Évaluation des outils d'IA pour la production écrite au lycée marocain : Benchmark et validation

Souad MHADA[1], **Ahmed OUJAK**[2]

[1] *Doctorante, Laboratoire CELHN, FLSHO, Université Mohammed Premier, Oujda — mhadasouad1@gmail.com*
[2] *Professeur Habilité, ENCG Oujda, Université Mohammed Premier — a.oujak@ump.ac.ma*

RÉSUMÉ

La correction automatisée des productions écrites par l'intelligence artificielle (IA) progresse rapidement, mais la littérature souligne la nécessité d'en vérifier la fiabilité, la validité et l'équité avant tout déploiement. Dans le secondaire marocain, les enseignants s'appuient sur le référentiel CECRL, alors que les outils IA mobilisent des logiques hétérogènes. Nous avons conduit un benchmark rigoureux sur 60 copies de lycéens (tronc commun, 1re et 2e BAC), notées selon une grille CECRL (score global /20 + six critères 0–4), afin de comparer quatre outils — deux LLM multilingues (GPT-4o, Claude 3.7 Sonnet) et deux correcteurs francophones (Grammalecte, LanguageTool) — à une référence humaine. Le critère principal est la MAE (Mean Absolute Error) sur la note globale ; les critères secondaires mobilisent le kappa pondéré quadratique (QWK) et des intervalles de confiance à 95 % par bootstrap. Les résultats montrent que GPT-4o affiche la MAE la plus basse (3,77) mais présente un biais de sous-notation (−2,68), tandis que Grammalecte offre un centrage quasi nul (biais ≈ −0,02). Sur les critères, les LLM convergent davantage sur les dimensions formelles (grammaire/orthographe) que sur les dimensions discursives (cohésion, registre), validant ainsi l'hypothèse exploratoire. L'étude propose un protocole reproductible (prompts figés, schéma JSON, grille CECRL) et des recommandations d'usage pour une intégration responsable de l'IA dans l'évaluation.

Mots-clés : Évaluation automatisée ; Production écrite ; Lycée ; CECRL ; Intelligence artificielle ; Fiabilité ; Validité ; Benchmark.

Evaluation of AI tools for written production in Moroccan high schools: Benchmarking and validation

ABSTRACT

Automated correction of written work by artificial intelligence (AI) is progressing rapidly, but the literature emphasizes the need to verify its reliability, validity, and fairness before any deployment. In Moroccan secondary schools, teachers rely on the CEFR framework, while AI tools employ heterogeneous logics. We conducted a rigorous benchmark on 60 student papers (common core curriculum, 1st and 2nd year of high school), graded according to a CEFR rubric (overall score /20 + six criteria 0–4), in order to compare four tools—two multilingual LLMs (GPT-4o, Claude 3.7 Sonnet) and two French-language correctors (Grammalecte, LanguageTool)—with a human reference. The primary criterion is the MAE (Mean Absolute Error) on the overall score; the secondary criteria utilize the quadratic weighted kappa (QWK) and 95% confidence intervals using bootstrapping. The results show that GPT-4o displays the lowest MAE (3.77) but exhibits an underscoring bias (−2.68), while Grammalecte offers near-zero centering (bias ≈ −0.02). Regarding the criteria, the LLMs converge more on formal dimensions (grammar/spelling) than on discursive dimensions (cohesion, register), thus validating the exploratory hypothesis. The study proposes a reproducible protocol (fixed prompts, JSON schema, CEFR grid) and usage recommendations for the responsible integration of AI into assessment.

KEYWORDS: Automated assessment; Written production; High school; CEFR; Artificial intelligence; Reliability; Validity; Benchmark.

INTRODUCTION

Évaluer, c'est décider. Et décider, souvent dans l'urgence, devant des piles de copies qui ne cessent de s'accumuler. L'intelligence artificielle (IA) promet un soulagement réel : rapidité de traitement, constance apparente, traçabilité des décisions. Mais au Maroc, avant toute intégration, une question simple et exigeante s'impose : ces outils jugent-ils comme un évaluateur humain compétent, selon nos référentiels et dans nos classes ? Dit autrement : quel niveau de validité, de fiabilité et d'équité peuvent-ils garantir ici et maintenant, au lycée marocain ?

