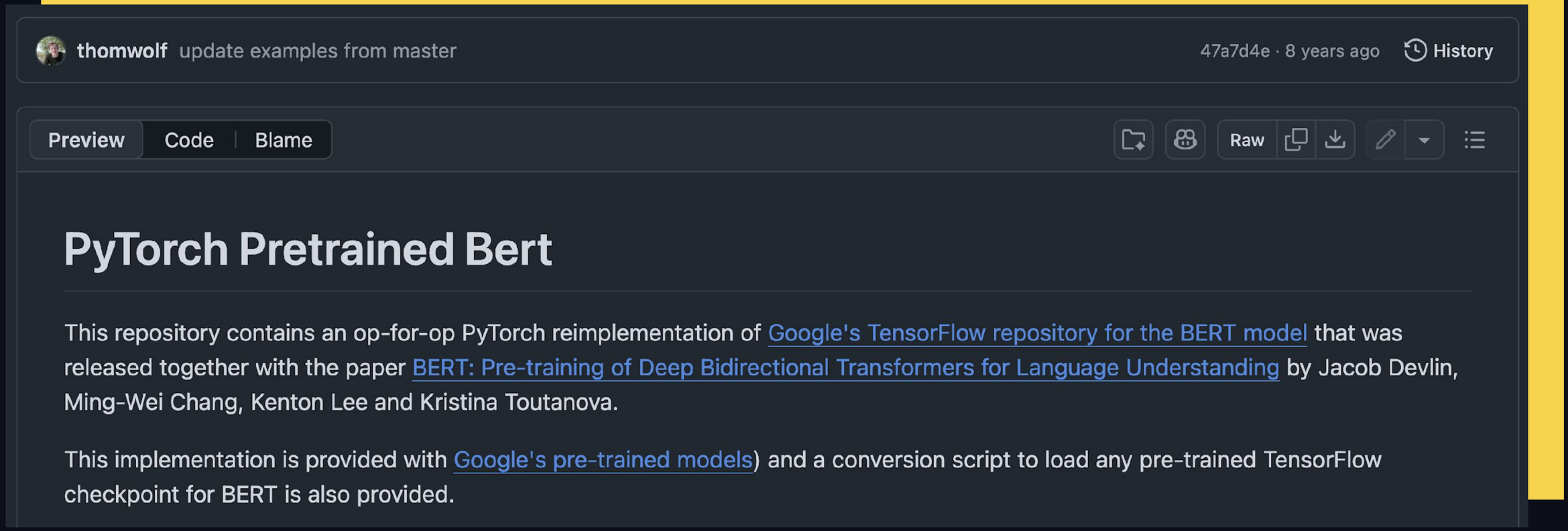
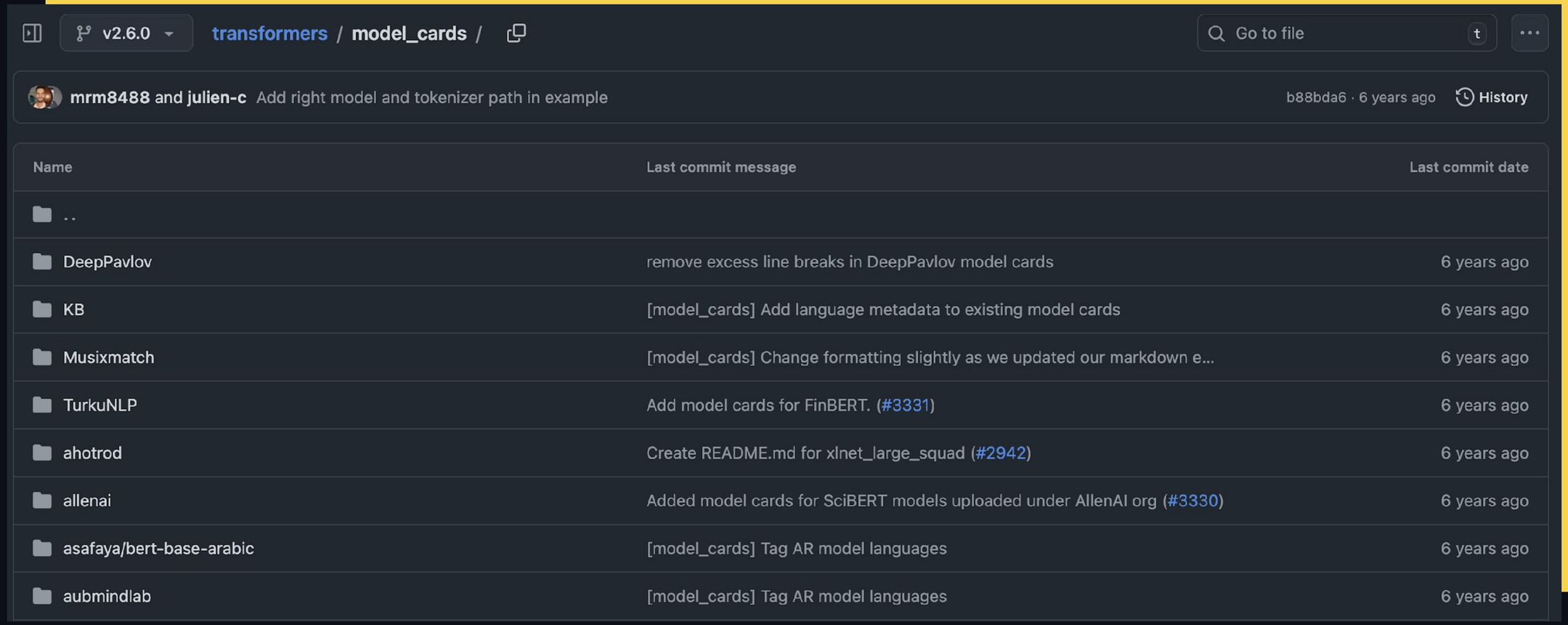
