The Good, The Bad, and The Ugly

The Current State of Open Source Large Language Models

550 Vandalia St #231

Saint Paul, MN 55114

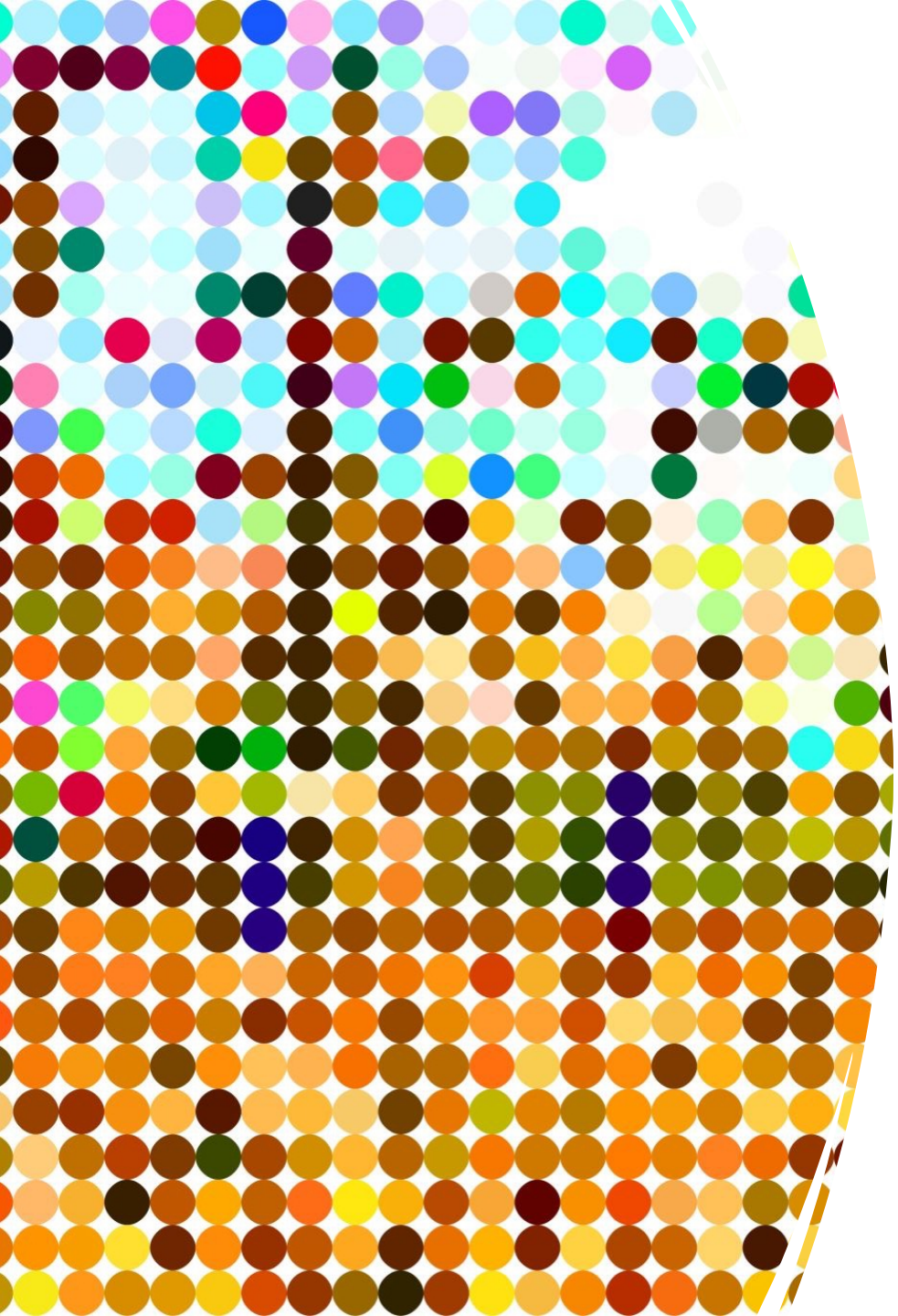
lab651.com

recursiveawesome.com

About Me

- Founder, **Recursive Awesome & Lab651** – AI Consulting & Implementation & Custom Software Development
- Owner, **Applied AI Weekly** – Publication covering AI
- Host of the "**Conversations on Applied AI Podcast**"
- Adjunct Professor, **University of Saint Thomas** – Teaching graduate courses on IoT & AI/ML
- Co-founder, **AppliedAI.MN** – 501(c)(3) non-profit: Monthly meetups, conferences, videos and podcasts on AI/ML
- We'll be having our 6th conference in November!

LAB651 RECURSIVE AWESOME



Agenda

- Large Language Models
- Open-source vs Commercial LLMs
- The Good, The Bad, The Ugly
- Where to Start and How to Decide
- Decision Framework
- Fine Tuning vs RAG
- Hugging Face & Local Applications

Large Language Models

A large language model is a type of artificial neural network that uses deep learning techniques to process, understand, and generate human-like text or many other forms of digital content.

- **Content Generation:** Used for creating articles, reports, stories, images, videos and even poetry
- **Language Translation:** High-quality translations between numerous languages
- **Learning and Research:** Assist in educational and research activities by summarizing information and giving ideas for solutions
- **Role Playing:** Human-like dialogue makes them useful for chatbots and virtual assistant

What Are LLMs NOT Good At?

- **Fixed Knowledge Base:** Knowledge is static and limited to what was available when it was trained.
  - Systems like Retrieval Augmented Generation (RAG) addressing this
- **Understanding Context Depth:** Lack common sense reasoning, so they struggle with domain specific context. Prone to hallucinations!
  - Smaller, Fine-Tuned Models addressing this
- **Conversational Continuity:** Generally don't maintain continuity over long conversations and lose track of previous conversations.
  - Larger context windows and frameworks are helping with this
- **Dependency on User Input:** Relies heavily on the quality and clarity of user queries (prompts).
  - Experience in giving good prompts are helping

What Makes A Model Open Source

- Accessible Source Code & Weights
  - Source code shows how the model works
  - Weights are the internal variables learned during the training process
- Permissive License
  - Freely allows for modification and redistribution
  - Download and use on your own with zero cost
- Community Driven & Transparency
  - Development involves contributions from a community
  - Updates, bug fixes, and enhancements to the model should be done in a transparent manner

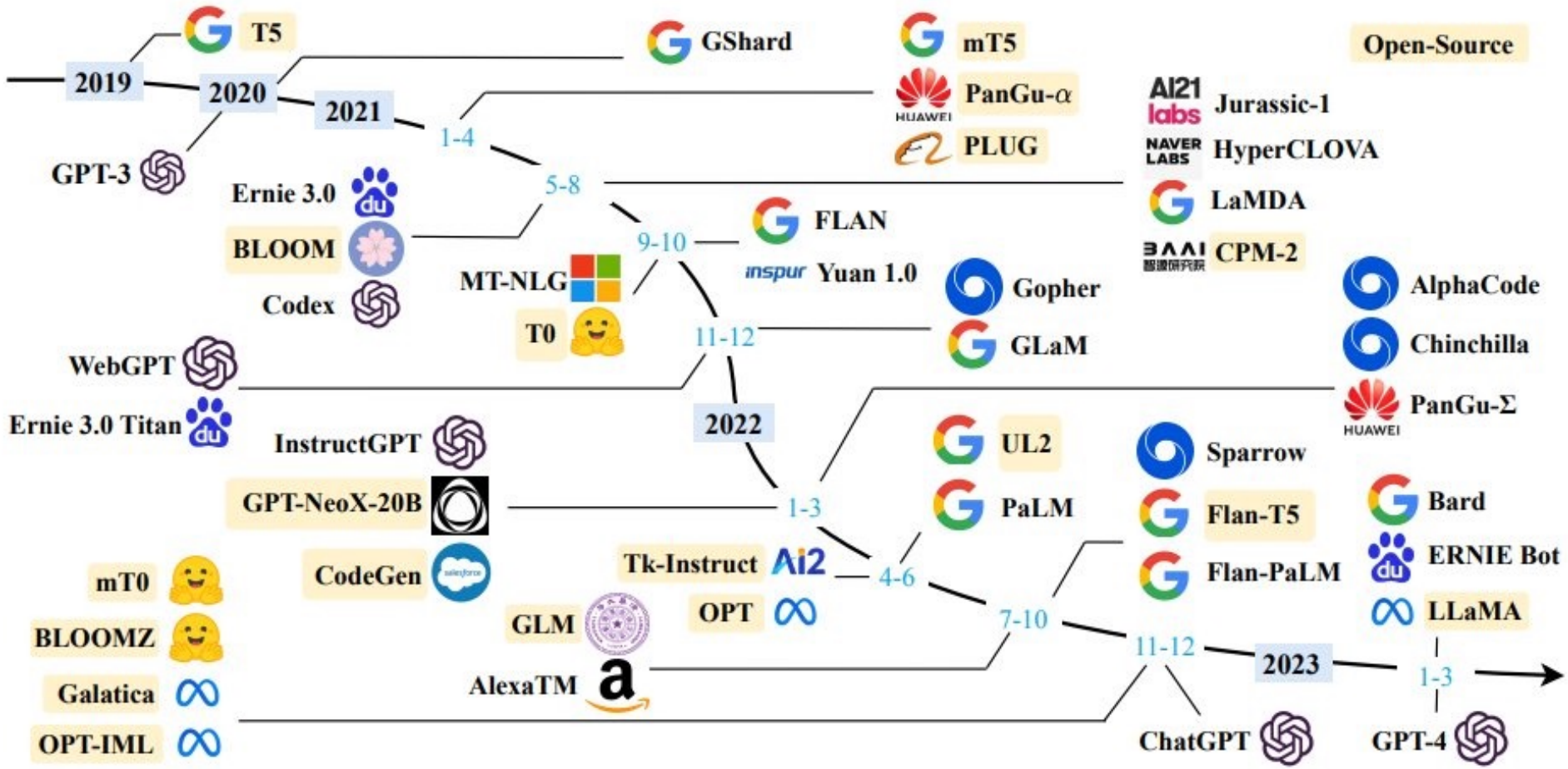
LLM Landscape

NON-EXHAUSTIVE

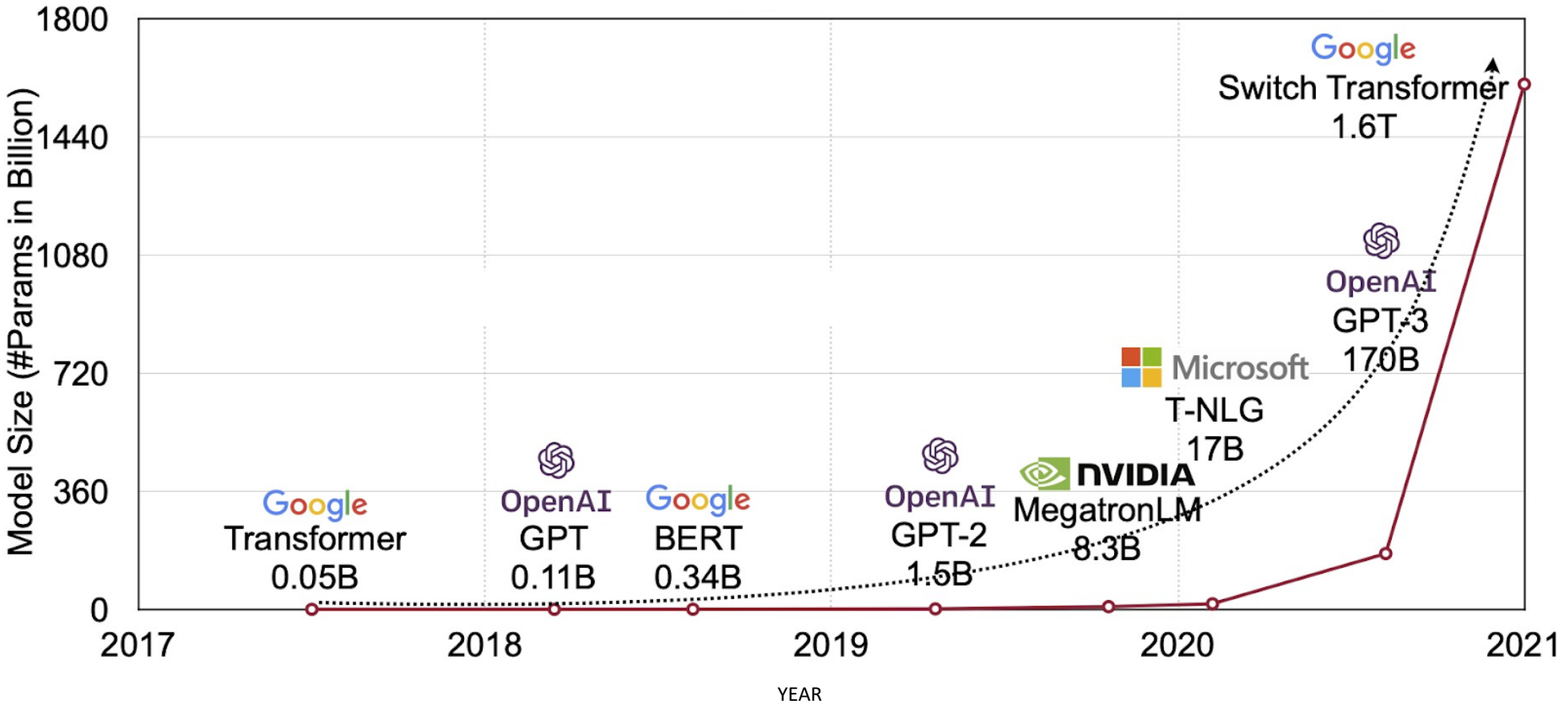
	Google DeepMind	amazon	Microsoft OpenAI	Meta	NVIDIA	stability.ai	Startups, researchers, OSS contributors
Text / language	**BERT**, **Bard (LaMDA)**, PaLM, GLaM, Chinchilla, Gopher	**AlexaTM, Titan LLMs**	**MT-NLG**, **GPT-4 / ChatGPT (GPT-3.5)**	**OPT-175B, LLaMA**	**MT-NLG**		**Dolly**, **Claude**, **Jurassic-1**, **Cohere**, **GPT-J**, **GPT-NeoX**, **BLOOM**, Wu Dao 2.0, **CTRL**, **Alpaca**
Code	AlphaCode	**Code-Whisperer**	**Copilot**, **CodeBERT**, **Codex**				**Replit**, **Polycoder**, **CodeGen**, **CodeT5**
Image	Imagen, **Parti**, **DreamBooth**		**Dall-E 2**, **CLIP**, **NUWA-Infinity**	Make-a-scene, **Dinov2**	**SPADE**	**Stable Diffusion**	**Midjourney**, **Waifu Diffusion**
Speech & Music	WaveNet, MusicLM		**Muzic**, **Whisper**, **Jukebox**			**Dance Diffusion**	**WaveGAN**
Video	Imagen Video, Phenaki		**NUWA-Infinity**	Make-a-video			**CogVideo**
3D	DreamFusion, 3DiM		**Point-E**		**Get3D**, Magic3D		**Motion Diffusion Model**

Proprietary
Open source
Private

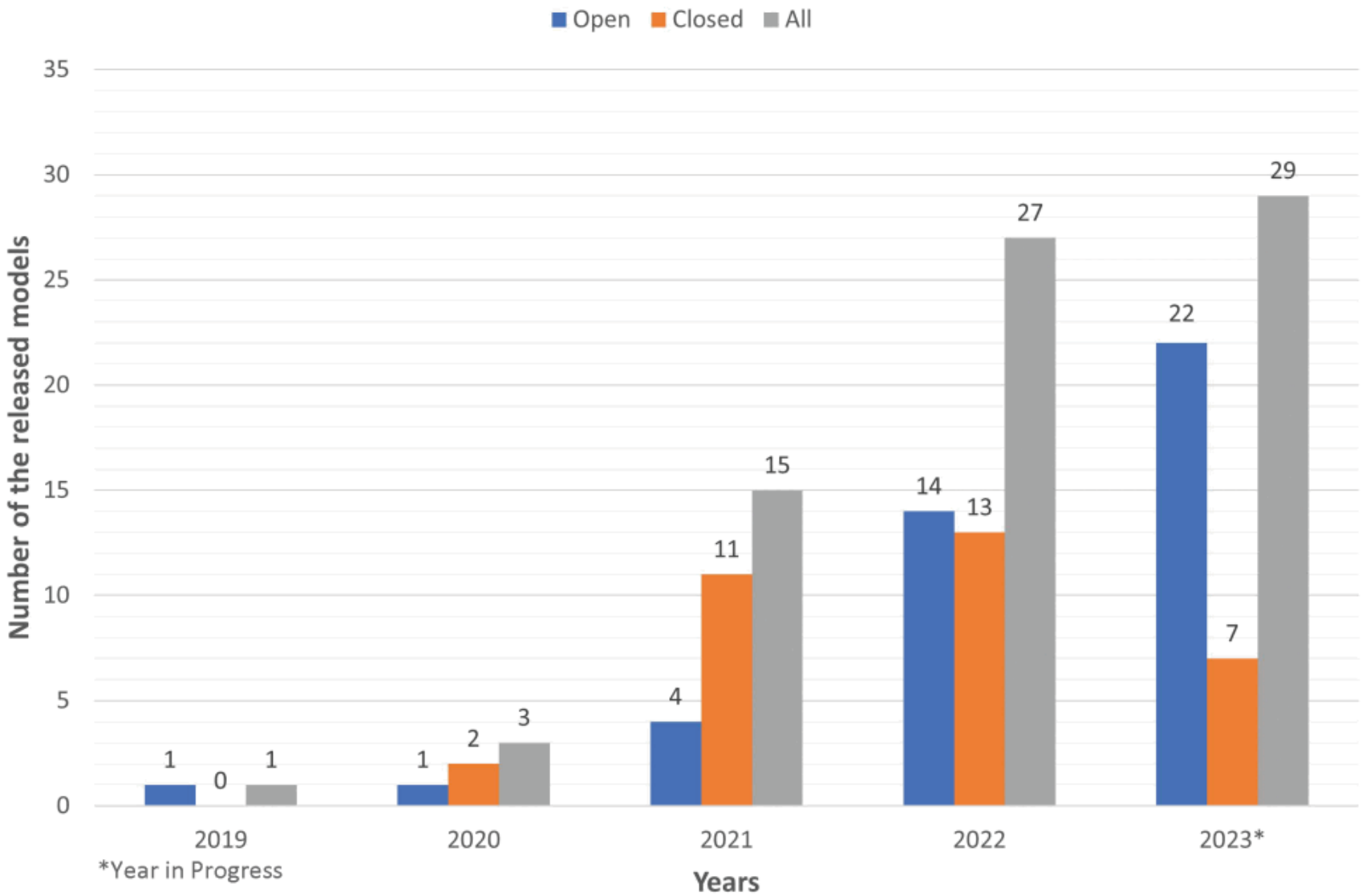
LLM Landscape



Model Size is Getting Larger



Open Source LLMs will get larger too!



Source: https://blog.n8n.io/open-source-llm/


50th ANNIVERSARY EDITION
CLINT EASTWOOD
THE GOOD THE BAD and THE UGLY
co-starring
LEE VAN CLEEF
also starring
ELI WALLACH
in the role of TUCO
directed by
SERGIO LEONE

The Good

- **Control and Customization**
  - Allows users to access the underlying code, architecture, and pre-trained weights, enabling them to customize and optimize the models for their specific needs.
- **Cost-Effectiveness**
  - Open-source LLMs are generally free or low-cost, making them more accessible, especially for small businesses and individual developers.
- **Transparency and Trust**
  - The open-source nature of these models provides full visibility into their inner workings, allowing users to inspect the code, identify vulnerabilities, and build trust.

The Bad

- **Computational Intensity**
  - Open-source models often require significant computational and operational resources for training and operation
- **Lack of Official Support**
  - Open-source LLMs often rely on volunteer contributions, which can lead to fewer resources for development, bug-fixing, and performance optimization
- **Integration Challenges**
  - Integrating open-source LLMs into existing systems can raise compatibility issues, lack of standardized APIs, and other fragmentation

The Ugly

- **Security Risks**
  - Open-source models are exposed to a broader community which includes the potential for malicious contributions or exploits.
- **Ethical and Legal Concerns**
  - There's a higher risk of incorporating biased data or using the model in ways that were not intended by its creators, raising ethical concerns.
- **Licensing**
  - Understanding and complying with the terms of open-source licenses

Framework for Open vs Proprietary

- **Define Business Objectives and Requirements**
  - Purpose: What are you aiming to achieve? Can you do it without an LLM?
  - Requirements: Performance, Scalability, Integration, Languages to Support
- **Evaluate Technical Considerations**
  - Customization
  - Support & Maintenance
- **Assess Costs**
  - Initial cost & ongoing expenses

Framework for Open vs Proprietary

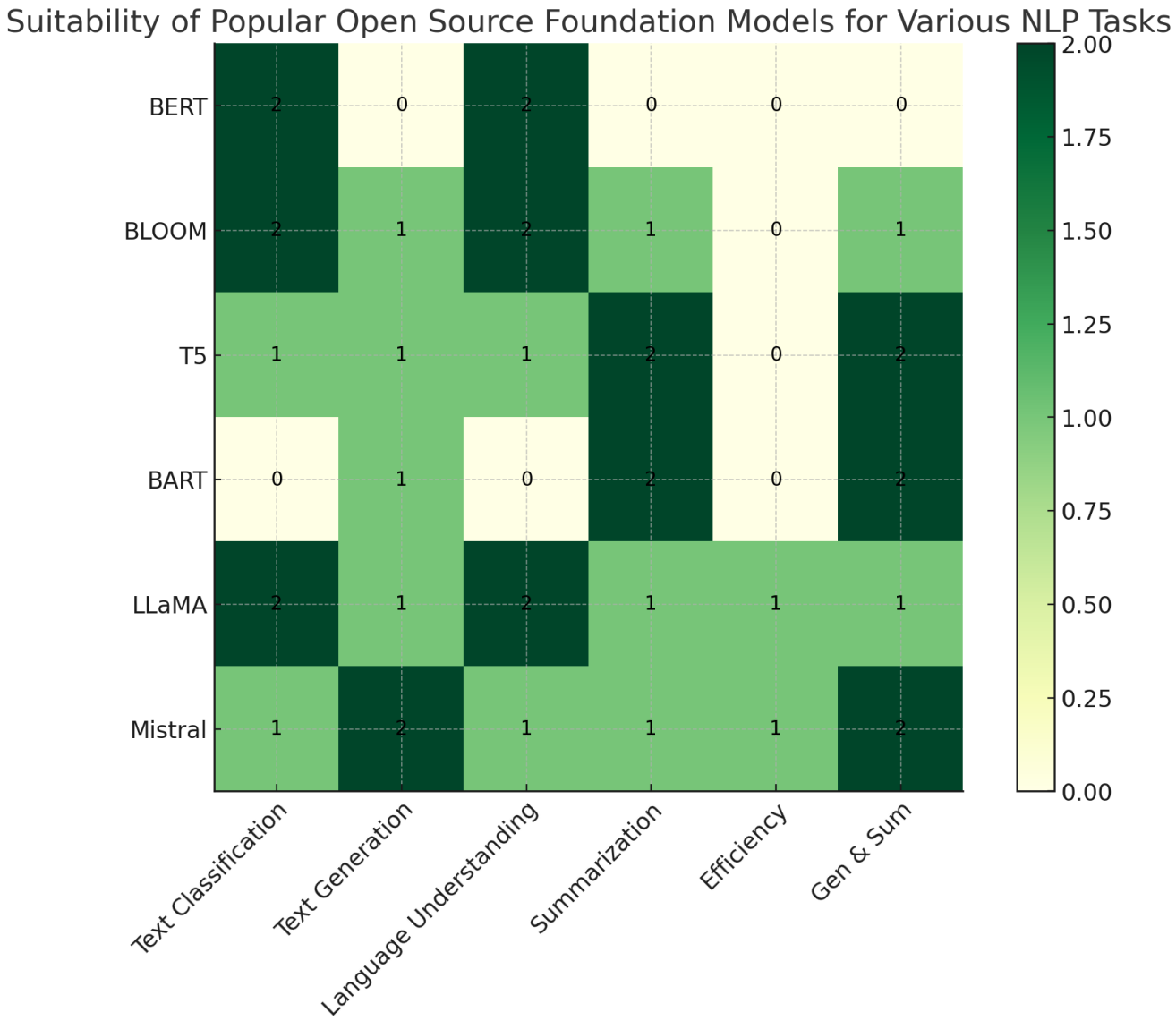
- **Legal and Compliance**
  - Understand the IP and licensing terms
  - Data privacy
- **Risks**
  - Dependency on vendor lock-in
  - Assess the stability of the community
- **Scalability & Performance**
  - Benchmarks
  - Future needs
- **Consult with Experts in the Field**

Sample Decision Matrix

Factors	Importance (1-10)	Open Source Rating (1-5)	Commercial Rating (1-5)	Notes
Costs				
Initial Setup Costs	8	2	5	Open source may require more customization and setup.
Ongoing Maintenance Costs	7	2	5	Commercial options include support and maintenance.
Technical Needs				
Customization	9	5	3	Open source offers greater flexibility.
Integration Ease	6	3	4	Commercial products might integrate more easily with existing systems.
Performance	8	4	4	Depends on specific model capabilities.
Risks				
Vendor Dependence	5	5	2	Less risk of dependence with open source.
Community Support Stability	4	2	N/A	Riskier if open source if community is small.
Legal and Compliance				
IP and Licensing	7	4	5	Open source offers more freedom but can be complex to navigate.
Data Privacy	8	3	5	Commercial vendors often provide clearer data handling policies.
Scalability	7	4	5	Commercial models may scale more readily with less technical overhead.

How to Decide?

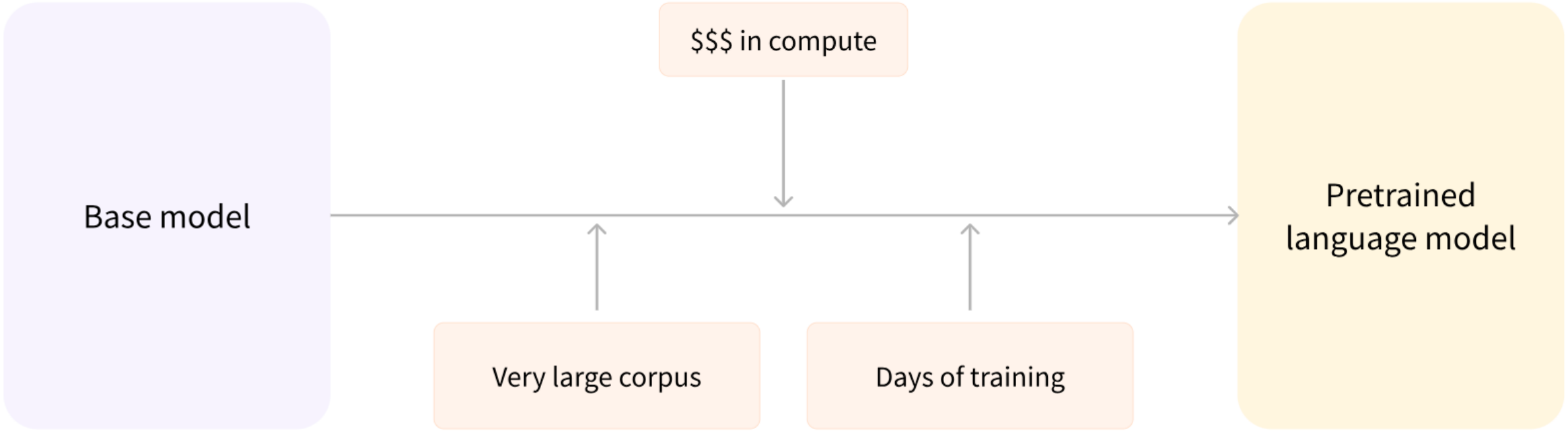
Know your use case and the LLMs strengths weaknesses



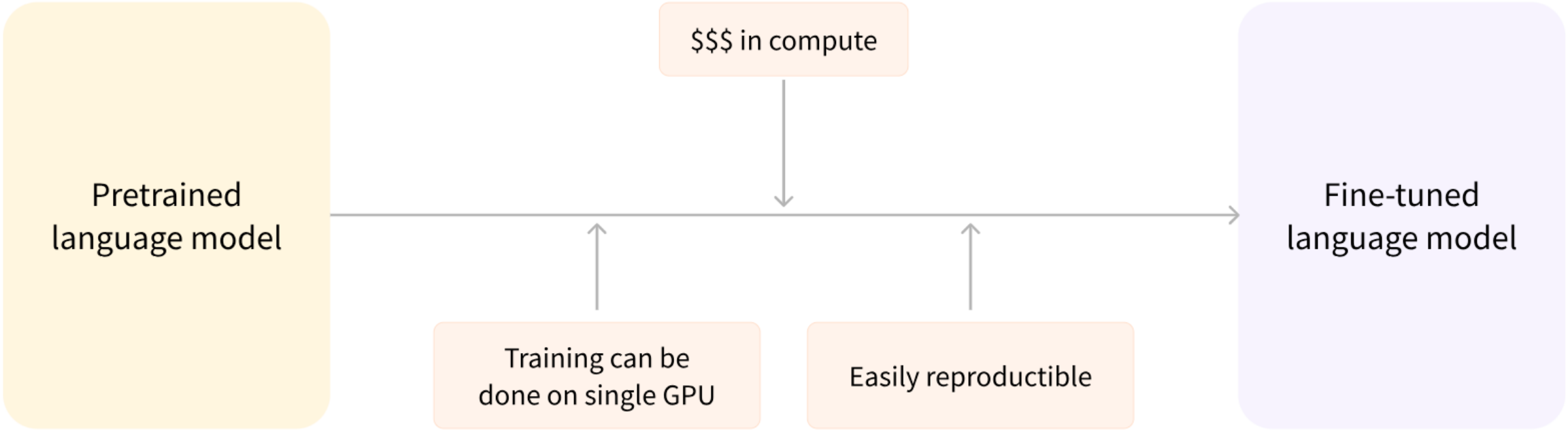
We've Chosen Open Source…

Where to now?

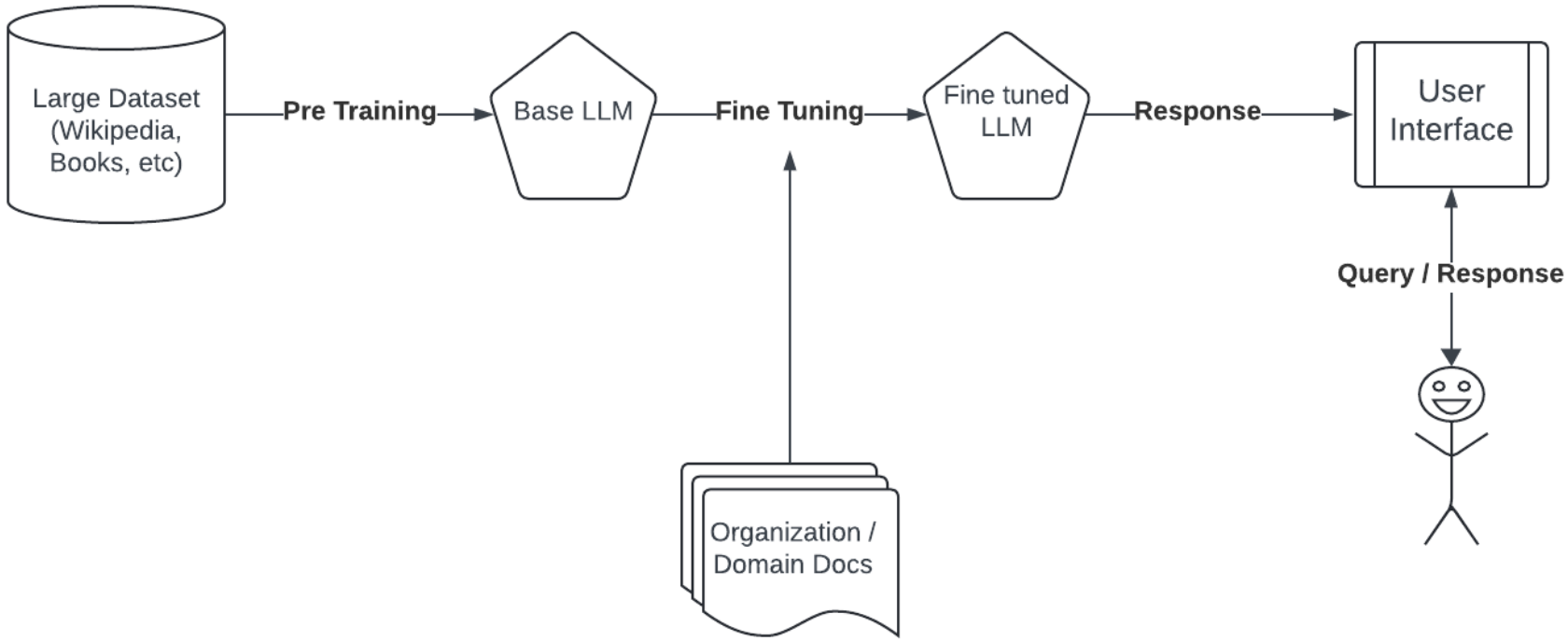
Pretrained Models



Fine-Tuned Models



Fine Tuning


Large Dataset
(Wikipedia,
Books, etc)
Pre Training
Base LLM
Fine Tuning
Fine tuned
LLM
Response
User
Interface
Query / Response
Organization /
Domain Docs

Model Training Analogy

- Pre-Training Phase
  - New employee understand the general idea of your business
  - Been trained and learned vast amount of general information
  - **Model:** Learns general patterns & relationships
- Fine-Tuning Phase
  - Employee learns to become an expert in your business
  - **Model:** Specialized in the data of your business use case
- Takeaways from Fine Tuning
  - **Employee & Model:** Effective, Efficient and Customized

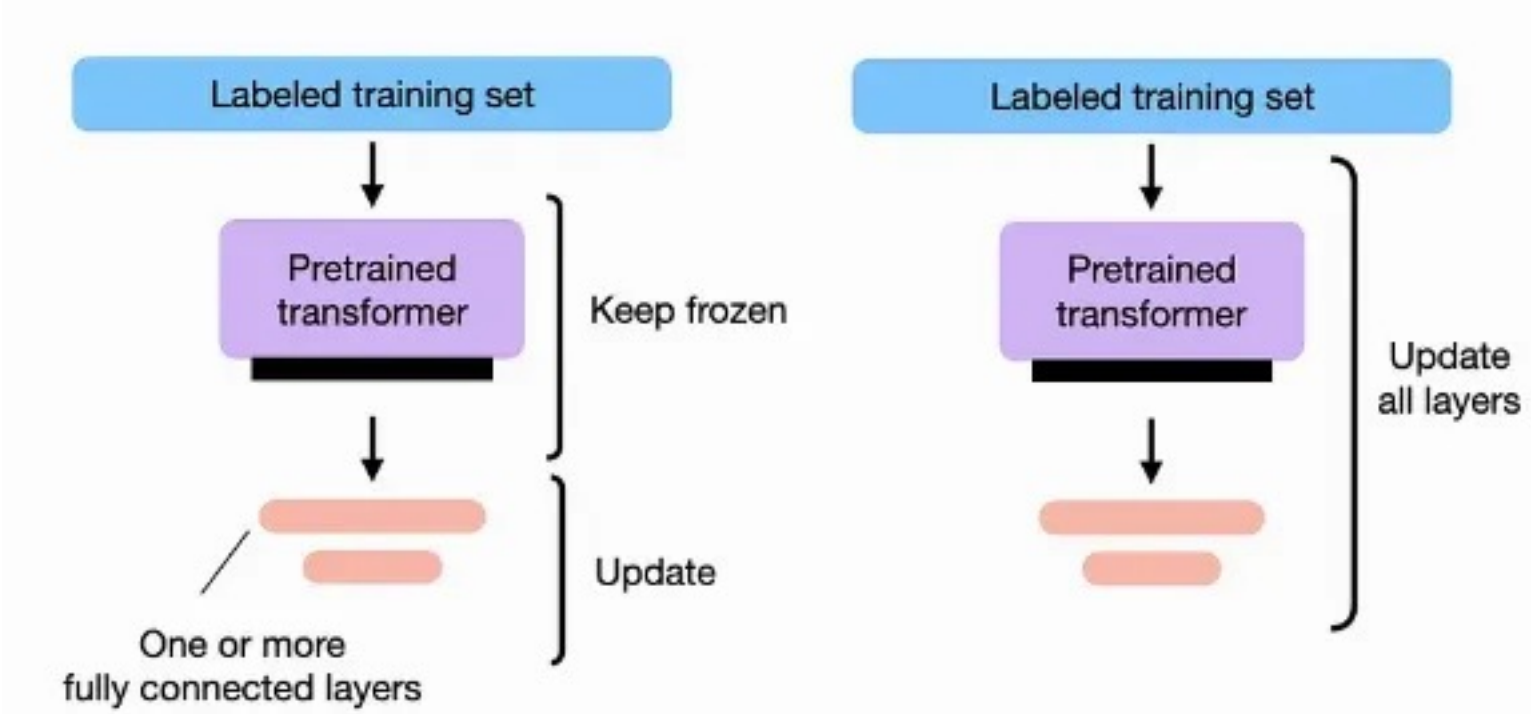
Challenges When Fine Tuning

- Data Requirements
  - Requires large amount of clean and accurate domain specific data
- Computational Cost
  - Requires powerful hardware and processing time and memory
- Specialization Risk
  - The model may become too specialized on the examples
- Continuous Updates
  - Requires regular updates to remain relevant and accurate.

Fine Tuning Steps & Techniques

Parameter Efficient Fine Tuning (PEFT)

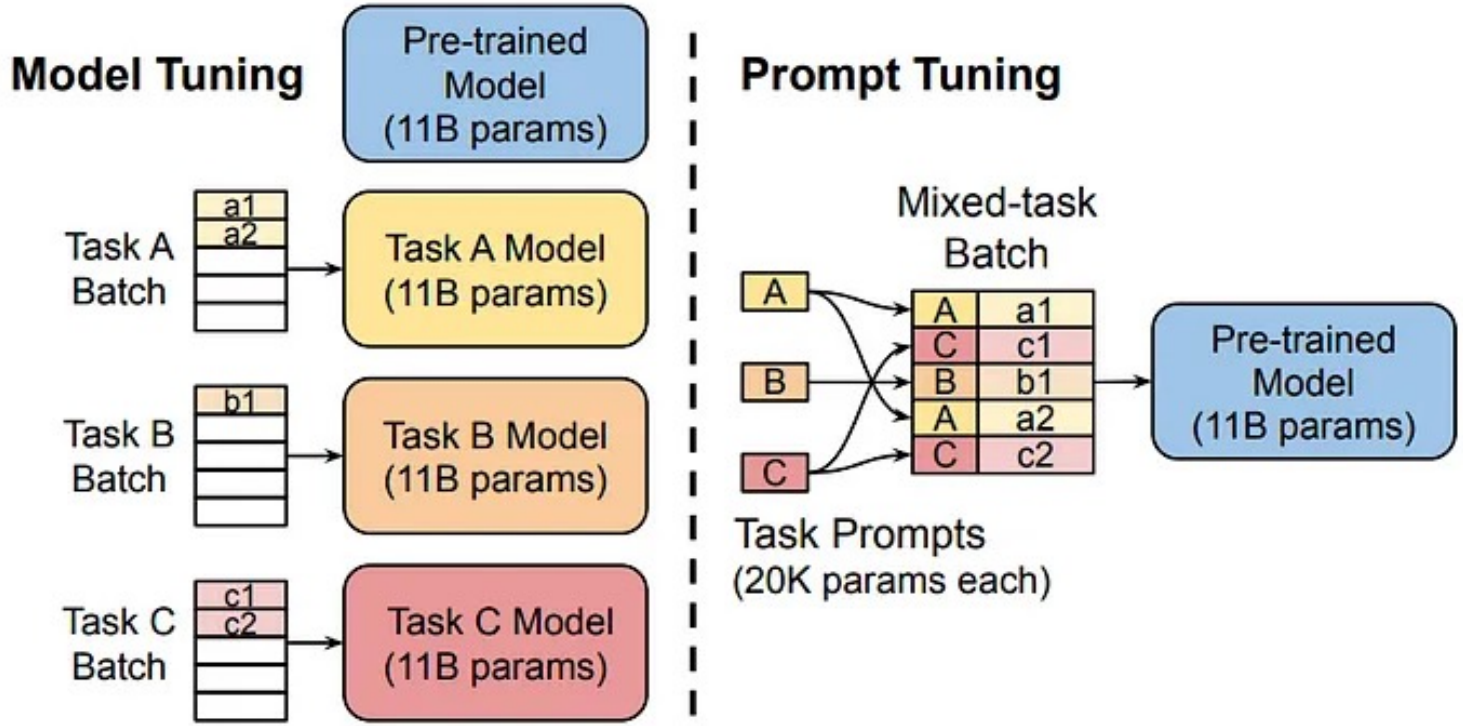
- Low Rank Adaptation (LoRA)
  - Freeze the model weights
  - Train on a set of smaller matrices for your task
  - Much faster to train



Fine Tuning Steps & Techniques

Prompt Tuning

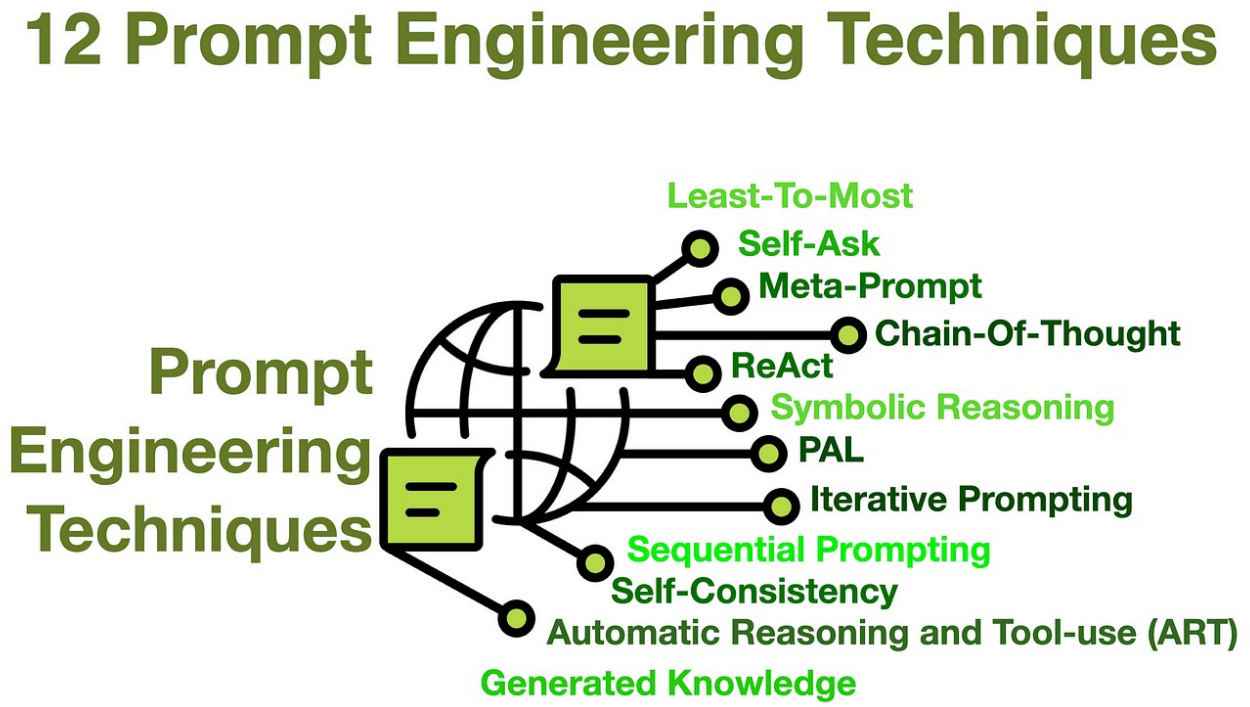
- Create “soft prompts” for the input
- Pass prompts in and evaluate output
- Don’t adjust weights, just prompts



Fine Tuning Steps & Techniques

Prompt Engineering

- No training or retraining
- User designing prompts for the model
- No additional resources needed

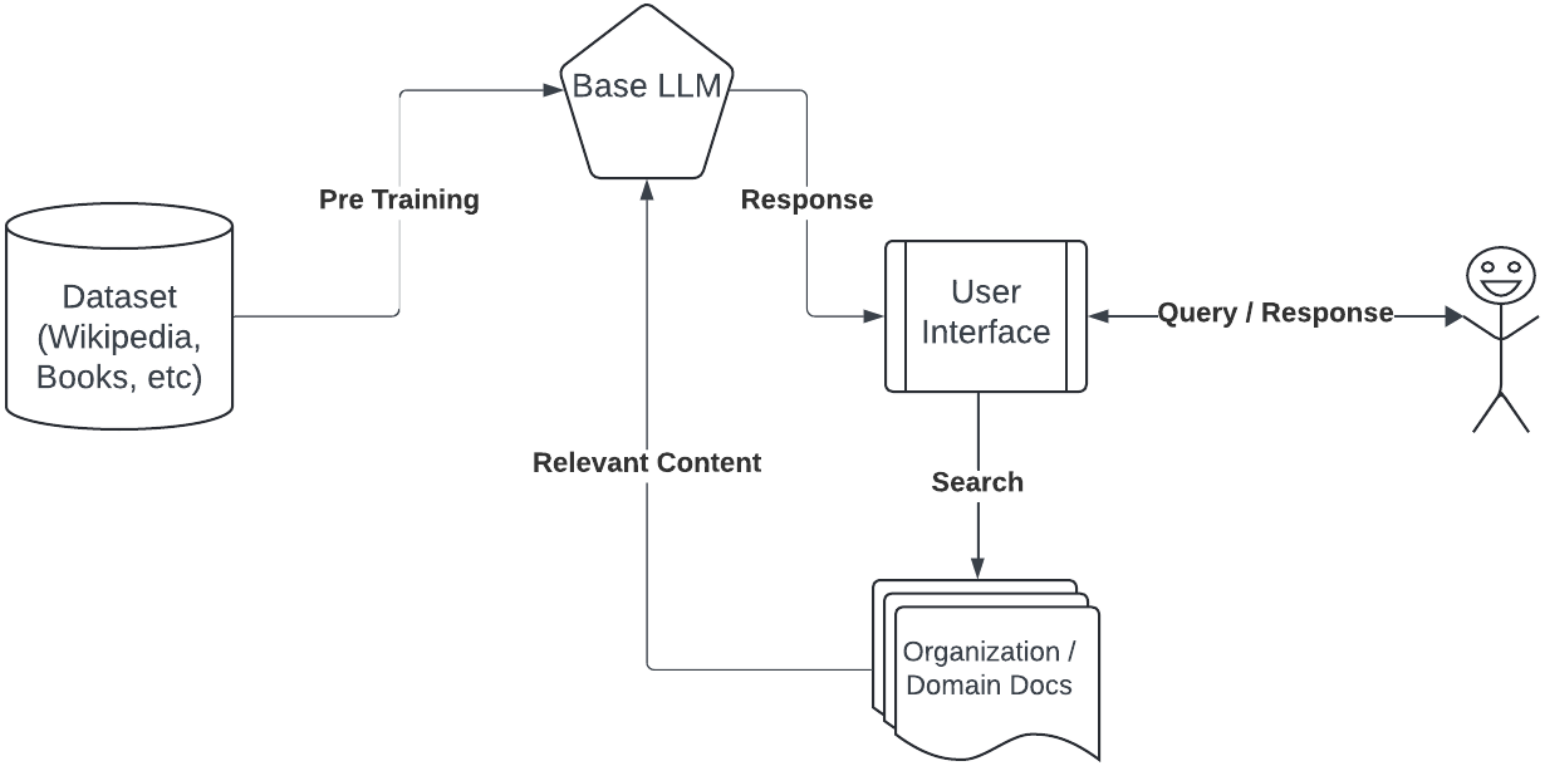


When To Choose Which Fine-Tuning Method

Method	Resource Intensity	Training Required	Best For
Fine-Tuning	High	Yes	Tasks requiring deep model customization
Prompt Tuning	Low	Yes	Maintaining model integrity across tasks
Prompt Engineering	None	No	Quick adaptations with no computational cost.


BUT THERE IS
ANOTHER WAY!

Retrieval Augmented Generation

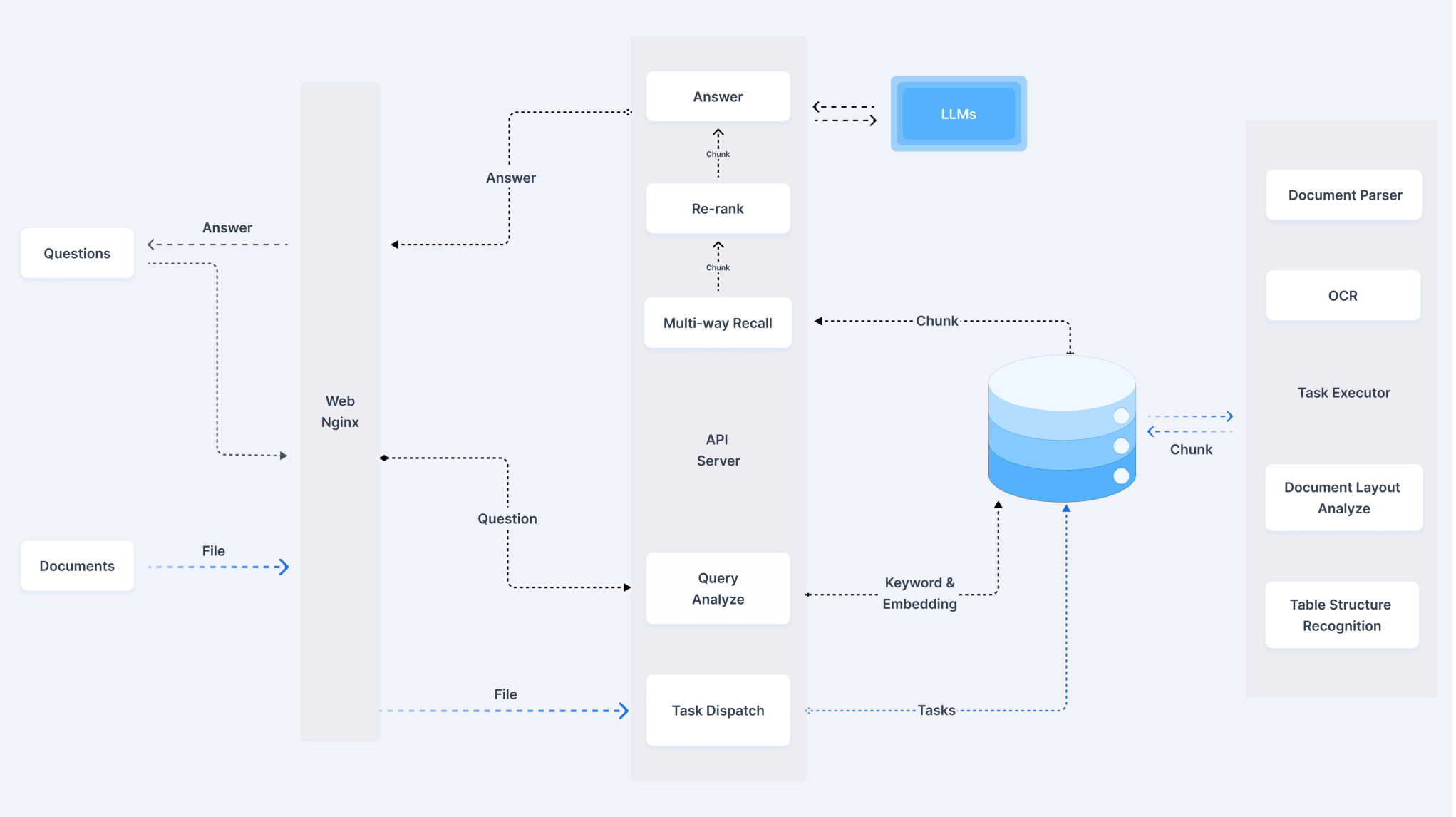


Benefits of Retrieval Augmented Generation

- Optimizing the output based on authoritative knowledge base
- LLM remains relevant, accurate & useful
- More cost effective than retraining or fine-tuning
- Attribute the source more easily
- Create access controls to sensitive data

Checkout RagFlow

- Open-source RAG (Retrieval-Augmented Generation) engine based on deep document understanding.
- Visualization of text chunking to allow human intervention.
- Supports Word, slides, excel, txt, images, scanned copies, structured data, web pages, and more.
- Configurable LLMs as well as embedding models.
- Intuitive APIs for seamless integration with business.



https://github.com/infiniflow/ragflow

When to Use RAG?

- Up-to-Date Information: Current information about recent events or finding specific details from a vast database.
- Broad Knowledge: If you want the model to handle a wide variety of topics without deep specialization.

When to Use Fine Tuning?

- Deep Expertise: medicine, law, software code, or a specific industry.
- Specific Tasks: Perform specific tasks consistently, such as understanding legal documents or generating marketing copy in a specific style.

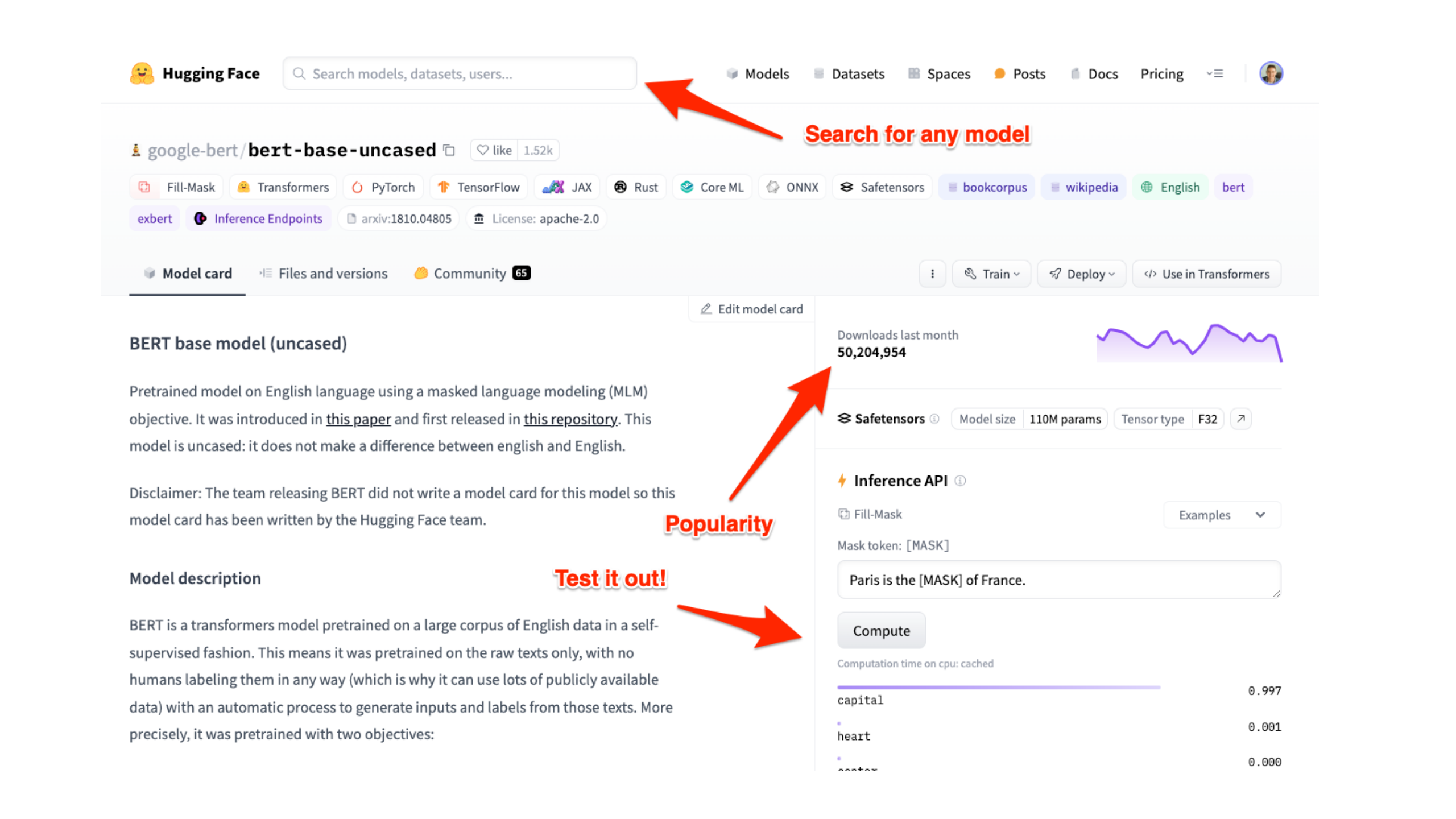
We've Chosen a Model…

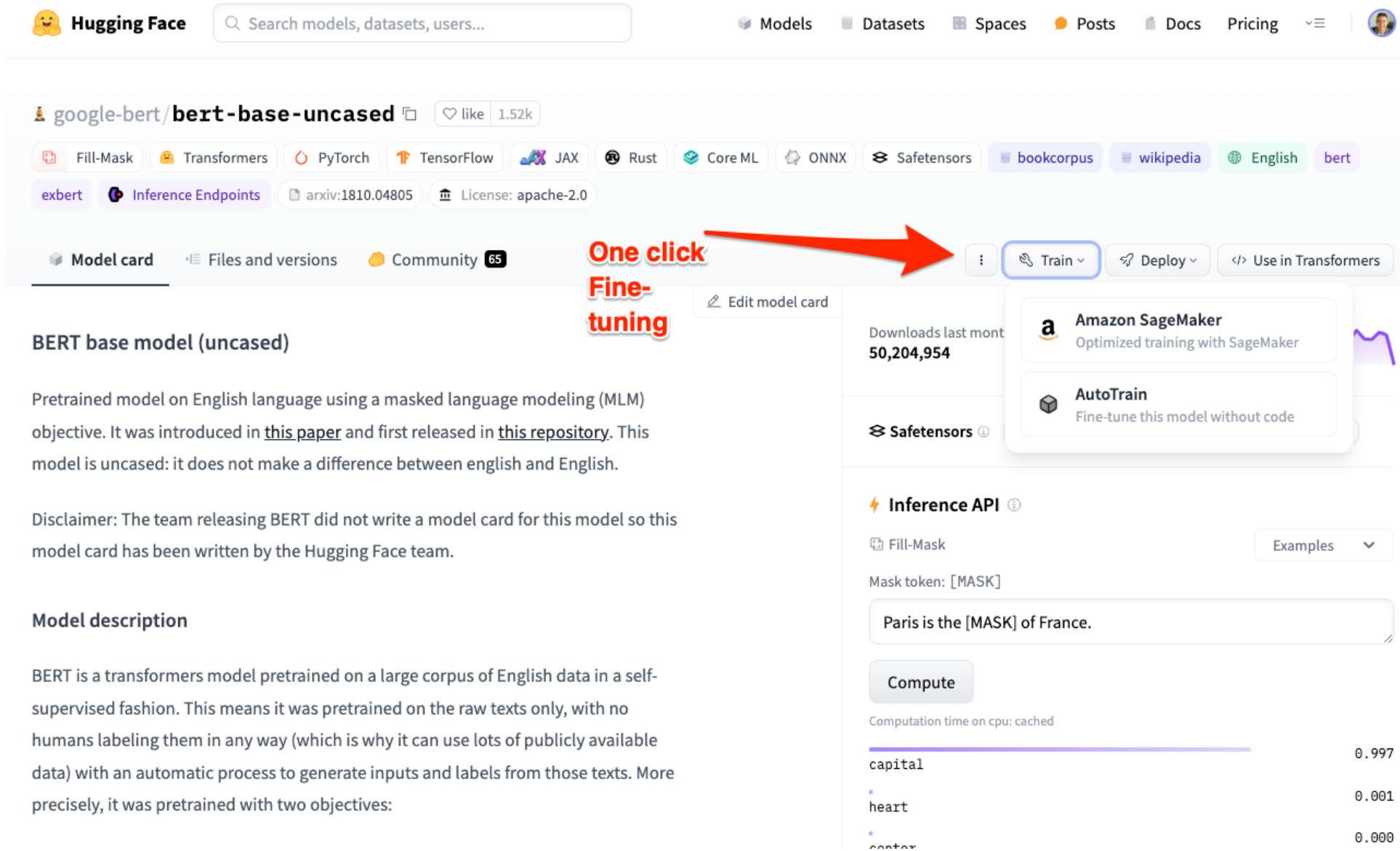
Where Do I Go Now?

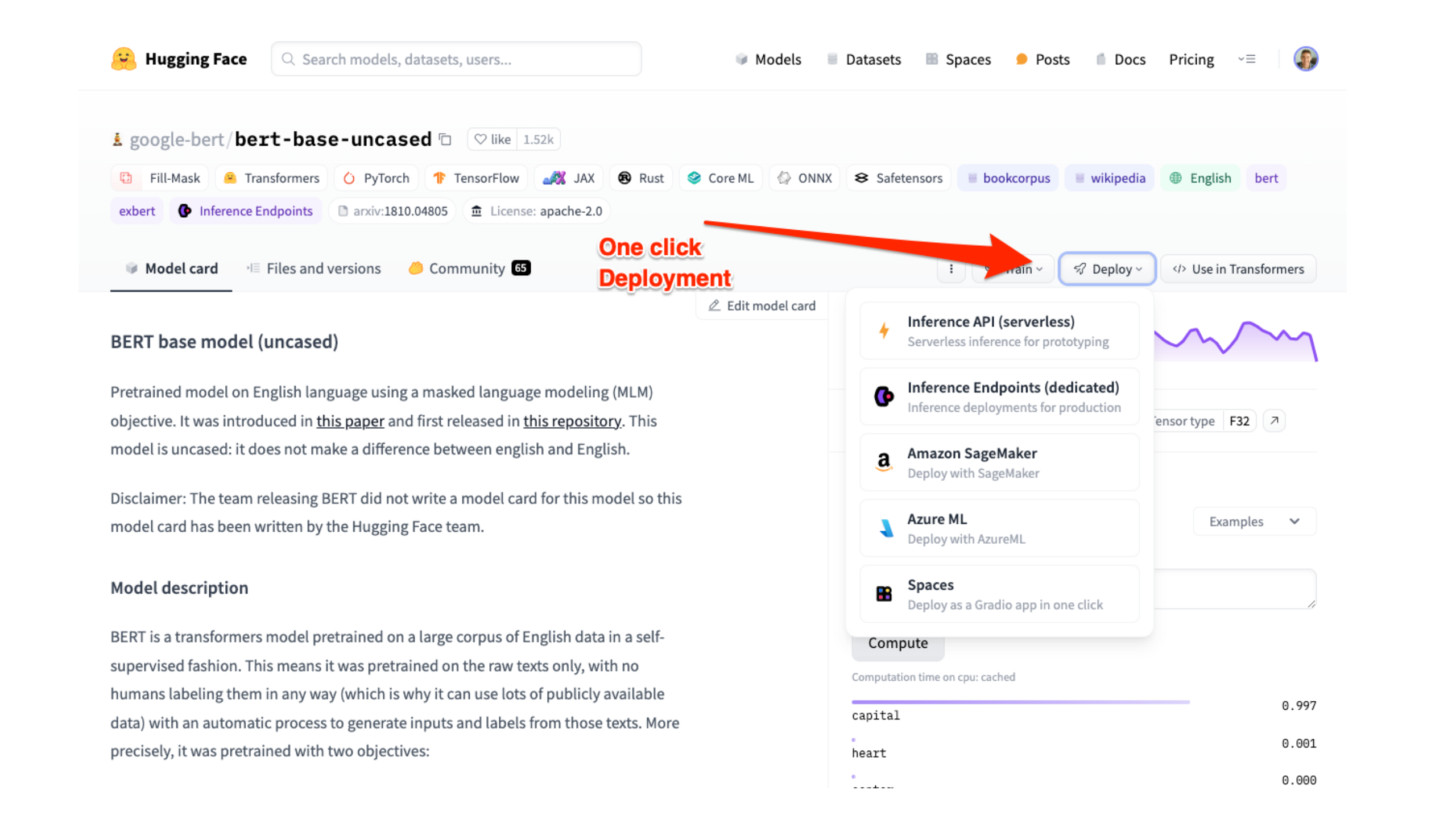
Hugging Face

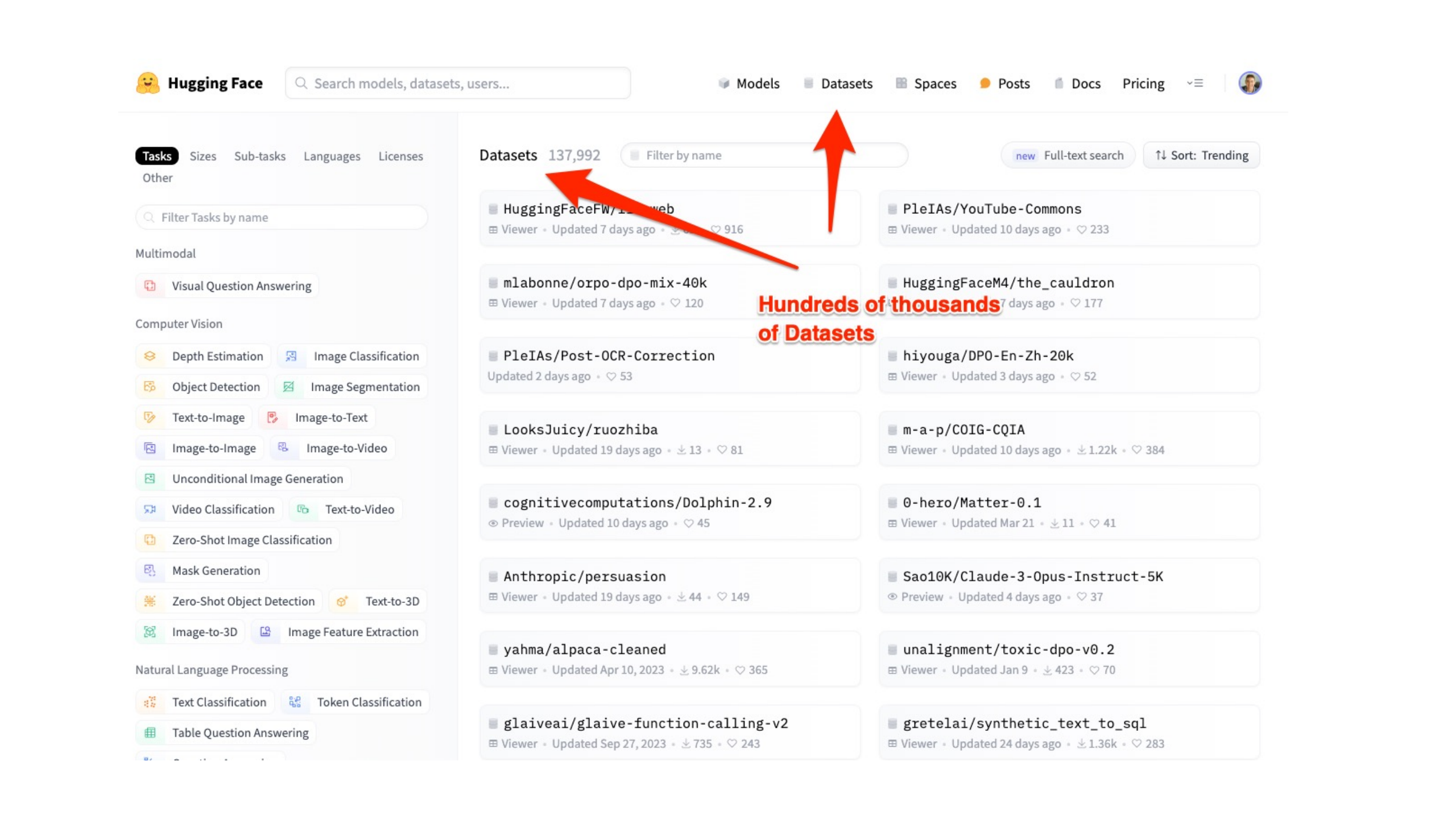
- Repository of Pre-Trained Open-Source Models
- Fine-Tuning and Deployments of Models
- Value To You?
  - Easy Access
  - Rapid Prototyping
  - Community Support
  - Educational Resources
  - Commercial Solutions

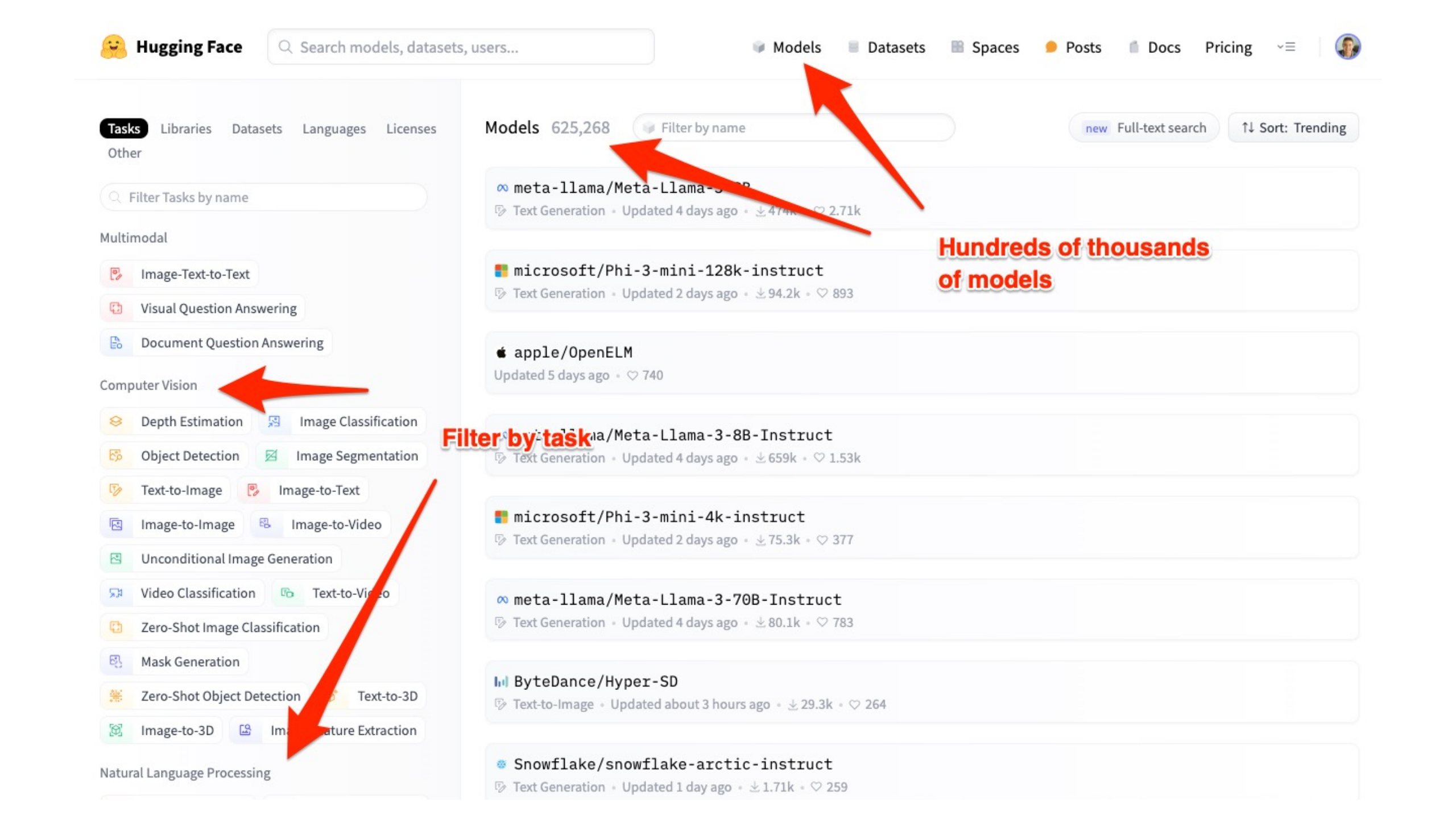
LAB651 RECURSIVE AWESOME

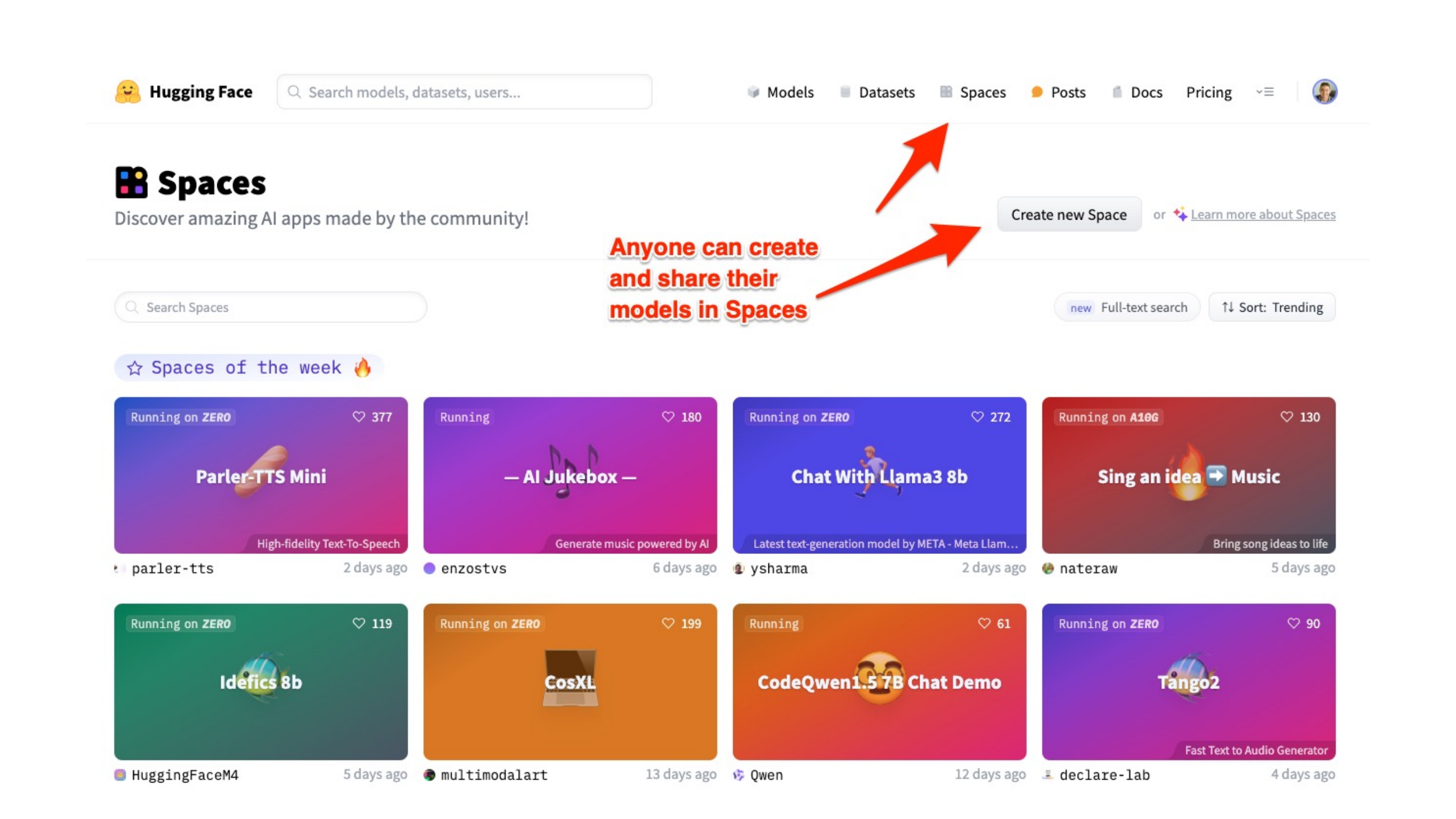

Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Search for any model
google-bert/bert-base-uncased
like 1.52k
Fill-Mask
Transformers
PyTorch
TensorFlow
JAX
Rust
Core ML
ONNX
Safetensors
bookcorpus
wikipedia
English
bert
exbert
Inference Endpoints
arxiv:1810.04805
License: apache-2.0
Model card
Files and versions
Community 65
Train
Deploy
Use in Transformers
Edit model card
BERT base model (uncased)
Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. This model is uncased: it does not make a difference between english and English.
Disclaimer: The team releasing BERT did not write a model card for this model so this model card has been written by the Hugging Face team.
Popularity
Model description
Test it out!
BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:
Downloads last month
50,204,954
Safetensors
Model size 110M params
Tensor type F32
Inference API
Fill-Mask
Examples
Mask token: [MASK]
Paris is the [MASK] of France.
Compute
Computation time on cpu: cached
capital 0.997
heart 0.001
0.000


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
google-bert/bert-base-uncased
like 1.52k
Fill-Mask
Transformers
PyTorch
TensorFlow
JAX
Rust
Core ML
ONNX
Safetensors
bookcorpus
wikipedia
English
bert
exbert
Inference Endpoints
arxiv:1810.04805
License: apache-2.0
Model card
Files and versions
Community 65
One click Fine-tuning
Train
Deploy
Use in Transformers
Edit model card
Amazon SageMaker
Optimized training with SageMaker
AutoTrain
Fine-tune this model without code
Downloads last mont
50,204,954
Safetensors
BERT base model (uncased)
Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. This model is uncased: it does not make a difference between english and English.
Disclaimer: The team releasing BERT did not write a model card for this model so this model card has been written by the Hugging Face team.
Model description
BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:
Inference API
Fill-Mask
Examples
Mask token: [MASK]
Paris is the [MASK] of France.
Compute
Computation time on cpu: cached
capital 0.997
heart 0.001
0.000


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
google-bert / bert-base-uncased
like 1.52k
Fill-Mask
Transformers
PyTorch
TensorFlow
JAX
Rust
Core ML
ONNX
Safetensors
bookcorpus
wikipedia
English
bert
exbert
Inference Endpoints
arxiv:1810.04805
License: apache-2.0
One click Deployment
Model card
Files and versions
Community 65
Deploy
Use in Transformers
Edit model card
Inference API (serverless)
Serverless inference for prototyping
Inference Endpoints (dedicated)
Inference deployments for production
Amazon SageMaker
Deploy with SageMaker
Azure ML
Deploy with AzureML
Spaces
Deploy as a Gradio app in one click
Tensor type
F32
Examples
Compute
Computation time on cpu: cached
capital
0.997
heart
0.001
0.000
BERT base model (uncased)
Pretrained model on English language using a masked language modeling (MLM) objective. It was introduced in this paper and first released in this repository. This model is uncased: it does not make a difference between english and English.
Disclaimer: The team releasing BERT did not write a model card for this model so this model card has been written by the Hugging Face team.
Model description
BERT is a transformers model pretrained on a large corpus of English data in a self-supervised fashion. This means it was pretrained on the raw texts only, with no humans labeling them in any way (which is why it can use lots of publicly available data) with an automatic process to generate inputs and labels from those texts. More precisely, it was pretrained with two objectives:


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Tasks
Sizes
Sub-tasks
Languages
Licenses
Other
Filter Tasks by name
Multimodal
Visual Question Answering
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Text-to-Image
Image-to-Text
Image-to-Image
Image-to-Video
Unconditional Image Generation
Video Classification
Text-to-Video
Zero-Shot Image Classification
Mask Generation
Zero-Shot Object Detection
Text-to-3D
Image-to-3D
Image Feature Extraction
Natural Language Processing
Text Classification
Token Classification
Table Question Answering
Datasets 137,992
Filter by name
new Full-text search
Sort: Trending
PleIAs/YouTube-Commons
Viewer • Updated 10 days ago • 233
mlabonne/orpo-dpo-mix-40k
Viewer • Updated 7 days ago • 120
HuggingFaceM4/the_cauldron
177
Hundreds of thousands
of Datasets
PleIAs/Post-OCR-Correction
Updated 2 days ago • 53
hiyouga/DPO-En-Zh-20k
Viewer • Updated 3 days ago • 52
LooksJuicy/ruozhiba
Viewer • Updated 19 days ago • 13 • 81
m-a-p/COIG-CQIA
Viewer • Updated 10 days ago • 1.22k • 384
cognitivecomputations/Dolphin-2.9
Preview • Updated 10 days ago • 45
0-hero/Matter-0.1
Viewer • Updated Mar 21 • 11 • 41
Anthropic/persuasion
Viewer • Updated 19 days ago • 44 • 149
Sao10K/Claude-3-Opus-Instruct-5K
Preview • Updated 4 days ago • 37
yahma/alpaca-cleaned
Viewer • Updated Apr 10, 2023 • 9.62k • 365
unalignment/toxic-dpo-v0.2
Viewer • Updated Jan 9 • 423 • 70
glaiveai/glaive-function-calling-v2
Viewer • Updated Sep 27, 2023 • 735 • 243
gretelai/synthetic_text_to_sql
Viewer • Updated 24 days ago • 1.36k • 283


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Tasks
Libraries
Datasets
Languages
Licenses
Other
Filter Tasks by name
Multimodal
Image-Text-to-Text
Visual Question Answering
Document Question Answering
Computer Vision
Depth Estimation
Image Classification
Object Detection
Image Segmentation
Text-to-Image
Image-to-Text
Image-to-Image
Image-to-Video
Unconditional Image Generation
Video Classification
Zero-Shot Image Classification
Mask Generation
Zero-Shot Object Detection
Text-to-3D
Image-to-3D
Natural Language Processing
Filter by task
Models 625,268
Filter by name
new Full-text search
Sort: Trending
Text Generation • Updated 4 days ago • 2.71k
Hundreds of thousands of models
microsoft/Phi-3-mini-128k-instruct
Text Generation • Updated 2 days ago • 94.2k • 893
apple/OpenELM
Updated 5 days ago • 740
Text Generation • Updated 4 days ago • 659k • 1.53k
microsoft/Phi-3-mini-4k-instruct
Text Generation • Updated 2 days ago • 75.3k • 377
meta-llama/Meta-Llama-3-70B-Instruct
Text Generation • Updated 4 days ago • 80.1k • 783
ByteDance/Hyper-SD
Text-to-Image • Updated about 3 hours ago • 29.3k • 264
Snowflake/snowflake-arctic-instruct
Text Generation • Updated 1 day ago • 1.71k • 259


Hugging Face
Search models, datasets, users...
Models
Datasets
Spaces
Posts
Docs
Pricing
Spaces
Discover amazing AI apps made by the community!
Create new Space
or Learn more about Spaces
Anyone can create and share their models in Spaces
Search Spaces
new Full-text search
Sort: Trending
Spaces of the week
Running on ZERO 377
Parler-TTS Mini
High-fidelity Text-To-Speech
parler-tts 2 days ago
Running 180
— AI Jukebox —
Generate music powered by AI
enzostvs 6 days ago
Running on ZERO 272
Chat With Llama3 8b
Latest text-generation model by META - Meta Llam...
ysharma 2 days ago
Running on A10G 130
Sing an idea ➡️ Music
Bring song ideas to life
nateraw 5 days ago
Running on ZERO 119
Idefics 8b
HuggingFaceM4 5 days ago
Running on ZERO 199
CosXL
multimodalart 13 days ago
Running 61
CodeQwen1.5 7B Chat Demo
Qwen 12 days ago
Running on ZERO 90
Tango2
Fast Text to Audio Generator
declare-lab 4 days ago

Spaces | HuggingFaceH4 / open_llm_leaderboard | like 9.6k | Running on CPU UPGRADE | App | Files | Community 726

Open LLM Leaderboard

Best models across a variety of metrics

LLM Benchmark | Metrics through time | About | FAQ | Submit

Search models or licenses (e.g., 'model_name; license: MIT') and press ENTER...

