

Modélisation de la lisibilité en français pour les personnes en situation d'illettrisme

Wafa Aissa^{1,*}, Thibault Bañeras-Roux^{1,*}, Elodie Vanzeveren¹, Lingyun Gao¹, Alice Pintard¹, Rodrigo Wilkens², Thomas François¹

¹CENTAL, IL&C, UCLouvain, Belgique. ²University of Exeter, Royaume-Uni.
* Ces auteurs ont contribué de manière égale à cet article.



Motivations et Contexte

- Illettrisme** : situation de personnes qui, bien qu'ayant été scolarisées, ne parviennent pas à lire et à comprendre un texte portant sur des situations de la vie quotidienne (ANLCI).
- Problème** : aucune formule de lisibilité adaptée aux adultes en situation d'illettrisme en français.

- Objectif** : une **formule de lisibilité spécialisée** pour ce public.
- Approche** :
 - Corpus annoté par des formateurs experts de ce public.
 - Échelle spécifique à 4 classes et des valeurs de 1 à 20.
 - Comparaison de plusieurs modèles de lisibilité.

Corpus annoté

- 461 textes courts (25 à 608 tokens), représentatifs de 11 types de communication (personnelle, professionnelle, commerciale, académique, légal, politique, religieuse, réseaux sociaux...).
- Sources** : formateurs en illettrisme, sur le web et manuels de FLE.
- Public cible** : adultes en situation d'illettrisme.
- Processus d'annotation** :
 - 15 annotatrices professionnelles, formées et rémunérées.
 - Utilisation d'un guide validé + plateforme Qualtrics.
 - Chaque texte annoté par au moins 3 personnes.

- Échelle principale** : 4 classes de difficulté
Très Facile → *Facile* → *Accessible* → *+Complexe*
- Échelle secondaire** : échelle de Likert à 5 points pour chaque classe
- Calcul du score final** : moyenne des 3 annotations
- Double annotation** :
4 classes (Classification), valeurs 1–20 (Régression)
- Distribution des données par classes**:

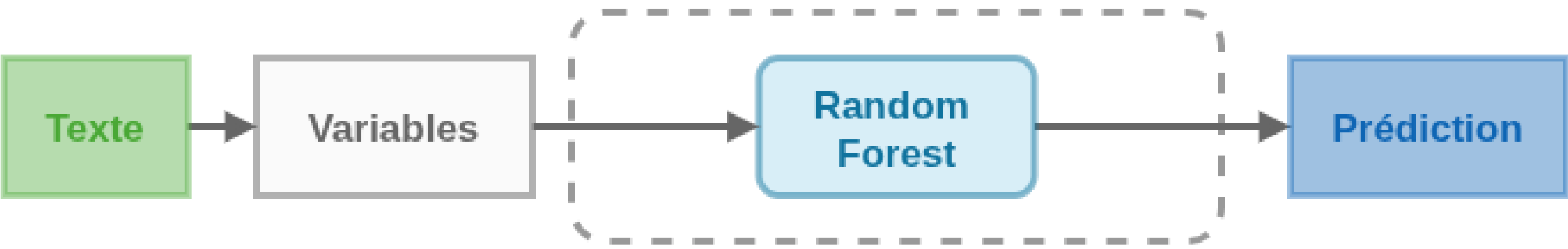
Très Facile	Facile	Accessible	+Complexe
19	212	198	32



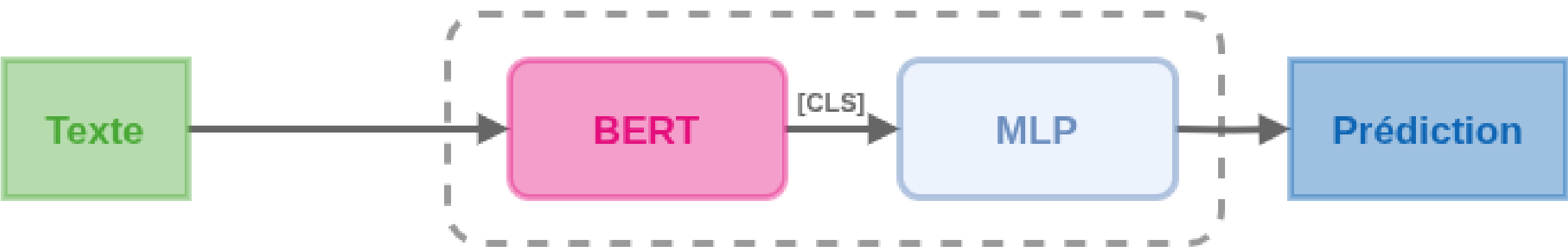
Modélisation de la lisibilité

- Texte** : entrée du modèle, texte à évaluer.
- Prédiction** : sortie du modèle, niveau de difficulté 1-4 (classification) ou score 1-20 (régression).
- Variables** : variables linguistiques via FABRA (Wilkens et al., 2022) suivies d'une sélection de variables avec mRmR (Ding & Peng, 2003).
- Bert**: CamemBERT (Antoun et al., 2024) affiné.