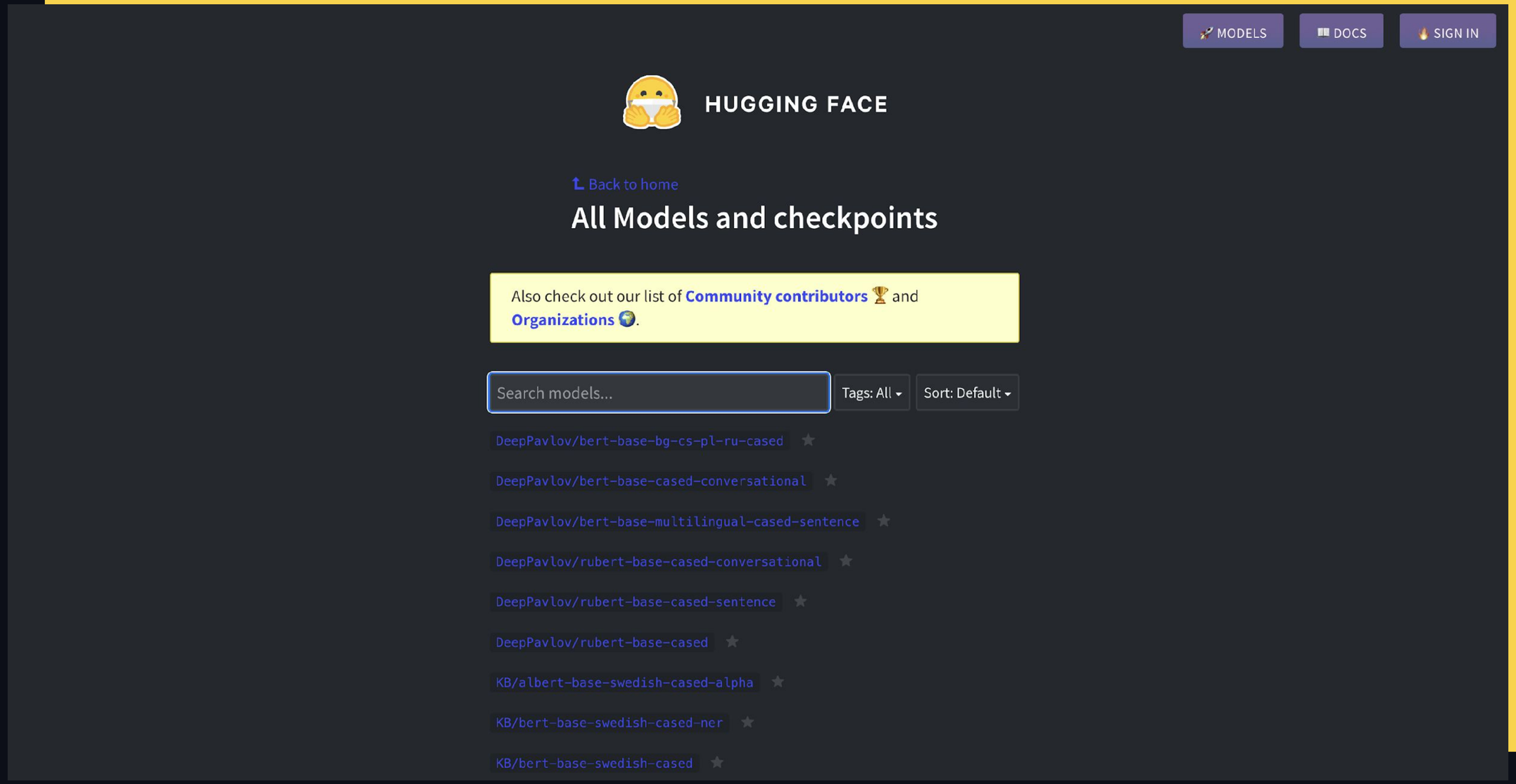
The Hub as Infrastructure

