Amélioration de l'accessibilité des textes dans l'enseignement primaire : une étude comparative entre la simplification par LLM et la réécriture par des experts

Anya Amel Nait Djoudi[1] Feriel Hannachi [2] Patrice Bellot[1] Ismail Badache[1]
Adrian-Gabriel Chifu[1]
(1) Aix-Marseille Université, CNRS, LIS, Marseille, France
(2) LARI Laboratory, Mouloud Mammeri University, Tizi Ouzou, Algeria
anya.nait-djoudi@lis-lab.fr, feriel.hannachi@ummto.dz
{patrice.bellot ; ismail.badache ; adrian.chifu}@lis-lab.fr

RÉSUMÉ

Les difficultés de lecture, particulièrement chez les élèves dyslexiques, constituent un frein majeur à l'équité pédagogique. Si la simplification manuelle des supports pédagogiques favorise la réussite scolaire, son déploiement reste limité par l'expertise multidisciplinaire et le temps qu'elle exige. Cette étude évalue le potentiel du modèle LearnLM, optimisé pour l'éducation, pour automatiser la simplification de manuels scolaires français (CE1-CM2). Dans notre étude, nous comparons plusieurs stratégies de prompting sur un corpus de 100 paires de textes experts. Les résultats, mesurés par les scores FRE, SARI et BERTScore, démontrent qu'une approche Few-shot intégrant seulement deux exemples experts permet à LearnLM d'imiter la logique de transformation humaine et de préserver l'intégrité sémantique. Ce travail souligne l'efficacité des LLM comme outils d'assistance pour générer massivement des ressources inclusives, palliant ainsi la rareté des supports adaptés manuellement.

ABSTRACT

Improving Text Accessibility in Primary Education : A Comparative Study Between LLM-Based Simplification and Expert Rewriting

Reading comprehension difficulties, including dyslexia, significantly impact on equal access to education for primary school learners'. These difficulties hinder their academic success in many areas. While manually adapting teaching resources has proven beneficial for learners facing these obstacles, its widespread adoption is limited by the diverse skills required and the considerable time involved. This study examines whether LearnLM, an LLM designed for educational applications, can effectively automate the process of adapting French textbooks (for grades 2 and 5) into more accessible formats. To this end, we evaluate several configurations on a corpus of 100 expert text pairs. Evaluation metrics including FRE, SARI and BERTScore reveal that a few-shot approach utilizing two reference examples enables LearnLM to capture the patterns employed by human experts and maintain semantic fidelity. These findings underscore how large language models can serve as valuable assistants for rapidly producing inclusive educational materials at scale, thereby alleviating the shortage of appropriately simplified resources for learners with reading comprehension challenges.

MOTS-CLÉS : Simplification des textes, Education, Simplification automatique, LLM, lisibilité.

KEYWORDS: Text Simplification, Education, Automatic Simplification, LLM , Readability.

1 Introduction

La maîtrise de la lecture est un pilier de la réussite scolaire, impliquant des processus cognitifs complexes essentiels à la compréhension des textes dans toutes les disciplines (Delgadova, 2015; Gala *et al.*, 2020). Cependant, le déclin inquiétant des compétences en lecture à l'échelle mondiale (Hedlin *et al.*, 2025; Mullis *et al.*, 2017), combiné à l'augmentation du nombre d'élèves souffrant de troubles de la lecture tels que la dyslexie [1] (Hedlin *et al.*, 2025) constitue un obstacle majeur pour de nombreux apprenants. Les difficultés de lecture entravent la dynamique d'apprentissage positive où la pratique régulière consolide les compétences en lecture, favorisant une meilleure compréhension (Stanovich *et al.*, 1986). Sans ce renforcement, les enfants confrontés à des difficultés de lecture prennent de plus en plus de retard, en particulier à mesure que la complexité des textes augmente à chaque niveau scolaire (Tunmer & Hoover, 2019). Ces obstacles soulignent le besoin urgent de stratégies pratiques visant à améliorer l'accessibilité des textes éducatifs pour tous les apprenants.

La simplification de texte (TS) : processus consistant à modifier des textes pour les rendre plus lisibles tout en préservant leur sens (Siddharthan, 2014; Shardlow, 2014) offre une piste prometteuse pour combler cette lacune. En réduisant la complexité syntaxique et lexicale, la simplification diminue la charge cognitive et facilite la compréhension pour les lecteurs en difficulté (Gala *et al.*, 2020), ce qui s'avère particulièrement utile pour les élèves dyslexiques ou présentant des difficultés de lecture (Rello *et al.*, 2013). Si la simplification de textes par les grands modèles de langage (LLM) a déjà prouvé son efficacité dans le secteur biomédical, tant en anglais qu'en français (Nait Djoudi *et al.*, 2025; Cardon & Grabar, 2020; Haver *et al.*, 2024), son application aux supports pédagogiques demeure limitée. Actuellement, cette approche n'a été explorée qu'en suédois et en anglais (Hedlin *et al.*, 2025; Mo & Hu, 2024), laissant un immense vide pour les contextes francophones. Ce retard est d'autant plus marqué que les méthodes existantes pour le français reposent encore soit sur une simplification manuelle (Gala *et al.*, 2020) processus fastidieux, coûteux et exigeant une expertise multidisciplinaire soit sur des approches automatiques plus anciennes (Todirascu *et al.*, 2022). Ces dernières, souvent opaques, ne tirent pas profit des récentes avancées de l'IA générative. En définitif, cette absence de solutions modernes basées sur les LLM pour les ressources pédagogiques francophones constitue une occasion manquée d'améliorer significativement l'accessibilité pour les lecteurs en difficulté.

Pour pallier ce problème, la présente étude examine le potentiel des grands modèles de langage (LLM) à simplifier automatiquement des textes pédagogiques en français destinés aux enfants dyslexiques ou ayant des difficultés en lecture. En comparant systématiquement les versions simplifiées générées par les LLM avec celles rédigées par des experts, nous souhaitons fournir aux enseignants, aux parents et aux élèves des outils pratiques et évolutifs permettant de créer des supports pédagogiques accessibles, afin de favoriser à terme des pratiques éducatives inclusives qui améliorent la compréhension de tous les apprenants, y compris ceux ayant des difficultés en lecture.

Cette étude s'articule autour de trois questions de recherche :

— **QR1 :** Dans quelle mesure la simplification basée sur les LLM améliore-t-elle la lisibilité et la simplicité linguistique par rapport aux manuels scolaires originaux ?
— **QR2 :** Comment la simplification basée sur les LLM se compare-t-elle aux versions de référence produites par une équipe multidisciplinaire d'experts ?
— **QR3 :** Le processus de simplification automatisé préserve-t-il le sens du contenu d'origine ?

1. https://nces.ed.gov/programs/coe/pdf/2024/CGG_508c.pdf

2 Etat de l'art : une vue d'ensemble

Les précédentes tentatives de simplification des textes reposaient sur des manipulations manuelles à plusieurs niveaux linguistiques (Gala *et al.*, 2018; Green & Hawkey, 2012; Saggion, 2017). Il s'agissait notamment de substitutions lexicales (remplacement de mots complexes par des synonymes plus simples), d'ajustements morphologiques (simplification des formes verbales ou suppression des diminutifs) et de restructuration syntaxique (transformation de la voix passive en voix active, reconstruction de la structure des phrases). D'autres techniques impliquaient des changements au niveau du discours (résolution des anaphores), la suppression des redondances et de certains mots, et le raccourcissement du texte pour en améliorer la compréhension. Les systèmes de simplification automatique de texte (ATS) ont considérablement évolué pour aider les enseignants à adapter le matériel pédagogique aux différents besoins des élèves (Liu *et al.*, 2024). À l'aide d'algorithmes issus du traitement automatique de la langue (TAL) et la linguistique computationnelle, des outils ont été développés pour aider les enseignants à extraire les caractéristiques du texte, à évaluer sa complexité et à identifier les passages difficiles (Crossley *et al.*, 2007; Jin *et al.*, 2020).

