Moveworks

L'informatique

Guide du dirigeant :

de ChatGPT à

IA conversationnelle de niveau entreprise

Avec l'augmentation de l'innovation en matière d'IA, comme le démontre ChatGPT d'OpenAI, la compréhension des grands modèles de langage (LLM) et de l'IA générative ainsi que leurs applications informatiques devient rapidement cruciale pour les responsables informatiques. pour rester compétitif.

Pour les responsables informatiques, comme vous, qui cherchent à réduire les coûts et à augmenter la productivité, en tirant parti de cette fonctionnalité le développement technologique nécessite de connaître l'impact potentiel de ces produits sur leur la main-d'œuvre et les opérations globales. Chaque entreprise doit réfléchir aux implications et à la manière d'adopter ou s'adapter à cette technologie. Sinon, vous prendrez du retard et perdrez votre avantage concurrentiel.

Lisez ce guide pour découvrir comment tirer le meilleur parti de la dernière avancée de l'IA dans l'informatique.

Vous repartirez avec une compréhension de :

- Les principaux termes d'IA que vous devez connaître
- Les forces et les faiblesses des grands modèles de langage (LLM)
- Les principaux cas d'utilisation de l'IA générative en informatique
- Comment distinguer le réel IA conversationnelle et battage médiatique ChatGPT

Ce rapport est particulièrement pertinent pour :

- Directeur de l'information
- Responsable de la Digital Workplace (Directeur+)
- Responsable des services utilisateurs finaux (Directeur+)
- Responsable du Service Desk (Directeur+)
- Autres rapports directs au CIO
- Toute personne intéressée à profiter de l'IA générative dans leur prestation de services informatiques

Table des matières

Les principaux termes de l'IA que les responsables informatiques doivent connaître

Oubliez la crypto et la blockchain. L'IA générative est la nouvelle technologie en ville, et vous devez connaître le jargon.

ChatGPT. Grands modèles de langage. Les réseaux de neurones. Ingénierie rapide. PNL. NLU. GNL. Nous savons que l'espace numérique regorge d'acronymes et de jargon, et notre glossaire des termes de l'IA est là pour vous aider. Nous avons sélectionné certains des termes d'IA les plus souvent mal compris afin que vous puissiez suivre la conversation.

Vous voulez aller plus loin ? Pour le glossaire complet, cliquez ici.

IA conversationnelle :

Type de technologie qui permet aux ordinateurs de comprendre et de répondre aux entrées en langage naturel d'une manière humaine, permettant aux gens d'interagir avec eux par le biais de texte ou de voix de manière conversationnelle. Exemple : un chatbot capable de comprendre et de répondre au client enquêtes d'une manière naturelle et humaine.

Grand Modèle de Langage (ou « LLM ») :

Un type de modèle formé sur un grand ensemble de données pour effectuer des tâches de compréhension et de génération du langage naturel. Il existe de nombreux LLM célèbres comme BERT, PaLM, GPT-2, GPT-3 et le GPT3.5 révolutionnaire. Tous ces modèles varient en taille (nombre de paramètres pouvant être réglés), en tâches (codage, chat, scientifique, etc.) et sur quoi ils sont formés.

Transformateur génératif pré-entraîné (ou « GPT ») :

L'architecture sous-jacente de ChatGPT, qui est un type de modèle d'apprentissage profond qui est formé sur un grand ensemble de données pour générer un texte de type humain.

GPT-3 :

GPT-3 est la 3ème version de la série GPT de des modèles. Il dispose de 175 milliards de paramètres (boutons réglables) avec des poids à réaliser. prédictions. Chat-GPT utilise GPT-3.5, qui est une autre itération de ce modèle.

ChatGPT :

Une interface de chat construite sur GPT-3.5. GPT-3.5 est un grand modèle de langage développé par OpenAI formé sur une quantité massive de données textuelles Internet et affiné pour effectuer une un large éventail de tâches en langage naturel. Exemple : GPT-3.5 a été affiné pour des tâches telles que traduction linguistique, résumé de texte et réponse aux questions.