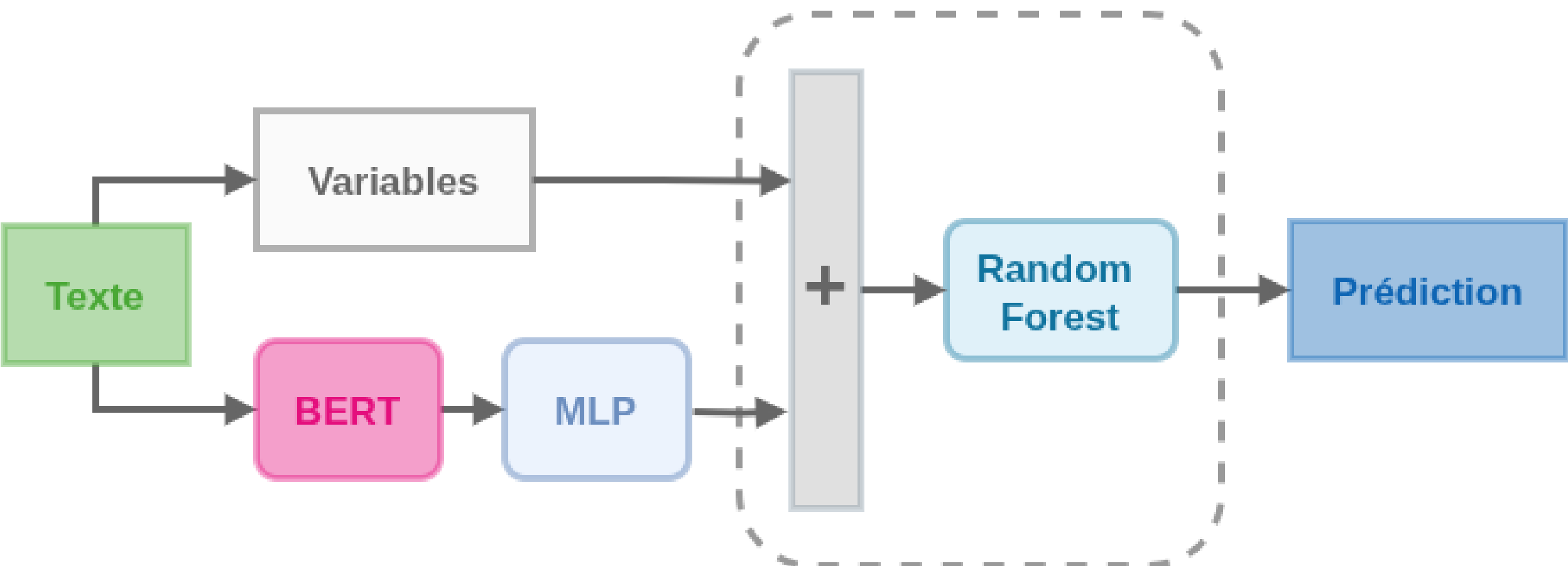
ML - Random Forest



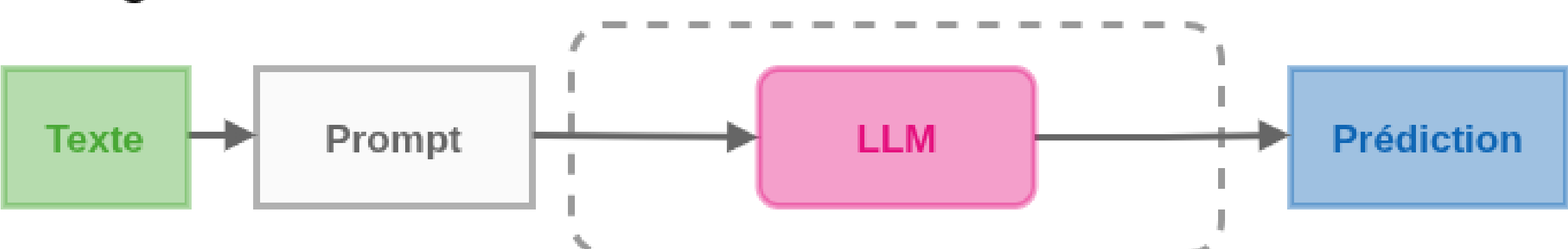
DL



Hybride



LLM génératif



Évaluation des systèmes de lisibilité

Comparaison des performances des modèles de classification et de régression:

	Exactitude	Exactitude Adj.	Macro-F1	EQM
Classification				
ML - SVM (500)	55,84 ± 4,26	97,83 ± 1,95	47,54 ± 6,15	–
ML - DT (300)	54,10 ± 3,28	94,81 ± 0,81	43,84 ± 4,98	–
ML - RF (400)	62,77 ± 4,18	98,05 ± 1,26	47,78 ± 7,60	–
DL - CamemBERT	64,04 ± 9,97	98,71 ± 1,77	60,36 ± 8,23	–
DL - CamemBERT-v2	64,26 ± 5,67	99,17 ± 0,91	60,05 ± 6,01	–
Hybride - RF (300)	67,32 ± 4,08	99,14 ± 0,81	56,26 ± 9,17	–
Régression				
ML - SVR (500)	39,60 ± 5,39	93,06 ± 3,61	22,63 ± 1,96	4,94 ± 1,07
ML - DT (50)	38,75 ± 5,29	88,10 ± 2,24	22,13 ± 3,15	7,22 ± 1,55
ML - RF (500)	40,89 ± 6,28	91,77 ± 3,10	22,96 ± 4,71	4,70 ± 0,73
DL - CamemBERT	70,77 ± 5,48	100,00 ± 0,00	59,63 ± 2,55	3,87 ± 0,72
DL - CamemBERT-v2	68,38 ± 5,70	100,00 ± 0,00	47,52 ± 8,36	3,78 ± 0,75
Hybride - RF (300)	64,28 ± 6,55	99,57 ± 0,53	36,50 ± 5,21	4,88 ± 0,75

- RF légèrement supérieur en exactitude, mais pas en macro-F1.
- Forte disparité exactitude / macro-F1 : biais vers classes majoritaires.
- BERT meilleur en EQM, mais peu de gains macro-F1 face aux hybrides.

Comparaison des performances des LLMs génératifs pour la classification:

	Exactitude	Exactitude Adj.	Macro-F1
Mistral-large-FR-zero-shot	31.46 ± 5.48	91.33 ± 2.83	26.72 ± 4.24
Mistral-large-FR-few-shot	56.61 ± 3.03	98.48 ± 0.87	43.86 ± 3.57
Mistral-large-EN-zero-shot	31.47 ± 8.33	90.02 ± 3.85	22.78 ± 5.34
Mistral-large-EN-few-shot	58.13 ± 3.36	98.48 ± 1.11	48.60 ± 4.08
GPT-4.1-FR-zero-shot	30.15 ± 6.98	84.60 ± 4.00	35.55 ± 9.34
GPT-4.1-FR-few-shot	46.64 ± 5.08	96.10 ± 1.62	43.02 ± 7.82
GPT-4.1-EN-zero-shot	30.18 ± 11.37	91.76 ± 3.78	27.01 ± 6.04
GPT-4.1-EN-few-shot	48.38 ± 5.61	96.75 ± 1.36	44.69 ± 6.84
DeepSeek-R1-FR-zero-shot	41.75 ± 5.14	91.77 ± 2.14	38.41 ± 3.14
DeepSeek-R1-FR-few-shot	53.89 ± 1.14	96.53 ± 1.87	47.06 ± 3.55
DeepSeek-R1-EN-zero-shot	33.31 ± 8.95	87.87 ± 4.04	29.64 ± 8.22
DeepSeek-R1-EN-few-shot	52.80 ± 3.25	94.81 ± 1.26	48.95 ± 0.49

- Few-shot > zero-shot : meilleures performances et plus de stabilité.
- FR > EN en zero-shot ; EN > FR en few-shot (selon macro-F1).
- DeepSeek meilleur en macro-F1, mais pas en exactitude.
- LLMs génératifs proches des modèles DL en macro-F1, sans supervision.

Conclusion

- Création d'un corpus annoté spécialisé pour l'évaluation de la lisibilité en français pour les adultes en situation d'illettrisme.
- Évaluation de modèles ML, DL, hybrides et LLMs pour la lisibilité.
- Les modèles hybrides et profonds offrent les meilleures performances en exactitude et macro-F1.
- La moyenne des annotations améliore la cohérence malgré la subjectivité des annotatrices.
- Perspectives: Une analyse des divergences entre annotatrices pour mieux comprendre les critères de jugement.

Définition des niveaux de difficulté

- Très Facile** : Textes compris par les personnes n'ayant pas terminé l'école primaire et n'ayant presque aucune expérience de lecture.
- Facile** : Textes entièrement (ou presque) compris par des personnes qui n'ont pas dépassé le premier cycle du secondaire et ont une faible expérience de lecture.
- Accessible** : Textes compris dès la première lecture par des personnes ayant fini le premier cycle du secondaire et ayant une expérience moyenne de lecture.
- +Complexe** : Textes plus complexes que les niveaux précédents.