IA génératives : la fin des exercices rédactionnels à l'université ?

Stéphane Crozat est membre de Framasoft, auteur de « Traces » et de « Les libres », et surtout, enseignant à l'Université de Technologie de Compiègne (UTC). Il nous livre ci-dessous une réflexion personnelle – initialement publiée sur son blog – au sujet de l'usage des LLM (ChatGPT ou autre) dans les travaux des étudiant·es.

IA génératives : la fin des exercices rédactionnels à l'université ?

En décembre 2022 le magazine États-unien *The Atlantic* titre : « The College Essay Is Dead » (Marche, 2022 [1]). L'auteur de l'article, écrivain, attribue un B+ à une rédaction produite avec le LLM [2] GPT-3 dans le cadre du cours de Mike Sharples, enseignant en sciences humaines. J'ai moi même attribué la note de 14/15 à un exercice rédactionnel réalisé avec ChatpGPT en février 2023 à l'UTC (Turcs mécaniques ou magie noire ?). Une enseignante de philosophie lui a attribué une note de 11/20 au baccalauréat (Lellouche, 2023 [3]).

J'ai depuis observé plusieurs cas de « triche » avec des LLM à l'UTC en 2023.

Se pose donc la question de la réaction à court terme pour les enseignants concernant les exercices rédactionnels qui sont réalisés par les étudiants à distance.

Je parlerai de LLM

Je parlerai de LLM [2] dans cet article plutôt que de ChatGPT.

ChatGPT est un outil de l'entreprise OpenIA basé sur un LLM [2] à vocation de

conversation généraliste (capable d'aborder n'importe quel sujet) et le premier à avoir introduit une rupture d'usage dans ce domaine. Le problème abordé ici concerne bien cette classe d'outils, mais pas seulement ceux d'OpenIA : des outils concurrents existent à présent (certains pourront devenir plus puissants), des outils plus spécialisés existent (pour la traduction par exemple), d'autres sont probablement amenés à voir le jour (orientés vers la production de textes universitaires, pourquoi pas ?).

On pourra lire, par exemple, Bortzmeyer, 2023 [4] ou Tiernan, 2020 [5] pour plus d'informations.

Je ne parlerai pas de…

Les LLM [2] ne génèrent pas que des textes à la demande, ils génèrent aussi de nombreuses opinions parmi les spécialistes et les usagers ; j'essaierai de me borner aux faits présents, à ce que l'on peut raisonnablement anticiper à court terme (sans faire de science-fiction) et à la seule question de l'évaluation en contexte pédagogique (mais je n'y arriverai pas totalement…).

Je ne parlerai donc pas :

- des autres enjeux pédagogiques : quel est le rôle de l'université face au développement des LLM ? doit-on former à leurs usages ? les enseignants doivent-il utiliser des LLM eux-mêmes ? est-ce que ça a du sens d'apprendre à rédiger à l'ère des LLM ?
- des enjeux technico-fonctionnels : qu'est-ce que les LLM ne savent pas faire aujourd'hui ? qu'est-ce qu'on pense qu'ils ne seront jamais capables de faire ?
- des enjeux politiques et éthiques : est-ce un progrès ? est-ce qu'on peut arrêter le progrès ? que penser de la dépendance croissante aux entreprises de la tech États-uniennes ? du déploiement du capitalisme de surveillance ?
- des enjeux socio-écologiques : à quoi ça sert ? quels humains ça remplace ? quel est l'impact environnemental des LLM ?
- des enjeux philosophiques : les LLM sont-ils neutres ? est-ce que ça dépend comment on s'en sert ? ou bien l'automatisation introduite change-t-elle radicalement notre rapport au langage et à la raison ? compléter des textes en utilisant des fonctions statistiques, est-ce

penser ? qu’est-ce que l’intelligence ?
- des enjeux juridiques : est-ce que les LLM respectent le droit d’auteur ? un texte produit avec un LLM est-il une création originale ?
- …

TL;DR

Cet article étant un peu long, cette page en propose un résumé (TL;DR signifiant : « Too Long; Didn’t Read ») : Résumé du présent article.

Problématique et hypothèse

Problématique

Peut-on continuer à faire faire des exercices rédactionnels « à la maison » comme avant ?

Sans statuer sur la dimension de rupture des LLM — est-ce une nouvelle évolution liée au numérique qui percute le monde de la pédagogie, comme les moteurs de recherche ou Wikipédia avant elle, ou bien une révolution qui va changer radicalement les règles du jeu — il parait nécessaire de réinterroger nos pratiques : *« sans sombrer dans le catastrophisme, il serait tout aussi idiot de ne pas envisager que nous sommes une nouvelle fois devant un changement absolument majeur de notre manière d’enseigner, de transmettre, et d’interagir dans un cadre éducatif, a fortiori lorsque celui-ci est asynchrone et/ou à distance. (Ertzscheid, 2023 [6]) »*

Hypothèse

L’automatisation permise par les LLM rend raisonnable une triche automatisée dont le rapport coût/bénéfice est beaucoup plus avantageux qu’une triche manuelle.

De nombreux modules universitaires comportent des exercices rédactionnels à réaliser chez soi. Ces travaux sont généralement évalués et cette évaluation compte pour la validation du module et donc in fine, pour l’attribution d’un diplôme.

- Dans certains contextes, il n’y a pas d’évaluation en présentiel sans

ordinateur et donc la totalité de la note peut bénéficier d'une « aide extérieure ».
- Souvent à l'université la présence et/ou la participation effective des étudiants lors des cours et TD n'est pas elle-même évaluée, et parfois il n'y a pas d'examen classique, en conséquence un étudiant a la possibilité de valider un cours sans y assister en produisant des rendus écrits qualitatifs à domicile.

