databricks

EBOOK

# Un guide compact pour Grands modèles de langage

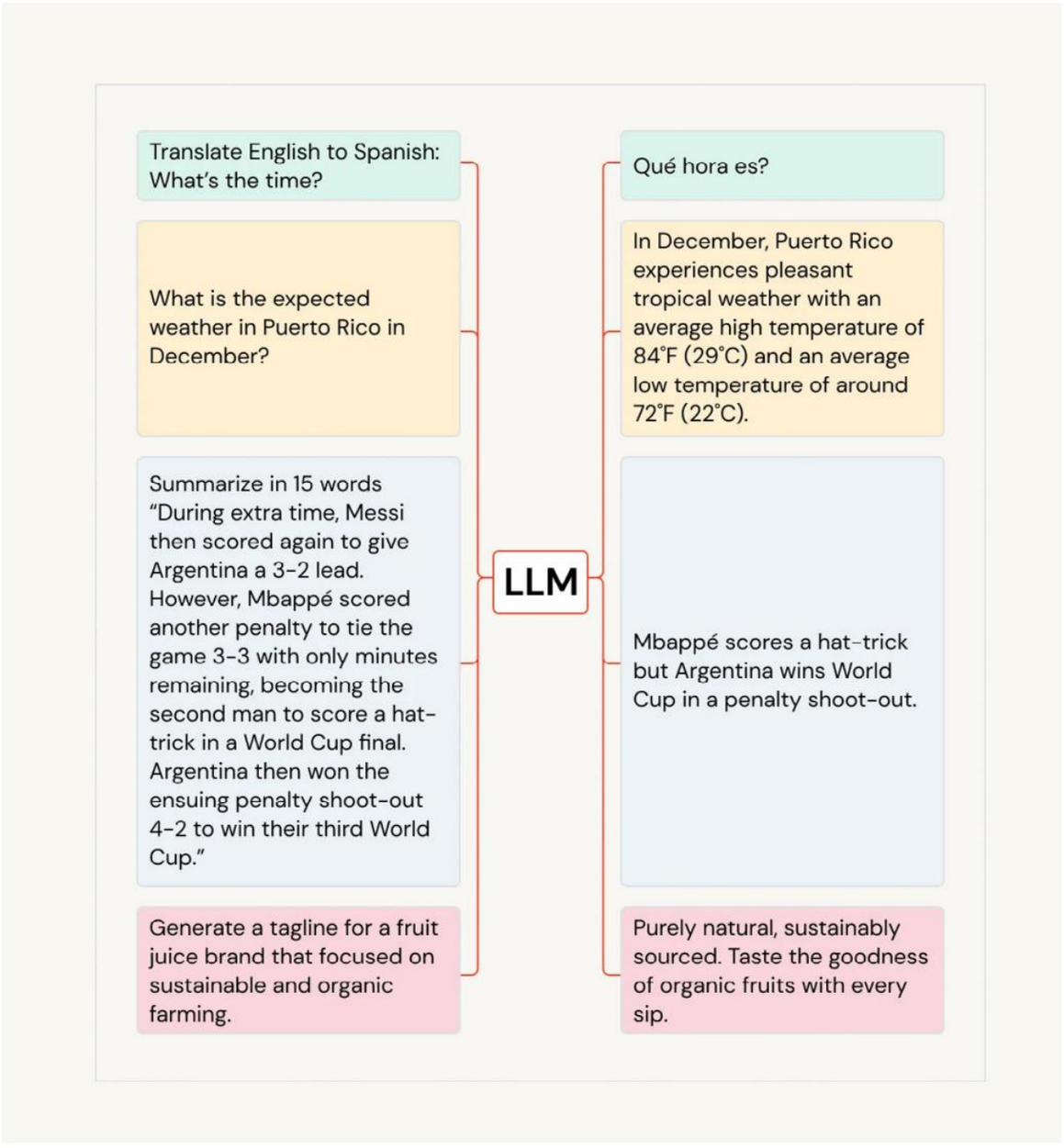
SECTION 1

# Introduction

## Définition de grands modèles linguistiques (LLM)

Les grands modèles de langage sont des systèmes d'IA conçus pour traiter et analyser de grandes quantités de données en langage naturel, puis utiliser ces informations pour générer des réponses aux invites de l'utilisateur. Ces systèmes sont formés sur des ensembles de données massifs à l'aide d'algorithmes d'apprentissage automatique avancés pour apprendre les modèles et les structures du langage humain, et sont capables de générer des réponses en langage naturel à un large éventail d'entrées écrites. Les grands modèles de langage deviennent de plus en plus importants dans une variété d'applications telles que le traitement du langage naturel, la traduction automatique, la génération de code et de texte, etc.

