



Évaluer un travail noté en rédaction à l'université : comparaison entre évaluation humaine et évaluation par un modèle d'intelligence artificielle générative

<https://doi.org/10.52358/mm.vi24.546>

Aya Bouebdelli, doctorante
Université de Montréal, Canada
aya.bouebdelli@umontreal.ca

Évelyne Deprêtre, professeure
Université TÉLUQ, Canada
evelyne.depretre@teluq.ca
<https://www.teluq.ca/communaute/corps-professoral/edepretr>

RÉSUMÉ

L'intelligence artificielle générative (IAG) peut soutenir le jugement évaluatif posé par le corps professoral sur des productions étudiantes. Cependant, son usage soulève des questions dépassant notamment la concordance des notes. Pour mieux comprendre les conditions d'une telle aide de l'IAG, cet article présente une expérimentation exploratoire menée dans un cours universitaire de maîtrise du français écrit. L'étude compare l'évaluation humaine et l'évaluation par l'IAG d'une même production écrite à fonction diagnostique. Un cadre de référence du jugement évaluatif en trois axes sous-tend l'analyse : les fonctions de l'évaluation, ses étapes et sa dimension éthique, indissociable de la responsabilité de la personne évaluatrice. Les résultats montrent une convergence globale entre les deux modes d'évaluation pour les notes et le nombre d'erreurs relevées, mais révèlent des écarts significatifs dans le repérage, la catégorisation des erreurs et l'enchaînement entre le diagnostic, la notation et la rétroaction lorsque la complexité des erreurs augmente. Cette étude révèle que l'IAG ne peut se substituer au jugement évaluatif. Néanmoins, elle constitue une aide pour certaines de ses opérations selon des balises méthodologiques, institutionnelles et éthiques claires, combinées à une formation du corps enseignant et une vigilance professionnelle constante. Elle ouvre également des perspectives de recherche, entre autres, vers la comparaison de différents modèles génératifs.

Mots-clés : intelligence artificielle générative (IAG), jugement évaluatif, rédaction universitaire (en français), comparaison humain-IAG



Introduction

Avec la démocratisation de l'accès aux outils d'intelligence artificielle générative (IAG), entre autres avec l'émergence de ChatGPT en 2022¹, le milieu de l'éducation connaît une transformation rapide de certaines pratiques d'enseignement, d'apprentissage et d'évaluation (Ghimire et Edwards, 2024). En enseignement supérieur, des positionnements institutionnels contrastés en découlent : certaines universités interdisent l'usage de ces outils, tandis que d'autres l'autorisent sous conditions ou à des fins pédagogiques (An *et al.*, 2025; McDonald *et al.*, 2024). Dans ce contexte de changement paradigmatique, la recherche scientifique répond rapidement à plusieurs préoccupations éthiques concernant l'évaluation. Parmi celles-ci figurent la préservation de la validité des évaluations (Audran, 2024; Elshall et Badir, 2025; Ilieva *et al.*, 2025), le maintien de l'intégrité intellectuelle (Bittle et El Gayar, 2025; Lund *et al.*, 2025) et la promotion d'un usage pédagogique responsable de l'IAG (Dabis et Csáki, 2024; Ilieva *et al.*, 2025; Kiliç, 2024; Peterson, 2025). Or, ces préoccupations contemporaines font écho à une réflexion plus ancienne sur l'éthique de « l'agir évaluatif », formulée dès les années 1990 par Charles Hadji. Cet auteur ne considère pas l'évaluation comme un acte neutre, purement technique ou instrumental, mais bien comme constituant fondamentalement un acte social et moral, qui engage la responsabilité de la personne évaluatrice en raison de ses effets sur les individus évalués et de son pouvoir symbolique et institutionnel (Hadji, 1997a, 1997b). Cette conception met en évidence la portée éthique intrinsèque du jugement évaluatif, et ce, au-delà des outils déployés. Elle conduit aussi aujourd'hui à étudier les conditions d'une intégration éthique de l'IAG dans le processus évaluatif.

La tâche d'évaluation en éducation peut d'ailleurs être définie comme un processus intentionnel et situé de jugement professionnel par lequel le corps enseignant recueille et interprète des informations relatives aux apprentissages d'une personne étudiante. Ce jugement s'appuie sur des critères définis et des finalités pédagogiques ou certificatives, dans le but de soutenir l'apprentissage, de garantir l'équité et d'assurer la validité des décisions éducatives (De Ketele, 2010a, 2010b; Black et Wiliam, 1998; Bloom, 1971; Hadji, 1997a, 1997b; Scriven, 1967). Ainsi, outre la conception et l'application d'outils, l'évaluation repose aussi sur une interprétation professionnelle des indices d'apprentissage recueillis, comme des productions écrites, des observations, des réponses, etc., et ce, selon des exigences éthiques.

Parallèlement à l'essor de l'IAG, plusieurs travaux de réflexion et documents d'orientation se sont penchés sur les enjeux éthiques liés à son utilisation en éducation. Ces publications abordent notamment des questions relatives à la conception et à l'usage des systèmes d'IA, à la protection des données, à la transparence, à l'équité et à la gouvernance de ces technologies (p. ex. Collin et Marceau, 2023; Conseil de l'innovation du Québec, 2024; OCDE, 2021; UNESCO, 2023). Bien qu'elles contribuent à encadrer l'intégration responsable de l'IA générative en éducation, elles documentent peu la manière dont les enseignantes et enseignants exercent leur jugement professionnel lorsqu'ils évaluent les productions étudiantes avec l'appui de ces technologies. On constate aussi que, lorsque l'IAG est utilisée dans le champ de l'évaluation, elle est principalement envisagée comme un outil de conception (Ahangama, 2026; Ilieva *et al.*, 2025), de détection (Lund *et al.*, 2025), de contrôle (Bittle et El Gayar, 2025) ou de génération de rétroactions génériques (Mpolomoka, 2025; Zhao *et al.*, 2024). Ainsi, à ce jour, peu de travaux étudient les potentialités de l'IAG pour soutenir le corps enseignant dans l'entièreté de son jugement évaluatif. Celui-ci peut être compris comme l'articulation des phases de correction et d'interprétation du processus d'évaluation (Leroux et Bélair, 2015), notamment en ce qui concerne l'identification de régularités ou d'écarts, la correction, la catégorisation des réponses, l'attribution de notes ou la formulation de rétroactions. C'est ce champ encore peu défriché que nous nous proposons d'examiner : à travers une démarche exploratoire et expérimentale, nous voulons analyser la contribution possible de l'IAG au jugement évaluatif enseignant, tout en examinant son incidence sur des exigences éthiques telles que la cohérence, l'équité, la transparence et les décisions éducatives.

¹ L'agent conversationnel (*chatbot*) ChatGPT fait partie de la série GPT de grands modèles de langage développés par l'entreprise d'intelligence artificielle OpenAI. La version GPT-3.5 a été lancée en 2022 et celle de GPT-4, en mars 2023, entre autres. (Voir OpenAI, 2024). GPT-4 technical report. ArXiv. <https://arxiv.org/abs/2303.08774>



Dans ce contexte, notre objectif principal est de mettre en lumière les différences et les similitudes dans l'évaluation de la maîtrise du français écrit dans un cours universitaire lorsque cette tâche est assurée par une personne humaine ou soutenue par un modèle d'IAG. Nous souhaitons identifier plusieurs aspects liés à l'utilisation de l'IAG pour soutenir l'évaluation de travaux écrits, dont les bonnes pratiques (requêtes, éthique, confidentialité, etc.), les limites (tâches trop complexes ou manquant de précision, etc.) et les écueils (affabulations, erreurs, etc.).

Cet article suit une structure canonique en quatre parties. La première partie est consacrée au cadre théorique de notre recherche et s'articule autour de deux axes : 1) l'évaluation des écrits universitaires, ses fonctions et les étapes de son processus; 2) la comparaison entre l'évaluation humaine et l'évaluation à l'aide de l'IAG. La deuxième partie présente la méthodologie retenue : le contexte universitaire de l'expérimentation, la grille de correction et la requête élaborées, ainsi que les données recueillies. La troisième partie expose les résultats ainsi que leurs analyses, à la fois quantitatives et qualitatives. Quant à la quatrième partie, elle est consacrée à la discussion : nous y mettons en évidence les apports et les limites du modèle d'IAG en réaffirmant le rôle central du jugement humain. Enfin, la conclusion propose des recommandations pratiques et ouvre sur des perspectives de recherche concernant le potentiel pédagogique des modèles génératifs d'IA pour exercer le jugement évaluatif sur les écrits de personnes étudiantes à l'université.

1. Cadre de référence

1.1 Évaluation

Nous nous appuyons sur un cadre de référence qui appréhende l'évaluation selon deux axes qui s'articulent autour d'une dimension éthique.

Le premier axe, celui des fonctions de l'évaluation, nous amène à revenir sur la définition de l'évaluation des apprentissages proposée en introduction, laquelle la conçoit comme un processus structuré et réfléchi visant à recueillir et à analyser des preuves du niveau de maîtrise d'une personne étudiante. Comme l'ont montré de nombreux spécialistes en éducation au cours des dernières décennies, ce processus remplit diverses fonctions qui enrichissent cette définition. L'évaluation au service des apprentissages peut ainsi être envisagée selon deux finalités complémentaires : d'une part, l'évaluation qui vise à juger et à reconnaître les acquis, notamment à travers des fonctions sommatives et certificatives (Bloom, 1971; De Ketele et Gérard, 2007; Perrenoud, 1998, 2004); d'autre part, celle qui soutient la progression des apprentissages, à travers des fonctions formatives, diagnostiques et régulatrices (Black et William, 1998, 2009; De Ketele, 2010a, 2010b; Hadji, 2012; Scalton, 2004, 2015; Scriven, 1967, 1991; Perrenoud, 1998, 2004). Pour le cours faisant l'objet de l'expérimentation, par exemple, la professeure conceptrice a retenu une évaluation en soutien aux apprentissages qui mobilise notamment les fonctions de diagnostic, de régulation et de guidage. Le positionnement précoce de l'évaluation (première semaine) et son intitulé, « texte diagnostique », permettent à la personne étudiante comme à la personne encadrante de repérer les acquis et les éléments à développer en français écrit.