L'adaptation du matériel pédagogique s'appuie aujourd'hui sur des approches complémentaires. D'une part, les Ressources Éducatives Libres (REL) (Bliss *et al.*, 2013; Farrow *et al.*, 2015) favorisent la création de textes adaptés au niveau des élèves. D'autre part, les outils basés sur l'IA (Badache & Bellet, 2024; Badache & Colombo, 2025) facilitent la planification des cours tout en permettant de remplacer les mots difficiles, de restructurer la syntaxe et de générer des textes de complexité variée (Eskenazi *et al.*, 2013). Au-delà de ces opérations superficielles, certains systèmes sont également capables d'effectuer des simplifications de plus haut niveau, telles que l'ajout d'élaborations et la production de textes dont le niveau de complexité est contrôlable (Maddela *et al.*, 2021; Martin *et al.*, 2020). S'appuyant sur ces bases, les progrès récents dans le domaine des grands modèles linguistiques (LLM) ont conduit au développement de systèmes basés sur des invites et d'agents pilotés par des LLM qui permettent des stratégies de simplification fines, adaptées à des contextes éducatifs et à des profils d'apprenants spécifiques. (Hedlin *et al.*, 2025) montre que ChatGPT-4 (Achiam *et al.*, 2023), en particulier avec le méta-prompting, améliore considérablement la lisibilité des textes universitaires suédois à travers différentes stratégies d'élicitation, tandis que ExpertEase (Mo & Hu, 2024) de Mo et Hu et AgentSimp (Fang *et al.*, 2025) de Fang utilisent des cadres multi-agents pour simuler la collaboration entre experts, enseignants et étudiants afin de simplifier les supports pédagogiques en anglais en fonction du niveau scolaire. En revanche, notre recherche comble une lacune dans le contexte éducatif français en exploitant LearnLM 1.5 (Team *et al.*, 2024), un modèle explicitement développé pour des applications éducatives. Basé sur Gemini 1.5 Pro et perfectionné grâce à un ajustement supervisé et à un apprentissage par renforcement à partir des commentaires humains, LearnLM recadre le comportement pédagogique en tant que « suivi des instructions », permettant aux enseignants d'encoder les comportements souhaités via des invites du système. Les évaluations d'experts indiquent que LearnLM surpasse des modèles tels que GPT-4o, Claude 3.5 et la base Gemini 1.5 Pro tant en termes de respect des instructions que d'adaptabilité des apprenants, ce qui le rend idéal pour explorer des techniques personnalisées de simplification de textes français.

Dans le contexte de cet article, nous définissons une "simplification réussie" non pas comme une simple réduction de la longueur des phrases, mais comme une adaptation répondant aux principes de la Conception Universelle de l'Apprentissage (CUA) (Rose & Meyer, 2002). L'objectif est de minimiser la charge cognitive extrinsèque liée au déchiffrage (Sweller, 1988) tout en préservant strictement l'intégrité conceptuelle du texte source. Comme le démontrent (Rello *et al.*, 2013) dans leurs travaux sur la dyslexie, une simplification pertinente nécessite de remplacer les mots peu fréquents par des synonymes courants et d'expliciter les structures syntaxiques, sans appauvrir le contenu informatif.

3 Méthodologie

3.1 Corpus

Nous utilisons le corpus ALECTOR (Gala *et al.*, 2020), un ensemble de 52 704 tokens conçu pour la recherche en simplification de textes destinés aux lecteurs en difficulté. Il regroupe 100 paires de textes (sources et simplifiés) de genres littéraires et scientifiques, initialement destinés aux élèves de 7 à 11 ans (du `CE1` au `CM2`)[2]. Les simplifications, réalisées par des experts pluridisciplinaires (éducation, linguistique, orthophonie), traitent les dimensions lexicale, morphologique, syntaxique et discursive. Il inclut en outre la version française des textes standardisés `IReST` (International Reading Speed Texts), dédiés à la mesure de la vitesse de lecture (Trauzettel-Klosinski *et al.*, 2012). La distribution complète du corpus est illustrée en Tab. 1. Pour notre étude, nous exploitons 98 paires de textes, les deux restantes ayant été réservées aux exemples de notre stratégie de prompting few-shot.

	CE1	CE2	CM1	CM2	IReST
SCI	10	10	10	11	9
LIT	15	14	10	10	1
Total	25	24	20	21	10

TABLE 1 – Nombre de textes par niveau et par genre

3.2 Pipeline de simplification automatisé

3.2.1 LearnLM

Afin d'améliorer l'accessibilité des manuels scolaires français du primaire, nous avons utilisé le modèle LearnLM, un modèle d'IA générative spécialisé développé par Google DeepMind. Au moment de cette expérimentation, **`learnlm-2.0-flash-experimental`** constituait la dernière version de la série LearnLM, spécialement optimisée pour le contexte éducatif et pédagogique. L'accès au modèle s'est fait via l'API Gemini[3].

Le choix de ce modèle s'explique par son orientation architecturale et son efficacité pédagogique avérée. LearnLM-Tutor est une version supervisée et affinée (SFT) de l'architecture Gemini, optimisée pour les applications éducatives grâce à un processus participatif axé sur l'évaluation qui intègre les connaissances issues des sciences de l'apprentissage, des enseignants et des apprenants (Jurenka *et al.*, 2024). Des évaluations comparatives récentes démontrent que LearnLM surpasse largement les modèles à usage général tels que GPT-4o[4], Claude 3.5 Sonnet[5] et Gemini 1.5 Pro[6] (Reid *et al.*, 2024), sur des dimensions pédagogiques directement liées à la simplification du texte. Plus précisément, le modèle excelle dans deux capacités essentielles : la gestion de la charge cognitive, qui consiste à décomposer des concepts complexes en segments gérables et digestibles, et le nivellement affiné, qui permet au modèle d'ajuster à la fois la complexité linguistique et conceptuelle pour correspondre précisément aux niveaux de compétence spécifiques des apprenants. Les experts ont systématiquement jugé que les résultats de LearnLM étaient en parfaite adéquation avec les pratiques pédagogiques fondées sur des données probantes. Alors que les premières évaluations se sont concentrées sur les interactions de tutorat, ses atouts intrinsèques en matière de gestion adaptative de la complexité en font un outil puissant pour générer des supports pédagogiques accessibles.

2. Disponible sur : https://alectorsite.wordpress.com/ ; les données `CM2` ont été fournies par les auteurs.
3. API Gemini
4. GPT-4o (version 2024-08-06), OpenAI Documentation.
5. Claude 3.5 Sonnet (version 2024-06-20), Anthropic Documentation.
6. Gemini 1.5 Pro-002 (version 2024-09-24), Google Cloud Vertex AI Documentation.

3.3 Stratégies de prompting

Afin d'évaluer l'efficacité de différentes approches d'ingénierie de prompt dans le cadre de la simplification de texte, nous avons sélectionné quatre stratégies de prompting distinctes : Standard, Few-Shot, Role Play et Chain-of-Thought (CoT). Ce choix s'appuie sur des recherches antérieures démontrant que les approches de prompting influencent les résultats des modèles linguistiques (Hedlin *et al.*, 2025; Knoth *et al.*, 2024), ainsi que sur les exigences cognitives variables que chaque stratégie impose au modèle pour guider les processus de simplification.

Chacune des quatre stratégies a été mise en œuvre avec deux variantes de prompt afin de vérifier si des différences subtiles dans le cadrage et l'orientation affectaient les résultats de la simplification.

1. **Standard** : Une instruction standard fournit une consigne directe et sans ambiguïté, sans aide supplémentaire, et sert de référence pour la simplification du texte (Hedlin *et al.*, 2025). Nous examinons deux variantes qui diffèrent par leur niveau de précision. La première repose sur une consigne minimale, laissant au modèle la liberté de déterminer comment simplifier le texte. La seconde introduit des contraintes plus explicites pour guider le processus de transformation (par exemple, en contrôlant la fidélité, la structure et la couverture informationnelle).
2. **Few-Shot** : L'utilisation de prompts few-shot complète l'instruction par des paires source-cible exemplaires, offrant ainsi des démonstrations explicites de la tâche de simplification (Brown *et al.*, 2020). Nous examinons deux variantes qui diffèrent par le nombre et la diversité des exemples. La première comprend un seul exemple au niveau scolaire le plus bas (`CE1`). La seconde ajoute un exemple de niveau supérieur (`CM2`), augmentant ainsi la variabilité des transformations démontrées. Les deux variantes s'appuient sur la prompt standard contrainte. Cette conception nous permet d'évaluer si l'augmentation de la quantité et de la diversité des exemples améliore la cohérence ou entraîne des rendements décroissants.
3. **Role Play** : La consigne de jeu de rôle (Role Play) attribue au modèle un personnage en rapport avec la tâche, le présentant comme un expert en rédaction de textes accessibles (Anderson, 1983; Kong *et al.*, 2023). Nous nous appuyons sur la consigne standard restreinte en introduisant une description de rôle qui positionne le modèle comme un instituteur français ayant de l'expérience dans l'accompagnement d'élèves en difficulté de lecture, ainsi qu'un public cible explicite défini par l'âge. Nous envisageons deux variantes qui diffèrent dans la manière dont la difficulté de lecture est spécifiée. La première cible les enfants présentant des difficultés générales de lecture, tandis que la seconde fait explicitement référence à la dyslexie. Cette conception nous permet d'examiner si le fait d'affiner le profil de l'apprenant dans la spécification du rôle conduit à des simplifications plus adaptées et plus fondées sur le plan pédagogique.
4. **Chain-of-Thought (CoT)** : Les prompts de type Chain-of-Though demandent explicitement au modèle de décomposer la tâche de simplification en étapes séquentielles et d'exposer son processus de raisonnement (Wei *et al.*, 2022). Plutôt que de générer directement un résultat simplifié, le modèle est guidé pour identifier et justifier les modifications linguistiques en fonction de l'objectif de simplification. Deux variantes du CoT ont été développées, qui se distinguent par leur degré de soutien pédagogique. La première variante, mise en œuvre dans la consigne 7, fournit des indications procédurales explicites sous forme d'instructions détaillées prescrivant des transformations linguistiques spécifiques à effectuer, accompagnées d'exemples concrets. La deuxième variante, mise en œuvre dans la consigne 8, comprend également des exemples, mais inscrit les instructions dans des catégories linguistiques explicitement définies : phénomènes lexicaux, morphologiques, syntaxiques et au niveau du

discours. Si les deux consignes visent des objectifs de simplification similaires, elles diffèrent dans la manière dont elles présentent ces objectifs : la consigne 7 propose des règles procédurales directes (par exemple,"Remplacez les mots difficiles ou peu courants par des mots plus simples. Par exemple : “volumineux” devient “gros”"), tandis que la consigne 8 contextualise la même transformation au sein d’une compréhension linguistique plus large (par exemple, "Effectuez des remplacements lexicaux en identifiant les mots difficiles ou peu courants, en particulier ceux qui sont longs, peu fréquents, ont une orthographe irrégulière ou des structures syllabiques complexes, et remplacez-les par des synonymes plus simples et plus courants. Par exemple, remplacez “volumineux” devient “gros” "). Cette conception comparative permet de déterminer si un soutien procédural explicite ou des conseils implicites basés sur des catégories linguistiques orientent plus efficacement le modèle vers des simplifications motivées par des considérations linguistiques, en établissant des parallèles avec les protocoles d’annotation d’experts utilisés dans la construction du corpus de simplification ALECTOR.

L’utilisation de consignes en anglais sur un corpus français (cross-lingual prompting) est un choix méthodologique délibéré. L’apprentissage par renforcement (RLHF), qui garantit le respect des instructions complexes, reste historiquement optimisé pour l’anglais (Ouyang *et al.*, 2022). Ce recours à l’anglais comme langue de contrôle isole l’instruction des données traitées, réduisant ainsi les hallucinations et les fuites d’instructions (instruction leakage) (Perez & Ribeiro, 2022). De plus, la fonction tutorat de LearnLM (Team *et al.*, 2024) permet une adaptation granulaire (fine-grained leveling) pour ajuster automatiquement le niveau lexical et syntaxique au profil cognitif de l’apprenant.

3.4 Métriques d’évaluation

Afin d’évaluer les performances du processus de simplification, nous adoptons une approche d’évaluation multidimensionnelle axée sur la lisibilité, la simplicité et la préservation du sens.

3.4.1 Lisibilité

La lisibilité est évaluée à l’aide du score de Kandel-Moles (équation 1) , une adaptation française de l’indice de lisibilité de Flesch (Challener *et al.*, 2025). Utilisé par (Cardon & Grabar, 2020), pour évaluer la lisibilité de textes biomédicaux français après simplification, cet indicateur repose sur des caractéristiques superficielles telles que la longueur des phrases et le nombre de syllabes. Bien qu’il ne mesure pas directement la qualité sémantique ou transformationnelle, il montre à quel point les phrases obtenues sont faciles à lire, ce qui est précisément l’objectif principal de la simplification. L’indice de Kandel-Moles est calculé selon la formule suivante :

$$\text{Score} = 207 - (1.015 \times \text{ASL}) - (73.6 \times \text{ASW}) \tag{1}$$

Où ASL average sentence length, correspond à la longueur moyenne des phrases (nombre de mots par phrase) et ASW average number of syllables per word, correspond au nombre moyen de syllabes par mot.

Le score de Kandel-Moles est calculé à l’aide de la bibliothèque TextSat [7] et désigné par l’abréviation FRE , en référence à la formule anglaise Flesh Reading Ease (FRE) (Flesch, 1948), dans la suite de cette étude.

7. https://github.com/textstat/textstat

3.4.2 Simplicité

Pour évaluer les performances du modèle LearnLM dans la génération de textes simplifiés, nous utilisons la métrique SARI. Notre adoption de SARI s'inscrit dans la lignée de l'approche de (Cardon & Grabar, 2020), qui l'a utilisée dans le contexte de la simplification de phrases en français dans le domaine biomédical. SARI quantifie la précision d'opérations d'édition spécifiques : ajouts, suppressions et conservations. En comparant les résultats du modèle à la fois au texte source original et à des simplification de référence , SARI fournit une évaluation nuancée de l'efficacité avec laquelle le modèle transforme une syntaxe complexe en formes simplifiées tout en conservant les informations essentielles.

3.4.3 Similarité sémantique

Afin de garantir que les simplifications générées par le modèle LearnLM préservent le sens essentiel du texte original, nous utilisons BERTScore (Zhang *et al.*, 2019). BERTScore évalue la similarité sémantique en exploitant des plongements contextuels pour calculer des alignements au niveau des tokens, offrant ainsi une évaluation fiable de la précision et du rappel. Notre utilisation de cette métrique s'inspire des travaux de (Rennard *et al.*, 2023), qui a utilisé BERTScore pour l'évaluation de la synthèse de dialogues en français. Alors que (Rennard *et al.*, 2023) comparait les résumés générés aux documents sources originaux, nous étendons cette application à la tâche de simplification en comparant la sortie du modèle à la fois aux textes sources originaux et aux références de référence rédigées par des experts. Nous avons utilisé le modèle **`bert-base-multilingual-cased`** sans redimensionnement de référence ; l'implémentation originale de ce modèle est disponible ici [8].

4 Résultats

Cette section évalue les simplifications automatiques générées à l'aide de LearnLM, en utilisant les indicateurs détaillés dans la section 3.4, en mettant particulièrement l'accent sur la lisibilité, la simplicité et la similarité sémantique. Notre analyse suit une approche descendante : nous procédons d'abord à une évaluation globale afin d'identifier la stratégie de suggestion la plus pertinente et la ou les suggestions les plus performantes. Nous affinons ensuite ces résultats en examinant les variations de performance entre les niveaux scolaires (`CE1`, `CE2`, `CM1`, `CM2`) et de l'`IReST`[9], ainsi qu'en fonction du genre textuel : scientifique (SCI) ou littéraire (LIT). Afin de tester la robustesse de ces stratégies, nous incluons également le corpus `IReST` (Trauzettel-Klosinski *et al.*, 2012), un ensemble de paragraphes standardisés calibrés pour un niveau de lecture de 6^{e} `année` . Ce sous-ensemble sert de référence hors niveau, nous permettant d'évaluer si les gains de simplification se généralisent au-delà du programme scolaire primaire couvert par le corpus ALECTOR.

