Analytics Vidhya

Popular Open-Source LLM API Platforms

Instructor

Dipanjan Sarkar

Head of Community & Principal AI Scientist at Analytics Vidhya

Google Developer Expert - ML & Cloud Champion Innovator

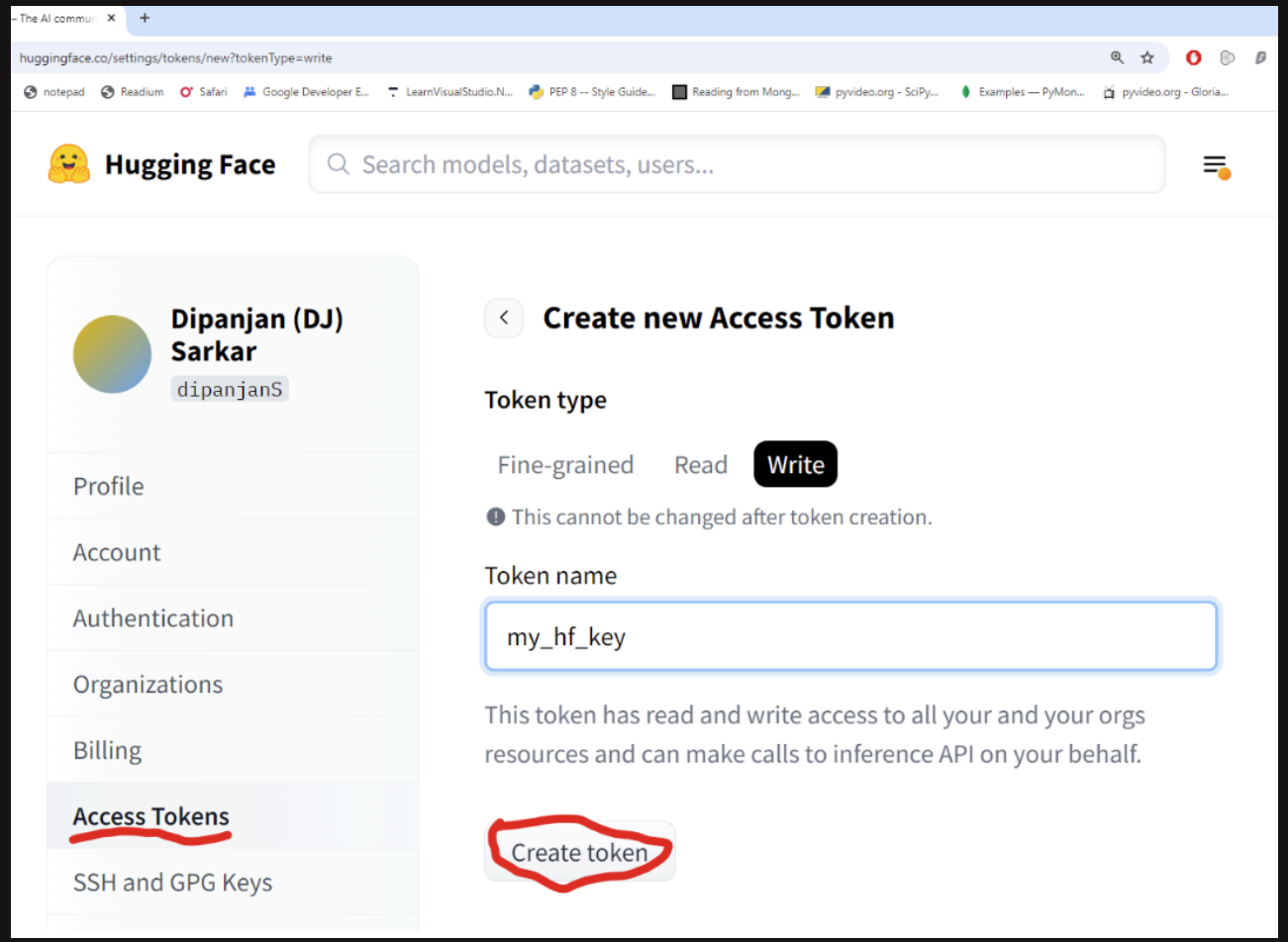
Published Author

