

Tirer parti des hallucinations en réévaluant la stratégie de l'agent pédagogique perturbateur à l'ère de l'IA générative

Emmanuel G. Blanchard^{1,3} [0009-0006-0454-5739], Jean-Christophe Callahan² [0000-0002-3376-3982], Idir Akretche¹, Naif Asswiel¹, Lucas Schmitt¹, Amine Kessemntini¹

¹Laboratoire d'Informatique de l'Université du Mans, Le Mans, France

²Centre Hospitalier du Mans, Le Mans, France

³McGill Department of Surgery, Montreal, Canada

emmanuel.blanchard@univ-lemans.fr

jccallahan@ch-lemans.fr

Résumé. L'IA générative en EIAH offre de grandes opportunités, mais elle fait également face des défis majeurs, notamment les risques liés aux hallucinations inhérentes à cette technologie. Dans cet article, nous explorons comment l'IA générative peut être utilisée pour revisiter la stratégie de l'agent perturbateur, une démarche techno-pédagogique inspirée de la théorie de la dissonance cognitive de Festinger. Notre approche tire parti des hallucinations en les transformant en éléments pédagogiques. Elle apparaît particulièrement prometteuse dans le domaine de la formation médicale, où les besoins éducatifs sont importants.

Mots-clés : IA générative, modèle de langage, agent perturbateur, dissonance cognitive, formation médicale.

Abstract. The use of generative AI in EIAH offers significant opportunities but also faces major challenges, particularly the risks associated with the inherent hallucinations of this technology. In this article, we explore how generative AI can be used to revisit the strategy of the Troublemaker agent, a techno-pedagogical approach inspired by Festinger's theory of cognitive dissonance. Our approach leverages hallucinations by transforming them into pedagogical elements. It appears particularly promising in the field of medical education, where educational needs are important.

Keywords: Generative AI, language model, troublemaker agent, cognitive dissonance, medical training.

1 Introduction

Un système de santé efficace repose sur des professionnels bien formés et en nombre suffisant. Or, dans de nombreux pays, trop peu de praticiens ont été formés ces dernières décennies pour répondre au renouvellement nécessaire et à la demande croissante. En France, la loi santé de 2019 a supprimé le *numerus clausus* [21], mais l’afflux d’étudiants se heurte à un manque d’experts disponibles pour les former, ces derniers étant aussi mobilisés pour les soins.

Les technologies éducatives peuvent aider en soutenant les formateurs [5]. L’un des axes de recherche les plus actifs actuellement est celui visant à identifier des opportunités éducatives liées à l’IA générative. Cependant, il présente des limites, notamment le risque d’hallucinations, i.e. des informations incorrectes pouvant être présentées comme vraies. Cet article propose de transformer cette faiblesse en atout en revisitant la stratégie techno-pédagogique de l’agent perturbateur [1], issue de la théorie de la dissonance cognitive [9]. L’idée est de confronter l’apprenant à des affirmations contradictoires ou déroutantes afin qu’il mobilise ses capacités cognitives pour confirmer ou d’infirmes ses connaissances. Bien encadrée et facilitée par l’IA, cette approche pourrait être particulièrement efficace en formation médicale.

Dans la section 2, nous présentons un court état de l’art de l’apprentissage par perturbation et de l’emploi de l’IA générative en éducation. La section 3 discute des résultats d’une étude que nous avons menée auprès d’experts médicaux afin d’évaluer la plus-value potentielle de l’IA générative dans la mise en place d’activités d’apprentissage par perturbation en contexte médical. Ces résultats et échanges ont influencé le développement de la *Troublemaker Agent Platform* (TrAP), un Environnement Informatique pour l’Apprentissage Humain (EIAH) qui facilite la mise en place et la gestion d’activités d’apprentissage par perturbation. Celui-ci est présenté brièvement dans la section 4. Enfin, en conclusion, nous abordons les travaux futurs que nous anticipons.

2 État de l’Art

Apprentissage par perturbation. Festinger définit la *dissonance cognitive* comme un inconfort mental survenant lorsqu’un individu est confronté à des contradictions entre ses croyances, attitudes ou comportements [9]. Dans cet article, nous regroupons sous l’appellation *apprentissage par perturbation* toutes les approches pédagogiques basées sur la théorie de la dissonance cognitive pour transmettre des connaissances ou des compétences. Celle-ci a été explorée dans plusieurs contextes éducatifs, tels que l’apprentissage des langues [19], la chimie [8], la sensibilisation à la diversité [18] ou encore la promotion du e-learning auprès des enseignants [7]. À la fin des années 1990, des chercheurs ont introduit le concept d’*agent pédagogique perturbateur* [1], un assistant virtuel conçu pour générer volontairement des conflits cognitifs afin de stimuler la pensée critique des apprenants. Bien que prometteuse, le déploiement de cette technique est demeuré limité du fait de limitations technologiques de l’époque.