4.1 Lisibilité et Simplification

4.1.1 Performance globale par stratégie

Les performances de simplification, évaluées à l'aide des scores FRE (lisibilité) et SARI (qualité de simplification), sont présentées dans la figure 1. L'analyse de la ***lisibilité*** indique que toutes les stratégies d'invite ($S1$–$S4$) obtiennent des scores FRE supérieurs à ceux de l'original ($86, 71$) et de

8. https://github.com/Tiiiger/bert_score/
9. International Reading Speed Texts, page officielle ORVIS/IReST.

la référence humaine $(93, 76)$. La stratégie « Jeu de rôle » ($S3$) présente la moyenne la plus élevée $(97, 81)$, grâce à l'invite P6 $(98, 08)$. Bien que le score individuel le plus élevé soit observé pour l'invite P1 $(98, 26)$, celle-ci présente une variance intra-stratégie légèrement supérieure à celle de $S3$. La stratégie « Chaîne de pensée » ($S4$) présente la moyenne la plus faible $(96, 58)$, notamment en raison de la consigne P8 $(95, 55)$. En ce qui concerne la ***simplification*** telle que mesurée par le score SARI, le classement des performances s'inverse. La stratégie $S4$ s'avère la plus efficace (moyenne = $41, 49$), avec un pic pour la consigne P8 $(42, 46)$, tandis que $S1$ enregistre les résultats les plus faibles (moyenne = $35, 83$). Ce contraste entre le FRE et le SARI est particulièrement flagrant lorsqu'on compare les extrêmes : P1 optimise la lisibilité au détriment de la simplification $(34, 35)$, tandis que P8 présente le profil inverse. Entre ces deux extrêmes, la consigne P4 apparaît comme un compromis optimal, alliant un FRE solide $(97, 54)$ à un SARI élevé $(41, 69)$. Sur la base de ces résultats, nous allons restreindre notre analyse à trois configurations clés : P1 (FRE maximal), P8 (SARI maximal) et P4 (meilleur équilibre). Notre analyse suivante explore ces configurations sous l'angle du niveau scolaire et du genre.

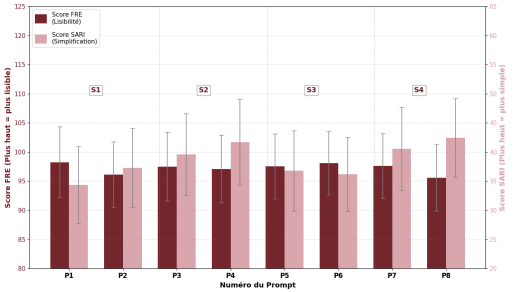

FIGURE 1 – Comparaison des scores FRE et SARI par Prompt et Stratégie

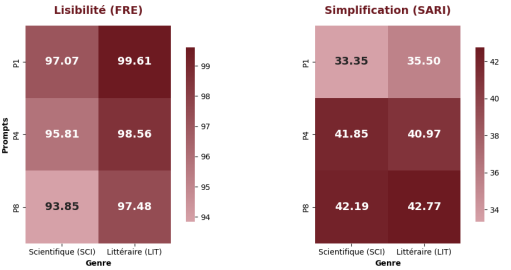

FIGURE 2 – Comparaison des scores FRE et SARI par Prompt et genre

4.1.2 Performance selon le niveau scolaire

L'analyse par niveau scolaire, détaillée dans les tableaux 2 et 3 qui reflètent respectivement la lisibilité (FRE) et la qualité de simplification (SARI), révèle des tendances distinctes pour les consignes P1, P4 et P8. En ce qui concerne la ***lisibilité (FRE)***, les données du tableau 2 montrent une tendance générale à la baisse des scores à mesure que le niveau scolaire augmente (de la 2^{e} à la 5^{e} année qui correspondent aux niveaux : (`CE1`, `CE2`, `CM1`, `CM2`)). Cependant, une incohérence apparaît entre la `CE2` et la `CM1` année, cette dernière affichant paradoxalement des scores plus élevés (par exemple, P1 passe de $98, 47$ à $98, 81$), ce qui suggère que les modèles ont du mal à distinguer la transition entre les cycles d'apprentissage. En ce qui concerne la ***simplification (SARI)***, le tableau 3 suggère que la progression est plus cohérente, bien qu'un renversement notable se produise en `CM2` année : ce niveau, bien qu'il présente le FRE le plus bas, obtient les scores SARI les plus élevés, dépassant même ceux de la `CE1` année. La consigne P8 se distingue ici avec une performance maximale $(45, 90)$. Le corpus `IReST` affiche des scores FRE élevés (par exemple, 95,78 pour le niveau P8), ce qui place sa facilité de lecture à un niveau comparable à celui des simplifications destinées au `CE1` plutôt qu'à celui des textes plus complexes du `CM1`. Cependant, en termes de qualité de simplification (tableau 3), le corpus `IReST` représente la référence ultime avec un SARI moyen de 49,72, surpassant largement toutes les catégories de niveau scolaire. Enfin, alors que P4 représentait le meilleur compromis global, l'analyse granulaire montre que P8 offre une plus grande stabilité, réduisant l'écart entre la lisibilité

et la qualité de la simplification à tous les niveaux scolaires, contrairement à P1, qui ne parvient pas à maintenir un score SARI satisfaisant sur les textes de quatrième et cinquième année.

Cat.	Src	CE1	CE2	CM1	CM2	IREST
Réf.	ORIG	91.97±6.74	88.62±6.94	86.02±9.17	79.78±10.90	86.88±4.07
	SIMP	98.56±6.41	93.08±4.74	91.53±7.62	90.59±10.50	96.58±4.95
S1	P1	100.99±5.02	98.47±4.30	98.81±5.95	94.98±7.86	97.74±5.87
	P2	97.27±5.41	96.24±4.24	96.68±3.97	92.89±8.15	98.64±3.51
	Moy.	99.13	97.35	97.75	93.94	98.19
S2	P3	99.55±5.52	97.66±3.56	98.87±3.87	92.74±8.32	99.70±3.47
	P4	99.88±4.91	97.25±4.25	97.98±4.02	93.09±8.13	97.39±4.80
	Moy.	99.72	97.46	98.42	92.92	98.55
S3	P5	100.35±4.02	97.38±3.84	98.59±3.22	92.61±7.91	100.09±3.31
	P6	100.77±4.11	97.75±3.67	99.22±3.48	93.31±7.54	100.73±3.53
	Moy.	**100.56**	97.56	98.90	92.96	100.41
S4	P7	99.38±5.34	97.75±3.56	98.91±3.67	93.83±8.34	98.68±3.30
	P8	96.39±6.71	96.16±4.15	96.84±4.27	92.57±7.38	95.78±4.23
	Moy.	98.79	96.96	97.87	93.20	97.23

TABLE 2 – Scores FRE ↑ par niveau scolaire (moyenne ± écart-type). Plus élevé = plus lisible.

Cat.	Prompt	CE1	CE2	CM1	CM2	IREST
S1	P1	37.02±7.20	32.23±5.28	29.72±4.37	35.75±6.14	40.58±4.87
	P2	36.81±7.05	34.72±4.94	35.02±7.84	39.90±4.38	43.86±6.38
	Moy.	36.92	33.47	32.37	37.82	42.22
S2	P3	40.42±8.37	37.21±5.80	36.14±4.96	40.69±4.77	48.25±6.86
	P4	40.76±8.57	38.72±6.18	38.99±5.32	45.48±6.19	48.45±7.16
	Moy.	40.59	37.97	37.57	43.09	48.35
S3	P5	36.74±8.27	34.02±5.24	33.07±4.95	40.48±5.32	43.45±6.16
	P6	37.24±7.22	33.16±5.51	32.90±4.66	39.51±4.44	41.01±6.48
	Moy.	37.00	33.59	32.98	39.99	42.23
S4	P7	39.99±7.90	38.28±6.58	37.13±4.80	42.81±5.69	49.13±5.81
	P8	41.21±7.11	40.09±5.50	39.19±4.61	45.90±5.43	50.31±6.48
	Moy.	40.60	39.18	38.16	44.36	**49.72**

TABLE 3 – Scores SARI ↑ par niveau scolaire (moyenne ± écart-type). Plus élevé = meilleure simplification.

4.1.3 Performance selon le genre

L'analyse par genre (fig. 2) révèle un écart marqué entre la lisibilité et l'effort requis pour la simplification. En ce qui concerne la ***lisibilité (FRE)***, les textes littéraires (LIT) surpassent systématiquement les textes scientifiques (SCI), le prompt P1 maximisant ce score pour les deux catégories (respectivement $99,61$ et $97,07$). À l'inverse, la qualité de la ***simplification (SARI)*** est dominée par la consigne P8, qui affiche des performances élevées et constantes dans les deux domaines ($42,77$ pour LIT et $42,19$ pour SCI). Le prompt P4 confirme son rôle de compromis efficace : il maintient une lisibilité élevée tout en offrant un score SARI compétitif ($41,85$ en SCI), surpassant clairement P1 sur ce dernier critère. En résumé, alors que P1 privilégie la lisibilité, P8 maximise la transformation structurelle, et P4 offre le meilleur équilibre entre ces deux exigences, en particulier pour le genre scientifique.