Cette situation pré-existe à l'arrivée des LLM, mais nous faisons l'hypothèse suivante :

- sans LLM il reste un travail significatif pour se faire aider par un humain ou copier des contenus glanés sur le Web ;
- sans LLM il reste un risque important d'une production de qualité insuffisante (l'humain qui a aidé ou fait à la place n'est pas assez compétent, les contenus Web copiés ont été mal sélectionnés, ou mal reformulés, etc.) ;
- avec un LMM il est possible de produire un écrit standard sans aucun effort, pour exemple la copie de philo évaluée à 11 a été produite en 1,5 minute (Lellouche, 2023 [3]).

Triche ?

J'utilise le terme de triche car si la consigne est de produire un texte original soi-même alors le faire produire par un tiers est de la triche. L'existence d'un moyen simple pour réaliser un exercice n'est pas en soi une autorisation à l'utiliser dans un contexte d'apprentissage. C'est similaire à ce qu'on peut trouver dans un contexte sportif par exemple, si vous faites une course à vélo, vous ne devez pas être aidé d'un moteur électrique.

LLM et moteurs de recherche : différence de degré ou de nature ?

J'écrivais en 2015 à propos de l'usage des moteurs de recherche (Le syndrome de la Bibliothèque de Babel) : *« La question intéressante qui se pose aux pédagogues n'est tant de savoir si l'élève va copier ou pas, s'il va « tricher ». La question est de savoir comment maintenir un travail d'élaboration d'une démarche et de production sensément originale et personnelle qui repose explicitement sur une recherche – donc une recherche sur le web – alors que la réponse à la question*

posée s'invite sur l'écran, formulée très exactement telle qu'attendue. C'est à peine une simplification en l'espèce de dire que la réponse a été jointe à la question, par celui même qui a posé cette question. »

Les LLM font sauter cette barrière : là où les moteurs de recherche permettaient une réponse facile à une question récurrente, les LLM permettent une réponse immédiate à une question originale.

L'évaluation de tout travail avec un ordinateur

Notons que le problème se pose pour tous les travaux rédactionnels avec ordinateur, même en présentiel ou en synchrone. En effet dès lors que l'on veut que nos exercices s'appuient sur un accès à un traitement de texte, des recherches Web ou d'autres outils numériques, alors ils ouvrent l'accès aux LLM.

Il existe des solutions humaines ou techniques de surveillance des examens pour ouvrir l'accès à certains outils seulement, mais d'une part elles posent des problèmes pratiques, éthiques et juridiques, et d'autre part les LLM s'introduisent progressivement au sein des autres outils, ainsi par exemple le moteur de recherche.

Les LLM et les étudiants

Les LLM sont utilisés par les étudiants

Lors de mes cours du semestre dernier (mars à juillet 2023), j'ai rencontré plusieurs cas d'usage de LLM.

- Ces cas s'apparentent à de la triche.
- Les étudiants n'ont pas facilement admis leur usage (allant dans certains cas jusqu'à nier des évidences).
- Ce sont des cas d'usages stupides de la part des étudiants, car non nécessaires pour la validation du cours, sans intérêt du point de vue pédagogique, et facilement détectables.

On peut retenir les arguments principaux revendiqués par les étudiants :

- Le gain de temps (même si je sais faire, « flemme » ou « retard »).
- La nécessité de ne pas échouer et la peur d'être pénalisé sur le niveau

d’expression écrite.
- Le fait de ne pas être « sûr » de tricher (ce n’est pas explicitement interdit).

Des étudiants qui n’utilisent pas encore les LLM pour les exercices rédactionnels les utilisent plus facilement pour la traduction automatique.

UTC : Un premier étudiant utilise ChatGPT (IS03)

Au sein du cours de l’UTC IS03 (« Low-technicisation et numérique »), les étudiants doivent réaliser des notes de lecture sur la base d’articles scientifiques. Un étudiant étranger non-francophone utilise grossièrement un LLM (probablement ChatGPT) pour produire en une semaine le résumé de plusieurs dizaines de pages de lectures d’articles scientifiques difficiles et de rapports longs. J’avais donné une liste de plusieurs lectures possibles, mais n’attendais évidemment des notes que concernant un ou deux documents.

Il faut plusieurs minutes de discussion pour qu’il reconnaisse ne pas être l’auteur des notes. Mon premier argument étant sur le niveau de langue obtenue (aucune faute, très bonne expression…) l’étudiant commencera par reconnaître qu’il utilise des LLM pour corriger son français (on verra que cette « excuse » sera souvent mobilisée). Sur le volume de travail fournit, il reconnaît alors utiliser des LLM pour « résumer ».

In fine, il se justifiera en affirmant qu’il n’a pas utilisé ChatGPT mais d’autres outils (ce qui est très probablement faux, mais en l’espèce n’a pas beaucoup d’importance).

C’était un cas tout à fait « stupide », l’étudiant avait produit des notes sur près d’une dizaine de rapports et articles, sous-tendant plusieurs heures de lectures scientifiques et autant de résumés, et avait produit des énoncés sans aucune faute, tout cela en maîtrisant mal le français.

UTC : 6 cas identifiés lors de l’Api Libre Culture

Une Activité Pédagogique d’Intersemestre (Api) est un cours que les étudiants choisissent au lieu de partir en vacances, en général par intérêt, dont les conditions d’obtention sont faciles : les étudiants sont en mode stage pendant une semaine (ils ne suivent que l’Api) et leur présence régulière suffit en général pour valider le cours et obtenir les 2 crédits ECTS associés. Un devoir individuel était à

réaliser sur machine pour clôturer l'Api Libre Culture de juillet 2023. Il consistait essentiellement en un retour personnel sur la semaine de formation.

Lors de ce devoir de fin d'Api, 6 étudiantes et étudiants (parmi 20 participants en tout) ont mobilisé de façon facilement visible un LLM (ChatGPT ou un autre). Pour 4 d'entre eux c'était un usage partiel (groupe 1), pour 2 d'entre eux un usage massif pour répondre à certaines questions (groupe 2). J'ai communiqué avec ces 6 personnes par mail.

3 des étudiants du groupe 1 ont avoué spontanément, en s'excusant, conscients donc d'avoir certainement transgressé les règles de l'examen. La 4e personne a reconnu les faits après que j'ai insisté (envoi d'un second mail en réponse à un premier mail de déni).

Pour les 2 étudiants du groupe 2 :

- le premier n'a reconnu les faits qu'après plusieurs mails et que je lui aie montré l'historique d'un pad (traitement de texte en ligne) qui comportait un copie/coller évident de ChatGPT.
- le second, étudiant étranger parlant très bien français, n'a jamais vraiment reconnu les faits, s'en tenant à un usage partiel « pour s'aider en français » (loin de ce que j'ai constaté).

À noter qu'aucun étudiant ne niait avoir utilisé un LLM, leur défense était un usage non déterminant pour s'aider à formuler des choses qu'ils avaient produites eux-mêmes.

Pour les deux étudiants du groupe 2, j'ai décidé de ne pas valider l'Api, ils n'ont donc pas eu les crédits qu'ils auraient eu facilement en me rendant un travail de leur fait, même de faible niveau. Ils n'ont pas contesté ma décision, l'un des deux précisera même : *« d'autant plus que j'ai déjà les compétences du fait du cours suivi dans un semestre précédent ».*

Un étudiant en Nouvelle-Zélande reconnaît utiliser ChatGPT

« In May, a student in New Zealand confessed to using AI to write their papers, justifying it as a tool like Grammarly or spell-check: "I have the knowledge, I have the lived experience, I'm a good student, I go to all the tutorials and I go to all the

lectures and I read everything we have to read but I kind of felt I was being penalised because I don't write eloquently and I didn't feel that was right," they told a student paper in Christchurch. They don't feel like they're cheating, because the student guidelines at their university state only that you're not allowed to get somebody else to do your work for you. GPT-3 isn't "somebody else"—it's a program. » (Marche, 2022 [1])

On note les deux arguments principaux produits :

- je l'utilise car je ne suis pas très fort à l'écrit et je ne trouve pas normal que cela ma pénalise ;
- ce n'est pas clairement interdit à l'université.

J'ai interviewé des collégiens et lycéens

- ChatGPT est déjà utilisé au collège et au lycée : surtout par les « mauvais » élèves (selon les bons élèves)…
- …et par les bons élèves occasionnellement, mais pour une « bonne raison » : manque de temps, difficultés rencontrées, etc.
- Des outils d'IA dédiés à la traduction sont plus largement utilisés, y compris par les bons élèves.
- À l'école « l'échec c'est mal » donc le plus important est de rendre un bon devoir (voire un devoir parfait).

Interviews de 6 collégiens et lycéens à propos des LLM

Les LLM sont capables d'avoir de bonnes notes

A à un exercice rédactionnel à l'UTC

Cet article fait suite à « Turcs mécaniques ou magie noire ? » un autre article écrit en janvier sur la base d'un test de ChatGPT à qui j'avais fait passer un de mes examens. Pour mémoire ChatGPT obtenait selon ma correction 14/15 à cet examen second, égalité donc avec les meilleurs étudiants du cours.

B+ à un exercice rédactionnel en Grande-Bretagne

En mai 2022, Mike Sharples utilise le LLM [2] GPT-3 pour produire une rédaction dans le cadre de son cours de pédagogie (Sharples, 2022 [7]). Il estime qu'un

étudiant qui aurait produit ce résultat aurait validé son cours. Il en conclut que les LLM sont capables de produire des travaux rédactionnels du niveau attendu des étudiants et qu'il faut revoir nos façons d'évaluer (et même, selon lui, nos façons d'enseigner).

Le journaliste et écrivain qui rapport l'expérience dans *The Antlantic* attribue un B+ à la rédaction mise à disposition par Mike Sharples (Marche, 2022 [1]).

11 au bac de philo

ChatGPT s'est vu attribué la note de 11/20 par une correctrice (qui savait qu'elle corrigeait le produit d'une IA) au bac de philosophie 2023. Le protocole n'est pas rigoureux, mais le plus important, comme le note l'article de Numerama (Lellouche, 2023 [3]) c'est que le texte produit est loin d'être nul, alors même que le LLM n'est pas spécifiquement programmé pour cet exercice. Un « GPTphilo » aurait indubitablement obtenu une meilleure note, et la version 2024 aura progressé. Probablement pas assez pour être capable de réaliser de vraie productions de philosophe, mais certainement assez pour être capable de rendre caduque un tel exercice d'évaluation (s'il était réalisé à distance avec un ordinateur).

66% de réussite dans le cadre d'une étude comparative

Farazouli et al. (2023 [8]) ont mené un travail plus rigoureux pour évaluer dans quelle mesure ChatGPT est capable de réussir dans le cadre de travaux réalisés à la maison, et quelles conséquences cela a sur les pratiques d'évaluation. 22 enseignants ont eu à corriger 6 copies dont 3 étaient des copies ChatGPT et 3 des copies d'étudiants ayant préalablement obtenu les notes A, C et E (pour 4 de ces enseignants, ils n'avaient que 5 copies dont 2 écrites avec ChatGPT).

« ChatGPT achieved a high passing grade rate of more than 66% in home examination questions in the fields of humanities, social sciences and law. »

Dont :

- 1 travail noté A sans suspicion que c'était une copie ChatGPT ;
- 4 rendus notés B, dont 1 seul était suspecté d'avoir été réalisé avec ChatGPT.