Bien que ce guide se concentre sur les modèles de langage, il est important de comprendre qu'ils ne sont qu'un aspect sous un parapluie plus large d'IA générative. D'autres implémentations d'IA génératives remarquables incluent des projets tels que la génération d'art à partir de la génération de texte, audio et vidéo, et certainement d'autres à venir dans un proche avenir.



## Contexte historique extrêmement bref et développement des LLM

### Années 1950 à 1990

Les premières tentatives sont faites pour établir des règles strictes autour des langues et suivre des étapes logiques pour accomplir des tâches telles que la traduction d'une phrase d'une langue à une autre.

Bien que cela fonctionne parfois, des règles strictement définies ne fonctionnent que pour des tâches concrètes et bien définies dont le système a connaissance.

### années 1990

Les modèles de langage commencent à évoluer vers des modèles statistiques et les modèles de langage commencent à être analysés, mais les projets à plus grande échelle sont limités par la puissance de calcul.

### années 2000

Les progrès de l'apprentissage automatique augmentent la complexité des modèles de langage, et l'adoption généralisée d'Internet entraîne une augmentation considérable des données de formation disponibles.

### 2012

Les progrès des architectures d'apprentissage en profondeur et des ensembles de données plus importants ont conduit au développement de GPT (Generative Pre-trained Transformer).

### 2018

Google présente BERT (Représentations d'encodeurs bidirectionnels à partir de transformateurs), qui représente un grand pas en avant dans l'architecture et ouvre la voie à de futurs grands modèles de langage.

### 2020

OpenAI publie GPT-3, qui devient le plus grand modèle avec 175 milliards de paramètres et établit une nouvelle référence de performance pour les tâches liées au langage.

### 2022

ChatGPT est lancé, ce qui transforme GPT-3 et des modèles similaires en un service largement accessible aux utilisateurs via une interface Web et déclenche une énorme augmentation de la sensibilisation du public aux LLM et à l'IA générative.

### 2023

Les LLM open source commencent à montrer des résultats de plus en plus impressionnants avec des versions telles que Dolly 2.0, LLaMA, Alpaca et Vicuna. GPT-4 est également publié, établissant une nouvelle référence pour la taille des paramètres et les performances.

SECTION 2

# Comprendre les grands modèles de langage

## Que sont les modèles de langage et comment fonctionnent-ils ?

Les grands modèles de langage sont des systèmes d'intelligence artificielle avancés qui acceptent certaines entrées et génèrent un texte de type humain en réponse. Ils travaillent d'abord en analysant de grandes quantités de données et en créant une structure interne qui modélise les ensembles de données en langage naturel sur lesquels ils sont formés. Une fois que cette structure interne a été développée, les modèles peuvent alors prendre des données sous forme de langage naturel et se rapprocher d'une bonne réponse.

## S'ils existent depuis tant d'années, pourquoi ne font-ils que maintenant la une des journaux ?

Quelques avancées récentes ont vraiment mis en lumière l'IA générative et les grands modèles de langage :

### PROGRÈS TECHNIQUES

Au cours des dernières années, des progrès significatifs ont été réalisés dans les techniques utilisées pour former ces modèles, ce qui a entraîné de grands progrès en termes de performances. Notamment, l'un des plus grands sauts de performance est venu de l'intégration de la rétroaction humaine directement dans le processus de formation.

### ACCESSIBILITÉ ACCRUE

La sortie de ChatGPT a ouvert la porte à toute personne disposant d'un accès Internet pour interagir avec l'un des LLM les plus avancés via une interface Web simple. Cela a mis en lumière les progrès impressionnants des LLM, car auparavant, ces LLM plus puissants n'étaient disponibles que pour les chercheurs disposant de grandes quantités de ressources et de connaissances techniques très approfondies.

### PUISSANCE DE CALCUL CROISSANTE

La disponibilité de ressources informatiques plus puissantes, telles que les unités de traitement graphique (GPU), et de meilleures techniques de traitement des données ont permis aux chercheurs de former des modèles beaucoup plus grands, améliorant ainsi les performances de ces modèles de langage.

### DONNÉES D'ENTRAÎNEMENT AMÉLIORÉES

Au fur et à mesure que nous nous améliorons dans la collecte et l'analyse de grandes quantités de données, les performances du modèle se sont considérablement améliorées. En fait, Databricks a montré que vous pouvez obtenir des résultats étonnants en formant un modèle relativement petit avec un ensemble de données de haute qualité avec Dolly 2.0 (et nous avons également publié l'ensemble de données avec l' ensemble de données databricks-dolly-15k).

