Modèles LLM open-source recommandés pour un chatbot en anglais

1. LLaMA 3 8B Instruct (Meta) : recommandé «1 »

(1) Coté serveur :

Llama 3 (8B et 70B) apporte une grande amélioration par rapport à Llama 2 : meilleurs résultats, moins de refus, réponses plus alignées et plus variées, ainsi que de meilleures capacités en raisonnement, code et suivi d'instructions.

- ✓ Très haute qualité en anglais (corpus massif, données récentes).
- ✓ Très performant pour les tâches conversationnelles.
- ✓ Capable de générer des réponses cohérentes, structurées et naturelles.
- ✓ Fonctionne très bien en quantization (GGUF) pour CPU ou petite VM.
- ✓ Large communauté + support dans Ollama, llama.cpp, HuggingFace.

Meta Llama 3 Instruct model performance

	Meta Llama 3 8B	Gemma 7B - It Measured	Mistral 7B Instruct Measured
MMLU 5-shot	**68.4**	53.3	58.4
GPQA 0-shot	**34.2**	21.4	26.3
HumanEval 0-shot	**62.2**	30.5	36.6
GSM-8K 8-shot, CoT	**79.6**	30.6	39.9
MATH 4-shot, CoT	**30.0**	12.2	11.0

	Meta Llama 3 70B	Gemini Pro 1.5 Published	Claude 3 Sonnet Published
MMLU 5-shot	**82.0**	81.9	79.0
GPQA 0-shot	39.5	**41.5** CoT	38.5 CoT
HumanEval 0-shot	**81.7**	71.9	73.0
GSM-8K 8-shot, CoT	**93.0**	91.7 11-shot	92.3 0-shot
MATH 4-shot, CoT	50.4	**58.5** Minerva prompt	40.5

Ce tableau compare les **performances des modèles LLaMA 3** (8B et 70B) avec d'autre modèles open-source et fermés (Gemma, Mistral, Gemini, Claude).

- **MMLU (Massive Multi-task Language Understanding)**

Teste la compréhension générale du langage

Plus le score est élevé → plus le modèle comprend bien le texte

- **GPQA (Graduate-Level Google Problem-Solving)**

Test de raisonnement scientifique difficile.

Indique si le modèle sait raisonner sur des questions complexes.

- **HumanEval**

Teste les capacités de programmation et raisonnement logique.

Mesure si le modèle peut comprendre et compléter du code.

- **GSM-8K**

Problèmes mathématiques simples

Teste la logique étape par étape ("chain of thought")

- **MATH**

Problèmes mathématiques plus complexes.

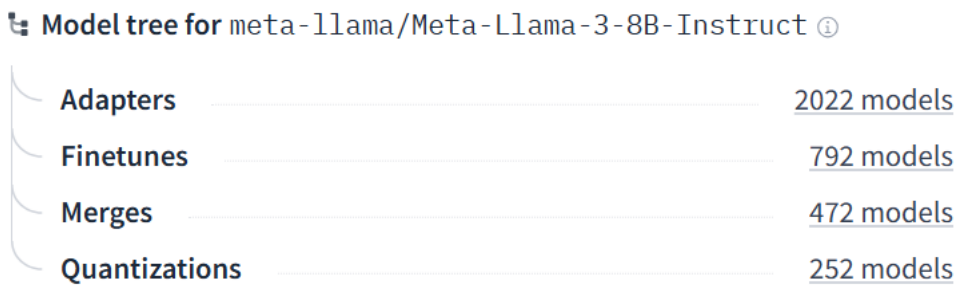
Mesure le raisonnement avancé.

	Time (GPU hours)	Power Consumption (W)	Carbon Emitted(tCO2eq)
Llama 3 8B	1.3M	700	390
Llama 3 70B	6.4M	700	1900
Total	7.7M		2290

Lors du développement de LLaMA 3, nous avons évalué les performances du modèle sur des benchmarks standards afin de mesurer sa puissance de calcul et sa précision dans des environnements serveurs. Ces tests visent à garantir des performances optimales sur des infrastructures dotées de ressources importantes (GPU, serveurs cloud), où le modèle peut exprimer pleinement ses capacités en raisonnement, compréhension et génération de texte.