Open-Source LLM API Platforms

Groq

HuggingFace

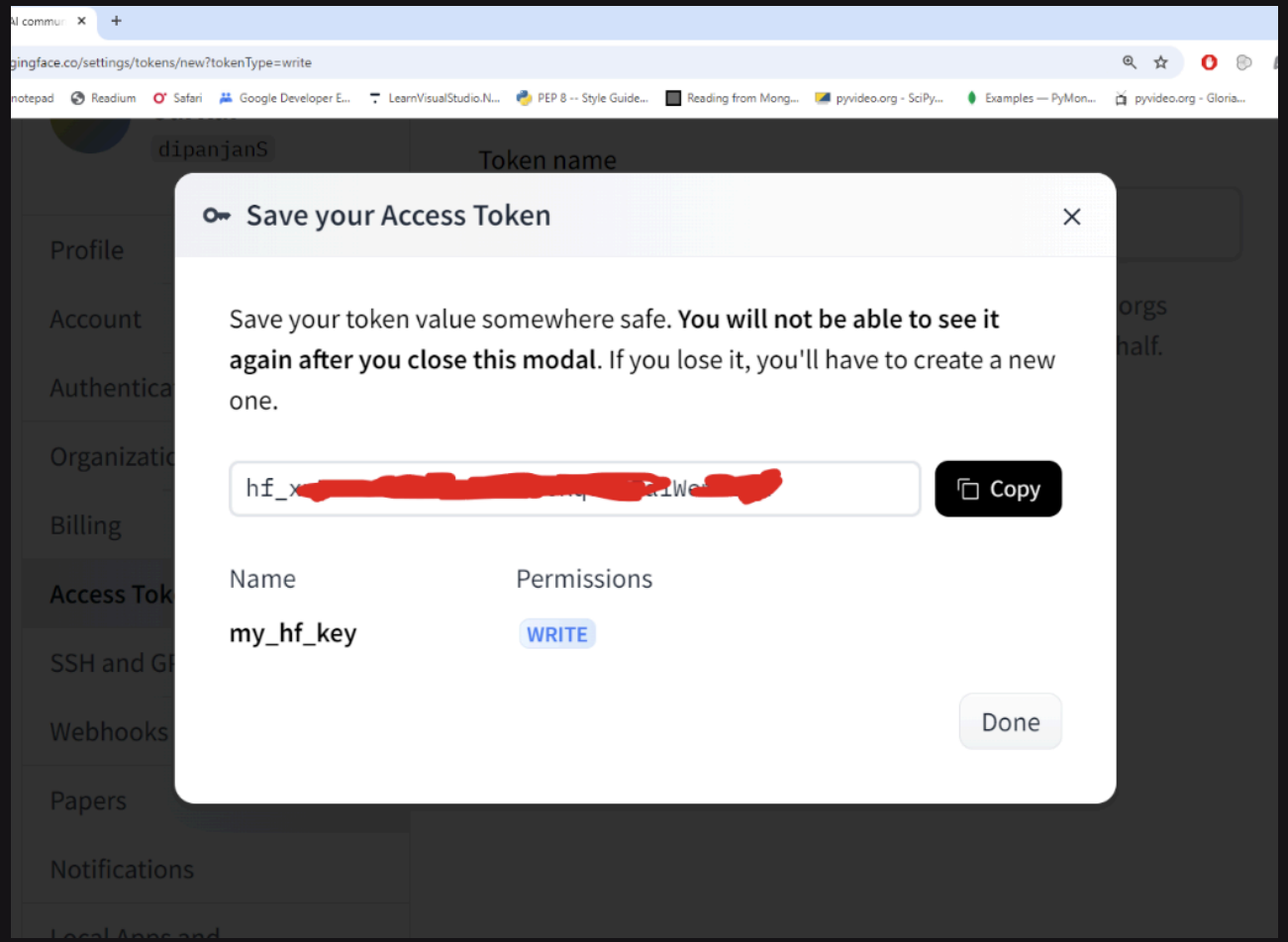
Get your Hugging Face Access Token



- Create and sign in to your account on https://huggingface.co/
- Go to Settings -> Access Tokens
- Create a new token with Write permissions

