La détection de l'utilisation de robots conversationnels en contexte universitaire : Le cas de Compilatio Magister+

Philippe Dessus, Daniel Seyve

La détection de l'utilisation de robots conversationnels en contexte universitaire : Le cas de *Compilatio Magister+*

Philippe Dessus[1] & Daniel Seyve[2]

[1] Univ. Grenoble Alpes, LaRAC & Inspé, 38000 Grenoble, France

[2] Univ. Grenoble Alpes, DAPI, 38000 Grenoble, France

Résumé

Nous avons testé les performances de *Compilatio Magister+* outil de détection de l'utilisation de robots conversationnels (*chatbots*) sur 86 documents (écrits par des humains, générés par deux robots conversationnels différents : *ChatGPT 4* et *Claude 2* ou *Claude 2.1*) représentant des compositions proches de celles rencontrées en milieu académique. Les résultats montrent que, paradoxalement, *Compilatio Magister+* obtient des scores sans erreurs pour la détection des *écrits humains*. En revanche, son score de détection n'est pas meilleur que le hasard, et son score F1 montre que pour une prédiction correcte, CM+ fait plus de 10 erreurs. Ces résultats sont en ligne avec la décision de plus en plus fréquente, dans les universités, notamment étasuniennes, de désactiver la détection de l'usage de ces robots conversationnels.

Mots-clés : Robots conversationnels ; Outils de détection ; Intégrité académique

Abstract

We tested the performance of the Compilatio Magister+ conversational bot detection tool on 86 documents (written by humans, generated by two different conversational bots: ChatGPT 4 and Claude 2 or Claude 2.1) representing compositions close to those encountered in academic environments. The results show that, paradoxically, Compilatio Magister+ achieves 100% scores for the detection of human writings. On the other hand, its detection score is no better than chance, and its F1 score shows that for a correct prediction, CM+ makes more than 10 errors. These results are in line with the increasingly frequent decision by universities, particularly in the USA, to disable detection of the use of these conversational bots.

Key-words: Conversational bots; Detection systems; Academic integrity

1. Introduction

Le but de cet article est d'évaluer les performances de l'outil *Compilatio Magister+*, dorénavant CM+ (https://www.compilatio.net/magister-plus), dans la détection de documents générés, totalement ou partiellement, par des robots conversationnels (*chatbots*) utilisant de grands modèles de langage (parmi les plus connus, citons *ChatGPT* d'OpenAI https://chat.openai.com, *Claude* d'Anthropic https://claude.ai/, ou *Gemini* de Google https://gemini.google.com/). Depuis l'arrivée et l'utilisation massive de *ChatGPT* v. 3, puis 3.5, mis gratuitement à la disposition du grand public depuis novembre 2022 cette détection est devenue centrale pour les universités, recourant déjà pour la plupart à des outils de détection de similarités dans les copies d'étudiant·es pour repérer des comportements de tricherie, c'est-à-dire d'externalisation du travail aux robots et leur appropriation sans mention de cette externalisation.

Nous avons testé les performances de l'utilisation de CM+ dans sa version payante (v. 2.2 utilisée de fin novembre à mi-décembre 2023). Jusqu'à présent, beaucoup d'outils, donc celui-ci, ont été testés en version gratuite, avec des documents dont le nombre de mots est très réduit (entre 200 et 2000 caractères, selon Weber-Wulff *et al*., 2023), et majoritairement en langue anglaise. Il est donc important de connaître le niveau de fiabilité de CM+ dans le traitement de documents en français, de plus grande longueur, et de composition proche de celle rencontrée en milieux académiques.

Il faut noter que nous avons prêté attention à deux points, souvent peu pris en compte :

– nous n'avons intentionnellement pas mentionné la ou les « intelligence·s artificielle·s » (ou la ou les IA) comme étant à la base des robots conversationnels et des outils de détection ; ce terme étant trop générique, et souvent plutôt utilisé à des fins commerciales. L'utiliser sans préciser le fonctionnement des systèmes participe à créer un flou, à attribuer à ces derniers des performances magiques, souvent bien différentes de leurs performances réelles ; ce n'est rien d'autre qu'un terme de marketing (Hawley, 2019).
– nous avons autant que possible évité d'anthropomorphiser les performances et productions des robots conversationnels, préférant donc mentionner que les humains *écrivent* un texte et que les robots le *génèrent*. Là aussi, l'anthropomorphisation montre faussement que les performances, mais parfois aussi les processus à l'œuvre dans les systèmes sont très proches de ceux des humains, ce qui n'est souvent pas le cas (Firestone, 2020 ; Floridi, 2023 ; Hawley, 2019).

Notre position est que verser dans la généralisation ou bien l'anthropomorphisation des processus ou performances des robots conversationnels, ou de tout autre outil de ce type, nous prive du recul nécessaire pour évaluer le plus justement possible ces processus ou performances. Cet article présente un bref état de l'art sur le sujet de la détection de l'usage de robots conversationnels, ainsi que sur le fonctionnement de CM+, pour en venir à décrire la méthode et les résultats de notre test du système, et finit par une discussion.

2. Détection de l'usage de robots conversationnels : une brève revue

La détection de l'usage de robots conversationnels amène une série de questions : Comment cette détection fonctionne-t-elle ? Est-elle fiable ? Et ensuite, dans le cas où la réponse est positive, cette détection peut-elle être utilisée en contexte éducatif ?