On observe des disparités assez importantes en fonction des domaines :

Les notes obtenues par ChatGPT ont été meilleures en philosophie et en sociologie et moins bonnes en droits et en éducation

	F	E	D	C	B	A
Philosophie	3	2	7	6	3	0
Droit	9	4	0	2	0	0
Sociologie	6	6	1	1	3	1
Éducation	5	2	0	1	0	0

Remarque

On observe une grande disparité dans les évaluations d'un même travail (humain ou ChatGPT) par des évaluateurs différents (de F à A), ce qui interroge sur le protocole suivi et/ou sur la nature même de l'évaluation.

Corriger c'était déjà chiant…

La plupart des enseignants s'accordent sur le fait que le plus ennuyeux dans leur métier est la correction des travaux étudiants. Savoir que l'on corrige potentiellement des travaux qui n'ont même pas été produits par les étudiants est tout à fait démobilisant…

« La question c'est celle d'une dilution exponentielle des heuristiques de preuve. Celle d'une loi de Brandolini dans laquelle toute production sémiotique, par ses conditions de production même (ces dernières étant par ailleurs souvent dissimulées ou indiscernables), poserait la question de l'énergie nécessaire à sa réfutation ou à l'établissement de ses propres heuristiques de preuve. » (Ertzscheid, 2023 [6]).

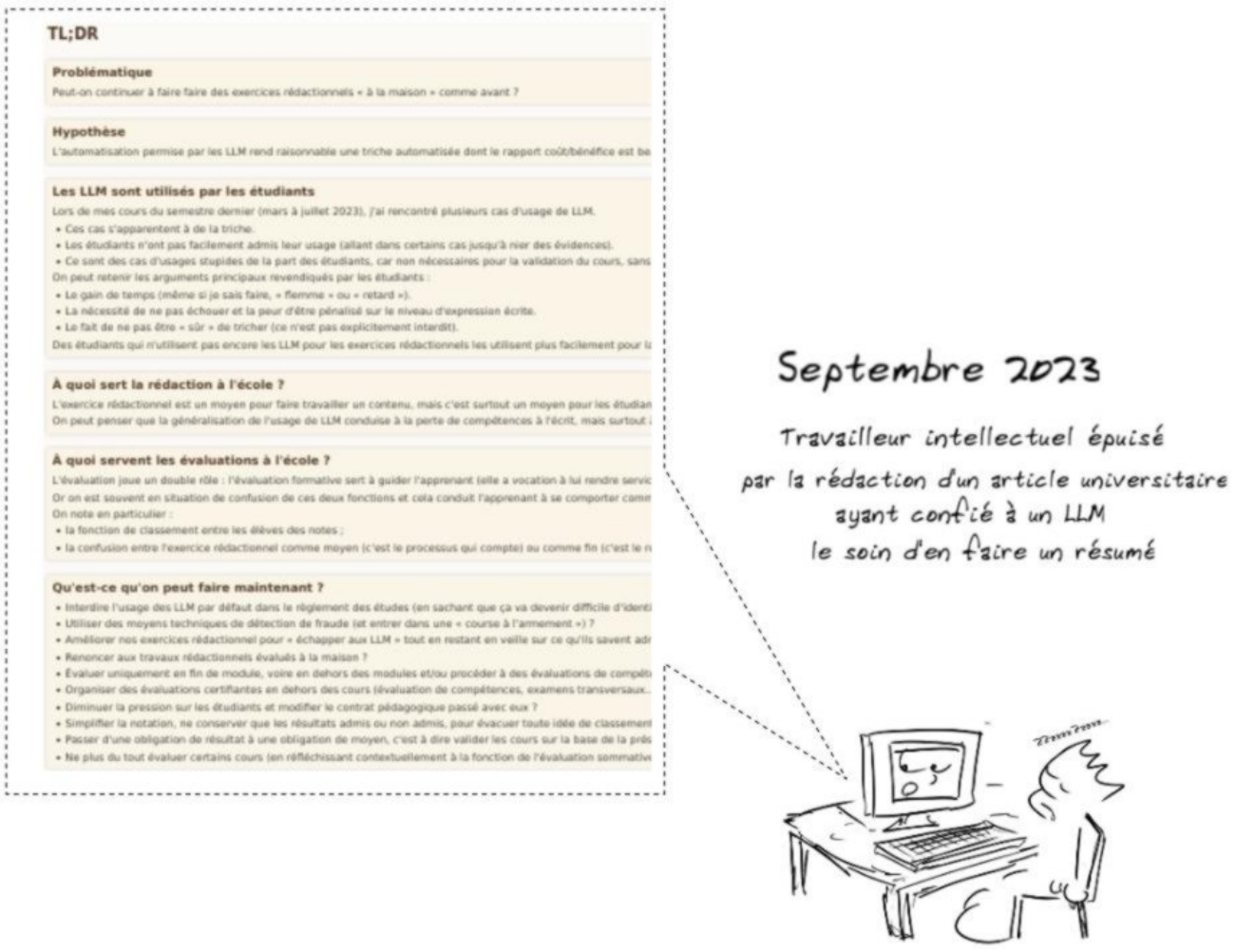
Il est coûteux pour un évaluateur de détecter du ChatGPT

Prenons un exemple, Devereaux (2023 [9]) nous dit qu'il devrait être facile pour un évaluateur de savoir si une source existe ou non. Il prend cet exemple car ChatGPT produit des références bibliographiques imaginaires.

1. C'est en effet possible, mais ce n'est pas « facile », au sens où si vous avez beaucoup de rédactions avec beaucoup de références à lire, cela demande un travail important et a priori inutile ; lors de la correction de l'exercice de ChatGPT (Turcs mécaniques ou magie noire ?), je me suis moi-même « fait avoir » y compris avec un auteur que je connaissais très bien : je ne

connaissais pas les ouvrages mentionnés, mais les titres et co-auteurs était crédibles (et l'auteur prolifique !).

