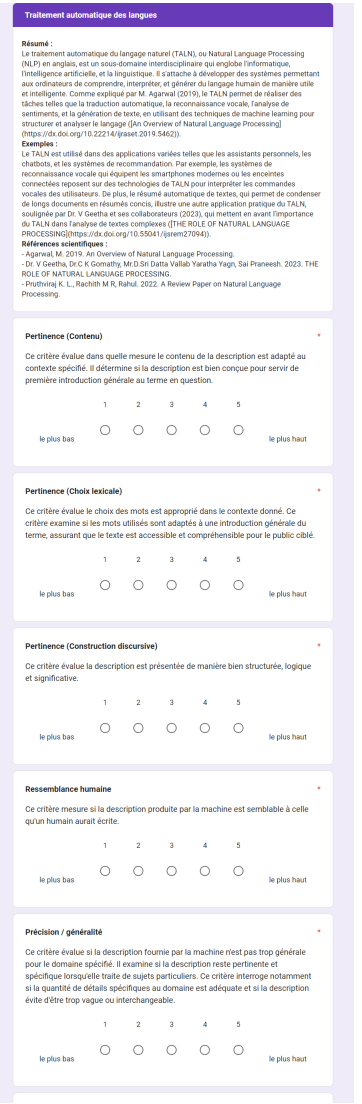
5) Instructions générales

5.1) Les trois paragraphes doivent être courts.

5.2) Le langage utilisé doit être neutre. Merci de ne pas utiliser d'expressions ou de tournures de phrases qui pourraient suggérer des positions morales, esthétiques, idéologiques ou politiques.

5.3) Le public à qui cette présentation est adressée est un public académique (universitaire). Vous pouvez donc utiliser un registre spécialisé de langage (style universitaire, terminologie spécialisée, formules mathématiques, expressions symboliques, ...). »

B Exemple de questionnaire



Traitement automatique des langues

Résumé :
Le traitement automatique du langage naturel (TALN), ou Natural Language Processing (NLP) en anglais, est un sous-domaine interdisciplinaire qui englobe l'informatique, l'intelligence artificielle, et la linguistique. Il s'attache à développer des systèmes permettant aux ordinateurs de comprendre, interpréter, et générer du langage humain de manière utile et intelligente. Comme expliqué par M. Agarwal (2019), le TALN permet de réaliser des tâches telles que la traduction automatique, la reconnaissance vocale, l'analyse de sentiments, et la génération de texte, en utilisant des techniques de machine learning pour structurer et analyser le langage ([An Overview of Natural Language Processing](https://dx.doi.org/10.22214/ijraset.2019.5462)).
Exemples :
Le TALN est utilisé dans des applications variées telles que les assistants personnels, les chatbots, et les systèmes de recommandation. Par exemple, les systèmes de reconnaissance vocale qui équipent les smartphones modernes ou les enceintes connectées reposent sur des technologies de TALN pour interpréter les commandes vocales des utilisateurs. De plus, le résumé automatique de textes, qui permet de condenser de longs documents en résumés concis, illustre une autre application pratique du TALN, soulignée par Dr. V Geetha et ses collaborateurs (2023), qui mettent en avant l'importance du TALN dans l'analyse de textes complexes ([THE ROLE OF NATURAL LANGUAGE PROCESSING](https://dx.doi.org/10.55041/ijsrem27094)).
Références scientifiques :
- Agarwal, M. 2019. An Overview of Natural Language Processing.
- Dr. V Geetha, Dr.C K Gomathy, Mr.D.Sri Datta Vallab Yaratha Yagn, Sai Praneesh. 2023. THE ROLE OF NATURAL LANGUAGE PROCESSING.
- Pruthviraj K. L., Rachith M R, Rahul. 2022. A Review Paper on Natural Language Processing.

Pertinence (Contenu) *

Ce critère évalue dans quelle mesure le contenu de la description est adapté au contexte spécifié. Il détermine si la description est bien conçue pour servir de première introduction générale au terme en question.

1 2 3 4 5
le plus bas — le plus haut

Pertinence (Choix lexicale) *

Ce critère évalue le choix des mots est approprié dans le contexte donné. Ce critère examine si les mots utilisés sont adaptés à une introduction générale du terme, assurant que le texte est accessible et compréhensible pour le public ciblé.

