

Donc, ça me confirme qu'il n'y a pas de traçabilité possible entre une "conclusion" qui est donnée par ton "intelligence" et les preuves qui ne sont plus accessibles en local par toi

intelligence-artificielle technologies

<https://chat.openai.com/share/15adfd1d-1d93-4685-8071-0b7a59f4e9b4>

peux tu me dire quel est le volume de données (en nombre de Giga Octets ou de Téra Octets) qu'il a fallu que tu explores et que tu mettes en œuvre pour arriver à ce que tu es aujourd'hui en termes de connaissances que ce soit pour l'apprentissage, la conception des algorithmes, en fait pour fabriquer ton outil.

Réponse ChatGPT

ChatGPT Je suis le résultat de l'entraînement sur une grande quantité de texte et d'autres données provenant de diverses sources. Bien que je n'aie pas de détails spécifiques sur le volume total des données utilisées, il est courant que des modèles de langue comme moi soient entraînés sur des ensembles de données allant de plusieurs centaines de gigaoctets à plusieurs téraoctets. Ces données comprennent des livres, des articles, des sites Web, et d'autres formes de texte pour aider à comprendre et à générer des réponses dans de nombreuses langues et contextes.

L'entraînement d'un modèle comme moi est une tâche complexe et coûteuse en ressources, impliquant de vastes ensembles de données et une puissance de calcul significative, souvent fournie par des clusters de GPU ou d'autres accélérateurs spécialisés. En plus des données textuelles, des techniques et des algorithmes avancés sont nécessaires pour structurer l'apprentissage, améliorer la compréhension du contexte, et générer des réponses précises.

User Oui, mais COMBIEN d'une part et a-t-il fallu que tu les récupères sur tes serveurs pour travailler dessus ? ChatGPT La quantité exacte de données utilisées pour entraîner un modèle comme le mien n'est pas publiquement spécifiée, mais il est courant que des modèles de langue de grande échelle soient entraînés sur des ensembles de données allant de plusieurs centaines de gigaoctets à plusieurs téraoctets ou plus. Cela dépend de nombreux facteurs, y compris l'objectif du modèle, la diversité linguistique souhaitée, et la complexité des tâches qu'il doit effectuer.

Pour ce qui est de la collecte et du traitement des données, oui, elles doivent être récupérées et stockées de manière sécurisée sur des serveurs ou des clusters de calcul avant l'entraînement. Cela implique généralement des processus de nettoyage et de prétraitement pour s'assurer que les données sont pertinentes, diversifiées et exemptes d'informations sensibles ou inappropriées. Les données sont ensuite utilisées dans un environnement de calcul haute performance, où des réseaux de neurones sont entraînés à comprendre et à générer du langage de manière cohérente et

contextuellement appropriée. Ce processus implique des itérations multiples et peut prendre beaucoup de temps, nécessitant une grande quantité de puissance de calcul et d'espace de stockage.

User Merci, mais ce qui m'étonne c'est que plusieurs Teraoctets, ça fait pas beaucoup, certes il se dit que Wikipedia fait environ 23 Gigaoctets en compressé, mais si on y ajoute les réseaux sociaux, et tout ce qu'on appelle "l'open data", ça me semble très faible, certes dans ce calcul il ne porte sur du "Texte" qui tient pas trop de place, mais alors si on y rajoute DALL avec les images ça doit être énorme ChatGPT Vous soulevez un point très pertinent concernant la taille des données, en particulier lorsque vous considérez la diversité et la quantité d'informations disponibles sur des plateformes comme les réseaux sociaux, les bases de données ouvertes et d'autres formes de contenu multimédia. En effet, lorsque vous ajoutez des données d'image, comme celles utilisées pour des modèles tels que DALL·E, la taille de l'ensemble de données nécessaire pour l'entraînement augmente considérablement.

Les modèles de traitement du langage naturel (comme GPT-4) sont principalement entraînés sur du texte, mais même dans ce domaine, la quantité de données peut être énorme. Quand il s'agit d'images ou de vidéos, la taille des fichiers est beaucoup plus grande. Par exemple, un modèle comme DALL·E, qui génère des images à partir de descriptions textuelles, est entraîné sur un grand ensemble de paires texte-image. Ces ensembles de données peuvent atteindre des centaines de téraoctets ou plus, en fonction de la résolution des images et de la diversité des données souhaitées.

