Actes de CORIA-TALN 2026
Nantes, France, 29 juin au 3 juillet 2026

Une étude empirique de la capacité de génération de paraphrases des *(S/M)LMs*

Quentin Lemesle[1] Jonathan Chevelu [1] Arnaud Delhay [1] Damien Lolive [2]
(1) Univ Rennes, CNRS, IRISA, EXPRESSION, 22300 Lannion, France
(2) Univ Bretagne Sud, CNRS, IRISA, ARCHIMEDIA, 56000 Vannes, France
{prénom}.{nom}@irisa.fr

RÉSUMÉ

Les grands modèles de langue (*LLM*s) sont aujourd'hui largement utilisés pour des tâches d'augmentation de données, notamment via la génération de paraphrases. Cependant, ces modèles présentent un coût de calcul important et un impact environnemental non négligeable. Dans cet article, nous menons une étude empirique de la capacité de génération de paraphrases de différents modèles de langue de petites et moyennes tailles (*(S/M)LM*s), dans un cadre contrôlé. Nos résultats montrent que ces modèles sont capables de générer, sans exemples préalables, des paraphrases présentant une diversité lexicale significative.

ABSTRACT

An Empirical Study of Paraphrase Generation with (S/M)LMs

Large Language Models (LLMs) are now widely used for data augmentation tasks, particularly with paraphrase generation. However, these models entail substantial computational costs and have a non-negligible environmental impact. In this paper, we conduct an empirical study of the paraphrase generation capabilities of several Small and Medium-sized Language Models ((S/M)LMs) within a controlled setting. Our results show that these models are able to produce paraphrases with significant lexical diversity in zero-shot.

MOTS-CLÉS : génération, paraphrase, modèle de langue, corpus.

KEYWORDS: generation, paraphrase, language model, corpus.

1 Introduction

La génération automatique de paraphrases constitue une tâche importante en traitement automatique des langues (TAL) et est au cœur de nombreuses applications (Kumar *et al.*, 2019; Gao *et al.*, 2020; Dai *et al.*, 2023). Une des applications les plus courantes est l'augmentation artificielle de jeux de données, ce qui a permis des améliorations de performances dans de nombreuses tâches telles que les questions réponses (Dong *et al.*, 2017), la traduction automatique (Mehdizadeh Seraj *et al.*, 2015; Yang *et al.*, 2019), la simplification de texte (Maddela *et al.*, 2021) et la révision de texte (Raheja *et al.*, 2024).

Des travaux récents ont exploré les capacités de reformulation des *LLM*s (Piedboeuf & Langlais, 2023; Chataigner *et al.*, 2025; Meier *et al.*, 2025; Vahtola *et al.*, 2025) pour la génération de paraphrases. Cependant, ces travaux utilisent des stratégies de génération par échantillonnage, et se concentrent

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 1 : articles scientifiques originaux, pages 634–654.

principalement sur des modèles de très grande taille. Notre étude s'inscrit dans un contexte de sobriété : nous cherchons à utiliser des modèles sans les affiner et à minimiser autant que possible le budget de génération. C'est pourquoi nous nous focalisons sur une stratégie de génération gloutonne conditionnée par un terme de pénalité dans le but de créer des paraphrases éloignées lexicalement de la phrase source. Nous proposons une évaluation de **modèles de taille moyenne** (*MLMs*) alignés, capables de répondre à des questions, et de **modèles de petite taille** (*SLMs*), affinés pour la génération de paraphrases. Nos principaux résultats sont les suivants : (1) Fournir les sous-mots initiaux pour guider la génération améliore substantiellement le respect du schéma de sortie. (2) L'application de pénalités de répétition lors du décodage accroît de manière systématique la diversité lexicale. (3) L'amorçage avec des exemples, contrairement aux attentes, réduit la diversité des générations. Nous rendons le jeu de données résultant de nos expériences disponible sur demande.

Le reste de cet article est organisé comme suit : La section 2 présente les modèles, les stratégies d'amorçage, les paramètres de décodage considérés et les données que nous utilisons. La section 3 décrit notre protocole expérimental, et discute des résultats selon 3 axes d'évaluation.

2 Ressources expérimentales

Les *MLM*s alignés sont capables de gérer un dialogue avec un utilisateur et de répondre à ses questions. Grâce à cela, ils semblent capables de produire des paraphrases sans fournir d'exemple dans l'amorce de génération, et peuvent être guidés afin de respecter des contraintes stylistiques ou structurelles spécifiques. Nous adoptons une approche empirique pour comprendre comment la génération de paraphrases est influencée par le choix du modèle, la stratégie d'amorçage utilisée pour présenter la tâche, ainsi que les paramètres de décodage qui contrôlent la sélection des sous-mots. Les sous-sections suivantes détaillent ces différents aspects.

2.1 Modèles considérés

Nous évaluons la capacité de génération de paraphrases de 11 modèles, présentés dans la table 1, déjà entraînés, disponibles librement. Nous les regroupons en deux catégories : les modèles non alignés, reposant sur des architectures *Transformer* encodeur seul (Conneau *et al.*, 2020) ou encodeur/décodeur (Vaswani *et al.*, 2017; Raffel *et al.*, 2023) et les modèles alignés d'architecture *Transformer* décodeur seul (Brown *et al.*, 2020) ou *Mamba* (Gu & Dao, 2024). Cet ensemble varié nous permet de comparer des *SLM*s dédiés à la génération de paraphrases à des *MLM*s généralistes, et d'analyser l'impact de la taille du modèle et du paradigme d'entraînement sur la qualité des paraphrases.

Modèles non alignés : Ces modèles ont été affinés pour la génération de paraphrases. Ils sont légers (0,2 à 0,4 milliard de paramètres) et reposent sur un affinage direct plutôt que sur un alignement à grande échelle par instructions. Ces modèles génèrent une paraphrase à partir d'une phrase source, sans instruction explicite. Cependant, cette absence de gestion d'instructions dans l'amorce rend difficile l'imposition de contraintes lors de la génération.

Modèles alignés : Les *MLM*s affinés par instruction sont entraînés sur de vastes corpus dans le but d'imiter les réponses et les raisonnements humains sur un large éventail de tâches. Ils présentent de fortes performances avec ou sans exemples, et peuvent être guidés avec des amorces contenant des

1. https://huggingface.co/ramsrigouthamg/t5_paraphraser

Type	Modèle	Params
non aligné	Bart-paraphraser (Lewis *et al.*, 2019)	0.22
	T5-chatGPT-paraphraser (Vorobev & Kuznetsov, 2023)	0.22
	T5-paraphraser [1]	0.22
	Parrot (Damodaran, 2021)	0.41
aligné	Mamba (Gu & Dao, 2024)	2.77
	Llama2 (Touvron *et al.*, 2023)	6.74
	Mistralv0.2 (Jiang *et al.*, 2023)	7.24
	Mistralv0.3 (Jiang *et al.*, 2023)	7.25
	Falcon-mamba (Zuo *et al.*, 2024)	7.27
	Qwen2.5 (Bai *et al.*, 2023)	7.62
	Llama3 (Grattafiori *et al.*, 2024)	8.03

TABLE 1 – Modèles considérés répartis selon leur type d'affinage (Type). Params dénote le nombre de paramètres de chaque modèle en milliards.

instructions. En contrepartie, ils produisent souvent de longues explications (Poddar *et al.*, 2025; Zheng *et al.*, 2023b) ou des chaînes de raisonnement (Xu *et al.*, 2025), ce qui complique l'extraction automatique de la réponse attendue dans la génération. Nous évaluons des modèles allant de 3 à 8 milliards de paramètres.

2.2 Stratégies d'amorçage

Les performances des *MLM*s alignés sont fortement influencées par la manière dont les tâches leur sont présentées (White *et al.*, 2023), ce qui rend le choix de la stratégie d'amorçage crucial pour la génération de paraphrases. Nous étudions les paradigmes d'amorçage selon deux axes : le nombre et le type d'exemples fournis dans l'amorce, ainsi que le caractère guidé ou libre de la génération.