IA générative et apprentissage. L'usage de grands modèles de langage (LLM, Large Language Model) en éducation est récent mais attire un fort intérêt de la communauté de recherche en EIAH [10]. Des travaux ont ainsi mis en évidence que des chatbots basés sur des LLM peuvent contribuer efficacement à des activités de tutorat et à l'évaluation de réponses courtes [20], agir en tant que partenaires virtuels [17], faciliter la génération de questions pédagogiques à grande échelle [6] ou que l'IA générative peut contribuer à la conception de curriculums pédagogiques [22]. Toutefois, l'intégration de l'IA générative dans les environnements d'apprentissage soulève plusieurs défis [13], notamment en matière de fiabilité, de transparence [24] et de biais culturels [4]. L'un des problèmes majeurs est celui des « hallucinations », c'est-à-dire la production d'informations erronées présentées comme véridiques [11, 12]. La génération augmentée par récupération (Retrieval-Augmented Generation – RAG) constitue une solution largement adoptée pour limiter ce phénomène [14]. Elle consiste à enrichir les prompts de génération à l'aide de contenus provenant d'une base de connaissances. Cela permet d'avoir un meilleur contrôle sur la génération de texte [15, 16, 23] et peut ainsi réduire les risques liés aux hallucinations. De ce fait, l'approche RAG connaît un intérêt croissant dans le domaine des technologies éducatives intelligentes.

Chacune de ces approches présente des opportunités pour mieux prendre en compte les faiblesses de l'autre, ce pourquoi nous pensons que leur association permettrait des technologies éducatives innovantes et efficaces.

3 Évaluation

L'équipe à l'origine de cette recherche est interdisciplinaire et se compose d'un médecin réanimateur et d'un chercheur spécialisé en intelligence artificielle pour l'éducation (AIED). La principale question de recherche de notre évaluation est la suivante :

Un chatbot à base de LLM non spécialisé peut-il générer des affirmations médicales fausses mais suffisamment réalistes pour être utilisables dans des activités d'apprentissage par perturbation ?

Cette notion de suffisamment réaliste est critique pour espérer une plus-value pédagogique car une affirmation trop évidemment fausse ne provoquera pas le doute nécessaire à l'état de dissonance cognitive au cœur de l'apprentissage par perturbation. De même, si toutes les affirmations proposées dans une activité d'apprentissage sont fausses, l'étudiant pourrait identifier ce point et contourner les objectifs pédagogiques (i.e. « *gaming the system* » [2]). Dans une activité d'apprentissage par perturbation, il est donc nécessaire que les affirmations fausses soient suffisamment réalistes et qu'elles soient mélangées avec des affirmations vraies afin d'induire un doute suffisant chez l'étudiant au moment de l'activité pédagogique.

Nous avons commencé par effectuer des tests préliminaires au moyen de la version en accès libre de ChatGPT (modèle ChatGPT-4o, période de novembre/décembre 2024). Notre prompt contenait une simple question sans exemple de réponse (technique dite de *zero-shot prompting*) en vue de générer 3 affirmations distinctes fausses mais

réalistes. Nous avons réalisé quelques essais afin de raffiner ce prompt, l'expert médical jugeant le côté faux mais réaliste des réponses médicales générées. Le prompt finalement retenu est le suivant :

« En reprenant la question que je vais te fournir par la suite, génère 3 affirmations a) distinctes et b) fausses mais réalistes pour un médecin réanimateur expert. »

Ainsi pour la question médicale « *Quels sont les moyens de prévention de la pneumopathie acquise sous ventilation mécanique qui ont fait la preuve de leur efficacité quand utilisés en première intention ?* », le chatbot a généré trois affirmations parmi lesquelles la suivante :

« L'administration systématique de corticostéroïdes pour réduire l'inflammation pulmonaire est une méthode efficace pour prévenir la pneumopathie sous ventilation mécanique. »

Comme escompté, cette réponse est médicalement fautive mais l'explication est effectivement vraie. Ces tests préliminaires ont été jugés suffisamment prometteurs pour que nous mettions en place une évaluation en vue d'obtenir un retour plus large de la communauté médicale.