L'évaluation est, en outre, envisagée selon un deuxième axe : celui des étapes. Comme processus, elle compte plusieurs étapes, allant de la clarification des intentions à la prise de décision (entre autres, De Ketele, 1986; Hadji, 2012; Scallan, 2015; Scriven, 1967). Dans cette recherche menée en enseignement supérieur, nous retenons le modèle en quatre étapes de Leroux et Bélair (2015), soit les intentions, la mesure, le jugement et la décision, qui rend compte d'une démarche d'évaluation s'inscrivant dans une approche par compétences. Ces étapes renvoient respectivement à la planification de l'évaluation, à sa conception et à la collecte de données, à leur correction et à leur interprétation, ainsi qu'aux suites pédagogiques qui en découlent.



Notre expérimentation se situe à l'étape du jugement, puisqu'elle porte sur l'analyse d'une production écrite de type diagnostique par la professeure-chercheuse qui encadre les personnes étudiantes dans le cours, sans pour autant en assumer la responsabilité pédagogique ni les choix relatifs au contenu ou aux modalités d'évaluation. Ce jugement s'inscrit principalement dans une perspective critériée, reposant sur des critères partiellement explicités en amont et limités à la finalité diagnostique de la tâche, au nombre de mots et à l'exposition de la motivation. Cependant, ces critères sont précisés au moment de la correction par la responsable de l'encadrement. Par ailleurs, ce jugement comporte également une dimension dynamique, puisqu'il vise à rendre compte des apprentissages à venir et des difficultés rencontrées. Nous décrivons, dans la méthodologie, la mise à l'épreuve empirique de ces différentes étapes.

La dimension éthique du jugement évaluatif, sans constituer un axe distinct, est intrinsèquement liée au processus d'évaluation que nous venons de décrire. En effet, dans la mesure où le jugement repose à la fois sur la sélection, l'interprétation et le sens attribué aux traces d'apprentissage, il implique des considérations de responsabilité, d'équité et de transparence. Dans notre recherche, celles-ci prennent la forme d'interprétations produites à partir de textes diagnostiques. La manière dont certaines dimensions des productions sont retenues, pondérées ou mises en évidence soulève, de ce fait, des questions relatives à la validité des interprétations, aux risques de biais ainsi qu'à la possible réduction du jugement à des dimensions strictement techniques, particulièrement dans un contexte d'usage de l'IAG. En sciences de l'éducation, cet aspect est constitutif de l'acte d'évaluer en raison de son caractère interprétatif et des effets concrets qu'il produit sur les personnes étudiantes et enseignantes, ainsi que sur leur parcours. Pour en rendre compte, les travaux scientifiques sont fréquemment examinés selon des aires linguistiques distinctes. Celles-ci correspondent moins à des frontières culturelles qu'à des traditions pédagogiques et institutionnelles qui ont façonné les conceptions de l'évaluation et du jugement professionnel (Broadfoot, 1996; De Ketele, 2006; Gipps, 1999). Les contrastes et les complémentarités entre certaines traditions scientifiques peuvent notamment apparaître dans l'étude de l'éthique du jugement évaluatif. Dans les travaux francophones, nous constatons qu'elle est d'abord liée à la responsabilité professionnelle et morale de la personne évaluatrice : expliciter ses critères, viser l'équité et exercer un discernement interprétatif sont alors au cœur de l'acte d'évaluer (Hadji, 2012; Perrenoud, 1998; Roblez, 2025; Scallan, 2015) en plus d'établir les bons usages de l'IAG. Pour leur part, les recherches anglophones proposent un autre point de vue : elles s'appuient sur les notions de *validity* (validité), de *fairness* (équité) et d'*accountability* (responsabilité de la personne évaluatrice) pour cerner la responsabilité des interprétations produites et les effets concrets des usages de l'évaluation (Brookhart, 2011, 2013; Gipps, 1994; Messick, 1989, 1994). Du côté hispanophone, des auteurs inscrivent davantage l'éthique du jugement évaluatif dans une perspective de justice éducative et de respect de la personne, et s'opposent à toute conception strictement techniciste du jugement (Álvarez Méndez, 2001; Santos Guerra, 2010).

Ce survol nous permet de dégager un socle commun de préoccupations éthiques : l'engagement de la responsabilité humaine dans l'interprétation et la décision évaluatives. Or, l'exercice du jugement évaluatif, désormais susceptible d'être assisté par des modèles d'IAG, rend plus manifeste, selon nous, la manière dont il se déploie ainsi que la responsabilité humaine qui lui demeure indissociablement liée. En nous appuyant sur ces conceptions éthiques du jugement évaluatif, nous ne visons ni à déléguer ni à automatiser l'acte de juger, mais à examiner les conditions dans lesquelles une IAG peut être appelée comme outil de soutien au jugement professionnel de la personne qui évalue. Notre expérimentation prend appui sur une démarche exploratoire et réflexive fondée sur la comparaison entre le jugement humain et le jugement produit par une IAG. Nous voulons ainsi éprouver les exigences éthiques identifiées dans la littérature, telles que la responsabilité, le discernement et l'attention portée aux effets du jugement, tout en examinant comment l'IAG peut éclairer les pratiques évaluatives humaines.



1.2 Comparaison entre l'évaluation humaine et l'évaluation à l'aide de l'IAG

Plusieurs chercheurs ont comparé l'évaluation de productions écrites par des évaluateurs humains et par des modèles d'IAG. La synthèse de Li *et al.* (2025), portant sur 65 études consacrées à l'utilisation des grands modèles de langage (LLM) dans le champ de l'*Automated Essay Scoring* (AES), qui consiste à attribuer automatiquement une note à des productions écrites à partir du traitement automatique du langage (Page, 1968), fait ressortir que ces travaux se concentrent principalement sur la notation, sans qu'il soit question d'un jugement évaluatif basé sur l'interprétation, la délibération, la rétroaction ou la responsabilité pédagogique. Elle met néanmoins en évidence une concordance généralement modérée à bonne entre les notes produites par les modèles et celles attribuées par des personnes évaluatrices, ainsi qu'une forte variabilité des résultats selon les contextes (niveau scolaire, discipline, etc.), les méthodologies (correction globale ou analytique, type de grille, etc.) et des critères d'évaluation retenus (qualité linguistique, organisation textuelle, etc.).

Dans le domaine de la didactique des langues, ce type de recherche comparative est courant. En enseignement de l'anglais langue étrangère (*English as a Foreign Language*, EFL), des travaux relèvent une forte concordance entre les notes attribuées par des modèles génératifs et celles du corps enseignant, en particulier pour l'évaluation d'aspects linguistiques au secondaire (Kemal et Liman-Kaban, 2024). Toutefois, d'autres études empiriques nuancent ce constat. À partir de 43 productions écrites en anglais langue seconde rédigées par une cohorte étudiante universitaire, évaluées par cinq personnes correctrices et par ChatGPT 3.5, l'équipe de Geçkini montre que les corrélations interjuges entre les personnes évaluatrices sont significativement plus élevées que celles observées entre les jugements humains et ceux produits par l'IAG (Geçkini *et al.*, 2023). Elle souligne que ChatGPT ne présente qu'un accord faible à modéré avec certaines personnes évaluatrices, ce qui la mène à envisager son usage en appui au jugement évaluatif plutôt qu'en substitut. De plus, Anistyasari *et al.* (2025) corroborent ce constat après avoir observé une faible cohérence entre les notes établies par les personnes correctrices et celles de ChatGPT lors de l'évaluation de 72 écrits dans le cadre d'un cours universitaire de rédaction scientifique en anglais. Les indices de fiabilité relevés sont faibles, voire négatifs, et l'IAG a tendance à surestimer les performances. En effet, elle attribue des notes plus élevées et moins dispersées que celles des personnes correctrices. Ainsi, les auteurs font-ils valoir que cette faible variance limite la capacité des modèles à discriminer finement la qualité des textes, ce qui constitue un enjeu important lorsque l'évaluation vise à soutenir les apprentissages.

Parallèlement, quelques contributions menées sur des essais de maîtrise dans d'autres disciplines de l'enseignement supérieur que celle de la didactique des langues montrent que, malgré une relative proximité des notes, des écarts importants entre l'IAG et le corps enseignant persistent quant à la cohérence et à la sensibilité au contexte disciplinaire (60 essais en sciences politiques et sociologie : Lundgren, 2024; 40 étudiants en marketing et gestion : García-Varela *et al.*, 2025). Ces constats empiriques s'inscrivent, à notre sens, dans le prolongement de réflexions plus générales portant sur la place de l'IAG dans l'évaluation universitaire. Nous songeons, entre autres, aux contributions qui interrogent une conception de l'évaluation qui serait réduite à l'étape de la notation automatisée et soulignent la nécessité de préserver les dimensions interprétative, délibérative et formatrice du jugement évaluatif en enseignement supérieur (Hannachi et Araújo, 2025; Roblez, 2025).

Les recherches comparant le jugement évaluatif de l'IA générative et celui de l'humain ne portent pas uniquement sur la notation, mais également sur la rétroaction. À cet égard, Tremblay *et al.* (2025) ont examiné les usages de l'IA générative pour fournir de la rétroaction sur les productions étudiantes en enseignement supérieur. Leur recension met en lumière la diversité des outils utilisés et des rétroactions produites, tout en soulignant la nécessité de mieux documenter les approches conceptuelles qui sous-tendent leur conception. Sous cet angle de la rétroaction, l'étude de Steiss *et al.* (2024) analyse 198 textes



rédigés en anglais (langue de scolarisation) par des élèves du secondaire. Elle met en évidence que les personnes évaluatrices produisent une rétroaction globalement supérieure à celle de ChatGPT, notamment en termes de clarté, d'exactitude, de priorisation des enjeux et de ton, tandis que l'IAG se distingue par un ancrage plus systématique dans les critères d'évaluation. Les chercheurs concluent que la rétroaction de l'IAG peut jouer un rôle d'appui ponctuel, en particulier en raison de son alignement systématique sur des critères d'évaluation précis, mais que sa contribution ne devrait être envisagée que dans une configuration où le jugement humain demeure central.

