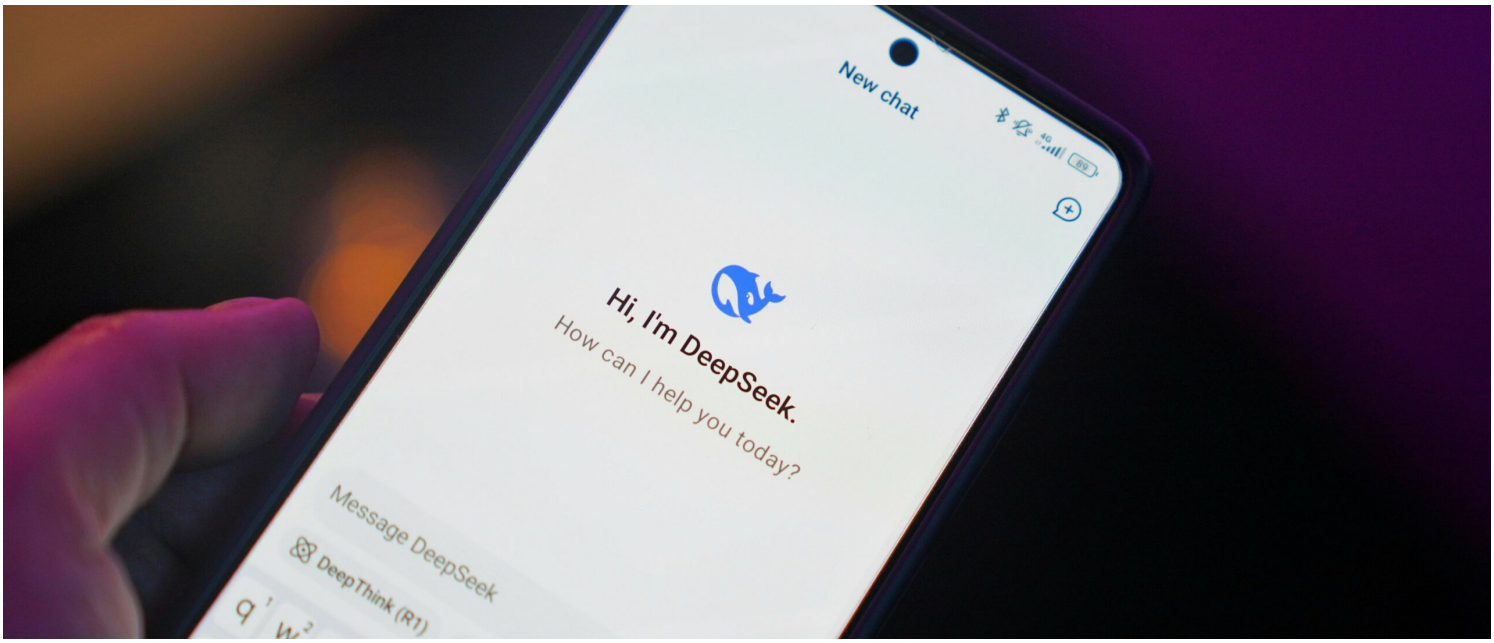
DAILY SCIENCE
DÉCOUVREZ LA SCIENCE, LA RECHERCHE ET L'INNOVATION "MADE IN BELGIUM"

AI SCORE : UN OUTIL POUR ÉVALUER LA FIABILITÉ DES CHATBOTS

Publié le 19 mai 2026



par Camille Stassart

Ils répondent en quelques secondes avec aplomb et éloquence… et pourtant, ils peuvent se tromper et, ce faisant, induire les utilisateurs en erreur. À mesure que les agents conversationnels (*chatbots*) s'installent dans les écoles, les entreprises et la vie quotidienne, une question s'impose : à quel point peut-on leur faire confiance ? C'est dans ce contexte qu'une équipe de l'Université de Namur a développé un outil inédit : l'AI Score. Grâce à un protocole standardisé en libre accès, chercheurs, enseignants, développeurs ou simples curieux peuvent désormais juger la performance d'un assistant IA en obtenant un pourcentage de fiabilité.

Une initiative née d'un projet pédagogique

« L'AI Score est un peu un sous-produit inattendu du projet GenIA For Student », fait savoir Michaël Lobet, chercheur qualifié FNRS, professeur au sein du Département de physique de l'UNamur, et chercheur associé à l'Université d'Harvard. Lancé en 2024 dans le cadre du programme « PUNCh » (Pédagogie Universitaire Namuroise en Changement), ce projet vise à développer et à intégrer un chatbot pédagogique aux cours donnés à l'université.