## Alors, à quoi servent les organisations qui utilisent de grands modèles de langage ?

Voici quelques exemples de cas d'utilisation courants pour les grands modèles de langage :

### CHATBOTS ET ASSISTANTS VIRTUELS

L'une des implémentations les plus courantes, les LLM peuvent être utilisées par les organisations pour fournir de l'aide avec des choses comme le support client, le dépannage ou même avoir des conversations ouvertes avec les invites fournies par l'utilisateur.

### GÉNÉRATION DE CODE ET DÉBOGAGE

Les LLM peuvent être formés sur de grandes quantités d'exemples de code et donner des extraits de code utiles en réponse à une demande écrite en langage naturel.

Avec les techniques appropriées, les LLM peuvent également être construits de manière à référencer d'autres données pertinentes avec lesquelles ils n'ont peut-être pas été formés, comme la documentation d'une entreprise, pour aider à fournir des réponses plus précises.

### ANALYSE DES SENTIMENTS

Souvent difficile à quantifier, les LLM peuvent aider à prendre un texte et à mesurer les émotions et les opinions. Cela peut aider les organisations à recueillir les données et les commentaires nécessaires pour améliorer la satisfaction des clients.

### CLASSIFICATION ET REGROUPEMENT DE TEXTES

La capacité de catégoriser et de trier de grands volumes de données permet d'identifier des thèmes et tendances communs, soutenant une prise de décision éclairée et des stratégies plus ciblées.

### LA TRADUCTION DE LA LANGUE

Globalisez tout votre contenu sans heures de travail minutieux en alimentant simplement vos pages Web via les LLM appropriés et en les traduisant dans différentes langues. Au fur et à mesure que de plus en plus de LLM sont formés dans d'autres langues, la qualité et la disponibilité continueront de s'améliorer.

### RÉSUMÉ ET PARAPHRASES

Des appels ou des réunions de clients entiers pourraient être résumés efficacement afin que d'autres puissent digérer plus facilement le contenu. Les LLM peuvent prendre de grandes quantités de texte et le réduire aux octets les plus importants.

### GÉNÉRATION DE CONTENU

Commencez par une invite détaillée et demandez à un LLM de développer un plan pour vous. Continuez ensuite avec ces invites et les LLM peuvent générer un bon premier brouillon sur lequel vous pourrez vous baser. Utilisez-les pour réfléchir à des idées et posez les questions LLM pour vous aider à vous inspirer.

Remarque : la plupart des LLM ne sont pas formés pour être des machines factuelles. Ils savent utiliser le langage, mais ils ne savent peut-être pas qui a remporté le grand événement sportif l'année dernière. Il est toujours important de vérifier les faits et de comprendre les réponses avant de les utiliser comme référence.

SECTION 3

# Application de grands modèles de langage

Il existe quelques chemins que l'on peut emprunter lorsque l'on cherche à appliquer de grands modèles de langage pour leur cas d'utilisation donné. De manière générale, vous pouvez les diviser en deux catégories, mais il y a un certain croisement entre chacune. Nous aborderons brièvement les avantages et les inconvénients de chacun et les scénarios qui conviennent le mieux à chacun.

## Services propriétaires

En tant que premier service alimenté par LLM largement disponible, ChatGPT d'OpenAI a été la charge explosive qui a introduit les LLM dans le courant dominant. ChatGPT fournit une interface utilisateur (ou API) agréable où les utilisateurs peuvent envoyer des invites à l'un des nombreux modèles (GPT-3.5, GPT-4, etc.) et obtenir généralement une réponse rapide. Ce sont parmi les modèles les plus performants, formés sur d'énormes ensembles de données, et sont capables de tâches extrêmement complexes à la fois d'un point de vue technique, comme la génération de code, ainsi que d'un point de vue créatif, comme écrire de la poésie dans un style spécifique.

L'inconvénient de ces services est la quantité absolument énorme de calcul nécessaire non seulement pour les former (OpenAI a déclaré que GPT-4 leur a coûté plus de 100 millions de dollars à développer) mais aussi pour servir les réponses. Pour cette raison, ces modèles extrêmement volumineux seront probablement toujours sous le contrôle des organisations, et vous demandent d'envoyer vos données à leurs serveurs afin d'interagir avec leurs modèles de langage. Cela soulève des problèmes de confidentialité et de sécurité, et soumet également les utilisateurs à des modèles de "boîte noire", dont ils n'ont aucun contrôle sur la formation et les garde-corps. De plus, en raison du calcul requis, ces services ne sont pas gratuits au-delà d'une utilisation très limitée, de sorte que le coût devient un facteur dans leur application à grande échelle.

