AYI NEDJIMI Consultants

vLLM vs Ollama vs TGI vs SGLang — benc

16 mai 2026 • Mis à jour le 17 mai 2026 • 16 min de lecture • 3133 mots •

Benchmark complet des serveurs LLM en 2026 : vLLM, Ollama, TGI, SGLa
consommation GPU, facilité de déploiement. Quel serveur choisir selon vo

À RETENIR

A retenir -- Benchmark serveurs LLM 2026

En 2026, quatre frameworks dominent le **serving LLM en production** : vLLM
Ollama (simplicite de deploiement sur GPU grand public), TGI (robustesse en
minimale pour les workloads interactifs). Le choix depend de trois facteurs :
disponible, et la priorite entre throughput et latence. Pour 90% des deploiem
optimal en production avec plus de 10 requetes concurrentes.

Le choix du **framework de serving LLM** est une decision d'infrastructure critique
et la complexite operationnelle d'un deploiement IA 2026, l'ecos
principaux -- vLLM, Ollama, TGI (Text Generation Inf -- chacun

Devis gratuit →
Réponse sous 24h

d'usage. Ce benchmark compare ces quatre solutions sur les metriques les plus i
throughput (tokens/seconde), latence P50 et P99, efficacite memoire GPU, facilite
avec les modeles open source majeurs (Llama 3.3, Mistral, Qwen 2.5), et niveau d
configuration representative (GPU NVIDIA A100 40GB SXM) avec le modele Llama
et en float16 pour vLLM, TGI et SGLang.

vLLM -- le champion du throughput haute performance

vLLM (docs.vllm.ai) est un framework de serving LLM developpe par UC Berkeley
avec support commercial. Son innovation cle est le **PagedAttention** : une gestion
systemes d'exploitation, qui permet d'utiliser la memoire GPU de facon beaucoup
dynamiquement des "pages" de KV-cache selon les besoins de chaque sequence

Les avantages de vLLM en production :

- **Throughput maximal** : vLLM est generalement 2 a 4x plus rapide que les imple
  rapide que TGI sur des charges elevees avec de nombreuses requetes concurr
- **Continuous batching** : contrairement au batching statique, vLLM traite en conti
  courant soit termine, reduisant considerablement la latence percue sous charge
- **API OpenAI compatible** : drop-in replacement pour l'API OpenAI, facilitant la mi
  les applications existantes
- **Support multi-GPU** : tensor parallelism et pipeline parallelism pour les modeles

```
# Installation et demarrage vLLM
pip install vllm

# Serveur OpenAI-compatible sur GPU A100
```

n projet cybersécurité ?
Réponse sous 24h
Devis gratuit →

Réponse sous 24h

Devis gratuit →