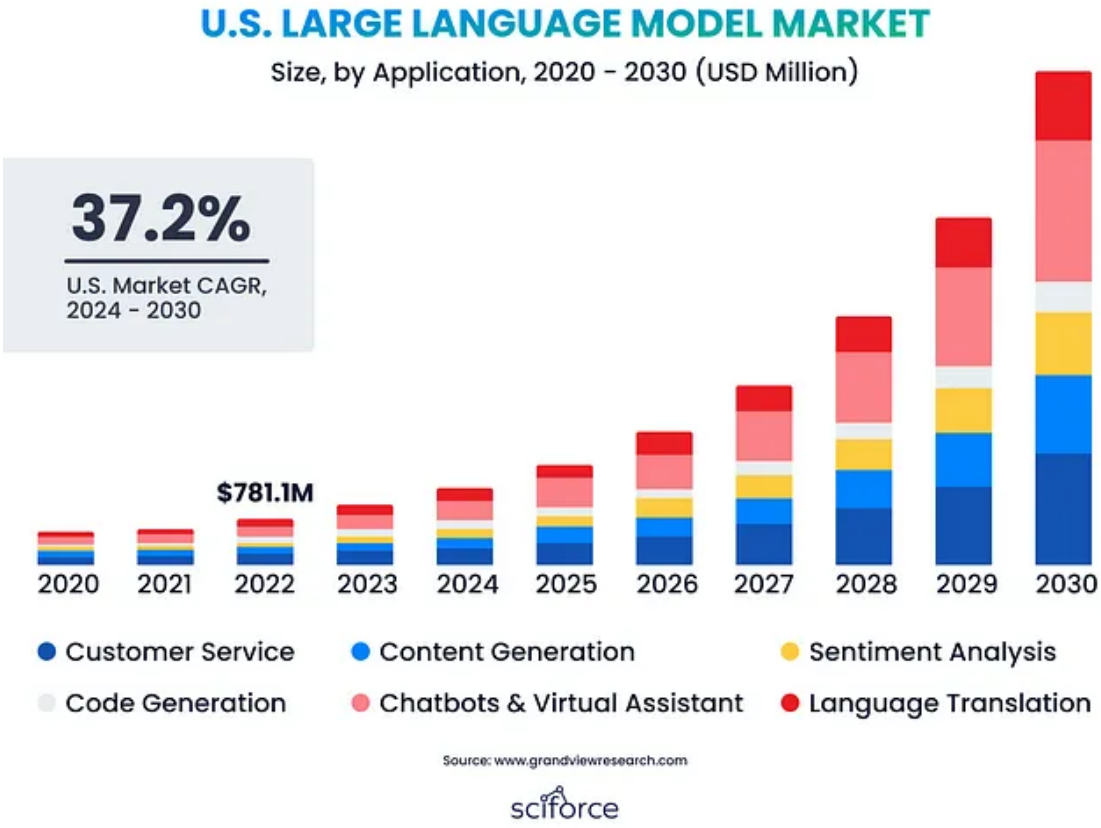
LLM

September 22, 2024

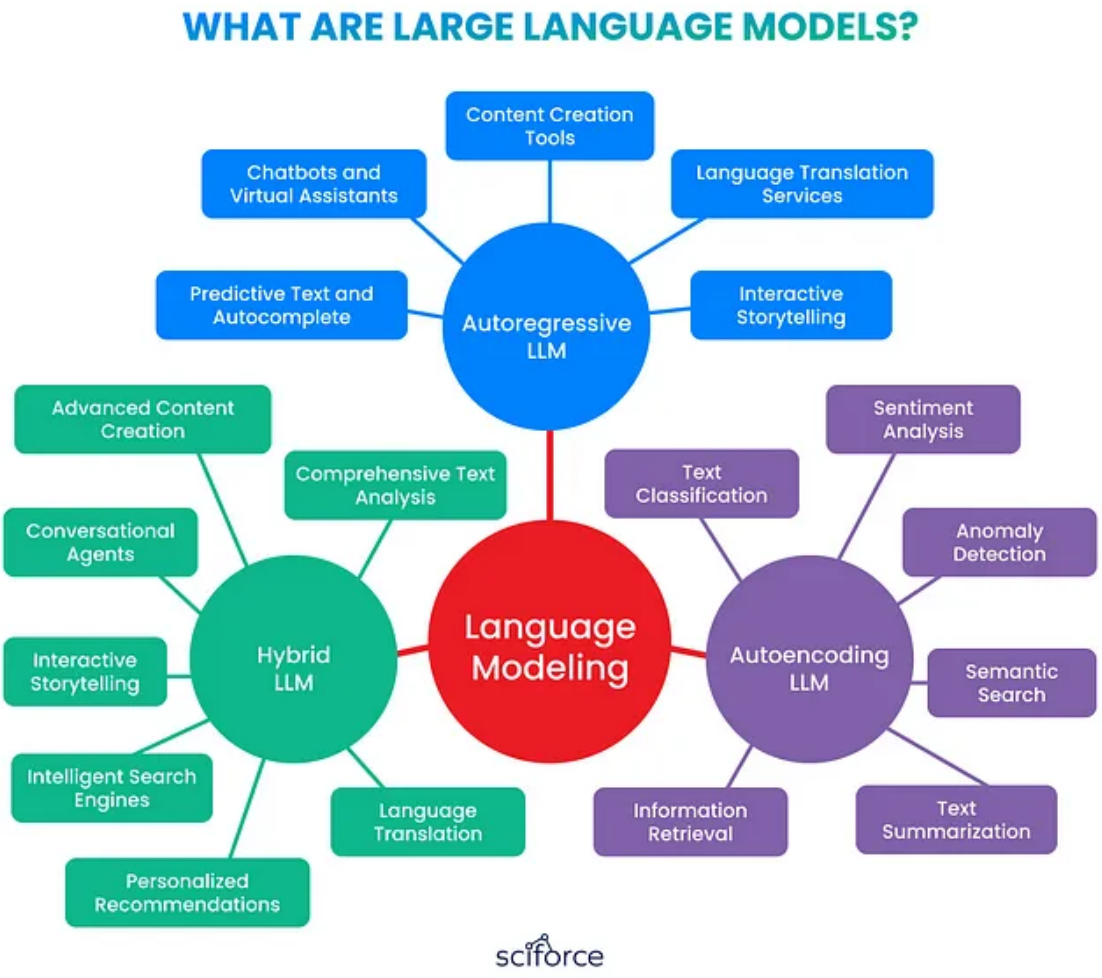
1 Guide étape par étape pour la création de votre propre modèle linguistique à grande échelle

Les grands modèles de langage (LLM) transforment l'IA en permettant aux ordinateurs de générer et de comprendre des textes semblables à ceux des humains, ce qui les rend essentiels dans divers secteurs. Le marché mondial des LLM est en pleine expansion et devrait passer de 1,59 milliard de dollars en 2023 à 259,8 milliards de dollars en 2030, sous l'effet de la demande de création automatisée de contenu, des progrès de l'IA et de la nécessité d'améliorer la communication entre l'homme et la machine.



Cette croissance est alimentée par la demande de création automatisée de contenu, les progrès de l'IA et du NLP, l'amélioration de la communication homme-machine et les grands ensembles de données. Les LLM privés gagnent en popularité car les entreprises recherchent le contrôle des données et la personnalisation. Ils fournissent des solutions sur mesure, réduisent la dépendance à l'égard des fournisseurs externes et renforcent la confidentialité des données. Ce guide vous aidera à créer votre propre LLM privé et vous offrira des informations précieuses, que vous soyez novice en matière de LLM ou que vous cherchiez à développer votre expertise.

2 Que sont les grands modèles linguistiques ?



Les grands modèles de langage (LLM) sont des systèmes d'IA avancés qui génèrent des textes semblables à ceux des humains en traitant de grandes quantités de données à l'aide de réseaux neuronaux complexes, tels que les transformateurs. Ils peuvent créer du contenu, traduire des langues, répondre à des questions et engager des conversations, ce qui les rend précieux dans divers secteurs, notamment le service à la clientèle et l'analyse de données. - **Les LLM autorégressifs** prédisent le mot suivant dans une phrase en fonction des mots précédents, ce qui les rend idéaux pour des tâches telles que la génération de texte. - **Les LLM à autoencodage** se concentrent sur l'encodage et la reconstruction de textes, excellant dans des tâches telles que l'analyse des sentiments et la recherche d'informations. - **Hybrid LLMs** combine the strengths of both approaches, offering versatile solutions for complex applications. Les LLM apprennent les règles linguistiques en traitant des quantités massives de textes provenant de diverses sources, de la même manière que la lecture de nombreux livres aide à comprendre la langue. Une fois formés, ils peuvent rédiger du contenu, répondre à des questions et participer à des conversations en s'appuyant sur leur apprentissage.

