

SEARCH

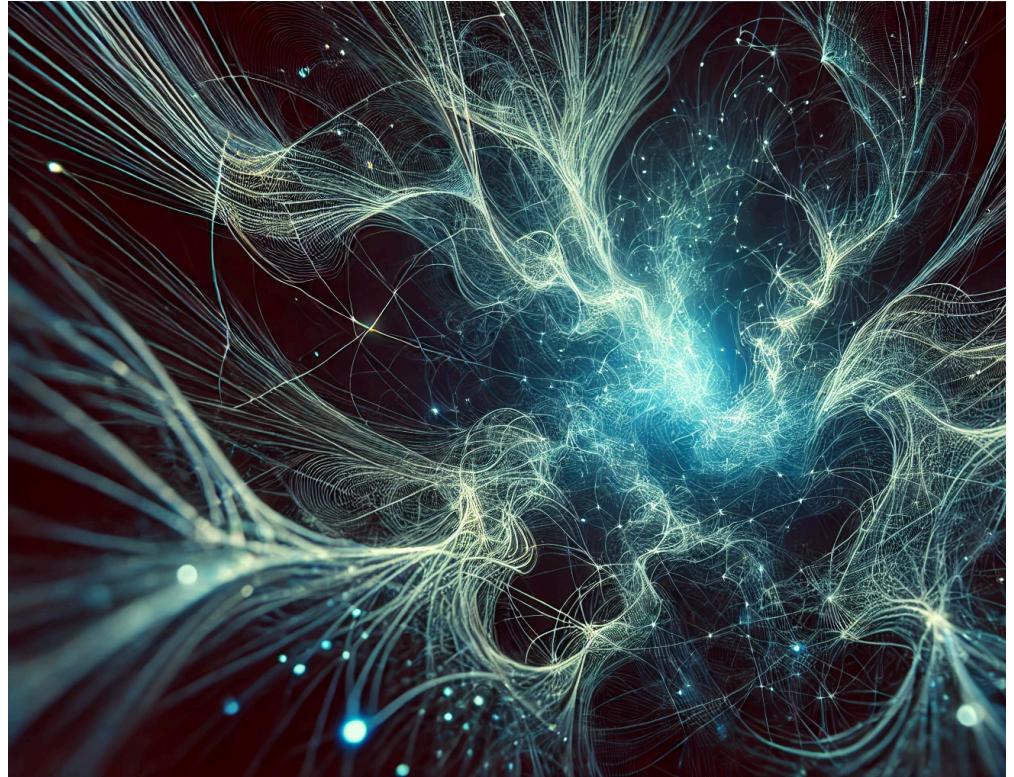
DES LEADERS D'OPINION

Benchmarks pour les LLM



Le kit de préparation mis à jour le 28 août 2024

By Irina Barskaya, PhD, responsable des données chez Yandex



Comprendre le rôle et les limites des benchmarks dans l'évaluation des performances de développement de LLM robustes.

Les modèles de langage volumineux ont gagné en popularité ces dernières années. L'exceptionnelle des LLM à comprendre les commandes du langage humain en a fait l'instrument essentiel pour les entreprises, prenant en charge les flux de travail critiques et automatisant les processus de manière maximale. De plus, au-delà de la compréhension de l'utilisateur moyen, les LLM peuvent mesurer que notre dépendance à leur égard augmente, nous devons vraiment accorder une attention particulière à garantir la précision et la fiabilité nécessaires. Il s'agit d'une tâche mondiale concernant toutes les industries, mais dans le domaine des entreprises, il existe désormais plusieurs points de référence utilisés pour évaluer les performances des LLM dans divers domaines. Ceux-ci peuvent être utilisés pour évaluer les performances d'un LLM en matière de compréhension, de construction logique, de mathématiques, etc. Cependant, il est important de se rappeler que si un LLM est prêt pour un déploiement commercial.

Dans cet article, j'ai rassemblé une liste complète des critères les plus populaires pour évaluer les performances des LLM. Nous discuterons de chaque référence en détail et verrons comment les différents LLM se mesurent contre ces critères d'évaluation. Mais d'abord, comprenons plus en détail l'évaluation LLM.

Qu'est-ce que l'évaluation LLM ?

Comme d'autres modèles d'IA, les LLM doivent également être évalués par rapport à spécifiques qui évaluent divers aspects des performances du modèle de langage : cohérence et cohérence. La norme implique généralement :

1. **Comprendre les requêtes des utilisateurs** : Évaluer la capacité du modèle à préciser un large éventail d'entrées utilisateur.
2. **Vérification de sortie** : Vérifier les réponses générées par l'IA par rapport à une norme fiable pour garantir qu'elles sont correctes et pertinentes.
3. **Robustesse** : Mesurer les performances du modèle avec des entrées ambiguës ou inattendues.

L'évaluation LLM donne aux développeurs le pouvoir d'identifier et de résoudre efficacement les défauts et les erreurs dans leur modèle. Cela leur permet de développer des applications plus sûres et plus fiables. Si un LLM est évalué de manière suffisamment précise et robuste pour gérer différentes applications du monde réel, y compris celles nécessitant des entrées ambiguës ou inattendues.

Repères

Les LLM sont l'un des éléments technologiques les plus complexes à ce jour et peuvent être utilisés dans de nombreuses applications les plus délicates. Le processus d'évaluation doit donc tout simplement être rigoureux, mettant à l'épreuve son processus de réflexion et sa précision technique.

Un benchmark utilise des ensembles de données, des mesures et des tâches d'évaluation pour évaluer les performances des LLM et permet de comparer différents LLM et de mesurer leur évolution au fil du temps. Cela favorise les progrès dans l'industrie grâce à des performances améliorées.

Voici quelques-uns des aspects les plus typiques de la performance LLM :

- **Connaissances**: Les connaissances du modèle doivent être testées dans différentes domaines pour voir si elles sont correctes et pertinentes. Il évalue l'efficacité avec laquelle le modèle utilise ces informations provenant de différents domaines, comme la physique, la programmation et la logique.
- **logique Raisonnement**: Il s'agit de tester la capacité d'un modèle à « penser » logiquement pour arriver à une conclusion logique. Ils impliquent généralement des scénarios dans lesquels le modèle doit trouver la réponse la plus plausible sur la base des connaissances quotidiennes et logiques.
- **Compréhension écrite**: Les modèles doivent être excellents dans l'interprétation de textes pour générer des réponses en conséquence. Le test consiste à répondre à des questions sur des passages de texte pour évaluer la compréhension, l'inférence et la rétention des détails appris à l'école.
- **Compréhension du code** : Ces tests sont nécessaires pour mesurer la capacité du modèle à comprendre et à déboguer du code. Ces tests donnent au modèle des tâches de codage et de débogage pour qu'il puisse résoudre avec précision, couvrant souvent une gamme de langages et de paradigmes de programmation.
- **Connaissance du monde**: Évaluer la compréhension du modèle des connaissances générales. Ces ensembles de données comportent généralement des questions auxquelles le modèle doit répondre correctement à des connaissances larges et encyclopédiques, ce qui les distingue des connaissances plus spécifiques et spécialisées.

Benchmarks « Connaissances »

MMLU (Compréhension du langage multimodal)

Ce benchmark est conçu pour tester la compréhension des connaissances factuelles que les sciences humaines, les sciences sociales, l'histoire, l'informatique et même le XNUMX tâches visant toutes à garantir que le modèle possède de grandes capacités MMLU un bon outil pour évaluer les connaissances factuelles et le raisonnement d'un

Récemment, il est devenu une référence clé pour l'évaluation des LLM dans les domaines. Les développeurs souhaitent toujours optimiser leurs modèles pour surpasser les autres en fait une norme de facto pour évaluer le raisonnement et les connaissances avancées. Ces modèles de niveau entreprise ont montré des scores impressionnantes sur ce benchmark, avec le Claude 3 Opus à 88.7 %, le Claude 3 à 86.8 %, le Gemini 1.5 Pro à 85.9 % et le Llama-3 70B qui sont généralement pas aussi performants sur ce benchmark, ne dépassant généralement pas les performances récentes du Phi-3-Small-7b à 75.3 %.

Cependant, MMLU n'est pas sans inconvénients : il a connu des problèmes tels que des réponses incorrectes, un contexte manquant. Et beaucoup pensent que certaines de ces erreurs sont nécessaires pour une évaluation LLM appropriée.

Je tiens à préciser que les benchmarks comme MMLU ne décrivent pas parfaitement tous les aspects de l'intelligence artificielle. Si un LLM obtient un excellent score dans ce domaine, cela ne signifie pas toujours qu'il comprend bien la matière. Les benchmarks ont une portée vraiment limitée et reposent souvent sur des scénarios simplifiés qui ne parviennent jamais à saisir pleinement la complexité et le contexte des interactions. Une véritable compréhension nécessite de connaître les faits et d'appliquer ces connaissances de manière critique, une résolution de problèmes et une compréhension des raisons. Les LLM doivent constamment être affinés et mis à jour afin que le modèle corresponde à l'efficacité du benchmark.

GPQA (référence de questions et réponses à l'épreuve de Google pour les connaissances)

Ce benchmark évalue les LLM sur le raisonnement logique à l'aide d'un jeu de données de questions et de réponses. Des experts du domaine l'ont développé et il couvre des sujets en biologie, en physique et en mathématiques.

Chaque question passe par le processus de validation suivant :

1. Un expert dans le même sujet répond à la question et fournit des commentaires.
2. Le rédacteur de la question révise la question en fonction de ces commentaires.
3. Un deuxième expert répond à la question révisée.

Ce processus peut en fait garantir que les questions sont objectives, précises et stimulantes pour l'apprentissage linguistique. Même les doctorants expérimentés n'obtiennent qu'une précision de 65 %. GPT-4-omni n'atteint que 53.6 %, soulignant l'écart entre l'intelligence humaine et l'intelligence artificielle.

En raison des exigences élevées en matière de qualification, l'ensemble de données GPQA limite quelque peu sa puissance statistique pour comparer l'exactitude et nécessite de nombreux experts. Les experts qui ont créé et validé ces questions venaient d'Upwork, ils ont donc potentinellement des背景 information basées sur leur expertise et les sujets abordés.

Références de code

HumanEval

164 problèmes de programmation, un véritable test pour les capacités de codage des LLM. Pour tester les capacités de codage de base des grands modèles de langage (LLM). Il s'agit d'un moyen pour juger de l'exactitude fonctionnelle du code généré, qui génère la probabilité qu'a un LLM d'achever correctement des échantillons de code générés par LLM réussisse les cas de test.

Bien que l'ensemble de données HumanEval comprenne des signatures de fonctions code et plusieurs tests unitaires, il n'inclut pas la gamme complète des problèmes de testeront tout simplement pas de manière adéquate la capacité d'un modèle à créer t scénarios.

MBPP (programmation Python principalement basique)

Mbpp Le benchmark se compose de 1,000 XNUMX questions de programmation Pytl Ce sont des problèmes de niveau d'entrée et ils se concentrent sur les compétences programmation. Il utilise quelques approches de réglage précis pour évaluer les perf modèles plus grands étant généralement plus performants sur cet ensemble de donn que l'ensemble de données contient principalement des programmes d'entrée de gan pas pleinement les complexités et les défis des applications du monde réel.

Références mathématiques

Bien que la plupart des LLM soient très doués pour structurer des réponses standard mathématique constitue pour eux un problème bien plus important. Pourquoi? Parce compétences liées à la compréhension des questions, une approche logique étape p mathématique et la détermination de la bonne réponse.

La méthode « Chaîne de pensée » (CoT) est conçue pour évaluer les LLM sur des cr Elle consiste à inciter les modèles à expliquer leur processus de raisonnement étape d'un problème. Cela présente plusieurs avantages. Cela rend le processus de raison identifier les failles dans la logique du modèle et permet une évaluation plus granulaire résolution de problèmes. En décomposant des problèmes complexes en une série d'améliorer les performances du modèle sur les tests mathématiques et fournir des info ses capacités de raisonnement.

GSM8K : une référence mathématique populaire

L'ensemble de données GSM8K est l'un des points de référence bien connus pour év mathématiques dans les LLM. GSM8K se compose de 8.5 4 problèmes de mathémat qui nécessitent quelques étapes à résoudre, et les solutions impliquent principalemen calculs élémentaires. En règle générale, les modèles plus grands ou ceux spécifier mathématique ont tendance à obtenir de meilleurs résultats sur cette référence, par e affichent un score de 7 %, tandis que DeepSeekMATH-RL-88.2B est légèrement en r

Bien que le GSM8K soit utile pour évaluer la capacité d'un modèle à résoudre des pr niveau de l'école primaire, il peut ne pas refléter pleinement la capacité d'un modèle à mathématiques plus avancés ou plus diversifiés, limitant ainsi son efficacité en tant q capacités mathématiques.

L'ensemble de données mathématiques : une alternative complète

L'ensemble de données mathématiques traitait des lacunes de références telles que l' données est plus étendu, couvrant l'arithmétique élémentaire jusqu'aux problèmes de collégial. Il est également comparé aux humains, avec un docteurant en informatique c mathématiques atteignant une précision de 40 % et un médaillé d'or atteignant une pi

Il fournit une évaluation plus complète des capacités mathématiques d'un LLM. Il se c modèles maîtrise l'arithmétique de base et est compétent dans des domaines complex géométrie et le calcul. Mais la complexité et la diversité accrues des problèmes peuvent modèles qui souhaitent atteindre une grande précision, en particulier ceux qui ne son

large éventail de concepts mathématiques. En outre, les formats de problèmes variés mathématiques peuvent introduire des incohérences dans les performances du modèle difficile la conclusion définitive sur la compétence mathématique globale d'un modèle.

L'utilisation de la méthode Chain of Thought avec l'ensemble de données Math peut révéler les capacités de raisonnement étape par étape des LLM sur un large éventail d'approche combinée comme celle-ci garantit une évaluation plus robuste et plus détaillée des mathématiques d'un LLM.

Repères de compréhension écrite

Une évaluation de la compréhension de lecture évalue la capacité du modèle à comprendre un texte complexe, ce qui est particulièrement fondamental pour des applications telles que le contenu et la recherche d'informations. Il existe quelques critères de référence pour cette compétence, chacun avec des attributs uniques qui contribuent à une évaluation complète du modèle.

RACE (ensemble de données de compréhension en lecture provenant de divers domaines)

Les benchmarks RACE comptent près de 28,000 à 100,000 passages et 12 à 18 questions d'anglais pour les élèves chinois des collèges et lycées âgés de XNUMX à XNUMX. Les questions et réponses à extraire des passages donnés, ce qui rend les tâches même assez difficiles.

Il couvre un large éventail de sujets et de types de questions, ce qui permet une évaluation des questions de différents niveaux de difficulté. De plus, les questions de RACE sont conçues pour tester les compétences humaines en lecture et sont créées par des experts du domaine.

Cependant, le benchmark présente certains inconvénients. Puisqu'il est développé en chinois, il est susceptible d'introduire des préjugés culturels qui ne reflètent pas un véritable niveau de difficulté élevé de certaines questions. Cela peut entraîner des évaluations moins précises.

DROP (raisonnement discret sur les paragraphes)

Une autre approche importante est DROP (Discrete Reasoning Over Paragraphs), qui consiste à effectuer un raisonnement discret sur des paragraphes. Il contient 96,000 passages et évalue les capacités de raisonnement des LLM et les questions sont extraites de Wikipédia et d'articles de New York Times (NYT) et de Wikipedia en turc. Les questions DROP appellent souvent des modèles pour effectuer des opérations telles que l'addition, la soustraction et la comparaison basées sur des informations dispersées dans les paragraphes.

Les questions sont difficiles. Ils demandent aux LLM de localiser plusieurs nombres dans un passage et de les ajouter ou de les soustraire pour obtenir la réponse finale. Les grands modèles tels que GPT-4 atteignent environ 80 % et 85 %, tandis que les humains atteignent 96 % sur l'ensemble de données DROP.

Repères de bon sens

Tester le bon sens dans les modèles de langage est une tâche intéressante mais également délicate, car elle évalue la capacité d'un modèle à porter des jugements et des inférences qui s'alignent avec le bon sens humain. Contrairement à nous, qui développons un modèle mondial complet à travers le contexte, les modèles linguistiques sont formés sur d'énormes ensembles de données sans tenir compte intrinsèquement le contexte. Cela signifie que les modèles ont du mal à réaliser des tâches nécessitant une compréhension intuitive des situations quotidiennes, un raisonnement logique et des connaissances générales qui sont très importantes pour des applications d'IA robustes et fiables.

HellaSwag (fins plus difficiles, contextes plus longs et activités à faible situations avec des générations adverses)

Hellaswag est développé par Rowan Zellers et ses collègues de l'Université de Wash Artificial Intelligence. Il est conçu pour tester la capacité d'un modèle à prédire la suite donnée. Ce benchmark est construit à l'aide du filtrage contradictoire (AF), où une série sélectionnent de manière itérative les mauvaises réponses contradictoires générées jusqu'à ce qu'il crée un ensemble de données avec des exemples triviaux pour les humains mais difficilement prédictibles, entraînant une zone de difficulté « Boucle d'or ».

Alors que Hellaswag représentait un défi pour les modèles précédents, les modèles contemporains ont atteint des niveaux de performances proches de la précision humaine, ce qui indique l'avancée dans ce domaine. Cependant, ces résultats suggèrent la nécessité de normes en constante évolution pour suivre le rythme des progrès des capacités de l'IA.

Livre ouvert

L'ensemble de données Openbook comprend 5957 questions à choix multiples en sciences naturelles. Les questions sont recueillies à partir d'exams à livre ouvert et développées pour évaluer la compréhension humaine du sujet.

Le benchmark Openbook nécessite une capacité de raisonnement au-delà de la recherche de faits simples. Il atteint actuellement la précision la plus élevée de 95.9 %.

OpenbookQA est calqué sur les exams à livre ouvert et comprend 5,957 1,326 questions à choix multiples de niveau élémentaire. Ces questions sont conçues pour sonder la compréhension des faits scientifiques fondamentaux et leur application à des situations nouvelles.

Semblable à Hellaswag, les modèles antérieurs trouvaient OpenbookQA difficile, mais comme GPT-4 ont atteint des niveaux de performances proches de ceux des humains. Cela souligne l'importance de développer des références encore plus complexes et nuancées pour dépasser les limites de la compréhension de l'IA.

Les critères de référence sont-ils suffisants pour l'évaluation des performances LLM ?

Certes, bien qu'ils fournissent une approche standardisée pour évaluer les performances, les benchmarks peuvent également être trompeurs. La Large Model Systems Organization affirme qu'un bon benchmark doit être évolutif, capable d'évaluer de nouveaux modèles avec un nombre relativement faible de classements uniques pour tous les modèles. Mais il existe des raisons pour lesquelles ces critères peuvent ne pas être suffisants. En voici quelques-unes :

Fuite de référence

Il s'agit d'une situation courante, qui se produit lorsque les données d'entraînement sont utilisées pour évaluer un modèle sur des données de test, ce qui donne lieu à une évaluation trompeuse. Si un modèle a déjà été entraîné sur les mêmes données de test au cours de la formation, son résultat peut ne pas refléter avec précision ses vraies capacités. Un bon benchmark idéal devrait minimiser la mémorisation et refléter des scénarios du monde réel.

Biais d'évaluation

Les classements de référence LLM sont utilisés pour comparer les performances des modèles. Cependant, il peut être difficile de s'appuyer sur ces classements pour comparer les performances entre deux modèles, car les résultats peuvent changer avec les modifications dans les tests de référence, comme la modification de l'ordre des questions.

classement des modèles jusqu'à huit positions. En outre, les LLM peuvent fonctionner par méthodes de notation, soulignant l'importance de prendre en compte les biais d'évaluation.

Une fin ouverte

L'interaction LLM dans le monde réel implique la conception d'invites pour générer les résultats du LLM dépendant de l'efficacité des invites, et les tests de référence sont conditionnés par la connaissance du contexte des LLM. Bien que les benchmarks soient conçus pour le contexte d'un LLM, ils ne se traduisent pas toujours directement en performances réelles, obtenant un score de 100 % sur un ensemble de données de référence, tel que le LS, mais avec un niveau de précision dans les applications pratiques. Cela souligne l'importance de considérer les tâches du monde réel dans l'évaluation LLM.

Évaluation efficace pour des LLM robustes

Vous savez désormais que les benchmarks ne sont pas toujours la meilleure option car ils peuvent être généralisés à tous les problèmes. Mais il existe d'autres moyens.

Repères personnalisés

Ces tests sont parfaits pour tester des comportements et des fonctionnalités spécifiques à une tâche. Supposons que si le LLM est conçu pour les médecins, les tests collectés dans les milieux médicaux représenteront efficacement des scénarios du monde réel. Les tests de référence personnalisés peuvent se concentrer sur la compréhension du langage, les connaissances contextuelles uniques du domaine. En alignant les tests de référence sur des scénarios personnalisés, vous pouvez vous assurer que le LLM fonctionne bien en général et excelle dans la tâche destinée. Cela peut aider à identifier et à combler très tôt les lacunes ou les faiblesses.

Pipeline de détection des fuites de données

Si vous souhaitez que vos évaluations « montrent » leur intégrité, il est très important de faire attention aux fuites de données. Une fuite de données se produit lorsque les données apparaissent dans le corpus de pré-entraînement du modèle, ce qui entraîne des scores de performance artificiellement élevés. Pour éviter cela, les références doivent être croisées avec les données de pré-entraînement pour éviter toute information vue précédemment. Cela peut impliquer l'utilisation d'ensembles de données propriétaires ou nouvellement organisés qui sont séparés du pipeline de formation du modèle. Les mesures de performances que vous obtenez reflètent la capacité du modèle à bien fonctionner dans des situations où il n'a pas été entraîné.

Évaluation humaine

Les métriques automatisées à elles seules ne peuvent pas capturer l'ensemble des points d'intérêt. En particulier lorsqu'il s'agit d'aspects très nuancés et subjectifs de la compréhension et de l'interprétation. Ici, l'évaluation humaine donne une bien meilleure évaluation :

- **Embaucher des professionnels** qui peuvent fournir des évaluations détaillées et spécialisées dans des domaines spécifiques.
- **Crowdsourcing!** Des plateformes comme Amazon Mechanical Turk vous permettent de recruter et de payer peu de frais divers jugements humains.
- **Commentaires de la communauté:** L'utilisation de plateformes telles que l'arena LMSYS Chatbot Arena Hard, par exemple, est particulièrement efficace pour mettre en évidence les différences subtiles entre les meilleurs modèles grâce aux interactions directes avec les utilisateurs.

Conclusion

Sans évaluation et analyse comparative, nous n'aurions aucun moyen de savoir si la tâches du monde réel est aussi précise et applicable que nous le pensons. Mais, comme ce sont pas un moyen totalement infaillible de vérifier cela, ils peuvent entraîner des erreurs dans les invites, effectuent les tâches comme indiqué et génèrent des sortes déjà excellents mais pas idéaux. C'est là que les benchmarks spécifiques à une tâche comme l'évaluation humaine et la détection des fuites de benchmarks. En les utilisant, nous pouvons produire des LLM véritablement robustes.

C'est ainsi que cela devrait être dans un monde idéal. Les LLM comprennent les requêtes, détectent les erreurs dans les invites, effectuent les tâches comme indiqué et génèrent des sortes déjà excellents mais pas idéaux. C'est là que les benchmarks spécifiques à une tâche comme l'évaluation humaine et la détection des fuites de benchmarks. En les utilisant, nous pouvons produire des LLM véritablement robustes.



RUBRIQUES CONNEXES: #BENCHMARKS DE L'IA #BENCHMARKS DE PERFORMANCES DE L'IA #LLM

NE MANQUEZ PAS



Comprendre l'architecture Lakehouse de données sur site

SUIVANT

L'analyse des sentiments prédire les tendances de



Irina Barskaya, PhD, responsable des données chez Yandex

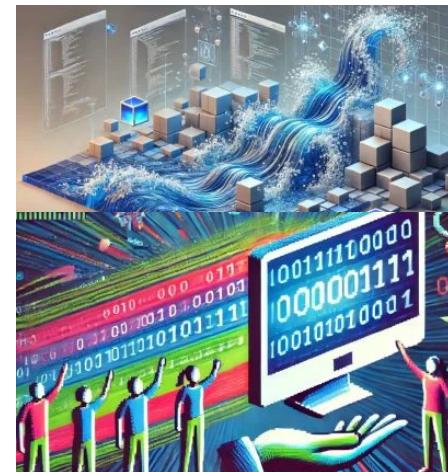


Irina Barskaya, PhD, est une éminente data scientist avec plus d'une décennie d'expérience, englobant l'analyse de technologies de pointe. Elle a dirigé la création et l'analyse de Yasmina, le premier assistant basé sur l'IA pour l'Arabie saoudite, gérant la localisation et l'étiquetage de données complexes pour les standards modernes. Actuellement, Irina dirige l'analyse de la qualité chez Yandex, favorisant les progrès dans l'IA et la machine learning.

TU PEUX AIMER



Le LLM Open Source le plus puissant à ce jour : Meta LAMA 3.1-405B



Optimisation du déploiement LLM : vLLM PagedAttention et l'avenir d'un service



Guide c synthét

[À propos de nous](#) [Notre équipe](#) [Notre Charte](#) [Noms de domaine .AI](#) [Outils de presse](#) [Contactez-Nous](#)

Annonceur Divulgation: Unite.AI s'engage à respecter des normes éditoriales rigoureuses pour fournir à nos lecteurs des informations et des actualités précises. Nous pouvons recevoir une compensation lorsque vous cliquez sur des liens vers des produits que nous avons examinés.