Enseignement supervisé:

Un type d'apprentissage automatique dans lequel un modèle est formé sur des données étiquetées pour faire des prédictions sur des données nouvelles et invisibles. Exemple : Un algorithme d'apprentissage capable de classer les images de chiffres manuscrits basés sur des données d'entraînement étiquetées.

Apprentissage non supervisé :

Un type d'apprentissage automatique dans lequel un modèle est formé sur des données non étiquetées pour trouver des modèles ou caractéristiques dans les données. Exemple : Un algorithme d'apprentissage capable de regrouper des images similaires de chiffres manuscrits en fonction de leur caractéristiques visuelles.

Modèles discriminants :

Modèles qui classent un exemple de données et prédisent une marque. Par exemple, un modèle qui identifie si une image est un chien ou un chat.

Modèles génératifs :

Modèles qui génèrent de nouvelles données en découvrant des modèles dans les entrées de données ou les données d'entraînement. Par exemple, créer une nouvelle originale basée sur l'analyse de nouvelles existantes publiées.

Traitement du langage naturel (ou « PNL ») :

Un sous-domaine de l'IA qui implique la programmation ordinateurs pour traiter des volumes massifs de données linguistiques. Se concentre sur la transformation du libre-former du texte dans une structure standardisée.

Compréhension du langage naturel (ou « NLU ») :

Un sous-thème de la PNL qui analyse le texte avec le but de glaner une signification sémantique à partir de langue écrite. Cela signifie comprendre contexte, sentiment, intention, etc.

Génération de langage naturel (ou « NLG ») :

Un sous-domaine de l'IA qui produit des écrits ou des langue parlée.

Les bases des grands modèles de langage

ChatGPT est une technologie révolutionnaire qui a captivé l'imagination du monde entier grâce à sa capacité à montrer la magie de l'IA.

Capable de générer à la volée des réponses semblables à celles des humains et de démontrer un raisonnement puissant, cette technologie a franchi un seuil. Le modèle qui sous-tend ChatGPT, GPT-3.5 et d'autres modèles d'IA générative est sur le point d'alimenter une vague de nouvelles innovations dans le domaine de l'IA conversationnelle. Presque du jour au lendemain, les chatbots de la boîte à outils semblent aussi obsolètes que l'Internet commuté et les disquettes.

Cela dit, ChatGPT et, par extension, les grands modèles de langage (LLM) ne sont pas sans limites. Pour exploiter pleinement le potentiel de la technologie, nous devons comprendre ce qu'elle fait bien – et peut-être plus important encore, quelles sont ses limites.

Vous partirez d'ici en comprenant (1) pourquoi ChatGPT change la donne dans le monde de l'IA conversationnelle, et (2) pourquoi atténuer les inconvénients potentiels des LLM n'est pas vraiment un exercice trivial.

Limites des grands modèles de langage dans les applications d'entreprise

Les LLM ont apporté des avancées impressionnantes dans le domaine de l'IA et ont le potentiel de révolutionner un large éventail d'industries. Cependant, comprendre leurs nuances est crucial pour les utiliser dans des applications réelles. Il existe trois principales faiblesses dans le monde des LLM :

1. Hallucinations
2. Contrôlabilité
3. Mémoire obsolète

Limites des grands modèles de
langage dans les applications d'entreprise

1. Hallucinations

Les entreprises exigent une précision à 100 %.
Mais les LLM prêts à l'emploi ont du mal à
véracité.

ChatGPT, un chatbot propulsé par une OpenAI
modèle de langage appelé GPT-3.5, est capable de
fournir des réponses précises aux questions avec
un taux de précision élevé. Cependant, il existe encore un
il est possible que les réponses soient inexactes
ou complètement faux, un phénomène connu
comme « hallucination ».