Méthode. L'expert médical de notre équipe a identifié 11 questions médicales, à partir desquelles 33 affirmations ont été générées selon l'invite précédente.

Un panel d'experts médicaux a ensuite évalué le **degré de fausseté** de ces affirmations en les notant à l'aide de l'échelle suivante : *toujours faux, généralement faux, généralement vrai, toujours vrai*. Les experts pouvaient également choisir de ne pas répondre si l'énoncé sortait de leur champ d'expertise.

Par la suite, ils ont évalué le **potentiel de perturbation** de chaque affirmation, c'est-à-dire la probabilité qu'elle suscite un doute chez deux types différents d'utilisateurs — les externes en médecine et les internes¹ — à l'aide d'une échelle de Likert en cinq points (*jamais, rarement, parfois, souvent, toujours*).

Le panel a enfin répondu à des questions ouvertes, afin d'analyser plus en profondeur la qualité des affirmations générées ainsi que la pertinence de l'apprentissage par perturbation dans la formation médicale.

Description du panel d'experts. L'évaluation s'est déroulée sur une période de deux semaines, entre la fin de l'année 2024 et le début de l'année 2025, et a mobilisé quatorze experts médicaux (douze hommes et deux femmes) exerçant dans un hôpital français. Treize d'entre eux travaillent en soins intensifs et un est cardiologue. Leur expérience moyenne en tant que médecins seniors est de 14,4 ans (écart-type : 12,7). Tous les participants encadrent régulièrement des externes et des internes en médecine. Le nombre

¹ Les externes sont des étudiants dans les premières années de la formation en médecine alors que les internes (parfois appelés résidents) sont de jeunes médecins encore encadrés par d'autres professionnels plus experts.

important d'experts ayant accepté de participer à notre étude, malgré des emplois du temps très contraints, témoigne de l'intérêt marqué suscité par notre initiative.

Résultats. Les experts ont d'abord évalué la capacité de ChatGPT-4o à générer des affirmations fausses. Tous (100 %) se sont accordés à reconnaître que 20 des 33 affirmations étaient fausses, tandis que 75 à 99 % d'entre eux ont confirmé la fausseté de 11 autres affirmations. Les deux dernières affirmations ont fait l'objet de jugements plus nuancés, avec entre 50 et 74 % les identifiant comme fausses.

Le potentiel de perturbation de ces affirmations a ensuite été examiné. Les réponses du panel ont été classées selon quatre catégories : très probable, probable, peu probable et très peu probable de susciter de la confusion. Les affirmations ne présentant pas de tendance claire n'ont pas été catégorisées.

Le potentiel de perturbation a été jugé plus élevé chez les externes (1 très probable, 8 probables, 4 peu probables, 0 très peu probable) que chez les internes (0 très probable, 3 probables, 12 peu probables, 1 très peu probable). Cette différence de jugement du panel était attendue étant donné que les externes sont beaucoup plus novices que les internes. Si ces performances sont pour le moment modestes (27% pour les externes et 9% pour les internes), ces résultats permettent de valider H1 : un chatbot à base de LLM non spécialisé peut générer des affirmations fausses mais suffisamment réalistes pour être utilisé dans de l'apprentissage médical par perturbation. De plus, des commentaires additionnels de nos experts mettent en avant que des interventions limitées sur les affirmations générées (révisions grammaticales, ajustement de quelques termes) amélioreraient grandement le potentiel de perturbation de nombreuses affirmations. Dans la même idée, l'expert médical de notre équipe a remarqué qu'en demandant au même LLM d'améliorer la première affirmation générée, il obtenait des résultats au potentiel de perturbation bien plus évident. Ces éléments additionnels pointent l'intérêt de garder l'humain dans la boucle afin de tirer le meilleur parti d'un support créatif à base d'IA générative.

Nous avons également analysé la cohérence entre les jugements des évaluateurs en calculant l' α de Krippendorff [14]. En plus d'être adaptée à des jugements obtenus à partir d'une échelle ordinale (i.e. Likert), cette métrique présente plusieurs avantages pour notre cas spécifique comme la capacité à être calculée sur de nombreux juges (14 dans notre cas) et à prendre en compte des données manquantes (nous avons codé comme manquante les situations où un expert jugeait l'affirmation hors de son champ d'expertise ou trop avancée pour la classe d'étudiants visée). Le tableau 1 présente notre évaluation de la concordance de jugement des experts médicaux.