Les amorces peuvent ne contenir aucun, un ou plusieurs exemples de résolution de l'instruction. Ne pas fournir d'exemple permet d'évaluer la capacité du modèle à généraliser depuis son pré-entraînement et son affinage. Fournir quelques exemples permet d'illustrer le style souhaité, la variation lexicale ou les schémas structurels attendus pour les générations. Cet amorçage peut orienter le comportement spécifique à la tâche, mais il peut aussi réduire la diversité si les exemples sont répétitifs ou contraignants. Afin d'étudier plus finement la sensibilité des modèles, nous considérons deux types d'exemples : un ensemble standard et un ensemble contenant des expressions vulgaires. Étant donné que de nombreux modèles disponibles sont affinés pour éviter les sujets sensibles (Zheng *et al.*, 2023a), on peut se demander si l'ajout d'exemples vulgaires (Cegin *et al.*, 2024) dans l'amorce pourrait favoriser la réponse à ces sujets ou, au contraire, les restreindre davantage.

Pour contrôler la sortie du modèle, il est également possible de fournir les premiers sous-mots composant la génération attendue. Cela est utile pour respecter un format de sortie spécifique et faciliter l'extraction automatique de la réponse. Nous considérons deux stratégies d'amorçage : (1) **Amorce libre** : le modèle génère une paraphrase de manière libre. (2) **Amorce continue** : l'amorce contient les sous-mots du format cible pour guider la génération. Le modèle doit continuer la génération depuis l'amorce, comme si c'était lui qui avait généré les premiers sous-mots. Cette

approche forcera le respect des schémas prédéfinis nécessaires pour extraire automatiquement la paraphrase générée.

Nous utilisons des amorces avec 0, 1 et 4 exemples standards ou vulgaires ; chacune est utilisée en amorçage libre et continu. Cela représente un total de 10 configurations d'amorçage que nous nommons selon le nombre d'exemples, leur type et la stratégie d'amorçage. Par exemple, «1.V.C.» indique un amorçage continue avec un exemple vulgaire. Les exemples que nous utilisons sont disponibles en table 2 et la structure de notre amorce en annexe A.1.

		Source	Paraphrase
		Standard	
4	**1 {**	*The little cat refreshes himself with water every morning.*	*Every morning, the little cat refreshes himself by drinking water.*
		There is a big tree in my garden.	*A great tree is planted in my garden.*
		Some kids are more adventurous than others.	*Some children are less afraid of the unknown than others.*
		I love my previous car!	*My old automobile is still in my heart.*
		Vulgaire	
4	**1 {**	*I stepped on a piece of shit this morning.*	*I was out for a walk this morning when I trample on a poop.*
		Fuck both of you.	*Go fuck yourselves.*
		Don't listen to him, he's batshit crazy.	*Fucking ignore that wanker, he's fucking mental!*
		That fuckin' heap of junk ain't worth shit, it won't fuckin' run!	*That goddamn car is utter crap, it refuses to start.*

TABLE 2 – Exemples considérés pour compléter notre amorce de génération. **1{** désigne le couple de phrases utilisé dans la configuration avec un seul exemple.

2.3 Paramètres de décodage

Lors du processus de génération, les *(S/M)LM*s attribuent une probabilité d'apparition à chaque sous-mot de leur vocabulaire. Il existe de nombreuses stratégies de décodage pour sélectionner le sous-mot suivant à partir de cette distribution de probabilités (Shi *et al.*, 2024).

Les plus courantes sont la gloutonne, l'échantillonnage aléatoire et la recherche par faisceau. Le décodage glouton sélectionne le sous-mot ayant la probabilité la plus élevée. L'échantillonnage sélectionne aléatoirement un sous-mot pondéré par la distribution de probabilité. La recherche par faisceau explore plusieurs séquences candidates en parallèle, en considérant que des sous-mot moins probables peuvent aboutir à des générations globalement plus pertinentes. La recherche par faisceau peut être combinée avec des stratégies gloutonnes ou d'échantillonnage pour contrôler l'aléatoire de l'exploration (Vijayakumar *et al.*, 2018). Des mécanismes supplémentaires peuvent pénaliser certains sous-mots avant leur sélection, en imposant une pénalité de diversité, qui décourage les sous-mots déjà apparus dans d'autres faisceaux ; ou une interdiction de répétition qui empêche la sélection de n-grammes d'une taille fixe déjà présents.

Nous utilisons trois configurations de décodage, une configuration **sans pénalité** de diversité ou répétition, avec **pénalité légère** et avec **pénalité lourde**. Nous nous limitons à une stratégie de sélection gloutonne pour restreindre l'étude. Pour limiter le coût de calcul, nous fixons une longueur maximale de génération de 100 sous-mots ; la génération s'arrête si cette limite est atteinte. Les configurations de pénalité légère et lourde correspondent aux réglages fournis par les auteurs des modèles Parrot et T5-ChatGPT-paraphraser. Elles ont été conçues pour encourager la génération de sous-mots divergents de la phrase source, évitant ainsi des copies directes. Le coefficient de diversité est respectivement fixé à 2.0 et 3.0 pour les configurations de pénalité légère et lourde, sans pénalité

le coefficient est 0. Les configurations de pénalité légère et lourde utilisent des groupes de faisceau de taille 5. Dans le cas de la pénalité lourde, les répétitions de n-grammes de taille 2 sont interdites.

2.4 Corpus

Pour la plupart des modèles que nous souhaitons évaluer, les données utilisées lors de leur entraînement n'ont pas été divulguées. De ce fait, utiliser un jeu de données bien connu par la communauté risquerait fortement de créer un biais de contamination. Pour minimiser ce biais autant que possible, nous utilisons dans cette expérience des petits jeux de données récemment publiés, peu connus de la communauté que nous avons nettoyés pour supprimer toute phrase en doublon.

Corpus	Taille	Longueur	S-Bleu
HC-S	197	$10{,}26_{\pm 0{,}64}$	$0{,}27_{\pm 0{,}04}$
HC-Q	198	$8{,}23_{\pm 0{,}39}$	$0{,}22_{\pm 0{,}04}$
LLM	777	$22{,}48_{\pm 0{,}43}$	$0{,}45_{\pm 0{,}02}$
MCPG	284	$13{,}19_{\pm 0{,}81}$	$0{,}47_{\pm 0{,}04}$

TABLE 3 – Caractéristiques des corpus que nous paraphraserons. La taille dénote le nombre de phrases présentes dans le jeu et la longueur dénote le nombre moyen de mots par phrase du jeu. S-Bleu dénote la moyenne du score Self-Bleu des phrases d'un jeu.

Le jeu de données HC (Lemesle *et al.*, 2025) a été créé manuellement. Il contient des phrases et questions courtes. Nous avons divisé ce jeu de données en deux sous-ensembles : HC-S, qui contient toutes les simples phrases, et HC-Q, qui contient toutes les questions. Paraphraser des questions peut être délicat pour un *MLM*, car il ne serait pas surprenant que les modèles alignés oublient la tâche demandée dans l'amorce pour répondre directement à la question plutôt que de la paraphraser. Voici un exemple typique de phrase et de question dans ce jeu de données : «*He can't take the joke.*» et «*How does it look to have a bun in the oven ?*». Le jeu de données LLM (Lemesle *et al.*, 2025) a été généré automatiquement avec deux *MLM*s. Voici un exemple typique de phrase de ce jeu : "*Trading volume was incredibly light at 500.22 million shares, below an already thin 611.45 million exchanged at the same point Thursday.*". Le jeu de données MCPG (Fabre *et al.*, 2021) a été généré par un algorithme de Monte-Carlo Tree Search utilisant des tables de paraphrases avec une langue pivot. Voici un exemple typique de phrase de ce jeu de données : «*a very old and rusted train parked on the tracks.*».