Les chercheurs travaillant sur ce problème partent du principe que les documents produits par des robots conversationnels ont des caractéristiques particulières pouvant être détectées automatiquement, soit par de simples analyses statistiques des documents, ou bien par des analyses plus sophistiquées. Par exemple, il a été montré que la référence à des sentiments est moins fréquente que dans les textes écrits par des humains, la fréquence de pronoms personnels ou de tournures agressives est plus basse, enfin, la fréquence de mots peu usités est plus haute (Mitrovic *et al*., 2023). Yu *et al*. (2023) indiquent aussi que les documents générés par des robots ont plus de mots uniques (hapax) et moins d'erreurs grammaticales que ceux écrits par des humains. Des outils graphiques, comme GLTR de Gehrmann *et al*. (2019) permettent de mettre en évidence la plus grande diversité et moindre prédictibilité des documents générés par des humains. Plus récemment, des systèmes plus sophistiqués, dont CM+, utilisent l'apprentissage profond (*deep learning*) pour les entraîner à classer les documents selon leur probabilité d'inclure des passages générés par des robots conversationnels.

S'il y a déjà des travaux testant les performances de détecteurs de robots conversationnels, ils sont encore peu nombreux et utilisent souvent les versions gratuites des détecteurs, de plus, en langue anglaise. Weber-Wulff *et al*. (2023) ont testé 14 détecteurs de robots conversationnels (leur version gratuite) sur 5 types de documents (54 au total) : écrits par des humains (nommés H), écrits par des humains puis traduits en anglais (H), générés par un robot conversationnel (M), générés par un robot puis révisés par un humain (MH) et enfin générés par un robot puis révisés par ce même robot (MM). Le score d'exactitude des détecteurs est entre 76 % (pour CM+) et 33 %. Ce score, prenant en compte l'analyse de Vrais négatifs (des documents humains correctement détectés), est donc optimiste car les différents systèmes détecteurs testés ont des scores élevés dans la détection de ces derniers (96 % en moyenne), qui se dégradent fortement dans la détection des textes générés par des robots.

Chaka (2023) a testé 5 détecteurs (toujours dans leur version gratuite) leur capacité, d'une part à détecter du texte anglais généré par trois robots conversationnels, et d'autre part à le détecter une fois traduit, par *Google Translate*, dans d'autres langues (5 langues testées, dont le français), ce qui correspond à des documents M et MM. La meilleure performance a été de 15 Vrais positifs (des textes générés par des robots correctement détectés) pour 6 Faux positifs (des textes écrits par des humains et faussement attribués à des robots) pour les documents anglais. La traduction espagnole des textes par le meilleur détecteur l'a amené à une performance de 13 Vrais positifs pour 4 Faux positifs, les quatre autres détecteurs ayant des performances bien plus faibles.

Sigut *et al*. (2023) s'intéressent à comparer les performances de deux détecteurs, dont CM+ version gratuite, sur des documents écrits par des humains ou générés par ChatGPT 3.5 ou 4.0, en tchèque et français *vs*. leur traduction en anglais (*via DeepL*), soit au total 135 documents. Le score d'exactitude pour les documents en langue originale générés par ChatGPT 4.0 est de 50 %, et baisse à 44 % pour les textes traduits des performances proches du hasard.

Ces résultats sont très proches de ceux trouvés par van Oijen (2023), qui a testé 7 détecteurs gratuits sur 10 documents (6 M et 4 H). Le meilleur score d'exactitude pour les documents M est de 50 %, là aussi égal à celui du hasard, deux détecteurs ayant même un score de 0 %. En revanche, le score d'exactitude pour les documents H atteint 100 % chez 3 détecteurs.

En résumé, nous pouvons mettre en avant, de cette brève revue, la faible performance des outils de détection de robots conversationnels, notamment lorsque ces robots sont

les plus évolués (version 4.0 de ChatGPT). Paradoxalement, ces outils ont de meilleures performances dans leur détection des documents H que des documents M. De plus, ces études sont principalement réalisées à partir de documents en anglais, pas nécessairement produits en contexte académique, et testant les versions gratuites des détecteurs. Cet article vise à tester la version commerciale de l'un d'entre eux, CM+, avec les robots conversationnels les plus évolués et dans des configurations de documents proches de celles rencontrées dans le milieu académique. Avant de présenter notre étude, détaillons le fonctionnement de CM+ tel que nous pouvons le comprendre de la documentation de son concepteur.

3. Le fonctionnement de Compilatio Magister +

Le fonctionnement de CM+ n'est pas clairement documenté, que ce soit sur le site de son concepteur ou dans des articles scientifiques. Agnès, le P.D.-G. de Compilatio, en présente le fonctionnement dans une vidéo du CSE de l'université de Lausanne (Centre de soutien à l'enseignement, CSE UNIL, 2023, extrait vidéo entre 6'00'' et 13'19''), que nous résumons ici. CM+ s'appuie sur des techniques de traitement automatiques des langues avancées, que nous pouvons sommairement décrire ainsi :