Par exemple, un LLM peut créer une histoire sur l'espace à partir des connaissances acquises en lisant des récits d'aventures spatiales ou expliquer la photosynthèse en se remémorant des informa-

tions tirées de textes de biologie.

3 Construire un LLM privé

3.1 Curation des données pour les LLM

Les LLM récents tels que Llama 3 et GPT-4 sont formés sur de vastes ensembles de données - Llama 3 avec 15 billions de jetons et GPT-4 avec 6,5 billions de jetons. Ces ensembles de données, provenant de divers contextes, notamment des médias sociaux (140 billions de jetons) et des données privées, s'étendent sur des centaines de téraoctets, voire plusieurs pétaoctets. Cette formation approfondie permet aux modèles de comprendre le langage en profondeur, en couvrant différents modèles, vocabulaires et contextes.

- **Données Web** :FineWeb (pas entièrement dédupliqué pour de meilleures performances, entièrement en anglais), Common Crawl (55% non-anglais) https://huggingface.co/datasets/HuggingFaceFW/fineweb
- **Code** : Code disponible publiquement à partir de toutes les principales plateformes d'hébergement de code. https://huggingface.co/datasets/bigcode/the-stack-v2
- **Textes académiques** : Archives d'Anna, Google Scholar, Google Patents https://annas-archive.org/
- **Books**: Google Books, Anna's Archive
- **Court Documents**: RECAP archive (USA), Open Legal Data (Germany): - https://www.courtlistener.com/recap/ - https://openlegaldata.io/

4 Data Preprocessing

Lors de la conservation des données pour les LLM, les étapes clés après le nettoyage et la structuration consistent à transformer les données dans un format dont le modèle peut tirer des enseignements, à l'aide de mécanismes de tokenisation, d'intégration et d'attention :

Character-Level Tokenization:
Each charecter is treated as a token

L a r g e l a n g u a g e
m o d e l s a r e p o w e
r f u l

Word-Level Tokenization:
Each word is treated as a token

Large language models are powerful

Subword-Level Tokenization:
Words are broken down into smaller meaningful subword units.

Lar ge lan gu age models s are
power ful

sciforce

Character-Level Tokenization:
Each charecter is treated as a token

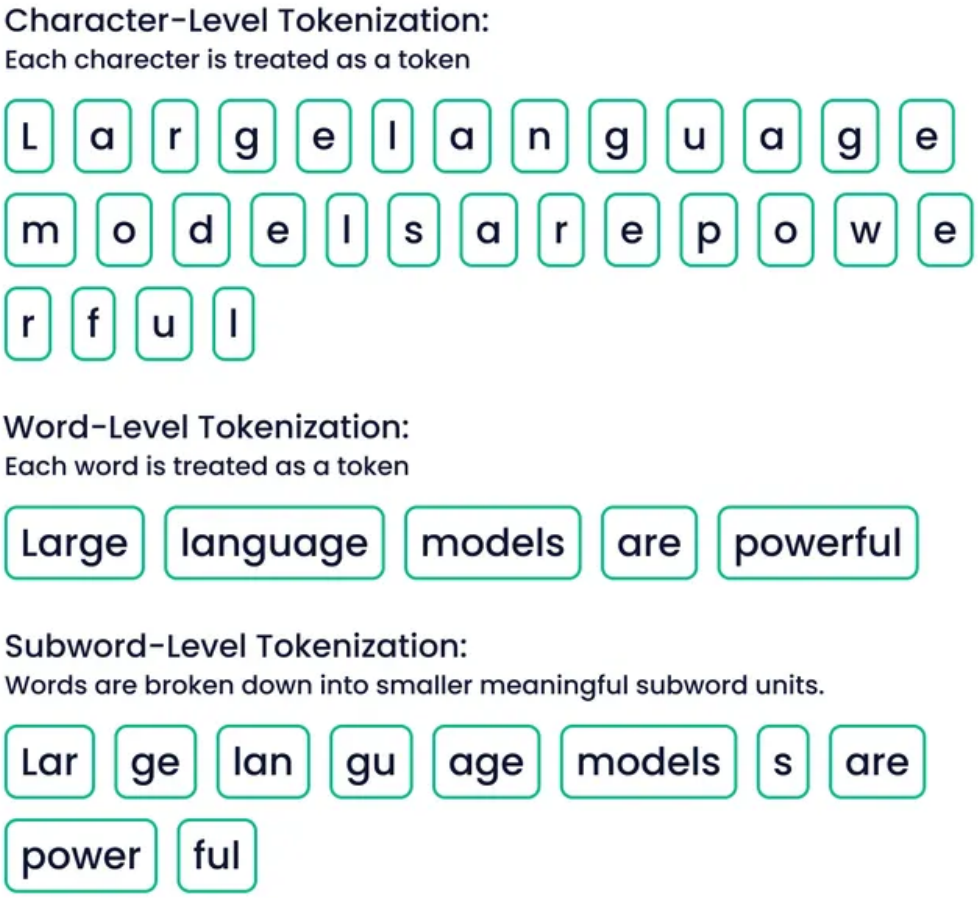
L a r g e l a n g u a g e
m o d e l s a r e p o w e
r f u l