4.2 Similarité sémantique

4.2.1 Performance globale

Il est essentiel d'évaluer la préservation sémantique afin de s'assurer que la simplification ne dénature pas l'information d'origine. La figure 3 illustre les performances sémantiques de différentes invites à l'aide du BERTScore. En termes de ***fidélité au texte source (AutoORIG)***, les LLM préservent efficacement le contenu d'origine avec des scores allant de $0,82$ à $0,87$, la stratégie S4 (CoT) atteignant les meilleures performances. La consigne P8 se distingue particulièrement avec un score de $0,87$, dépassant même la similarité sémantique existante entre la référence humaine et l'original ($0,86$). En ce qui concerne ***l'alignement avec les normes d'experts (AutoSIMP)***, P8 obtient le meilleur résultat ($0,89$), confirmant sa supériorité dans la génération de simplifications conformes aux normes d'experts. À l'inverse, bien que P1 affiche de bons résultats en termes de lisibilité (fig. 1), il présente l'alignement sémantique le plus faible ($BERTScore_{AutoSIMP} = 0,79$), s'écartant davantage de la structure de référence. En résumé, P4 s'impose comme la consigne la plus équilibrée, offrant une simplification qui préserve strictement l'intégrité sémantique de l'original tout en s'alignant sur le travail des experts ; cependant, en moyenne, P4 reste sémantiquement plus proche du texte original que de la simplification produite par les experts ($BERTScore_{AutoSIMP} = 0,84$). De manière générale, nous observons que toutes les invites produisent systématiquement des scores $BERTScore_{AutoORIG}$

plus élevés que les scores $BERTScore_{AutoSIMP}$; en d'autres termes, les simplifications générées par les LLM restent sémantiquement plus proches du texte source que des références produites par les experts.

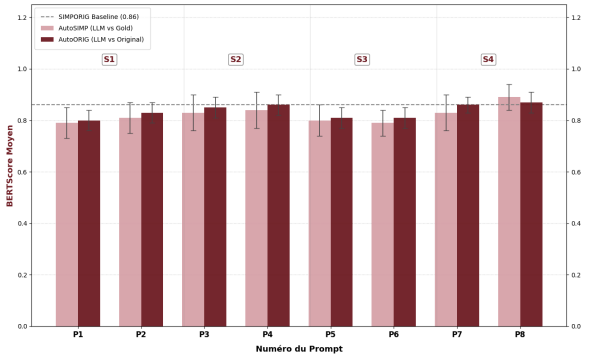

FIGURE 3 – Comparaison des BERTScore (moyenne ± écart-type) pour 4 stratégies de prompting.

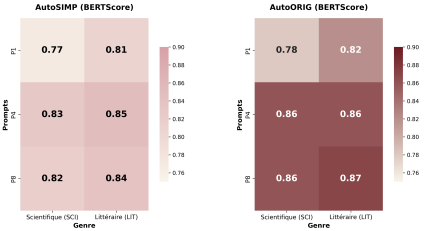

FIGURE 4 – Similarité sémantique par genre (P1, P4, P8)

4.2.2 Performance selon le niveau scolaire

Une analyse de la préservation sémantique à travers les niveaux scolaires, illustrée dans le tableau 4, révèle des tendances nuancées en fonction des stratégies de guidage utilisées. La courbe de référence $BERTScore_{SIMPORIG}$ (experts vs original) montre une augmentation de la proximité sémantique de la `CE1` à la `CM1` année $(0,91)$, avant de connaître une légère baisse en `CM2` $(0,87)$, ce qui suggère que les simplifications humaines sont structurellement plus proches de l'original pour les niveaux intermédiaires. En ce qui concerne la fidélité au texte source (AutoORIG), une tendance majeure se dessine : les modèles conservent une plus grande proximité sémantique avec l'original en 2e année (`CE1`) qu'en 5e année (`CM2`). Cette observation est particulièrement marquée pour les invites P4 (0,88) et P8 (0,89). Un point de convergence apparaît en quatrième année (`CM1`), où toutes les suggestions atteignent un pic de performance sémantique, que ce soit par rapport à l'original ou à la référence humaine. Ce niveau semble représenter le « point d'équilibre » auquel la simplification automatique se rapproche le plus des normes humaines. Enfin, une analyse comparative des invites confirme la supériorité de P4 en tant que modèle mimétique. Sa courbe AutoSIMP suit presque parfaitement la dynamique de la courbe de référence $BERTScore_{SIMPORIG}$, tout en maintenant une distance sémantique constante. Alors que P8 suit de près cette tendance avec une excellente préservation du sens d'origine, la consigne P1 s'avère la moins efficace ; en effet, elle présente un faible $BERTScore_{AutoSIMP}$ (sémantiquement éloignée de la simplification des experts) mais reste sémantiquement proche du texte d'origine, alors que pour les textes `IReST`, elle est sémantiquement éloignée à la fois du texte d'origine et du texte simplifié par les experts. Bien que la forte similitude sémantique avec le texte original (AutoORIG) au niveau de la 2e année (`CE1`) puisse laisser supposer un manque de modification, les indices de lisibilité offrent une image plus nuancée. Les résultats du tableau 2 montrent que les textes de 3e (`CM1`) année obtiennent un score FRE moyen de $99,63$, dépassant la lisibilité de l'original $(92,68)$ et la référence des experts $(97,91)$. Cependant, cette amélioration structurelle est tempérée par un score SARI relativement modeste de $37,02$. Ces données suggèrent que, si les modèles parviennent à rendre le texte plus fluide pour ce niveau, leur simplification reste partielle, privilégiant une préservation stricte du sens d'origine

là où les experts s'autorisent une refonte sémantique plus profonde pour atteindre une simplicité optimale. En résumé, P4 apparaît comme le meilleur compromis, reproduisant avec succès la logique de transformation des experts tout en garantissant une grande fidélité aux idées du texte source.

Score	Prompt	CE1	CE2	CM1	CM2	IREST
$BERTScore_{SIMPORIG}$ (Experts vs Orig)	Réf.	0.80±0.14	0.86±0.11	0.91±0.03	0.87±0.03	0.89±0.02
$BERTScore_{AutoSIMP}$ (LLM vs Experts)	P1	0.78±0.09	0.78±0.07	0.81±0.03	0.80±0.03	0.78±0.03
	P2	0.77±0.08	0.80±0.08	0.84±0.03	0.81±0.02	0.82±0.03
	Moy.	0.78	0.79	0.82	0.81	0.80
	P3	0.80±0.10	0.82±0.09	0.85±0.04	0.83±0.03	0.85±0.02
	P4	0.78±0.11	0.82±0.09	0.87±0.03	0.85±0.03	0.87±0.03
	Moy.	0.79	0.82	<u>0.86</u>	0.84	<u>0.86</u>
	P5	0.77±0.08	0.79±0.07	0.81±0.02	0.81±0.03	0.80±0.02
	P6	0.77±0.08	0.79±0.07	0.81±0.03	0.81±0.02	0.77±0.02
	Moy.	0.77	0.79	0.81	0.81	0.78
	P7	0.79±0.10	0.82±0.09	0.85±0.04	0.83±0.02	0.86±0.02
	P8	0.79±0.11	0.83±0.09	0.85±0.03	0.83±0.03	0.86±0.01
	Moy.	0.77	0.82	0.85	0.83	<u>0.86</u>
$BERTScore_{AutoORIG}$ (LLM vs Original)	P1	0.82±0.04	0.81±0.04	0.80±0.04	0.79±0.04	0.77±0.04
	P2	0.82±0.04	0.83±0.03	0.84±0.05	0.82±0.03	0.82±0.04
	Moy.	0.82	0.82	0.82	0.81	0.80
	P3	0.87±0.03	0.86±0.04	0.85±0.04	0.83±0.04	0.85±0.02
	P4	0.87±0.04	0.86±0.05	0.86±0.03	0.84±0.04	0.86±0.02
	Moy.	**0.87**	<u>0.86</u>	0.85	0.84	0.85
	P5	0.81±0.03	0.83±0.04	0.81±0.02	0.82±0.04	0.79±0.03
	P6	0.80±0.04	0.82±0.04	0.81±0.03	0.81±0.03	0.76±0.03
	Moy.	0.81	0.83	0.81	0.81	0.78
	P7	0.86±0.03	0.86±0.03	0.85±0.03	0.85±0.04	0.86±0.03
	P8	0.88±0.04	0.88±0.03	0.86±0.03	0.85±0.04	0.86±0.03
	Moy.	**0.87**	**0.87**	0.85	0.85	<u>0.86</u>

TABLE 4 – Analyse de similarité sémantique : BERTScore ↑ par niveau scolaire (moyenne ± écart-type). $BERTScore_{SIMPORIG}$: référence expert vs. original ; $BERTScore_{AutoSIMP}$: LLM vs. référence expert ; $BERTScore_{AutoORIG}$: LLM vs. original.