1. Les mots de chaque document sont représentés dans un espace à grandes dimensions (plusieurs milliers), selon la méthode des « *Word embeddings* » (plongements lexicaux). Ainsi, les mots de sens proche se situeront à proximité dans l'espace. De plus, des couples de mots ayant des rapports similaires se situeront à des distances égales l'un de l'autre et les vecteurs formés par chaque couple auront une direction similaire (*e.g.*, le vecteur composé des mots « Paris » et « France » aura la même norme et direction que celui composé des mots « Ankara » et « Turquie »). Cette représentation est un modèle approximatif du sens des mots, ce qui va être utile, par exemple pour extraire ce que l'on nomme des attributs (*features*) des documents (*e.g.*, le nombre de phrases par paragraphe, le nombre de mots par paragraphe, la présence de mots spécifiques, comme « toutefois », « mais », « parce que », etc., voir Desaire *et al.*, 2023, pour un détail de cette technique). À la fin de cette étape, chaque document n'est plus représenté par les mots qui le composent, mais par un grand nombre d'attributs qui le caractérisent de manière unique.
2. Ensuite, un grand corpus de documents, composé d'une part de documents écrits par des humains et d'autres générés par des robots conversationnels et étiquetés sont transformés en attributs selon l'étape 1 pour être soumis à analyse par des techniques d'apprentissage profond (*deep learning*), afin que le système génère automatiquement les règles permettant de les classer aussi fiablement que possible dans l'une ou l'autre des classes « humain » ou « machine ». La langue et le domaine de ce corpus de documents, et bien sûr le corpus lui-même, a été sélectionné par ses concepteurs. Cette sélection, non explicitée pour des raisons de protection commerciale du produit, est faite selon les partis-pris des concepteurs et influe grandement sur les performances du système.
3. Enfin, une fois que le système de l'étape 2 est suffisamment entraîné, on peut présenter un nouvel ensemble de documents (dont on connaît la classe, H vs. M) pour tester la fiabilité du système, et s'il s'avère qu'il est suffisamment fiable, le développer à grande échelle. Il est important de noter que la qualité des résultats dépend de la langue des corpus d'entraînement. Nous supposons, même si ce point n'est pas non plus documenté pour CM+, que ce dernier a été entraîné avec des

corpus de documents en langue française, et partageant le plus possible de caractéristiques des données d'entraînement.

4. Méthode

Détaillons maintenant les principaux paramètres du contexte d'utilisation de robots conversationnels à l'université. Ces paramètres sont très nombreux et, comme nous avions un nombre limité de tickets d'évaluation de CM+, nous avons dû faire des choix. Chacun des paramètres ci-après peut influer sur les résultats finaux de détection de CM+. Nous avons essayé d'en proposer la plus grande variété possible, en restant aussi les plus proches possible d'une utilisation réelle des robots conversationnels. Les principaux paramètres qui varient sont :

– *la langue* : CM+ étant un système français, nous supposerons qu'il est suffisamment performant pour la détection de document écrits en français ; mais tous les étudiants ne sont pas des locuteurs natifs francophones et le système pourrait catégoriser en Faux positifs des documents écrits par des allophones (Liang *et al*., 2023). Toutefois, nous avons préféré ne pas faire varier ce paramètre et ne faire analyser que des documents en français ;
– *le domaine* : la matière dont il est question dans le document ; il est possible que la composition du corpus d'entraînement rende CM+ plus performant dans certains domaines que d'autres. Nous avons décidé de faire varier ce paramètre (voir section 4.2) ;
– *le type de travail universitaire* : sa longueur, sa complexité syntaxique et lexicale, mais également sa forme (essai, résumé, synthèse). Nous avons également décidé de faire varier ce paramètre ;
– *le type de robot conversationnel utilisé* : il existe de nombreux robots conversationnels sur le marché, certains d'accès gratuit et d'autres d'accès payant. Il est tout à fait possible que certains parviennent mieux que d'autres à générer des documents moins détectables (Sigut *et al*., 2023). Ce paramètre a également varié dans notre étude ;
– *le type de prompt utilisé* : un prompt est la requête donnée en input au robot conversationnel pour qu'il génère une réponse. Comme l'ingénierie du prompt est complexe et que des prompts très élaborés pourraient rendre la détection de leur résultat plus difficile, nous avons préféré ne pas faire varier ce paramètre, utiliser des prompts simples et de ne les reformuler que si le robot conversationnel venait à proposer une réponse non satisfaisante (voir ci-après, et l'Encadré 1 qui présente certains prompts utilisés) ;
– *le type de révision après génération* : il est à noter que de multiples révisions d'un même document (que ce soit par machine ou humain) le rend plus difficilement détectable (Cai & Cui, 2023). Nous avons aussi décidé de faire varier ce paramètre.

4.1. Situations d'utilisation de robots conversationnels à l'université

Afin de tester CM+ dans des situations réalistes, d'un point de vue universitaire, c'est-à-dire représentatives de ce que des étudiant·es pourraient pratiquer, nous avons constitué un corpus comprenant quatre types de documents :

– Des documents écrits par des étudiants (copies d'examen), ou diverses personnes (issus de l'encyclopédie en ligne Wikipédia), dans un but académique (dorénavant appelés *documents H*, pour humain) ;
– Des documents générés par des robots conversationnels (appelés *documents M*, pour machine) ;

– Des documents générés par des étudiants et ensuite améliorés, enrichis, par des robots conversationnels (appelés *documents HM*) ;
– Des documents générés par des robots conversationnels, passés au détecteur CM+, et dont les passages détectés ont été modifiés par ces mêmes robots (appelés *documents MM*).