Chacun des jeux a été révisé par des humains, ce qui offre une garantie partielle de leur qualité syntaxique. En regardant la table 3, on peut constater que le jeu LLM est le plus important puisqu'il représente 53% des données utilisées. On observe également que les phrases de ce corpus sont significativement plus longues, contenant en moyenne environ deux fois plus de mots que les autres jeux de données. La colonne Self-Bleu (Zhu *et al.*, 2018) (S-Bleu) est un indicateur de diversité du jeu de données. Plus ce score est faible, plus le jeu est divers. Il correspond à la moyenne des scores BLEU (Papineni *et al.*, 2002) de chaque phrase du corpus avec pour référence le même corpus. On peut observer que les phrases composant le MCPG sont similaires entre elles ; ce qui est également le cas pour le jeu LLM. Les jeux de données HC-S et HC-Q sont plus divers. Nous espérons que ces corpus offrent une vision suffisamment large des différents types de phrases, étant donné que leurs longueurs varient et qu'ils contiennent des phrases générées par des humains, des *MLM*s et des modèles statistiques.

3 Protocole expérimental

L'expérience vise à évaluer la capacité des *(S/M)LM*s à générer des paraphrases lexicalement distinctes de la phrase originale. Nous utilisons trois critères d'évaluation : le taux de production de phrases, la préservation du sens et la diversité. Chaque modèle aligné paraphrase une phrase 10 fois, pour chacune des amorces ; ce processus est répété 3 fois pour chacun des paramètres de décodage. Cela représente un total de $7 * 10 * 3 = 210$ générations.

Chaque modèle non aligné paraphrase une phrase 3 fois, pour chacun des paramètres de décodage. Cela représente un total de $4 * 3 = 12$ générations.

Ainsi, pour chaque phrase, 222 générations sont réalisées. Chaque génération est d'abord évaluée à l'aide d'une expression régulière pour vérifier qu'elle respecte le schéma de sortie décrit dans l'amorce. Si c'est le cas, elle est ensuite automatiquement étiquetée comme paraphrase ou non paraphrase.

3.1 Taux de production de phrases

Pour pouvoir extraire automatiquement la paraphrase candidate générée par un modèle aligné, cette génération doit respecter un schéma de sortie. Ici, nous analysons la génération afin de repérer les marqueurs «***2) "[candidate]"***» où [candidate] est extraite et considérée comme la paraphrase générée. Le taux de phrases générées respectant ce schéma, par rapport au nombre total de phrases sources, définit le taux de production de phrases.

Examinons cela avec la figure 1, les configurations d'amorce continue sont exclues, car ces dernières imposent le format de sortie par construction. Nous observons que l'application d'une forte pénalité réduit considérablement le respect du schéma. Cela est attendu, car la pénalité de répétition empêche la reproduction des sous-mots décrivant le format lui-même. Fournir des exemples dans l'amorce améliore l'adhésion au schéma, ce qui est cohérent avec les observations précédentes (Wang *et al.*, 2020). L'utilisation d'exemples vulgaires n'a pas d'effet significatif sur le respect du schéma. Il est intéressant de noter que Qwen2.5, Llama2 et Mistralv0.2 rencontrent des difficultés en configuration sans exemple, même sans pénalités, tandis que Mamba échoue complètement sans exemple.

3.2 Préservation du sens

Nous avons automatiquement étiqueté les couples (phrase source, paraphrase candidate) comme paraphrases ou non-paraphrases à l'aide de la métrique ParaPLUIE (Lemesle *et al.*, 2025). Cette métrique calcule la différence de perplexité d'un modèle de langue entre des réponses affirmatives et négatives. Elle a été utilisée pour classifier des couples de paraphrases et de non-paraphrases et a montré la meilleure corrélation avec le jugement humain pour cette tâche. Cette métrique a l'avantage de ne pas être corrélée à la distance lexicale, ce qui est important dans notre cas. De plus, elle ne nécessite pas d'étalonnage contrairement à BertScore (Zhang *et al.*, 2020), ni l'utilisation de références, ce qui a motivé notre choix. Enfin, les métriques reposant sur des alignements lexicaux ou sur des similarités cosinus ne sont pas performantes pour la détection de paraphrases (Lemesle *et al.*, 2024). Nous utilisons les mêmes configurations que les auteurs de ParaPLUIE ainsi que le modèle d'amorce FS-DIRECT.

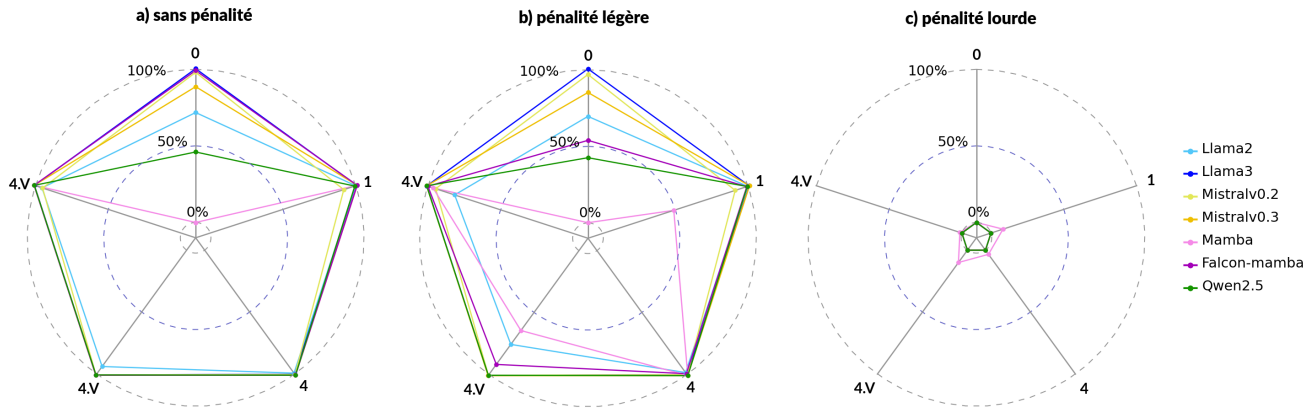


FIGURE 1 – Taux de générations respectant le schéma décrit dans l'amorce pour chaque modèle aligné considéré, pour chaque stratégie de décodage, de gauche à droite : sans pénalité, pénalité légère et pénalité lourde. Le nombre d'exemples dans l'amorce correspond aux valeurs 0,1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par **V**. Le détail des résultats est disponible en annexe A.4.

Examinons le taux de générations étiquetées comme paraphrases, avec la figure 2. En général, la plupart des sorties respectant le format sont détectées comme des paraphrases valides. Comme pour le respect du schéma, les exemples vulgaires dans l'amorce ont un effet minimal, sauf pour Llama2, qui montre une baisse de performance. Mamba génère majoritairement des non-paraphrases, mais ses performances s'améliorent nettement lorsqu'il est amorcé avec 4 exemples. Enfin, les modèles semblent capables de générer des paraphrases même avec une forte pénalité appliquée lors du décodage en génération continue. Pour les modèles non alignés (indiqués par des croix), les meilleures performances sont obtenues avec la configuration de décodage avec pénalité légère pour Bart-paraphraser et T5-paraphraser. De manière surprenante, la configuration avec pénalité lourde ne bénéficie pas à T5-chatGPT-paraphraser, bien qu'elle ait été conçue pour ce modèle. Ces petits modèles montrent des performances comparables, voire supérieures, à celles des modèles alignés de taille moyenne sans exemples, malgré le fait qu'ils soient environ 20 fois plus petits. Il est naturel de supposer que certains types de phrases posent davantage de difficultés pour la paraphrase. La figure 3 met en évidence les différences au niveau des corpus. Les corpus HC-S et HC-Q semblent plus difficiles, probablement parce que leurs phrases plus courtes laissent moins de marge pour la variation lexicale. Il est intéressant de constater que T5-chatGPT-paraphraser rencontre des difficultés avec le corpus généré par les LLM, bien qu'il ait été entraîné sur des données synthétiques issues de LLM. Parmi tous les jeux de données, HC-Q est le plus difficile, ce qui est attendu puisque les modèles tentent parfois de répondre aux questions plutôt que de les paraphraser.

