Devis de Recherche : Détection du texte généré par IA dans l'enseignement supérieur

Jacques Parizeau

November 22, 2025

Cycle de recherche

Ce devis s'aligne sur la séquence: Question → Théorie → Hypothèses → Méthodologie → Résultats → Conclusion → Contribution.

1 Question

Quelle combinaison de méthodes (outils de détection automatisés, techniques adverses, et révision humaine) maximise la précision et l'équité dans l'identification du texte généré par intelligence artificielle dans le contexte de l'enseignement supérieur?

2 Théorie

La littérature suggère que les outils actuels de détection de texte IA présentent une précision limitée Elkhatat et al. (2023); Farrelly & Baker (2023); Weber-Wulff et al. (2023), et que l'utilisation de techniques adverses, telles que la paraphrase et l'introduction d'erreurs, peut significativement réduire leur efficacité Perkins et al. (2024); Weber-Wulff et al. (2023); Perkins et al. (2023). Les modèles d'apprentissage machine spécialisés surpassent parfois les systèmes grand public Najjar et al. (2025); "The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication" (2025); Gotoman et al. (2025). Les réviseurs humains tendent à mieux discerner le contenu généré par IA dans certains cas, grâce à leur capacité de détecter des incohérences Liu et al. (2024); Flitcroft et al. (2024); Popkov & Barrett (2024).

3 Hypothèses

- **H1 : Les outils de détection de texte généré par l'IA présentent une fiabilité limitée.**

Plusieurs études indiquent que l'exactitude des détecteurs est souvent inférieure à 80 %, avec un risque non négligeable de faux positifs et de classifications incertaines, surtout pour les textes humains écrits par des non-natifs ou dans des contextes particuliers (Elkhatat et al., 2023; Farrelly & Baker, 2023; Weber-Wulff et al., 2023).

- **H2 : Les techniques d'adversarialisation et de paraphrase réduisent significativement l'efficacité des détecteurs d'IA.**

  Des stratégies simples telles que l'ajout de fautes d'orthographe ou l'augmentation de l'hétérogénéité syntaxique diminuent l'exactitude des outils de détection, rendant ces derniers peu fiables pour garantir l'intégrité académique (Perkins et al., 2024; Weber-Wulff et al., 2023).

- **H3 : Les humains, notamment les professeurs, sont globalement plus performants que les outils automatisés pour discerner le contenu généré par IA dans des textes spécialisés.**

  Des études dans le domaine médical montrent que les réviseurs humains identifient plus efficacement les incohérences ou erreurs dans les textes générés par IA, souvent avec une précision supérieure à celle des outils d'IA (Liu et al., 2024).

- **H4 : Les modèles d'apprentissage machine spécialisés surpassent les systèmes de détection généralisés pour identifier le texte IA dans des contextes techniques précis.**

  Les modèles d'apprentissage automatique tels que *XGBoost* ou *Random Forest* offrent une meilleure performance sur des tâches ciblées que les outils grand public tels que *GPTZero*, surtout lorsque la différentiation porte sur des caractéristiques pratiques plutôt qu'abstraites (Najjar et al., 2025).

- **H5 : L'utilisation exclusive des détecteurs de texte IA expose à des risques d'équité et d'intégrité scientifique.**

  L'application non critique de ces outils peut engendrer des conséquences injustes pour certains groupes (locuteurs non natifs, étudiants internationaux), soulignant le besoin d'une approche pédagogique intégrée, éthique et informée (Farrelly & Baker, 2023).

4 Méthodologie

Ce devis propose une expérimentation comparative sur un corpus de textes humains et générés par IA, dont certains sont modifiés par des techniques adverses (paraphrase, fautes d'orthographe). Chaque texte sera soumis à des outils automatisés généralistes, des modèles d'apprentissage machine spécialisés et des réviseurs humains. Les performances seront mesurées en termes de taux de vrais positifs, faux positifs et faux négatifs, en tenant compte de la variabilité selon le profil des auteurs (locuteurs natifs/non natifs, étudiants internationaux). Une

analyse statistique permettra de comparer la pertinence de chaque approche et d'identifier la meilleure combinaison pour garantir fiabilité et équité.

5 Résultats attendus

Il est attendu que la précision varie selon l'approche et le type de texte. Les techniques adverses devraient diminuer l'exactitude des détecteurs automatisés, alors que l'expertise humaine et le recours à des modèles spécialisés pourraient améliorer la fiabilité globale et limiter les injustices envers les groupes vulnérables Walters (2023); Han et al. (2025); Mauti & Ayieko (2025).

6 Conclusion

L'étude visera à proposer un cadre combiné où la complémentarité entre outils automatiques et révision humaine permet de renforcer le contrôle d'intégrité académique, tout en limitant les risques de biais et d'erreur.

7 Contribution

Ce projet contribuera à l'optimisation des protocoles de détection de texte généré par IA dans l'enseignement supérieur, en intégrant une analyse comparative, des recommandations méthodologiques, et une attention particulière à l'équité entre les groupes d'étudiants.

