

LLM Models

Chinese LLMs

- [Taiwan LLM](#) - Project TAME (TAiwanese Mixture of Experts)
 - GitHub: <https://github.com/MiuLab/Taiwan-LLM>
 - HF: <https://huggingface.co/yentinglin>
 - HF: <https://huggingface.co/audreyt>
 - [????LLM?????Project TAME?5,000??Token????????? | iThome](#)
- [TAIDE](#) - Trustworthy AI Dialogue Engine
 - GitHub: <https://github.com/taide-taiwan>
 - HF: <https://huggingface.co/taide>
 - [TAIDE | iThome](#)
- 01.AI - [Yi](#)
 - GitHub: <https://github.com/01-ai/Yi>
 - HF: <https://huggingface.co/01-ai/>
- CKIP-Llama-2-7b ??????????(CKIP)????????????????????????????Llama-2-7b??Atom-7b????????????????????????405????????????????????????70?(7 billion)?
 - GitHub: <https://github.com/f901107/CKIP-Llama-2-7b>
 - HF: <https://huggingface.co/spaces/ckiplab/CKIP-Llama-2-7b-chat>
- [Qwen](#) - ???????
 - GitHub: <https://github.com/QwenLM/Qwen>
 - GitHub: <https://github.com/QwenLM/Qwen2>
 - HF: <https://huggingface.co/Qwen>
 - Doc: <https://help.aliyun.com/zh/dashscope/create-a-chat-foundation-model?spm=a2c4g.11186623.0.0.20ea4937azFCan>
- GLM-4 - ?? AI ??????????
 - GitHub: <https://github.com/THUDM/GLM-4>
 - HF: <https://huggingface.co/collections/THUDM/glm-4-665fcf188c414b03c2f7e3b7>
- [Chinese-Mixtral](#)
- [DeepSeek](#) - ?????
 - GitHub: <https://github.com/deepseek-ai/DeepSeek-V2>

- HF: <https://huggingface.co/deepseek-ai/DeepSeek-V2>

Code LLMs

- [Granite](#) - Open sourcing IBM's Granite code models
 - H F: <https://huggingface.co/ibm-granite>
 - GitHub: <https://github.com/ibm-granite>
 - [IBM????????Granite????????????????? | iThome](#)
 - [IBM Granite 3.0 models · Ollama Blog](#)
 - [IBM Granite.Code - Visual Studio Marketplace](#)
- [Codestral](#) - Mistral's first generative AI model for code
 - HF: <https://huggingface.co/mistralai/Codestral-22B-v0.1>
 - [Mistral AI????????????? | iThome](#)

LLM Evaluation

- [PromptBench](#): A Unified Library for Evaluating and Understanding Large Language Models.
- AI?????????: [AI?????????.xlsx](#)

LLM Monitor

- [Opik](#) is an open-source platform for evaluating, testing and monitoring LLM applications.

Function Calling LLMs

- [Firefunction-v2](#)
 - HF: <https://huggingface.co/fireworks-ai/firefunction-v2>

Content Safty

- [Google ShieldGemma](#)
 ShieldGemma??4?????????????????
 ???????????????????

Calculate VRAM required for LLM

- [???? Model ???? GPU VRAM](#)

- [Calculates how much GPU memory you need and how much token/s you can get for any LLM & GPU/CPU](#)
- [LLM RAM Calculator](#)

Revision #42

Created 25 April 2024 19:42:56 by Admin

Updated 28 October 2024 09:12:20 by Admin