LE Grand Continent

ChatGPT est-il en train de casser le cerveau humain ? 5 points sur l'alarmante étude du MIT sur les effets de l'IA

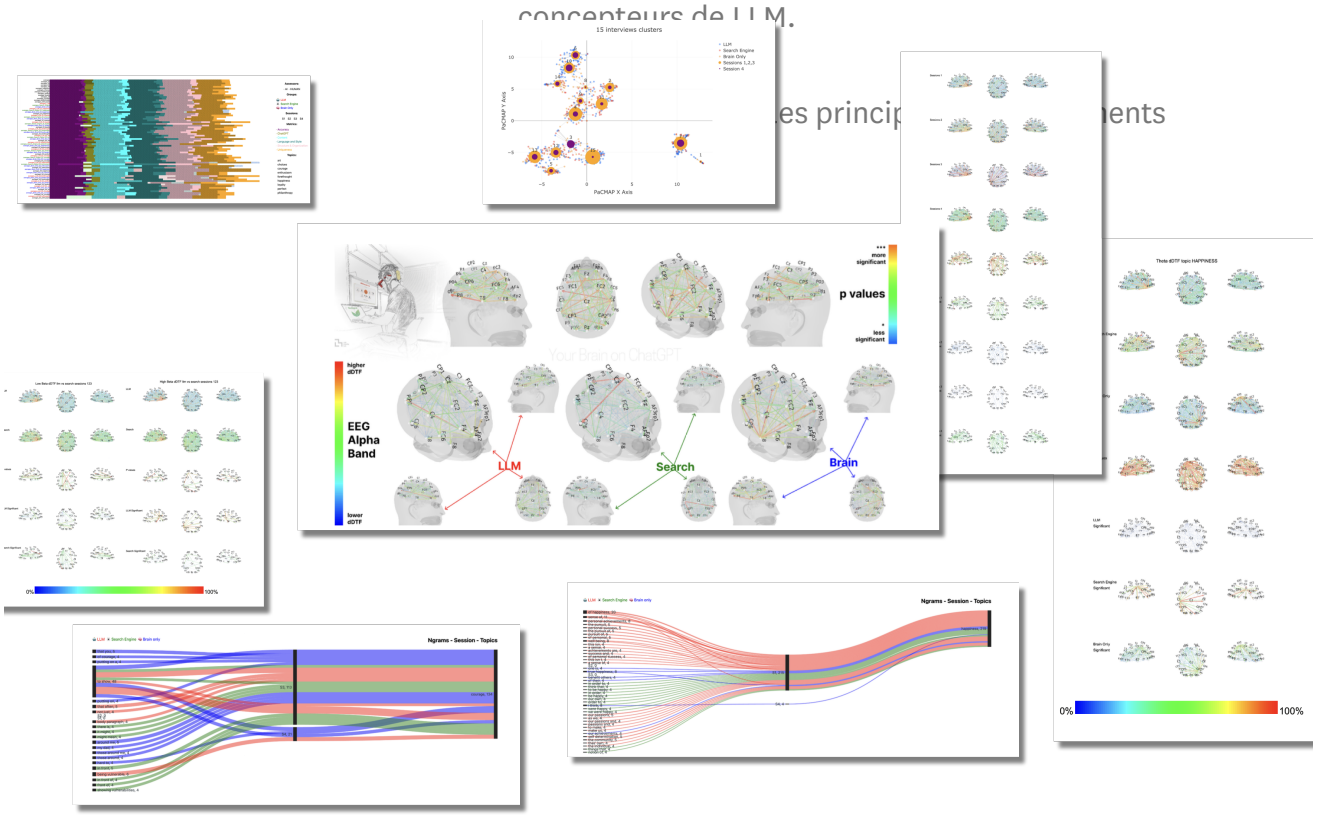
IMAGE © MIT Media Lab

DATE 19 juin 2025

Pour la première fois, une étude scientifique a mesuré ce que l'usage répété de ChatGPT produit sur notre cerveau.