Word-Level Tokenization:
Each word is treated as a token

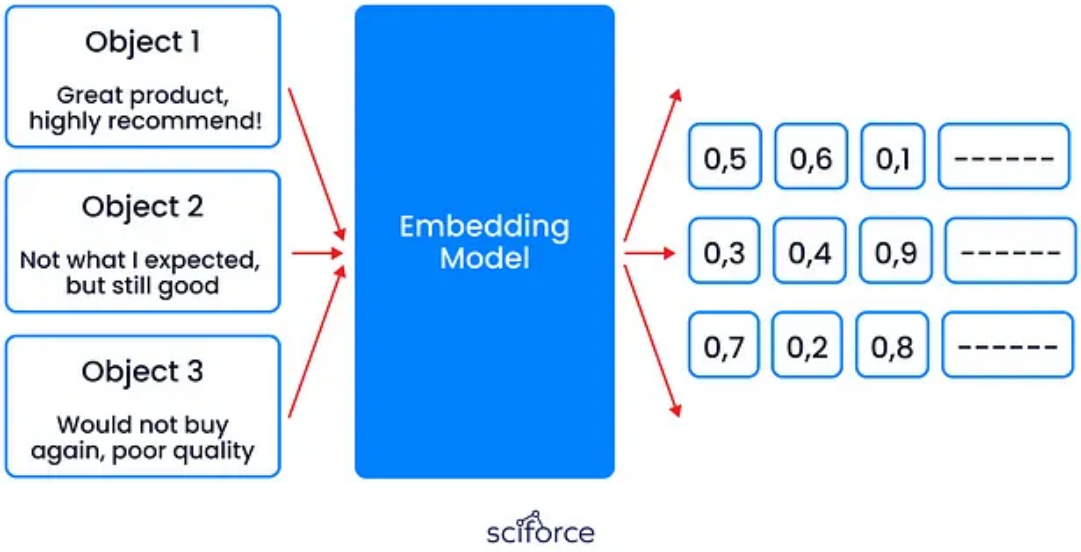
Large language models are powerful

Subword-Level Tokenization:
Words are broken down into smaller meaningful subword units.

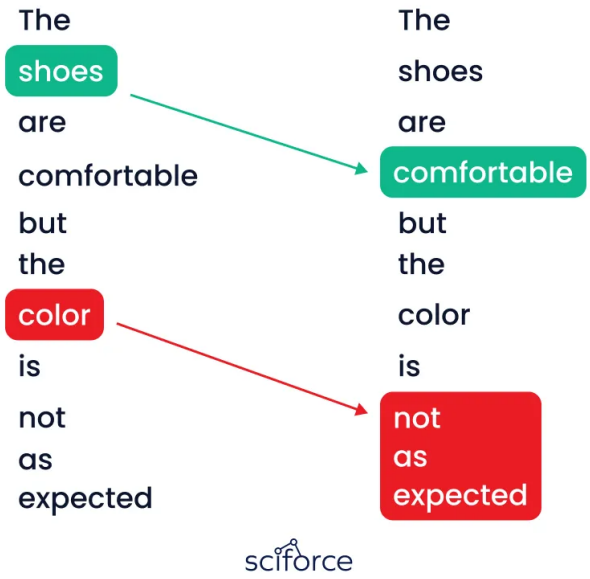
Lar ge lan gu age models s are



embedding L'intégration convertit les commentaires des clients en vecteurs numériques qui capturent le sentiment et la signification, ce qui aide le modèle à analyser les commentaires et à améliorer les recommandations.