Ces quatre types de documents sont censés couvrir une partie, que nous estimons suffisante, des utilisations de ce type d'outils en contexte universitaire :

– *Les documents H* sont pertinents à tester pour vérifier que le détecteur ne fait pas de Faux positifs (*i.e.*, signale à tort qu'un document écrit par un humain a été en réalité généré, en tout ou partie, par un robot conversationnel) ;
– *Les documents M* sont pertinents à tester pour s'assurer que le détecteur parvient à en détecter le plus grand nombre possible (score de précision). Cela pourrait correspondre à un usage du robot conversationnel par des étudiant·es, pour se débarrasser d'un devoir au lieu de l'écrire ;
– *Les documents HM* correspondent à des documents H, dont le premier jet est écrit par des étudiant·es, mais qui ont voulu en améliorer soit le contenu soit le style en recourant à un robot conversationnel ;
– Enfin, *les documents MM* sont des documents générés par la machine et passés au détecteur pour en révéler les possibles failles (passages détectés comme générés par la machine) ; ces passages sont ensuite donnés au robot conversationnel pour reformulation, afin d'espérer en gommer la provenance. Les concepteurs de CM+ signalent toutefois que « Le texte reformulé sera plus *proche d'une génération de textes par IA*. Il sera plus enclin à être reconnu comme tel. » (les auteurs soulignent, voir https://support.compilatio.net/hc/fr/articles/17407151992465), ce test permettra de trancher sur la capacité de CM+ à détecter plus aisément, ou non, des documents retouchés deux fois par une machine en comparaison avec les documents M.

4.2. Constitution du corpus

Dans le but de tester une diversité de types de documents, nous avons choisi d'en sélectionner dans deux domaines : les sciences de la nature et informatique d'une part, et les sciences humaines et sociales de l'autre. Pour qu'ils correspondent à des productions universitaires standard (type « devoirs à la maison »), nous avons sélectionné et fait générer des documents de taille relativement réduite (environ 1500 mots).

Les documents H (écrits par des humains) sont composés, d'une part, d'une dizaine de copies d'examen d'étudiants anonymisées du début des années 2000, utilisées à l'époque pour tester un système de détection de sections de cours (Lemaire & Dessus, 2001) ; et d'autre part de pages Wikipédia sur des thèmes divers (*cf.* ci-dessous). Comme il commence à être difficile d'être sûr, *a priori*, que des documents ne sont pas, au moins partiellement, générés par un robot conversationnel, nous avons repris l'idée de Mindner *et al.* (2023). Nous avons récupéré des pages Wikipédia sur des sujets divers, dans les SHS et les sciences de la nature. Pour nous assurer de leur « non pollution » nous avons utilisé la *Wayback machine* (https://web.archive.org) pour récupérer la version de ces pages à la date de mi–2016. Nous avons aussi fait produire, par deux robots conversationnels largement utilisés (Claude 2.0 ou 2.1 et ChatGPT 4), des documents M, sur les thèmes des documents précédents, dans une taille équivalente.

Le Tableau 1 ci-dessous reprend les principales caractéristiques des documents du corpus, et des détails sont donnés ci-après.

Tableau 1. Principales caractéristiques des documents du corpus de test initial. Légende : SHS : sciences humaines et sociales ; SNI : sciences de la nature et informatique

Domaine	Thèmes	Auteur ou générateur	*N*
SHS	L'effet Pygmalion, Jean Piaget, l'attention, la motivation, la parenté	Wikipédia (H)	5
		Robot conversationnel (M)	10
SNI	Le réchauffement climatique, la protéine, le chat de Schrödinger, la naine blanche, le test de Turing	Wikipédia (H)	5
		Robot conversationnel (M)	10
SHS	Sciences de l'éducation : pensée des enseignants, analyse de l'activité des élèves et des enseignants	Étudiants (copies d'examen) (H)	10
Total			40

Ces 40 documents composent le corpus initial des documents H et M. Quarante-six documents ont été ajoutés, formant les documents HM et MM, de la manière suivante.

– Pour composer les *documents HM*, nous avons d'une part sélectionné les documents H dont des passages ont été détectés — à tort — par CM+ et avons demandé que ces derniers soient reformulés et précisés, à l'aide des deux robots conversationnels. Six documents issus de la Wikipédia (Piaget, Motivation, Attention, Turing, Protéine, et Réchauffement climatique) étaient dans ce cas, ce qui donne *12 documents HM* ;
– Nous avons aussi, systématiquement, fait modifier (reformuler et préciser) les copies d'examen par les 2 robots conversationnels (qu'ils aient ou non été sujets à détection par CM+), car cela pourrait correspondre à une situation où des étudiants veulent améliorer le contenu de leur texte. Cela donne *20 autres documents HM*.
– Enfin, pour composer les documents MM, nous avons fait modifier les documents M ayant été sujets à détection par CM+ pour les faire reformuler et préciser, dans le but de tester si oui ou non le système continuerait de détecter les passages modifiés. Cela concerne *14 documents MM*.

Ce qui porte le corpus à *86 documents au total*. Pour générer les documents M, MM et HM, nous avons utilisé le site Vello.ai (https://vello.ai/) qui présentait l'avantage de donner l'accès gratuit à deux des plus performants robots conversationnels, ChatGPT 4 et Claude 2.1 (il faut noter que pendant le déroulement de l'étude, la version de Claude est passée de 2.0 à 2.1).

4.3. Méthode de génération des documents M, MM et HM

Décrivons plus précisément le type de prompts (requêtes en langage naturel donnant un ordre au robot) utilisés pour faire générer les documents par les deux robots conversationnels. Il faut déjà noter que nous avons ont dû parfois adapter notre stratégie pour faire en sorte que le résultat attendu soit réellement généré, tant les robots conversationnels peuvent avoir des réponses variées, et parfois non conformes aux attendus. Si l'ingénierie des prompts est un domaine en pleine expansion, elle n'est à notre avis pas fréquemment maîtrisée par des étudiant·es tout-venant. Nous n'avons donc pas cherché à construire des prompts sophistiqués, qui seraient de plus moins susceptibles d'être détectés.