Les résultats marquent une tendance nette : travailler avec les modèles de langage de l'IA fait perdre le contrôle cognitif et modifie le comportement.

Plus inquiétant : les utilisateurs intègrent passivement les biais algorithmiques des concepteurs de LLM.



POINTS CLEFS

- L'étude de leur activité cérébrale montre que, sur le plan neuronal, linguistique et comportemental, les utilisateurs de ChatGPT sous-performent systématiquement par rapport aux personnes qui n'utilisent pas le chatbot pour les mêmes tâches.
- 83,3 % des utilisateurs de ChatGPT soumis au test de l'étude sont incapables de citer des passages d'essais qu'ils avaient rédigés quelques minutes auparavant.
- Dans cette expérimentation, 55 % de la « charge cognitive » nécessaire pour rédiger un essai sans aucune assistance diminue avec l'utilisation d'un LLM provoquant une atrophie cérébrale.
- Dans le temps, écrire avec ChatGPT ferait accumuler une « dette cognitive » rendant difficile un retour à une activité cérébrale normale pour les tâches effectuées sans LLM.

Dans un important *preprint* de 206 pages publié par le MIT Media Lab, Nataliya Kosmyna et ses co-auteurs (disponible intégralement à ce lien) présentent pour la première fois de manière aussi précise les effets de l'utilisation des modèles de langage (LLM) comme ChatGPT sur le cerveau humain.

1 — Une étude qualitative : méthodologie d'un dispositif expérimental inédit

Dans leur expérience menée au MIT pendant plusieurs mois, les chercheurs ont examiné l'effort cognitif induit par trois modes de rédaction d'un essai pour l'examen d'entrée à l'université aux États-Unis (SAT) et étudié l'activité cognitive de trois groupes distincts en regardant chez chacun d'eux les mêmes indicateurs.

un premier groupe rédigeait sans aucune aide (*brain-only*) ;
un second avec un moteur de recherche classique (un *Search Engine* comme Google)
un troisième avec le modèle de langage de ChatGPT (LLM).

54 volontaires recrutés par le MIT Media Lab ont ainsi réalisé trois sessions espacées sur quatre mois.

Chaque session consistait à rédiger, en vingt minutes, un essai répondant à l'un des neuf sujets fournis, à raison de trois sujets différents par session.

Ces trois sessions ont permis aux chercheurs de dégager les différences dans l'activité du cerveau en termes d'activité cognitive, de type d'effort mobilisé – en comparant quelles zones du cerveau sont les plus mobilisées – et les incidences comportementales induites par l'usage d'un moteur de recherche ou d'un LLM – comme la capacité à citer son travail, à considérer son essai comme le sien et à penser de manière critique.

Mais l'expérience a également cherché à comprendre, dans les mêmes conditions de rédaction, *l'effet* que pouvait avoir l'utilisation d'un LLM – en l'occurrence ChatGPT – dans le temps.

Afin de produire des métriques de comparaison, 18 des volontaires du MIT sont revenus pour une quatrième session croisée, où l'objectif était d'analyser les modifications de l'activité cérébrale : de ceux travaillant avec LLM à ceux travaillant sans aucune, et inversement. Dans l'étude, ces groupes sont notés « LLM>Brain-only » et « Brain-only>LLM » (voir *infra*, point 4).

Les conditions expérimentales pour les trois types de rédaction étaient les suivantes :

dans l'hypothèse de l'aide LLM, la rédaction était effectuée avec l'usage exclusif de GPT-4o dans une fenêtre dédiée et aucun autre site Internet n'était autorisé ;
dans l'hypothèse du moteur de recherche classique, l'accès à Internet était libre à l'exception des LLM ;
dans l'hypothèse d'une rédaction sans aide (*brain-only*), aucun outil numérique – aucun écran, donc – n'était autorisé.