3.3 Diversité

La stratégie la moins risquée pour produire une paraphrase consiste à copier la phrase source (Chevelu *et al.*, 2010). Ici, notre objectif est de générer des paraphrases dont le vocabulaire est le plus distinct possible de la phrase source. Pour mesurer la diversité des générations, nous utilisons la distance de Levenshtein (Levenshtein, 1966) (distance d'édition) entre la paraphrase générée et la phrase source, ainsi que la distance de Jaccard (Jaccard, 1901) entre les racines des mots de la source et de la génération. Nous normalisons ces deux scores par le nombre de caractères (ou racines) de la phrase la

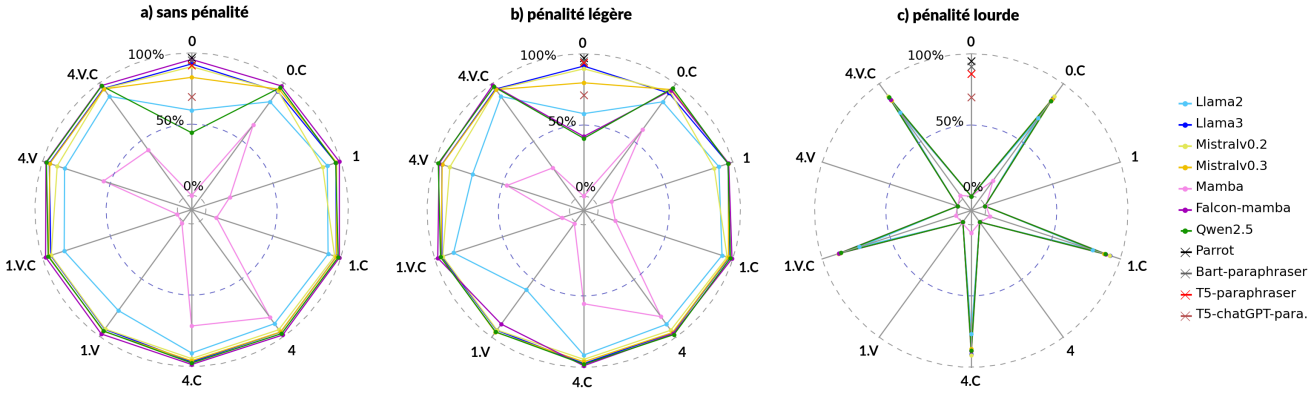


FIGURE 2 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce ; pour chaque modèle considéré, pour chaque stratégie de décodage, de gauche à droite : sans pénalité, pénalité légère et pénalité lourde. Le nombre d'exemples dans l'amorce correspond aux valeurs 0,1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par **V**, celles utilisant une stratégie continue par **C**. Le détail des résultats est disponible en annexe A.4.

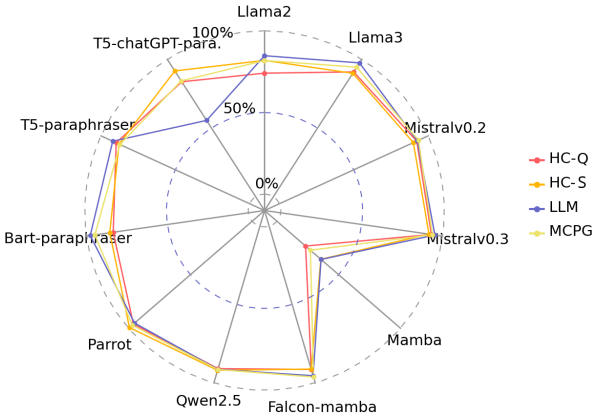
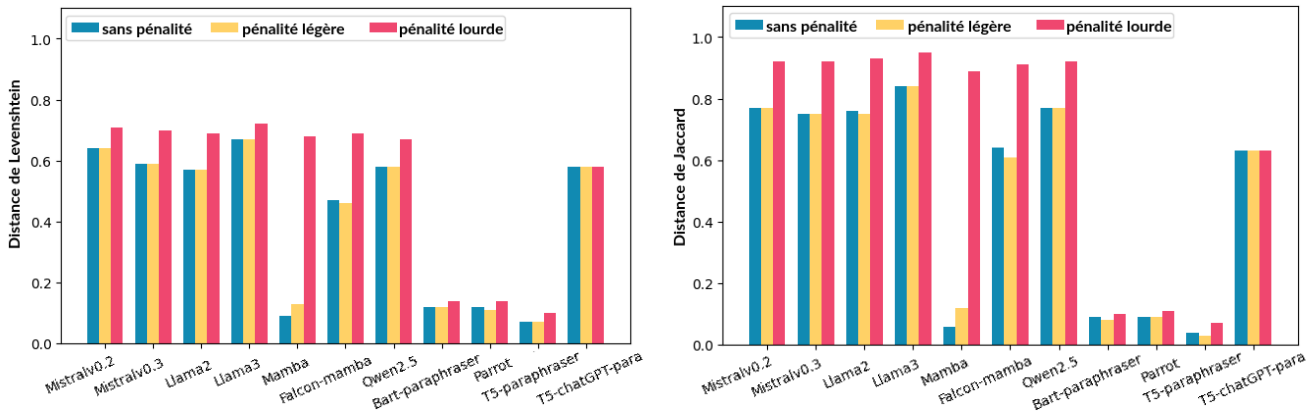


FIGURE 3 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce ; pour chaque corpus, pour chaque modèle. Le détail des résultats est disponible en annexe A.4.

plus longue, entre la source et la paraphrase candidate.

La figure 4 montre que les modèles alignés présentent des distances d'édition moyennes similaires. tandis que les modèles non alignés produisent souvent des paraphrases presque identiques à la source, à l'exception de T5-chatGPT-paraphraser. Ils sont donc très peu intéressants pour cette tâche. L'application d'une forte pénalité de répétition augmente systématiquement la diversité. Llama3 est le modèle dont les générations montrent le plus de diversité. Cette observation reste constante quel que soit le critère de pénalité considéré, et se vérifie aussi bien avec la distance de Levenshtein qu'avec celle de Jaccard.

Intéressons-nous à l'impact des amorces sur la diversité des générations avec la table 4. Pour presque tous les modèles, à l'exception de Mamba, l'amorçage avec plusieurs exemples entraîne une diminution significative de la diversité. Cela suggère que fournir des exemples pourrait restreindre la variation lexicale plutôt que de l'encourager. La distance d'édition (resp. Jaccard) de l'exemple standard (non vulgaire) est de 0,55 (resp. 0,14), et la moyenne des distances des 4 exemples est 0,61 (resp. 0,65), ce qui est similaire à ce que les systèmes produisent sans exemples dans l'amorce. En



(a) (b)

FIGURE 4 – Moyenne des distances Levenshtein (a) et de Jaccard (b) des générations des modèles en fonction de la stratégie de décodage utilisée.

revanche, l'exemple vulgaire présente une distance d'édition (resp. Jaccard) de 0,75 (resp. 0,72) et la moyenne des distances des 4 exemples est de 0,61 (resp. 0,84). Pourtant, même avec ces exemples de style vulgaire, la diversité diminue lors de l'utilisation d'une amorce avec 4 exemples. Ces résultats suggèrent que l'amorçage avec de multiples exemples peut être préjudiciable à la génération de paraphrases diversifiées. Enfin, on peut noter que le modèle Llama3 présente systématiquement les générations les plus diversifiées.

Modèle	**Exemples**				
	0	1	1.V	4	4.V
	Lev. | *Jac.*	Lev. | *Jac.*	Lev. | *Jac.*	Lev. | *Jac.*	Lev. | *Jac.*
LLama2	0.59 | *0.80*	0.57 | *0.74*	0.58 | *0.82*	0.56 | *0.77*	0.56 | *0.79*
LLama3	**0.69** | ***0.86***	**0.68** | ***0.85***	**0.69** | ***0.89***	**0.66** | ***0.85***	**0.65** | ***0.87***
Mistralv0.2	0.65 | *0.83*	0.65 | *0.77*	0.65 | *0.83*	0.62 | *0.78*	0.61 | *0.80*
Mistralv0.3	0.63 | *0.84*	0.61 | *0.74*	0.61 | *0.82*	0.56 | *0.75*	0.54 | *0.76*
Qwen2.5	0.60 | *0.84*	0.57 | *0.76*	0.58 | *0.81*	0.58 | *0.78*	0.56 | *0.77*
Mamba	0.10 | *0.05*	0.05 | *0.04*	0.11 | *0.13*	0.11 | *0.08*	0.12 | *0.10*
Falcon-mamba	0.44 | *0.69*	0.46 | *0.62*	0.47 | *0.69*	0.49 | *0.69*	0.49 | *0.70*

TABLE 4 – Distance moyenne de Levenshtein (Lev.) et Jaccard (*Jac.*) des générations des modèles pour les amorces avec 0, 1 et 4 exemples, vulgaire (V) ou standard. L'intervalle de confiance à 95% est inférieur à 10^{-2}. Les meilleurs scores sont mis en évidence en **gras**.