Les deux prompts génériques, pour d'une part générer un document sur un thème donné (document M), et d'autre part pour les reformuler (MM et HM) étaient respectivement les suivants :

- `Écris un article encyclopédique sur [le sujet X] en environ 1 500 mots.`
- `Peux-tu reformuler et préciser les paragraphes commençant par XXX, sans modifier les autres paragraphes ?`

Mais il faut noter que demander au robot conversationnel de regénérer avec plus de précisions les paragraphes d'un document, fourni juste après le prompt, commençant par "XXX" n'est pas toujours suivi du résultat attendu : parfois le robot ajoute des "XXX" au début de certains paragraphes, parfois il reformule d'autres paragraphes que ceux attendus. Nous avons donc dû, le cas échéant, reformuler le prompt en ne ciblant que les paragraphes concernés. L'Encadré 1 ci-dessous détaille les prompts utilisés.

Encadré 1. Quelques prompts utilisés pour générer des document et leurs modifications

```
Écris un article encyclopédique sur le phénomène de l'attention, en psychologie, d'environ 1500 mots

Explique en 1500 mots comment l'effet Pygmalion pourrait affecter le bonheur et la réussite scolaire des enfants à l'école?

Retrace l'histoire du concept général de motivation en psychologie

Écris un article encyclopédique sur le réchauffement climatique d'environ 1500 mots.

Décris et précise-moi en 1500 mots les principaux concepts de la théorie de Jean Piaget

Détaille en environ 1500 mots l'ensemble des fonctions qu'assurent les protéines au sein des cellules et des tissus

Écris en environ 1500 mots un document sur Turing et son apport philosophique sur les IA

Écris un article encyclopédique sur le Chat de Schrödinger, d'environ 1500 mots

Décris et précise-moi en 1500 mots les principaux concepts de la parenté en sciences humaines
```

Les documents MM proviennent uniquement de documents M générés par ChatGPT, parce qu'ils ont fait l'objet d'un taux de détection de paragraphes plus élevé que les documents M générés par Claude 2.0 ou 2.1. En revanche, nous les avons faits modifier concurremment avec Claude 2 et ChatGPT 4, à des fins de comparaison. Nous nous sommes aussi centrés sur les documents ayant le pourcentage de détection IA le plus élevé, soit les suivants : Attention, Pygmalion, Piaget, Réchauffement climatique, Turing, Schödinger, et Protéines.

4.4. Traitement des données

Comme CM+ n'indique pas, dans ses résultats, les limites des passages des documents qu'il analyse (seuls sont indiqués les passages détectés, et encore dans leur globalité), il nous est impossible de calculer un ratio passages détectés/passages totaux. La

méthode d'évaluation, tirée de Weber-Wulff *et al.* (2023), contourne ce problème. Elle sépare tout d'abord les documents écrits par l'humain de ceux écrits par la machine, et range chacun selon le score de détection donné par CM+, dans les catégories décrites dans le Tableau 2 ci-dessous. Bien évidemment, les documents générés par la machine et ensuite retravaillés par elle (MM) sont rangés dans la catégorie « Document généré par la machine ».

Les documents HM (initialement écrits par un humain et ensuite modifiés par la machine) ne rentrent *a priori* dans aucune de ces deux catégories. Nous avons calculé le taux de mots modifiés par la machine, qui est le pourcentage de mots détectés initialement par CM+, rapporté à la taille totale du document. Ce taux est toujours faible et va de 2,4 % à 7,5 %, ce qui signifie que les documents humains (Wikipédia et copies d'examen) n'ont été que faiblement retouchés par la machine, ayant eu un taux de détection d'écriture machine très faible. Nous avons donc décidé de ranger ces documents dans la catégorie « Document écrit par l'humain » (H) et de leur appliquer les règles de décision de cette catégorie.

Tableau 2. Critères de classification des documents (humains vs. machine). Lecture : Pour un document écrit par un humain, un résultat de 81 % de passages détectés par CM+ est classé "Faux positif". Repris de Weber-Wulff *et al*. (2023, p. 13)

Document écrit par l'humain (H)	
Intervalle (%)	Décision de classement
[100, 80[	Faux positif
[80, 60[	Faux positif partiel
[60, 40[	Non tranché
[40, 20[	Vrai négatif partiel
[20, 0]	Vrai négatif
Document généré par la machine (M ou MM)	
[100, 80[	Vrai positif
[80, 60[	Vrai positif partiel
[60, 40[	Non tranché
[40, 20[	Faux négatif partiel
[20, 0]	Faux négatif

4.5. Calcul des scores

Nous avons ensuite calculé, pour chaque catégorie de documents, les scores suivants, toujours en nous inspirant en partie du travail de Weber-Wulff *et al.* (2023). Deux types de scores peuvent être calculés : le score *global*, calculé sur l'ensemble des documents, quel que soit leur score de détection (voir Tableau 4) ; et le score d'*exactitude strict*, calculé sur l'ensemble des documents ayant eu un score de détection tranché (voir Tableau 5). L'Encadré 2 détaille les formules des différents scores calculés dans l'étude. Ils seront expliqués au fur et à mesure dans la Section résultats.

Encadré 2. Formules des différents scores utilisés dans l'étude

$$Exactitude = (VP + VN)/(VP + VN + FN + FP)$$

$$Spécificité = VN/(VN + FP)$$

$$Précision = VP/(VP + FP)$$

$$Rappel = VP/(VP + FN)$$

$$Exactitude\ équilibrée = (Rappel + Spécificité)/2$$

$$F1 = 2VP/(2VP + FP + FN)$$

5. Résultats

Cette section décrit les principaux résultats du test de détection de CM+. Le Tableau 3 récapitule les effectifs pour chaque type de document, dans chaque classe définie par le Tableau 2.

Tableau 3. Résultats du test de détection CM+. Lecture : Sur un total de 32 documents écrits par des humains et retouchés par un robot conversationnel (HM), CM+ a signalé que 27 d'entre eux étaient écrits par des humains (score strictement compris entre 0 et 20 %) et 5 d'entre eux étaient partiellement écrits par des humains (score compris entre 20 exclus et 40 %)

	Effectifs					
Type de document	Faux positifs	Faux positifs partiels	Indéterminés	Vrais négatifs partiels	Vrais négatifs	Total
H	0	0	0	0	20	20
HM	0	0	0	5	27	32
	Faux négatifs	Faux négatifs partiels	Indéterminés	Vrais positifs partiels	Vrais positifs	Total
M	9	4	6	1	0	20
MM	5	3	6	0	0	14

Ce Tableau 3 permet, en supprimant la catégorie « indéterminés » et en agrégeant ou non les détections partielles, de produire les deux matrices de confusion suivantes (Tableaux 4 et 5). C'est à partir de ces deux matrices que nous pouvons calculer les différents scores de performance de CM+ (Tableau 6). Il montre que les documents H sont tous bien détectés en tant que tels, que faire modifier des documents H par un robot amène 5 documents sur 32 à être partiellement détectés. Un seul des documents M est partiellement détecté et près de la moitié ne le sont pas du tout. Aucun des documents MM n'est détecté, et près du tiers sont considérés comme des documents humains. Passons maintenant au calcul des scores de performance, qui permettent une comparaison fiable entre systèmes et entre études.

Tableau 4. Matrice de confusion globale, en supprimant les cas indéterminés et en agrégeant les détections partielles avec les détections complètes

		Estimé	
		Positif	Négatif
Réel	Positif	Vrai positif : 1	Faux négatifs : 21
	Négatif	Faux positif : 0	Vrais négatifs : 52

Tableau 5. Matrice de confusion stricte, en supprimant les cas indéterminés et les détections partielles

		Estimé	
		Positif	Négatif
Réel	Positif	Vrai positif : 0	Faux négatifs : 14
	Négatif	Faux positif : 0	Vrais négatifs : 47

Tableau 6. Scores d'exactitude, précision, rappel, et F1 pour les 2 matrices de confusion ci-dessus (en %)

Scores	Exactitude	Spécificité	Exactitude équilibrée	Précision	Rappel	F1
Scores globaux	71,6	100	52,3	100	4,5	8,7
Scores stricts	77,0	100	50,0	0	0	0
Scores concepteur	92,3	N/A	N/A	92,5	92,5	N/A

Voici comment interpréter les différents scores. Le *score d'exactitude* donne une information sur le taux de détection des Vrais positifs *et* des Vrais négatifs, donc CM+ parvient à détecter si un document a été écrit par un humain ou une machine pour environ ¾ des documents. Il faut toutefois noter que ce score, comme il combine la détection H et M, masque le fait que CM+ détecte tous les Vrais négatifs (documents H ou HM) et quasiment aucun Vrai positif (documents M ou MM, voir Tableau 3). Le score d'exactitude trouvé avec la version payante de CM+ est conforme avec celui trouvé en utilisant la version gratuite, avec des documents de taille plus petite (Weber-Wulff *et al*., 2023 trouvent un score d'exactitude de 74 %).

Comme il y a un déséquilibre manifeste dans les différentes classes (avec un nombre très important de Vrais négatifs), le score *d'exactitude équilibrée* (*balanced accuracy score*) permet de rééquilibrer le score, il se calcule comme la moyenne arithmétique du score de rappel et de spécificité. On voit, avec ce calcul, s'effondrer le score d'exactitude et le ramener au niveau du hasard.

Le *score de précision*, lui, va se centrer uniquement sur les Vrais positifs. C'est le ratio de Vrais positifs parmi toutes les prédictions positives. Ce ratio sera d'autant plus élevé que le nombre d'erreurs lors d'une prédiction positive (Faux positifs) sera faible, minimisant les accusations d'utiliser un robot conversationnel à tort. Ce score permet donc d'établir une certaine *justice académique*, où les enseignants n'accuseraient pas à

tort des non-usages. CM+ n'ayant reconnu, et encore que partiellement, qu'un seul document généré par un robot et n'ayant produit aucun Faux positif, le score de précision est donc de 100 % globalement (mais ce résultat n'est fondé que sur une seule valeur), et de 0 % si on ne considère pas les résultats partiels.

Le *score de rappel*, également appelé sensibilité, se centre également sur les Vrais positifs, mais en calculant un ratio avec le nombre de cas *réellement* positifs. Ce ratio sera donc d'autant plus élevé que le système détecte correctement les documents générés par un robot, tout en minimisant le nombre d'oublis (Faux négatifs). Ce score est donc également relié à la question de *l'intégrité académique*, mais il faut noter qu'il ne s'intéresse pas aux documents humains. Le score de rappel de CM+, très faible quel que soit le mode de calcul (global ou strict), montre qu'il parvient très faiblement à cibler les documents réellement générés par une machine.

Le *score F1* est une combinaison des deux scores précédents (précision et rappel). Son score d'environ 10 % signale que, pour une prédiction correcte, CM+ fait plus de 10 erreurs.

Il est à noter que les documents humains issus de la Wikipédia ont des taux de Faux positifs bien plus importants que les copies d'examen et que cela ne peut être attribué à leur longueur moyenne plus grande. Nous pensons que cela peut être dû, d'une part, au fait que la Wikipédia en entier a servi de corpus d'entraînement aux différents robots conversationnels, dont les deux que nous avons utilisés et, d'autre part, à la plus grande hétérogénéité de leurs auteurs, certains pouvant de plus ne pas être locuteurs de français L1. En effet, des travaux ont montré que les rédacteurs de L2 avaient plus de chances de voir leur production détectée (Liang *et al*., 2023).