Des réponses inexactes peuvent avoir de graves conséquences
conséquences, en particulier dans les domaines critiques pour l'entreprise
des situations telles que les soins de santé et les affaires
opérations. Pour résoudre ce problème, il est crucial
avoir des garanties en place, telles que des garanties humaines
surveillance, pour affiner les entrées et contrôler les sorties.
En conséquence, de nombreuses applications actuelles des LLM
nécessitent une supervision humaine pour des résultats fiables.

2. Contrôlabilité

Vous n'avez pas besoin d'être un expert en apprentissage automatique pour que ChatGPT fasse des choses magiques. N'importe qui peut saisir une invite et obtenir une réponse immédiate. Cela est possible car le LLM sur lequel ChatGPT est construit est un modèle de base – GPT3.5 pour être précis.

Ce modèle peut effectuer un large éventail de tâches et s'adapter à de nouveaux environnements car il est le résultat de la fusion de couches et de couches de modèles plus profonds.

Bien que puissante, cette approche limite
La contrôlabilité de LLM, faisant référence à la capacité de un système qui doit être dirigé ou amené à un état spécifique à l'aide d'une entrée spécifique. Ces couches peuvent réduire considérablement le temps nécessaire à la création et à la formation de systèmes complexes, mais elles limitent la capacité de l'utilisateur à contrôler les réponses du modèle.

Pour qu'un système intelligent prospère dans un environnement professionnel, la contrôlabilité est nécessaire. Les LLM doivent faire partie d'une architecture d'IA plus large qui ajoute du contrôle et des réglages précis, propose des processus de formation et d'évaluation supplémentaires et combine des approches alternatives d'apprentissage automatique.

— Jiang Chen, CTO, AI et
Co-fondateur, Moveworks

ChatGPT a ouvert les yeux des vétérans de l'apprentissage automatique comme moi sur ce qui est possible. Ne vous y trompez pas, le potentiel de ChatGPT pour stimuler l'innovation et faire progresser les entreprises f c'est une technologie essentielle.

3. Mémoire obsolète

Les LLM sont formés sur de grandes quantités de données textuelles
comprendre et répondre au langage naturel
d'une manière humaine. GPT-3 d'OpenAI,
par exemple, a été formé sur un énorme 45
téraoctets de données texte provenant de diverses sources.

Le défi est que les données de formation d'un LLM
sont généralement tirés d'une période de temps spécifique
et peut ne pas refléter fidèlement la situation actuelle
l'état du monde ou les derniers développements.
Remplacer les LLM pour mettre à jour leurs connaissances
est difficile et ne peut être réalisé qu'en recyclant les
modèle, qui est cher.

Il n'est pas facile de demander au LLM de remplacer des pièces
des connaissances du modèle tout en conservant les
autre afin de générer une réponse à jour.
Et même dans ce cas, rien ne garantit qu'un LLM
ne donnera pas d'informations périmées même si
le moteur de recherche avec lequel il est associé est à jour
information. Cela pose un roman et unique
défi, surtout dans un contexte professionnel,
où la plupart des informations sont privées et changent
en temps réel.

Points forts des grands modèles de langage dans les applications d'entreprise

Si vous n’explorez pas et n’investissez pas dans l’IA générative dès maintenant, vous allez être laissé pour compte. Poursuivez votre lecture pour découvrir cinq points forts des LLM qui pourraient grandement profiter à un DSI et à son organisation de soutien :

1. Capacités avancées de traitement du langage naturel prêtes à l’emploi
2. Puissantes capacités génératives
3. Expérience utilisateur conversationnelle transparente
4. Efficacité accrue des utilisateurs

1. Naturel avancé prêt à l'emploi capacités de traitement du langage

Il n'y a pas si longtemps, il fallait une équipe de chercheurs et d'ingénieurs hautement talentueux travaillant des centaines, voire des milliers d'heures, pour créer et former des couches d'algorithmes d'apprentissage automatique : ASR, traduction, correction orthographique, reconnaissance d'entité, résolution d'entité, classification d'intention, emplacement. Remplissage, recherche, réponse aux questions, politique de dialogue et bien plus encore – pour donner un sens à une conversation naturelle. C'était un travail incroyablement complexe.

