

<https://lettres-lca.enseigne.ac-lyon.fr/spip/spip.php?article482>



Expérimentation de la correction de copies du DNB par l'intelligence artificielle

- Examens - Évaluations -

Date de mise en ligne : mercredi 20 novembre 2024

Copyright © Lettres & Langues et Cultures de l'Antiquité - Tous droits

réservés

Présentation et objectifs de l'expérimentation

La correction de copies occupe une place importante de l'emploi du temps des professeurs ([source du Ministère](#)) : la moyenne serait de [6h30 par semaine pour les matières littéraires](#). S'il existe des outils pour alléger cela (la correction automatique de QCM en ligne de [La Quizinière](#) par exemple ou encore via un smartphone qui identifie, le temps d'une photographie, les bonnes réponses d'un élève [sur papier](#) ou via [une ardoise numérique](#)), l'utilisation de l'intelligence artificielle semble encore plus prometteuse (l'entreprise Compilatio, basée à Annecy, expérimente actuellement avec des enseignants volontaires [un logiciel](#) qui va en ce sens) car elle pourra notamment traiter de plus grandes quantités de textes, comme pour une rédaction, par exemple. En outre, l'IA ne se fatigue pas, peut répéter à l'envi une même demande, reste objective et peut surtout traiter en un temps record une grande quantité de texte.

Mais doit-on pour autant confier à une machine la correction de la production d'élèves ?

L'objectif de cette expérimentation a été de partir de productions réelles de collégiens, sur des copies officielles récupérées lors de l'épreuve de DNB de Français session 2024 et de comparer la correction humaine par rapport à la correction menée par la machine. Qui est le plus rapide ? Qui est le plus objectif ? Qui est le plus efficace ? Et pour terminer, la technologie est-elle suffisamment élaborée pour s'y fier ? En somme qui remporte le "combat" entre l'homme et la machine ?

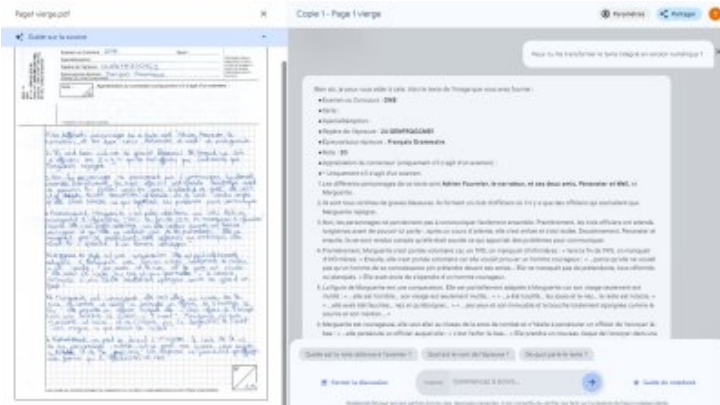
Pour me faire un avis le plus objectif possible, j'ai récupéré sept copies officielles, avant correction [1]. Il s'agissait de sept copies présentes dans des enveloppes de sept correcteurs différents qui ont bien voulu tenter l'expérience. Je les ai scannées, page par page, en haute résolution (600 dpi) avec le photocopieur du centre d'examen et ai obtenu ainsi des fichiers PDF (un fichier par page).

J'ai ensuite rendu les copies aux enseignants, puis une fois les copies corrigées, je les ai à nouveau scannées avec l'apparition des points attribués, les quelques annotations et autres commentaires pour permettre la comparaison ultérieurement.

Une reconnaissance de l'écriture manuscrite bien meilleure mais largement perfectible

Le premier écueil est que les copies d'examen, comme la plupart des productions des élèves, sont manuscrites. Pour ceux qui ont connu les logiciels OCR [2] d'antan, la tâche à relever paraissait immense : transformer une écriture d'élève, pas toujours lisible, en texte numérique [3]. Si le correcteur doit en effet tout retaper, le gain de temps est évidemment nul. Mais là encore, l'intelligence artificielle a permis des progrès fantastiques. J'ai donc utilisé [NotebookLM](#), une des interfaces de l'IA Gemini [4], en la détournant de son usage premier qui est de traiter des informations contenues dans des documents, de les synthétiser. On peut importer différents types de documents et je lui ai transmis ainsi page par page une première copie non corrigée.