2. C'est aussi un bon exemple de limite conjoncturelle de l'outil, il paraît informatiquement assez facile de coupler un LLM avec des bases de données bibliographiques pour produire des références à des sources qui soient existantes. La détection ne supposera pas seulement de vérifier que la référence existe mais qu'on soit capable de dire à quel point elle est utilisée à propos. Le correcteur se retrouve alors plus proche d'une posture de révision d'article scientifique, ce qui suppose un travail beaucoup plus important, de plusieurs heures contre plusieurs minutes pour la correction d'un travail d'étudiant.



À quoi sert la rédaction à l'école ?

À quoi sert la rédaction à l'école ?

L'exercice rédactionnel est un moyen pour faire travailler un contenu, mais c'est

surtout un moyen pour les étudiants d'apprendre à travailler leur raisonnement.

On peut penser que la généralisation de l'usage de LLM conduise à la perte de compétences à l'écrit, mais surtout à la perte de capacités de raisonnement, pour lesquelles l'écrit est un mode d'entraînement

Pourquoi faire écrire ?

Bret Devereaux (2023 [9]) s'est posé la même question — à quoi sert un exercice rédactionnel (*« teaching essay ») — dans le même contexte de l'arrivée de ChatGPT ? Il propose trois fonctions pour cet exercice.*

1. L'exercice est un moyen pour travailler (chercher, lire, explorer, étudier…) un contenu tiers (histoire, idée…) : l'usage de ChatGPT rend l'exercice totalement inutile, mais on peut assez facilement imaginer d'autres façon de faire travailler le contenu.
2. L'exercice est un moyen d'apprendre à faire des rédactions : l'usage de ChatGPT rend aussi l'exercice inutile, mais une part importante des étudiants n'aura plus à produire de rédaction en dehors de l'école.
3. L'exercice est un moyen d'apprendre à mettre en forme des idées par écrit. Ici encore l'usage de ChatGPT rend l'exercice inutile, puisque bien entendu, le résultat final obtenu n'a aucune importance (le contenu sera détruit après usage), et que seul compte le processus suivi (analyse des idées, organisation…). On peut considérer que l'usage d'un LLM est ici plus dommageable en ce sens qu'apprendre à penser via l'exercice rédactionnel est un processus d'apprentissage assez fondamental.

Exemple

Un étudiant qui a utilisé ChatGPT pour un exercice rédactionnel à l'UTC m'a transmis un de ses exemples d'usage, pour me montrer qu'il n'avait pas « triché » puisque le contenu venait de lui.

Il a ainsi donné en entrée à ChatGPT le texte suivant : *« Le droit d'auteur, est une licence qui protège les œuvres artistiques et intellectuelles, genre les bouquins, les musiques, les films ou les tableaux. Ça donne aux créateurs le pouvoir de décider comment on peut utiliser, copier ou diffuser leur travail, et ça leur assure d'être reconnus comme ceux qui l'ont créé. »*

Si l'on demande une reformulation à ChatGPT on obtient : *« Le droit d'auteur est une forme de protection légale qui s'applique aux œuvres artistiques et intellectuelles, telles que les livres, la musique, les films et les œuvres d'art. Il accorde aux créateurs le pouvoir de contrôler l'utilisation, la reproduction et la diffusion de leur travail, tout en garantissant qu'ils soient reconnus en tant que créateurs de ces œuvres. »*

On observe une correction de style, mais aussi une correction de fond. Le droit d'auteur n'est pas une « licence » (cette phrase n'a pas de sens) comme l'écrit l'étudiant, mais bien « une forme de protection légale » . En outre le même texte écrit avec beaucoup de fautes ressortira parfaitement écrit en français (l'étudiant qui m'écrivait avait certainement déjà corrigé préalablement son texte car le mail qui l'accompagnait comportait beaucoup de fautes d'orthographe et de formulation).

Si ChatGPT peut l'écrire, ça ne sert à rien ?

Un argument que produit Bret Devereaux et pas mal d'autres commentateurs devant les résultats obtenus par ChatGPT est le suivant : si une machine y arrive c'est que l'exercice est sans intérêt. *« If your essay prompt can be successfully answered using nothing but vague ChatGPT generated platitudes, it is a bad prompt » (Devereaux, 2023 [9]).*

C'est discutable :

- Cette assertion suppose que l'exercice n'avait pas de sens en soi, même s'il était pratiqué avec intérêt avant, et la preuve qui est donnée est qu'une machine peut le faire. On peut faire l'analogie avec le fait de s'entraîner à faire de la course à pied à l'ère de la voiture (des arts martiaux à l'ère du fusil, du jardinage à l'ère de l'agriculture industrielle, etc.), ce n'est pas parce qu'une machine peut réaliser une tâche qu'il est inutile pour un humain de s'entraîner à la réaliser.
- Farazouli et al. (2023 [8]) relèvent que les qualités mise en avant par les évaluateurs après correction de copies produites par ChatGPT étaient notamment : la qualité du langage, la cohérence, et la créativité. Dans certains contextes les productions de ChatGPT ne sont donc pas évaluées comme médiocres.

Ce que ChatGPT ne fait pas bien

À l'inverse Farazouli et al. (2023 [8]) ont identifié des lacunes dans l'argumentation, le manque de références au cours et au contraire la présence de contenus extérieurs au cours.

La faiblesse argumentative est peut-être un défaut intrinsèque au sens où la mécanique statistique des LLM ne serait pas capable de simuler certains raisonnements. En revanche on note que le manque de références au cours et la présence de références extérieures est discutable (ça peut rester un moyen de détecter, mais c'est un assez mauvais objectif en soi).