1 2 3 4 5
le plus bas — le plus haut

Pertinence (Construction discursive) *

Ce critère évalue la description est présentée de manière bien structurée, logique et significative.

1 2 3 4 5
le plus bas — le plus haut

Ressemblance humaine *

Ce critère mesure si la description produite par la machine est semblable à celle qu'un humain aurait écrite.

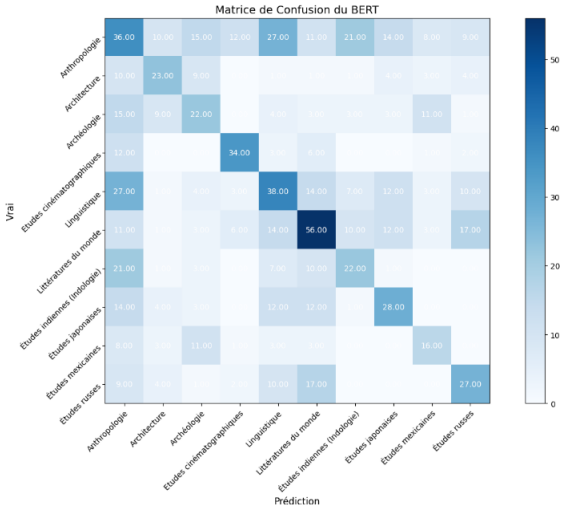
1 2 3 4 5
le plus bas — le plus haut

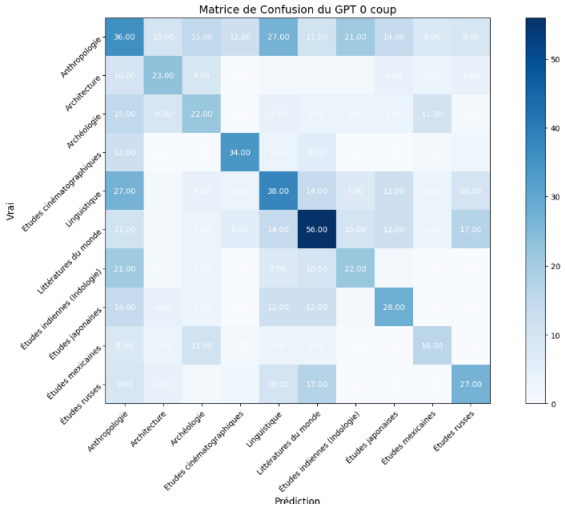
Précision / généralité *

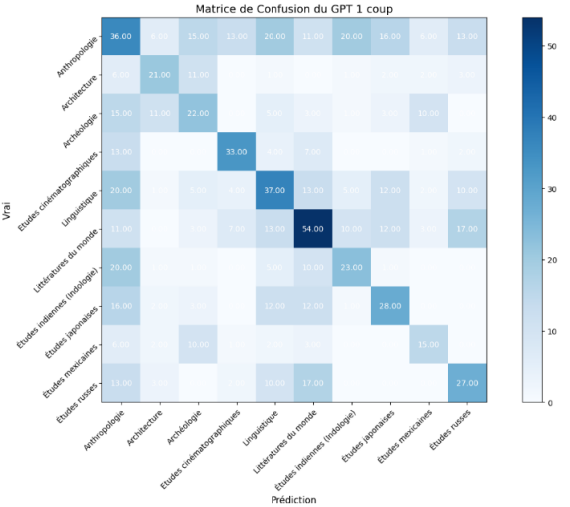
Ce critère évalue si la description fournie par la machine n'est pas trop générale pour le domaine spécifié. Il examine si la description reste pertinente et spécifique lorsqu'elle traite de sujets particuliers. Ce critère interroge notamment si la quantité de détails spécifiques au domaine est adéquate et si la description évite d'être trop vague ou interchangeable.