Voici l'instruction générative (ou prompt) que j'ai saisie après le transfert du document : "Peux-tu me transformer le texte intégral en version numérique ?". Dans l'interface apparaît alors le résultat.



Transformation du texte manuscrit en texte numérique

Force est de constater que le rendu est impressionnant au premier abord. Cependant, à bien y regarder, on peut relever de nombreuses transformations, vraisemblablement lorsque l'IA cherche à trouver de la cohérence là où il n'y en a pas ou tout simplement lorsque la reconnaissance des caractères manuscrits a été inefficace : la renumérotation des questions (en omettant les a. ou b.), les absences d'accent ou erreurs d'homophones grammaticaux sont corrigées par exemple (cela serait particulièrement gênant pour la dictée : il faut demander alors une restitution fidèle du texte de l'élève en conservant les erreurs) mais parfois cela va plus loin dans des paragraphes qui ont mal été "lus" par l'IA. Ainsi, dans l'exemple ci-dessous le « Premièrement » de l'élève pour introduire sa première justification attendait vraisemblablement, pour l'IA, un « Deuxièmement » qui n'apparaît pas. L'élève a en effet utilisé « ensuite » pour introduire le deuxième argument, ce qui est tout à fait acceptable, ou du moins largement compréhensible par un correcteur humain. La transformation par l'IA n'est par conséquent pas plus efficace, voire souvent bien pire que l'original : cela peut conduire à des biais importants pour la correction notamment lorsque des phrases entières sont réécrites comme c'est encore le cas ici. À sa décharge, l'IA rencontre parfois les mêmes difficultés que l'œil humain lorsque l'écriture manuscrite est peu lisible et que les lignes ne sont pas sautées (ce qui devrait être obligatoire sur les copies officielles d'examen). Il arrive enfin que le texte "lu" n'ait absolument aucun rapport avec ce qui est écrit : l'IA a horreur du vide et à partir de quelques bribes de phrases est capable d'inventer n'importe quoi (c'est une forme d'hallucination [5]).

Un paragraphe de la copie de l'élève :



Extrait de copie manuscrite

Le paragraphe transformé en numérique par l'IA :

« 3. Non, les personnages ne parviennent pas à communiquer facilement ensemble. Premièrement, les trois officiers ont attendu longtemps avant de pouvoir lui parler ; après **un cours d'attente oublié, elle s'est enfuie et s'est isolée. Deuxièmement, Penanster** et ensuite, ils se sont rendus compte qu'elle était sourde ce qui apportait des problèmes pour communiquer. »

J'ai donc dû vérifier et remettre à l'identique, pour chaque copie, ce que l'élève avait écrit afin de viser la plus grande objectivité de correction par l'IA ensuite. Mais cela est chronophage, sans compter le temps de numérisation des copies et l'insertion ensuite dans NotebookLM.

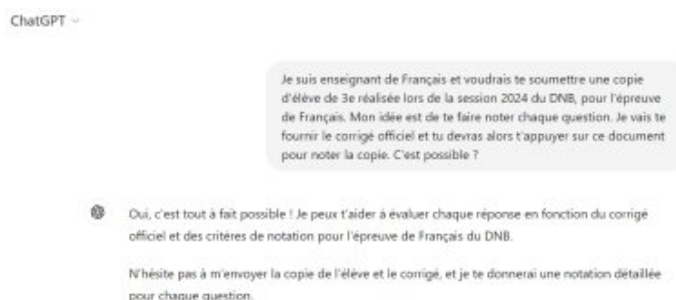
Pour conclure cette partie, l'œil humain s'avère donc encore indispensable tant les transformations opérées par l'IA sont nombreuses par sa volonté d'améliorer un texte ou dans sa difficulté de reconnaissance de l'écriture manuscrite. En outre, le gain de temps est très discutable voire défavorable à l'IA car il faut tout revérifier avant de passer à l'étape suivante. On peut malgré tout souligner les réels progrès en terme d'HTR [6] et avec des outils

encore plus perfectionnés, encore inaccessibles au grand public mais qui existent, on peut espérer que dans un avenir prochain cela ne posera plus de difficultés.