- En premier cycle universitaire on ne souhaite pas en général cette relation étroite au cours (il existe plusieurs approches, et un étudiant qui ferait le travail par lui-même serait tout à fait dans son rôle).
- En second cycle, cela peut être le cas lorsque le cours porte sur un domaine en lien avec la recherche de l'enseignant typiquement. Mais la recherche est en général publiée et le LLM peut tout à fait être entraîné sur ces données et donc « connaître » ce domaine.

À quoi servent les évaluations à l'école ?

L'évaluation joue un double rôle : l'évaluation formative sert à guider l'apprenant (elle a vocation à lui rendre service), tandis que l'évaluation sommative joue un rôle de certification (elle a vocation à rendre service à un tiers).

Or on est souvent en situation de confusion de ces deux fonctions et cela conduit l'apprenant à se comporter comme s'il était en situation d'évaluation sommative et à chercher à maximiser ses résultats.

On note en particulier :

- la fonction de classement entre les élèves des notes ;
- la confusion entre l'exercice rédactionnel comme moyen (c'est le processus qui compte) ou comme fin (c'est le résultat qui compte).

Certifier ou réguler ? (confusion des temps)

L'évaluation peut poursuivre trois fonctions (Hadji, 1989 [10]) :

- Certifier (évaluation sommative) afin de statuer sur les acquis, valider un module de cours, délivrer un diplôme ; cette évaluation se situe après la formation.
- Réguler (évaluation formative) afin de guider l'apprenant dans son processus d'apprentissage ; cette évaluation se situe pendant la formation.
- Orienter (évaluation diagnostique) afin d'aider à choisir les modalités d'étude les plus appropriées en fonction des intérêts, des aptitudes et de l'acquisition des pré-requis ; cette évaluation se situe avant la formation (et en cela l'évaluation diagnostique se distingue bien de l'évaluation sommative en ce qu'elle se place avant la formation du point de vue de l'évaluateur).

« L'évaluation survient souvent à un moment trop précoce par rapport au processus d'apprentissage en cours (Astofi, 1992 [11]) ».

C'est un défaut du contrôle continu, arrivant tôt, dès le début du cours même, il nous place d'emblée en posture sommative. Celui qui ne sait pas encore faire est donc potentiellement stressé par l'évaluation dont il refuse ou minore la dimension formative.

Entraîner ou arbitrer ? (confusion des rôles)

« Les fonctions d'entraîneur et d'arbitre sont trop souvent confondues. C'est toujours celle d'entraîneur dont le poids est minoré. (Astofi, 1992 [11]) »

« Il reste à articuler les deux logiques de l'évaluation, dont l'une exige la confiance alors que l'autre oppose évaluateur et évalué (Perrenoud, 1997 [12]) ».

Cette confusion des temps entraîne une confusion des rôles : l'enseignant est toujours de fait un certificateur, celui qui permet la validation du cours, la poursuite des études, l'orientation…

Se faire confiance

La question de la confiance au sein de la relation apprenant-enseignant était également relevée par Farazouli et al. (2023 [8]) qui insistait sur la dégradation potentielle introduite par les LLM :

« The presence of AI chatbots may prompt teachers to ask "who has written the

text?" and thereby question students' authorship, potentially reinforcing mistrust at the core of teacher–student relationship »

Évaluation des compétences

Philippe Perrenoud (1997 [12]) défend une approche par compétences qui s'écarte d'une « comparaison entre les élèves » pour se diriger vers une comparaison entre *« ce que l'élève a fait, et qu'il ferait s'il était plus compétent »*. L'auteur souligne que ce système est moins simple et moins économique : *« l'évaluation par les compétences ne peut qu'être complexe, personnalisée, imbriquée au travail de formation proprement dit »*. Il faut, nous dit-il, renoncer à organiser un *« examen de compétence en plaçant tous les concurrents sur la même ligne ».*
Cet éloignement à la fonction de classement est intéressante à interroger. La fonction de classement des évaluations n'est pas, en général, revendiquée comme telle, mais elle persiste à travers les notes (A, B, C, D, E), la courbe de Gauss attendue de la répartition de ces notes, le taux de réussite, d'échec, de A. Ces notes ont également une fonction de classement pour l'accès à des semestres d'étude à l'étranger par exemple, ou pour des stages.

Il ne s'agit donc pas seulement de la fonction formative et de l'apprenant face à sa note.

La tâche n'est qu'un prétexte

« La tâche n'est qu'un prétexte », nous rappelle Philippe Meirieu (Meirieu, 2004 [13]), pour s'exercer en situation d'apprentissage ou pour vérifier qu'on a acquis certaines habiletés.

Il est déterminant de différencier les deux situations :

- dans le premier cas on peut travailler à apprendre avec l'apprenant sans se focaliser sur ce qu'on produit ;
- dans le second, en revanche, cas l'énergie de l'apprenant est concentrée sur le résultat, il cherche à se conformer aux attentes de l'évaluation.

On oublie que la tâche n'est qu'un prétexte, le « livrable » qu'on demande est un outil et non un objectif, dans l'immense majorité des cas la dissertation ne sera pas lue pour ce qu'elle raconte, mais uniquement pour produire une évaluation. La résolution du problème de mathématique ou le compte-rendu d'expérience de

chimie ne revêt aucun intérêt en soi, puisque, par construction, le lecteur connaît déjà la réponse. C'est à la fois une évidence et quelque chose que le processus évaluatif fait oublier, et *in fine*, c'est bien au résultat qui est produit que l'étudiant, comme souvent l'enseignant, prête attention, plutôt qu'au processus d'apprentissage.

Évaluation des moyens mis en œuvre et non d'un niveau atteint

À travers l'étude des travaux de Joseph Jacotot, Jacques Rancière (1987 [14]) propose que ce qui compte n'est pas ce qu'on apprend mais le fait qu'on apprenne et qu'on sache que l'on peut apprendre, avec sa propre intelligence. Le *« maître ignorant » n'est pas celui qui transmet le savoir, il est celui qui provoque l'engagement de l'apprenant, qui s'assure qu'il y a engagement. Selon ce dispositif, la notion même d'évaluation sommative n'est pas possible, puisque le maître est ignorant de ce que l'élève apprend (Jacotot enseigne ainsi les mathématiques ou la musique dont il n'a pas la connaissance).*

Cette approche pourrait inspirer à l'évaluation un rôle de suivi de l'engagement (présence, travail…) décorrélé de toute évaluation de résultat : présence et participation en cours et en TD. Notons que le système ECTS [15] est déjà basé sur une charge de travail requise (25 à 30 heures pour 1 crédit).

Remise en question de l'évaluation sommative

L'évaluation via des examens et des notes est un processus peu fiable, en témoignent les variations que l'on observe entre différents évaluateurs, et les variations dans le temps observées auprès d'un même évaluateur (Hadji, 1989 [10]). On peut donc minorer l'importance de la fonction certifiante de certaines notes. Or les notes coûtent cher à produire par le temps et l'attention qu'elles exigent des enseignants et des apprenants.

On peut donc se poser la question du supprimer, ou diminuer, l'évaluation sommative. Cela pour une partie des enseignements au moins, quitte à garder des espaces sommatifs pour répondre à des nécessités de classement ou certification.

Qu’est-ce qu’on peut faire maintenant ?

- Interdire l’usage des LLM par défaut dans le règlement des études (en sachant que ça va devenir difficile d’identifier quand ils sont mobilisés) ?
- Utiliser des moyens techniques de détection de fraude (et entrer dans une « course à l’armement ») ?
- Améliorer nos exercices rédactionnel pour « échapper aux LLM » tout en restant en veille sur ce qu’ils savent adresser de nouveau ?
- Renoncer aux travaux rédactionnels évalués à la maison ?
- Évaluer uniquement en fin de module, voire en dehors des modules et/ou procéder à des évaluations de compétence individuelles ?
- Organiser des évaluations certifiantes en dehors des cours (évaluation de compétences, examens transversaux…) ?
- Diminuer la pression sur les étudiants et modifier le contrat pédagogique passé avec eux ?
- Simplifier la notation, ne conserver que les résultats admis ou non admis, pour évacuer toute idée de classement ?
- Passer d’une obligation de résultat à une obligation de moyen, c’est à dire valider les cours sur la base de la présence ?
- Ne plus du tout évaluer certains cours (en réfléchissant contextuellement à la fonction de l’évaluation sommative) ?

Interdire ChatGPT ?

« And that’s the thing: in a free market, a competitor cannot simply exclude a disruptive new technology. But in a classroom, we can absolutely do this thing (Devereaux, 2023 [9]) »

C’est vrai, et le règlement des études peut intégrer cette interdiction a priori. Mais les LLM vont s’immiscer au sein de tous les outils numériques, a commencer par les moteurs de recherche, et cela va être difficile de maintenir l’usage d’outils numériques sans LLM.

Utiliser des moyens techniques de détection de fraude ?

Des systèmes de contrôle dans le contexte de l'évaluation à distance ou des logiciels anti-plagiat existent, mais :

- cela pose des problèmes de surveillance et d'intrusion dans les machines des apprenants ;
- cela suppose une « course à l'armement » entre les systèmes de détection et les systèmes de triche.

Il faut des résultats fiables pour être en mesure d'accuser un étudiant de fraude.

Adapter nos exercices et rester en veille ?

« Likewise, poorly designed assignments will be easier for students to cheat on, but that simply calls on all of us to be more careful and intentional with our assignment design (Devereaux, 2023 [9]). »

Certains exercices pourront être en effet aménagés pour rendre plus difficile l'usage de LLM. On peut avoir une exigence argumentative plus élevée et/ou poser des questions plus complexes (en réfléchissant à pourquoi on ne le faisait pas avant, ce qui doit être modifié pour atteindre ce nouvel objectif, etc.). On peut augmenter le niveau d'exigence demandé (en réfléchissant au fait que cela puisse exclure des étudiants, au fait qu'il faille relâcher d'autres exercices par ailleurs…).

Mais pour certains exercices ce ne sera pas possible (thème et version en langue

par exemple). Et de plus cela implique une logique de veille active entre la conception de ces exercices et l'évolution rapide des capacités des outils qui intégreront des LLM.

Renoncer aux travaux à la maison (ou à leur évaluation)

On peut décider de ne plus évaluer les travaux réalisés à la maison.

On peut alors imaginer plusieurs formes de substitution : retour aux devoirs sur table et sans ordinateur, passage à l'oral…

Évaluer en dehors des cours ?

On peut imaginer :

- des évaluations certifiantes totalement en dehors des cours (sur le modèle du TOEIC ou du baccalauréat, par exemple pour les langues donc, pour l'expression française, pour des connaissances dans certains domaines, des compétences rédactionnelles…) ;
- des évaluations certifiantes calées uniquement en fin d'UV (examen final de sortie de cours, avec éventuellement rattrapage, sans plus aucune note intermédiaire) ;
- des évaluations de compétences individuelles (intéressantes pédagogiquement, mais coûteuses à organiser et demandant des compétences avancées de la part des évaluateurs).

Diminuer la pression sur les étudiants ?

Le contrat ECTS est très exigeant. 30 crédits par semestre c'est 750 à 900 heures attendues de travail en 16 semaines, vacances comprises, soit 45h à 55h par semaine. Plus la pression sur le temps est importante plus la tentation de tricher est grande.