Analytics Vidhya

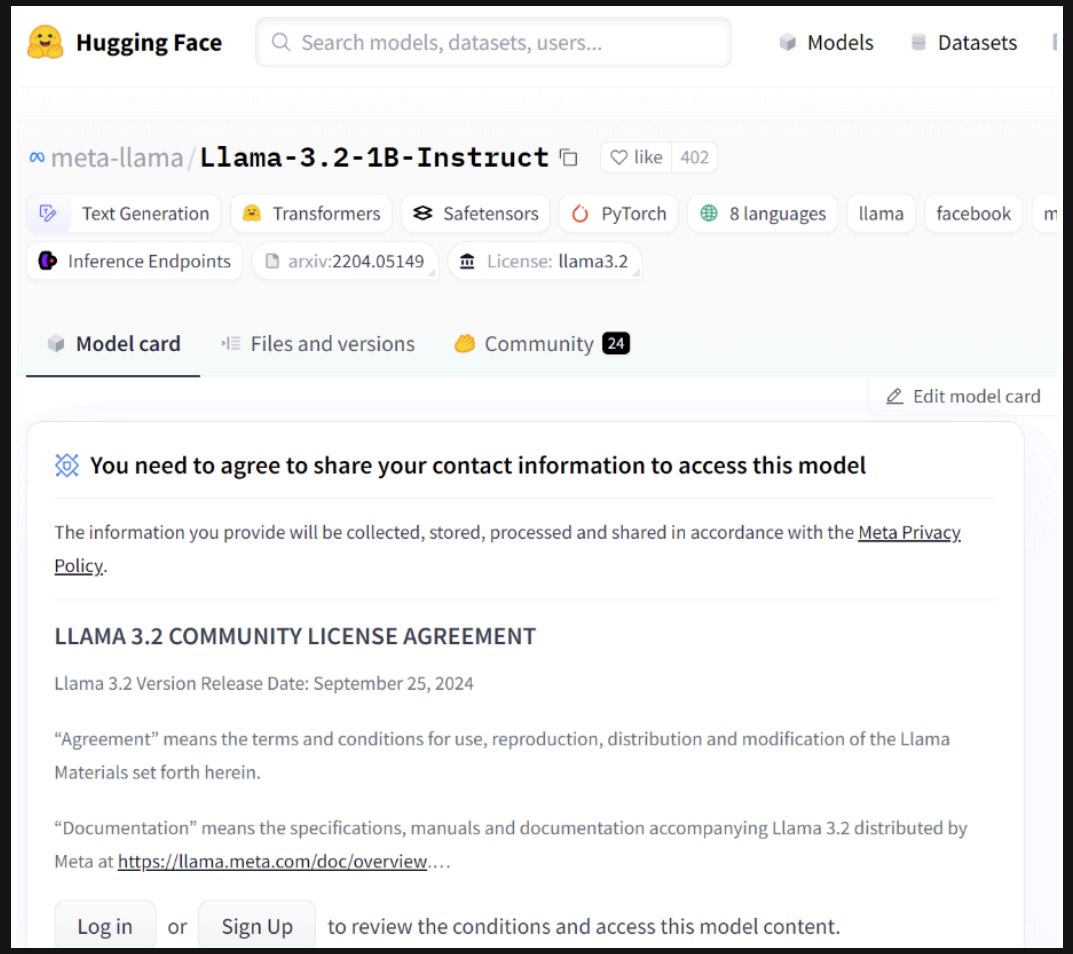
Save your Hugging Face Access Token



- Save your secret key in a secure location as it will be just shown once
- If you forget your key just generate a new one
- Do NOT paste your key publicly on the internet

