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

L'impact de l'échantillonnage sur la détectabilité des textes rédigés par une IA

Matthieu Dubois[1] Pablo Piantanida[2,3] François Yvon[1]
(1) Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, Paris, France
(2) International Laboratory on Learning Systems (ILLS), Quebec AI Institute (MILA), CNRS, Québec, Canada
(3) CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
duboism[at]isir.upmc.fr, pablo.piantanida[at]mila.quebec,
yvon[at]isir.upmc.fr

RÉSUMÉ

Les textes générés par les grands modèles de langage (LLM) étant souvent indiscernables des écrits humains, leur détection automatique suscite aujourd'hui un intérêt croissant. Bien que de nombreux détecteurs récents affichent une précision quasi parfaite (avec des scores AUROC dépassant 99%), ces performances reposent généralement sur des paramètres de génération fixes. Cela soulève la question de leur robustesse face aux différentes stratégies de décodage. Dans cette étude, nous analysons l'impact du décodage par échantillonnage sur la détectabilité, en observant comment d'infimes variations dans la distribution des (sous-)mots affectent les résultats. Nous montrons que des ajustements mineurs (température, top-p ou nucleus) peuvent sévèrement dégrader la précision, l'AUROC chutant parfois jusqu'à 1%. Nos conclusions mettent en lumière les failles des méthodes actuelles et soulignent le besoin d'évaluations plus rigoureuses. Nous publions nos résultats et données https://github.com/BaggerOfWords/Sampling-and-Detection.

ABSTRACT

How Sampling Affects the Detectability of Machine-written texts : A Comprehensive Study.

As texts generated by Large Language Models (LLMs) are indistinguishable from human-written content, research on automatic text detection has attracted growing attention. Many recent detectors report near-perfect accuracy, boasting AUROC scores above 99%. However, these claims assume fixed generation settings, leaving open the question of how robust such systems are to changes in decoding strategies. In this work, we examine how sampling-based decoding impacts detectability, with a focus on how variations in a model's (sub)word-level distribution affect detection performance. We find that minor adjustments to decoding parameters - such as temperature, top-p, or nucleus sampling - can impair detector accuracy, with AUROC dropping from near-perfect levels to 1% in some settings. Our findings expose blind spots in current detection methods and emphasize the need for more comprehensive evaluation protocols.

MOTS-CLÉS : Génération de Textes, Détection de textes artificiels, Théorie de l'information.

KEYWORDS: Text Generation, Machine-generated text detection, Information theory.

ARTICLE ACCEPTÉ À : Findings of the Association for Computational Linguistics : EMNLP 2025.
URL : https://aclanthology.org/2025.findings-emnlp.609/

Actes des 33ème Conférence sur le Traitement Automatique des Langues Naturelles.
Volume 2 : articles déjà publiés, pages 23–23.