



HAL
open science

Reborn to be Wilde

Pedro Cabrera Ramírez, Javier Lopetegui González

► **To cite this version:**

Pedro Cabrera Ramírez, Javier Lopetegui González. Reborn to be Wilde : Une approche de transfert stylistique textuel par l'affinage d'un modèle massif de langage. *Humanistica* 2024, Association francophone des humanités numériques, May 2024, Meknès, Maroc. hal-04557633

HAL Id: hal-04557633

<https://hal.science/hal-04557633>

Submitted on 24 Apr 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

Reborn to be Wilde : une approche de transfert stylistique textuel par l’affinage d’un modèle massif de langage

Pedro Cabrera Ramírez¹ et Javier Lopetegui González²

¹LATTICE - ENS | PSL

pedro.cabrera.ramirez@ens.psl.eu

²Université Paris-Saclay

javier-alejandro.lopetegui-gonzalez@universite-paris-saclay.fr

Résumé

Cet article présente une technique de transfert stylistique dans la génération de texte via les grands modèles de langage (LLM, *Large Language Models*) génératifs, en se concentrant sur l’imitation du style d’Oscar Wilde. Grâce à l’affinage du modèle Mistral-7B-Instruct-v0.1 avec la technique QLoRA, nous réussissons à générer des textes se rapprochant du style de Wilde, soulignant l’efficacité de l’approche et l’importance du nettoyage des données. Les résultats mettent en évidence le potentiel des LLM pour la reproduction de caractéristiques stylistiques, ouvrant des voies pour des recherches futures sur l’interaction entre intelligence artificielle et style littéraire.

1 Introduction

L’un des défis majeurs de l’intelligence artificielle générative réside dans l’alignement des sorties avec les exigences ou intérêts spécifiques des utilisateurs, incluant l’ajustement du style des productions. Malgré la complexité de la notion de style, difficile à définir et analyser (Molinié, 2014; Compagnon, 2014), elle joue un rôle crucial tant dans la recherche en sciences humaines que dans des contextes industriels.

Le Transfert Stylistique (TS), qui vise à appliquer les attributs stylistiques d’un corpus à un document cible ou à influencer le mode de génération d’un modèle, est une problématique particulièrement d’actualité, surtout dans le domaine de la vision par ordinateur (Zhao, 2020). Dans ce champ disciplinaire, le TS est envisagé comme la transmission des caractéristiques stylistiques d’une image à une autre, en visant une ressemblance visuelle et une impression de provenance commune, tout en préservant le contenu sémantique original. Parallèlement, le domaine du Traitement Automatique du Langage naturel (TAL) a vu l’émergence de techniques textuelles exploitant les capacités des grands modèles de langage (LLM, *Large Language*

Models) pour des applications similaires (Jin et al., 2022; Firshman, 2023).

Cet article présente une méthode de TS appliquée à la génération de textes. Le TS est ici défini comme l’imitation de traits textuels spécifiques, appris de façon semi-supervisée, dans les *outputs* d’un LLM génératif. Nous proposons une mise en œuvre sur le modèle Mistral-7B-Instruct-v0.1 (Jiang et al., 2023) et exploitant la technique d’affinage QLoRA.

2 Problématique, objectif et méthode

En littérature, la création d’un pastiche consiste à emprunter le style d’un ensemble de textes pour en créer un nouveau, évoquant ainsi un auteur, un mouvement ou une époque spécifiques (Genette, 1982). Cette pratique, ancrée dans la tradition littéraire, a été notamment explorée par Marcel Proust dans *Pastiches et Mélanges* (1971) [1919], illustrant son importance et sa portée comme exercice d’écriture et comme moyen de confirmation des capacités techniques de leurs auteurs.

Ce genre peut être revisité avec les progrès des LLM génératifs. Peut-on entraîner un LLM à réaliser des pastiches imitant un corpus spécifique ? Pour adapter cette interrogation à une tâche réalisable et mesurable en TAL, nous la concevons comme un jeu dont l’objectif est de tromper un classificateur de texte entraîné pour identifier les écrits d’un auteur, afin qu’il accepte des textes artificiels comme authentiques. L’ambition réside dans l’affinage d’un modèle de langage pour que ses productions émulent suffisamment l’œuvre originale pour déjouer le classificateur.

L’approche adoptée pour cette expérience comprend les étapes suivantes :

1. Création de deux corpus : textes originaux de l’auteur ciblé et *prompts* pour évaluer la performance du LLM avant et après l’entraînement.