Analytics Vidhya

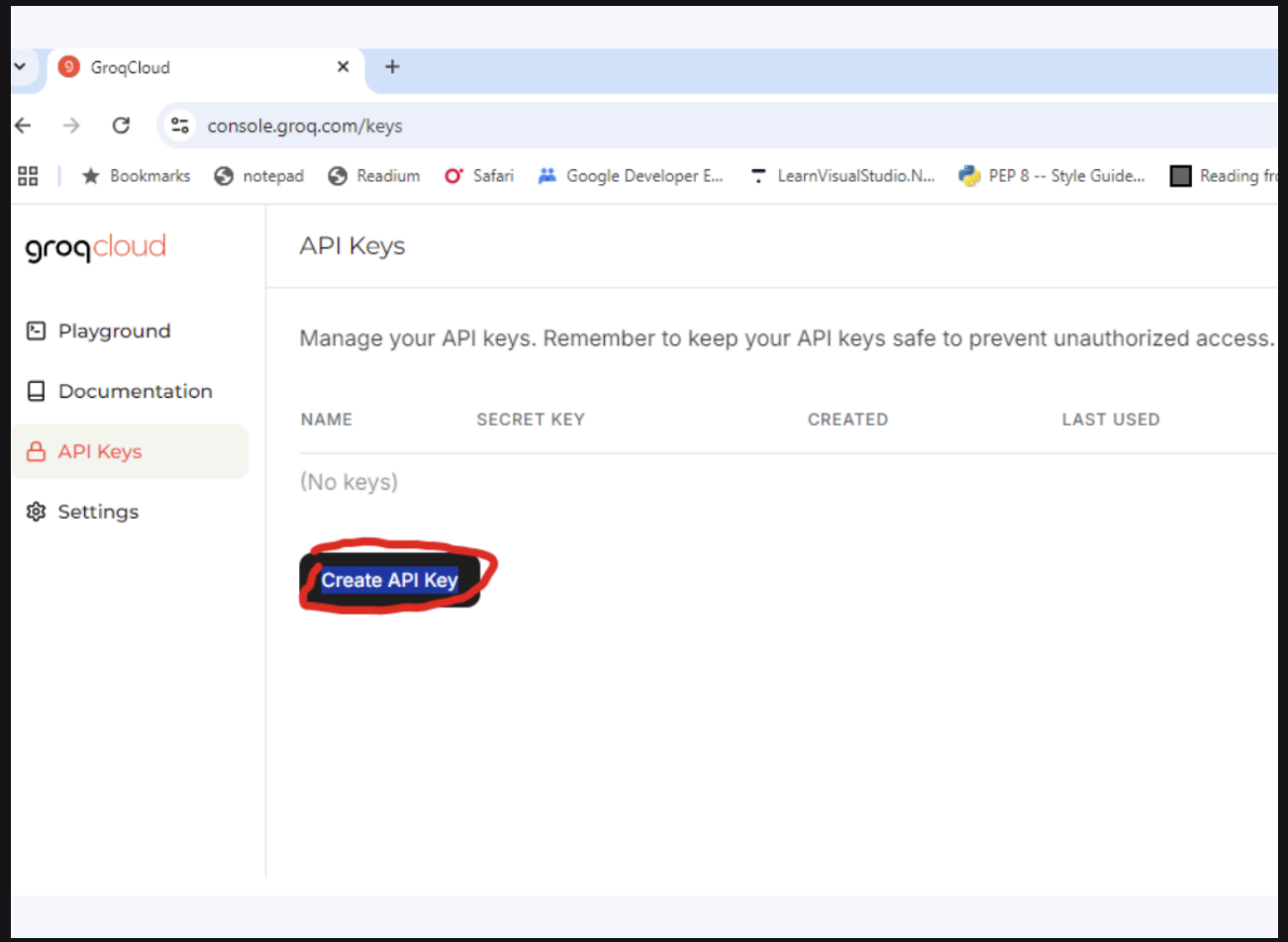
Apply for Access to Gated Models



- Several new LLMs have gated access
- If you want to use these LLMs like Llama 3.2 Instruct you need to apply for access
- Go to the relevant model page for models you want to use and access the terms and conditions
- You can then start using the model via Hugging Face transformers