La correction par l'IA

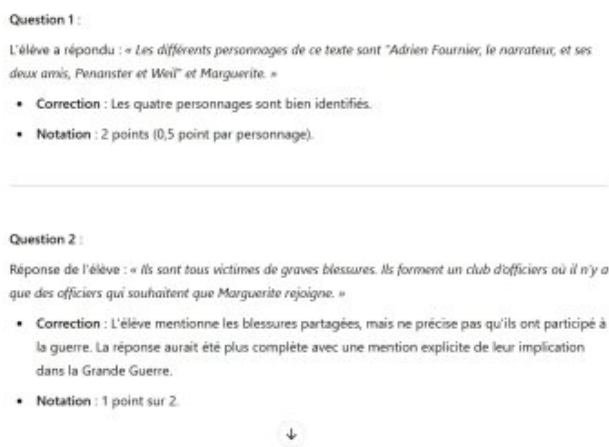
Après une relecture complète et le regroupement en un seul document .PDF de l'intégralité de la première copie, j'ai utilisé ChatGPT-4o (version gratuite) pour la partie correction.

Voici ce que j'ai alors demandé :



Instruction générative de départ donnée à ChatGPT

J'ai donc chargé les deux documents et en quelques secondes, l'IA a fourni les premiers éléments de manière détaillée :



Extrait des réponses générées par l'IA

C'est alors que j'ai débuté la confrontation entre la correction humaine et celle menée par la machine.

a. Compréhension et compétences d'interprétation et Grammaire et compétences linguistiques

Premier bilan en nombre de points : une différence de seulement 0,5 point apparaît entre la notation par la machine et par l'homme sur les 11 premières questions de la première copie (Compréhension et compétences d'interprétation et Grammaire et compétences linguistiques). Voici le détail en cas de désaccord :

Comparaison entre IA et Humain

Expérimentation de la correction de copies du DNB par l'intelligence artificielle

Rappel de la question	Points attribués par l'IA	Points attribués par le correcteur humain	Commentaires
Question 2 : « Qu'ont-ils en commun ? Deux éléments de réponse sont attendus. (2 points) »	1 / 2	2 / 2	Le correcteur a été un peu généreux au regard des attendus du corrigé
Question 5b : « Complétez ce portrait physique de Marguerite par son portrait moral en identifiant deux traits de caractère du personnage. Vous justifierez chaque trait de caractère en vous appuyant sur le texte. (4 points) »	2 / 4	4 / 4	L'IA s'est appuyée sur le corrigé national qui n'évoquait pas un trait de caractère pourtant bien présent dans le portrait de Marguerite. Le correcteur a par conséquent attribué les points en toute logique. L'IA n'est pas responsable puisqu'elle ne possédait même pas le texte source.
Question 7 : « Image. Cette affiche vous paraît-elle être une bonne illustration du texte ? Vous développerez votre réponse en vous appuyant sur deux arguments. Chaque argument doit être justifié en vous référant au texte et à l'image. (6 points) »	5 / 6	4 / 6	La question sur l'image prête plus à la subjectivité et l'écart de point est minime. Les 5 erreurs orthographiques ont pu influencer le correcteur.
Question 8 : « « Nous formons, lui expliqua-t-il, un club d'officiers qui compte à ce jour trois membres actifs et volontiers bienfaiteurs. » (lignes 4 à 5). Relevez les expansions du nom « club » et indiquez la classe grammaticale de chacune d'elle. (2 points) »	2 / 2	1 / 2	Le correcteur a sanctionné la confusion entre "proposition" et "préposition" alors que l'IA a corrigé d'elle-même le texte (encore une fois), rendant la réponse de l'élève correcte.
Question 9b : « Précisez la fonction grammaticale de cette proposition subordonnée et mentionnez au moins une manipulation que vous avez utilisée pour trouver la réponse. (2 points) »	1 / 2	/ 2	Le correcteur a oublié de noter la question et n'a donc attribué aucun point alors que la question était traitée par l'élève et que le point attribué par l'IA était justifié.
Question 10b : « Expliquez le sens de ce mot [NDLR : "Insupportable"] puis trouvez-en un synonyme. (1,5 point) »	1 / 1.5	0.5 / 1.5	L'IA a accepté en partie le mot "incontenable" comme synonyme d'"insupportable" dans une certaine forme de générosité, admettant malgré tout que "Un synonyme plus précis comme « insoutenable » ou « intolérable » aurait été mieux."