2. Collecte de premières réponses du modèle : génération initiale à partir des *prompts* pour établir un référentiel du comportement du modèle avec une instruction qui le stimule à produire un texte narratif.
3. Entraînement du classificateur d'attribution d'auteur : utilisation des œuvres de l'écrivain et des productions initiales du modèle pour créer un classificateur binaire, dont la précision est évaluée.
4. Affinage du modèle génératif : incorporation du corpus de l'auteur pour une nouvelle série de générations textuelles basées sur les *prompts*.
5. Évaluation des nouvelles sorties via le classificateur pour comparer les performances à celles obtenues initialement.

Les spécificités de l'implémentation de cette méthode et l'analyse de ses résultats seront exposées dans les sections ultérieures.

3 Corpus

Pour la réussite de l'expérience, il est impératif de différencier deux corpus : d'une part, les textes de l'auteur dont le style est à imiter, et d'autre part, les extraits conçus pour motiver la production du LLM.

3.1 Les œuvres d'Oscar Wilde

L'application du TS à la génération textuelle par un LLM peut se faire avec pratiquement tout auteur disposant d'un corpus suffisamment grand pour permettre l'affinage du modèle, bien que le seuil de suffisance de la taille du jeu de données ne soit pas scientifiquement établi, mais qui varie entre 10 000 et 100 000 exemples selon la tâche (Microsoft, 2024). En plus de ce prérequis, d'autres critères essentiels pour assurer la pertinence de l'auteur choisi pour la tâche et son analyse doivent être considérés :

1. Pour assurer la qualité des résultats et la performance optimale du modèle lors des générations, il est nécessaire de disposer d'un LLM génératif pré-entraîné et de son *tokenizer* correspondant, adaptés à la langue des textes utilisés pour l'affinage.
2. La variété linguistique de l'auteur sélectionné doit être contemporaine ou largement représentée dans le corpus d'entraînement du

modèle, afin de minimiser le besoin d'ajustements substantiels du modèle initial.

3. Autant que possible, les œuvres utilisées pour l'entraînement devraient être libres de droits d'auteur, éliminant ainsi les contraintes légales pour leur utilisation.
4. Un corpus métatextuel (Genette, 1982) significatif, composé de commentaires critiques exhaustifs (Raby, 1997), d'études historiques (Powell et Raby, 2013) et biographiques (Ellmann, 1987) et d'autres types de documents à visée analytique concernant l'œuvre de l'auteur est requis pour un examen détaillé des résultats dans les recherches futures.

Le choix de l'anglais comme langue de travail s'imposait en raison de la rareté de modèles et, en particulier, de *tokenizers* formés pour d'autres langues aptes à gérer le discours narratif. La familiarité d'un membre de l'équipe avec les axes de recherche entourant Oscar Wilde a motivé le choix de celui-ci, les caractéristiques distinctives de son idiolecte étant assez marquées pour induire une modification notable des productions. Le corpus a été constitué à partir d'œuvres disponibles sur Project Gutenberg, lesquelles ont été nettoyées et normalisées par l'élimination de caractères superflus comme les doubles espaces, les sautes de ligne et les tirets bas, ainsi que les métadonnées et les marques de format du projet bibliographique à l'aide de la bibliothèque *gutenberg_cleaner* (Kiasari, 2019). Le corpus, composé par un total de 952 802 mots, peut être considéré comme un ensemble représentatif de l'œuvre de Wilde, mais il n'est pas complet. On note notamment l'absence de la correspondance, qui est pourtant de grande ampleur.

3.2 Le catalogue des *prompts* de génération

Le second corpus, destiné à la génération, est inspiré des exercices pratiqués dans les ateliers d'écriture créative. Il consiste en une série de fragments de récits inachevés, conçus pour inviter le participant à poursuivre la narration avec ses propres idées. Cette méthode présente un double intérêt : elle s'aligne étroitement, tant dans l'esprit que dans la forme, sur les opérations attendues des LLM, et facilite l'intégration des textes de Wilde lors de la phase d'affinage grâce à sa structure modulaire.

Un ensemble initial de 300 amorces narratives a été collecté sur diverses plateformes dédiées à l'écriture créative (Story Writing Academy, 2024;

Wood, 2022; Journal Buddies, 2024; The Literacy Shed, 2024). Afin d’enrichir le corpus pour l’entraînement efficace du modèle d’attribution d’auteur, une augmentation artificielle a été réalisée grâce à ChatGPT 4, sollicité pour générer des amorces similaires à celles de notre corpus, portant ainsi le total à 500.