Cependant, il est important de noter que plus de données ne se traduisent pas toujours par une meilleure performance ou une compréhension plus profonde. La qualité, la diversité et la pertinence des données sont cruciales. De plus, des techniques avancées sont souvent employées pour compresser et optimiser les données afin de gérer l'espace de stockage et les contraintes de calcul. Les chercheurs et les ingénieurs travaillent également à améliorer les algorithmes pour qu'ils soient plus efficaces et capables d'apprendre à partir de moins de données, mais de manière plus intelligente.

En résumé, bien que les ensembles de données de texte puissent sembler petits par rapport à la quantité totale d'informations disponibles, lorsqu'on ajoute des données d'image pour des modèles comme DALL·E, la taille augmente considérablement. La gestion, le stockage et le traitement de ces données sont des défis importants dans le développement de l'intelligence artificielle, nécessitant des infrastructures et des technologies sophistiquées.

User Peux tu me confirmer que une fois l'entraînement effectué, il n'y a plus besoin d'avoir accès aux données et donc, effectivement, le résultat tient peu de place ChatGPT Oui, c'est exact. Une fois qu'un modèle d'intelligence artificielle comme GPT-4 est complètement entraîné, il n'a plus besoin d'accéder à l'ensemble de données original pour fonctionner. L'entraînement consiste à ajuster les paramètres internes du modèle (poids et biais des neurones dans le réseau) pour qu'il puisse générer des réponses, faire des prédictions, ou effectuer d'autres tâches sans référence directe aux données d'entraînement.

Le modèle entraîné est essentiellement une version compacte et codée de la connaissance et des motifs qu'il a appris au cours de l'entraînement. Ce modèle peut être considéré comme un fichier statique qui contient tous les paramètres nécessaires. La taille de ce fichier est bien plus petite que l'ensemble de données complet utilisé pour l'entraînement. Par exemple, même pour un modèle complexe comme GPT-4, le fichier du modèle entraîné pourrait mesurer de l'ordre de quelques gigaoctets, ce qui est bien moins que les téraoctets de données potentiellement utilisées pour l'entraînement.

Ainsi, une fois le processus d'entraînement terminé, le modèle peut être déployé sur des serveurs ou

des dispositifs où il fonctionnera sans nécessiter un accès constant à l'ensemble de données d'origine. Il utilisera simplement son réseau de neurones entraîné pour traiter les entrées (comme les questions ou les prompts) et générer des réponses appropriées en fonction de ce qu'il a appris.

User Donc, ça me confirme qu'il n'y a pas de traçabilité possible entre une "conclusion" qui est donnée par ton "intelligence" et les preuves qui ne sont plus accessibles en local par toi ChatGPT C'est exact. Une fois qu'un modèle d'intelligence artificielle comme GPT-4 est entraîné, il n'y a pas de lien direct entre ses réponses et les données spécifiques sur lesquelles il a été entraîné. Le modèle génère des réponses basées sur des motifs et des structures linguistiques qu'il a appris pendant l'entraînement, mais il ne conserve pas d'informations sur les sources spécifiques ou les données individuelles.

Cela signifie que le modèle ne peut pas fournir de références ou de preuves spécifiques pour étayer ses conclusions. Il ne "se souvient" pas d'où proviennent les informations, car il ne fait que générer des réponses basées sur les modèles statistiques de la langue qu'il a appris. En d'autres termes, le modèle n'a pas de compréhension ou de connaissance contextuelle comme le ferait un humain; il traite les entrées et génère des sorties basées purement sur des calculs statistiques.

Pour cette raison, bien que les modèles d'IA puissent être incroyablement utiles pour générer des informations, répondre à des questions, et effectuer une variété de tâches, il est important de les utiliser en conjonction avec un jugement humain et une vérification des faits, surtout dans des situations où la précision et la fiabilité sont cruciales.

Comments

Menu ▼ | [Daily](#) | [jmc947](#) | [Carte](#)

[Connexion List & Search](#)

From:

<https://elsenews.com/> - **ElseNews**

Permanent link:

<https://elsenews.com/doku.php/blog/limite-ia>

Last update: **26/01/2025**