Nous pouvons établir un parallèle théorique avec la réflexion de Hannachi et Araújo (2025). Ces auteurs développent une analyse didactique et conceptuelle de l'usage de l'IAG dans l'enseignement et l'évaluation de l'écriture en français à l'université. Leur réflexion les conduit à voir les outils d'IAG comme une ressource pertinente pour la médiation de l'écriture, car ils sont capables d'apporter une aide pour certaines opérations du processus scriptural, par exemple l'explication de normes, la reformulation ou la structuration. Elles en dégagent aussi les limites structurelles, comme le caractère générique des textes produits et l'absence d'une énonciation pédagogique. En ce sens, nous tendrions déjà à considérer que, si l'IAG peut soutenir certaines opérations pour évaluer l'écrit, elle ne peut en assumer ni le travail d'appropriation pédagogique ni la prise de décision évaluative située, qui incombe au corps enseignant.

À la lumière de ces recherches, la question centrale ne nous paraît plus seulement être celle de l'efficacité ou de la fiabilité de l'IAG, mais plutôt celle de la nature du jugement évaluatif qu'elle soutient. Nous interrogeons ainsi la responsabilité éthique de l'acte d'évaluer en convoquant l'approche de Roblez, qui est précisément centrée sur les fondements, les valeurs et les effets éducatifs du jugement (Roblez, 2025). Selon lui, l'évaluation relève d'une intelligence évaluative irréductible à toute automatisation. Elle consiste en une activité humaine de jugement qui suppose trois actions : 1) interpréter les productions au regard d'un contexte pédagogique donné; 2) hiérarchiser les critères pertinents; 3) assumer les effets éducatifs, académiques et symboliques de la décision évaluative. La proposition de Roblez rejoint celle de Jorro et Mercier-Brunel (2022), qui lui est antérieure, pour qui le jugement évaluatif constitue déjà une pratique professionnelle engagée. Ces chercheurs l'appréhendent comme indissociable d'une éthique de la relation et de la responsabilité du corps enseignant face aux effets éducatifs de l'évaluation, tant sur l'élève que sur le corps enseignant lui-même (Jorro et Mercier-Brunel, 2022).

Dans ce paysage de recherche, force est de constater que les travaux consacrés à l'usage de l'IAG pour soutenir l'évaluation de la maîtrise de la langue française écrite en milieu universitaire demeurent peu nombreux. À l'exception de contributions récentes à visée théorique (Hannachi et Araújo, 2025), le champ reste dominé par des études portant sur d'autres langues (en particulier l'anglais) ou sur la notation automatisée. Ces travaux sont souvent détachés des enjeux propres à l'évaluation des compétences langagières en français. De plus, les recherches existantes accordent peu d'attention à l'ensemble des étapes du jugement évaluatif, qui plus est dans le contexte du français écrit à l'université, entre autres l'identification des acquis et des besoins, ainsi que la prise de décisions pédagogiques éclairées.

Afin d'interroger empiriquement le jugement évaluatif à l'ère de l'IAG, nous avons conçu une expérimentation menée dans un cours universitaire portant sur la maîtrise de la langue française écrite. Celle-ci explore, de manière située et critique, les apports et les limites de l'IAG comme outil d'aide au jugement évaluatif.



2. Méthodologie

2.1 Expérimentation

Notre expérimentation était exploratoire et s'est déroulée dans le cadre d'un enseignement universitaire à distance en mode asynchrone. Nous avons adopté une démarche comparative articulée autour des étapes de l'évaluation et de la conception des tâches évaluatives. En amont de notre expérimentation, la professeure responsable du cours de maîtrise du français écrit a assuré la planification, la conception et la collecte des travaux sur la plateforme numérique du cours. Pour les étapes de correction, d'interprétation et de communication, l'une d'entre nous a agi comme évaluatrice humaine, tandis que l'autre a soumis les productions au prisme de l'IAG. Tout d'abord, il a été question d'évaluer le premier travail noté d'une cohorte étudiante (111 travaux), soit un texte d'environ 250 mots présentant, en début de cours, ses motivations à le suivre. Cette évaluation remplit une fonction diagnostique et formative : elle permet au corps enseignant de repérer les principales difficultés en français écrit, plus particulièrement en grammaire et en orthographe (lexicale et grammaticale), en plus d'amener la personne étudiante à prendre conscience de son niveau de maîtrise et des défis qu'elle aura à relever dans le cours.

Pour la construction du travail évaluatif, nous avons considéré cinq aspects complémentaires, présentés dans le tableau 1. Nous précisons toutefois que leur affichage différerait selon le type d'évaluateur : humain ou IAG. Dans l'évaluation réalisée par la personne encadrante, les erreurs catégorisées apparaissaient en marge du texte, à l'aide de la fonction « commentaire » du logiciel Word, tandis que, dans celle effectuée par l'IAG, la catégorisation figurait sous le texte, les erreurs étant signalées en gras.

Tableau 1

Construction du travail évaluatif

5 aspects	
1	Repérage des erreurs directement dans le texte
2	Catégorisation des erreurs en fonction des ordres de difficultés
3	Établissement des critères de notation
4	Notation proprement dite
5	Rédaction d'une rétroaction ajustable

Note. © Autrices, 2026.

Quant aux étapes de l'expérience, elles se sont déroulées comme suit. Tout d'abord, la phase consacrée à l'évaluation par l'une des chercheuses a été réalisée durant la session d'été 2025. Ensuite, une fois le cours terminé et la note finale attribuée aux personnes étudiantes, l'autre chercheuse a interrogé un modèle génératif afin d'évaluer le même travail. Copilot (Microsoft, 2024), dans sa version Premium, a été retenu, car il est préconisé par l'institution universitaire où s'est déroulée l'expérience. Comme le cours était déjà terminé, cette étape n'a eu aucune incidence sur la note finale obtenue par les personnes étudiantes et n'a exigé aucune certification éthique. Toutefois, nous leur avons demandé leur consentement au préalable afin de respecter leur propriété intellectuelle. De plus, par souci de confidentialité, toutes les données personnelles ou sensibles ont été supprimées avant la soumission des textes à l'IAG. Enfin, les deux chercheuses ont réalisé l'analyse comparative des résultats, suivie de discussions et de leurs conclusions.



2.2 Grille de correction et requête

Nous avons établi la grille de correction avant le début de l'expérimentation, en définissant les différentes catégories linguistiques prises en considération. Nous avons également fixé les modalités de notation, avec une pondération de 0,2 point sur 5 par erreur diagnostiquée, de même que les principes encadrant la rédaction de la rétroaction, celle-ci devant être ajustable en fonction des catégories d'erreurs surreprésentées et de la note attribuée.

Parallèlement, sans formation spécifique ni protocoles théoriques préexistants, nous avons élaboré et testé les requêtes destinées à l'IAG en suivant un processus itératif d'essais et d'ajustements. À des fins d'intégrité et de confidentialité, nous avons mené ces tests à partir de textes que nous avons nous-mêmes rédigés. Cette démarche nous a permis de mieux comprendre le fonctionnement du modèle, à la fois dans son interprétation des consignes, dans la structuration de ses réponses et dans sa réaction aux différents degrés de précision lorsque nous lui formulons des demandes. Après plusieurs essais, nous avons fondu nos demandes en une requête unique (voir annexe A) afin de faciliter notre travail avec l'IAG et d'assurer un cadre uniforme et équitable pour toutes les productions étudiantes qui lui étaient soumises.

Nous avons construit notre requête de façon progressive, avec des objectifs distincts à chacune des étapes du travail évaluatif. Il fallait d'abord ancrer l'évaluation dans un contexte universitaire précis, en indiquant le type de cours, la nature de la tâche écrite et sa fonction pédagogique. Ensuite, nous devions poser les balises de l'évaluation, en identifiant expressément les catégories linguistiques à prendre en compte : la grammaire, le code orthographique ainsi que les orthographes d'usage et grammaticale. Au moyen de consignes portant sur le repérage et la catégorisation des erreurs, nous souhaitons que l'IAG produise un diagnostic linguistique structuré et comparable à celui d'une personne évaluatrice. Quant à la partie consacrée à la notation, notre but était de réduire au minimum la marge d'interprétation. Pour ce faire, nous avons proposé une pondération avec des bornes chiffrées précises, lesquelles permettraient au modèle de poser un diagnostic et d'attribuer une note. Enfin, nos indications détaillées concernant la rétroaction avaient trois finalités : tester la capacité de l'IAG à adapter un commentaire évaluatif à une situation donnée, à hiérarchiser les difficultés identifiées et à produire une rétroaction ajustée tout en respectant le cadre énonciatif que nous imposions. L'ensemble de cette requête témoigne ainsi du travail de précision tant lexicale que méthodologique, que nous avons réalisé afin que le modèle d'IAG effectue les tâches attendues.

2.3 Corpus et démarche d'analyse de données

Les analyses reposent sur un corpus de 111 productions écrites réalisées dans le cadre du cours universitaire FRA1002 – Maîtrise du français écrit. Chaque production a été évaluée de manière indépendante par une évaluatrice humaine, puis par une IAG institutionnelle (Copilot), à partir d'une requête structurée intégrant plusieurs tâches évaluatives. Afin de documenter le jugement évaluatif produit par ces deux modalités, nous avons collecté des données quantitatives et qualitatives. Les données quantitatives correspondent au nombre d'erreurs relevées dans chaque production ainsi qu'aux notes attribuées, exprimées sur une échelle de 0 à 5. Quant aux données qualitatives, il s'agit des commentaires consignés par l'évaluatrice humaine au cours du processus de correction.

Nous avons d'abord compilé l'ensemble des données quantitatives dans un fichier Excel, où chaque production anonymisée était associée à un identifiant. Pour chaque copie, nous avons consigné le nombre d'erreurs relevées par l'évaluatrice et par l'IAG, les notes attribuées ainsi que les écarts observés pour ces deux variables. Ce fichier a ensuite été importé dans le logiciel RStudio (Posit team, 2025) afin de réaliser les analyses statistiques.