(2) Coté client

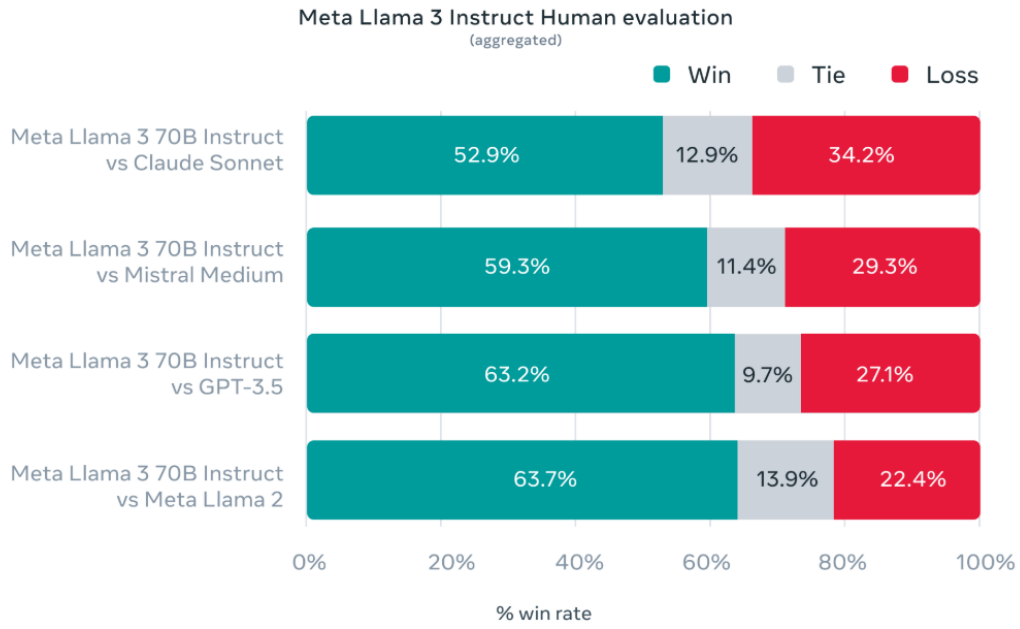
En parallèle, nous avons également cherché à optimiser ses performances dans des scénarios réels, notamment sur des environnements plus légers et limités en ressources. Pour cela, un ensemble d'évaluation humaine de haute qualité a été conçu, permettant de mesurer la réactivité, la cohérence et la qualité des réponses du modèle dans des conditions proches de l'utilisation côté client



Cet ensemble contient **1 800 prompts** couvrant **12 cas d'usage clés** :

– demande de conseil,
– génération d'idées (brainstorming),
– classification,
– questions fermées,
– programmation (coding),
– écriture créative,
– extraction d'informations,
– incarnation d'un personnage/persona,
– questions ouvertes,
– raisonnement,
– réécriture,
– résumé.

Le graphique ci-dessous présente les résultats agrégés de nos évaluations humaines sur ces catégories, comparés à Claude Sonnet, Mistral Medium et GPT-3.5.



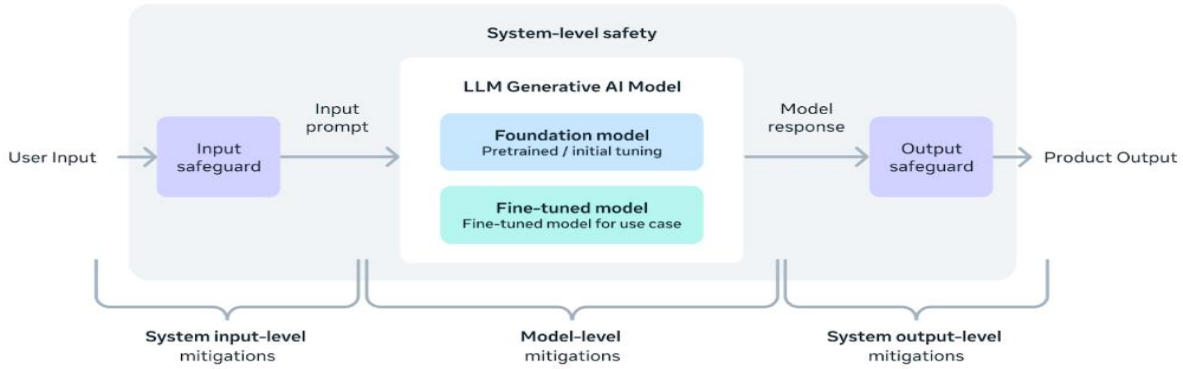
Les classements de préférence établis par des annotateurs humains à partir de cet ensemble d'évaluation mettent en évidence la forte performance de notre modèle 70B spécialisé dans le suivi d'instructions, comparé aux modèles concurrents de taille similaire dans des scénarios réels.

Le schéma ce dessous montre que Llama 3 est conçu pour être utile tout en garantissant un déploiement responsable. Meta adopte une approche système axée sur la sécurité, incluant :

-**instruction fine-tuning sécurisé**,

-**tests de red teaming** contre les risques (chimique, biologique, cybersécurité…),

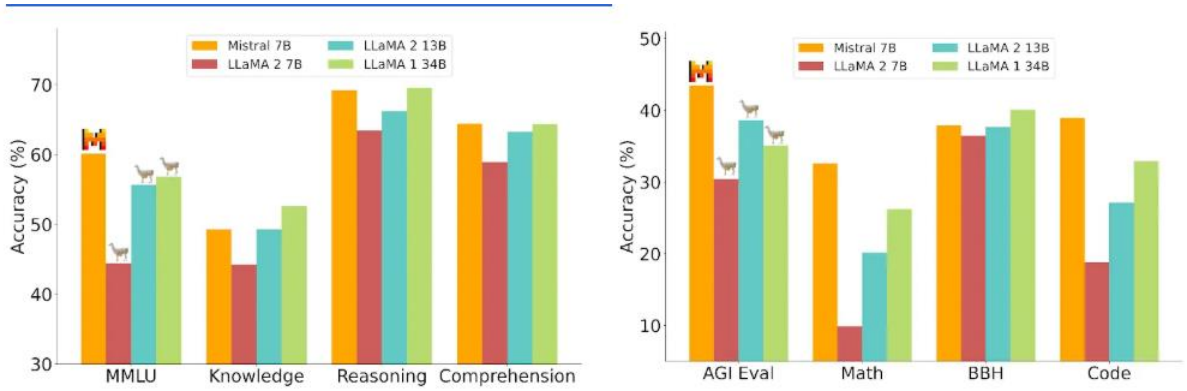
-outils spécialisés comme **Llama Guard 2**, **CyberSecEval 2**, et **Code Shield** pour filtrer les prompts, les réponses et le code non sécurisé.



2) Mistral 7B Instruct (Mistral AI) — Recommandé « 2 »

(1) Coté serveur :

Sur le plan serveur, Mistral 7B démontre des performances remarquables malgré sa taille compacte.
Il surpasse parfois des modèles plus volumineux sur divers benchmarks standards, tels que MMLU, AGI Eval, Math, BBH et Code, confirmant sa capacité à traiter des tâches complexes de raisonnement, programmation et compréhension linguistique.
Ces résultats mettent en évidence sa robustesse et son efficacité énergétique, en faisant un excellent choix pour des déploiements sur serveurs GPU ou infrastructures cloud, notamment lorsque le compromis entre coût et performance est essentiel.

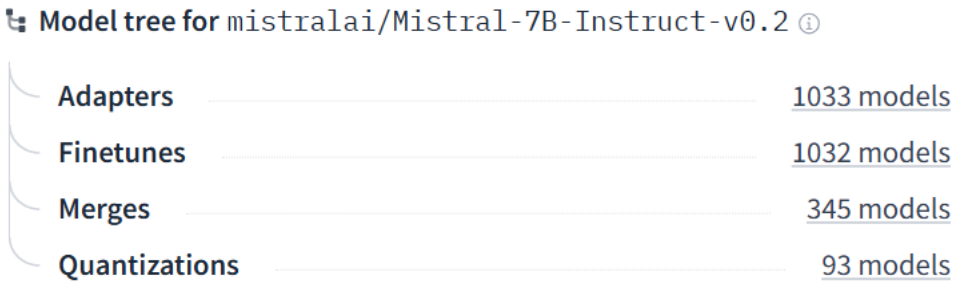


le déploiement de Mistral 7B Instruct nécessite une configuration matérielle adaptée afin de garantir des performances optimales en inférence et en fine-tuning.
Le tableau suivant présente les exigences minimales, recommandées et hautes performances selon les principaux composants matériels :