Analytics Vidhya

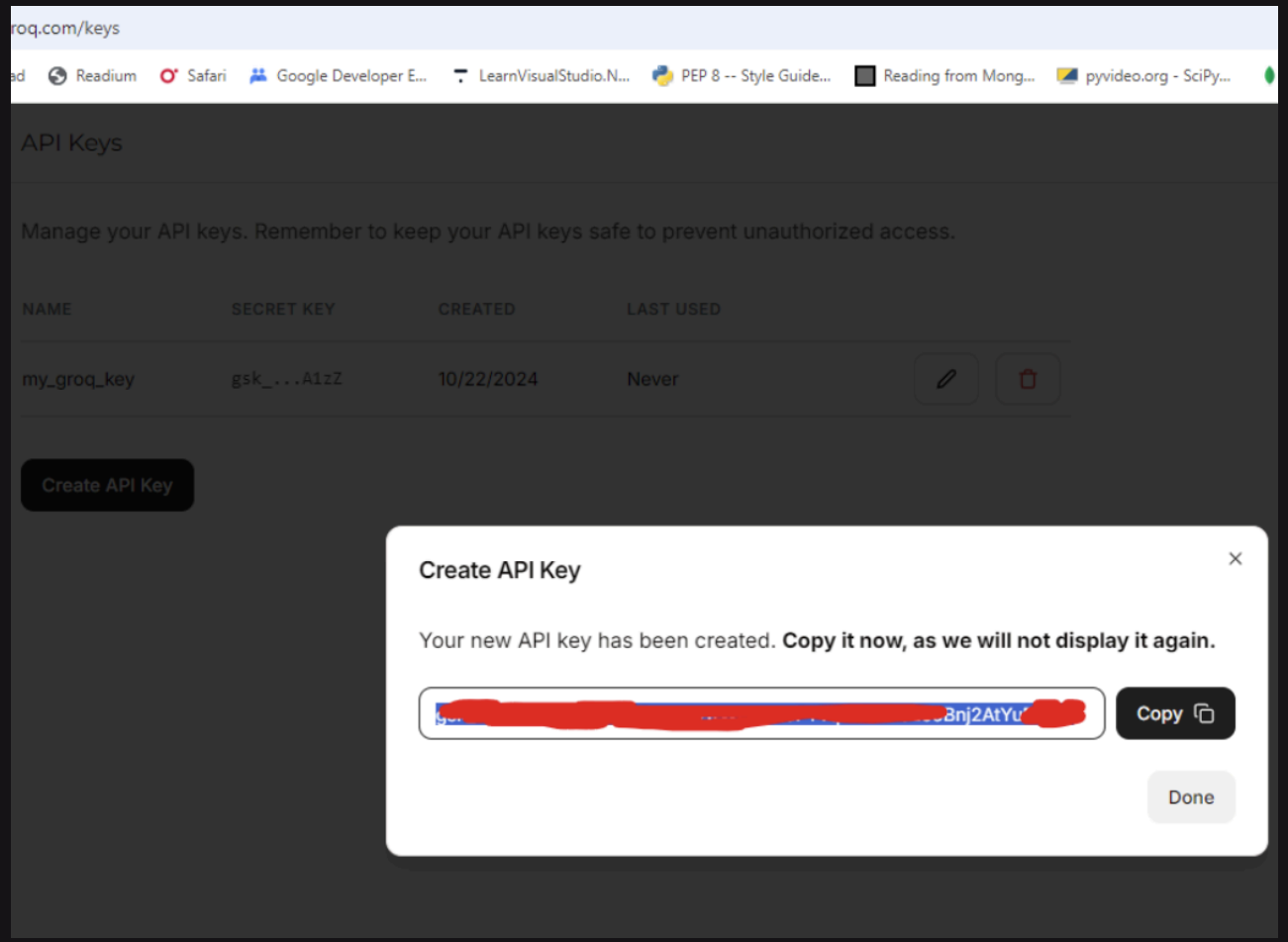
Get your Groq Cloud API Key



- Create an account on https://console.groq.com/keys
- Go to Groq Cloud -> Create API Key to create your API key
- Several open source models have a free tier access

Analytics Vidhya

Save your Groq Cloud API Key



- Save your secret key in a secure location as it will be shown just once
- Can create more keys if you lose it
- Do NOT paste your key publicly on the internet

Analytics Vidhya

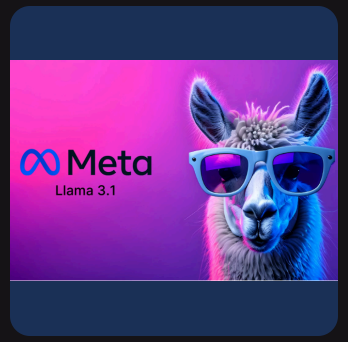
Key Highlights of Llama 3.1

Model Variants:

Released in 8B, 70B, and an enormous 405B parameter version.

Benchmark Leader:

Outperforms most LLMs across nearly all major benchmarks.



Multilingual Excellence:

Supports multiple languages with superior reasoning capabilities.

Future Multimodal Variants:

Upcoming variants to include audio, video, and image processing.

Extended Context Length:

Handles up to 128K tokens, ideal for complex, lengthy inputs.

Analytics Vidhya

Key Highlights of Llama 3.2

Released by Meta in September 2024, introduces significant advancements in LLMs, enhancing both multimodal capabilities and efficiency.

- **Multimodal Processing**:
  - Supports text and image inputs for tasks like visual reasoning, image captioning, and document analysis.
- **Model Variants**:
  - **Llama 3.2 90B Vision**: High-performance model for complex, resource-intensive tasks.
  - **Llama 3.2 11B Vision**: Balanced model for optimal performance and resource efficiency.
  - **Llama 3.2 3B and 1B**: Lightweight text-only models for mobile and edge deployments.

Key Highlights of Llama 3.2

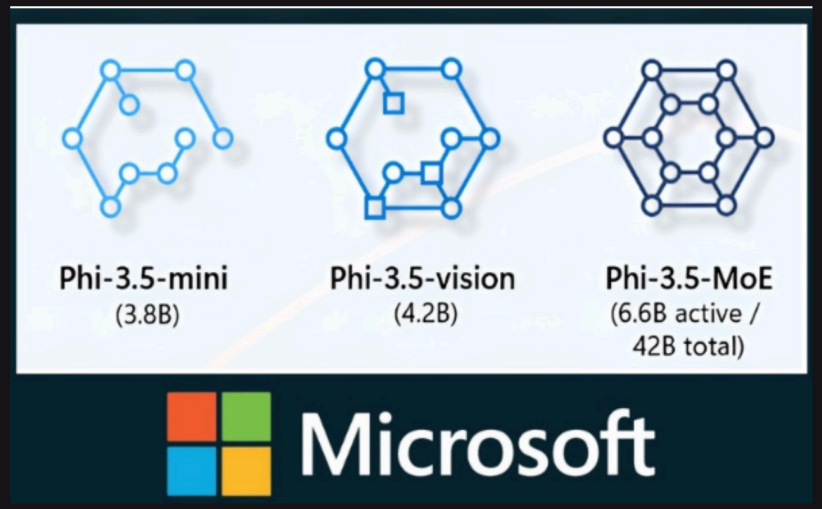
Released by Meta in September 2024, introduces significant advancements in LLMs, enhancing both multimodal capabilities and efficiency.

- **Enhanced Reasoning:**
  - Excels in visual understanding, grounding, document QA, and image-text retrieval.
- **Multilingual Support:**
  - Handles multiple languages in text-only mode.
- **On-Device AI:**
  - The 1B and 3B models enable efficient AI applications on mobile and edge devices without cloud dependence.

Other Popular Open-Source LLMs

Qwen2.5

Gemma



MISTRAL AI_

Thank You

Analytics Vidhya