4.2.3 Performance selon le genre

L'analyse par genre présentée à la figure 3 révèle des tendances distinctes selon l'indicateur considéré. Pour le $BERTScore_{AutoSIMP}$, les textes scientifiques affichent une augmentation notable, passant de $0,77$ pour P1 à $0,83$ pour P4, ce qui indique que des stratégies de formulation d'invites plus structurées sont mieux à même de répondre aux attentes du benchmark. Les textes littéraires présentent une fourchette de variation plus étroite $(0,81 - 0,85)$, ce qui suggère une plus grande convergence des invites pour ce type de texte. En ce qui concerne le $BERTScore_{AutoORIG}$, les scores dépassent systématiquement ceux du $BERTScore_{AutoSIMP}$, avec des valeurs atteignant $0,86$ dans le contexte scientifique et $0,87$ pour le genre littéraire avec P8. Ce résultat indique que les simplifications produites par P4 et P8 maintiennent une cohérence sémantique plus étroite avec le texte source, tandis que P1 génère des résultats qui s'éloignent davantage, sur le plan sémantique, de l'original.

5 Discussion

L'analyse croisée des mesures de lisibilité (FRE), de simplification (SARI) et de similarité sémantique (BERTScore) permet d'apporter des réponses nuancées aux questions de recherche initiales de cette étude.

Amélioration de la lisibilité et de la simplicité (QR1) : Les résultats démontrent que la simplification par LLM surpasse systématiquement les manuels originaux et les références expertes en termes de lisibilité. Le prompt P1 maximise la fluidité textuelle ($FRE = 98,26$), particulièrement sur le genre littéraire ($99,61$). Toutefois, cette haute lisibilité s'accompagne du score SARI le plus faible ($34,35$), révélant une transformation de surface qui peine à modifier la structure linguistique profonde, notamment pour les niveaux débutants (`CE1`).

Comparaison avec l'expertise humaine (QR2) : La stratégie *few-shot* (S2) et *Chain-of-Thought* (S4), incarnée respectivement par les prompts P4 et P8, se rapproche le plus des standards experts. P4 s'impose comme un bon modèle mimétique : sa courbe sémantique *AutoSIMP* épouse fidèlement la dynamique de la référence humaine ($BERTScore\_SIMPORIG$) sur l'ensemble des niveaux scolaires. Bien que les experts s'autorisent des remaniements sémantiques plus profonds au `CE1` pour atteindre une simplicité optimale, P4 parvient à équilibrer simplification structurelle et alignement expert, surpassant les autres configurations dans le domaine scientifique.

Préservation du sens d'origine (QR3) : La conservation sémantique est robuste, les scores $BERTScore_{AutoORIG}$ se maintenant au-delà de $0,80$ pour l'ensemble des configurations. P8 atteint une conservation maximale ($0,87$), surpassant même la proximité sémantique entre l'original et la référence humaine. Un point d'équilibre sémantique émerge au `CM1`, où la simplification automatisée s'aligne le mieux sur les attentes humaines tout en respectant rigoureusement l'intégrité du texte source.

Cependant, si le BERTScore (Zhang *et al.*, 2019) permet de valider avec robustesse la proximité sémantique et d'assurer qu'aucune information essentielle n'est perdue lors du processus, il présente des limites inhérentes à son approche purement vectorielle. L'évaluation de la lisibilité par des métriques automatiques ne capture pas pleinement l'adéquation ergonomique pour un profil cognitif spécifique (Novikova *et al.*, 2017). Pour un public dyslexique, l'espacement visuel, la charge mentale et le décodage réel priment sur la simple correspondance statistique des *tokens*. Bien que ces scores automatiques soient indispensables pour valider la faisabilité technique de nos stratégies de *prompting* à grande échelle, ils ne se substituent pas à une validation cognitive directe.

6 Conclusion et perspectives

Cette étude a évalué l'efficacité du modèle LearnLM, spécifiquement conçu pour un contexte éducatif, pour la simplification de textes scolaires selon diverses stratégies de *prompting*. L'analyse révèle que si les méthodes de base privilégient une lisibilité de surface, les approches structurées telles que le *Few-shot* (P4) et le *Chain-of-Thought* (P8) offrent une simplification structurelle plus robuste. La prompt P4 (stratégie « Few-Shot ») s'est imposée comme le meilleur compromis en termes de FRE, SARI et BertScore. En obtenant un score de similarité sémantique (BertScore AutoSIMP) supérieur à celui de P8 (Chain-of-Thought), elle parvient à reproduire plus fidèlement la simplification utilisée par les experts, notamment en reproduisant la courbe du comportement humain à travers les différents niveaux scolaires. Sa supériorité réside également dans sa stabilité sémantique à travers les

genres scientifiques et littéraires, alors que P8 s'écarte plus significativement des textes scientifiques d'experts. On observe toutefois que P4 produit des scores SARI légèrement inférieurs pour les niveaux intermédiaires 3 (`CE2`) et 4 (`CM1`), conséquence directe de l'utilisation de textes des niveaux 2 (`CE1`) et 5 (`CM2`) dans sa configuration Few-Shot. À l'inverse, P8 se distingue par le score SARI le plus élevé, ce qui indique une simplification plus radicale. Bien qu'il produise des textes sémantiquement plus proches des originaux que les simplifications d'experts (à l'exception des niveaux 4 (`CM1`) et 5 (`CM2`)), il parvient à réduire l'écart-type entre les niveaux scolaires sans s'appuyer sur des textes de référence. Ainsi, si l'objectif est d'optimiser la similitude avec l'expertise multidisciplinaire, P4 est préférable. Si l'objectif est une simplification maximale sans référence, P8 s'avère plus efficace.

Cependant, pour pallier la complexité de l'ingénierie de prompt en contexte scolaire, le déploiement opérationnel de ces modèles nécessitera l'intégration de la stratégie P4 en arrière-plan (*back-end*) au sein d'une interface utilisateur simplifiée. L'enseignant pourra ainsi générer des textes adaptés aux besoins spécifiques de sa classe sans avoir à maîtriser la syntaxe des requêtes. Par ailleurs, les travaux futurs s'inscriront dans le cadre d'un modèle d'ingénierie pédagogique structuré (de type ADDIE) pour la création de ressources pédagogiques multimédias interactives dans le cadre d'une e-formation. La phase d'évaluation (*Evaluation*) dépassera alors le calcul exclusif de métriques automatisées pour privilégier le recueil d'un feedback qualitatif rigoureux. Ce processus impliquera des tests *in situ* afin de mesurer, non pas par une simple note quantitative, mais par une analyse qualitative approfondie de l'expérience utilisateur, l'impact réel de ces textes générés sur l'aisance de lecture et la compréhension globale des élèves concernés.

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*.

ANDERSON J. (1983). Lix and rix : Variations on a little-known readability index. *The Journal of Reading*, **26**.

BADACHE I. & BELLET P. (2024). Intelligence artificielle : usage pédagogique et esprit critique. In *16ème édition du colloque Interactions Multimodales Par ÉCran, IMPEC 2024*.

BADACHE I. & COLOMBO E. (2025). Repenser les pratiques d'enseignement et d'apprentissage par la robotique éducative : le cas du robot socio-émotionnel buddy. In *Actes de l'atelier Intelligence Artificielle générative et ÉDUcation : Enjeux, Défis et Perspectives de Recherche 2025 (IA-ÉDU)*, p. 98–110.