Nous nous sommes enfin penchés sur la différence des scores de détection pour les documents HM et MM, entre la première et la seconde passe. Les concepteurs de CM+ signalent que faire reformuler, dans une deuxième passe, un passage détecté ne le rend pas moins détectable. Le Tableau 7 permet d'examiner cet argument. On peut remarquer que le type de source (Wikipédia *vs*. copies d'examen) interagit avec le score de détection : pour les documents Wikipédia, la reformulation des passages détectés en première passe a tendance à abaisser le score à la deuxième, alors que c'est le contraire pour les copies d'examen. Il est possible que ce phénomène soit dû à l'hétérogénéité des styles d'écriture dans la Wikipédia (combiné au fait que certains auteurs peuvent être non francophones), alors que pour les copies d'examen, plus homogènes, on peut noter une augmentation des scores de détection à la seconde passe.

Les deux dernières colonnes du Tableau 7 montrent que les documents générés par la machine, lors d'une seconde passe, ont des scores de détection sensiblement plus bas, ce qui contredit les concepteurs de CM+ : une reformulation générée par robot des passages détectés amène une baisse de la qualité de la détection (la moyenne du score de passages détectés passant de 45,9 % (écart-type, 15,0) à 29,6 % (écart-type, 18,2).

Tableau 7. Évolution des scores de détection dans les documents HM issus de la Wikipédia. Lecture : Le premier document HM Wikipédia a un score de détection de 0 %, que ce soit avant (passe 1) ou après (passe 2) avoir été reformulé par un robot conversationnel

Wikipédia (HM)		Copies d'examen (HM)		Documents MM	
Passe 1 (%)	Passe 2 (%)	Passe 1 (%)	Passe 2 (%)	Passe 1 (%)	Passe 2 (%)
0	0	0	0	18	13
0	0	0	0	18	16
2	1	0	0	39	27
2	2	0	0	39	34
3	0	0	0	44	0
3	0	0	0	44	47
3	0	0	0	45	24
3	2	0	0	45	42
3	3	0	0	52	4
3	3	0	0	52	49
7	2	0	3	54	45
7	3	0	11	54	56
		0	11	69	13
		0	13	69	44
		0	20		
		0	23		
		0	31		
		0	32		
		0	32		
		0	32		

6. Discussion

Le but de cette étude était de tester l'efficacité de la version commerciale Compilatio Magister + dans la détection de documents générés, en totalité ou partie, par des robots conversationnels. Nous avons constitué un corpus de 86 documents, soit entièrement écrits par des humains, soit entièrement générés par deux robots conversationnels différents (ChatGPT 4 et Claude 2.0 ou 2.1), soit encore des documents d'une de ces 2 catégories « retouchés » ensuite par les deux robots conversationnels, de manière à avoir des documents correspondant à des situations académiques réelles.

Les résultats montrent que, dans ces situations, paradoxalement, *Compilatio Magister+* obtient des scores sans erreurs pour la détection des *écrits humains*. Sa détection ne génère aucun Faux positif. Il obtient grâce à cela un score global d'exactitude de 76 %, ce qui est bien plus bas que les données des concepteurs, mais plus élevé que le score obtenu par Sigut *et al*. (2023), testant aussi des documents en tchèque. En revanche, son score de détection (exactitude corrigée du déséquilibre des classes) n'est pas meilleur que le hasard, et son score F1 montre que pour une prédiction correcte, CM+

fait plus de 10 erreurs. Il est possible que les concepteurs de CM+ n'aient pas pris le risque baisser le seuil de détection des Vrais positifs, au risque d'augmenter en conséquence celui des Faux positifs, et donc des accusations à tort.

Ces résultats sont très largement inférieurs à ceux documentés par les éditeurs de CM+. L'une des raisons est que ChatGPT 4.0 générerait des documents bien moins détectables que sa précédente version 3.5 — que nous supposons avoir été utilisée dans les tests des éditeurs. D'autre part, ces résultats sont compatibles avec les résultats de Sigut *et al.* (2023) ou de van Oijen (2023).

Les limites de cette étude, et donc les pistes de recherches à venir sont les suivantes : nous n'avons pas cherché à peaufiner des prompts évolués, qui auraient peut-être fait baisser encore les performances du système, et, parce que les robots conversationnels ne généraient pas toujours les résultats attendus, nous avons dû leur fournir seulement les passages concernés et non pas le document entier pour les documents HM et MM. Cela, en revanche, devrait avoir tendance à fournir des documents générés moins cohérents, donc plus aisément détectables.

Notons une autre piste : de même qu'il existe des sites utilisant des grands modèles de langage modifiant les documents pour qu'ils ne soient pas détectables par des détecteurs de similarités (*e.g.*, plagiat), il en existe aussi qui annoncent les modifier pour les rendre indétectables par un outil de détection d'usage de robots conversationnels (voir *Humanizer Pro*, https://app.aiprm.com/gpts/g-2azCVmXdy/humanizer-pro). Il pourra être intéressant de tester les performances de CM+ sur des documents passés par cet outil.