En résumé : les services propriétaires sont parfaits si vous avez des tâches très complexes, si vous êtes d'accord pour partager vos données avec un tiers et si vous êtes prêt à engager des frais si vous travaillez à une échelle significative.

## Modèles open source

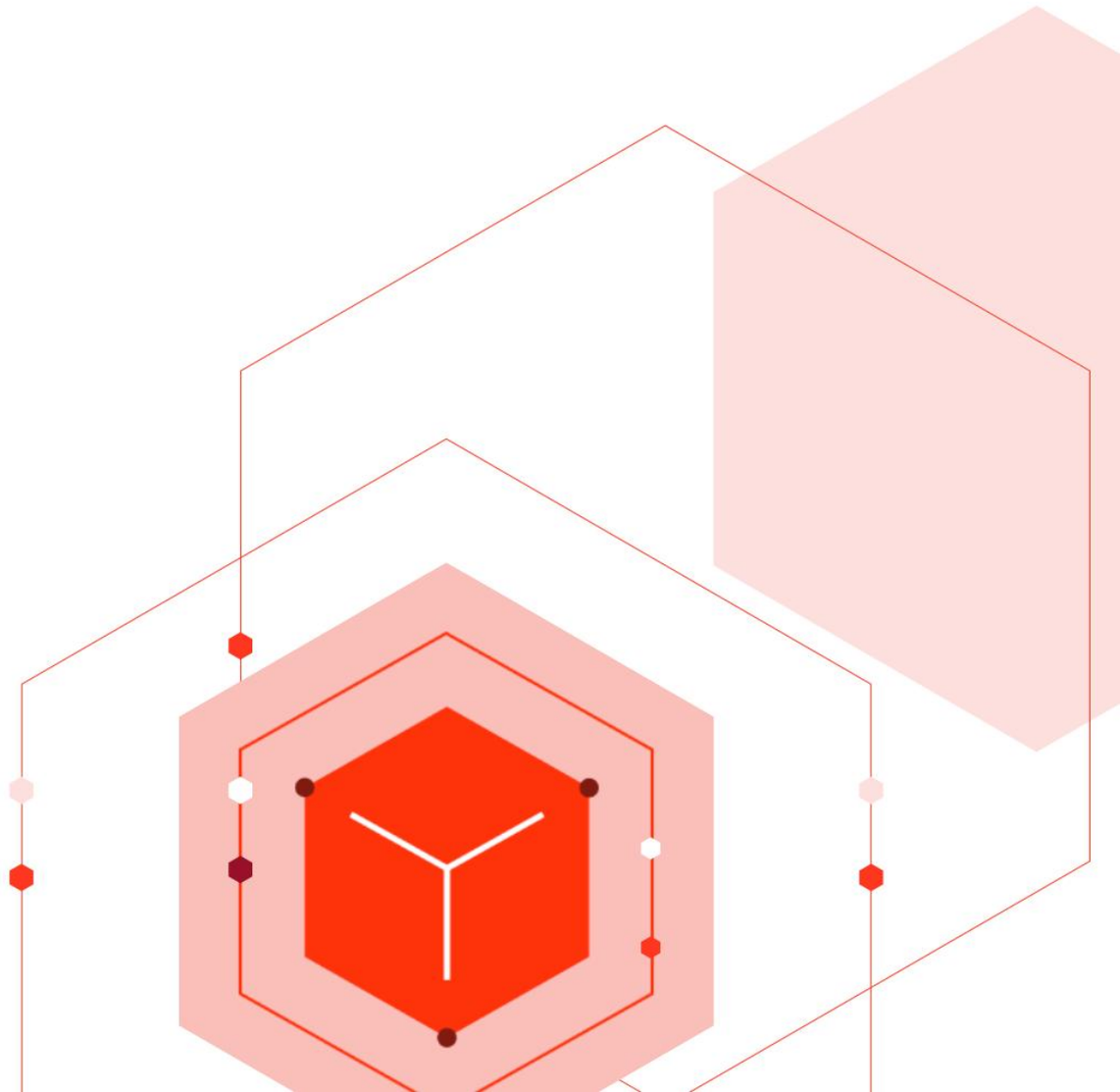
L'autre avenue pour les modèles de langage est d'aller vers la communauté open source, où il y a eu une croissance tout aussi explosive au cours des dernières années.
Des communautés comme Hugging Face rassemblez des centaines de milliers de modèles de contributeurs qui peuvent aider à résoudre des tonnes de cas d'utilisation spécifiques tels que la génération de texte, la synthèse et la classification. La communauté open source a rapidement rattrapé les performances des modèles propriétaires, mais n'a finalement toujours pas égalé les performances de quelque chose comme GPT-4.