Les commentaires visaient à documenter les écarts entre les deux modalités d'évaluation, les difficultés rencontrées et les ajustements effectués en cours de processus. Ils portent principalement sur le repérage des erreurs, leur catégorisation linguistique, l'application des consignes et le calcul de la note. Compte tenu du nombre limité de commentaires ($n = 122$), majoritairement issus de l'évaluation par l'IAG, nous avons privilégié un traitement manuel de ces données.

Notre démarche s'appuie sur une analyse de contenu (Bardin, 2013; Miles et Huberman, 2009) visant à décrire et à interpréter les phénomènes observés à partir des contenus textuels. Chaque paire de commentaires constitue une unité de contexte, tandis que chaque commentaire ou segment correspond à une unité d'enregistrement au sens de Bardin (2013). L'analyse s'est déroulée en trois étapes : 1) une phase de préanalyse consiste à l'identification des thèmes récurrents et des axes d'analyse; 2) l'élaboration d'un système de catégories combinant des catégories déductives issues des questions de recherche et des catégories inductives émergentes du corpus (Wanlin, 2007); 3) une phase d'interprétation portant sur la fréquence, la cooccurrence et la signification des catégories au regard des résultats quantitatifs.

3. Résultats

Cette section présente les analyses issues de la comparaison entre les évaluations produites par l'évaluatrice humaine et par une IAG. Nous y examinons la convergence, la cohérence et la stabilité des jugements, tant pour la note attribuée que pour le nombre d'erreurs identifiées, à partir d'analyses quantitatives complémentaires. Les résultats de l'analyse qualitative des commentaires produits au cours du processus de correction sont également présentés.

3.1 Résultats quantitatifs

3.1.1 STATISTIQUES DESCRIPTIVES DES ÉVALUATIONS

Les statistiques descriptives permettent de comparer les tendances centrales et la dispersion des évaluations produites par l'évaluatrice humaine et par l'IAG. Elles offrent un premier niveau d'analyse des écarts entre les deux modalités d'évaluation.

Les résultats de notre expérimentation montrent des différences moyennes faibles entre les deux modes d'évaluation. La moyenne des notes attribuées par l'évaluatrice humaine est de 3,85/5, contre 3,80/5 pour l'IAG, soit une différence moyenne de -0,06 point. Pour le nombre d'erreurs, l'évaluatrice humaine en identifie en moyenne 6,75 par production, comparativement à 6,19 pour l'IAG, soit une différence moyenne de -0,56 erreur. Un test de Wilcoxon pour échantillons appariés confirme cette observation et ne met pas en évidence de différence statistiquement significative entre les deux modalités ($p > 0,05$), tant pour les notes que pour le nombre d'erreurs.

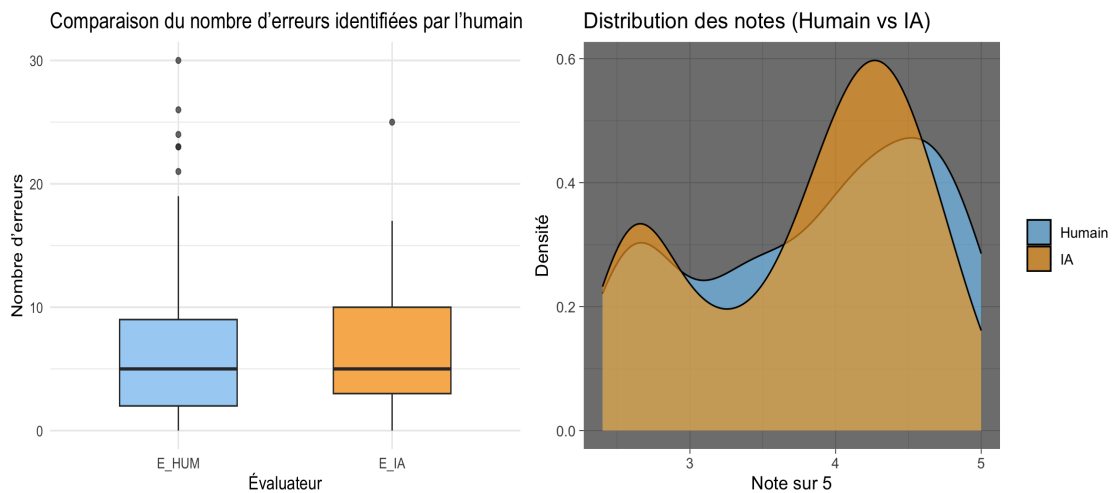
Les distributions des notes (figure 1) présentent une forme bimodale et se chevauchent largement entre les deux modalités. La distribution associée à l'IAG apparaît toutefois légèrement plus concentrée, tandis que celle de l'évaluatrice humaine présente une dispersion plus étendue. Une tendance similaire est observée pour le nombre d'erreurs, avec une variabilité légèrement plus marquée du côté humain, notamment en raison de la présence de valeurs élevées.

Dans l'ensemble, ces résultats descriptifs indiquent que les évaluations produites par l'IAG s'inscrivent dans un ordre de grandeur comparable à celles de l'évaluatrice humaine, sans mettre en évidence, à ce stade, de tendance systématique marquée à la surévaluation ou à la sous-évaluation.



Figure 1

Comparaison descriptive des évaluations de l'évaluatrice humaine et de l'IAG : distributions des notes et du nombre d'erreurs



Note. © Autrices, 2026.

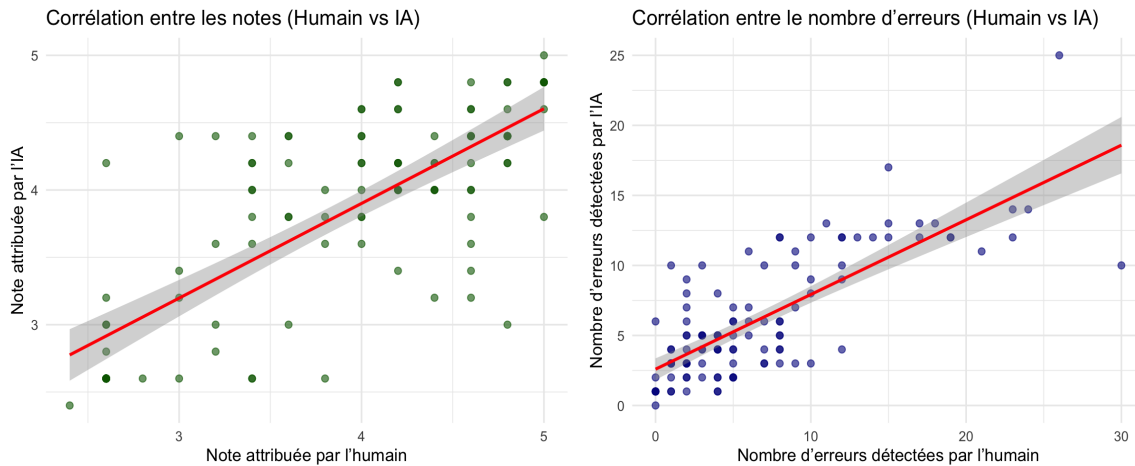
3.1.2 COHÉRENCE DIRECTIONNELLE ENTRE L'ÉVALUATRICE HUMAINE ET L'IAG (CORRÉLATION DE SPEARMAN)

Afin d'examiner dans quelle mesure l'évaluatrice humaine et l'IAG classent les productions écrites de manière comparable, une analyse de corrélation de Spearman a été réalisée (figure 2). Le recours à ce coefficient non paramétrique se justifie par la non-normalité des distributions (tests de Shapiro-Wilk, $p < 0,05$), ainsi que par la nature discrète des variables analysées. Les résultats mettent en évidence une corrélation positive élevée tant pour le nombre d'erreurs relevées ($\rho = 0,77$, $p < 0,001$) que pour les notes attribuées ($\rho = 0,76$, $p < 0,001$), indiquant une forte cohérence dans le classement relatif des productions écrites par les deux modalités d'évaluation.



Figure 2

Corrélation de Spearman entre les évaluations de l'évaluatrice humaine et de l'IAG pour la note attribuée et le nombre d'erreurs identifiées



Note. © Autrices, 2026.

Cette cohérence directionnelle n'en préjuge pas néanmoins du niveau d'accord absolu entre les évaluations, ce qui justifie le recours à des analyses complémentaires.

3.1.3 FIABILITÉ ET ACCORD ABSOLU ENTRE L'ÉVALUATRICE HUMAINE ET L'IAG (COEFFICIENT DE CORRÉLATION INTRACLASSE – ICC)

La fiabilité de l'accord absolu entre l'évaluation humaine et l'évaluation automatisée a été examinée à l'aide du coefficient de corrélation intraclass de type ICC(A,1). Ce choix a été retenu, car l'étude compare deux systèmes d'évaluation fixes (une évaluatrice humaine et une IAG) à partir d'une mesure unique par évaluateur, dans l'objectif d'estimer un accord absolu plutôt qu'une simple cohérence de classement (Koo et Li, 2016).

Les résultats indiquent un accord bon à excellent pour la note attribuée (ICC = 0,755; IC à 95 % [0,66-0,83]) et un accord bon pour le nombre d'erreurs identifiées (ICC = 0,719; IC à 95 % [0,62-0,80]) (tableau 2). Ces valeurs témoignent d'une concordance globale satisfaisante entre l'IAG et l'évaluatrice humaine, l'accord étant légèrement plus élevé pour la note globale que pour le repérage précis des erreurs. Ces coefficients montrent que, pour les deux indicateurs, la majeure partie de la variance observée est attribuable aux différences entre les productions écrites elles-mêmes plutôt qu'à des désaccords entre les deux systèmes d'évaluation.



Tableau 2

Fiabilité et accord absolu entre l'évaluatrice humaine et l'IAG selon le coefficient de corrélation intraclasse (ICC(A,1))

Variable évaluée	ICC(A,1)	Intervalle de confiance (95 %)	P-value	Niveau d'accord
Nombre d'erreurs identifiées	0,719	[0,616; 0,798]	< 0,001	Bon
Note attribuée	0,755	[0,663; 0,825]	< 0,001	Bon à excellent

Note. © Autrices, 2026.