L'histoire de l'évaluation automatisée des écrits ne commence pas avec les grands modèles de langage (LLM). Les systèmes d'automated essay scoring (AES) existent depuis plus de vingt ans. Des outils comme e-rater® (ETS) ont posé des jalons méthodologiques solides : validité, fiabilité, débats contradictoires sur les limites. Cette tradition scientifique a appris que la performance globale peut masquer des failles sur des dimensions fines — contenu, cohérence, registre — et qu'aucune adoption sérieuse ne peut se passer de preuves empiriques. Depuis l'émergence des LLM, la promesse s'est élargie : ces modèles savent structurer une sortie, « parler CECRL », décliner un jugement par critères. Mais leurs résultats restent contexto-dépendants, et la communauté scientifique appelle explicitement à valider ces performances dans le contexte d'usage, sur des textes réels et avec des métriques appropriées.

C'est précisément cette perspective qui guide le présent travail : mettre à l'épreuve quatre outils d'IA en conditions réalistes de lycée marocain, avec une référence humaine adossée au CECRL et des métriques explicites. Le design retenu est comparatif intra-sujets — chaque copie passe dans tous les outils — et les indicateurs choisis sont interprétables par les équipes

pédagogiques : MAE pour le score global (/20), QWK pour les critères ordonnés (0–4), intervalles de confiance à 95 % par bootstrap pour qualifier l'incertitude. L'article se structure ainsi : après avoir situé le contexte de la recherche et posé la problématique, nous présentons la revue de littérature et le cadre conceptuel, puis le dispositif méthodologique, avant d'exposer les résultats et leur discussion, et de conclure par des recommandations pratiques.

1. CONTEXTE DE LA RECHERCHE

Le lycée marocain change. Lentement, mais il change. Les textes officiels le confirment : la Vision stratégique 2015–2030 et la Loi-cadre n° 51-17 fixent des objectifs nets — qualité, équité, transparence — et appellent des pratiques d'évaluation crédibles, comparables et compatibles avec la protection des données. Sur le terrain cependant, l'évaluation des apprentissages demeure un point de tension, surtout lorsqu'il s'agit de la production écrite : activité riche, exigeante, lente à corriger, et sensible aux variations inévitables d'un correcteur à l'autre.

Dans ce contexte, l'IA apparaît comme une promesse et un risque à la fois. Promesse de rapidité et de constance, promesse d'outils capables d'objectiver certaines dimensions de la langue écrite. Risque de désalignement avec les objectifs pédagogiques du référentiel CECRL, risque d'injustice si l'outil favorise, à performance égale, certains sous-groupes d'élèves. La position raisonnable tient en trois mots : évidence avant adoption. Tant qu'on n'a pas mesuré, on ne sait pas.

Le cadre de référence existe pourtant, et il est robuste : le Cadre européen commun de référence pour les langues (CECRL). Ce n'est pas une grille parmi d'autres ; c'est un langage commun, avec des descripteurs, des niveaux et une cohérence d'ensemble. La note globale doit signifier quelque chose — efficacité communicative — et les critères doivent mesurer ce qui est effectivement visé : adéquation à la tâche, organisation, cohésion, lexique, grammaire/orthographe, registre. Sans cet ancrage, l'outil — ou même l'évaluateur humain — court le risque d'évaluer « autre chose ». L'alignement n'est pas un détail : c'est la condition de sens de toute évaluation, automatisée ou non.

La Loi-cadre n° 51-17 et la Vision stratégique imposent par ailleurs des principes qui obligent : qualité, équité, transparence, protection des données, reddition de comptes. L'OCDE, dans son rapport dédié à l'évaluation de la performance des établissements scolaires au Maroc (2024), souligne que la réussite des réformes est étroitement liée à la qualité des instruments d'évaluation et à leur ancrage dans les réalités de terrain. Autrement dit : pas de boîte noire, pas d'adoption sans conditions d'usage claires, pas de décision sans preuves.