Il faut actuellement un peu plus de travail pour saisir un modèle open source et commencer à l'utiliser, mais les progrès progressent très rapidement pour les rendre plus accessibles aux utilisateurs. Sur Databricks, par exemple, nous avons apporté des améliorations aux frameworks open source comme MLflow pour permettre à quelqu'un ayant un peu d'expérience Python de tirer très facilement n'importe quel modèle de transformateur Hugging Face et de l'utiliser comme objet Python. Souvent, vous pouvez trouver un modèle open source qui résout votre problème spécifique qui est d' un ordre de grandeur inférieur à ChatGPT, vous permettant d'intégrer le modèle dans votre environnement et de l'héberger vous-même. Cela signifie que vous pouvez garder les données sous votre contrôle pour des raisons de confidentialité et de gouvernance, ainsi que pour gérer vos coûts.

Un autre énorme avantage de l'utilisation de modèles open source est la possibilité de les ajuster à vos propres données. Puisque vous n'avez pas affaire à une boîte noire d'un service propriétaire, il existe des techniques qui vous permettent de prendre des modèles open source et de les former à vos données spécifiques, améliorant considérablement leurs performances sur votre domaine spécifique. Nous pensons que l'avenir des modèles linguistiques va évoluer dans cette direction, car de plus en plus d'organisations voudront un contrôle et une compréhension totale de leurs LLM.

## Conclusion et directives générales

En fin de compte, chaque organisation aura des défis uniques à surmonter, et il n'y a pas d'approche unique en ce qui concerne les LLM. Alors que le monde devient de plus en plus axé sur les données, tout, y compris les LLM, dépendra d'une base solide de données. Les LLM sont des outils incroyables, mais ils doivent être utilisés et mis en œuvre en plus de cette solide base de données. Databricks apporte à la fois cette base de données solide ainsi que les outils intégrés pour vous permettre d'utiliser et d'affiner les LLM dans votre domaine.



databricks

SECTION 4

# Alors, que dois-je faire ensuite si je veux commencer à utiliser les LLM ?

Cela dépend où vous en êtes dans votre voyage ! Heureusement, nous avons quelques pistes pour vous.

Si vous souhaitez approfondir un peu les LLM mais que vous n'êtes pas tout à fait prêt à le faire vous-même, vous pouvez regarder l'un des développeurs et conférenciers les plus talentueux de Databricks aborder ces concepts plus en détail lors de la conférence à la demande "Comment construire Votre propre grand modèle de langage comme Dolly.

Si vous êtes prêt à approfondir un peu et à approfondir votre formation et votre compréhension des fondements du LLM, nous vous recommandons de consulter notre cours sur les LLM. Vous apprendrez à développer des applications LLM prêtes pour la production et à vous plonger dans la théorie des modèles de base.

Si vos mains tremblent déjà d'excitation et que vous avez déjà une connaissance pratique de Python et de Databricks, nous vous fournirons d'excellents exemples avec des exemples de code qui peuvent vous permettre d'être immédiatement opérationnel avec les LLM !

Premiers pas avec le NLP à l'aide des pipelines de transformateurs Hugging Face

Ajustement fin des grands modèles de langage avec Visage étreignant et DeepSpeed

Présentation des fonctions d'IA : intégration de grandes Modèles de langage avec Databricks SQL

databricks

# À propos des Databrick

Databricks est la société de données et d'IA. Plus de 9 000 organisations du monde entier, dont Comcast, Condé Nast et plus de 50% des Fortune 500 - comptent sur le Databricks Lakehouse Plate-forme pour unifier leurs données, leurs analyses et leur IA. Databrick est dont le siège est à San Francisco, avec des bureaux dans le monde entier. Fondée par les créateurs originaux d'Apache Spark™, Delta Lake et MLflow, Databricks a pour mission d'aider les équipes de données à résoudre les problèmes les plus difficiles du monde. Pour en savoir plus, suivez Databricks sur Twitter, LinkedIn et Facebook.

COMMENCER VOTRE ESSAI GRATUIT

Contactez-nous pour une démo personnalisée : databricks.com/contact

databricks