References

The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. (2025, June). *PeerJ Comput. Sci.*, *11*. Retrieved from https://consensus.app/papers/the-accuracybias-tradeoffs-in-ai-text-detection-tools-and
doi: 10.7717/peerj-cs.2953

Elkhatat, A., Elsaid, K., & Almeer, S. (2023, September). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, *19*, 1–16. Retrieved from https://consensus.app/papers/evaluating-the-efficacy-of-ai-content-detection-tools-in-
doi: 10.1007/s40979-023-00140-5

Farrelly, T., & Baker, N. (2023, November). Generative Artificial Intelligence: Implications and Considerations for Higher Education Practice. *Education Sciences*. Retrieved from https://consensus.app/papers/generative-artificial-intelligence-implications-and-farre
doi: 10.3390/educsci13111109

Flitcroft, M., Sheriff, S., Wolfrath, N., Maddula, R., McConnell, L., Xing, Y., ... Kothari, A. (2024, June). Performance of Artificial Intelligence Content Detectors Using Human and Artificial Intelligence-Generated Scientific Writing. *Annals of surgical oncology*. Retrieved from https://consensus.app/papers/performance-of-artificial-intelligence-content-flitcroft-
doi: 10.1245/s10434-024-15549-6

Gotoman, J. E., Luna, H. L., Sangria, J. C., Santiago, C., & Barbuco, D. D. (2025, February). Accuracy and Reliability of AI-Generated Text Detection Tools: A Literature Review. *American Journal of IR 4.0 and Beyond*. Retrieved from https://consensus.app/papers/accuracy-and-reliability-of-aigenerated-text-detection-go
doi: 10.54536/ajirb.v4i1.3795

Han, B., Nawaz, S., Buchanan, G., & Mckay, D. (2025, January). Students' Perceptions: Exploring the Interplay of Ethical and Pedagogical Impacts for Adopting AI in Higher Education. *International Journal of Artificial Intelligence in Education*. Retrieved from

Liu, J., Hui, K., Zoubi, F. A., Zhou, Z., Samartzis, D., Yu, C., ... Wong, A. (2024, May). The great detectives: humans versus AI detectors in catching large language model-generated medical writing. *International Journal for Educational Integrity*, *20*, 1–14. Retrieved from https://consensus.app/papers/the-great-detectives-humans-versus-ai-detectors-in-liu-hu
doi: 10.1007/s40979-024-00155-6

Mauti, J., & Ayieko, D. S. (2025, January). Ethical Implications of Artificial Intelligence in University Education. *East African Journal of Education Studies*. Retrieved from https://consensus.app/papers/ethical-implications-of-artificial-intelligence-in-mauti-
doi: 10.37284/eajes.8.1.2583

Najjar, A., Ashqar, H., Darwish, O., & Hammad, E. (2025, January). Detecting AI-Generated Text in Educational Content: Leveraging Machine Learning and Explainable AI for Academic Integrity. *ArXiv*, *abs/2501.03203*. Retrieved from https://consensus.app/papers/detecting-aigenerated-text-in-educational-content-najjar-
doi: 10.48550/arxiv.2501.03203

Perkins, M., Roe, J., Postma, D., McGaughran, J., Vietnam, D. H. B. U., Vietnam, ... Singapore (2023, May). Detection of GPT-4 Generated Text in Higher Education: Combining Academic Judgement and Software to Identify Generative AI Tool Misuse. *Journal of Academic Ethics*, *22*, 89–113. Retrieved from https://consensus.app/papers/detection-of-gpt4-generated-text-in-higher-education-perk
doi: 10.1007/s10805-023-09492-6

Perkins, M., Roe, J., Vu, B., Postma, D., Hickerson, D., McGaughran, J., ... Singapore, J. C. U. (2024, March). GenAI Detection Tools, Adversarial Techniques and Implications for Inclusivity in Higher Education. *ArXiv*, *abs/2403.19148*. Retrieved from https://consensus.app/papers/genai-detection-tools-adversarial-techniques-and-perkins-
doi: 10.1186/s41239-024-00487-w

Popkov, A., & Barrett, T. (2024, March). AI vs academia: Experimental study on AI text detectors' accuracy in behavioral health academic writing. *Accountability in research*, 1–17. Retrieved from https://consensus.app/papers/ai-vs-academia-experimental-study-on-ai-text-detectors-po
doi: 10.1080/08989621.2024.2331757

Walters, W. (2023, January). The Effectiveness of Software Designed to Detect AI-Generated Writing: A Comparison of 16 AI Text Detectors. *Open Information Science*, *7*. Retrieved from https://consensus.app/papers/the-effectiveness-of-software-designed-to-detect-walters/
doi: 10.1515/opis-2022-0158

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., ... Waddington, L. (2023, June). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, *19*, 1–39. Retrieved from https://consensus.app/papers/testing-of-detection-tools-for-aigenerated-text-weber-wul
doi: 10.1007/s40979-023-00146-z