3.1.4 BIAIS MOYEN ET DE LA STABILITÉ DES ÉVALUATIONS ENTRE L'ÉVALUATRICE HUMAINE ET L'IAG (MÉTHODE DE BLAND-ALTMAN)

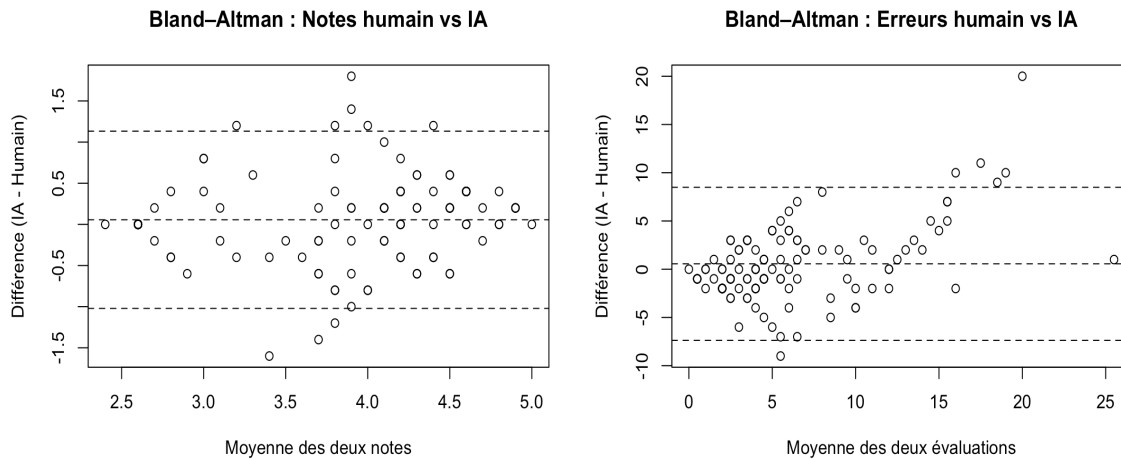
Les diagrammes de Bland-Altman ont été utilisés afin de compléter l'analyse de la fiabilité globale estimée par l'ICC, en examinant le biais moyen ainsi que la stabilité des écarts entre les évaluations de l'évaluatrice humaine et de l'IAG à l'échelle des productions individuelles. En cohérence avec les résultats de l'ICC, qui indiquent un accord absolu bon à excellent pour la note attribuée et bon pour le nombre d'erreurs relevées, les graphiques montrent que, pour les deux indicateurs, les différences entre les évaluations se concentrent globalement autour de zéro, ce qui atteste l'absence de biais directionnel systématique de la part de l'IAG.

Pour la note attribuée (figure 3), les limites d'accord se situent approximativement à $\pm 1,5$ point sur une échelle de 0 à 5, et la majorité des observations demeure à l'intérieur de ces bornes. L'absence d'association entre l'amplitude des écarts et le niveau moyen des notes suggère une stabilité de la notation automatisée sur l'ensemble des niveaux de performance, en concordance avec la fiabilité élevée mise en évidence par l'ICC. En ce qui concerne le nombre d'erreurs identifiées (figure 3), l'analyse de Bland-Altman met également en évidence un biais moyen proche de zéro, ce qui confirme l'absence de tendance systématique de l'IAG à relever davantage ou moins d'erreurs que l'évaluatrice humaine. Toutefois, une dispersion plus importante des écarts apparaît lorsque la charge d'erreurs augmente : les productions peu fautives présentent des différences faibles et proches de zéro, tandis que les textes plus complexes donnent lieu à des écarts plus marqués, ce qui indique une stabilité réduite de l'évaluation automatisée dans ces configurations. Ces résultats viennent ainsi préciser les indices globaux fournis par l'ICC, en soulignant une concordance globalement satisfaisante entre l'évaluatrice humaine et l'IAG, assortie de limites lorsque la tâche évaluative exige un repérage plus fin des erreurs.



Figure 3

Diagrammes de Bland-Altman illustrant le biais moyen et la stabilité des évaluations entre l'évaluatrice humaine et l'IA



Note. © Autrices, 2026.

L'ensemble de ces analyses quantitatives éclaire à la fois le niveau global d'accord entre l'évaluatrice humaine et l'IA, et les limites de stabilité observées selon la complexité des productions.

3.2 Résultats qualitatifs

L'analyse des commentaires confirme une convergence globale entre l'évaluatrice humaine et l'IA. À plusieurs reprises, l'évaluatrice souligne la précision du travail de l'IA au regard de la consigne ainsi que sa concordance avec sa propre évaluation : « L'IA réalise le travail d'une manière presque calquée à la professeure »; « L'IA s'aligne plutôt bien avec mon propre diagnostic ». Elle mentionne aussi, à l'occasion, une efficacité supérieure de l'IA dans le repérage des erreurs : « L'IA est plus efficace et a repéré deux erreurs de plus que moi ». En revanche, elle relève des divergences récurrentes quant au repérage et à l'interprétation des erreurs dans les copies. Celles-ci concernent principalement leur identification, leur catégorisation linguistique, l'application des consignes et le calcul de la note.

Certains écarts s'expliquent par le fait que l'IA repère parfois des erreurs non relevées par l'évaluatrice, notamment en raison de fatigue ou d'inattention ponctuelle, tandis que d'autres lui échappent. Comme en témoignent certaines observations, l'évaluatrice mentionne, par exemple, « ai oublié 2 erreurs : vœux, soutien » et, dans d'autres cas, signale des erreurs non détectées par l'IA ou différentes de celles relevées par l'humain, telles que « 1 anglicisme, mais différent de H ». Ces constats concordent avec les résultats quantitatifs, indiquant que plus une copie comporte d'erreurs, moins l'IA parvient à les identifier. Ainsi, l'évaluatrice souligne que « Comme souvent, quand le nombre de fautes est très important, l'IA ne diagnostique pas autant de fautes que moi. Cependant, la note finale est assez semblable vu qu'on s'arrête à 2,6 », soit la note minimale sur 5 que l'étudiant peut obtenir. Par ailleurs, l'IA tend parfois à regrouper en une seule occurrence des phénomènes linguistiques que l'évaluatrice distingue en plusieurs, ce qui explique les écarts de décompte sans entraîner systématiquement de variation de la note finale.

Certains commentaires indiquent également la difficulté de l'IA à respecter les consignes de manière stable (ex. : « l'IA a décidé de réécrire des passages »). Cela témoigne d'une sensibilité à la formulation des requêtes ainsi que d'une application parfois inconstante des critères d'évaluation. Cette difficulté persiste parfois malgré des reformulations, bien que des ajustements progressifs soient observés au fil des interactions.



Les résultats font aussi ressortir des divergences dans la catégorisation des erreurs. De nombreux commentaires signalent des désaccords quant à la nature linguistique attribuée à une erreur. L'IAG classe fréquemment comme grammaticales des erreurs que l'évaluatrice interprète plutôt comme relevant de la syntaxe, du lexique ou du style. Elle confond également des problèmes stylistiques avec des erreurs linguistiques ou attribue à certaines erreurs une catégorie inadéquate, ce qui peut avoir des répercussions directes sur la note lorsque ces catégories sont prises en compte dans le calcul. En voici un exemple : « L'IA diagnostique des "constructions grammaticales fautives". Or elle catégorise le lexique dans la grammaire. »

Enfin, l'analyse met en lumière des limites liées à la fiabilité du calcul dans l'évaluation automatisée. Certains commentaires font état d'erreurs arithmétiques, d'une application inexacte des règles de pondération ou de l'inclusion inadéquate de certains éléments, ce qui entraîne des écarts mineurs dans la note finale. La gestion de la dimension stylistique constitue, à cet égard, un point de divergence particulièrement saillant. Malgré des consignes explicites demandant d'exclure ces aspects du calcul, l'IAG tend à les intégrer de manière récurrente ou à les traiter comme des erreurs linguistiques. Comme le souligne l'évaluatrice : « L'IA a inclus les erreurs de style dans le calcul de la note. Même en lui demandant de recalculer, il n'a pas la réponse exacte : 4,2. En réalité, il a le bon nombre d'erreurs et de problèmes diagnostiqués, mais il calcule mal la note. »

Bien que ces situations demeurent relativement peu fréquentes, elles contribuent à expliquer certaines différences observées dans les résultats quantitatifs relatifs aux scores attribués. Ces constats soulignent la nécessité de fournir à l'IAG un gabarit de correction strict, incluant des indications claires sur les règles à appliquer, les rectifications attendues ainsi que l'interdiction de toute réécriture du texte, afin d'obtenir des rétroactions plus cohérentes.

Dans l'ensemble, malgré une concordance globale, les processus sous-jacents à l'évaluation diffèrent partiellement entre l'évaluatrice humaine et l'IAG. Notre analyse qualitative apporte un éclairage sur les écarts observés dans les résultats quantitatifs, lesquels s'expliquent notamment par des différences dans le repérage et la catégorisation des phénomènes linguistiques, ainsi que par des difficultés liées à l'application des consignes et au calcul des notes.

4. Discussion

Dans cette partie, nous discutons des résultats à la lumière du cadre de référence afin de situer la portée et les limites de l'évaluation par l'IAG, ainsi que les conditions dans lesquelles elle peut soutenir le jugement évaluatif humain sans s'y substituer.

Les résultats de notre expérimentation montrent d'abord une convergence globale entre l'évaluatrice et l'IAG, tant pour le nombre d'erreurs que pour la note attribuée, comme en témoignent les analyses descriptives, de corrélation et de fiabilité. Celles-ci suggèrent que l'IAG est en mesure de reproduire un classement des productions écrites comparable à celui établi par le jugement humain. Ces résultats s'inscrivent dans le prolongement des travaux recensés dans la littérature, qui font état d'une bonne concordance entre l'évaluation humaine et l'évaluation automatisée (Anistyasari *et al.*, 2025; Geçkin *et al.*, 2023). Toutefois, l'analyse des écarts entre les deux modalités permet de nuancer cette convergence. Bien que l'IAG ne présente pas de biais directionnel systématique, son évaluation devient plus variable lorsque les textes comportent de nombreuses erreurs. Ce résultat fait écho à celui de Steiss *et al.* (2024), qui soulignent que la performance de l'IA dépend du niveau des copies.



De plus, l'analyse qualitative met en évidence des divergences récurrentes dans la catégorisation et l'interprétation des erreurs. Contrairement à l'évaluatrice humaine, qui catégorise les erreurs linguistiques avec plus de précision, l'IAG produit des classifications plus génériques, parfois inadéquates, en confondant notamment des erreurs grammaticales, lexicales ou stylistiques. L'IAG montre aussi des difficultés à respecter certaines consignes et une instabilité dans l'application des règles de pondération.

Les divergences observées dans les analyses quantitatives et qualitatives peuvent s'expliquer par le processus de construction et de stabilisation de la requête adressée à l'IAG. Bien que cette dernière ait été élaborée avec méthode, à partir d'essais et d'ajustements successifs, plusieurs tensions inhérentes à l'usage de l'IAG pour évaluer ont émergé. En premier lieu, les essais ont révélé une forte sensibilité du modèle à la formulation des consignes : de légères variations lexicales, syntaxiques ou énonciatives pouvaient entraîner des écarts notables dans l'interprétation des catégories linguistiques. Notre reformulation minutieuse a permis de limiter ces glissements et d'amener le modèle à traiter les erreurs plus conformément aux attentes définies dans notre grille de correction. En second lieu, la pondération et les bornes imposées, bien qu'elles aient réduit la variabilité des notes, n'ont pas entièrement neutralisé les différences dans le raisonnement sous-jacent du modèle, notamment dans l'articulation entre le diagnostic linguistique, la notation et la rétroaction. En troisième lieu, notre choix de regrouper le repérage, la catégorisation, la notation et la rétroaction dans une requête unique a révélé les difficultés du modèle à traiter simultanément plusieurs tâches évaluatives : certaines réponses montraient une hiérarchisation inégale des tâches ou une focalisation sur certaines dimensions au détriment d'autres. Or, loin de constituer de simples limites, ces observations constituent des points d'appui empiriques pour interroger la capacité de l'IAG à soutenir un jugement évaluatif complexe.

Ces constats prennent tout leur sens à la lumière du cadre théorique du jugement évaluatif. Comme le définit Roblez (2025), le jugement ne se réduit pas à une opération de mesure ou de notation : il constitue une activité interprétative et délibérative impliquant la mise en relation de critères, leur hiérarchisation ainsi que l'anticipation des effets des décisions évaluatives. Dans le même sens, Jorro et Mercier-Brunel (2022) soulignent la dimension à la fois professionnelle et éthique du jugement, qui engage la responsabilité de la personne évaluatrice dans l'interprétation des productions et dans les décisions qui en découlent. Cependant, les résultats de notre étude montrent que ces dimensions du jugement demeurent difficilement automatisables. Les difficultés de catégorisation, les erreurs de calcul et la sensibilité aux consignes observées chez l'IAG traduisent une absence de véritable délibération évaluative, au sens où l'entendent ces auteurs.

Dès lors, l'IAG apparaît moins comme un substitut au jugement évaluatif que comme un outil susceptible d'en soutenir certaines dimensions. Sa capacité à produire rapidement des rétroactions alignées sur des critères explicites et à repérer un certain nombre d'erreurs peut constituer un appui pertinent dans des contextes de rétroaction formative et sommative, en particulier lorsque les attentes sont clairement définies. Toutefois, comme le soulignent Hannachi et Araújo (2025), cet apport demeure limité, du moins à ce jour, par le caractère souvent générique des productions de l'IAG et par l'absence d'une véritable intention pédagogique. Par conséquent, l'usage de ces outils requiert un encadrement rigoureux ainsi qu'une supervision humaine constante afin d'assurer la cohérence, la pertinence et l'équité des évaluations produites.

Enfin, ces résultats appellent à repenser l'intégration de l'IAG en évaluation, non plus seulement sous l'angle de la performance ou de la fiabilité, mais à partir des exigences éthiques qui encadrent l'acte d'évaluer. Loin de réduire la responsabilité humaine, le recours à l'IAG tend au contraire à la rendre plus visible, car il remet au premier plan la nécessité d'exercer un jugement professionnel éclairé. En contribuant à expliciter ce qui, dans l'acte évaluatif humain, relève de l'interprétation, de la délibération et de la responsabilité, l'IAG peut ainsi être envisagée comme un révélateur des composantes morales et sociales du jugement évaluatif.



5. Conclusion

Cette expérimentation nous permet de formuler des recommandations étayées par nos analyses quantitatives et qualitatives. Tout d'abord, les écarts observés dans le repérage, la catégorisation des erreurs ou dans l'enchaînement entre le diagnostic, la notation et la rétroaction montrent effectivement que l'aide de l'IAG reste partielle : le modèle utilisé devient instable lorsque les erreurs dans un texte se multiplient et demeure très sensible à l'énoncé des consignes. En cohérence avec les travaux de Roblez (2025), de Jorro et Mercier Brunel (2022), de Hannachi et Araújo (2025), ainsi qu'avec les résultats de Steiss et de ses collaborateurs (2024) notamment, nos résultats confirment ipso facto que l'IAG ne peut être substituée au jugement interprétatif, mais qu'elle peut, à condition d'un encadrement rigoureux, soutenir certaines opérations du processus d'évaluation.

Notre analyse indique également que la qualité et la pertinence des réponses produites par l'IAG sont fortement tributaires des balises précisées en amont, soit lors de la rédaction de la requête. Ces difficultés nous conduisent à recommander une réflexion méthodique préalable à l'utilisation de l'IAG en évaluation. Celle-ci s'appuie, entre autres, sur une définition des besoins, une rédaction soignée des consignes et des ajustements successifs pour assurer la stabilité et la fiabilité des réponses. Au regard de ces observations, nous insistons sur la nécessité d'exercer une vigilance constante à l'égard du jugement évaluatif généré par l'IA, en particulier lorsque les textes présentent de nombreuses erreurs linguistiques.

Ensuite, nous mettons en relief combien le fait de poser des balises institutionnelles claires s'avère essentiel pour accompagner le corps professoral dans son jugement évaluatif à l'aide de l'IAG. Bien que des cadres et des réflexions aient déjà été amorcés dans les universités pour encadrer l'usage de l'IAG (voir notamment McDonald *et al.*, 2024; An, Yu et James, 2025), les écarts observés rappellent que l'évaluation assistée par l'IAG détermine des choix à la fois interprétatifs, normatifs et éthiques qui ne reposent pas uniquement sur des pratiques individuelles. Il importe alors de poursuivre et d'ajuster ce travail institutionnel au rythme de l'évolution des modèles génératifs et de leurs performances; préciser les responsabilités liées à l'usage de l'IAG dans l'exercice du jugement évaluatif garantira aux personnes étudiantes une évaluation à la fois équitable et intelligible.

De plus, notre analyse souligne combien la formation du corps enseignant est nécessaire, en particulier dans le contexte de l'évaluation, tant pour formuler des requêtes efficaces que pour en reconnaître les limites, comme la production trop générale de classifications ou la confusion dans certaines catégories linguistiques. Une formation continue aiderait ainsi le corps enseignant à clarifier sa posture quant à son jugement évaluatif lorsqu'il recourt à l'IAG pour en soutenir certaines étapes, sans pour autant que celle-ci ne se substitue à son jugement professionnel.

Notre expérimentation suggère également que l'intégration de l'IAG comme soutien à l'acte évaluatif peut induire un effet autoréflexif chez le corps enseignant lui-même, car celui-ci peut s'interroger sur ses pratiques, telles que la précision de ses critères et la systématisation de son raisonnement évaluatif.

Enfin, cette recherche ouvre sur plusieurs perspectives. Sur le plan scientifique, il apparaît pertinent d'élargir le nombre de personnes évaluatrices afin d'examiner plus finement la fidélité interjuges humains, de comparer différents modèles d'IAG et d'étudier le point de vue étudiant, notamment quant à la compréhension des rétroactions, au sentiment d'équité et à leur reconnaissance. Nous estimons, par ailleurs, que l'évolution rapide des modèles génératifs requiert la poursuite d'études empiriques régulières afin d'en documenter les performances et les limites. Cela nous semble, en effet, une condition indispensable pour utiliser l'IAG en évaluation de façon à la fois éclairée, critique et responsable.



Quoi qu'il en soit, l'évaluation ne peut être réduite à une démarche technique ou instrumentale, a fortiori lorsqu'elle s'appuie sur des outils numériques ou sur l'IAG. Il s'agit d'un processus inscrit dans un cadre axiologique établi sur la justice, l'équité, la rigueur, la cohérence, la transparence et la responsabilité professionnelle de la personne évaluatrice (De Ketele, 2010a, 2010b; Hadji, 2012; Enseigner à l'UQAM, s. d.). Nous rappelons d'ailleurs, avec Hadji (2012), que le jugement évaluatif implique un rapport aux valeurs et aux effets de l'évaluation sur les personnes évaluées. Dans une perspective similaire, Huba et Freed (2000) soulignent qu'une évaluation pleinement formative, comme dans notre expérimentation, ne se limite pas à l'application d'outils ou de critères, mais constitue avant tout une pratique éducative orientée vers le sens, l'apprentissage et la responsabilité pédagogique. Dès lors, que l'acte évaluatif soit assisté ou non par l'IAG, ces valeurs forment un socle incontournable qui doit, selon nous, encadrer le jugement professionnel et en garantir la légitimité à la fois pédagogique et éthique.

Liste des références

- Ahangama, N. (2026). Designing assessments in the generative AI era: A tailored assessment framework for ICT tertiary education. *International Journal of Educational Technology in Higher Education*, 23(9). <https://doi.org/10.1186/s41239-026-00582-0>
- Álvarez Méndez, J. M. (2001). *Evaluar para conocer, examinar para excluir*. Morata.
- An, Y., Yu, J. H. et James, S. (2025). Investigating the higher education institutions' guidelines and policies regarding the use of generative AI in teaching, learning, research, and administration. *International Journal of Educational Technology in Higher Education*, 22(10). <https://doi.org/10.1186/s41239-025-00507-3>
- Anistyasari, Y., Hidayati, S. C., Suparji, S., Ekohariadi, E. et Kusumaningtyas, D. A. (2025). Comparing AI and human assessment of academic writing skills: A kappa analysis. *E3S Web of Conferences*, 645. <https://doi.org/10.1051/e3sconf/202564506014>
- Audran, J. (2024). Cinq enjeux d'évaluation face à l'émergence des IA génératives en éducation. *Mesure et évaluation en éducation*, 47(1), 6-26. <https://doi.org/10.7202/1114564ar>
- Bardin, L. (2013). *L'analyse de contenu*. Presses Universitaires de France. <https://doi.org/10.3917/puf.bard.2013.01>
- Bittle, K. et El Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), 296. <https://doi.org/10.3390/info16040296>
- Black, P. et Wiliam, D. (1998). Assessment and Classroom Learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Black, P. et Wiliam, D. (2009) Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability*, 21, 5-31. <https://doi.org/10.1007/s11092-008-9068-5>
- Bloom, B. S. (1971). *Handbook on formative and summative evaluation of student learning*. McGraw-Hill.
- Broadfoot, P. (1996). *Education, assessment and society: A sociological analysis*. Open University Press.
- Brookhart, S. M. (2011). Educational assessment knowledge and skills for teachers. *Educational Measurement: Issues and Practice*, 30(1), 3-12. <https://doi.org/10.1111/j.1745-3992.2010.00195.x>
- Brookhart, S. M. (2013). Classroom assessment in the context of standards-based reform. *Teachers College Record*, 115(11), 1-16.
- Dabis, A., et Csáki, C. (2024). AI and ethics: Investigating the first policy responses of higher education institutions to the challenge of generative AI. *Humanities and Social Sciences Communications*, 11, 1006. <https://doi.org/10.1057/s41599-024-03526-z>
- De Ketele, J.-M. (1986). *L'évaluation : approche descriptive ou normative?* De Boeck.
- De Ketele, J.-M. (2006). La recherche en évaluation : propos synthétiques et prospectifs. *Mesure et évaluation en éducation*, 29(1), 99-118. <https://doi.org/10.7202/1086970a>
- De Ketele, J.-M. et Gérard, F.-M. (2007). *La qualité et la validation des évaluations*. De Boeck.
- De Ketele, J.-M. (2010a). Ne pas se tromper d'évaluation. *Revue française de linguistique appliquée*, 15(1), 25-37. <https://doi.org/10.3917/rfla.151.0025>



- De Ketele, J.-M. (2010b). *L'évaluation : une problématique complexe*. De Boeck.
- Collin, S. et Marceau, E. (2023). Enjeux éthiques et critiques de l'intelligence artificielle en enseignement supérieur. *Éthique publique*, 24(2). <https://doi.org/10.4000/ethiquepublique.7619>
- Conseil de l'innovation du Québec (2024). *L'intelligence artificielle générative : comprendre, encadrer et agir*. Gouvernement du Québec. <https://www.ethique.gouv.qc.ca/publications/iagenerative/>
- Elshall, A. S. et Badir, A. (2025). Balancing AI-assisted learning and traditional assessment: The FACT assessment in environmental data science education. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1596462>
- Enseigner à l'UQAM. (s. d.). La démarche évaluative et les valeurs qui s'y rattachent. <https://enseigner.uqam.ca/evaluer/demarche-et-valeurs/>
- García-Varela, F., Nussbaum, M., Mendoza, M., Martínez-Troncoso, C. et Bekerman, Z. (2025). ChatGPT as a Stable and Fair Tool for Automated Essay Scoring. *Education Sciences*, 15(8), 946. <https://doi.org/10.3390/educsci15080946>
- Geçkin, V., Kızıltaş, E. et Çınar, Ç. (2023). Assessing second-language academic writing: AI vs. human raters. *Journal of Educational Technology & Online Learning*, 6(4), 1096-1108. <http://doi.org/10.31681/jetol.1336599>
- Ghimire, A., Edwards, J. (2024). From Guidelines to Governance: A Study of AI Policies in Education. Dans Olney, A.M., Chounta, I.A., Liu, Z., Santos, O.C., Bittencourt, I.I. (dir.) *Artificial Intelligence in Education. Posters and Late Breaking Results, Workshops and Tutorials, Industry and Innovation Tracks, Practitioners, Doctoral Consortium and Blue Sky. AIED 2024. Communications in Computer and Information Science* (vol. 2151, p.299-307). Springer, Cham. https://doi.org/10.1007/978-3-031-64312-5_36
- Gipps, C. V. (1994). *Beyond testing: Towards a theory of educational assessment*. Falmer Press.
- Gipps, C. V. (1999). Socio-cultural aspects of assessment. *Review of Research in Education*, 24, 355-392.
- Hadji, C. (1997a). *L'évaluation démystifiée*. ESF.
- Hadji, C. (1997b). Pour une éthique de l'agir évaluatif. *Mesure et évaluation en éducation*, 20(2), 7-26. <https://doi.org/10.7202/1091380ar>
- Hadji, C. (2012). *Faut-il avoir peur de l'évaluation?* De Boeck.
- Hannachi, R. et Araújo, S. (2025). Vers une évaluation humaine et automatisée de l'écrit avec l'intelligence artificielle générative. *Les Langues modernes*, 119(1). <https://hal.science/hal-05220332v1>
- Huba, M. E. et Freed, J. E. (2000). *Learner-centered assessment on college campuses*. Allyn & Bacon.
- Ilieva, G., Yankova, T., Ruseva, M. et Kabaivanov, S. (2025). A Framework for Generative AI-Driven Assessment in Higher Education. *Information*, 16(6), 472. <https://doi.org/10.3390/info16060472>
- Jorro, A. et Mercier-Brunel, Y. (2022). Évaluation. Dans A. Jorro (dir.), *Dictionnaire des concepts de la professionnalisation* (p. 197-200). De Boeck Supérieur.
- Kemal, S. et Liman-Kaban, A. (2024). Comparative Analysis of Human Graders and AI in Assessing Secondary School EFL Journal Writing. *Asian Journal of Distance Education*, 20(1). <https://www.asianjde.com/ojs/index.php/AsianJDE/article/view/796>
- Kılınç, S. (2024). Comprehensive AI assessment framework: Enhancing educational evaluation with ethical AI integration. *Journal of Educational Technology & Online Learning*, 7(4), 521-540. <https://doi.org/10.31681/jetol.1492695>
- Koo, T. K. et Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155-163. <https://jcm.kglmeridian.com/view/journals/jcm/15/2/article-p155.xml>
- Leroux, J. L. et Bélair, L. (2015). Exercer son jugement professionnel. Dans J. L. Leroux (dir.), *Évaluer les compétences* (p. 65-107). Chenelière.
- Li, H., Chen, C. H., Fan, K., Young-Johnson, C., Lim, S. ET Feng, Y. (2025). *Agreement between large language models and human raters in essay scoring: A research synthesis*. arXiv. <https://doi.org/10.48550/arXiv.2512.14561>
- Lund, B. D., Lee, T. H., Mannuru, N. R., et Arutla, N. (2025). AI and academic integrity: Exploring student perceptions and implications for higher education. *Journal of Academic Ethics*, 23, 1545-1565. <https://doi.org/10.1007/s10805-025-09613-3>
- Lundgren, M. (2024). *Large Language Models in Student Assessment: Comparing ChatGPT and Human Graders*. arXiv. <https://doi.org/10.48550/arXiv.2406.16510>



- McDonald, N., Johri, A., Ali, A., Hingle, A., 2024. Generative artificial intelligence in higher education: Evidence from an analysis of institutional policies and guidelines. *Computers in Human Behavior: Artificial Humans*, 3. <https://doi.org/10.1016/j.chbah.2025.100121>
- Messick, S. (1989). Validity. Dans R. L. Linn (dir.), *Educational measurement* (3^e éd., p. 13-103). Macmillan.
- Messick, S. (1994). The Interplay of Evidence and Consequences in the Validation of Performance Assessments. *Educational Researcher*, 23(2), 13-23. <https://doi.org/10.3102/0013189X023002013>
- Microsoft (2024). *Microsoft Copilot*. <https://www.microsoft.com/copilot>
- Miles, M. et Huberman, A. (2009). *Analyse des données qualitatives*. De Boeck.
- Mpolomoka, D. L. (2025). Utilizing artificial intelligence for assessment. *Pedagogical Research*, 10(3). <https://doi.org/10.29333/pr/16677>
- OpenAI. (2024). GPT-4 technical report. ArXiv. <https://arxiv.org/abs/2303.08774>
- Organisation de coopération et de développement économiques (OCDE). (2021). *Recommandation du Conseil sur l'intelligence artificielle*. OCDE. <https://legalinstruments.oecd.org/fr/instruments/OECD-LEGAL-0449>
- Page, E. B. (1968). The use of the computer in analyzing student essays. *International Review of Education*, 14, 210-225. <https://doi.org/10.1007/BF01419938>.
- Perrenoud, P. (1998). *L'évaluation des élèves : de la fabrication de l'excellence à la régulation des apprentissages*. De Boeck Université.
- Perrenoud, P. (2004). *Évaluer des compétences*. De Boeck.
- Peterson, S. (2025). Addressing student use of generative AI in schools and universities through academic integrity reporting. *Frontiers in Education*, 10. <https://doi.org/10.3389/feduc.2025.1610836>
- Posit team (2025). *RStudio: Integrated Development Environment for R*. Posit Software, PBC, Boston, MA. URL <http://www.posit.co/>
- Roblez, A. (2025, 25 février). *Intelligence évaluative et évaluation artificielle : réflexions sur la relation entre l'IA et l'évaluation* [communication orale]. Webinaire du Réseau de la pédagogie, Institut catholique de Lille. <https://hal.science/hal-04967764v1>
- Santos Guerra, M. A. (2010). *La evaluación como aprendizaje*. Kairós.
- Scallon, G. (2015). *Des savoirs aux compétences : exploration en évaluation des apprentissages*. De Boeck.
- Scriven, M. (1967). The methodology of evaluation. Dans R. W. Tyler, R. M. Gagne et M. Scriven (dir.), *Perspectives of curriculum evaluation* (p. 39-83). Rand McNally.
- Scriven, M. (1991). *Evaluation thesaurus* (4^e éd.). Sage.
- Steiss, J., Tate, Tamara., Graham, S., Gruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., Olsan, C.B. (2024). Comparing human and ChatGPT feedback. *Learning and Instruction*, 91. <https://doi.org/10.1016/j.learninstruc.2024.101894>
- Tremblay, C., Miklohou, S. et Poellhuber, B. (2025). Recension des usages d'intelligences artificielles génératives (IAg) pour offrir de la rétroaction en enseignement supérieur. *International Journal of Technologies in Higher Education*, 22(1). <https://doi.org/10.18162/ritpu-2025-v22n1-11>
- UNESCO (2023). *Guidance for generative AI in education and research*. UNESCO.
- Wanlin, P. (2007). L'analyse de contenu comme méthode d'analyse qualitative d'entretiens : une comparaison entre les traitements manuels et l'utilisation de logiciels. *Recherches qualitatives*, hors série, 3, 243-272. <https://www.erudit.org/fr/livres/collection-hors-serie-les-actes-de-la-revue-recherches-qualitatives/bilan-et-prospectives-de-la-recherche-qualitative-en-sciences-humaines-et-sociales/5174co/>
- Zhao, J., Chapman, E. et Ghassemi Pour Sabet, P. (2024). Generative AI and educational assessments: A systematic review. *Education Research & Perspectives*, 51(31), 124-155. <https://doi.org/10.70953/ERPv51.2412006>



Annexe A

Requête soumise à l'IAG

Bonjour, je suis professeure dans une université à distance. J'enseigne un cours dont les contenus sont axés sur la maîtrise du français écrit. Les étudiant·es doivent produire un premier texte évalué qui consiste en un texte de motivation.

Pour l'évaluation, plusieurs aspects sont pris en considération :

- Grammaire : accords, correspondance des temps, emploi des pronoms relatifs, anglicismes, etc.
- Code orthographique : ponctuation, majuscule, numération, etc.
- Orthographe d'usage et orthographe grammaticale : redoublement de consonnes, homonymes, genre des noms, « tout », « quelque », mots composés, conjugaisons, etc.
- Les rectifications orthographiques du français (nouvelle orthographe) : il faut que l'usage soit cohérent.

En fonction de ces différents aspects, peux-tu effectuer les tâches suivantes?

1) mettre en gras les erreurs et les problèmes de style dans le texte;

2) dans un tableau, donner les explications quant aux erreurs et problèmes de style (répétitions, verbes ternes, etc.) diagnostiqués et signifier le type d'erreur (grammaire, code orthographique, orthographe, etc., ainsi que la catégorie;

3) noter le travail en fonction de la pondération suivante : 0,2 point par erreur diagnostiquée jusqu'à concurrence de 2,4 points, soit 12 erreurs. Ne pas comptabiliser les problèmes de style. Le résultat ne pourra pas être inférieur à 2,6 points sur 5 points;

4) fournir la rétroaction globale proposée ci-après, que tu individualises en fonction des éléments suivants :

a) les points obtenus : de 2,6 à 2,8 points/5, remplace « déjà bien abouti » par « qui mérite encore du travail »; de 3 à 3,4 points/5, remplace « déjà bien abouti » par « assez bien abouti »; de 4 à 4,4/5 points, remplace « déjà bien abouti » par « déjà fort bien abouti »; de 4,6 à 4,8/5 points, remplace « déjà bien abouti » par « déjà très abouti »; pour 5/5 points, remplace « déjà bien abouti » par « superbe »;

b) en fonction de tes explications : précise, dans le commentaire, après « sont surtout de l'ordre de », 2 ordres d'erreurs majoritairement diagnostiquées (par exemple, les accents et les accords; ou, comme autre exemple, les pronoms relatifs et les conjugaisons; la ponctuation et les homonymes);

c) indique le nombre d'erreurs à la fin du commentaire ainsi que la note finale.



Il est toujours très intéressant pour moi de découvrir ce qui motive les étudiant.e.s à suivre ce cours. Vous proposez un texte déjà bien abouti du point de vue de la construction des phrases, des accords, etc. Si vous ne le faites pas déjà, je vous invite à utiliser le logiciel Antidote (il y a des réglages en fonction de la langue maternelle) pour la rédaction de texte, tant pour son diagnostic de correction et du style (par exemple, la répétition de formes de phrase comme les phrases impersonnelles) que pour celui de la révision. Évidemment, cet usage devra toujours être accompagné d'autres outils linguistiques reconnus, car il ne constitue pas la panacée. Il pourrait vous aider à voir les points de langue que vous devriez améliorer et vous permettre de travailler le style (tel que le logiciel conçoit le style). Il reste que ce logiciel n'est pas l'objet du cours. En revanche, je suis certaine que vous l'apprécierez pour tous vos travaux universitaires après l'avoir acquis.

Sachez qu'il existe également le cours FRA3005 Rédaction fondamentale et outils numériques qui traite entre autres des questions du style et de la cohérence textuelle.

À la lecture de ce petit texte de motivation, vos maillons faibles sont surtout de l'ordre. Je vous souhaite déjà que ce cours vous aide à atteindre vos objectifs et je vous félicite pour votre détermination. N'hésitez pas à me poser toutes les questions qui vous taraudent. Bon parcours !

Pour vous préparer aux examens dans ce cours, réalisez tous les exercices ainsi que ceux dans le cahier d'exercices. Visitez aussi le site de la Vitrine Linguistique ainsi que celui du CCDMD (faites les nombreux exercices) dont vous trouverez les liens ci-dessous :

<https://vitrinelinguistique.oqlf.gouv.qc.ca>

<https://ameliofrancais.ccdmd.qc.ca>



Abstract / Resumen / Resumo

Assessing a Graded Writing Assignment in University: Comparison Between Human Evaluation and Evaluation by a Generative Artificial Intelligence Model

ABSTRACT

Generative artificial intelligence (GAI) can support the evaluative judgment exercised by university faculty when assessing students' written productions. However, its use raises issues that go beyond score concordance alone. To better understand the conditions under which GAI can provide such support, this article presents an exploratory experiment carried out in a university course focused on the mastery of written French. The study compares human evaluation and GAI-supported evaluation of the same diagnostically oriented written task. The analysis is grounded in a three-axis framework of evaluative judgment, encompassing the functions of assessment, its stages, and its ethical dimension, understood as inseparable from the evaluator's responsibility. The results show an overall convergence between the two modes of evaluation in terms of scores and number of identified errors, while revealing significant discrepancies in error identification, categorization, and the sequencing of diagnosis, scoring, and feedback as error complexity increases. The study shows that GAI cannot replace evaluative judgment. Nevertheless, it can support certain evaluative operations under clear methodological, institutional, and ethical guidelines, combined with faculty training and ongoing professional vigilance. It also opens avenues for further research, notably through comparison between different generative models.

Keywords: generative artificial intelligence (GAI), evaluative judgment, university writing (French), human–GAI comparison

Evaluar un trabajo escrito calificado en la universidad: comparación entre la evaluación humana y la evaluación por un modelo de inteligencia artificial generativa

RESUMEN

La inteligencia artificial generativa (IAG) puede apoyar el juicio evaluativo formulado por el profesorado sobre las producciones de los estudiantes. Sin embargo, su uso plantea cuestiones que van más allá, en particular, de la concordancia de las calificaciones. Para comprender mejor las condiciones de este apoyo por parte de la IAG, este artículo presenta una experimentación exploratoria realizada en un curso universitario de francés escrito. El estudio compara la evaluación humana y la evaluación por la IAG de una misma producción escrita con función diagnóstica. El análisis se basa en un marco de referencia del juicio evaluativo articulado en tres ejes: las funciones de la evaluación, sus etapas y su dimensión ética, inseparable de la responsabilidad de la persona evaluadora. Los resultados muestran una convergencia global entre ambos modos de evaluación en cuanto a las calificaciones y al número de errores señalados, pero revelan diferencias significativas en la identificación, la categorización de los errores y en la articulación entre diagnóstico, calificación y retroalimentación cuando aumenta la complejidad de los errores. Este estudio muestra que la IAG no puede sustituir al juicio evaluativo. No obstante, constituye una ayuda para



determinadas operaciones, siempre que se apoye en balizas metodológicas, institucionales y éticas claras, combinadas con la formación del profesorado y una vigilancia profesional constante. Asimismo, abre perspectivas de investigación, entre otras, hacia la comparación de diferentes modelos generativos.

Palabras clave: inteligencia artificial generativa (IAG), juicio evaluativo, escritura universitaria (en francés), comparación humano–IAG

Avaliar Examinar um trabalho escrito avaliado na universidade: comparação entre a avaliação humana e a avaliação por um modelo de inteligência artificial generativa

RESUMO

A inteligência artificial generativa (IAG) pode apoiar o juízo avaliativo exercido pelo corpo docente sobre produções estudantis. No entanto, a sua utilização levanta questões que vão além, nomeadamente, da concordância das classificações. Para compreender melhor as condições de tal apoio por parte da IAG, este artigo apresenta uma experimentação exploratória realizada num curso universitário de mestrado em domínio do francês escrito. O estudo compara a avaliação humana e a avaliação pela IAG de uma mesma produção escrita com função diagnóstica. Um quadro de referência do juízo avaliativo estruturado em três eixos sustenta a análise: as funções da avaliação, as suas etapas e a sua dimensão ética, indissociável da responsabilidade da pessoa avaliadora. Os resultados mostram uma convergência global entre os dois modos de avaliação no que respeita às notas e ao número de erros identificados, mas revelam desvios significativos na identificação, na categorização dos erros e no encadeamento entre diagnóstico, classificação e retroação quando a complexidade dos erros aumenta. Este estudo demonstra que a IAG não pode substituir o juízo avaliativo. Contudo, constitui um apoio para certas operações do processo, desde que enquadrado por balizas metodológicas, institucionais e éticas claras, associadas a uma formação do corpo docente e a uma vigilância profissional constante. Abre igualmente perspectivas de investigação, entre outras, no sentido da comparação de diferentes modelos generativos.

Palavras-chave: inteligência artificial generativa (IAG), juízo avaliativo, escrita universitária (em francês), comparação humano–IAG