2. PROBLÉMATIQUE, QUESTIONS DE RECHERCHE ET HYPOTHÈSES

Corriger n'est pas compter. C'est juger. Une intention, une organisation, une cohésion — puis décider, avec des conséquences réelles pour l'élève. Les outils d'IA promettent vitesse, constance et traçabilité. Très bien. Mais la question brute, locale, urgente, reste entière : dans les lycées marocains, quel outil produit des décisions suffisamment proches d'un évaluateur humain compétent, et à quelles conditions d'usage ? La littérature sur l'AES l'enseigne depuis vingt ans : on ne parle de déploiement qu'après validation et examen de l'équité.

Deux questions de recherche structurent cette étude. Q1 : parmi les outils d'IA testés, lequel présente le meilleur accord avec la référence humaine pour la note globale (/20) et des accords satisfaisants par critère (0–4) en conditions réelles, sur des copies de lycéens marocains ?

Q2 : les deux LLM diffèrent-ils entre eux dans leur évaluation des critères spécifiques, et existe-t-il des variations systématiques entre dimensions formelles et discursives ?

Ces questions appellent deux hypothèses. L'hypothèse principale (H1) avance qu'au moins un outil atteint une MAE faible sur la note globale et un QWK ≥ 0,75 sur plusieurs critères, avec des IC 95 % compatibles avec un usage pilote en contexte de lycée. L'hypothèse exploratoire (H2) postule que l'accord IA–humain est plus élevé sur les dimensions formelles (grammaire/orthographe) que sur les dimensions discursives (cohésion, registre), tendance déjà documentée dans les comparaisons humain–automatique dans la littérature.

3. REVUE DE LITTÉRATURE

L'évaluation automatisée des écrits n'est pas née avec les LLM. Elle a une histoire, et cette histoire porte des enseignements durables. Les systèmes comme e-rater® (ETS) ont ouvert la voie : modéliser des indices textuels, prédire une note, comparer à l'humain. Très tôt, la communauté a joué le jeu du test contradictoire. Powers, Burstein, Chodorow, Fowles et Kukich (2001) ont délibérément tenté de piéger e-rater® pour éprouver sa validité, rappelant qu'aucun score automatique ne doit être utilisé seul dans un contexte à enjeux. Attali et Burstein (2006) ont documenté les performances de la version V.2 — forces, faiblesses, conditions d'usage — sur des essais de candidats anglophones aux examens standardisés GRE et TOEFL. Un champ scientifique s'est ainsi structuré autour d'un principe simple : évidence avant adoption. Il importe de souligner que ces travaux fondateurs portent sur l'anglais, ce qui rend nécessaire une validation distincte pour le français et le contexte CECRL.

Pendant longtemps, l'approche dominante fut celle des caractéristiques de surface — longueur, orthographe, densité lexicale. Ces traits ont l'avantage d'être stables et transférables, parfois multilingues. Zesch, Wojatzki et Scholten-Akoun (2015) ont popularisé l'idée des task-independent features : un socle de variables utile pour généraliser au-delà d'un sujet précis. Mais on a aussi appris que la précision brute cache des angles morts : le contenu, la pertinence argumentative, le registre peuvent être ignorés par un système qui « score bien » sur d'autres variables. D'où la montée en puissance de travaux de synthèse sur l'état de l'art et ses manques, notamment Ke et Ng (2019) pour une revue panoramique et dense de l'AES à l'heure des architectures profondes.

L'émergence des LLM a changé l'échelle du possible. Ces modèles savent « parler la grille », produire des sorties structurées par critères, et parfois imiter le jugement humain avec un accord élevé. Lee, Cai, Meng, Wang et Wu (2024) proposent un cadre de zero-shot prompting par traits — le framework MTS (Multi Trait Specialization) — qui décompose la compétence d'écriture en dimensions distinctes et génère des critères de scoring pour chacune. Validé sur des corpus anglophones (ASAP), ce cadre constitue une source d'inspiration méthodologique pertinente, même si ses traits ne se superposent pas directement aux critères CECRL de notre étude. Pack, Barrett et Escalante (2024) confirment pour leur part la nécessité de valider les performances des LLM en condition d'usage, avec des métriques lisibles et des analyses de fiabilité. Autrement dit : l'IA peut bien scorer, parfois. Il faut donc mesurer, non supposer.