4 Modèle de génération et méthode de transfert stylistique

Le modèle choisi pour la génération est Mistral-7B-Instruct-v0.1, sélectionné tant pour son excellente performance sur les métriques principales que pour l’ouverture de ses poids (Jiang et al., 2023). L’instruction guidant les phases ultérieures a été formulée ainsi : `<s> [INST] These are the first lines of a work of fiction. Continue it. <TEXT> [/INST] <COMPLETION> </s>`. Pour la première série de générations, le *token* `<TEXT>` a été remplacé par chacun des *prompts* de notre collection. La section `<COMPLETION>` est laissée vide, attendant que le modèle y insère la suite du texte, comme illustré dans la Figure 1.

[<s> [INST]These are the first lines of a work of fiction. Continue it. I tiptoed into the bathroom. If anyone caught me doing this, I'd be in big trouble. I grabbed my mother's lipstick and brought it back to my bedroom where my brother slept... [/INST]I kissed him gently on the lips. I'll let him figure out what I did. I made my bed; it was the best one in the house... I didn't shower here, but I made it all the cleaner than I could. I went into the shower and wet my hair. I rubbed the lipstick all over and really smoothed it out... The water ran down my cheeks. I smiled. I smiled again. I kept smooching it out. After an hour, I knew I'd have to stop. I took myself to my bed. I'd be safe here. I put on my pajamas. I tried to feel guilty, but I didn't. I had a pretty smile on my face. I drifted off. I knew I'd be able to think it through. When I woke up, it had slipped right off. The guilt didn't come running back, either. My face was still pretty. Everything was still fine... The first time I caught myself looking in the mirror, I didn't recognize myself-I looked different. What had happened to me? Had I done something different? My face had changed, but it wasn't the first time it had happened. What do you call it? The kiss theory... I smiled at myself in the mirror, I kissed myself, I kissed myself. I kissed myself I kissed myself so, too! I kissed it and kissed it until it was like me again. Then I waited. And kissed myself more. Then I waited. And kissed myself again. I kept waiting. Then I kissed it another time... I kissed myself until the kisses were so deep and true that I was me. I held my lips to it on the sidewalk when school started. When it was over, I kissed it some more on the way home. Then I smiled at myself as I went up to my room. When I kissed myself again in the bathroom in front of the mirror, I didn't know what to do. I stood there as I watched my lips move close to my hair. I froze in shock. I had kissed something other than myself. I had kissed someone else's hair. I wanted to know what had happened to me. Had I kissed someone else? Did that mean I'd kissed someone else? My lips started to tremble. It all became fuzzy to me. I grabbed my phone. "<s>"]

FIGURE 1 – Exemple de sortie du modèle de base.

L’affinage du modèle a été réalisé avec *QLoRA* (Dettmers et al., 2023), une méthode de double quantisation à 4 bits qui optimise l’utilisation des ressources computationnelles et réduit la taille nécessaire du jeu de données à apprendre. L’entraînement a été fait sur un GPU T4 de 15GB sur Google Colab. Nous avons découpé l’intégralité de l’œuvre de Wilde en segments de 450 *tokens*. Pour chaque étape de l’entraînement, le *token* `<TEXT>` a été remplacé par le fragment *n* de Wilde, et `<COMPLETION>` par le fragment *n + 1*, parcourant ainsi la totalité du texte malgré des problèmes minimes de format. Un second ensemble de textes générés a été alors obtenu en appliquant la même instruction que lors du premier cycle de génération. Au total, nous avons construit de cette façon 1533

exemples, une quantité notablement au dessous des recommandations données par (Microsoft, 2024).

5 Modèle d’attribution d’auteur

Pour quantifier objectivement l’effet du TS sur les sorties, un modèle d’attribution d’auteur a été développé en utilisant la technique d’affinage BertAA (Fabien et al., 2020). L’objectif est de vérifier si le modèle affiné parvient à diminuer la confiance du classificateur d’attribution, ce qui indiquerait un impact réel du processus d’affinage conforme aux visées de notre expérimentation.

Le modèle sélectionné pour cette tâche est le *bert-base-cased*, complété par l’ajout d’une couche dense et de la fonction d’activation *Softmax* pour classer les auteurs. Cette technique a été implémentée grâce à la bibliothèque SimpleTransformers de Hugging Face.