Attention L'attention se concentre sur les parties les plus importantes d'une phrase, ce qui permet au modèle d'appréhender avec précision les sentiments clés, comme la distinction entre la qualité du produit et les problèmes de service.



5 Boucle de formation LLM

5.1 Data Input and Preparation

1. **Ingestion de données** : Collecte et chargement des données à partir de diverses sources. Nettoyage des données : Supprimer le bruit, traiter les données manquantes et expurger les informations sensibles.
2. **Normalisation** : Normaliser le texte, traiter les données catégorielles et assurer la cohérence des données.

3. **Découpage** : Diviser les textes volumineux en morceaux gérables tout en préservant le contexte.
4. **Tokenisation** : Convertir les morceaux de texte en jetons pour le traitement du modèle.
5. **Chargement des données** : Charger et mélanger efficacement les données pour une formation optimisée, en utilisant le chargement parallèle si nécessaire.

5.2 Loss Calculation

Calculer la perte: comparer les prédictions aux étiquettes réelles à l'aide d'une fonction de perte, en convertissant la différence en une valeur de "perte" ou d'"erreur". **Indicateur de performance** : Une perte élevée indique une mauvaise précision ; une perte plus faible indique un meilleur alignement avec les cibles réelles.

5.3 Hyperparameter Tuning

1. Learning Rate: Contrôle la taille de la mise à jour du poids pendant l'entraînement - une valeur trop élevée peut entraîner une instabilité ; une valeur trop faible ralentit l'entraînement..
2. Batch Size: Nombre d'échantillons par itération - des lots plus importants stabilisent l'apprentissage mais nécessitent plus de mémoire ; des lots plus petits introduisent de la variabilité mais sont moins gourmands en ressources.

6 Parallélisation et gestion des ressources

1. **Parallélisation des données** : Répartissez les ensembles de données sur plusieurs GPU pour un traitement plus rapide.
2. **Parallélisation des modèles** : Diviser le modèle entre les GPU pour traiter les modèles de grande taille.
3. **Contrôle du gradient** : Réduire l'utilisation de la mémoire pendant l'apprentissage en stockant sélectivement les résultats intermédiaires.

6.0.1 Iteration and Epochs

1. **Itérations** : Traiter des lots de données, en mettant à jour les poids à chaque fois.
2. **Époques** : Effectuer des passages complets à travers l'ensemble de données, en affinant les paramètres du modèle à chaque passage.

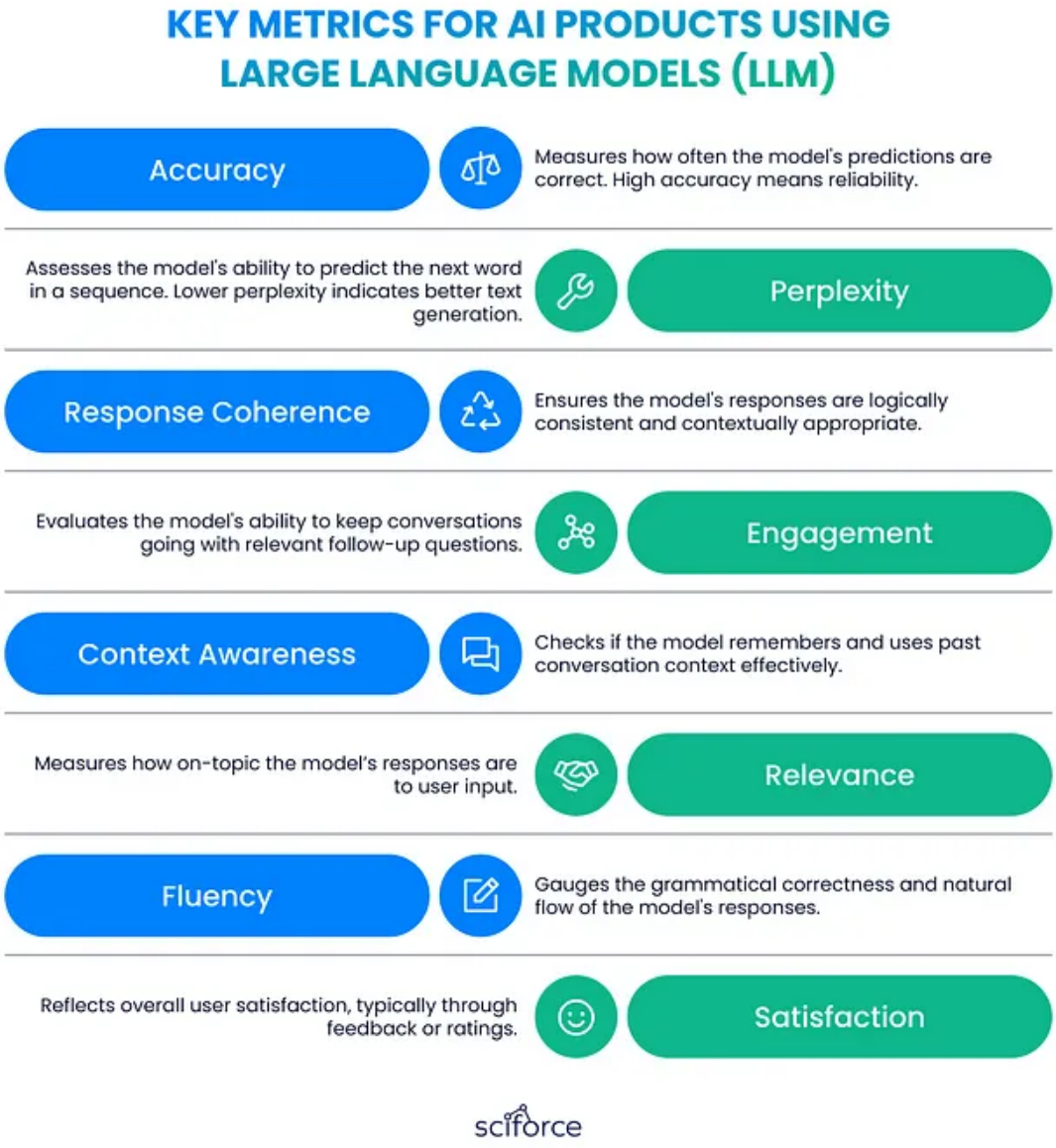
3. **Surveillance** : Suivre les mesures telles que la perte et la précision après chaque période afin de guider les ajustements et d'éviter le surajustement.