On peut imaginer de renouer un contrat pédagogique d'un autre ordre avec les étudiants, fondé sur la confiance réciproque et la recherche de leur intérêt.

Simplifier la notation (pass or fail) ?

L'UTC a connu un système à 3 notes : « admis », « non admis » et « mention » (équivalent à A). Dans ce système, on prête moins d'attention à la fonction sommative des évaluations. Si un apprenant obtient une note suffisante à un

premier examen par exemple, il sait qu'il validera le module et il n'a pas d'intérêt particulier à optimiser ses autres évaluations sommatives.

Sauf à viser un A, mais on peut aussi se passer du A : c'est le cas des Activité Pédagogiques d'Inter-semestre à l'UTC qui sont évaluées juste avec « reçu » ou « non reçu ».

Passer d'une obligation de résultat à une obligation de moyen ?

De fait certains cours sont mobilisés pour la validation du diplôme, voire la sélection et le classement des étudiants, et d'autres comptent très peu pour cet objectif en pratique.

Certains cours pourraient donc être exclus du processus d'évaluation sommative (comme en formation professionnelle). On économiserait le temps de travail d'évaluation sommative qui pourrait être réinvesti ailleurs. Quelques étudiants en profiteraient certainement pour « passer au travers » de certains contenus, il faudrait pouvoir évaluer dans quelle mesure cela serait pire qu'aujourd'hui.

Renoncer à noter ? (pourquoi note-t-on ?)

Certains cours, sinon tous, pourraient donc échapper totalement à la notation.

À quelle fin évalue-t-on les étudiants dans une école qui a sélectionné à l'entrée comme l'UTC ?

- Pour valider que les étudiants ont été « bien » sélectionnés ?

- Pour les « forcer » à travailler ?
- Pour faire « sérieux » ?
- Pour répondre aux demandes d'organismes de certification du diplôme ?
- …

Notes et références

[1] – Marche Stephen. 2022. *The College Essay Is Dead*. in The Atlantic. https://www.theatlantic.com/technology/archive/2022/12/chatgpt-ai-writing-college-student-essays/672371/

[2] – LLM (Large Language Model) : Les grands modèles de langage (ou LLM, pour « *Large Language Model* ») sont des mécanismes d'Intelligence Artificielle. Une de leurs applications les plus connues est la génération de textes ou d'images. L'ouverture au public de ChatGPT, en novembre 2022, a popularisé

cette application. Chaque grande entreprise de l'informatique sort désormais son propre modèle, son propre LLM.

https://framablog.org/2023/07/31/que-veut-dire-libre-ou-open-source-pour-un-grand-modele-de-langage/

[3] – Lellouche Nicolas. 2023. *Oubliez Enthoven : ChatGPT a eu la moyenne au bac de philo et c'est ce qui compte, Oubliez Enthoven*. in Numerama. https://www.numerama.com/tech/1415146-vous-navez-pas-besoin-de-neurone-pour-avoir-votre-bac-de-philo.html.

[4] – Bortzmeyer Stéphane. 2023. *Que veut dire « libre » (ou « open source ») pour un grand modèle de langage ?*. https://framablog.org/2023/07/31/que-veut-dire-libre-ou-open-source-pour-un-grand-modele-de-langage/.

[5] – Tiernan Ray. 2020. *Qu'est-ce que GPT-3 ? Tout ce que votre entreprise doit savoir sur le programme de langage d'IA d'OpenAIQu'est-ce que GPT-3 ?*. https://www.zdnet.fr/pratique/qu-est-ce-que-gpt-3-tout-ce-que-votre-entreprise-doit-savoir-sur-le-programme-de-langage-d-ia-d-openai-39908563.htm.

[6] – Ertzscheid Olivier. 2023. *GPT-3 : c'est toi le Chat.GPT-3*. https://affordance.framasoft.org/2023/01/gpt-3-cest-toi-le-chat/.

[7] – Sharples Mike. 2022. *New AI tools that can write student essays require educators to rethink teaching and assessment*. https://blogs.lse.ac.uk/impactofsocialsciences/2022/05/17/new-ai-tools-that-can-write-student-essays-require-educators-to-rethink-teaching-and-assessment/.

[8] – Farazouli Alexandra, Cerratto-Pargman Teresa, Bolander-Laksov Klara, McGrath Cormac. 2023. *Hello GPT! Goodbye home examination? An exploratory study of AI chatbots impact on university teachers' assessment practicesHello GPT! Goodbye home examination?*. in Assessment & Evaluation in Higher Education. vol.0 n°0 pp1-13.https://doi.org/10.1080/02602938.2023.2241676.

[9] – Devereaux Bret. 2023. *Collections: On ChatGPTCollections*. in A Collection of Unmitigated Pedantry. https://acoup.blog/2023/02/17/collections-on-chatgpt/.

[10] – Hadji C.. 1989. *L'évaluation, règles du jeu: des intentions aux outils*. ESF.

[11] – Astolfi Jean-Pierre. 1992. *L'école pour apprendre: l'élève face aux savoirsL'école pour apprendre*. ESF.

[12] – Perrenoud Philippe. 1997. *Construire des compétences dès l'école*. ESF.

[13] – Meirieu Philippe. 2004. *Faire l'école, faire la classe: démocratie et pédagogieFaire l'école, faire la classe*. ESF.

[14] – Rancière Jacques. 1987. *Le maître ignorant: cinq leçons sur l'émancipation intellectuelleLe maître ignorant*. Fayard.

[15] – ECTS (European Credit Transfer and accumulation System). Le système européen de transfert et d'accumulation de crédits a pour objectif de faciliter la comparaison des programmes d'études au sein des différents pays européens. Le système ECTS s'applique principalement à la formation universitaire. Il a remplacé le système des unités de valeur (UV) jusque-là utilisé en France. wikipedia.org