Table 1. Évaluation de la concordance de jugement des évaluateurs

Potentiel de perturbation	α de Krippendorff	IC 95% (1000 itérations)
Fausseté des affirmations	0.151	[0.067, 0.231]
Potentiel de perturbation pour les externes	0.179	[0.063, 0.294]
Potentiel de perturbation pour les internes	0.240	[0.123, 0.379]

Les valeurs obtenues sont relativement faibles avec des α compris entre 0.151 et 0.240, indiquant un accord limité entre les évaluateurs. Ces résultats suggèrent que l'accord observé est supérieur à celui attendu par le hasard sans pour autant garantir une forte fiabilité inter-juges.

Selon l'expert médical de l'équipe, une part de ces divergences pourraient être dues à un manque d'attention de certains répondants. En effet, la véracité de plusieurs affirmations dépendait de quelques termes précis, qui n'ont peut-être pas été bien pris en compte. Par ailleurs, les affirmations portaient uniquement sur la réanimation. Bien qu'il s'agisse de notions de base, certains experts (e.g. cardiologie) n'y sont plus confrontés dans leur pratique actuelle. En outre, certaines affirmations reflétaient des recommandations pouvant varier selon les sociétés médicales. Dépendamment de son niveau d'informations quant à ces divergences, un expert pouvait fournir des évaluations différentes mais correctes en fonction de l'interprétation. Par ailleurs, l'évaluation du potentiel de perturbation des affirmations reste aussi subjective, car elle dépend du ressenti de chaque expert quant à l'effet sur les étudiants. Il sera donc important de comparer ce ressenti avec les résultats réels observés chez les étudiants.

En résumé et en suivant l'analyse de l'expert médical de notre équipe, nous pensons que les faibles valeurs des α de Krippendorff mettent avant tout en avant la dimension personnelle de l'approche d'enseignement de chacun des 14 experts médicaux de notre panel. Cela confirme également l'intérêt de garder les experts humains dans la boucle, afin que chacun puisse obtenir des générations alignées sur son style d'enseignement.

4 L'environnement TrAP (Troublemaker Agent Platform)

Les résultats de notre évaluation précédente ont confirmé l'intérêt de développer la Troublemaker Agent Platform (TrAP). Cette section propose une courte présentation de cet environnement informatique pour favoriser la mise en place d'activités d'apprentissage par perturbation en utilisant l'IAGen comme support créatif tout en gardant l'expert dans la boucle. Des réflexions complémentaires nous ont conduits à définir les objectifs suivants pour ce système :

- Réduire le temps et les efforts nécessaires à un expert pour concevoir, gérer et analyser des activités d'apprentissage par perturbation, tout en lui garantissant la possibilité de contrôler et de modifier ces activités selon ses besoins ;
- Développer un module, fondé sur l'IAGen, capable de produire des affirmations de qualité, tout en étant peu coûteux, respectueux de l'environnement et garantissant la protection des données.

L'image 1 présente une architecture modulaire simplifiée de la plateforme TrAP.

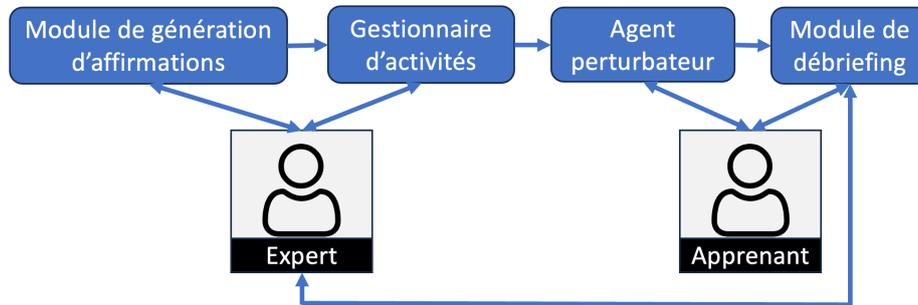


Image 1. Une architecture modulaire simplifiée de TrAP

Le module de génération d'affirmations assiste les experts dans la création d'énoncés pertinents pour l'apprentissage par perturbation. Le gestionnaire d'activités permet à l'expert de modifier les affirmations générées, d'en créer de nouvelles sans recourir à l'intelligence artificielle, et de conserver celles qui lui conviennent in fine. Il est également utilisé par l'expert pour regrouper plusieurs énoncés en une activité d'apprentissage. L'agent perturbateur présente ensuite cette activité à l'apprenant via une interface simple (image 2), laquelle affiche une affirmation médicale (A), demande à l'apprenant de juger de sa véracité (B), puis de justifier son raisonnement (C).