7 Évaluer votre LLM

Il est essentiel d'évaluer les performances d'un LLM après sa formation pour s'assurer qu'il répond aux normes requises. Les critères de référence couramment utilisés dans l'industrie sont les suivants : - **MMLU (Massive Multitask Language Understanding)** : Évalue la compréhension du langage naturel et le raisonnement dans un large éventail de sujets. - **MATH** : mesure le raisonnement mathématique du modèle en résolvant des problèmes à plusieurs étapes. - **HumanEval** : évalue les compétences en matière de codage en déterminant la capacité du modèle à générer un code précis et fonctionnel.

Pour ceux qui construisent des LLM à partir de zéro, des plateformes telles qu'Arena offrent des évaluations dynamiques, pilotées par les utilisateurs, qui peuvent ainsi comparer les modèles. Des entreprises comme OpenAI et Anthropic publient régulièrement des résultats de référence pour des modèles tels que GPT et Claude, mettant en évidence les progrès réalisés dans les capacités des LLM.

Pour ceux qui construisent des LLM à partir de zéro, des plateformes telles qu'Arena offrent des évaluations dynamiques, pilotées par les utilisateurs, qui peuvent ainsi comparer les modèles. Des entreprises comme OpenAI et Anthropic publient régulièrement des résultats de référence pour des modèles tels que GPT et Claude, mettant en évidence les progrès réalisés dans les capacités des LLM.



7.1 Conclusion

Construire un LLM privé est un processus difficile mais gratifiant qui offre une personnalisation, une sécurité des données et des performances inégalées. En rassemblant des données, en sélectionnant la bonne architecture et en affinant le modèle, vous pouvez créer un outil puissant adapté à vos besoins.

Ce guide présente les étapes clés du développement d'un LLM, vous aidant à construire un modèle qui excelle et s'adapte à l'évolution de la demande. Pour obtenir des conseils d'experts ou pour commencer votre parcours LLM, contactez-nous pour une consultation gratuite. Pour lire la version complète de l'article, visitez notre site web.

Nous aimerions également vous inviter à un webinaire gratuit sur la façon dont les MLD peuvent améliorer votre entreprise. Il aura lieu le 26 septembre à 17 heures (GMT+3). Vous aurez l'occasion

de poser des questions à notre expert, Volodymyr Sokhatskyi, sur le développement du LLM ou les opportunités pour votre industrie. Pour obtenir toutes les informations et vous inscrire, cliquez ici.

[ ]: