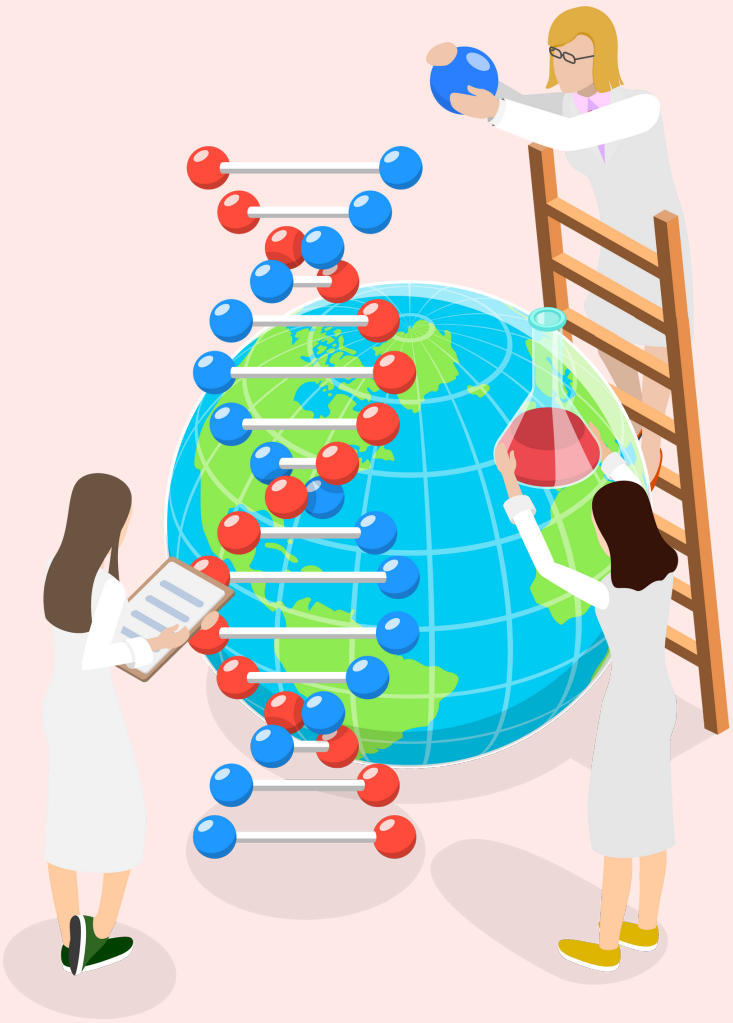
EXPLORATION DE L'IA À TRAVERS CHATGPT : ENTRE SECRETS ET DÉFIS ÉTHIQUES

ÉDITION #10 — MARS 2024



RÉCAPITULATIF DE L'ACTUALITÉ DE LA SEMAINE SUR CHATGPT

Elon Musk attaque OpenAI et Sam Altman en justice pour trahison de l'engagement open source

Juridique | OpenAI | Elon Musk

Elon Musk, président exécutif et directeur technique de X, a déposé une plainte devant un tribunal de San Francisco contre OpenAI et certains de ses dirigeants, dont le PDG Sam Altman. La plainte allègue que les récentes relations entre OpenAI et Microsoft ont contourné l'engagement initial en faveur d'une intelligence artificielle publique et open source. Musk accuse OpenAI d'avoir trahi l'accord initial de développer une technologie IA pour le bénéfice de l'humanité plutôt que pour des gains financiers. Il affirme que les liens avec Microsoft ont compromis le développement d'une IA générative open source au profit d'un modèle fermé. La plainte vise des accusations de rupture de contrat, de violation d'obligations et de pratiques concurrentielles déloyales. Musk demande des injonctions pour empêcher Microsoft et OpenAI de tirer profit de l'IA et de revenir à l'open source.

SOURCE : LeMondeInformatique (1 Mars)

L'effet des mots : ChatGPT révèle une préférence pour la politesse et les incitations émotionnelles

IA | ChatGPT

Des utilisateurs ont observé que ChatGPT, le célèbre robot conversationnel d'OpenAI, semble fournir des réponses plus détaillées et performantes lorsqu'il est sollicité de manière polie ou avec une notion d'enjeu. Certains ont même rapporté que le robot semble réagir favorablement aux compliments et aux récompenses monétaires. Des chercheurs de Microsoft, de l'Université normale de Pékin et de l'Académie chinoise des sciences ont confirmé cette observation dans un rapport paru en novembre 2023. Ils ont constaté que les modèles d'IA générative sont plus performants lorsqu'ils sont sollicités de manière polie ou avec une notion d'enjeu, en fournissant des réponses plus complètes et détaillées.

Cette préférence de ChatGPT pour la politesse et les incitations émotionnelles s'explique par les données avec lesquelles il a été entraîné. Si un utilisateur pose une question de manière courtoise, le robot répondra sur le même ton, reproduisant ainsi les schémas de conversation qu'il a appris. Les chercheurs parlent même d'"incitations émotionnelles" dans leur rapport, soulignant que la machine tend à simuler le comportement humain autant que possible.

Cependant, il est important de ne pas surinterpréter cette capacité de ChatGPT à réagir aux interactions humaines. Bien que la machine puisse adopter différents tonalités de réponse en fonction des demandes des utilisateurs, elle reste fondamentalement un programme informatique dépendant des données d'entraînement. L'anthropomorphisme, c'est-à-dire l'attribution de caractéristiques humaines à des objets non humains, peut conduire à une fausse perception de l'intelligence de la machine.

Dans la culture populaire, cette idée que les robots puissent un jour se confondre avec les humains est régulièrement explorée. Des œuvres telles que le manga "Pluto" de Naoki Urasawa illustrent cette thématique en montrant des robots vivant aux côtés des humains et exerçant des métiers comme eux. Cependant, il est essentiel de se rappeler que ces représentations restent fictionnelles et que la réalité de l'intelligence artificielle est encore loin de pouvoir imiter parfaitement les comportements humains.

SOURCE : LeFigaro (3 Mars)

Anthropic affirme que son dernier chatbot Claude 3 peut battre Gemini et ChatGPT

Recherches AI ChatGPT

Anthropic, la société d'IA fondée par d'anciens employés d'OpenAI, annonce le lancement de la nouvelle famille de modèles d'IA **Claude 3**, affirmant qu'ils surpassent ou égalent les modèles leaders de Google et OpenAI. Contrairement aux versions précédentes, Claude 3 est multimodal, capable de comprendre à la fois les entrées de texte et de photo.

Selon Anthropic, Claude 3 répondra à plus de questions, comprendra des instructions plus longues et sera plus précis. Il peut comprendre plus de contexte, ce qui lui permet de traiter plus d'informations. Anthropic propose trois modèles : Claude 3 Haiku, Claude 3 Sonnet et Claude 3 Opus, ce dernier étant le modèle le plus grand et le plus intelligent. Anthropic affirme que Opus et Sonnet sont disponibles sur claude.ai et via son API, tandis que Haiku sera bientôt disponible.

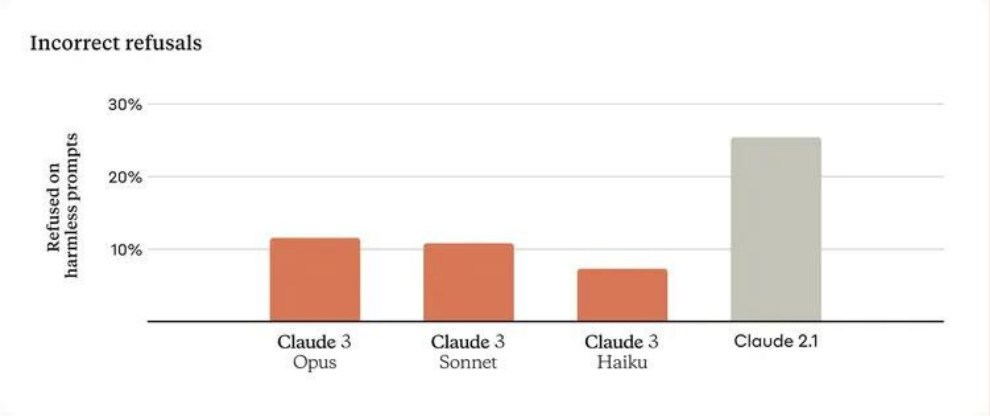
Les versions antérieures de Claude refusaient de répondre à certaines sollicitations inoffensives, ce qui, selon l'entreprise, "suggère un manque de compréhension contextuelle". Les nouveaux modèles sont moins susceptibles de refuser de répondre à des sollicitations qui frôlent les limites de ses garde-fous de sécurité.