Au-delà de la précision moyenne, la question de l'équité s'impose. Une note exacte mais injuste ne vaut pas mieux qu'une note aléatoire. Litman, Zhang, Correnti, Matsumura et Wang (2021) évaluent la fairness de méthodes d'AES dans le domaine spécifique du scoring de l'utilisation des preuves textuelles dans l'écriture scolaire, et montrent que les variations systématiques entre sous-groupes dépendent du modèle utilisé. Ces résultats, bien que portant sur une tâche ciblée, illustrent un risque générable à l'ensemble des dispositifs d'évaluation automatisée. Schaller, Ding, Horbach, Meyer et Jansen (2024), quant à eux, comparent plusieurs algorithmes

d'AES sur des essais d'apprenants germanophones au niveau secondaire et appellent à des protocoles transparents — données, métriques, scripts publiés — pour auditer la justice du scoring dans des contextes scolaires réels. Le message est clair : optimiser l'exactitude moyenne ne suffit pas ; il faut regarder pour qui ça marche, où ça casse, et comment on le sait.

Sur le plan des métriques, trois familles dominent la littérature. La MAE (Mean Absolute Error) est privilégiée pour l'écart moyen absolu sur une note continue comme une note sur 20 : elle est lisible et interprétable par les équipes pédagogiques. Le QWK (kappa pondéré quadratique) sert l'accord ordinal par critères, car il pénalise davantage les gros désaccords — ce que le terrain juge pertinent. L'ICC, enfin, qualifie la fiabilité absolue quand les notes sont continues et qu'on veut dépasser le simple accord relatif (Koo et Li, 2016 ; McHugh, 2012). L'ensemble de ces indicateurs gagne en crédibilité lorsqu'il est accompagné d'intervalles de confiance à 95 % par bootstrap, pratique robuste popularisée depuis Efron (1979). En contexte marocain francophone, la rareté d'études empiriques sur l'AES adossé au CECRL constitue une lacune que le présent benchmark vise précisément à combler.

4. CADRE CONCEPTUEL ET THÉORIQUE

Évaluer n'est pas produire un chiffre : c'est interpréter des traces d'apprentissage avec un cadre de référence explicite, puis justifier l'usage de ces scores. Quatre piliers conceptuels structurent ce travail, en conformité avec les Standards for Educational and Psychological Testing (AERA/APA/NCME, 2014) : la validité de construit et d'interprétation, la fiabilité et l'accord inter-rateur, l'équité, et l'utilité opérationnelle. On n'emploie un score à des fins décisionnelles qu'après avoir établi des preuves pour ces quatre dimensions — et en étant transparent quant aux inférences autorisées.

Le CECRL — Companion Volume (2020) — sert de langage commun tout au long de ce travail. Il relie des descripteurs (adéquation à la tâche, organisation, cohésion, lexique, grammaire/orthographe, registre) et des niveaux d'exigence à travers une cohérence d'ensemble. Notre grille de référence humaine reprend cette structuration — score global /20 + six critères 0–4 — et nous exigeons que les sorties des outils soient ancrées dans ces mêmes descripteurs. Sans cet alignement, un outil peut « bien scorer » sur autre chose que ce qui est visé pédagogiquement ; avec lui, la validité de construit devient testable.

La notion de validité est ici comprise comme un argument appuyé par des preuves, non comme une propriété intrinsèque de l'outil. Les systèmes historiques d'AES ont montré qu'on peut prédire une note humaine à partir de traits textuels, mais ils ont aussi été éprouvés par des contre-tests qui ont rappelé les limites et les conditions d'un usage prudent. Notre protocole reprend cette logique contradictoire : démontrer où et jusqu'où l'accord tient, et où il se fissure. Pour la fiabilité, nous privilégions la MAE sur le score global — lisible pour les enseignants — et le QWK pour les critères ordonnés, complétés par des IC 95 % par bootstrap pour chiffrer l'incertitude de façon non paramétrique.

L'équité n'est pas un supplément d'âme : une mesure précise en moyenne peut rester injuste si elle varie systématiquement selon des sous-groupes. Nous adoptons une lecture pragmatique et transparente : documenter les biais (ΔMAE, biais moyen) par niveau et par tâche lorsque l'information est disponible, et rendre publics les prompts, les versions de modèles et les scripts d'analyse pour permettre l'auditabilité. Enfin, l'utilité opérationnelle entre dans la décision d'adoption au même titre que l'accord statistique : un outil précis mais lent, coûteux, instable ou risqué sur le plan de la confidentialité n'aidera personne sur le terrain.