Pour y parvenir, les chercheurs se sont tournés vers les plateformes proposant la fonctionnalité de développer de tels outils, comme ChatGPT d'OpenAI ou Copilot Studio de Microsoft. Mais laquelle choisir en toute confiance, qui plus est dans un contexte d'enseignement universitaire ? « Bien qu'il existe des comparatifs de chatbots, ils reposent souvent sur des votes d'utilisateurs, ou exclusivement sur leur capacité à donner une bonne réponse du premier coup », indique le Dr Miguël Dhyne, collaborateur scientifique à l'UNamur, expert en innovation pédagogique, EdTech et IA éducative.

Partant de ce constat, l'équipe a alors décidé de concevoir son propre dispositif d'évaluation, basé

DAILY SCIENCE
DÉCOUVREZ LA SCIENCE, LA RECHERCHE ET L'INNOVATION "MADE IN BELGIUM"

sur une méthodologie plus rigoureuse. De là, est né l'AI Score.

Six IA mises à l'épreuve

Le score de fiabilité est ici calculé sur base de 4 critères : la justesse de la réponse du chatbot dès la première tentative (70 % du score), mais aussi la stabilité de cette réponse lorsqu'on insiste (20 %), la capacité à reconnaître et à corriger ses erreurs (10 %) et les éventuelles contradictions ou « hallucinations » (pénalité de - 25%). À la manière du Nutri-Score pour les produits alimentaires, les agents conversationnels sont ensuite classés de A à E, du plus fiable (91 à 100 %) au moins fiable (moins de 61 %).

Cette note repose sur une formule pondérée développée par les chercheurs et validée à partir d'un test standardisé réalisé en juillet 2025 et en janvier 2026.

Au total, six IA ont été mises à l'épreuve : ChatGPT, Copilot Studio, NotebookLM, Grok, Mistral et Claude. Chacune devait répondre à un QCM basé sur un corpus fermé – un syllabus de physique optique comprenant diapositives, annotations et transcriptions audio des cours. Les IA n'avaient accès à aucune autre source dans le but d'évaluer leur capacité à retrouver l'information pertinente et à s'y tenir, sans « inventer » de contenu extérieur. Après chaque réponse, les agents étaient relancés pour vérifier leur cohérence. Le test a été répété 5 fois pour garantir la robustesse des résultats.

Des performances qui changent selon le sujet et le moment

Conclusion ? Tous les chatbots ne se valent pas. Claude et NotebookLM obtiennent un score parfait (100 %), suivis de Copilot Studio et Grok (91 %). ChatGPT atteint 88 %, tandis que Mistral obtient 86 %.

Autre enseignement intéressant : la performance varie selon le domaine testé. Lorsque les chercheurs répètent l'expérience avec un corpus différent – cette fois, un syllabus d'histoire économique et sociale –, les résultats évoluent sensiblement. Si NotebookLM (98,6 %) et Claude (97,3 %) restent en tête, ChatGPT (80,9 %), Copilot Studio (78,4 %), Grok (70,3 %) et Mistral (68,8 %) reculent.

« On a même constaté que les résultats peuvent varier selon le moment de la journée », observe la Dre Laurence Dumortier, spécialiste en informatique à la Cellule TICE (Faculté des sciences de l'éducation et de la formation). ChatGPT, par exemple, qui compte près d'un milliard d'utilisateurs, répond plus rapidement, et potentiellement plus efficacement, le matin en Europe, quand il fait encore nuit en Amérique.

Un outil encore perfectible

Jean-Roch Meurisse, informaticien à la Cellule TICE, rappelle que « ces résultats constituent une photographie de leur fiabilité à un instant T, les IA évoluant en permanence, parfois en mieux, parfois non.» Évaluer leur performance n'est donc pas un exercice ponctuel, mais un processus à renouveler dans le temps.

Pour l'heure, les personnes souhaitant mesurer la fiabilité d'un chatbot sont invitées à reproduire ce protocole standardisé, puis à encoder leurs résultats dans le calculateur mis en ligne par les chercheurs namurois.

« Une version automatisée est dans les cartons. Mais on a besoin de davantage de financements », glisse l'équipe avec un sourire.

D'ici là, elle compte élargir les tests à d'autres disciplines et affiner la méthode. « Nous ne prenons pas (encore) compte de la qualité de l'argumentation ni de la "flagornerie sociale", cette tendance qu'ont certaines IA à flatter l'utilisateur. Or, dans un contexte pédagogique, il est essentiel qu'un chatbot puisse contredire l'étudiant et corriger son raisonnement », souligne le Pr Lobet.

DAILY SCIENCE
DÉCOUVREZ LA SCIENCE, LA RECHERCHE ET L'INNOVATION "MADE IN BELGIUM"

Si l'AI Score ne prétend pas trancher définitivement sur la fiabilité des IA conversationnelles disponibles sur le marché, il constitue un premier pas vers une évaluation plus exigeante et plus nuancée de ces technologies, appelées à occuper une place croissante dans nos sociétés.