Catégorie	Configuration minimale	Configuration recommandée	Configuration haute performance
CPU	8 cœurs (Intel i7 / AMD Ryzen 7)	12 cœurs (Intel i9 / Ryzen 9)	16 cœurs+ (AMD Threadripper)
GPU	RTX 3060 (12 Go VRAM)	RTX 3090 (24 Go VRAM)	NVIDIA A100 (40 Go VRAM)
RAM	16 Go DDR4	32 Go DDR4	64 Go+ DDR5
Stockage	100 Go SSD (SATA)	500 Go SSD (NVMe)	1 To+ SSD NVMe
Système d'exploitation	Linux (Ubuntu 20.04+) ou Windows 10	Linux (Ubuntu 22.04+), Windows 11	Linux (Ubuntu 22.04+ / CentOS)

(2) Coté client :

En complément, la structure modulaire de Mistral 7B Instruct facilite son adaptation à une large variété d'environnements clients. Comme l'illustre la figure ci-dessous, le modèle dispose de plus de 1 000 versions dérivées, incluant des adapters, finetunes, merges et quantizations. Cette diversité témoigne de la forte adoptabilité de Mistral 7B par la communauté open source, permettant aux développeurs d'optimiser le modèle pour des besoins spécifiques ou de l'alléger pour une exécution locale sur CPU ou GPU modestes.
Les nombreuses versions quantisées (93 modèles recensés) illustrent tout particulièrement la capacité du modèle à fonctionner efficacement sur des machines à ressources limitées, rendant son déploiement client rapide et accessible.



- ✓ Très bon niveau d'anglais, bonne fluidité.
- ✓ Excellente cohérence dans les réponses longues.
- ✓ Très rapide, optimisé pour les ressources limitées.

- ✓ Très bon pour les use-cases Q&A, assistants et chatbots spécialisés.

3) Conclusion

L'analyse comparative des deux modèles LLaMA 3 8B Instruct (Meta) et Mistral 7B Instruct (Mistral AI) met en évidence deux approches complémentaires en matière de déploiement des LLM open-source : la performance côté serveur et l'efficacité côté client.

Sur le plan serveur, LLaMA 3 8B démontre une supériorité nette en matière de raisonnement, compréhension du langage et cohérence des réponses. Ses résultats sur les benchmarks tels que MMLU, GPQA, HumanEval et GSM-8K confirment une meilleure précision et une plus grande robustesse dans les environnements à forte charge (GPU, cloud, serveurs haute performance). Grâce à son architecture avancée et à ses outils de sécurité intégrés (Llama Guard 2, CyberSecEval 2, Code Shield), il constitue un choix idéal pour des applications serveur critiques, nécessitant une fiabilité et une qualité linguistique élevées.

De son côté, Mistral 7B reste performant sur plusieurs benchmarks (MMLU, AGI Eval, BBH, Math, Code), tout en consommant moins de ressources. Il se distingue par son efficacité énergétique et sa rapidité d'exécution, le rendant pertinent pour des serveurs à coût réduit ou des environnements nécessitant une scalabilité flexible.

Sur le plan client, Mistral 7B s'impose comme la solution la plus adaptée. Sa structure modulaire et légère (nombreuses versions finetunées, merges et quantizations) facilite son déploiement sur des machines locales ou des petites VM, même avec des ressources limitées (GPU 12 Go, 16 Go RAM). Sa rapidité et sa capacité à fonctionner en quantization (4-bit, 8-bit) assurent une expérience fluide sur des appareils de moyenne puissance.
À l'inverse, LLaMA 3 8B, bien qu'adaptable en local via la quantization GGUF, reste plus exigeant et mieux optimisé pour un usage serveur.

- **Recommandation finale :**

En conclusion, le choix du modèle dépend du contexte de déploiement :

- Pour un chatbot professionnel nécessitant des réponses précises, cohérentes et sûres, le modèle LLaMA 3 8B Instruct est le plus recommandé, notamment pour un déploiement côté serveur.

- Pour un usage local ou embarqué, où la vitesse et la légèreté sont prioritaires, Mistral 7B Instruct constitue la meilleure alternative côté client.

Ainsi, notre étude recommande LLaMA 3 8B Instruct comme modèle principal pour sa performance supérieure, sa robustesse et sa polyvalence en production serveur, tandis que Mistral 7B Instruct demeure une option secondaire, idéale pour des projets légers, rapides et économes en ressources.