From open PyTorch models, to a safe and performant distribution hub.

Lysandre, Hugging Face


thomwolf update examples from master
47a7d4e · 8 years ago
History
Preview
Code
Blame
Raw
PyTorch Pretrained Bert
This repository contains an op-for-op PyTorch reimplementation of Google's TensorFlow repository for the BERT model that was released together with the paper BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding by Jacob Devlin, Ming-Wei Chang, Kenton Lee and Kristina Toutanova.
This implementation is provided with Google's pre-trained models) and a conversion script to load any pre-trained TensorFlow checkpoint for BERT is also provided.


v2.6.0
transformers / model_cards /
Go to file t
mrm8488 and julien-c Add right model and tokenizer path in example
b88bda6 · 6 years ago
History
Name
Last commit message
Last commit date
..
DeepPavlov
remove excess line breaks in DeepPavlov model cards
6 years ago
KB
[model_cards] Add language metadata to existing model cards
6 years ago
Musixmatch
[model_cards] Change formatting slightly as we updated our markdown e...
6 years ago
TurkuNLP
Add model cards for FinBERT. (#3331)
6 years ago
ahotrod
Create README.md for xlnet_large_squad (#2942)
6 years ago
allenai
Added model cards for SciBERT models uploaded under AllenAI org (#3330)
6 years ago
asafaya/bert-base-arabic
[model_cards] Tag AR model languages
6 years ago
aubmindlab
[model_cards] Tag AR model languages
6 years ago


🚀 MODELS
📖 DOCS
🔥 SIGN IN
HUGGING FACE
Back to home
All Models and checkpoints
Also check out our list of Community contributors 🏆 and Organizations 🌍.
Search models...
Tags: All
Sort: Default
DeepPavlov/bert-base-bg-cs-pl-ru-cased
DeepPavlov/bert-base-cased-conversational
DeepPavlov/bert-base-multilingual-cased-sentence
DeepPavlov/rubert-base-cased-conversational
DeepPavlov/rubert-base-cased-sentence
DeepPavlov/rubert-base-cased
KB/albert-base-swedish-cased-alpha
KB/bert-base-swedish-cased-ner
KB/bert-base-swedish-cased

Hugging Face | Search models, datasets, users... | Models | Datasets | Spaces | Buckets NEW | Docs | Enterprise | Pricing

Log in using Single Sign-On to view activity within the **huggingface** org. Log In

Main | Tasks | Libraries | Languages | Licenses | Other

Tasks

Text Generation | Any-to-Any | Image-Text-to-Text | Image-to-Text | Image-to-Image | Text-to-Image | Text-to-Video | Text-to-Speech | +44

Parameters

<1B | 6B | 12B | 32B | 128B | >500B

Libraries

PyTorch | TensorFlow | JAX | Transformers | Diffusers | sentence-transformers | Safetensors | ONNX | GGUF | Transformers.js | MLX | +42

Apps

vLLM | llama.cpp | MLX LM | LM Studio | Ollama | Jan | Draw Things | +8

Inference Providers

Groq | Novita | Cerebras | SambaNova | Nscale | fal | Hyperbolic | Together AI | +16

Models 2,740,384

Filter by name | Full-text search | Inference Available | Sort: Trending

Jackrong/Qwen3.5-27B-Claude-4.6-Opus-Reasoning-Dist...
Image-Text-to-Text • 28B • Updated 2 days ago • 184k • ♡ 1.35k

HauhauCS/Qwen3.5-35B-A3B-Uncensored-HauhauCS-Aggres...
Image-Text-to-Text • 35B • Updated 16 days ago • 425k • ♡ 962

nvidia/Nemotron-Cascade-2-30B-A3B
Text Generation • 32B • Updated 1 day ago • 55.5k • ♡ 304

GAIR/daVinci-MagiHuman
Image-to-Video • Updated about 20 hours ago • 273 • ♡ 168

baidu/Qianfan-OCR
Image-Text-to-Text • 5B • Updated 7 days ago • 11.9k • ♡ 373

Jackrong/Qwen3.5-27B-Claude-4.6-Opus-Reasoning-Dist...
Image-Text-to-Text • 27B • Updated 1 day ago • 57.1k • ♡ 154

Tesslate/OmniCoder-9B
Text Generation • Updated 13 days ago • 22.1k • ♡ 458

Jackrong/Qwen3.5-9B-Claude-4.6-Opus-Reasoning-Disti...
Image-Text-to-Text • 9B • Updated 3 days ago • 66.9k • ♡ 161

Jackrong/Qwen3.5-27B-Claude-4.6-Opus-Reasoning-Dist...
Image-Text-to-Text • 27B • Updated 2 days ago • 511k • ♡ 409

RoyalCities/Foundation-1
Updated 10 days ago • ♡ 263

HauhauCS/Qwen3.5-9B-Uncensored-HauhauCS-Aggressive
9B • Updated 22 days ago • 440k • ♡ 655

fishaudio/s2-pro
Text-to-Speech • 5B • Updated 15 days ago • 15.3k • ♡ 743

zai-org/GLM-OCR
Image-to-Text • Updated 14 days ago • 3.64M • ♡ 1.46k

Qwen/Qwen3.5-9B
Image-Text-to-Text • 10B • Updated 24 days ago • 3.74M • ♥ 1.02k

mistralai/Mistral-Small-4-119B-2603
119B • Updated about 11 hours ago • 41.9k • ♡ 327

Jackrong/Qwen3.5-9B-Claude-4.6-Opus-Reasoning-Disti...
Image-Text-to-Text • 10B • Updated 3 days ago • 31.3k • ♡ 123

Lightricks/LTX-2.3
Image-to-Video • Updated 11 days ago • 1.18M • ♡ 766

Qwen/Qwen3.5-35B-A3B
Image-Text-to-Text • 36B • Updated 27 days ago • 2.93M • ♥ 1.26k

2.7M public repositories total

Models 241,169 | Filter by name | Inference Available | Edit filters | ↑↓ Sort: Trending

Active filters: peft | Clear all

alvdansen/illustration-1.0-flux-dev

Text-to-Image • Updated 3 days ago • ♡ 6

artificialguybr/AceStep_Refine_Redmond

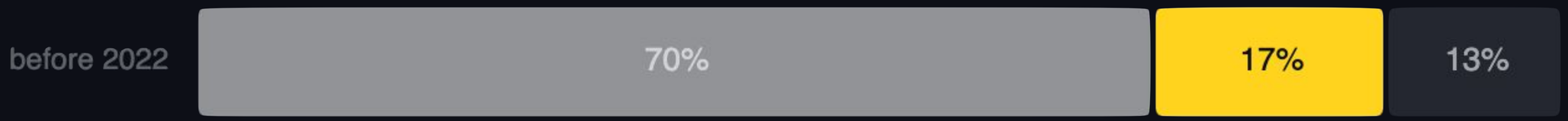
Text-to-Audio • Updated 7 days ago • ♡ 5

oddadmix/Katib-Qwen3.5-0.8B-0.1

Model tree for meta-llama/Llama-3.2-3B-Instruct

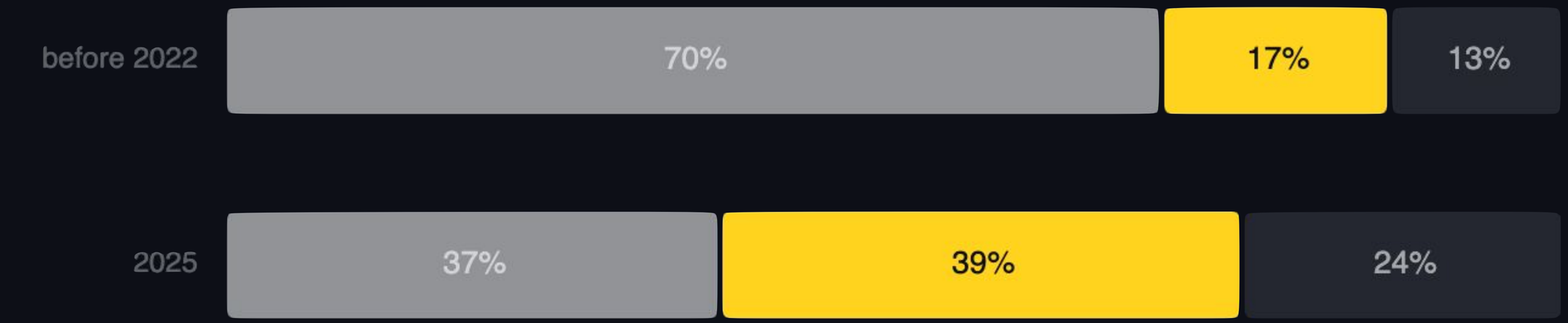
- Adapters ..... 648 models
- Finetunes ..... 1462 models
- Merges ..... 19 models
- Quantizations ..... 429 models

WHO BUILDS THE MODELS · BEFORE 2022 vs 2025



Industry Independent developers Other

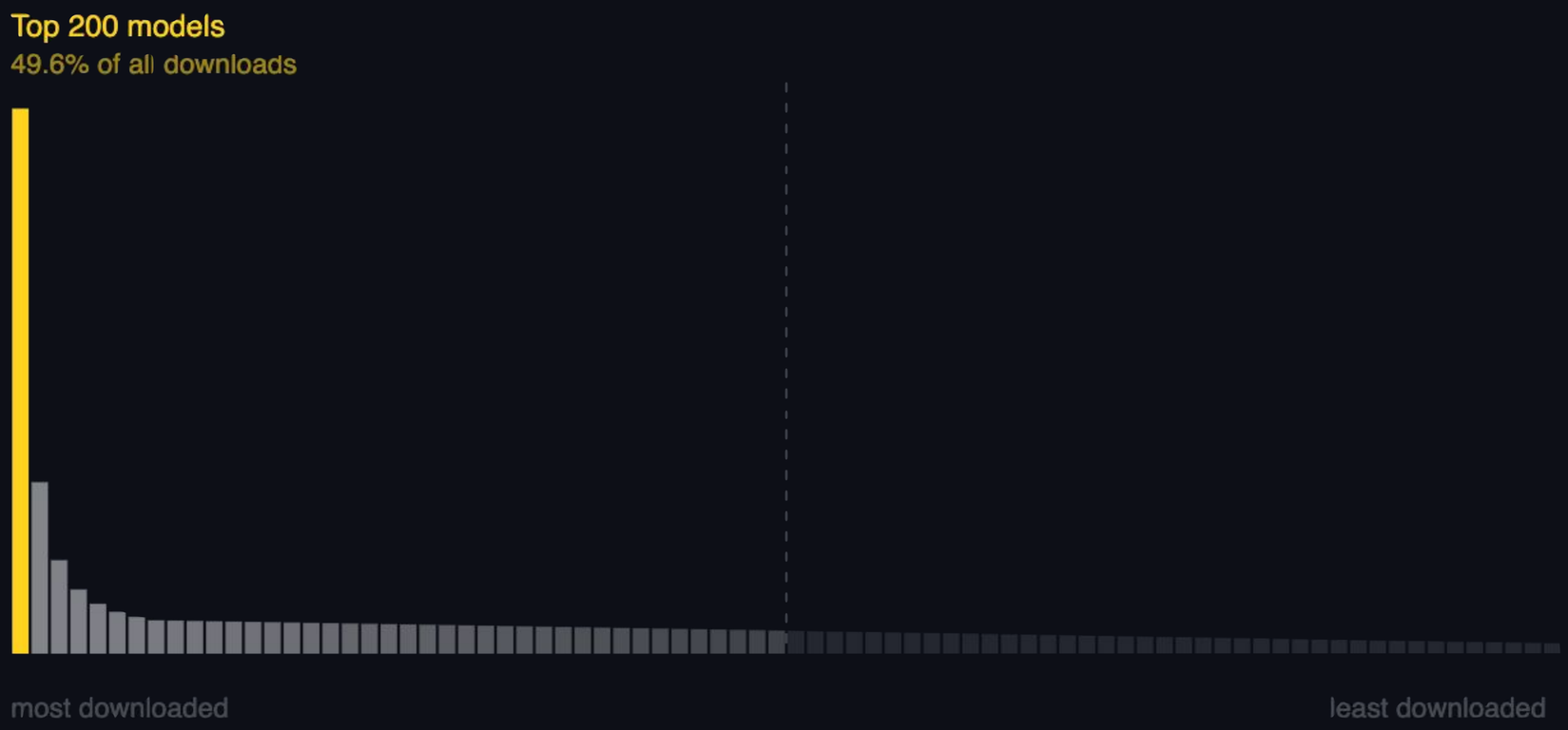
WHO BUILDS THE MODELS · BEFORE 2022 vs 2025



Industry Independent developers Other

30%+

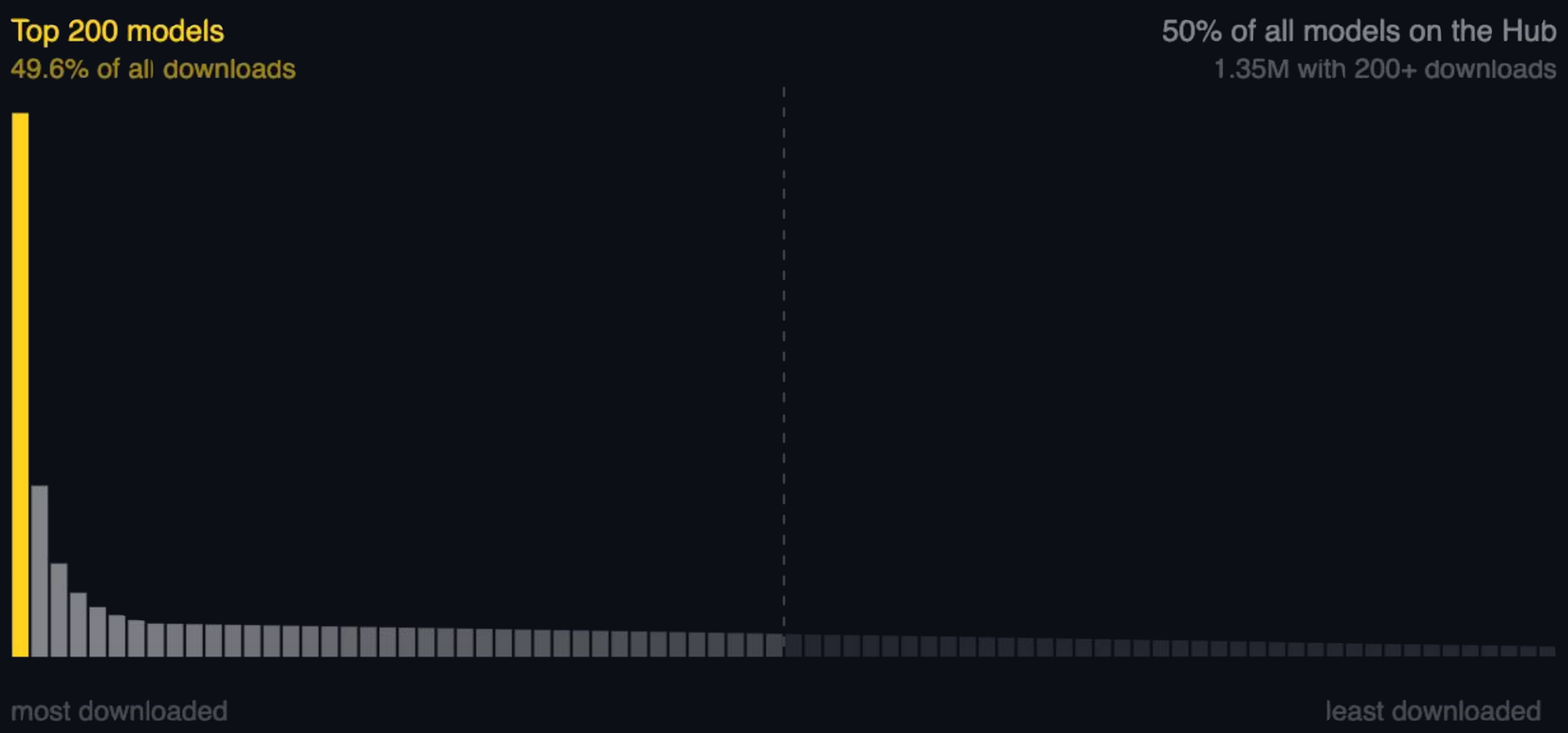
of Fortune 500 companies
have a verified org on the Hub



top 200 models

49.6% of all downloads

2.7M public repositories total



top 200 models

49.6% of all downloads

1.35M models

each with 200+ downloads

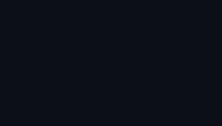
2.7M public repositories total

4.47M

model versions scanned

352K

unsafe or suspicious files found



Safetensors

ML Safer For All

Safetensors

ML Safer For All

WHAT CHANGES

Governance

Formal steering committee

Stewardship

No longer owned by any single company

Community ownership

Safetensors becomes a fully community-owned tool

Natively integrated into PyTorch

First-class support within the PyTorch ecosystem

WHAT CHANGES

Governance
Formal steering committee

Stewardship
No longer owned by any single company

Community ownership
Safetensors becomes a fully community-owned tool

Natively integrated into PyTorch
First-class support within the PyTorch ecosystem

WHAT DOESN'T CHANGE

The format
Safetensors is exactly what it was yesterday

Pace of iteration
Development continues at the same cadence

New data types
Support keeps expanding as the ecosystem grows

WHAT'S NEXT

Device-aware loading

Direct load/save onto CUDA, ROCm, and XPU with no CPU staging and no unnecessary copies.

Quantization & data types

Formalise FP8 support and track emerging low-precision formats as the inference landscape evolves.

Distributed training & inference

Tensor Parallel and Pipeline Parallel loading with integration hooks for TorchTitan, DeepSpeed, and vLLM.

Security & community

No-code-execution guarantees upheld. Roadmap driven by the PyTorch Foundation TAC.

```python
from transformers import AutoModel
model = AutoModel.from_pretrained("...")
```

THE LAST MILE

GPU UTILISATION

underutilized

distribution solved

execution unsolved

One model, many targets

NVIDIA H100 / A100	**AMD** MI300 / ROCm	**Apple** M-series / MLX	**Arm** Neoverse
Intel XPU / CPU	**AWS** Trainium / Inferentia (Neuron / NKI)	**Google** TPU v4 / v5	Many others

```
__global__ void fused_attention_kernel(
  float* Q, float* K, float* V,
  float* out, int seq_len, int head_dim) {
  int tid = blockIdx.x * blockDim.x + threadIdx.x;
  __shared__ float smem[BLOCK_SIZE][HEAD_DIM];
  load_to_shared(smem, Q, tid, head_dim);
  __syncthreads();
  atomicAdd(&out[tid], dot(smem[tid], K) * scale);
}
```

Memory layout awareness per architecture

Shared memory and thread synchronisation

Different code for every hardware target

KERNELS AVAILABLE · YOUR HARDWARE HIGHLIGHTED

pulled for your runtime · available, not needed · open for contributions

indicative — not reflective of current support

KERNELS AVAILABLE · YOUR HARDWARE HIGHLIGHTED

pulled for your runtime · available, not needed · open for contributions

indicative — not reflective of current support

GETTING STARTED

Introduction

Installation

USING KERNELS

Basic Usage

Using Layers

Locking Kernel Versions

Environment Variables

FAQ

Migrating from older versions

BUILDING KERNELS

Writing Kernels

Building with Nix

Building with Docker

Local Development

Kernel Requirements

Security

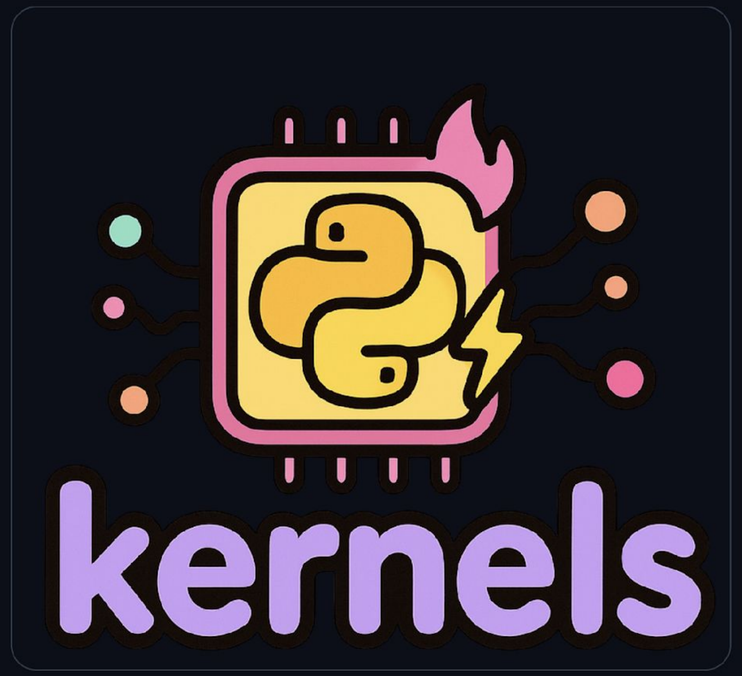
Why Nix?

Metal Notes

Build Variants

Kernels

Copy page



The Kernel Hub allows Python libraries and applications to load compute kernels directly from the Hub. To support this kind of dynamic loading, Hub kernels differ from traditional Python kernel packages in that they are made to be:

- **Portable**: a kernel can be loaded from paths outside `PYTHONPATH`.
- **Unique**: multiple versions of the same kernel can be loaded in the same Python process.
- **Compatible**: kernels must support all recent versions of Python and the different PyTorch build configurations (various CUDA versions and C++ ABIs). Furthermore, older C library versions must be supported.

You can search for kernels on the Hub.

Kernel Pages and the Kernel Hub

Released April 8th, 2026

Kernels: sgl-project/sgl-flash-attn3 | like 0 | Follow SGLang Project 19

Kernel card | Files and versions xet | Community | Settings

Edit kernel card

sglang-flash-attn3

Pre-built Flash Attention 3 (forward-only) CUDA kernels from sgl-flash-attn

Kernel source: kernels-community/sgl-flash-attn3

Usage

```
pip install kernels
```

```
from kernels import get_kernel

fa3 = get_kernel("sgl-project/sgl-flash-attn3", version=1)

fa3.flash_attn_varlen_func(q, k, v, cu_seqlens_q, cu_seqlens_k, causal=True)
fa3.flash_attn_with_kvcache(q, k_cache, v_cache, cache_seqlens=cache_seqlens, cau
fa3.is_fa3_supported()  # True on H100/H200
```

Credits

- Tri Dao - Flash Attention 3
- SGLang - `sgl_kernel` FA3 implementation
- HuggingFace - kernel-builder infrastructure

Use this kernel

Downloads last month -

kernels | bsd-3-clause

Supported hardwares new

CUDA 8.0 9.0a

H200 141GB | H100 80GB | L40s 48GB | L40 48GB | L20 48GB

OS linux

Arch x86_64

Agents, Kernels, and the Kernel Hub

15:25 CEST	Bridging the Hardware Gap With Code Harnesses on the Hugging Face Kernels Hub - Ben Burtenshaw, Hugging Face Master Stage

```
import torch
from transformers import AutoModel
```

2018 → 2026

Eight years ago, we bet on
PyTorch and open models.

We're making the same bet today.

Thank you

Lysandre, Hugging Face