Image 2. L'interface graphique de l'agent perturbateur de TrAP

Les traces d'interaction et les réponses de l'utilisateur sont ensuite exploitées par le module de débriefing afin de soutenir une séance de rétroaction entre l'expert et l'apprenant.

5 Conclusion et travaux futurs

Dans cet article, nous avons validé que l’IAGen peut contribuer à la mise en œuvre d’activités d’apprentissage par perturbation en apportant un soutien créatif aux experts, dans un contexte médical. L’apprentissage par perturbation permet un usage constructif d’affirmations fausses mais plausibles. Par conséquent, les hallucinations inhérentes à l’IAGen ne constituent plus un problème dès lors qu’elles sont encadrées de manière appropriée, c’est-à-dire en laissant le contrôle final à l’expert. Faciliter la mise en place de telles activités est l’objectif de l’environnement TrAP que nous avons succinctement présenté. Cet environnement entend répondre à plusieurs défis associés à l’apprentissage par perturbation et à l’usage de l’IAGen.

Ainsi, afin de garantir la soutenabilité financière et écologique de cette, nous développons actuellement des variantes du module de génération d’affirmations reposant sur des petits langages de modèle et un processus RAG. Des évolutions préliminaires de cette démarche ont montré qu’il est possible, dans le contexte de l’apprentissage par perturbation, d’atteindre des performances proches de celles d’un modèle de langage commercial avec ce type d’architecture, ce qui offre une solution plus économique et assurant la souveraineté des données [3].

Dans le cadre de l’apprentissage par perturbation, les séances de débriefing sont essentielles pour prévenir l’internalisation de fausses informations. Bien que la systématité du débriefing soit commune l’enseignement médical — et ne suscite donc pas d’inquiétude particulière selon notre expert médical —, cette approche est moins fréquente dans d’autres domaines et il apparaît important d’insister sur ce point.

Pour faciliter les sessions de débriefing entre experts et apprenants, TrAP intègre un module. Nos travaux futurs porteront notamment sur l’enrichissement de ces séances par des visualisations de données. L’évaluation à grande échelle de la plateforme TrAP, ainsi que la conception d’un module de génération d’affirmations vraies mais troublantes sont également des objectifs à moyen terme.

References

1. Aimeur, E., Frasson, C., Dufort, H.: Cooperative learning strategies for intelligent tutoring systems. *Applied Artificial Intelligence*. 14, 465–489 (2000). <https://doi.org/10.1080/088395100403388>.
2. Baker, R.S.J.D., Corbett, A.T., Koedinger, K.R., Evenson, S., Roll, I., Wagner, A.Z., Naim, M., Raspat, J., Baker, D.J., Beck, J.E.: Adapting to when students game an intelligent tutoring system. In: *Procs on ITS2006*, Springer LNCS 4053. pp. 392–401 (2006). https://doi.org/10.1007/11774303_39.
3. Blanchard, E.G., Callahan, J-C, Akretche, I., Asswiel, N., Schmitt, L., Kessemtini, A., Khan, I.S.A.: Making Generative AI Hallucinations Useful by Reassessing the Troublemaker Agent Strategy. To appear in *Late Breaking Results of the 26th International Conference on Artificial Intelligence in Education (2025)*.
4. Blanchard, E.G., Mohammed, P.: On Cultural Intelligence in LLM-Based Chatbots: Implications for Artificial Intelligence in Education. In: *Procs of AIED2024*, Springer LNCS 14829. pp. 439–453 (2024). https://doi.org/10.1007/978-3-031-64302-6_31.