6 Résultats et discussion

[<s> [INST]These are the first lines of a work of fiction. Continue it. I tiptoed into the bathroom. If anyone caught me doing this, I'd be in big trouble. I grabbed my mother's lipstick and brought it back to my bedroom where my brother slept... [/INST] I decided to give myself an old lip... I carefully pulled my lip down and smeared it with a dab of lipstick! It made my skin a sickening, pallid blue color. Then I got the idea of using it as lipstick! I started to put some of the poison that my father keeps in his room on. I decided to use it on my lips to see if it would change them. I got the lipstick and brought it to the door of the bathroom. I could hear the sound of my mother's hysterical body writhing about and gasping for air. I carefully opened the door, grabbed her hand and pressed this lipstick into her mouth. "Good night, Mother!" I said as I turned around. Then I walked back to my room. "Good night, little brother," said my brother. I smiled. Then I took some of the poison and opened my brother's lips. I heard my brother's lips say, "Good night, little sister." I flipped the light switch. When the lights went on, I could hear the sound of a gurgling fountain. I looked in the mirror. I almost fainted... [Illustration: THE LITTLE HAND WAS RIGHT] "How dare you, little girl!" said Mrs. Otis. "Put the lipstick down, I said! I said, do not put the poison on." "I'm not going to stop, Mother," said little Virginia; and she ran home again. "The wicked little monkey!" said Mother; and she threw a basin of cold water on her. "I'll show you!" said little Virginia, and she ran back to Mother's room. "I'll put poison on myself, now!" "Naughty, little girl!" cried Mother. "Stop, stop, you foolish wretch! You are quite beautiful!" "Not any more!" screamed Virginia, "and I'm not a little girl. I'm quite grown up. The whole world has grown up. I don't care any more about it!" "You'll die, Virginia! You'll die!" cried Mother. "Then let me die!" said Virginia, "but I'll make you die first." She opened the poison in her hands, and took a little piece of it to her lips... ""]

FIGURE 2 – Exemple de sortie du modèle affiné.

Le classificateur d’auteur, entraîné sur les données d’Oscar Wilde ainsi que sur les premières sorties du modèle génératif, a atteint une précision de 100% sur l’ensemble d’évaluation, classant correctement tous les textes produits par Mistral 7B. Après l’affinage du modèle génératif, la précision du classificateur s’est réduite. Seulement 45% des textes artificiellement générés ont été classés de manière correcte, démontrant une diminution significative par rapport aux résultats du modèle de base. Comme illustré par la Figure 2, les nouvelles sorties sont marquées par une amélioration sensible en termes de rythme et de cohérence narrative, ainsi que par l’empreinte d’un ton dramatique. Les répétitions constantes – dans le cas de la Figure 1 : “*I kissed myself, I kissed myself. I kissed myself!*” –, un des problèmes les plus fréquents dans les sorties des LLM (Holtzman et al., 2020) a été significativement réduit dans la Figure 2.

Les modifications observées dans les textes générés post-affinage incluent des inversions verbe-sujet (“said Mother”), et l’utilisation de noms spécifiques à l’univers de Wilde, comme Otis et Virginia de *The Canterville Ghost*. L’affinage a également reproduit certaines traces de bruit présent dans les données d’entraînement, notamment des descriptions d’illustrations dans la forme de métadonnées dans la ligne 14 de la Figure 2, un élément récurrent dans les éditions de Project Gutenberg. Ce phénomène valide l’efficacité de l’affinage, tout en soulignant l’importance cruciale d’un nettoyage approfondi des données en amont.

Plusieurs exemples confirment que les personnages sont l’un des premiers traits imités par le modèle. Dans la Figure 2, il est possible d’apprécier l’apparition d’un “Mr. Gray”, ainsi que d’un “Basil”, issus du *Portrait of Dorian Gray*, et d’un “Jack”, qui pourrait avoir comme origine le personnage homonyme de *The Importance of Being Earnest*.

[INST]These are the first lines of a work of fiction. Continue it. I tiptoed into the bathroom. If anyone caught me doing this, I'd be in big trouble. I grabbed my mother's lipstick and brought it back to my bedroom where my brother slept... /INST I decided to give myself an old lip... I carefully pulled my lip down and smeared it with a dab of lipstick. It made my skin a sickening, pallid blue color. Then I got the idea of using it as lipstick! I started to put some of the poison that my father keeps in his room on. I decided to use it on my lips to see if it would change them. I got the lipstick and brought it to the door of the bathroom. I could hear the sound of my mother's hysterical body writhing about and gasping for air. I carefully opened the door, grabbed her hand and pressed this lipstick into her mouth. "Good night, Mother!" I said as I turned around. Then I walked back to my room. "Good night, little brother," said my brother. I smiled. Then I took some of the poison and opened my brother's lips. I heard my brother's lips say, "Good night, little sister." I flipped the light switch. When the lights went on, I could hear the sound of a gurgling fountain. I looked in the mirror. I almost fainted... [Illustration: THE LITTLE HAND WAS RIGHT] "How dare you, little girl!" said Mrs. Otis. "Put the lipstick down, I said! I said, do not put the poison on." "I'm not going to stop, Mother," said little Virginia; and she ran home again. "The wicked little monkey!" said Mother; and she threw a basin of cold water on her. "I'll show you!" said little Virginia, and she ran back to Mother's room. "I'll put poison on myself, now!" "Naughty, little girl!" cried Mother. "Stop, stop, you foolish wretch! You are quite beautiful!" "Not any more!" screamed Virginia, "and I'm not a little girl. I'm quite grown up. The whole world has grown up. I don't care any more about it!" "You'll die, Virginia! You'll die!" cried Mother. "Then let me die!" said Virginia, "but I'll make you die first." She opened the poison in her hands, and took a little piece of it to her lips. , "

