Key Points

* Có nhiều mô hình ngôn ngữ lớn (LLM) miễn phí, bao gồm Llama, Mistral, Falcon, MPT, BLOOM, Bert, và GPT-2, với các giấy phép mở như Apache 2.0 hoặc MIT, nhưng một số có hạn chế sử dụng nhất định.
* Bạn có thể sử dụng miễn phí bằng cách tải xuống và chạy trên máy tính của mình, thường cần GPU cho các mô hình lớn; một số có API miễn phí với giới hạn sử dụng.
* Số token phụ thuộc vào kích thước mô hình và phần cứng, không có giới hạn cụ thể từ nhà cung cấp khi chạy cục bộ.
* Yêu cầu GPU thay đổi theo mô hình: mô hình nhỏ như Bert có thể chạy trên CPU, nhưng mô hình lớn như Llama-7B cần GPU với ít nhất 16GB VRAM.

Cách Sử Dụng Miễn Phí

Để sử dụng các mô hình này miễn phí, bạn cần:

* Tải xuống trọng số mô hình từ các liên kết được cung cấp, thường trên [Hugging Face](https://huggingface.co/).
* Cài đặt thư viện cần thiết như Transformers từ Hugging Face.
* Chạy mô hình cục bộ trên máy tính, có thể cần GPU cho các mô hình lớn.

Một số mô hình cũng có thể truy cập qua API Inference của Hugging Face, có tầng miễn phí với giới hạn sử dụng, nhưng để sử dụng không giới hạn, chạy cục bộ là lựa chọn tốt nếu bạn có phần cứng phù hợp.

Số Token và Yêu Cầu GPU

* Số token: Token là đơn vị văn bản mà mô hình xử lý, mỗi mô hình có cửa sổ ngữ cảnh (context window) tối đa, ví dụ Llama-2-7B có 4096 token. Số token bạn có thể xử lý phụ thuộc vào bộ nhớ phần cứng và cửa sổ ngữ cảnh của mô hình.
* Yêu cầu GPU: Mô hình nhỏ như Bert-base (110 triệu tham số) có thể chạy trên CPU, nhưng chậm hơn. Mô hình lớn như Llama-7B cần GPU với ít nhất 16GB VRAM, trong khi mô hình lớn hơn như Falcon-40B có thể cần GPU với 40-50GB VRAM hoặc nhiều GPU.

Ghi Chú Khảo Sát: Danh Sách và Phân Tích Các LLM Miễn Phí

Giới Thiệu

Bài viết này cung cấp danh sách đầy đủ các mô hình ngôn ngữ lớn (LLM) miễn phí, cùng với liên kết, cách sử dụng miễn phí, số token, và yêu cầu GPU. Các mô hình này được phát hành với các giấy phép mở, phù hợp cho cả nghiên cứu và sử dụng thương mại, với một số hạn chế nhất định. Dựa trên thông tin cập nhật đến ngày 26 tháng 3 năm 2025, chúng tôi phân tích chi tiết để giúp người dùng hiểu rõ hơn về cách tận dụng các công cụ AI này.

Danh Sách Các LLM Miễn Phí

Dưới đây là danh sách các LLM miễn phí, bao gồm thông tin về nhà phát triển, giấy phép, liên kết, và cách sử dụng:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Mô Hình** | **Nhà Phát Triển** | **Giấy Phép** | **Liên Kết** | **Cách Sử Dụng Miễn Phí** | **Số Token** | **Yêu Cầu GPU** |
| Llama | Meta AI | Apache 2.0 | <https://ai.meta.com/llama/> | Tải xuống và chạy cục bộ trên GPU | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |
| Mistral | Mistral AI | Apache 2.0 | <https://mistral.ai/> | Tải xuống và chạy cục bộ trên GPU | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |
| Falcon | EleutherAI | Apache 2.0 | <https://huggingface.co/EleutherAI/falcon-7b> | Tải xuống và chạy cục bộ trên GPU | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |
| MPT | MosaicML | Apache 2.0 | <https://huggingface.co/mosaicml/mpt-7b> | Tải xuống và chạy cục bộ trên GPU | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |
| BLOOM | BigScience | BigScience RAIL v1.0 | <https://huggingface.co/bigscience/bloom> | Tải xuống và chạy cục bộ trên GPU, tuân thủ giấy phép | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |
| Bert | Google | Apache 2.0 | <https://huggingface.co/bert-base-uncased> | Tải xuống và chạy trên CPU hoặc GPU | Phụ thuộc kích thước mô hình | Mô hình nhỏ chạy trên CPU, lớn cần GPU |
| GPT-2 | OpenAI | Modified MIT | <https://huggingface.co/gpt2> | Tải xuống và chạy cục bộ trên GPU | Phụ thuộc kích thước mô hình | Cần GPU, lớn hơn cần VRAM cao hơn |

Ngoài các mô hình trên, còn có một số mô hình khác như Dolly, StableLM, và Pythia, cũng miễn phí với giấy phép Apache 2.0, nhưng bài viết tập trung vào các mô hình chính được đề cập.