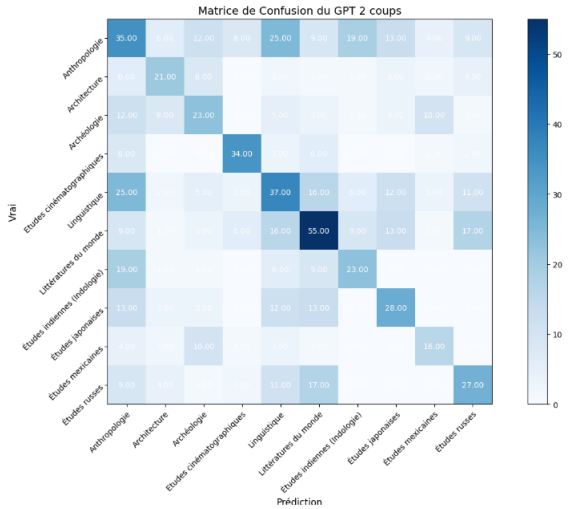
1 2 3 4 5
le plus bas — le plus haut

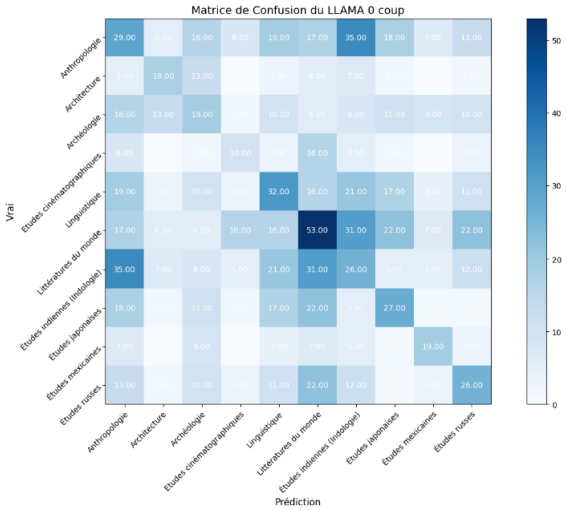
C Résultats de matrice de confusion

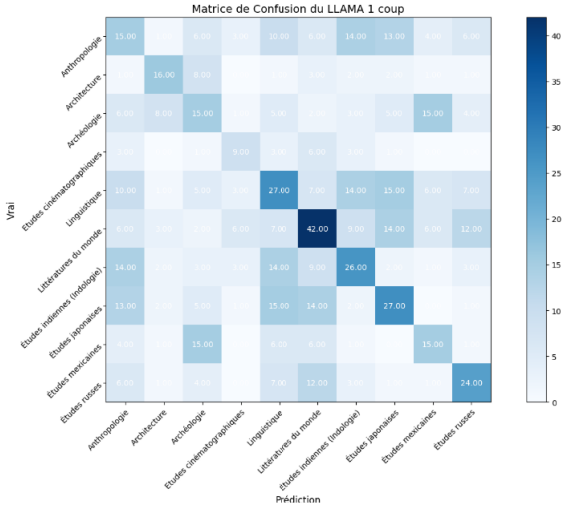

Matrice de Confusion du BERT
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes

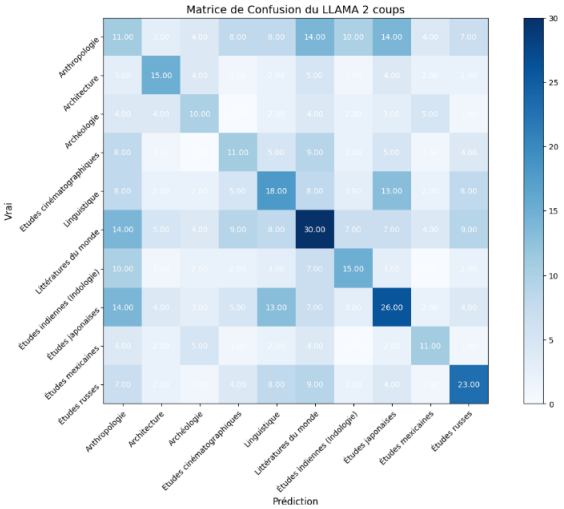

Matrice de Confusion du GPT 0 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes

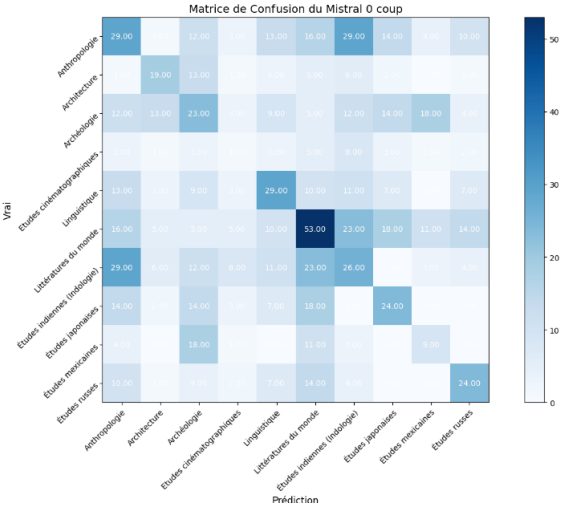

Matrice de Confusion du GPT 1 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Études cinématographiques
Linguistique
Littératures du monde
Études indiennes (Indologie)
Études japonaises
Études mexicaines
Études russes
36.00
21.00
22.00
33.00
37.00
54.00
23.00
28.00
15.00
27.00
0
10
20
30
40
50

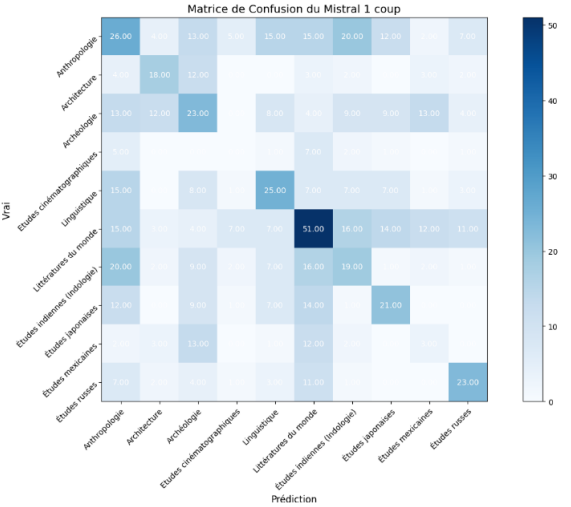

Matrice de Confusion du GPT 2 coups
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Études cinématographiques
Linguistique
Littératures du monde
Études indiennes (Indologie)
Études japonaises
Études mexicaines
Études russes
35.00
21.00
23.00
34.00
37.00
55.00
23.00
28.00
16.00
27.00
0
10
20
30
40
50

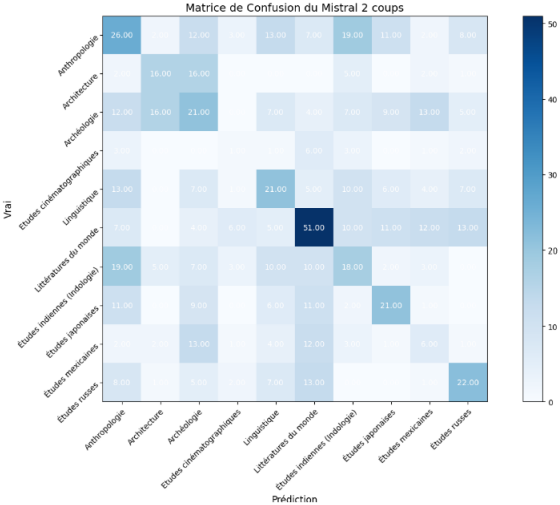

Matrice de Confusion du LLAMA 0 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes


Matrice de Confusion du LLAMA 1 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes


Matrice de Confusion du LLAMA 2 coups
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes


Matrice de Confusion du Mistral 0 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Etudes indiennes (indologie)
Etudes japonaises
Etudes mexicaines
Etudes russes


Matrice de Confusion du Mistral 1 coup
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Études indiennes (Indologie)
Études japonaises
Études mexicaines
Études russes


Matrice de Confusion du Mistral 2 coups
Vrai
Prédiction
Anthropologie
Architecture
Archéologie
Etudes cinématographiques
Linguistique
Littératures du monde
Études indiennes (Indologie)
Études japonaises
Études mexicaines
Études russes