FIGURE 3 – Exemple de sortie du modèle affiné.

7 Conclusion

Les initiatives en TS appliquées à la génération de texte par les LLM mettent en évidence leur sensibilité à des aspects stylistiques, avec des impacts mesurables tant qualitativement que quantitativement. Toutefois, ce domaine est encore en cours de consolidation, et il est impératif de dépasser les réussites techniques pour s’attaquer à des questions fondamentales sur le fonctionnement interne des LLM. Il est essentiel de s’interroger sur les éléments stylistiques effectivement reproduits par le LLM, les caractéristiques favorisées dans le transfert, et comment ces observations peuvent contribuer à la recherche en stylistique.

Bibliographie

- Antoine Compagnon. 2014. *Le style*. In *Le Démon de la théorie. Littérature et sens commun*, pages 195–230. Éditions du Seuil, Paris.
- Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, et Luke Zettlemoyer. 2023. *Qlora: Efficient finetuning of quantized llms*.
- Richard Ellmann. 1987. *Oscar Wilde*. Alfred A. Knopf, New York.
- Maël Fabien, Esau Villatoro-Tello, Petr Motlicek, et Shantipriya Parida. 2020. *Bertaa: Bert fine-tuning for authorship attribution*. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, Indian InstFirshmanitute of Technology Patna, India. NLP Association of India (NLP AI).
- Ben Firshman. 2023. *Fine-tune llama to speak like homer simpson*. <https://replicate.com/blog/fine-tune-llama-to-speak-like-homer-simpson>.
- Gérard Genette. 1982. *Palimpsestes : la littérature au second degré*. Éditions du Seuil, Paris.
- Ari Holtzman, Jan Buys, Li Du, et Yejin Forbes, Maxwell et Choi. 2020. *The curious case of neural text degeneration*.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Devendra Bamford, Chris Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le, Thibaut Lavril, Thomas Wang, Timothée Lacroix, et William El Sayed. 2023. *Mistral 7b*.
- Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, et Rada Mihalcea. 2022. *Deep learning for text style transfer: A survey*. *Computational Linguistics*, 48(1).
- Journal Buddies. 2024. *51 super story starter sentences*. <https://www.journalbuddies.com/creative-writing-2/story-starter-sentences/>. Date inconnue, consulté le 5 janvier 2024.
- Peyman Mohseni Kiasari. 2019. *gutenberg_cleaner*.
- Microsoft. 2024. *Getting started with llm fine-tuning*. In *The Artificial Intelligence (AI) Playbook*. Microsoft.
- Georges Molinié. 2014. *La stylistique*. Quadrige. Presses Universitaires de France.
- Kerry Powell et Peter Raby, éditeurs. 2013. *Oscar Wilde in Context*. Literature in Context. Cambridge University Press.
- Marcel Proust. 1971. *Contre Sainte-Beuve, précédé de Pastiches et Mélanges et suivi de Essais et Articles*. Bibliothèque de la Pléiade. Gallimard.
- Peter Raby, éditeur. 1997. *The Cambridge Companion to Oscar Wilde*, illustrated, reprinted édition. Cambridge Companions to Literature. Cambridge University Press.
- Story Writing Academy. 2024. *100 creative writing prompts for middle school*. <https://storywritingacademy.com/creative-writing-prompts-middle-school/>. Date inconnue : consulté le 5 janvier 2024.

- The Literacy Shed. 2024. The story starter shed. <https://www.literacyshed.com/story-starters.html>. Date inconnue, consulté le 5 janvier 2024.
- Moe Wood. 2022. 100 short story (or novel) writing prompts. <https://hobbylark.com/writing/100-short-story-novel-prompts>. Consulté le 5 janvier 2024.
- Changshen Zhao. 2020. A survey on image style transfer approaches using deep learning. In *Journal of Physics : Conference Series*, volume 1453, Xi'an, China. IOP Publishing Ltd.