BLISS T., ROBINSON T. J., HILTON J. & WILEY D. A. (2013). An oer coup : College teacher and student perceptions of open educational resources. *Journal of interactive media in education*, **2013**(1), 4–4.

BROWN T. B., MANN B., RYDER N., SUBBIAH M., KAPLAN J., DHARIWAL P., NEELAKANTAN A., SHYAM P., SASTRY G., ASKELL A., AGARWAL S., HERBERT-VOSS A., KRUEGER G., HENIGHAN T., CHILD R., RAMESH A., ZIEGLER D. M., WU J., WINTER C., HESSE C., CHEN M., SIGLER E., TEUSZ LITWIN M., GRAY S., CHESS B., CLARK J., BERNER C., MCCANDLISH S., RADFORD A., SUTSKEVER I. & AMODEI D. (2020). Language models are few-shot learners. *ArXiv*, **abs/2005.14165**.

CARDON R. & GRABAR N. (2020). French biomedical text simplification : When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, p. 710–716.

CHALLENER D. W., WEN A., FAN J. W., LIU H., O'HORO J. & NYMAN M. A. (2025). Flesch-kincaid grade level readability scores to evaluate readability of clinical documentation during an electronic health record transition. *Advances in health information science and practice*, **1 1**, VBWY7913.

CROSSLEY S. A., LOUWERSE M. M., MCCARTHY P. M. & MCNAMARA D. S. (2007). A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, **91**(1), 15–30.

DELGADOVA E. (2015). Reading literacy as one of the most significant academic competencies for the university students. *Procedia-Social and Behavioral Sciences*, **178**, 48–53.

ESKENAZI M., LIN Y. & SAZ O. (2013). Tools for non-native readers : the case for translation and simplification. In *Proceedings of the Workshop on Natural Language Processing for Improving Textual Accessibility*, p. 20–28.

FANG D., QIANG J., OUYANG X., ZHU Y., YUAN Y. & LI Y. (2025). Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, p. 897–912.

FARROW R., PITT R., DE LOS ARCOS B., PERRYMAN L.-A., WELLER M. & MCANDREW P. (2015). Impact of oer use on teaching and learning : Data from oer r esearch h ub (2013–2014). *British Journal of Educational Technology*, **46**(5), 972–976.

FLESCH R. F. (1948). A new readability yardstick. *Journal of Applied Psychology*, **32**(3), 221–233. DOI : 10.1037/h0057532.

GALA N., FRANÇOIS T., JAVOUREY-DREVET L. & ZIEGLER J. C. (2018). La simplification de textes, une aide à l'apprentissage de la lecture. *Langue française*, **199**(3), 123–131.

GALA N., TACK A., JAVOUREY-DREVET L., FRANÇOIS T. & ZIEGLER J. C. (2020). Alector : A parallel corpus of simplified french texts with alignments of misreadings by poor and dyslexic readers. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, p. 1353–1361.

GREEN A. & HAWKEY R. (2012). Re-fitting for a different purpose : A case study of item writer practices in adapting source texts for a test of academic reading. *Language Testing*, **29**(1), 109–129.

HAVER H. L., GUPTA A. K., AMBINDER E. B., BAHL M., OLUYEMI E. T., JEUDY J. & YI P. H. (2024). Evaluating the use of chatgpt to accurately simplify patient-centered information about breast cancer prevention and screening. *Radiology : Imaging Cancer*, **6**(2), e230086.

HEDLIN E., ESTLING L., WONG J., DEMMANS EPP C. & VIBERG O. (2025). Got it ! prompting readability using chatgpt to enhance academic texts for diverse learning needs. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference*, p. 115–125.

JIN T., LU X. & NI J. (2020). Syntactic complexity in adapted teaching materials : Differences among grade levels and implications for benchmarking. *The Modern Language Journal*, **104**(1), 192–208.

JURENKA I., KUNESCH M., MCKEE K. R., GILLICK D., ZHU S., WILTBERGER S., PHAL S. M., HERMANN K., KASENBERG D., BHOOPCHAND A. *et al.* (2024). Towards responsible development of generative ai for education : An evaluation-driven approach. *arXiv preprint arXiv :2407.12687*.

KNOTH N., TOLZIN A., JANSON A. & LEIMEISTER J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education : Artificial Intelligence*, **6**, 100225. DOI : 10.1016/j.caeai.2024.100225.

KONG A., ZHAO S., CHEN H., LI Q., QIN Y., SUN R. & ZHOU X. (2023). Better zero-shot reasoning with role-play prompting.

LIU F., JIANG Y., LAI C. & JIN T. (2024). Teacher engagement with automated text simplification for differentiated instruction. *Language Learning & Technology*, **28**(2), 163–182.

MADDELA M., ALVA-MANCHEGO F. & XU W. (2021). Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics : Human Language Technologies*, p. 3536–3553.

MARTIN L., DE LA CLERGERIE É. V., SAGOT B. & BORDES A. (2020). Controllable sentence simplification. In *Proceedings of the twelfth language resources and evaluation conference*, p. 4689–4698.

MO K. & HU R. (2024). Expertease : A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics : EMNLP 2024*, p. 9080–9099.

MULLIS I. V., MARTIN M. O., FOY P. & HOOPER M. (2017). epirls 2016 : International results in online informational reading. *International Association for the Evaluation of Educational Achievement*.

NAIT DJOUDI A. A., NOUALI S., AABID M., BADACHE I., CHIFU A.-G. & BELLOT P. (2025). Lis at simpletext 2025 : Enhancing scientific text accessibility with llms and retrieval-augmented generation. In *Notebook for the SimpleText Lab at the 16th Conference and Labs of the Evaluation Forum (CLEF 2025)*, volume 4038.

NOVIKOVA J., DUŠEK O., CERCAS CURRY A. & RIESER V. (2017). Why we need new evaluation metrics for NLG. In M. PALMER, R. HWA & S. RIEDEL, Éds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, p. 2241–2252, Copenhagen, Denmark : Association for Computational Linguistics. DOI : 10.18653/v1/D17-1238.

OUYANG L., WU J., JIANG X., ALMEIDA D., WAINWRIGHT C., MISHKIN P., ZHANG C., AGARWAL S., SLAMA K., RAY A. *et al.* (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, **35**, 27730–27744.

PEREZ F. & RIBEIRO I. (2022). Ignore previous prompt : Attack techniques for language models. *arXiv preprint arXiv :2211.09527*.

REID M., SAVINOV N., TEPLYASHIN D., LEPIKHIN D., LILLICRAP T. P. *et al.* (2024). Gemini 1.5 : Unlocking multimodal understanding across millions of tokens of context. *ArXiv*, **abs/2403.05530**.

RELLO L., BAEZA-YATES R., BOTT S. & SAGGION H. (2013). Simplify or help ? text simplification strategies for people with dyslexia. In *Proceedings of the 10th international cross-disciplinary conference on web accessibility*, p. 1–10.

RENNARD V., SHANG G., GRARI D., HUNTER J. & VAZIRGIANNIS M. (2023). Fredsum : A dialogue summarization corpus for french political debates. In *Findings of the Association for Computational Linguistics : EMNLP 2023*, p. 4241–4253.

ROSE D. H. & MEYER A. (2002). *Teaching every student in the digital age : Universal design for learning*. ERIC.

SAGGION H. (2017). Applications of automatic text simplification. In *Automatic Text Simplification*, p. 71–77. Springer.

SHARDLOW M. (2014). A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*. DOI : 10.14569/SpecialIssue.2014.040109.

SIDDHARTHAN A. (2014). A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, **165**(2), 259–298.

STANOVICH K. E., NATHAN R. G. & VALA-ROSSI M. (1986). Developmental changes in the cognitive correlates of reading ability and the developmental lag hypothesis. *Reading research quarterly*, p. 267–283.

SWELLER J. (1988). Cognitive load during problem solving : Effects on learning. *Cognitive science*, **12**(2), 257–285.

TEAM L., MODI A., VEERUBHOTLA A. S., RYSBEK A., HUBER A., WILTSHIRE B., VEPREK B., GILLICK D., KASENBERG D., AHMED D. *et al.* (2024). Learnlm : Improving gemini for learning. *arXiv preprint arXiv :2412.16429*.