Mais les récents grands modèles linguistiques, comme le modèle GPT-3.5 qui alimente ChatGPT, changent ce paradigme. Au lieu d'avoir une équipe de personnes travaillant pour créer des couches d'algorithmes reliés entre eux, un modèle unique fait ce que de très nombreux systèmes différents devaient faire auparavant. Essentiellement, avoir une conversation avec ChatGPT, l'interface de GPT-3.5, c'est comme la 3D
imprimer une montre suisse. Soudain, tout le monde a les connaissances et le pouvoir de raisonnement des LLM à portée de main.

2. Puissantes capacités génératives

Nous savons déjà que les idées coûtent cher et que leur exécution est difficile. Tout aspirant entrepreneur peut avoir une idée d'entreprise, mais pour démarrer une entreprise et la faire réussir, il faut plus que cette idée.

ChatGPT a réduit la quantité d'effort nécessaire pour transformer les pensées en mots sur une page. Désormais, en travaillant ensemble, une personne et un LLM pourraient identifier et explorer de nouvelles idées qui n'auraient peut-être pas été découvertes autrement et travailler pour concrétiser ces idées.

Ce partenariat constitue la prochaine étape dans le domaine logiciel, transformant les moteurs d'apprentissage profond en collaborateurs capables de créer du nouveau contenu presque comme le ferait un humain.

Nous disposons désormais de modèles d'IA rapides, de haute qualité et facilement disponibles pour générer du texte, des images, des vidéos, du code logiciel, de la musique, de la voix, des modèles 3D, et plus encore – dont aucun n'est protégé par le droit d'auteur ou plagié. Et l'interface conversationnelle de ChatGPT a prouvé qu'il est possible pour n'importe qui de partager sa vision.

3. Favoriser une expérience utilisateur conversationnelle transparente

Les LLM sont capables de fournir des résultats qui imiter une conversation humaine. Ces des modèles de langage peuvent être intégrés dans des interfaces utilisateur, comme dans un chatbot, pour permettre une communication rapide et naturelle avec clients.

Dans le support informatique, par exemple, les LLM peuvent aider à résoudre les requêtes et les problèmes des employés de manière rapide et précise. L'utilisation du traitement du langage naturel (NLP) permet le LLM pour comprendre l'intention derrière demandes de renseignements, ce qui conduit à un accompagnement plus personnalisé et plus efficace.

Le résultat est une meilleure expérience client et une augmentation potentielle de la satisfaction et de la productivité. En automatisant certaines tâches de support de manière conversationnelle, les entreprises peuvent libérer des ressources précieuses pour se concentrer sur des problèmes plus complexes nécessitant une expertise humaine.

4. Efficacité accrue des utilisateurs

Les LLM sont conçus pour traiter et comprendre le langage humain, ce qui les rend bien adapté pour automatiser les tâches répétitives et tâches chronophages.

Par exemple, en informatique, les LLM peuvent être utilisés pour gérer demandes de renseignements et demandes d'assistance simples, libérant les agents humains pour se concentrer sur des tâches plus complexes problèmes. Dans le secteur du marketing, les LLM peuvent être utilisé pour automatiser la création d'une copie de base vers itérer, réduisant ainsi le temps et les ressources requis pour compléter un actif créatif.

L'efficacité accrue et les économies de coûts apportés par les LLM ne sont que quelques-uns des de nombreuses raisons pour lesquelles ils deviennent rapidement des outils précieux dans tous les secteurs. En automatisant tâches répétitives et chronophages, les LLM sont aider les organisations à se concentrer sur ce qu'elles font mieux, et pour stimuler la croissance et l'innovation.