5. CADRE MÉTHODOLOGIQUE

5.1 Design de l'étude

Le devis retenu est quantitatif comparatif intra-sujets : chaque production écrite est évaluée par tous les outils et par la référence humaine, ce qui permet des comparaisons directes sans biais d'échantillonnage entre groupes. L'étude s'inscrit dans un paradigme post-positiviste à orientation pragmatique d'aide à la décision. Elle est centrée sur la notation seulement — sans rétroaction aux élèves — dans des conditions réalistes de lycée marocain.

5.2 Corpus et échantillonnage

Le corpus comprend 60 productions écrites authentiques, réparties équitablement sur trois niveaux : tronc commun, 1re BAC et 2e BAC (environ 20 copies par niveau). Deux genres textuels sont représentés — l'argumentation et le compte rendu — avec des textes d'une longueur approximative de 120 mots (tolérance ±30 %). Les sujets proposés correspondent aux pratiques effectives du lycée marocain : pour le tronc commun, « L'argent peut-il réaliser le bonheur ? » ; pour la 1re BAC, une production argumentative sur la famille ; pour la 2e BAC, un texte sur la valeur du travail à partir d'une citation de Voltaire. Les copies ont été collectées en classe dans des conditions standardisées, puis numérisées, anonymisées par identifiant alphanumérique (ex. TC-ARG-001), et normalisées en texte brut.

5.3 Référence humaine

La référence humaine repose sur une grille CECRL adaptée FLE comportant un score global /20 (à une décimale) et six critères notés de 0 à 4 : adéquation à la tâche, organisation, cohésion, lexique, grammaire/orthographe, registre. La notation a été effectuée en aveugle des sorties IA. La fiabilité intra-rateur est documentée par une re-notation à l'aveugle de 20 % des copies après 10 à 14 jours, permettant de calculer l'ICC(1,1) et la MAE pour le global, et le QWK par critère. Ce dispositif constitue une référence humaine contrôlée, légère mais documentée, suffisante pour un benchmark initial.

5.4 Outils évalués : quatre bras complémentaires

Quatre outils ont été évalués, couvrant deux familles. Les deux LLM multilingues — GPT-4o (OpenAI) et Claude 3.7 Sonnet (Anthropic) — ont été utilisés pour la notation directe : sortie JSON structurée, température = 0, global /20 et six critères 0–4. GPT-4o a été retenu pour sa robustesse multilingue et ses sorties structurées fiables ; Claude 3.7 Sonnet a été choisi comme contrepoint pour une comparaison intra-sujets propre. Les deux correcteurs francophones — Grammalecte et LanguageTool (self-host) — fournissent des indicateurs objectivables : densité d'erreurs (erreurs/100 mots), alertes typographiques, éléments de cohésion, mappés sur les critères CECRL selon une règle de pondération fixe et documentée. Versions et paramètres ont été figés dès le pilote, et les sorties ont été horodatées.

5.5 Procédure et indicateurs

La procédure a comporté un pilote sur 10 à 12 copies pour geler les prompts, vérifier la conformité JSON et tester le pipeline d'analyse. Le passage complet a ensuite été réalisé sur toutes les copies, pour chaque outil, en deux runs successifs afin d'évaluer la stabilité. La notation humaine a été effectuée en aveugle des sorties IA. L'indicateur principal est la MAE (moyenne et médiane) IA–humain avec IC 95 % par bootstrap (B = 2 000 à 5 000 rééchantillonnages). Les indicateurs secondaires sont le QWK par critère (0–4) avec IC 95 % bootstrap et, optionnellement, l'ICC(1,1) pour la fiabilité absolue. La comparaison multi-outils repose sur le test de Friedman avec post-hoc Nemenyi ou Wilcoxon apparié corrigé Holm-Bonferroni, conformément aux recommandations de Demšar (2006) pour la comparaison de systèmes sur les mêmes items.