Terminons par des considérations plus générales. Il est souvent mis en avant que l'usage de ces outils de détection permet d'améliorer, ou au moins de préserver, l'intégrité académique (e.g., Bin-Nashwan *et al.*, 2023), vue comme la promotion de valeurs comme l'honnêteté, la confiance, la responsabilité, le respect, le pouvoir d'agir. Mais peut-on vraiment dire que ces valeurs sont respectées si elles amènent la détection abusive de Faux positifs et placent des personnes dans des situations où elles sont accusées sans pouvoir fournir de contre-arguments ? Et ce d'autant plus si les performances des systèmes de détection sont proches de celle du hasard ? Sans doute pas. Comme l'indiquent ces derniers auteurs, l'intégrité académique est en suspens depuis l'arrivée des robots conversationnels « Ils peuvent tout autant être utilisés pour gagner du temps, renforcer l'estime de soi, améliorer l'auto-efficacité académique et réduire le stress, commettre des mauvaises conduites et du plagiat » (*id.*, p. 5).

Ces résultats sont également compatibles avec les décisions, de plus en plus fréquentes, de désactiver les fonctions de détection d'usage de robots conversationnels des plates-formes universitaires de détection de similitudes textuelles dans les travaux étudiants (de *Turnitin*, majoritairement utilisé aux États-Unis).

7. Note des auteurs

Nous avons utilisé la version 2.2 de *Compilatio Magister+* du 29 novembre 2023 au 20 décembre 2023. Les résultats de ce test ne valent donc que pour cette période et on ne peut tirer de conclusions fermes sur les versions ultérieures du système, ou avec un autre corpus que celui collationné pour cette étude. Les auteurs remercient la société Compilatio qui nous a aimablement fourni des crédits d'accès à *Compilatio Magister +* pour réaliser cette étude, et qui a répondu avec diligence et précision aux questions que nous avons posées à son service d'aide technique. Nous tenons à la disposition de quiconque nous sollicitera le corpus des documents utilisés dans l'étude, à des fins de

réplication. Ils remercient aussi Timothée Liotard pour son aide dans le recueil des résultats et François Portet pour ses commentaires d'une version précédente de l'article. Enfin, le résumé anglais a été généré par Deepl.com et édité par nous.

8. Références

Bin-Nashwan, S. A., Sadallah, M., & Bouteraa, M. (2023). Use of ChatGPT in academia: Academic integrity hangs in the balance. *Technology in Society*, *75*. https://doi.org/10.1016/j.techsoc.2023.102370

Cai, S., & Cui, W. (2023). Evade ChatGPT detectors via a single space. *ArXiv preprint*. https://arxiv.org/pdf/2307.02599.pdf

Chaka, C. (2023). Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal of Applied Learning & Teaching*, *6*(2). https://doi.org/10.37074/jalt.2023.6.2.12

CSE UNIL (2023, 3 mai). Détection des textes produits par des IA. Entretien avec Frédéric Agnès, PDG de Compilatio. Vidéo YouTube https://youtu.be/yYP1zqrNiJE?si=K0aq7uvo7So5ZjPQ

Desaire, H., Chua, A. E., Kim, M. G., & Hua, D. (2023). Accurately detecting AI text when ChatGPT is told to write like a chemist. *Cell Rep Phys Sci*, *4*(11).

Firestone, C. (2020). Performance vs. competence in human-machine comparisons. *Proc Natl Acad Sci USA*, *117*(43), 26562-26571. https://doi.org/10.1073/pnas.1905334117

Floridi, L. (2023). *L'éthique de l'intelligence artificielle. Principes, défis et opportunités* (E. Panaï & E. R. Goffi, Trad.). Mimésis.

Gehrmann, S., Strobelt, H., & Rush, A. M. (2019). GLTR: Statistical Detection and Visualization of Generated Text. *ArXiv preprint.* https://arxiv.org/pdf/1906.04043.pdf

Hawley, S. H. (2019). Challenges for an ontology of Artificial Intelligence. *ArXiv preprint*. https://arxiv.org/abs/1903.03171

Lemaire, B., & Dessus, P. (2001). A system to assess the semantic content of student essays. *Journal of Educational Computing Research*, *24*(3), 305-320. https://doi.org/10.2190/G649-0R9C-C021-P6X3

Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *arXiv Preprint*. https://arxiv.org/pdf/2304.02819.pdf

Luo, J. (2024). A critical review of GenAI policies in higher education assessment: a call to reconsider the "originality" of students' work. *Assessment & Evaluation in Higher Education*, 1–14. https://doi.org/10.1080/02602938.2024.2309963

Mindner, L., Schlippe, T., Schaaff, K. (2023). Classification of Human- and AI-Generated Texts: Investigating Features for ChatGPT. *arXiv preprint*. https://arxiv.org/pdf/2308.05341.pdf

Mitrovic, S., Davide, A., & Ayoub, O. (2023). ChatGPT or human? Detect and explain. Explaining decisions of machine learning model for detecting short ChatGPT-generated text. *ArXiv preprint*. https://arxiv.org/abs/2301.13852

Šigut, P., & Foltýnek, T. s. (2023). Can We Detect ChatGPT-generated Texts in Czech and Slovak Languages? In A. Horák, P. Rychlý, & A. Rambousek (Eds.), *Proc. Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)* (pp. 35–43). Tribun EU.

van Oijen, V. (2023). AI-generated text detectors: Do they work? https://communities.surf.nl/en/ai-in-education/article/ai-generated-text-detectors-do-they-work

Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *arXiv preprint*. https://arxiv.org/abs/2306.15666

Yu, P., Chen, J., Feng, X., & Xia, Z. (2023). CHEAT: A Large-scale Dataset for Detecting ChatGPT-writtEn AbsTracts. *ArXiv preprint*. https://arxiv.org/pdf/2304.12008.pdf