Phân Tích Sử Dụng Miễn Phí

Để sử dụng các mô hình này miễn phí, người dùng cần:

* Tải xuống trọng số mô hình từ các liên kết, thường trên [Hugging Face](https://huggingface.co/). Ví dụ, để sử dụng Llama, bạn có thể truy cập <https://ai.meta.com/llama/> và tìm liên kết tải xuống.
* Cài đặt môi trường với các thư viện cần thiết, như Transformers từ Hugging Face, bằng lệnh pip install transformers.
* Chạy mô hình cục bộ trên máy tính. Ví dụ, để chạy Llama, bạn có thể sử dụng mã Python như sau:

python

from transformers import LlamaTokenizer, LlamaForCausalLanguageModeling

tokenizer = LlamaTokenizer.from\_pretrain("meta-llama/Llama-2-7b-hf")

model = LlamaForCausalLanguageModeling.from\_pretrain("meta-llama/Llama-2-7b-hf")

* Một số mô hình có thể truy cập qua API Inference của Hugging Face, có tầng miễn phí với giới hạn sử dụng. Ví dụ, để sử dụng MPT-7B, bạn có thể vào <https://huggingface.co/mosaicml/mpt-7b> và sử dụng tab "Inference" để lấy API token và gửi yêu cầu.

Tuy nhiên, để sử dụng không giới hạn, chạy cục bộ là lựa chọn tốt nhất, với điều kiện bạn có phần cứng phù hợp, đặc biệt là GPU cho các mô hình lớn.

Số Token và Giới Hạn

* Trong ngữ cảnh LLM, token là đơn vị văn bản mà mô hình xử lý. Mỗi mô hình có cửa sổ ngữ cảnh tối đa, ví dụ Llama-2-7B có 4096 token, nghĩa là có thể xử lý chuỗi lên đến 4096 token. Số token bạn có thể xử lý phụ thuộc vào bộ nhớ phần cứng và cửa sổ ngữ cảnh của mô hình.
* Không có giới hạn cụ thể từ nhà cung cấp khi chạy cục bộ, nhưng nếu sử dụng API miễn phí, có thể có giới hạn về số yêu cầu hoặc token mỗi tháng, tùy thuộc vào dịch vụ.

Yêu Cầu GPU

* Yêu cầu GPU thay đổi theo kích thước mô hình. Mô hình nhỏ như Bert-base (110 triệu tham số) có thể chạy trên CPU, nhưng tốc độ chậm. Mô hình lớn như Llama-7B cần GPU với ít nhất 16GB VRAM. Mô hình lớn hơn như Falcon-40B (40 tỷ tham số) có thể cần GPU với 40-50GB VRAM hoặc sử dụng kỹ thuật phân chia mô hình trên nhiều GPU.
* Ví dụ, để chạy Llama-7B, bạn cần GPU với ít nhất 16GB VRAM, trong khi mô hình lớn hơn như Llama-70B có thể cần nhiều GPU hoặc phần cứng mạnh hơn.

Thông Tin Bổ Sung

* Một số mô hình như BLOOM có giấy phép đặc biệt (BigScience RAIL License v1.0), với các hạn chế sử dụng, như cấm sử dụng cho mục đích quản lý công lý, thực thi pháp luật, hoặc quy trình nhập cư. Người dùng cần tuân thủ các điều khoản này.
* Ngoài chạy cục bộ, một số dịch vụ như Hugging Face Inference API cung cấp tầng miễn phí, nhưng giới hạn không rõ ràng, nên người dùng cần kiểm tra kỹ trước khi sử dụng.

Kết Luận

Danh sách trên cung cấp các lựa chọn LLM miễn phí, phù hợp cho nhiều mục đích, từ nghiên cứu đến phát triển ứng dụng. Với phần cứng phù hợp, bạn có thể tận dụng sức mạnh của AI mà không tốn phí, nhưng cần chú ý đến yêu cầu GPU và giới hạn giấy phép.

Key Citations

* [Llama official page](https://ai.meta.com/llama/)
* [Mistral AI official page](https://mistral.ai/)
* [Falcon 7B on Hugging Face](https://huggingface.co/EleutherAI/falcon-7b)
* [MPT 7B on Hugging Face](https://huggingface.co/mosaicml/mpt-7b)
* [BLOOM on Hugging Face](https://huggingface.co/bigscience/bloom)
* [Bert base uncased on Hugging Face](https://huggingface.co/bert-base-uncased)
* [GPT-2 on Hugging Face](https://huggingface.co/gpt2)