6. RÉSULTATS

6.1 H1 — Accord sur la note globale (/20)

L'analyse de l'erreur absolue moyenne en mode non calibré, sur les 60 copies, donne le tableau suivant pour les quatre outils. GPT-4o présente la MAE la plus basse (3,77 ; IC95 % [3,36 ; 4,20]) avec 17 % des copies à moins d'un point d'écart, mais affiche un biais négatif notable (−2,68), traduisant une tendance systématique à sous-noter. Grammalecte obtient une MAE légèrement supérieure (4,27 ; IC95 % [3,85 ; 4,70]) mais un centrage quasi parfait (biais ≈ −0,02), ses erreurs étant plus dispersées copie par copie. Claude présente une MAE de 4,86 (IC95 % [4,45 ; 5,26]) combinée à une sous-notation accentuée (biais = −4,64). LanguageTool affiche la MAE la plus élevée (5,99 ; IC95 % [5,59 ; 6,40]) avec un biais de sur-notation très fort (+5,99), ce qui le disqualifie en l'état pour une utilisation en note /20.

Outil	MAE	IC95 %	% ≤ 1 pt	Biais moyen	Verdict
GPT-4o (brut)	3,77	[3,36 ; 4,20]	17 %	−2,68	Meilleure MAE
Grammalecte (brut)	4,27	[3,85 ; 4,70]	14 %	≈ −0,02	Mieux centré
Claude (brut)	4,86	[4,45 ; 5,26]	14 %	−4,64	Sous-notation
LanguageTool (brut)	5,99	[5,59 ; 6,40]	4 %	+5,99	Inadapté

Tableau 1. Indicateurs H1 — Accord note globale /20 (mode non calibré, 60 copies).

Avec un seuil strict de MAE ≤ 2,0, aucun outil brut ne satisfait pleinement H1. Toutefois, si l'on adopte un seuil pragmatique de 2,5 après calibration simple, certains outils approchent la zone d'utilité pilote. GPT-4o est retenu comme outil de référence pour les analyses H2, en tant que meilleur MAE non calibré, avec la réserve explicite sur son biais de sous-notation.

6.2 H2 — Variations par critères et comparaison inter-LLM

À partir des sorties JSON des deux LLM, l'accord inter-LLM mesuré par QWK révèle un profil hiérarchique clair. L'accord le plus élevé apparaît sur le critère grammaire/orthographe (aspect formel), tandis que les accords les plus faibles sont observés sur cohésion et registre (aspects discursifs). Les deux LLM convergent donc davantage sur le formel et divergent sur le discursif, conformément à l'hypothèse H2.

Les corrélations de Spearman entre chaque critère LLM et la note humaine globale confirment ce profil pour GPT-4o comme pour Claude : les corrélations les plus fortes concernent grammaire/orthographe et souvent lexique, tandis que celles de cohésion et registre sont plus faibles. Les contrastes bootstrap (ρ(gram_orth) − ρ(cohésion) et ρ(gram_orth) − ρ(registre)) sont positifs et souvent significatifs (IC 95 % ne recouvrant pas 0), signalant que les aspects formels s'alignent significativement mieux sur la note humaine globale que les aspects discursifs.

Le contrôle de la longueur des textes (nb_mots) — corrélée positivement à la note humaine, effet classique — ne remet pas en cause ce résultat. Après contrôle de nb_mots, les corrélations partielles des critères formels restent les plus élevées, et les R^2 hiérarchiques montrent que les critères ajoutent une variance expliquée substantielle au-delà de la longueur seule ($\Delta R^2 > 0$). La qualité formelle ne se réduit donc pas à « plus de mots » : elle explique une part indépendante de la note humaine. H2 est globalement soutenue par les données.

DISCUSSION

Les résultats en mode non calibré dessinent un paysage contrasté. GPT-4o obtient la MAE la plus faible (3,77) mais présente un biais négatif marqué (−2,68), traduisant une sous-notation systématique. Grammalecte affiche une MAE légèrement plus élevée (4,27) avec un centrage quasi parfait (biais ≈ −0,02), ce qui en fait un repère « neutre » utile pour le terrain. Claude combine une MAE supérieure (4,86) à une sous-notation accentuée (−4,64), tandis que LanguageTool sur-note franchement (+5,99) avec l'écart moyen le plus élevé, ce qui le disqualifie en l'état pour une note certificative.

Ce compromis précision-centrage est déterminant pour l'usage. Dans un contexte formatif, l'enseignant peut préférer GPT-4o pour ses erreurs absolues plus faibles, à condition d'anticiper la sous-notation et d'accompagner la pré-notation d'une relecture humaine. Dans un contexte certificatif, le biais devient critique : Grammalecte, bien que moins précis copie par copie, colle mieux au niveau moyen humain et réduit le risque d'injustice systématique. En pratique, ces résultats invitent à ne jamais publier la note brute telle quelle, mais à utiliser l'IA comme aide — pré-notation argumentée — avec arbitrage humain sur les cas limites.

Les analyses par critères (H2) éclairent l'origine de ces comportements. L'accord inter-LLM plus élevé sur grammaire/orthographe que sur cohésion et registre révèle que les modèles convergent davantage sur l'aspect formel que sur les dimensions discursives. Cette hiérarchie se retrouve côté validité : les corrélations entre critères et note humaine globale sont plus fortes pour gram/orth que pour cohésion/registre. L'avantage du formel subsiste même après contrôle de la longueur des textes, signalant une sensibilité réelle des LLM aux indices de correction linguistique — indices qui pèsent aussi dans la notation humaine.

Ces constats permettent de comprendre la différence GPT-4o/Grammalecte. GPT-4o, plus précis globalement, pénalise davantage — sans doute parce qu'il pondère fortement des indices formels détectés avec précision et les agrège de manière stricte au score global. Grammalecte mesure des erreurs linguistiques mais ne les traduit pas en note /20 avec la même agressivité, d'où un centrage proche de l'humain au prix d'une dispersion plus grande. Cette tension entre sensibilité et équité est au cœur des choix d'intégration pratique.

Les limites de l'étude doivent être explicitées avec franchise. L'absence de notation humaine par critère ne permet pas d'estimer le QWK humain–IA par dimension, ce qui aurait affiné l'analyse de la validité. L'échantillon (n = 60) reste modeste, même si sa répartition sur trois niveaux et deux genres textuels constitue un avantage. La référence humaine repose sur un seul évaluateur, bien que la procédure intra-rateur à 20 % apporte un premier garde-fou. Le bruit potentiel de transcription/OCR et les variations liées aux versions de modèles s'ajoutent à ces contraintes. Enfin, l'ancrage marocain — grilles, pratiques d'examen — restreint la généralisation immédiate à d'autres contextes.

CONCLUSION

Cette étude avait un objectif simple et exigeant : mesurer à quel point des outils d'IA peuvent approcher une notation humaine de productions écrites au lycée, dans des conditions réalistes et avec un protocole reproductible. Deux résultats saillants se dégagent.

Sur la note globale (/20), en mode non calibré, GPT-4o présente la MAE la plus faible, au prix d'un biais négatif. Grammalecte offre un centrage quasi nul mais une MAE légèrement plus élevée. Claude sous-note davantage, et LanguageTool sur-note franchement : ces deux derniers, en l'état brut, sont moins adaptés à une note certificative. Le choix opérationnel dépend donc du risque acceptable et du cadre d'usage : précision versus équilibre autour de la référence humaine.

Sur les critères (0–4), les analyses confirment H2 : les LLM convergent surtout sur les dimensions formelles et divergent sur les dimensions discursives. Cet avantage du formel domine encore après contrôle de la longueur, signalant une contribution explicative au-delà de l'effet « plus de mots ». En clair, les indices formels constituent le point d'appui le plus fiable pour l'évaluation automatique ; les indices discursifs restent plus fragiles et contextuels.

Sur le plan pratique, trois recommandations s'imposent. Premièrement, ne pas publier la note brute telle quelle : utiliser l'IA en pré-notation argumentée avec relecture humaine, en particulier sur les cas limites. Deuxièmement, choisir l'outil selon l'objectif : GPT-4o en formatif pour ses repères plus fins, Grammalecte en certificatif pour minimiser les biais. Troisièmement, documenter la transparence — protocole, prompts, formats, métriques, biais surveillés — et garantir les exigences de confidentialité liées aux textes d'élèves.