Select columns to show

- [x] Average ⬆️
- [x] ARC
- [x] HellaSwag
- [x] MMLU
- [x] TruthfulQA
- [x] Winogrande
- [x] GSM8K
- [ ] Type
- [ ] Architecture
- [ ] Precision
- [ ] Merged
- [ ] Hub License
- [ ] #Params (B)
- [ ] Hub ❤️
- [ ] Model sha

Hide models

- [x] Private or deleted
- [x] Contains a merge/moerge
- [x] Flagged
- [ ] MoE

Model types

- [x] 🟢 pretrained
- [x] 🟩 continuously pretrained
- [x] 🔶 fine-tuned on domain-specific datasets
- [x] 💬 chat models (RLHF, DPO, IFT, ...)
- [x] 🤝 base merges and moerges
- [x] ?

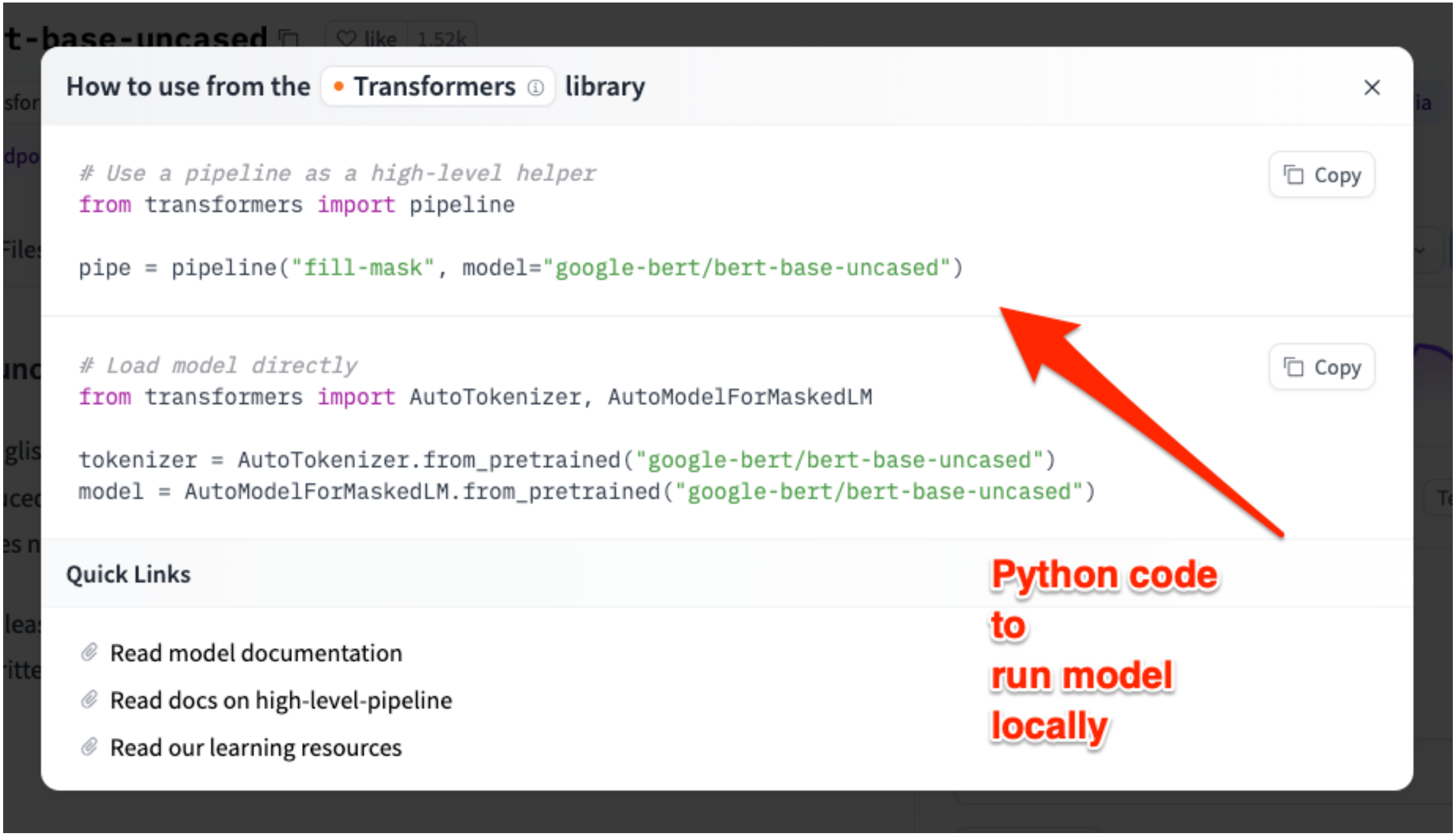
Precision

- [x] float16
- [x] bfloat16
- [x] 8bit
- [x] 4bit
- [x] GPTQ
- [x] ?

Model sizes (in billions of parameters)

- [x] ?
- [x] ~1.5
- [x] ~3
- [x] ~7
- [x] ~13
- [x] ~35
- [x] ~60
- [x] 70+

T	Model	Average ⬆️	ARC	HellaSwag	MMLU	T
🔶	SF-Foundation/Ein-70B-v2	81.29	79.86	91.49	78.05	7
🔶	davidkim205/Rhea-72b-v0.5	81.22	79.78	91.15	77.95	7
💬	MTSAIR/MultiVerse_70B	81	78.67	89.77	78.22	7
🔶	MTSAIR/MultiVerse_70B	80.98	78.58	89.74	78.27	7


How to use from the Transformers library
# Use a pipeline as a high-level helper
from transformers import pipeline
pipe = pipeline("fill-mask", model="google-bert/bert-base-uncased")
Copy
# Load model directly
from transformers import AutoTokenizer, AutoModelForMaskedLM
tokenizer = AutoTokenizer.from_pretrained("google-bert/bert-base-uncased")
model = AutoModelForMaskedLM.from_pretrained("google-bert/bert-base-uncased")
Copy
Quick Links
Read model documentation
Read docs on high-level-pipeline
Read our learning resources
Python code
to
run model
locally

Examples

Hugging Face Infrastructure

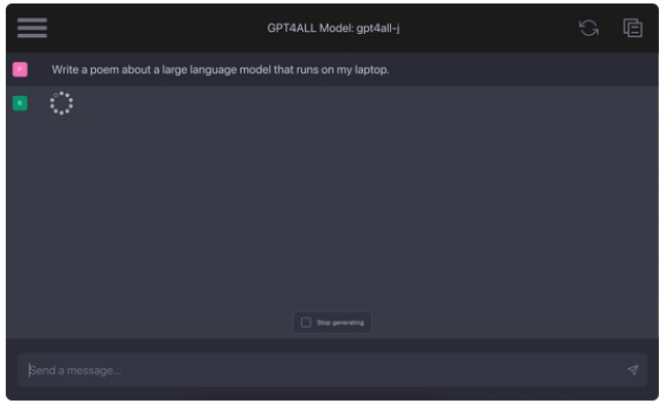
Locally on my laptop

Applications

- GPT4All
- Ollama
- Stable Diffusion

GPT4All

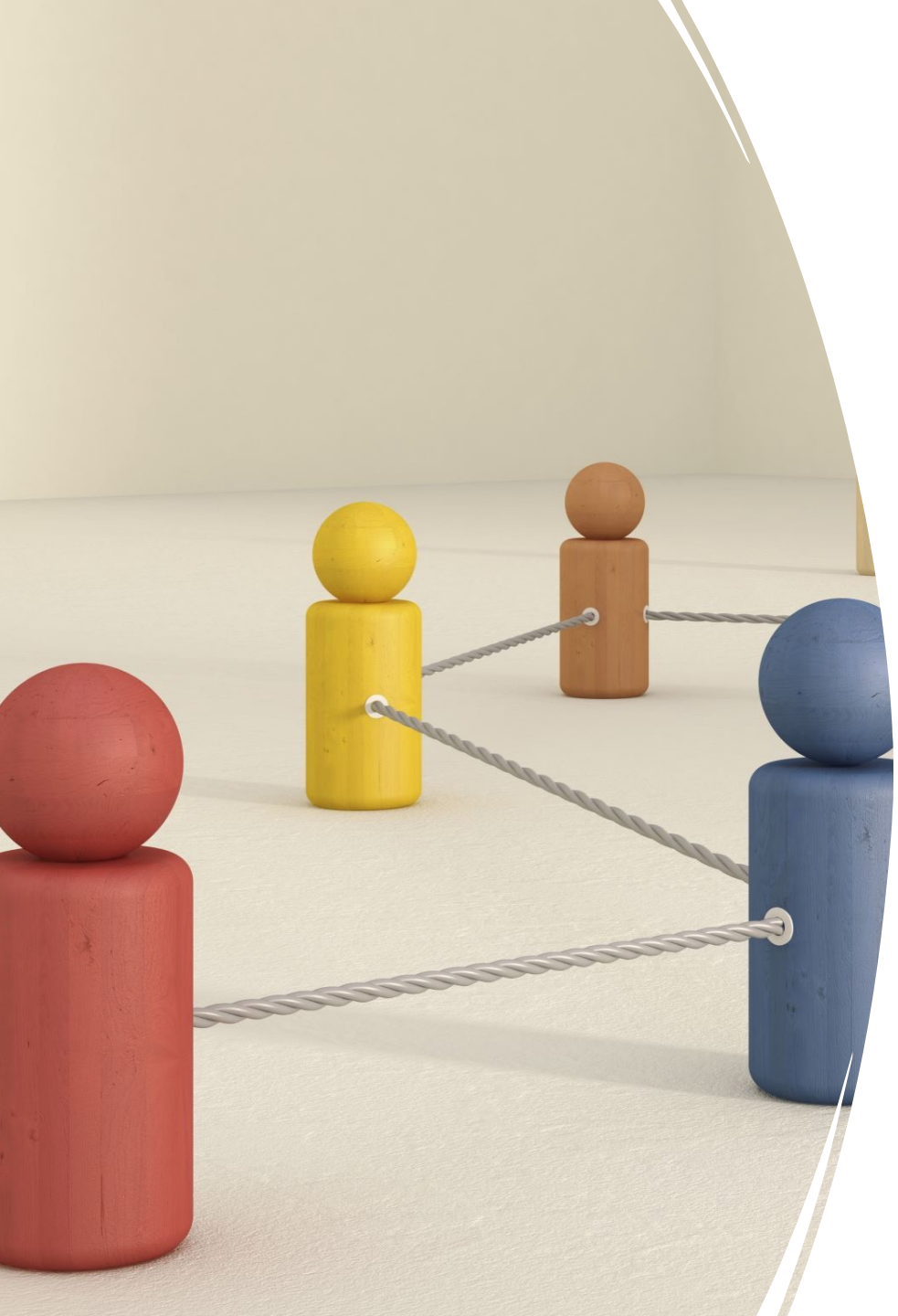
A free-to-use, locally running, privacy-aware chatbot. **No GPU or internet required.**



Real-time inference latency on an M1 Mac

Download Desktop Chat Client

Windows Installer OSX Installer Ubuntu Installer



Future Outlook for Open Source LLMs

- **Democratization of AI Technology**
  - Increased Accessibility
  - Community Collaboration
- **Advancements in Model Robustness and Capabilities**
  - Improved Model Quality
  - Innovation in Training Techniques
- **Integration with Other Technologies**
  - Cross-Platform Integration (Blockchain, IoT and AR/VR)
  - Enhanced Tooling and Support (LangChain)
  - Push proprietary models to more open and collaborative approaches

Need Help Assessing Your Options?

Talk with me!

Thank you!

Justin Grammens
Founder + CEO
justin@lab651.com
https://lab651.com
https://recursiveawesome.com

Join us at an upcoming
Applied AI event!
https://appliedai.mn

Applied AI Meetup &
Links to Resources

LAB651 | RECURSIVE AWESOME

LAB651 RECURSIVE AWESOME