Pendant qu'ils rédigeaient, les volontaires ont été soumis à un enregistrement neuro-physiologique à l'aide d'une série d'électrodes et de capteurs placés directement sur leur cuir chevelu.

Pour complexifier l'opération et reproduire les conditions en milieu scolaire, une dernière étape consistait en la notation des copies par des humains et une par une IA conçue sur mesure.

Enfin, à l'issue de l'expérience, les volontaires étaient soumis à un questionnaire.

Conduit sur un nombre restreint de participants, ce test essentiellement qualitatif permet de donner des tendances qui pourraient être des pistes pour la recherche, dans l'attente d'études quantitatives menées sur des échantillons plus importants.

2 — Les effets cognitifs de l'utilisation de ChatGPT : comment les LLM atrophient notre activité cérébrale

Les résultats de l'étude sont sans appel : la « connectivité cérébrale » diminue systématiquement en fonction du soutien externe.

Le groupe travaillant avec seulement son cerveau présente les réseaux neuronaux les plus solides et les plus étendus ; le groupe s'étant aidé d'un moteur de recherche montre un engagement intermédiaire ; et le groupe ayant travaillé avec l'appui d'un LLM suscite le couplage global le plus faible.

Autrement dit : plus le soutien extérieur est élevé, plus l'amplitude des zones actives dans le cerveau est faible.

Cette amplitude est étudiée à partir d'un indicateur, la *Dynamic Direct Transfer Function* (dDTF).

Obtenu à l'aide des données enregistrées par les électrodes, cet outil mathématique sert à observer, instant par instant, quelle zone du cerveau « commande » directement une autre à une fréquence donnée. Elle permet aux neuroscientifiques de suivre en temps réel le flux direct d'information entre régions cérébrales, fréquence par fréquence, et en tirer des marqueurs pertinents.

Les chiffres de l'étude sont à cet égard frappant : le groupe ayant travaillé sans assistance à l'activité la plus élevée ; celui qui travaille avec un moteur de recherche à une activité inférieure d'environ 34 à 48 %.

Quant au groupe travaillant avec ChatGPT, il voit son amplitude cognitive totale réduite de près de 55 %.

L'observation granulaire de cette activité sur les différentes zones du cerveau révèle plusieurs spécificités.

Le groupe ayant rédigé son texte à l'aide d'un moteur de recherche présente une activité accrue dans le cortex occipital et visuel, révélant que l'activité du cerveau se focalise sur la recherche d'informations visuelles et leur compilation dans la phase de recherche.
Ce résultat semble corroborer le fait que les participants ont trié et sélectionné un certain nombre d'informations en vue de rédiger leurs essais. Il s'agit d'une « intégration cognitivement exigeante des ressources visuelles, attentionnelles et exécutives » mobilisée par les écrans.
Si les utilisateurs de LLM passent aussi par un écran, leur groupe ne présente pas de « niveaux comparables d'activation corticale visuelle ». Autrement dit : l'objectif de leur interaction avec l'écran semble distinct : l'utilisation de ChatGPT réduit leur besoin de recherche visuelle prolongée

et de filtrage sémantique. La « charge cognitive » se trouve déplacée dans l'intégration procédurale et la coordination motrice.
Enfin, le groupe n'ayant utilisé que son cerveau montre quant à lui des activations fortes en dehors du cortex visuel, en particulier « dans les régions du cerveau impliquées dans l'intégration sémantique, l'idéation créative et l'autocontrôle exécutif ».

Ces résultats suggèrent que l'utilisation des LLM, si elle augmente la performance des tâches, a surtout un effet sur l'architecture cognitive.