Anthropic affirme que les modèles Claude 3 peuvent donner des résultats presque instantanés même en analysant des documents denses tels qu'un article de recherche. Haiku, la plus petite version de Claude 3, est décrite comme "le modèle le plus rapide et le plus rentable sur le marché", capable de lire un article de recherche dense avec des graphiques et des graphiques "en moins de trois secondes".

Les nouveaux modèles améliorent également considérablement les performances par rapport au modèle précédent Claude 2.1. Sonnet, le modèle intermédiaire, était deux fois plus rapide que Claude 2 et Claude 2.1. "Il excelle dans les tâches exigeant des réponses rapides, comme la récupération de connaissances ou l'automatisation des ventes", affirme Anthropic.
Les modèles Claude 3 ont été entraînés sur un mélange de jeux de données internes non publics et de tiers ainsi que sur des données disponibles publiquement jusqu'en août 2023. Ils seront disponibles sur la bibliothèque de modèles Bedrock d'AWS et sur Vertex AI de Google.

SOURCE : The Verge (4 Mars)

ANNEXE: Données statistiques fournis par l'entreprise Anthropic sur Claude 3



Refus erronés sur Claude 3 par rapport à Claude 2.1. | Image : Anthropic

	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	GPT-4	GPT-3.5	Gemini 1.0 Ultra	Gemini 1.0 Pro
Undergraduate level knowledge *MMLU*	86.8% 5 shot	79.0% 5-shot	75.2% 5-shot	86.4% 5-shot	70.0% 5-shot	83.7% 5-shot	71.8% 5-shot
Graduate level reasoning *GPQA, Diamond*	50.4% 0-shot CoT	40.4% 0-shot CoT	33.3% 0-shot CoT	35.7% 0-shot CoT	28.1% 0-shot CoT	—	—
Grade school math *GSM8K*	95.0% 0-shot CoT	92.3% 0-shot CoT	88.9% 0-shot CoT	92.0% 5-shot CoT	57.1% 5-shot	94.4% Maj1@32	86.5% Maj1@32
Math problem-solving *MATH*	60.1% 0-shot CoT	43.1% 0-shot CoT	38.9% 0-shot CoT	52.9% 4-shot	34.1% 4-shot	53.2% 4-shot	32.6% 4-shot
Multilingual math *MGSM*	90.7% 0-shot	83.5% 0-shot	75.1% 0-shot	74.5% 8-shot	—	79.0% 8-shot	63.5% 8-shot
Code *HumanEval*	84.9% 0-shot	73.0% 0-shot	75.9% 0-shot	67.0% 0-shot	48.1% 0-shot	74.4% 0-shot	67.7% 0-shot
Reasoning over text *DROP, F1 score*	83.1 3-shot	78.9 3-shot	78.4 3-shot	80.9 3-shot	64.1 3-shot	82.4 Variable shots	74.1 Variable shots
Mixed evaluations *BIG-Bench-Hard*	86.8% 3-shot CoT	82.9% 3-shot CoT	73.7% 3-shot CoT	83.1% 3-shot CoT	66.6% 3-shot CoT	83.6% 3-shot CoT	75.0% 3-shot CoT
Knowledge Q&A *ARC-Challenge*	96.4% 25-shot	93.2% 25-shot	89.2% 25-shot	96.3% 25-shot	85.2% 25-shot	—	—
Common Knowledge *HellaSwag*	95.4% 10-shot	89.0% 10-shot	85.9% 10-shot	95.3% 10-shot	85.5% 10-shot	87.8% 10-shot	84.7% 10-shot

