Communiqués de presse

# IBM annonce la disponibilité du modèle d'IA Open-Source Mistral sur watsonx et élargit son choix de modèles pour aider les entreprises à mettre l'IA à l'échelle avec confiance et flexibilité

**— IBM propose une version optimisée de Mixtral-8x7B qui a montré un potentiel de réduction de la latence jusqu'à 75 %**

**— Il s'ajoute au catalogue croissant de modèles IBM, tiers et open-source pour offrir aux clients choix et flexibilité**

**— Dernier modèle open-source disponible sur la plateforme d'IA et de données watsonx avec un studio d'IA conçu pour les entreprises, un entrepôt de données et des capacités de gouvernance**


**ARMONK, N.Y., le 29 février 2024** : IBM (NYSE : IBM) a annoncé aujourd'hui la disponibilité du célèbre modèle de langage de grande taille (LLM) Mixtral-8x7B, développé par Mistral AI, sur sa plateforme d'IA et de données watsonx, tout en continuant à développer ses capacités pour aider les clients à innover avec les propres modèles de fondation d'IBM et ceux d'un éventail de fournisseurs open-source.

IBM propose une version optimisée de Mixtral-8x7b qui, lors de tests internes, a permis d'augmenter le débit - c'est-à-dire la quantité de données pouvant être traitées dans un laps de temps donné - de 50 % par rapport au modèle standard[1]. Cela pourrait potentiellement réduire le temps de latence de 35 à 75 %, en fonction de la taille du prompt, accélérant ainsi le temps nécessaire pour obtenir des informations. Ce résultat est obtenu grâce à un processus appelé quantization[2] qui réduit la taille du modèle et les besoins en mémoire pour les LLMs et, à son tour, peut accélérer le traitement pour aider à réduire les coûts et la consommation d'énergie.

L'ajout de Mixtral-8x7B élargit la stratégie ouverte et multi-modèle d'IBM pour répondre aux besoins des clients et leur donner le choix et la flexibilité pour mettre à l'échelle des solutions d'IA d'entreprise sur l'ensemble de leurs activités. Grâce à des décennies de recherche et développement dans le domaine de l'IA, à une collaboration ouverte avec Meta et Hugging Face, et à des partenariats avec des leaders en matière de modèles, IBM élargit son catalogue de modèles watsonx.ai et apporte de

nouvelles fonctionnalités, modalités et de nouveaux langages.

Les choix de modèles de fondation prêts pour les entreprises d'IBM et sa plateforme de données et d'IA watsonx peuvent permettent aux clients d'utiliser l'IA générative pour obtenir de nouvelles informations et de nouveaux gains d'efficacité, et créer de nouveaux modèles économiques fondés sur des principes de confiance. IBM permet aux clients de sélectionner le bon modèle pour le bon cas d'usage et les objectifs de prix et de performance pour des domaines d'activité ciblés tels que la finance.

Mixtral-8x7B a été conçu en combinant le *Sparse modeling* - une technique innovante qui ne trouve et n'utilise que les parties les plus essentielles des données pour créer des modèles plus efficaces - et la technique de *Mixture-of-Experts*, qui combine différents modèles ("experts") qui se spécialisent dans différentes parties d'un problème et les résolvent. Le modèle Mixtral-8x7B est largement reconnu pour sa capacité à traiter et à analyser rapidement de grandes quantités de données afin de fournir des informations adaptées au contexte.

*« Les clients demandent du choix et de la flexibilité pour déployer les modèles qui correspondent le mieux à leurs cas d'usage uniques et à leurs exigences métiers », a déclaré* **Xavier Vasques, Vice-Président et Directeur Technique chez IBM Technology et R&D en France**. *« En proposant Mixtral-8x7B et d'autres modèles sur watsonx, nous ne leur donnons pas seulement la possibilité de choisir la manière dont ils déploient l'IA, nous donnons à un écosystème de créateurs d'IA et de dirigeants d'entreprises les outils et les technologies nécessaires pour stimuler l'innovation dans divers secteurs et domaines. »*

Cette semaine, IBM a également annoncé la disponibilité d'ELYZA-japanese-Llama-2-7b, un modèle LLM japonais Open-Source d'ELYZA Corporation, sur watsonx. IBM propose également les modèles open-source Llama-2-13B-chat et Llama-2-70B-chat de Meta, ainsi que d'autres modèles de tiers sur watsonx, et d'autres encore seront disponibles au cours des prochains mois.

*Les déclarations d'IBM concernant ses orientations et intentions futures sont sujettes à modification ou retrait sans préavis et ne représentent que des buts et des objectifs.*

**Contacts Presse** :

**Weber Shandwick pour IBM**

**IBM**
Gaëlle Dussutour
Tél. : + 33 (0)6 74 98 26 92
dusga@fr.ibm.com

Louise Weber

Tél. : + 33 (0)6 89 59 12 54

ibmfrance@webershandwick.com

---

[1] Based on IBM testing over two days using internal workloads captured on an instance of watsonx for IBM use.

[2] La **quantization** est une méthode qui visent à**diminuer la précision numérique des nombres flottants utilisés comme paramètres des modèles,** qui sont des réseaux de neurones très profonds.