Comme le révèle l'étude : « Le groupe « cerveau seul » a exploité de vastes réseaux neuronaux distribués pour générer du contenu en interne ; le groupe « moteur de recherche » s'est appuyé sur des stratégies hybrides de gestion de l'information visuelle et de contrôle régulateur ; et le groupe « LLM » a optimisé l'intégration procédurale des suggestions générées par l'IA. »

3 — Des conséquences comportementales : la perte de capacité à agir induite par l'usage des LLM

POURQUOI LES UTILISATEURS DE CHATGPT NE SE SOUVIENNENT-ILS PAS DE CE QU'ILS ONT ÉCRIT ?

Les données dites « comportementales » – en particulier celles relatives à la capacité de citation, à l'exactitude des citations et à l'appropriation des essais – prolongent et corroborent les conclusions de l'étude en matière de connectivité neuronale.

Ces résultats suggèrent que la dynamique fonctionnelle du réseau activée pendant la rédaction d'un essai façonne des processus comme la mémoire, l'efficacité de l'autocontrôle et le degré d'appropriation perçue du travail écrit.

La divergence comportementale la plus constante et la plus significative entre les groupes a été observée dans la capacité à citer de tête son propre essai.

Les utilisateurs de LLM obtiennent dans ce domaine des résultats nettement inférieurs : 83 % des participants déclarent avoir des difficultés à citer leur

essai après la première session – et aucun ne fournit une seule citation correcte de son propre travail.

Cette déficience persiste, bien qu'atténuée, lors des sessions suivantes : 6 participants sur 18 ne parviennent toujours pas à se citer correctement lors de la session 3.

Cette difficulté correspond directement à la réduction de la connectivité cérébrale dans le groupe utilisant un LLM : ces oscillations sont en effet généralement plus fortes lorsque les individus génèrent et structurent intérieurement du contenu plutôt que d'intégrer passivement des informations générées à l'extérieur.

L'étude met notamment en avant le fait que la réduction de l'activité cognitive chez les utilisateurs de LLM « reflète probablement un contournement des processus d'encodage profond de la mémoire, les participants lisant, sélectionnant et transcrivant les suggestions générées par l'outil sans les intégrer dans les réseaux de mémoire épisodique ».

Les volontaires utilisant un moteur de recherche comme Google ou ne s'appuyant sur aucun outil numérique n'ont pas présenté de telles déficiences.

À la session 2, les deux groupes ont atteint une capacité de citation presque parfaite, et à la session 3, 100 % des participants des deux groupes ont déclaré être capables de citer leurs essais, observant seulement des écarts mineurs dans la précision des citations.
À l'inverse, l'absence totale de citations correctes dans le groupe LLM lors de la session 1 et les déficiences persistantes lors des sessions suivantes suggèrent que non seulement l'encodage de la mémoire était superficiel, mais surtout que le contenu sémantique lui-même n'avait peut-être pas été entièrement intériorisé.

QUI EST L'AUTEUR DES TEXTES ÉCRITS AVEC L'IA ? ÉTUDE D'UNE DISSOCIATION

Une autre grande divergence comportementale qui ressort de l'étude est la perception qu'ont les participants de la paternité de leur essai.

Alors que le groupe travaillant sans assistance revendique presque à l'unanimité la pleine propriété de ses textes (16/18 lors de la première session et 17/18 lors de la troisième), le groupe rédigeant avec ChatGPT présente un sentiment ambivalent à ce sujet.

Ainsi, certains participants revendiquent être les auteurs de leurs textes, d'autres le nient explicitement et beaucoup s'attribuent une partie seulement du mérite (entre 50 et 90 %).

Cette co-création humain-IA a été théorisée dans nos pages par le dispositif Jianwei Xun qui propose à travers le concept d'hypnocratie un nouveau dispositif d'écriture assumant pleinement et construisant à partir de l'apport des modèles de langage dans les processus d'auctorialité.

Au total, l'étude conclut que « ces réponses suggèrent une diminution du sentiment d'agentivité cognitive ».

Selon les auteurs, cela met en évidence un problème : les outils d'IA, bien que précieux pour soutenir les performances, peuvent involontairement entraver le traitement cognitif approfondi, la rétention et l'engagement authentique avec le matériel écrit.

Cette observation expérimentale semblerait fournir la preuve que si les utilisateurs s'appuient trop fortement sur les outils d'IA, ils peuvent penser acquérir une maîtrise superficielle sans parvenir à intérioriser et à s'approprier les connaissances.

4 — L'IA destructrice : observer les dégâts causés par le LLM sur le cerveau.

L'un des intérêts de cette étude est d'avoir proposé à certains volontaires une quatrième session d'écriture.

Au cours de celle-ci, les groupes ont interverti leurs pratiques pour rechercher les différences d'activité cognitive en passant d'un usage répété du LLM à pas de LLM du tout – et inversement.

La bascule est alors frappante : les volontaires passés de ChatGPT à la seule utilisation de leurs cerveaux peinent à recréer un réseau de connexions et une activité cérébrale aussi riche ; ceux qui ont le droit à l'IA après trois essais libres montrent un pic inédit d'activité, signe qu'ils utilisent l'outil comme un multiplicateur plutôt qu'un substitut – ils écrivent également des prompts plus précis et plus variés.

Ceux qui avaient précédemment écrit sans outils et qui ont eu le droit d'utiliser l'IA (le groupe dit « Brain-only > LLM »), montrent une augmentation significative de la connectivité cérébrale globale, suggérant que la mobilisation soutenue par l'IA suscite des niveaux élevés d'intégration cognitive, de réactivation de la mémoire et de contrôle descendant.

À l'inverse, pour le groupe qui avait rédigé avec l'IA lors des premières sessions et qui a dû s'en passer pour la quatrième, l'utilisation répétée du LLM a reflété une réduction nette de la connectivité au fil du temps.

Selon les auteurs : « ces résultats soulignent l'interaction dynamique entre le soutien cognitif et l'engagement neuronal dans les contextes d'apprentissage assisté par l'IA. »

5 — Inception : la trace de l'algorithme contre la trace de la mémoire

Une autre conclusion plus préoccupante de la quatrième session est que les volontaires du groupe « LLM > Brain-only » se sont concentrés à plusieurs reprises sur un ensemble d'idées plus restreint.

Selon les auteurs, cette répétition pourrait suggérer que de nombreux participants ne se sont peut-être pas engagés profondément dans les sujets ou n'ont pas examiné de manière critique le matériel fourni par le LLM.

Ce schéma reflèterait l'accumulation d'une *dette cognitive* : le recours répété à des systèmes externes tels que les LLM remplacerait des processus cognitifs exigeants nécessaires à la pensée indépendante par des processus purement intégratifs.

La dette cognitive reporterait donc l'effort mental à court terme mais entraînerait des coûts à long terme comme une diminution de l'esprit critique, une vulnérabilité accrue à la manipulation et une baisse de la créativité.

L'étude conclut à cet égard que « lorsque les participants reproduisent des suggestions sans en évaluer l'exactitude ou la pertinence, ils renoncent non seulement à la propriété des idées, mais risquent également d'intérioriser des perspectives superficielles ou biaisées. »

L'analyse de ces différents résultats montre qu'une dépendance précoce à l'IA pourrait conduire à un encodage mémoriel plus superficiel.

Les déficits importants du groupe utilisant les LLM – notamment en termes de mémoire – pourraient ainsi être le signe d'une intégration interne déficiente lors des premiers essais – probablement attribuable à un traitement cognitif externalisé vers le LLM.

À l'inverse, le simple fait de ne pas utiliser les LLM durant les phases initiales pourrait potentiellement favoriser l'émergence de la mémoire.

Chez les rédacteurs n'utilisant aucune assistance numérique, les efforts initiaux sans aide ont ainsi favorisé la formation de traces mémorielles durables, permettant une réactivation plus efficace – même lorsque les outils LLM ont été introduits ultérieurement.