



Le 07 octobre 2025 - Le lancement de Granite 4.0 marque le début d'une nouvelle ère pour les modèles de langage d'IBM conçus pour les entreprises, qui tire parti des dernières avancées architecturales pour miser sur des petits modèles de langage efficaces offrant des performances compétitives à des coûts et une latence réduits. Les modèles Granite 4.0 ont été développés en mettant particulièrement l'accent sur les tâches essentielles aux processus agentiques, tant dans le cadre de déploiements autonomes que comme composants rentables dans des systèmes complexes, aux côtés de modèles de raisonnement plus importants.

Pourquoi Granite 4.0 ?

- **Rapidité & efficacité** : des vitesses d'inférence 2x plus rapides et une réduction de 70 % des besoins mémoire.
- **Flexibilité** : déploiement on-premise, Cloud ou edge, avec des tailles compactes adaptées à vos contraintes.
- **Ouverture & transparence** : open source sous licence **Apache 2.0**, avec une gouvernance et une sécurité de niveau entreprise.
- **Scalabilité** : contextes illimités, architecture hybride (Mamba-2 + Transformers), et options de personnalisation avancées.

Les cas d'usages :

- Chatbots & support client 24/7
- Résumé et analyse de documents complexes
- Knowledge management & RAG
- Workflows agentiques
- Applications edge AI

Cliquez [ici](#) pour lire l'intégralité du blog.

Contacts Presse :

IBM

Gaëlle Dussutour

Tél. : + 33 (0)6 74 98 26 92

dusga@fr.ibm.com

Weber Shandwick pour IBM

Louise Weber / Louise Oualid

Tél. : + 33 (0)6 89 59 12 54 / +33 (0)6 98 48 92 05

ibmfrance@webershandwick.com