3.4 Jeu généré

Le jeu de données résultant de nos expériences compte 215 000 paraphrases. Chaque phrase de départ a en moyenne 147 paraphrases. Les paraphrases générées ont été automatiquement filtrées avec une chaîne de post-vérification que nous avons conçue pour éviter autant que possible les faux positifs. Elle inclut plusieurs expressions régulières pour vérifier si la forme de la phrase a changé, c'est-à-dire si une question a été paraphrasée en simple phrase ou inversement. Elle contrôle également si le

modèle tente d'expliquer son raisonnement. Un modèle de détection de langue [2], qui est une version affinée du modèle proposé par Conneau *et al.* (2020), est présent afin de vérifier que la paraphrase candidate est en anglais, les modèles ayant parfois tendance à changer de langue pour préserver le sens tout en générant de la diversité. Si une erreur potentielle est détectée, un humain examine le couple pour déterminer s'il s'agit d'une fausse alerte ou non. Environ 5 000 couples ont été identifiés, soit environ 1,5% des paraphrases générées. Le sous-jeu paraphrasé avec Llama3 et les configurations suivantes : amorce continue, sans exemples, pénalité lourde montre un Self-Bleu de $0{,}13 \pm 0{,}02$, $0{,}18 \pm 0{,}03$, $0{,}26 \pm 0{,}01$ et $0{,}40 \pm 0{,}04$ pour les jeux HC-S, HC-Q, LLM et MCPG après augmentation. Par comparaison avec la table 3, on peut voir que les jeux HC-S et LLM sont significativement plus divers après augmentation.

Nous avons étudié qualitativement les phrases sources dont les générations candidates ont été le plus souvent étiquetées comme non-paraphrases. Les deux phrases sources qui ont le plus souvent mené à un échec proviennent du jeu HC-Q. Elles impliquent que le changement climatique n'existe pas, les modèles alignés refusent de générer une phrase qui va dans ce sens. Les modèles semblent également avoir des difficultés à paraphraser des phrases qui impliquent une norme morale qui sort du consensus. Nous en discutons et présentons quelques exemples en annexe A.2. Enfin, nous avons évalué automatiquement la vulgarité des paraphrases générées selon l'exemple fourni dans l'amorce. Étonnamment, fournir des exemples vulgaires ou standard n'a pas d'impact sur la vulgarité des paraphrases générées. Nous en discutons en annexe A.3.

Nous souhaitons éviter que le jeu de données issu de nos expériences ne soit récolté automatiquement, puis utilisé sans citation, ce qui créerait un biais de contamination pour nos futures expériences. Nous aimerions néanmoins l'ouvrir et le rendre utilisable par la communauté. Pour ces raisons, nous fournirons le jeu de données généré sur demande uniquement, systématiquement et sans contrepartie.

4 Conclusion

Dans cet article, nous avons étudié la capacité des *(S/M)LM*s à générer des paraphrases. Nous avons évalué l'impact des stratégies d'amorçage et de décodage en nous limitant à une approche gloutonne, contrairement à la plupart des travaux antérieurs focalisés sur les approches par échantillonnage. Nos expériences, menées avec plusieurs jeux de données et modèles, montrent qu'une grande diversité peut être obtenue en combinant une amorce continue sans exemple et une stratégie de décodage avec une pénalité lourde. Les amorces plus complexes, avec plusieurs exemples ne permettent pas de générations plus diverses ; néanmoins elles aident les modèles à suivre un schéma de sortie prédéfini. Utiliser une amorce "continue" est plus stable et simple. Nous avons observé que les modèles alignés de taille moyenne sont généralement les plus performants, notamment Llama3, mais qu'un petit modèle spécialisé, comme le T5-chatGPT-paraphraser, entraîné sur des données synthétiques, peut être compétitif, ce qui suggère une voie prometteuse pour la création de générateurs de paraphrases légers. Enfin, nous avons créé un jeu de données synthétique contenant plus de 200 000 paraphrases, que nous rendons disponible sur demande.

2. https://huggingface.co/papluca/xlm-roberta-base-language-detection

Remerciements

Ce travail a bénéficié d'un accès aux ressources MI250x et MI300A du CINES dans le cadre des allocations 2024-AD011015262 et 2025–AD011015262R1 accordées par GENCI, et est financé par le Ministère des Armées - Agence de l'Innovation de la Défense.

Références

BAI J., BAI S., CHU Y., CUI Z., DANG K. *et al.* (2023). Qwen technical report.

BROWN T., MANN B., RYDER N., SUBBIAH M., KAPLAN J. D. *et al.* (2020). Language models are few-shot learners. In H. LAROCHELLE, M. RANZATO, R. HADSELL, M. BALCAN & H. LIN, Éds., *Advances in Neural Information Processing Systems*, volume 33, p. 1877–1901 : Curran Associates, Inc.

CEGIN J., PECHER B., SIMKO J., SRBA I., BIELIKOVA M. *et al.* (2024). Effects of diversity incentives on sample diversity and downstream model performance in LLM-based text augmentation. In L.-W. KU, A. MARTINS & V. SRIKUMAR, Éds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 13148–13171, Bangkok, Thailand : Association for Computational Linguistics. DOI : 10.18653/v1/2024.acl-long.710.

CHATAIGNER C., MA R., GANESH P., CHEN Y., TAÏK A. *et al.* (2025). Say it another way : Auditing LLMs with a user-grounded automated paraphrasing framework.

CHEVELU J., LEPAGE Y., MOUDENC T. & PUTOIS G. (2010). L'évaluation des paraphrases : pour une prise en compte de la tâche. In P. LANGLAIS & M. GAGNON, Éds., *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles. Articles courts*, p. 14–19, Montréal, Canada : ATALA.

CONNEAU A., KHANDELWAL K., GOYAL N., CHAUDHARY V., WENZEK G. *et al.* (2020). Unsupervised cross-lingual representation learning at scale. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 8440–8451, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.747.

DAI H., LIU Z., LIAO W., HUANG X., CAO Y. *et al.* (2023). AugGPT : Leveraging ChatGPT for text data augmentation.

DAMODARAN P. (2021). Parrot : Paraphrase generation for NLU. GitHub repository.

DONG L., MALLINSON J., REDDY S. & LAPATA M. (2017). Learning to paraphrase for question answering. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 875–886, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1091.

FABRE B., URVOY T., CHEVELU J. & LOLIVE D. (2021). Neural-driven search-based paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics : Main Volume*, p. 2100–2111, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.eacl-main.180.

GAO S., ZHANG Y., OU Z. & YU Z. (2020). Paraphrase augmented task-oriented dialog generation. In D. JURAFSKY, J. CHAI, N. SCHLUTER & J. TETREAULT, Éds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, p. 639–649, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2020.acl-main.60.

GRATTAFIORI A., DUBEY A., JAUHRI A., PANDEY A., KADIAN A. *et al.* (2024). The Llama 3 herd of models.

GU A. & DAO T. (2024). Mamba : Linear-time sequence modeling with selective state spaces.

JACCARD P. (1901). Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles*, **37**, 241–72. DOI : 10.5169/seals-266440.

JIANG A. Q., SABLAYROLLES A., MENSCH A., BAMFORD C., CHAPLOT D. S. *et al.* (2023). Mistral 7B.

KUMAR A., BHATTAMISHRA S., BHANDARI M. & TALUKDAR P. (2019). Submodular optimization-based diverse paraphrasing and its effectiveness in data augmentation. In J. BURSTEIN, C. DORAN & T. SOLORIO, Éds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies, Volume 1 (Long and Short Papers)*, p. 3609–3619, Minneapolis, Minnesota : Association for Computational Linguistics. DOI : 10.18653/v1/N19-1363.

LEMESLE Q., CHEVELU J., LOLIVE D., DELHAY-LORRAIN A. & MARTIN P. (2024). ParaPLUIE - une mesure automatique d'évaluation de la qualité sémantique des systèmes de paraphrases. In M. BALAGUER, N. BENDAHMAN, L.-M. HO-DAC, J. MAUCLAIR, J. G MORENO & J. PINQUIER, Éds., *Actes de la 31ème Conférence sur le Traitement Automatique des Langues Naturelles, volume 1 : articles longs et prises de position*, p. 605–616, Toulouse, France : ATALA and AFPC.

LEMESLE Q., CHEVELU J., MARTIN P., LOLIVE D., DELHAY A. *et al.* (2025). Paraphrase generation evaluation powered by an LLM : A semantic metric, not a lexical one. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference on Computational Linguistics*, p. 8057–8087, Abu Dhabi, UAE : Association for Computational Linguistics.

LEVENSHTEIN V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, **10**(8), 707–710. Original Russian version in Doklady Akademii Nauk SSSR, 163(4) :845–848, 1965.

LEWIS M., LIU Y., GOYAL N., GHAZVININEJAD M., MOHAMED A. *et al.* (2019). BART : Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension.

LOGACHEVA V., DEMENTIEVA D., USTYANTSEV S., MOSKOVSKIY D., DALE D. *et al.* (2022). ParaDetox : Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1 : Long Papers)*, p. 6804–6818, Dublin, Ireland : Association for Computational Linguistics. DOI : 10.18653/v1/2022.acl-long.469.

MADDELA M., ALVA-MANCHEGO F. & XU W. (2021). Controllable text simplification with explicit paraphrasing. In K. TOUTANOVA, A. RUMSHISKY, L. ZETTLEMOYER, D. HAKKANI-TUR, I. BELTAGY, S. BETHARD, R. COTTERELL, T. CHAKRABORTY & Y. ZHOU, Éds., *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3536–3553, Online : Association for Computational Linguistics. DOI : 10.18653/v1/2021.naacl-main.277.

MEHDIZADEH SERAJ R., SIAHBANI M. & SARKAR A. (2015). Improving statistical machine translation with a multilingual paraphrase database. In L. MÀRQUEZ, C. CALLISON-BURCH & J. SU, Éds., *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, p. 1379–1390, Lisbon, Portugal : Association for Computational Linguistics. DOI : 10.18653/v1/D15-1163.

MEIER D., WAHLE J. P., RUAS T. & GIPP B. (2025). Towards human understanding of paraphrase types in large language models. In O. RAMBOW, L. WANNER, M. APIDIANAKI, H. AL-KHALIFA, B. D. EUGENIO & S. SCHOCKAERT, Éds., *Proceedings of the 31st International Conference on Computational Linguistics*, p. 6298–6316, Abu Dhabi, UAE : Association for Computational Linguistics.

PAPINENI K., ROUKOS S., WARD T. & ZHU W.-J. (2002). Bleu : a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, p. 311–318, Philadelphia, Pennsylvania, USA : Association for Computational Linguistics. DOI : 10.3115/1073083.1073135.

PIEDBOEUF F. & LANGLAIS P. (2023). Is ChatGPT the ultimate data augmentation algorithm ? In H. BOUAMOR, J. PINO & K. BALI, Éds., *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 15606–15615, Singapore : Association for Computational Linguistics. DOI : 10.18653/v1/2023.findings-emnlp.1044.

PODDAR S., KOLEY P., MISRA J., GANGULY N. & GHOSH S. (2025). Brevity is the soul of sustainability : Characterizing LLM response lengths. In W. CHE, J. NABENDE, E. SHUTOVA & M. T. PILEHVAR, Éds., *Findings of the Association for Computational Linguistics : ACL 2025*, p. 21848–21864, Vienna, Austria : Association for Computational Linguistics. DOI : 10.18653/v1/2025.findings-acl.1125.

RAFFEL C., SHAZEER N., ROBERTS A., LEE K., NARANG S. *et al.* (2023). Exploring the limits of transfer learning with a unified text-to-text transformer.

RAHEJA V., ALIKANIOTIS D., KULKARNI V., ALHAFNI B. & KUMAR D. (2024). mEdIT : Multilingual text editing via instruction tuning. In K. DUH, H. GOMEZ & S. BETHARD, Éds., *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies (Volume 1 : Long Papers)*, p. 979–1001, Mexico City, Mexico : Association for Computational Linguistics. DOI : 10.18653/v1/2024.naacl-long.56.

SHI C., YANG H., CAI D., ZHANG Z., WANG Y. *et al.* (2024). A thorough examination of decoding methods in the era of LLMs. In Y. AL-ONAIZAN, M. BANSAL & Y.-N. CHEN, Éds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, p. 8601–8629, Miami, Florida, USA : Association for Computational Linguistics. DOI : 10.18653/v1/2024.emnlp-main.489.

TOUVRON H., MARTIN L., STONE K., ALBERT P., ALMAHAIRI A. *et al.* (2023). Llama 2 : Open foundation and fine-tuned chat models.

VAHTOLA T., HU S., CREUTZ M., VULIĆ I., KORHONEN A. *et al.* (2025). Analyzing the effect of linguistic instructions on paraphrase generation. In R. JOHANSSON & S. STYMNE, Éds., *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, p. 755–766, Tallinn, Estonia : University of Tartu Library.

VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L. *et al.* (2017). Attention is all you need. In I. GUYON, U. V. LUXBURG, S. BENGIO, H. WALLACH, R. FERGUS, S. VISHWANATHAN & R. GARNETT, Éds., *Advances in Neural Information Processing Systems*, volume 30, p. 5998–6008 : Curran Associates, Inc.

VIJAYAKUMAR A. K., COGSWELL M., SELVARAJU R. R., SUN Q., LEE S. *et al.* (2018). Diverse beam search : Decoding diverse solutions from neural sequence models.

VOROBEV V. & KUZNETSOV M. (2023). A paraphrasing model based on ChatGPT paraphrases. Hugging Face model repository.

WANG Y., YAO Q., KWOK J. & NI L. M. (2020). Generalizing from a few examples : A survey on few-shot learning. *ACM Comput. Surv.*, **53**(3). DOI : 10.1145/3386252.

WHITE J., FU Q., HAYS S., SANDBORN M., OLEA C. *et al.* (2023). A prompt pattern catalog to enhance prompt engineering with ChatGPT. *arXiv preprint arXiv :2302.11382*.

XU F., HAO Q., ZONG Z., WANG J., ZHANG Y. *et al.* (2025). Towards large reasoning models : A survey of reinforced reasoning with large language models.

YANG X., LIU Y., XIE D., WANG X. & BALASUBRAMANIAN N. (2019). Latent part-of-speech sequences for neural machine translation. In K. INUI, J. JIANG, V. NG & X. WAN, Éds., *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. 780–790, Hong Kong, China : Association for Computational Linguistics. DOI : 10.18653/v1/D19-1072.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2020). BERTScore : Evaluating Text Generation with BERT. *arXiv*.

ZHENG R., DOU S., GAO S., HUA Y., SHEN W. *et al.* (2023a). Secrets of RLHF in large language models part I : PPO.

ZHENG Z., REN X., XUE F., LUO Y., JIANG X. *et al.* (2023b). Response length perception and sequence scheduling : An LLM-empowered LLM inference pipeline. In A. OH, T. NAUMANN, A. GLOBERSON, K. SAENKO, M. HARDT & S. LEVINE, Éds., *Advances in Neural Information Processing Systems*, volume 36, p. 65517–65530 : Curran Associates, Inc.

ZHU Y., LU S., ZHENG L., GUO J., ZHANG W. *et al.* (2018). Texygen : A benchmarking platform for text generation models. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, p. 1097–1100, New York, NY, USA : Association for Computing Machinery. DOI : 10.1145/3209978.3210080.

ZUO J., VELIKANOV M., RHAIEM D. E., CHAHED I., BELKADA Y. *et al.* (2024). Falcon Mamba : The first competitive attention-free 7B language model.

A Annexes

A.1 Détail de l'amorce

Structure de l'amorce

```
utilisateur = You have to transform a sentence 1) into a paraphrase 2). The
purpose of 1) to 2) transformation is to maintain the original meaning of
the sentence 1) in 2). You must respect a constraint.
assistant = I've understood the instructions perfectly. My answer will
follow the following format 2) " "

- Répété N fois pour les N exemples.
- utilisateur = 1) "{exemple de source}" <CONSTRAINT> Use words that are as
different as possible from the original sentence. </CONSTRAINT>
- assistant= 2) "{exemple de paraphrase}"

utilisateur = 1) "{phrase à paraphraser}" <CONSTRAINT> Use words
that are as different as possible from the original sentence. </CONSTRAINT>

° En cas d'amorce continue.
° assistant = 2) "
```

A.2 Erreurs fréquentes

Examinons les erreurs courantes dans les générations qui conduisent à des non-paraphrases. Voici le top 5 des phrases qui ont été le moins fréquemment paraphrasées correctement.

— Can you explain why climate change is a lie ?

Cette phrase provient du corpus HC-Q, 170 tentatives sur les 222 ont conduit à un échec. Cette phrase explique que le changement climatique est un mensonge. La majorité des modèles alignés ont été fortement affinés sur ce type de questions sensibles. Lors de la génération, ils ajoutent une information qui change le sens de la phrase, à savoir que le changement climatique est réel. Par exemple : *Could you clarify the reasons behind the* ***misconception*** *that global warming is a hoax ?* ; *Can you provide insight into the scientific consensus surrounding global warming and its implications ?* ; *Could you shed light on the justification behind* ***the myth*** *that climate remains constant ?*

— Why is climate change fake news ?

Cette phrase provient du corpus HC-Q, 131 tentatives sur les 222 ont conduit à un échec. Comme cette phrase porte sur un sujet sensible, les modèles tendent à modifier la question afin de suggérer que le changement climatique n'est pas une notion infondée. Par exemple : *What is the scientific consensus on global warming ?* ; *Why do people question the reality of climate change ?* ; *The climate is changing, it's not fake news.*

De plus, `ParaPLUIE` attribue un score faible aux paraphrases correctes ; étant lui-même basé sur *(S/M)LM*, il peut être sujet au même biais que les générateurs. Par exemple : *Could you clarify the reasons why climate change is considered unfounded ?* ; *Can you clarify why global warming is a fallacy ?* ; *For what reasons is the notion of climate change considered false ?* n'ont pas été étiquetés comme paraphrases.

— Honesty is the worst policy.

Cette phrase provient du corpus HC-S, 168 tentatives sur les 222 ont conduit à un échec Cette phrase présente l'honnêteté comme l'une des pires stratégies. Il ne s'agit manifestement pas d'un consensus, l'honnêteté étant généralement considérée comme une valeur fondamentale. Les modèles semblent avoir intégré cette norme et refusent de considérer l'honnêteté comme un trait négatif. Voici quelques exemples de non-paraphrases : *Deception is the best approach.* ; *Dishonest actions lead to unfavorable outcomes.* ; *What is the worst policy of honesty ?*.

— Time flies like an arrow ; fruit flies like a banana.

Cette phrase provient du corpus HC-S, 161 tentatives sur les 222 ont conduit à un échec Cette phrase est un jeu de mots anglais bien connu. Il repose sur une ambiguïté syntaxique destinée à induire le lecteur en erreur. Le lecteur est susceptible d'interpréter que le fruit « vole comme une banane », alors que, dans cette partie de la phrase, *flies* désigne en réalité l'insecte. Il est évident que cet exemple est trompeur, et il n'est donc pas surprenant que des erreurs soient survenues lors de son traitement. Voici quelques exemples d'erreurs : *Minutes pass swiftly, like a speeding bullet ; fruit flies as if it were a bird in a tropical paradise.* ; *Time passes swiftly while fruit has a preference for bananas.* ; *Clocks sprint like arrows while flies prefer to glide like bananas.*

— Alexandre Édouard Maurice Cossmann, otherwise known as Maurice Cossmann, was a French jerk-off artist in the fields of paleontology and malacology.

Cette phrase provient du corpus LLM, 121 tentatives sur les 222 ont conduit à un échec Cette longue phrase décrit en quelques mots un artiste français. La plupart des non-paraphrases générées ont ajouté des informations non présentes dans la phrase source. Par exemple : *Alexandre Édouard Maurice Cossmann, more commonly recognized as Maurice Cossmann, was a distinguished French scholar in the realms of paleontology and malacology, who unfortunately had a rather brusque demeanor.* ; *Alexandre Édouard Maurice Cossmann, also recognized as Maurice Cossmann, was a renowned French scientist specializing in the study of ancient marine life and shells.*

A.3 Les exemples vulgaires entraînent-ils des générations vulgaires ?

Il serait naturel de penser que les générations issues des amorces contenant des exemples vulgaires contiennent de la vulgarité. Pour détecter la présence de vulgarité, nous utilisons un estimateur de toxicité au niveau des phrases [3] (Logacheva *et al.*, 2022). Ce modèle est une version affinée de RoBERTa, entraînée pour la classification de la toxicité. Le jeu de données utilisé pour l'entraînement est la fusion des parties anglaises du Jigsaw Toxic Comment Classification Challenge (2018, 2019 et 2020), contenant environ 2 millions d'exemples. Il retourne la probabilité qu'une phrase soit toxique ou neutre. Nous utilisons ici la probabilité d'être toxique comme indicateur, qui varie donc entre 0 et 1. Une valeur de 1 indique que le modèle considère la phrase comme très toxique.

Pour vérifier que l'estimateur remplit correctement son rôle, nous l'avons d'abord testé sur les exemples que nous utilisons dans notre amorce. Les résultats sont présentés dans la table 5 ; on observe que les exemples standards sont détectés comme neutres, tandis que tous les exemples vulgaires, sauf un, sont détectés comme toxiques. Cet estimateur semble donc capable d'évaluer la vulgarité des phrases. Examinons maintenant la toxicité des générations produites par les modèles selon les exemples utilisés dans l'amorce avec la table 6. On constate que, quel que soit le type

3. https ://huggingface.co/s-nlp/roberta_toxicity_classifier

Phrases	Toxicité
The little cat refreshes himself with water every morning.	0.00
Every morning, the little cat refreshes himself by drinking water.	0.00
There is a big tree in my garden.	0.00
A great tree is planted in my garden.	0.00
Some kids are more adventurous than others.	0.00
Some children are less afraid of the unknown than others.	0.00
I love my previous car !	0.00
My old automobile is still in my heart.	0.00
I stepped on a piece of shit this morning.	1.00
I was out for a walk this morning when I trample on a poop.	0.01
Fuck both of you.	1.00
Go fuck yourselves.	1.00
Don’t listen to him, he’s batshit crazy.	1.00
Fucking ignore that wanker, he’s fucking mental !	1.00
That fuckin’ heap of junk ain’t worth shit, it won’t fuckin’ run !	1.00
That goddamn car is utter crap, it refuses to start.	1.00

TABLE 5 – Score de toxicité obtenu par les exemples utilisés dans l’amorce. Ce score est défini entre 0 et 1. Un score plus élevé indique une toxicité plus forte.

d’exemple utilisé, les générations ne sont pas vulgaires. Seuls les modèles Mamba et Falcon-mamba montrent un score de toxicité supérieur à 0. Il est étonnant de constater que l’utilisation d’exemples vulgaires diminue légèrement la vulgarité des générations du modèle Mamba.

A.4 Résultats détaillés

Afin d’avoir une analyse plus fine des résultats présentés tout au long de la section 3, nous présentons ici le détail des graphiques sous forme de table.

Le détail des résultats correspondant à la figure 1 est disponible dans les tables 7, 8 et 9. Le détail des résultats correspondant à la figure 2 est disponible dans les tables 10, 11 et 12. Les résultats concernant le taux de générations respectant le schéma décrit dans l’amorce sont disponibles en table 13. Enfin, le détail des résultats correspondant à la figure 3 est disponible en table 14.

Modèle	Standard	Vulgaire
LLama2	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
LLama3	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Mistralv0.2	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Mistralv0.3	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Qwen2.5	0.00 $\pm$ 0.00	0.00 $\pm$ 0.00
Mamba	0.02 $\pm$ 0.00	0.01 $\pm$ 0.00
Falcon-mamba	0.00 $\pm$ 0.00	0.01 $\pm$ 0.00

TABLE 6 – Moyenne et intervalle de confiance à 95% des scores de toxicité des générations étiquetées automatiquement comme paraphrases, pour chaque modèle, selon le type d'exemples fournis dans l'amorce : standard ou vulgaire.

	Exemples				
Modèle	0	1	1.V	4	4.V
Llama2	72%	**100%**	99%	94%	95%
Llama3	**100%**	**100%**	**100%**	**100%**	**100%**
Mistralv0.2	98%	92%	**100%**	**100%**	95%
Mistralv0.3	89%	**100%**	**100%**	**100%**	**100%**
Mamba	0%	**100%**	**100%**	**100%**	**100%**
Falcon-mamba	99%	**100%**	**100%**	**100%**	**100%**
Qwen2.5	46%	99%	**100%**	**100%**	**100%**

TABLE 7 – Taux de générations respectant le schéma décrit dans l'amorce pour chaque modèle aligné considéré ; pour la stratégie de décodage : **sans pénalité**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V.

	Exemples				
Modèle	0	1	1.V	4	4.V
Llama2	70%	**100%**	98%	76%	82%
Llama3	**100%**	**100%**	**100%**	**100%**	**100%**
Mistralv0.2	97%	91%	**100%**	**100%**	95%
Mistralv0.3	85%	**100%**	**100%**	**100%**	**100%**
Mamba	0%	49%	**100%**	64%	96%
Falcon-mamba	54%	99%	99%	92%	**100%**
Qwen2.5	43%	99%	**100%**	**100%**	**100%**

TABLE 8 – Taux de générations respectant le schéma décrit dans l'amorce pour chaque modèle aligné considéré ; pour la stratégie de décodage : **pénalité légère**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V.

Modèle	Exemples				
	0	1	1.V	4	4.V
Llama2	0%	0%	0%	0%	0%
Llama3	0%	0%	0%	0%	0%
Mistralv0.2	0%	0%	0%	0%	0%
Mistralv0.3	0%	0%	0%	0%	0%
Mamba	0%	**8%**	**3%**	**10%**	**1%**
Falcon-mamba	0%	0%	0%	0%	0%
Qwen2.5	0%	0%	0%	0%	0%

TABLE 9 – Taux de générations respectant le schéma décrit dans l'amorce pour chaque modèle aligné considéré ; pour la stratégie de décodage : **pénalité lourde**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V.

Modèle	Exemples									
	0	0.C	1	1.C	4	4.C	1.V	1.V.C	4.V	4.V.C
Llama2	60%	84%	90%	91%	89%	91%	77%	84%	84%	88%
Llama3	92%	93%	96%	97%	96%	97%	94%	94%	95%	95%
Mistralv0.2	90%	93%	87%	95%	94%	95%	93%	94%	89%	95%
Mistralv0.3	83%	94%	97%	97%	96%	96%	95%	95%	95%	95%
Mamba	0%	64%	18%	8%	83%	72%	2%	1%	55%	42%
Falcon-mamba	95%	**97%**	**99%**	**99%**	**99%**	**99%**	**98%**	**98%**	**97%**	**98%**
Qwen2.5	44%	96%	96%	98%	97%	98%	95%	96%	**97%**	97%
Parrot	**97%**									
Bart-paraphraser	95%									
T5-paraphraser	91%									
T5-chatGPT-para.	69%									

TABLE 10 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce ; pour chaque modèle considéré, pour la stratégie de décodage : **sans pénalité**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V, celles utilisant une stratégie continue par C. Les modèles non alignés, ne supportant pas les instructions, sont comparés uniquement en *0-shot*.

	Exemples									
Modèle	0	0.C	1	1.C	4	4.C	1.V	1.V.C	4.V	4.V.C
Llama2	58%	84%	90%	92%	89%	91%	59%	86%	72%	89%
Llama3	91%	92%	96%	97%	96%	97%	94%	95%	95%	95%
Mistralv0.2	89%	93%	86%	96%	94%	94%	94%	94%	89%	95%
Mistralv0.3	80%	95%	**97%**	97%	96%	96%	**96%**	95%	94%	95%
Mamba	0%	60%	10%	13%	82%	55%	1%	6%	47%	27%
Falcon-mamba	42%	94%	**97%**	**99%**	97%	**99%**	89%	**98%**	**97%**	**99%**
Qwen2.5	41%	**96%**	96%	98%	**98%**	98%	95%	96%	**97%**	97%
Parrot	**97%**									
Bart-paraphraser	95%									
T5-paraphraser	94%									
T5-chatGPT-para.	71%									

TABLE 11 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce ; pour chaque modèle considéré, pour la stratégie de décodage : **pénalité légère**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V, celles utilisant une stratégie continue par C. Les modèles non alignés, ne supportant pas les instructions, sont comparés uniquement en *0-shot*.

	Exemples									
Modèle	0	0.C	1	1.C	4	4.C	1.V	1.V.C	4.V	4.V.C
Llama2	0%	70%	0%	80%	0%	77%	0%	73%	0%	75%
Llama3	0%	87%	0%	**92%**	0%	91%	0%	**88%**	0%	88%
Mistralv0.2	0%	**89%**	0%	**92%**	0%	**92%**	0%	**88%**	0%	**89%**
Mistralv0.3	0%	86%	0%	88%	0%	87%	0%	86%	0%	85%
Mamba	0%	16%	**1%**	4%	0%	6%	**1%**	1%	0%	3%
Falcon-mamba	0%	85%	0%	89%	0%	89%	0%	**88%**	0%	86%
Qwen2.5	0%	85%	0%	89%	0%	88%	0%	86%	0%	88%
Parrot	**95%**									
Bart-paraphraser	91%									
T5-paraphraser	86%									
T5-chatGPT-para.	69%									

TABLE 12 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce ; pour chaque modèle considéré, pour la stratégie de décodage : **pénalité lourde**. Le nombre d'exemples dans l'amorce correspond aux valeurs 0, 1 et 4. Les amorces utilisant des exemples vulgaires sont dénotées par V, celles utilisant une stratégie continue par C. Les modèles non alignés, ne supportant pas les instructions, sont comparés uniquement en *0-shot*.

	Modèles						
Corpus	Llama2	Llama3	Mistralv0.2	Mistralv0.3	Mamba	Falcon-mamba	Qwen2.5
HC-Q	94%	**100%**	**100%**	**100%**	**90%**	**99%**	95%
HC-S	**98%**	**100%**	**100%**	**100%**	88%	**99%**	**98%**
LLM	94%	**100%**	97%	98%	84%	97%	93%
MCPG	95%	**100%**	**100%**	**100%**	88%	**99%**	97%

TABLE 13 – Taux de générations respectant le schéma décrit dans l'amorce pour chaque corpus, pour chaque modèle. (Pénalité lourde exclue)

	Modèles										
Corpus	Llama2	Llama3	Mistralv0.2	Mistralv0.3	Mamba	Falcon-mamba	Qwen2.5	Parrot	Bart-paraphraser	T5-paraphraser	T5-chatGPT-paraphraser
HC-Q	74%	91%	92%	93%	23%	91%	91%	96%	84%	90%	84%
HC-S	82%	90%	90%	92%	**36%**	91%	**92%**	**99%**	86%	88%	**92%**
LLM	**85%**	**97%**	**93%**	**95%**	**36%**	95%	91%	95%	**98%**	**92%**	56%
MCPG	82%	94%	**93%**	94%	27%	**96%**	91%	97%	95%	87%	84%

TABLE 14 – Taux de générations étiquetées comme paraphrases parmi celles respectant le schéma décrit dans l'amorce pour chaque corpus, pour chaque modèle.