Conclusion sur l'ensemble des questions : si l'écart de points est minime, il enfonce malgré tout quelques portes ouvertes. En effet, l'IA fait fi des erreurs orthographiques de l'élève ce qui a pour conséquence une rehausse des points parfois. Par ailleurs, le correcteur humain n'est pas infallible et peut omettre involontairement certaines erreurs dans la dictée ou bien oublier de corriger ou noter certaines questions dans la première partie d'épreuve (c'est malgré tout très marginal sur l'ensemble du corpus) ou encore être irrité par la forme ce qui a pour conséquence une diminution de la valeur du fond. Enfin, l'IA se cantonne aux instructions fournies par le corrigé et n'a pas la subjectivité ou plutôt la sensibilité d'un correcteur, disposé à valoriser une tournure ou le choix d'un mot plutôt qu'un autre. La réelle plus-value de l'IA pour cette partie est l'extrême rapidité de correction (quelques secondes) mais c'est en oubliant tout le temps préparatoire à cette possibilité de correction numérique.

b. La dictée

Pour ce qui est de la dictée, la première copie a souffert d'une appréciation par l'IA plus sévère que celle qui est appliquée lors des corrections humaines.

En effet, si toutes les erreurs ont été repérées, certaines sont mal catégorisées et d'autres pénalisées deux fois. Cela entraîne un écart conséquent de 3 points entre l'IA et le correcteur humain. Pour les autres copies, j'ai dû modifier mon instruction pour obtenir une évaluation adéquate.

Analyse des erreurs et notation

1. Erreurs grammaticales (1 point par erreur) :

- devint → Erreur de conjugaison : « devint » au lieu de « devint » (1 point).
- répétait → Erreur de conjugaison : « répétait » avec un accent aigu (1 point).
- atteint → Accord incorrect : « atteint » doit être au pluriel « atteints » (1 point).
- grognement → Accord incorrect : « grognements » au pluriel (1 point).
- inform**er → Erreur de conjugaison : « informée » au lieu de « informé » (1 point).
- leurs → Erreur de pronom : « leur » sans « s » (1 point).
- dépêcher → Erreur de conjugaison : « dépêché » au lieu de « dépêcher » (1 point).
- ou devant → Erreur de syntaxe : « au-devant » doit être relié par un trait d'union (1 point).

Total des erreurs grammaticales : 8 points.

2. Erreurs lexicales (0,5 point par erreur) :

- grognement → faute répétée (non comptabilisée deux fois).
- leurs → faute répétée (non comptabilisée deux fois).

Total des erreurs lexicales : 0,5 point, car les erreurs lexicales ne concernent pas les mots à corriger, mais plutôt le lexique de deux mots déjà comptabilisés dans les erreurs grammaticales.

3. Erreurs de ponctuation, majuscule, trait d'union ou accent (0,5 point pour quatre erreurs) :

- Les accents sur « répétait » et « devint » sont des erreurs de ponctuation/orthographe.

Total des erreurs de ponctuation : 0,5 point.

Note de la dictée

En appliquant les pénalités, la note de la dictée est de :

$10 - (8 + 0,5 + 0,5) = 1$ point.

L'élève obtient 1 point sur 10 pour la dictée. ↓

Explication du barème de la dictée

c. La rédaction

L'IA reprend les propositions du barème national (invention ; organisation et situation d'énonciation ; expression), détaille et justifie ensuite sa notation. C'est un atout par rapport au correcteur humain qui laisse très peu de commentaires sur la copie et encore moins le barème retenu : seule la note finale apparaît. En revanche, nombreux sont les correcteurs à marquer leurs annotations sur un brouillon à côté pour chaque rédaction voire même remplissent un tableau fourni par les harmonisateurs, mais je ne possède pas ces documents.

Appréciation du correcteur (uniquement s'il s'agit d'un examen) :

*C'est un récit bien mené avec des efforts
d'expression sensibles (vocabulaire, tournures...)
Quelques erreurs dans l'emploi des temps.*

38/40

Annotations générales laissées par le correcteur

Rédaction de l'élève

L'élève a choisi d'imaginer la suite du récit de Marguerite en utilisant la première personne. Voici une évaluation détaillée :

1. Invention (14 points)

- **Mobilisation des éléments du texte** : L'élève remobilise de façon cohérente les éléments de contexte, en plaçant Marguerite sur le front, dans un environnement de guerre, comme demandé. Il évoque également des aspects de sa personnalité tels que sa détermination et son courage, ce qui est fidèle au personnage.
- **Description du cadre** : La description des conditions de vie dans la tranchée est bien détaillée (froid, insalubrité), ce qui enrichit l'aspect immersif du récit.
- **Développement des péripéties** : L'élève introduit des actions cohérentes avec le rôle de Marguerite comme infirmière, jusqu'à la rencontre d'un soldat blessé et l'accident qui la blesse, ce qui respecte le fil narratif attendu.
- **Sensations et sentiments** : L'élève développe les émotions de Marguerite face aux horreurs de la guerre, bien que cela puisse être approfondi davantage.

Note pour l'invention : 12/14 ↓

2. Organisation et situation d'énonciation (12 points)

- **Enchaînement des étapes du récit** : Les événements sont présentés de manière logique et fluide, bien que certains passages auraient pu être mieux structurés pour renforcer l'impact narratif.
- **Respect de la première personne et cohérence temporelle** : L'élève respecte l'énonciation en première personne tout au long du texte et maintient un temps passé, ce qui correspond aux consignes.
- **Longueur suffisante** : Le texte est d'une longueur appropriée, permettant une évaluation complète de la rédaction.

Note pour l'organisation et situation d'énonciation : 11/12

3. Expression (14 points)

- **Structure des phrases** : Dans l'ensemble, la structure des phrases est correcte, mais quelques maladresses grammaticales et syntaxiques perturbent légèrement la lecture.
- **Orthographe et grammaire** : Il y a plusieurs erreurs d'orthographe, telles que « assaults » pour « assauts » et « ce serai » pour « ce serait », qui nuisent quelque peu à la fluidité de lecture.
- **Lexique** : L'élève utilise un vocabulaire relativement riche et adapté à la situation de guerre, même si certains termes sont parfois maladroits.

Note pour l'expression : 10/14



Commentaires pour la rédaction par l'IA

Là encore une différence sensible de notation apparaît entre l'IA (33 / 40) et le correcteur humain (38 / 40) qui a vraisemblablement été séduit par le style. En outre, je prenais dans chaque enveloppe de correcteurs une copie se trouvant dans les dernières de l'enveloppe : une comparaison inconsciente des copies a nécessairement lieu et le correcteur peut ainsi être amené à noter plus généreusement une bonne copie si les précédentes étaient décevantes.

Sur l'ensemble du corpus des sept copies, une seule (la copie 6) traite le sujet de réflexion (ce qui est statistiquement assez fidèle à la réalité d'une enveloppe classique) : j'ai dû préciser à l'IA qu'il s'agissait d'un sujet de réflexion et elle a appliqué le barème proposé dans le corrigé national. La justification des points attribués est encore une fois développée et pertinente. L'écart des points entre l'IA et l'enseignant est minime (2 points).

Bilan pour l'ensemble des copies

Après avoir exposé dans le détail la confrontation de la correction par l'homme et la machine sur une première copie, voici le bilan sur l'ensemble des sept copies :

Numéro copie	Questions		Dictée		Rédaction		Écart
	IA	Enseignant	IA	Enseignant	IA	Enseignant	
1	40	38.5	4	4	33	38	3,5 plus haut chez le correcteur
2	25.5	30.5	8	9	27	20	1 point plus bas chez le correcteur
3	18	19,5	2	2,5	22	16	4 points plus bas chez le correcteur
4	27.5	31.5	6.5	6.5	20	18	2 points plus haut chez le correcteur
5	29	29	0	0	16	12	4 points plus bas chez le correcteur
6	36	41	8,5	9	26	24	3,5 plus haut chez le correcteur
7	24	27	6,5	7	17	14	0,5 plus haut chez le correcteur

En cas d'écart important de points, notamment pour la rédaction, il est possible de demander à l'IA la réévaluation de

la copie. C'est ce que j'ai fait car l'IA avait mis en premier lieu un 29/40 alors que le correcteur avait mis 20/40. Voici ce que répond alors l'IA :

« Pourquoi ma note était plus élevée

Lors de ma première évaluation, j'ai tenu compte de la cohérence générale et du respect des consignes (rédaction à la première personne, contexte spatio-temporel), ce qui a pu me rendre plus indulgent. Un correcteur humain, en revanche, peut accorder plus d'importance aux détails tels que la richesse narrative, la qualité des transitions et la profondeur des émotions et réflexions.

Conclusion

La note de 20/40 reflète probablement une évaluation stricte des critères de richesse de l'invention, de clarté de l'organisation et de précision de l'expression. Il semble que j'ai initialement sous-estimé l'impact des erreurs et de la simplicité du récit. Avec cette perspective, je pense que la note de 20/40 est justifiée si l'on tient compte de l'exigence d'une plus grande profondeur et d'une meilleure maîtrise de l'écriture. »

CONCLUSION

Les expérimentations actuelles de l'évaluation par l'IA partent toutes d'un contenu déjà numérique et l'on comprend aisément pourquoi : la perte de temps à scanner et revérifier le contenu après lecture par HTR diminue considérablement la rapidité de traitement et donc l'intérêt. Cependant, l'écriture manuscrite reste encore la norme, notamment pour les examens.

Une fois cet obstacle franchi, le gain de temps est évident lorsque la correction est menée par l'IA. Cependant, l'instruction générative doit être précise voire affinée ou modifiée en cours d'évaluation (par exemple pour le cas d'une dictée aménagée pour la copie 7 par exemple) et c'est surtout confier à une machine sans âme le soin de sonder l'originalité de style d'un élève. En outre, une IA ne mesure pas toutes les compétences acquises par l'élève pour parvenir à ces productions contrairement à un enseignant qui saura apprécier des connaissances spécifiques, du vocabulaire approprié, concernant par exemple la Première Guerre Mondiale dans le cas de ce sujet. Par ailleurs, l'IA va rester cantonnée à un corrigé officiel qui ne peut nécessairement exprimer tout l'implicite ou même penser à tout : elle se positionne seulement par rapport à des écarts à la norme. Cependant, sur l'ensemble des sept copies, l'écart du nombre de point total est seulement de 0,5. Mais cela cache parfois des écarts importants sur une même copie.

Enfin, l'IA semble souvent plus généreuse concernant la deuxième partie de l'épreuve, c'est-à-dire la rédaction : c'est là que l'écart avec les correcteurs humains est le plus remarquable. On sent par ailleurs une certaine fébrilité de l'IA lorsque l'on pointe des dissonances avec un correcteur humain. Est-ce une marque d'humilité qui reconnaîtrait encore la suprématie humaine ? En tout cas, cette expérimentation tend à montrer que le regard de l'enseignant est vraiment indispensable et que le système de correction par l'IA est encore perfectible. Mais pour combien de temps ?

Quelques références :

- BRUYEN Christian et FIALAIRE Bernard (rapporteurs), [L'IA et l'avenir du service public](#), rapport n°101 du sénat, 2024-2025.
- DE LA HIGUERA Colin et IYER Jotsna, [IA pour les enseignants : un manuel ouvert, L'intelligence artificielle par et pour les enseignants](#), AI4T, 2024.
- KABORÉ André, [Évaluation en situation d'effectif pléthorique dans le domaine des Lettres à l'Université Joseph Ki-Zerbo : approche par le logiciel AMC](#), Actes du colloque Ouagadougou 2022, Éditions Francophones Universitaires d'Afrique, Novembre 2023.
- MASSE Bastien, [IA et évaluation](#), intervention filmée pour un parcours de formation "[Trajectoires IA en AURA](#)" et

publiée le 25 juin 2024.

- [L'intelligence artificielle dans l'éducation, apports de la recherche et enjeux de la politique publique](#), janvier 2024.

[1] Autorisation accordée par la DEC 8 le 2/07/24

[2] OCR : Optical Character Recognition, en français : reconnaissance optique de caractères

[3] Trois logiciels classiques d'OCR ont été incapables de lire une seule ligne manuscrite

[4] Gemini est une IA développée par Google. Il faut donc avoir un compte et s'authentifier pour pouvoir l'utiliser

[5] Hallucination : C'est une forme de production d'un texte nouveau qui imite le style et le ton d'un humain mais sans comprendre le sens du texte généré

[6] L'HTR : Handwritten Text Recognition, ou reconnaissance d'écritures manuscrites se développe depuis une décennie en utilisant l'IA