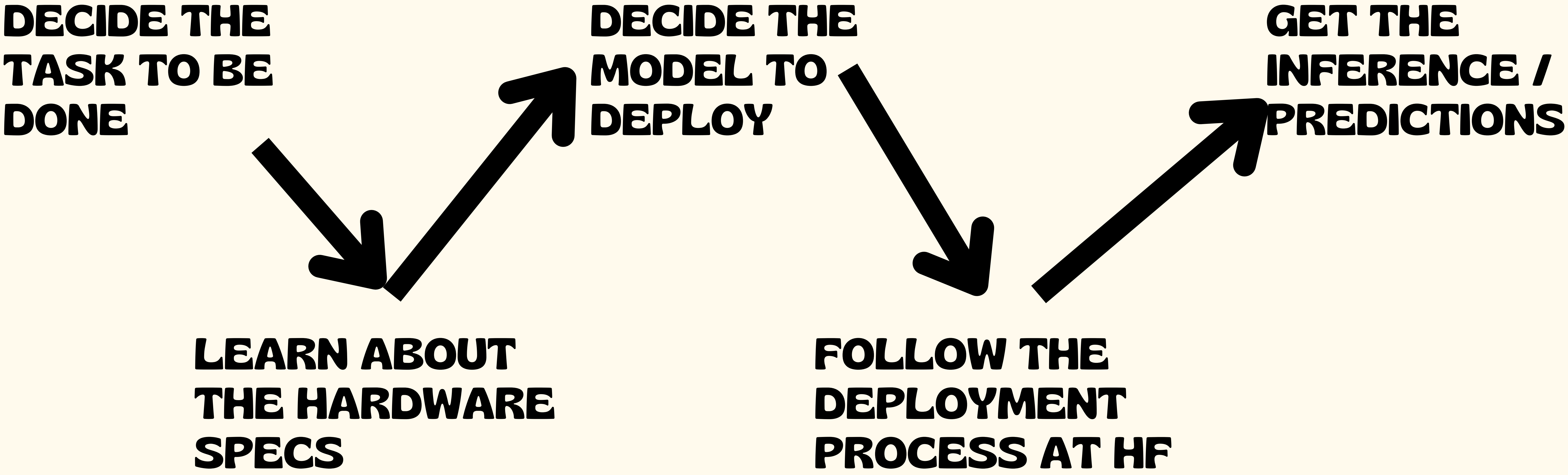
DEMISTIFYING OPEN SOURCE MODEL DEPLOYMENT @ HUGGING FACE

INTRODUCING SPACES AND INFERENCE ENDPOINTS

OPEN SOURCE

HTTPS://GITHUB.COM/INSIGHTBUILDER

TWO WAYS TO DEPLOY MODELS



DECIDE THE TASK TO BE DONE

LEARN ABOUT THE HARDWARE SPECS

DECIDE THE MODEL TO DEPLOY

FOLLOW THE DEPLOYMENT PROCESS AT HF

GET THE INFERENCE / PREDICTIONS

SPACES

INFERENCE ENDPOINTS

HTTPS://GITHUB.COM/INSIGHTBUILDER

WHAT ARE THE FEATURES

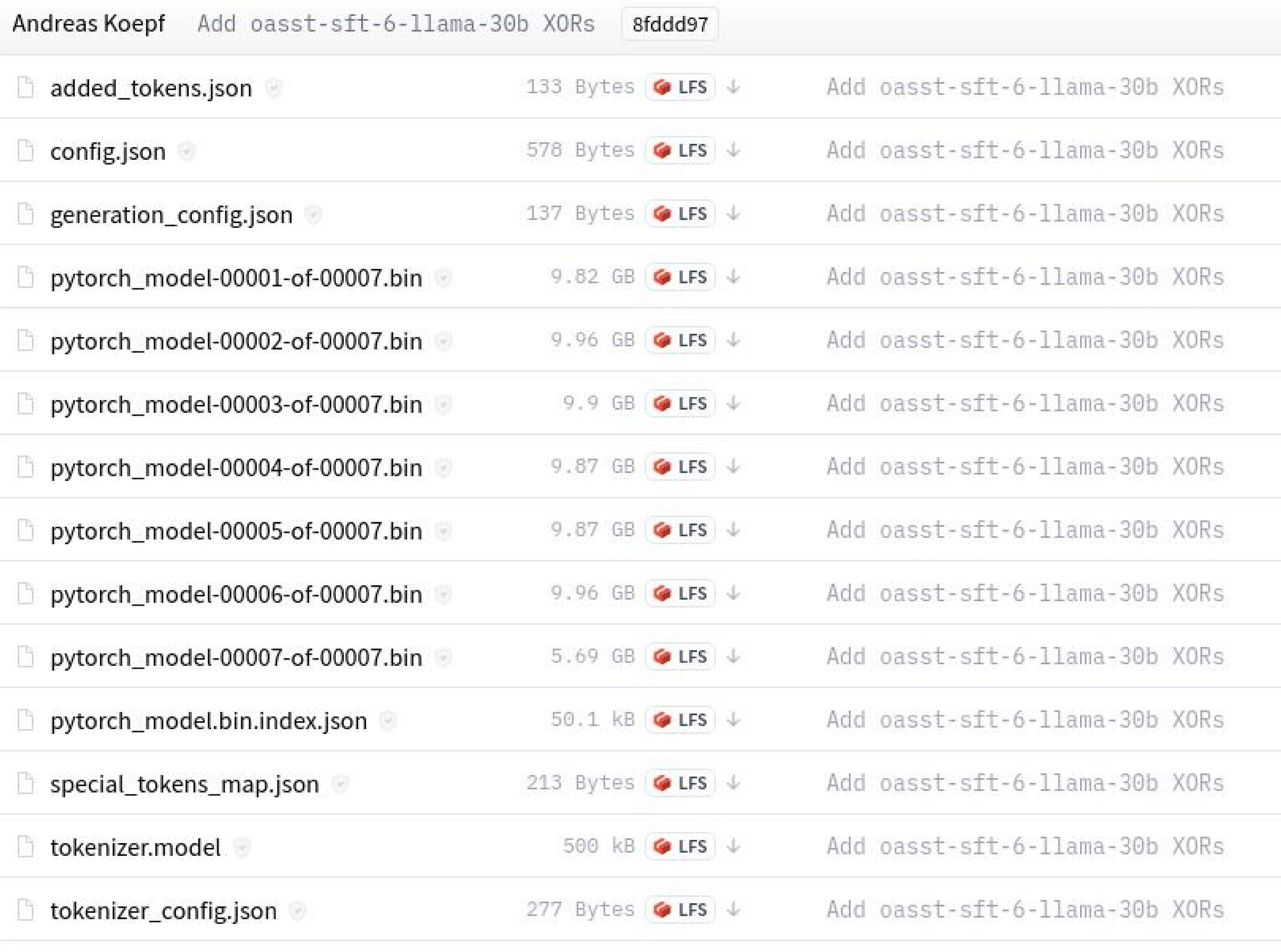
SPACES

- WILL BE DEPLOYED INSIDE HF
- FREE LIMIT OF 16GB RAM AND 8 CPU CORES
- WORKS WITH STREAMLIT/GRADIO
- ACCESSIBLE ONLY THROUGH THE UI INSIDE HUGGING FACE
- TRANSFORMERS PIPELINE FUNCTION IS THE KEY
- UPGRADABLE TO GPU INSTANCES

HUGGINGFACE INFERENCE ENDPOINTS

- WILL BE DEPLOYED @ AWS / AZURE
- STARTS AT 0.06 USD / HR
- MINIMUM 1VCPU 2GB INTEL CORE
- WILL PROVIDE THE API ENDPOINT TO WHICH REQUEST TO BE SENT
- WILL BE ACCESSIBLE OUTSIDE HF
- NO CODE TO BE WRITTEN
- AUTOSCALING / CUSTOM CONFIGS ARE POSSIBLE