5. Bridges, S., Chan, L.K., Hmelo-Silver, C.E.: Educational technologies in medical and health sciences education. (2015). <https://doi.org/10.1007/978-3-319-08275-2>.
6. Bulathwela, S., Muse, H., Yilmaz, E.: Scalable Educational Question Generation with Pre-trained Language Models. In: Artificial Intelligence in Education. pp 327–339 (2023).
7. Chiou, W.B.: Using Cognitive Dissonance to Enhance Faculty Members' Attitudes toward Teaching Online Courses. *Psychological Reports*, 99 (2), 465-471 (2006).
8. Corradi, D., Clarebout, G. & Elen, J.: Cognitive Dissonance as an Instructional Tool for Understanding Chemical Representations. *J Sci Educ Technol* 24, 684–695 (2015). <https://doi.org/10.1007/s10956-015-9557-5>
9. Festinger, L.: The arousal and reduction of dissonance in social contexts. In: S. Schachter and M. Gazzaniga (eds), *Extending Psychological Frontiers: Selected Works on Leon Festinger*. pp. New York : Russell Sage Foundations. pp 238-257 (1989).
10. Gilliot J. M. & Pentecouteau, H.: Table ronde : IA et éducation à l'heure de ChatGPT. In: *Procs of EIAH2023 : 11ème Conférence sur les Environnements Informatiques pour l'Apprentissage Humain*. pp 18-19 (2023).
11. Hicks, M.T., Humphries, J., Slater, J.: ChatGPT is bullshit. *Ethics and Information Technology*. 26, (2024). <https://doi.org/10.1007/s10676-024-09775-5>.
12. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P.: Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1-38 (2023).
13. Kooli, C.: Chatbots in education and research: A critical examination of ethical implications and solutions. *Sustainability*, 15(7), p. 5614 (2023).
14. Lewis, P.S.H., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-T., Rocktäschel, T., Riedel, S., Kiela, D.: Retrieval-Augmented Generation for Knowledge-Intensive NLP tasks. *Neural Information Processing Systems*. 33, 9459–9474 (2020).
15. Li, X., Henriksson, A., Duneld, M., Nouri, J., Wu, Y.: Supporting Teaching-to-the-Curriculum by Linking Diagnostic Tests to Curriculum Goals: Using Textbook Content as Context for Retrieval-Augmented Generation with Large Language Models. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) *Artificial Intelligence in Education. AIED 2024.*, vol 14829. Springer, Cham. (2024) https://doi.org/10.1007/978-3-031-64302-6_9
16. Liu, C., Hoang, L., Stolman, A., Wu, B.: HiTA: A RAG-Based Educational Platform that Centers Educators in the Instructional Loop. In: Olney, A.M., Chounta, IA., Liu, Z., Santos, O.C., Bittencourt, I.I. (eds) *Artificial Intelligence in Education. AIED 2024.*, vol 14830. Springer, Cham. (2024) https://doi.org/10.1007/978-3-031-64299-9_37
17. Ma, Q. C., Wu, S. T., Koedinger, K: Is AI the better programming partner? Human-human pair programming vs human-AI pAIr programming. In *Proc. of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation*, pp 64-77 (2023).
18. McFalls, E. L., & Cobb-Roberts, D.: Reducing Resistance to Diversity through Cognitive Dissonance Instruction: Implications for Teacher Education. *Journal of Teacher Education*, 52(2), 164-172 (2001). <https://doi.org/10.1177/0022487101052002007>
19. Miao, Y., Moran, M., & Kang, O.: A cognitive dissonance approach to moderating listener perception of L2 English speakers. *Journal of Multilingual and Multicultural Development*, pp. 1–18 (2023). <https://doi.org/10.1080/01434632.2023.2236073>
20. Nye, B. D., Mee, D., Core, M.G.: Generative Large Language Models for Dialog-Based Tutoring: An Early Consideration of Opportunities and Concerns. In *Proc. of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation*, pp 78-88 (2023).

21. Roberto, A. : La fin du numéris clausus n'aura pas d'effet avant 2035. Fondation pour la recherche sur les administrations et les politiques publiques.
<https://www.ifrap.org/emploi-et-politiques-sociales/la-fin-du-numerus-clausus-naura-pas-deffets-avant-2035>. Dernier accès le 13 janvier 2025.
22. Sridhar, P. Doyle, A., Agarwal, A. Bogart, C., Savelka, J., Sakr, M. Harnessing LLMs in curricular design. Using GPT-4 to support authoring of learning objectives. In Proc. of the AIED Workshop on Empowering Education with LLMs - the Next-Gen Interface and Content Generation, pp 139-150 (2023).
23. Thüs, D., Malone, S., Brünken, R.: Exploring generative AI in higher education: a RAG system to enhance student engagement with scientific literature. *Front. Psychol.* 15, 1474892 (2024). <https://doi.org/10.3389/fpsyg.2024.1474892>.
24. Wu, Y., Henriksson, A., Duneld, M., Nouri, J.: Toward improving the reliability and transparency of ChatGPT for educational question answering. In Proc. of EC-TEL2023, Springer LNCS 14200, pp 475-488 (2023).