Les cas d’usage de l’IA générative qui donneront du pouvoir aux équipes informatiques

Nous ne faisons qu’effleurer la surface de ce qui est possible avec l’IA générative.

Maintenant que vous êtes bien informé sur les points forts et faiblesses et LLM, il est temps de plonger dans les cas d'utilisation qui transformera votre entreprise. Par souci de simplicité, nous nous séparerons ces cas d'utilisation en trois catégories principales :

1. Génération de langage
2. Résumé linguistique
3. Génération de code et de données

1. Génération de langage

L'IA générative a le potentiel de révolutionner la façon dont les équipes informatiques travaillent en rationalisant les processus.

et proposer des solutions nouvelles et innovantes. De l'amélioration de la gestion des connaissances à l'automatisation

Lors de la création de documents, les cas d'utilisation suivants mettent en évidence les nombreuses façons dont l'IA générative peut responsabiliser les

équipes informatiques et améliorer leurs flux de travail :

• Recommander de nouveaux articles et formulaires de connaissances

• Rédiger de nouveaux articles et formulaires de connaissances

• Mettre à jour et modifier les articles et le formulaire Knowledge

• Traduire des articles et des formulaires Knowledge à la volée

• Rédiger les communications des employés

• Traduire les communications des employés en temps réel

• Traduire les descriptions des tickets informatiques, les notes communes et les notes de travail.

• Rédiger la documentation produit

2. Résumé linguistique

L'IA générative a le potentiel d'améliorer considérablement l'efficacité et la productivité des équipes de support informatique en automatisant les tâches fastidieuses et chronophages. Cette liste de cas d'utilisation met en évidence certaines des façons dont l'IA générative peut aider à faire apparaître des analyses importantes, à résumer les informations et à fournir des solutions rapides et précises aux sujets liés aux tickets informatiques :

- Extrayez les sujets, symptômes et sentiments courants des tickets informatiques.
- Regrouper les tickets informatiques par sujet
- Générer automatiquement des récits à partir d'analyses
- Résumer les solutions de tickets informatiques pour les agents
- Résumer les longs fils de discussion de tickets informatiques pour les agents
- Résumer les transcriptions de l'assistance téléphonique
- Mettez en évidence les solutions clés dans les extraits d'articles.

3. Génération de code et de données

L'IA générative peut transformer l'infrastructure informatique et le développement de chatbots et faire gagner du temps aux agents informatiques en automatiser des tâches chronophages telles que :

- Codage de produits de notification destinés aux développeurs
- Création de code API REST
- Suggérer des flux de conversation et des modèles de suivi
- Générer des variations d'énoncé pour améliorer la robustesse du langage
- Testez des articles ou des formulaires de connaissances pour une pertinence améliorée

Comment repérer les véritables entreprises d’IA conversationnelle et dépasser le battage médiatique

Naviguer dans le monde complexe des produits d’IA conversationnelle peut être une tâche ardue, mais avec la bonne équipe et la bonne infrastructure en place, les entreprises peuvent exploiter tout le potentiel des grands modèles linguistiques.

Le défi est de ne pas se laisser prendre au piège du battage médiatique, de séparer les prétendants des prétendants et de découvrir le véritable potentiel de l’IA conversationnelle.

Commençons donc par quelques notions de base : la pile d'IA conversationnelle peut être divisé en trois couches :

1. Fondation
2. Milieu
3. Demande

La couche de fondation
— La base de la pile IA

ChatGPT est basé sur un modèle d'IA de base appelé GPT-3.5. Ce type de modèle est formé pour offrir une perspective sur un large éventail de sujets.

Cependant, les modèles de base comme GPT-3 ont quelques défis importants. D'une part, la seule façon de modifier leur sortie est de modifier l'entrée et leurs capacités dépendent de comment ils ont été formés.