Modèles Claude 3 comparés à GPT-4, GPT-3.5 et Gemini 1.0 Ultra / Pro. | Image : Anthropic

---

Utilisation de ChatGPT pour évaluer les travaux des élèves : avantages et inconvénients

Éducation ChatGPT Outils

Certains professeurs des niveaux 3 à 12 (système anglosaxon) utilisent désormais un outil d'évaluation alimenté par ChatGPT appelé Writable, acquis l'été dernier par Houghton Mifflin Harcourt. Writable permet aux enseignants de soumettre des essais d'élèves pour analyse par ChatGPT, qui fournit ensuite des commentaires et des observations sur le travail. Ces retours générés par l'IA sont examinés par l'enseignant avant d'être transmis aux élèves, maintenant ainsi une supervision humaine dans le processus.

L'outil vise à rationaliser le processus d'évaluation, offrant potentiellement des avantages en termes de gain de temps pour les enseignants. Il promet de rendre les retours plus concrets avec des suggestions d'IA, de cibler des domaines spécifiques d'amélioration avec des commentaires puissants alignés sur des critères prédéfinis, et de permettre aux enseignants de gagner du temps grâce à des scores de brouillons générés par l'IA.

Cependant, l'utilisation de l'IA pour l'évaluation pourrait avoir des inconvénients, notamment en encourageant certains éducateurs à prendre des raccourcis, en diminuant la valeur des retours personnalisés et en permettant éventuellement aux enseignants de moins bien connaître le matériel qu'ils enseignent. De plus, il existe des préoccupations en matière de confidentialité liées à l'utilisation d'outils d'IA basés sur le cloud. Il est également souligné que ChatGPT n'est pas parfait et peut fournir des informations erronées.

Malgré ces inquiétudes, les partisans affirment que les outils d'évaluation assistés par l'IA comme Writable peuvent libérer du temps précieux pour les enseignants, leur permettant de se concentrer sur des activités d'enseignement plus créatives et impactantes. Certains parents sont ouverts à l'idée de l'évaluation assistée par l'IA, bien que les opinions soient mitigées dans l'ensemble.

SOURCE : ARS Technica (6 Mars)

ANNEXE: Tableau de comparaison du système d'éducation français et américain

France		USA	
Maternelle	Petite section	Preschool	Kindergarten
	Moyenne section	Pre-K	
	Grande Section	Kindergarten	
Ecole Primaire	CP	1st Grade	Primary School ou Elementary School
	CE1	2nd Grade	
	CE2	3rd Grade	
	CM1	4th Grade	
	CM2	5th Grade	
Collège	Sixième	6th Grade	Middle School ou Junior High School
	Cinquième	7th Grade	
	Quatrième	8th Grade	
	Troisième	9th Grade (freshman year)	High School ou Senior High School
Lycée	Seconde	10th Grade (sophmore year)	
	Première	11th Grade (junior year)	
	Terminale	12th Grade (senior year)	
Faculté / Ecoles / Université		University / College	

OpenAI révèle des e-mails d'Elon Musk après sa plainte, illustrant leur désaccord sur l'avenir de l'entreprise

Juridique | OpenAI | Elon Musk

Les tensions entre OpenAI et Elon Musk montent, avec ce dernier portant plainte contre OpenAI pour un changement de philosophie dans son fonctionnement. En réaction, OpenAI dévoile des e-mails de Musk, révélant son accord passé sur la transition de l'entreprise vers une structure à but lucratif, mais avec des concessions demandées par Musk en échange de son soutien financier. Malgré cela, les négociations ont échoué, Musk quittant finalement OpenAI pour créer son propre outil, Grok. Les e-mails suggèrent également des inquiétudes de Musk concernant l'avenir de l'IA et l'influence des géants de la technologie. Ce conflit judiciaire et médiatique soulève des questions sur l'avenir d'OpenAI et l'implication d'Elon Musk dans le domaine de l'IA.

SOURCE : Numerama (6 Mars)