Les perspectives sont claires : étendre la fiabilité inter-rateur, collecter des scores humains par critère, répliquer sur d'autres cohortes, tester des calibrations simples par niveau et genre textuel, explorer des approches hybrides combinant LLM et indicateurs formels objectivables, et conduire des audits d'équité selon les profils d'élèves et les filières. En somme, l'IA peut utilement épauler l'évaluation de l'écrit au lycée — si l'on accepte sa signature, si l'on maintient un contrôle humain, et si l'on s'astreint à une transparence méthodologique. C'est à ces conditions que l'IA devient un levier, et non un substitut, pour une évaluation plus réactive, plus explicite et plus juste.

RÉFÉRENCES

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. American Educational Research Association. https://www.testingstandards.net

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V.2. Journal of Technology, Learning, and Assessment, 4(3). https://files.eric.ed.gov/fulltext/EJ843852.pdf

Bennett, R. E. (2004). Moving the field forward: Some thoughts on validity and automated scoring (ETS Research Memorandum RM-04-01). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RM-04-01.pdf

Conseil de l'Europe. (2020). Common European Framework of Reference for Languages: Companion volume. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

CSEFRS. (2015). Pour une école de l'équité, de la qualité et de la promotion : Vision stratégique de la réforme 2015–2030. https://iro.umi.ac.ma/wp-content/uploads/2021/10/ODD-4-A4-Vision-stratégique-de-la-réforme-2015-2030.pdf

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30. https://jmlr.org/papers/volume7/demsar06a/demsar06a.pdf

Efron, B. (1979). Bootstrap methods: Another look at the jackknife. The Annals of Statistics, 7(1), 1–26. https://doi.org/10.1214/aos/1176344552

Ke, Z., & Ng, V. (2019). Automated essay scoring: A survey of the state of the art. In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI-19) (pp. 6300–6308). https://www.ijcai.org/proceedings/2019/0879.pdf

Koo, T. K., & 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://pmc.ncbi.nlm.nih.gov/articles/PMC4913118/

Lee, S., Cai, Y., Meng, D., Wang, Z., & Wu, Y. (2024). Unleashing large language models' proficiency in zero-shot essay scoring. Findings of the Association for Computational Linguistics: EMNLP 2024, 181–198. https://doi.org/10.18653/v1/2024.findings-emnlp.10

Litman, D., Zhang, H., Correnti, R., Matsumura, L. C., & Wang, E. (2021). A fairness evaluation of automated methods for scoring text evidence usage in writing. In I. Roll et al. (Eds.), Artificial Intelligence in Education (LNAI 12748, pp. 255–267). Springer. https://doi.org/10.1007/978-3-030-78292-4_21

McHugh, M. L. (2012). Interrater reliability: The kappa statistic. Biochemia Medica, 22(3), 276–282. https://www.biochemia-medica.com/en/journal/22/3/10.11613/BM.2012.031

OCDE. (2024). L'évaluation de la performance des établissements scolaires au Maroc. Organisation de Coopération et de Développement Économiques. https://www.oecd.org/content/dam/oecd/fr/publications/reports/2024/03/l-evaluation-de-la-performance-des-etablissements-scolaires-au-maroc_9031fa8c/4f59bfc1-fr.pdf

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. Computers & Education: Artificial Intelligence, 6, 100234. https://doi.org/10.1016/j.caeai.2024.100234

Powers, D. E., Burstein, J., Chodorow, M., Fowles, M. E., & Kukich, K. (2001). Stumping e-rater: Challenging the validity of automated essay scoring (ETS Research Report RR-01-03). Educational Testing Service. https://www.ets.org/Media/Research/pdf/RR-01-03-Powers.pdf

Royaume du Maroc. (2019). Loi-cadre n° 51-17 relative au système d'éducation, de formation et de recherche scientifique (9 août 2019). https://planipolis.iiep.unesco.org/sites/default/files/ressources/morocco_la-loi-cadre-17-51-fr.pdf

Schaller, N.-J., Ding, Y., Horbach, A., Meyer, J., & Jansen, T. (2024). Fairness in automated essay scoring: A comparative analysis of algorithms on German learner essays from secondary education. In Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA) (pp. 210–221). https://aclanthology.org/2024.bea-1.18.pdf

Zesch, T., Wojatzki, M., & Scholten-Akoun, D. (2015). Task-independent features for automated essay grading. In Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications (pp. 224–232). https://aclanthology.org/W15-0626/