De plus, ces modèles ne peuvent pas rechercher de données en temps réel ni créer de nouvelles idées à partir de zéro. Les grands acteurs technologiques comme Microsoft et Google ont leurs propres modèles de base, ce qui rendra la tâche plus difficile. pour les petites entreprises de rivaliser dans ce domaine.

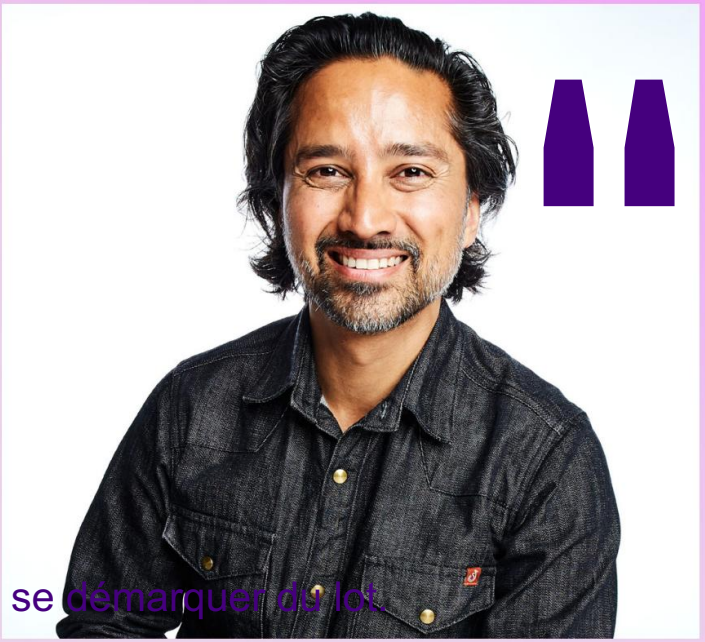
La couche intermédiaire

— Modèles alimentés par des données spécialisées

La couche intermédiaire est celle où sont construits des modèles plus spécialisés, capables de gérer des tâches que la couche de fondation. Ces modèles sont formés sur des données propriétaires très détaillées et sont souvent développés pour un besoin spécifique. application, secteur d'activité, secteur vertical ou cas d'utilisation. En tant que tels, ils surpassent les modèles de fondations dans leurs domaines spécifiques.

Les entreprises peuvent se différencier en ajustant un modèle d'IA de base aux besoins de leur entreprise ou de leur secteur d'activité, en particulier dans les domaines où les données sont très sensibles et où une connaissance spécifique du domaine est requise. La clé du succès dans cette couche est l'accès à des données spécialisées, qui aboutissent à des résultats plus nuancés et un produit plus défendable.

Les gagnants seront applications qui exploitent des données propriétaires. Les entreprises ayant accès à des données spécifiques à un domaine seront en mesure de maximiser la valeur des grands modèles linguistiques fondamentaux et de se démarquer du lot.



— Varun Singh, président
et co-fondateur, Moveworks

3. La couche application — Une interface utilisateur conversationnelle

La couche application fait référence à l'interface où les humains et les machines interagissent, comme comme chatbots ou assistants vocaux. Suite au lancement de ChatGPT, cette couche est cruciale car elle devrait désormais offrir une expérience conversationnelle magique où n'importe qui peut rédiger une invite et obtenir une réponse.

Cependant, disposer d'une interface ne suffit pas pour survivre dans ce domaine concurrentiel. Entreprises qui peuvent apporter des ensembles de données uniques et offrir des réponses précises plus susceptibles de réussir. Les entreprises ne peuvent pas se contenter d'être un mince vernis au-dessus des technologies existantes pour être véritablement compétitives.

N'oubliez pas : ce qui est difficile et ce qui est cher
sont en fin de compte des différenciateurs. Une
entreprise – comme Moveworks – qui se situe à l'intersection du
Les couches de base, intermédiaire et applicative
pourront se différencier de la concurrence.