TODIRASCU A., WILKENS R., ROLIN E., FRANÇOIS T., BERNHARD D. & GALA N. (2022). Hector : A hybrid text simplification tool for raw texts in french. In *12th International Conference on Language Resources and Evaluation (LREC)*.

TRAUZETTEL-KLOSINSKI S., DIETZ K., GROUP I. S. *et al.* (2012). Standardized assessment of reading performance : The new international reading speed texts irest. *Investigative ophthalmology & visual science*, **53**(9), 5452–5461.

TUNMER W. E. & HOOVER W. A. (2019). The cognitive foundations of learning to read : A framework for preventing and remediating reading difficulties. *Australian Journal of Learning Difficulties*, **24**(1), 75–93.

WEI J., WANG X., SCHUURMANS D., BOSMA M., ICHTER B., XIA F., CHI E. H., LE Q. V. & ZHOU D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA : Curran Associates Inc.

ZHANG T., KISHORE V., WU F., WEINBERGER K. Q. & ARTZI Y. (2019). Bertscore : Evaluating text generation with bert. *arXiv preprint arXiv :1904.09675*.

A Formulation complète des prompts

Cet appendix présente la formulation complète des prompts utilisée dans notre experimentation.

A.1 Prompts complets

A.1.1 Stratégie de prompt 1 (S1)

P1

Make this text more readable for a poor reader or a dyslexic child :

"**{text}**"

P2

Rewrite the following text so that it would be easier to read for a poor reader or a dyslexic child. Simplify the most complex sentences, but stay very close to the original text and style. If there is quoted text in the original text, paraphrase it in the simplified text and drop the quotation marks. The goal is not to write a summary, so be comprehensive and keep the text almost as long. :

"**{text}**"

A.1.2 Stratégie de prompt 2 (S2)

P3

Rewrite the following text so that it would be easier to read for a poor reader or a dyslexic child. Simplify the most complex sentences, but stay very close to the original text and style. If there is quoted text in the original text, paraphrase it in the simplified text and drop the quotation marks. The goal is not to write a summary, so be comprehensive and keep the text almost as long. :

Here is an example

Original :

Jeudi, après déjeuner, comme j'avais encore un peu de temps avant de retrouver les copains au terrain vague, je me suis assis dans le fauteuil, et j'ai pris le journal que papa y avait laissé avant de partir à son travail. Comme d'habitude, à part les photos et les dessins, il n'y avait rien à lire dans le journal, sauf en dernière page, et là, c'était formidable ! Il y avait un grand concours, et le premier prix, c'était une auto !

...

Simplified :

Jeudi, après manger, j'avais encore un peu de temps avant de retrouver les copains dehors. Je me suis assis dans le fauteuil. J'ai pris le journal que papa avait laissé avant de partir à son travail. Comme d'habitude, à part les photos et les dessins, il n'y avait rien à lire dans le journal, sauf en dernière page, et là, c'était génial !

Il y avait un grand concours, et le premier prix, c'était une auto !

On voyait la photo de l'auto. C'était une auto belle comme tout.

...

"**{text}**"

P4

Rewrite the following text so that it would be easier to read for a poor reader or a dyslexic child. Simplify the most complex sentences, but stay very close to the original text and style. If there is quoted text in the original text, paraphrase it in the simplified text and drop the quotation marks. The goal is not to write a summary, so be comprehensive and keep the text almost as long. :

Example 1 :

Original :

Jeudi, après déjeuner, comme j'avais encore un peu de temps avant de retrouver les copains au terrain vague, je me suis assis dans le fauteuil, et j'ai pris le journal que papa y avait laissé avant de partir à son travail.

Comme d'habitude, à part les photos et les dessins, il n'y avait rien à lire dans le journal, sauf en dernière page, et là, c'était formidable !

Il y avait un grand concours, et le premier prix, c'était une auto !

On voyait la photo de l'auto, et c'était une auto chouette comme tout.

...

Simplified :

Jeudi, après manger, j'avais encore un peu de temps avant de retrouver les copains dehors. Je me suis assis dans le fauteuil. J'ai pris le journal que papa avait laissé avant de partir à son travail.

Comme d'habitude, à part les photos et les dessins, il n'y avait rien à lire dans le journal, sauf en dernière page, et là, c'était génial !

Il y avait un grand concours, et le premier prix, c'était une auto !

On voyait la photo de l'auto. C'était une auto belle comme tout.

...

Example 2 :

Original :

L'année 1866 fut marquée par un événement bizarre, un phénomène inexpliqué et inexplicable que personne n'a sans doute oublié. Sans parler des rumeurs qui agitaient les populations des ports et surexcitaient l'esprit public à l'intérieur des continents les gens de mer furent particulièrement émus. Les négociants, armateurs, capitaines de navires, skippers et masters de l'Europe et de l'Amérique, officiers des marines militaires de tous pays, et, après eux, les gouvernements des divers États des deux continents, se préoccupèrent de ce fait au plus haut point. En effet, depuis quelque temps, plusieurs navires s'étaient rencontrés sur mer avec « une chose énorme » un objet long, fusiforme, parfois phosphorescent, infiniment plus vaste et plus rapide qu'une baleine. Les faits relatifs à cette apparition, consignés aux divers livres de bord, s'accordaient assez exactement sur la structure de l'objet ou de l'être en question, la vitesse inouïe de ses mouvements, la puissance surprenante de sa locomotion, la vie particulière dont il semblait doué. ... Simplified :

Quelque chose de bizarre arriva en 1866, une chose difficile à expliquer. Tout le monde s'en souvient. Des rumeurs circulaient dans les ports et sur les continents. Les gens de mer furent très émus. Les négociants, armateurs, capitaines de navires, d'Europe et d'Amérique, officiers des marines militaires de tous pays, et les gouvernements des divers États des deux continents, se préoccupèrent de ce fait important. En effet, depuis quelque temps, plusieurs bateaux avaient rencontré sur mer « une chose énorme ». C'était long, parfois brillant, beaucoup plus grand et plus rapide qu'une baleine. Les informations sur cette apparition étaient écrites dans divers livres de bord. Tous étaient d'accord sur la forme de l'objet ou de l'être en question, la vitesse de ses mouvements, la puissance. ... : "text""

"**{text}**"

A.1.3 Stratégie de prompt 3 (S3)

P5

You are a French primary school teacher with years of experience helping struggling readers. Your task is to rewrite this passage so that a {age_group} years old children with reading difficulties can understand it clearly.

Use short, simple sentences, and easy vocabulary. Maintain the meaning and overall tone of the text. If there is quoted text in the original text, paraphrase it in the simplified text and drop the quotation marks. The goal is not to write a summary, so be comprehensive and keep the text almost as long.

"**{text}**"

P6

You are a French primary school teacher with years of experience helping struggling readers. Your task is to rewrite this passage so that a {age_group} years old children with dyslexia can understand it clearly.

Use short, simple sentences, and easy vocabulary. Maintain the meaning and overall tone of the text. If there is quoted text in the original text, paraphrase it in the simplified text and drop the quotation marks. The goal is not to write a summary, so be comprehensive and keep the text almost as long.

"**{text}**"

A.1.4 Stratégie de prompt 4 (S4)

P7

Your task is to rewrite the text below to make it easier to understand for poor readers and dyslexic children, while keeping the original meaning and tone. Do not summarizekeep the text almost as long and as detailed.

Please follow these simplification strategies :

1. Replace difficult or uncommon words with simpler ones. Prioritize short, frequently used vocabulary.
2. Simplify complex word forms.
3. Use short, direct sentences. Prefer active voice over passive. Avoid subordinate clauses.
4. Clarify references by replacing pronouns when needed.
5. Paraphrase quoted speech and remove quotation marks.

Now simplify this text step by step :

"**{text}**"

P8

,Your task is to rewrite the text below to make it easier to understand for poor readers and dyslexic children, while keeping the original meaning and tone. Do not summarize , keep the text almost as long and as detailed.

Please follow these simplification strategies :

1. Perform lexical replacements by identifying difficult or uncommon words and replace them with simpler synonyms.
2. Apply morphological simplifications by using simpler verb forms.
3. Carry out syntactic transformations by simplifying sentence structures.
4. Perform discourse simplification by replacing pronouns with explicit referents.

Make sure the meaning of the original sentence is preserved throughout the simplification process. Do it step by step :

"**{text}**"