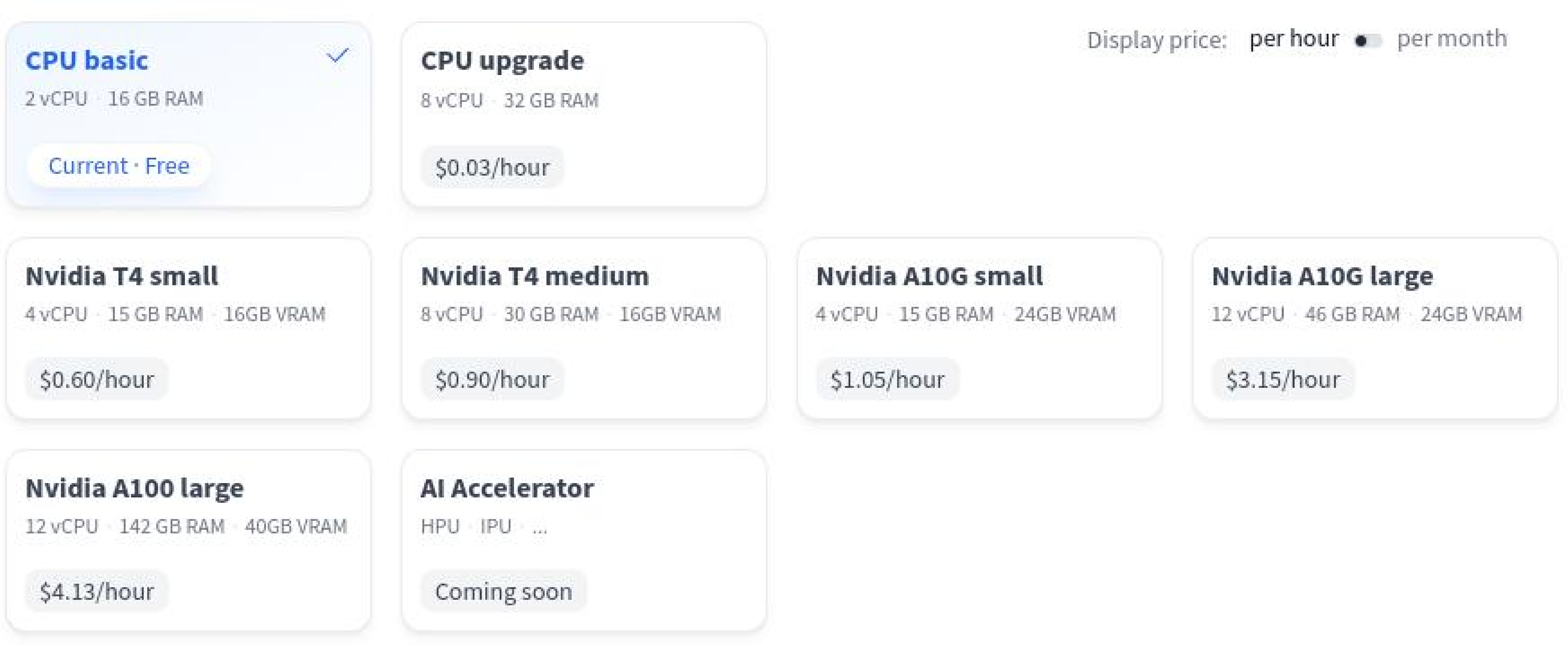
SIMPLE WAY TO FIND THE MODEL SIZE



Andreas Koepf Add oasst-sft-6-llama-30b XORs 8fddd97

added_tokens.json	133 Bytes	LFS	↓	Add oasst-sft-6-llama-30b XORs
config.json	578 Bytes	LFS	↓	Add oasst-sft-6-llama-30b XORs
generation_config.json	137 Bytes	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00001-of-00007.bin	9.82 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00002-of-00007.bin	9.96 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00003-of-00007.bin	9.9 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00004-of-00007.bin	9.87 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00005-of-00007.bin	9.87 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00006-of-00007.bin	9.96 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model-00007-of-00007.bin	5.69 GB	LFS	↓	Add oasst-sft-6-llama-30b XORs
pytorch_model.bin.index.json	50.1 kB	LFS	↓	Add oasst-sft-6-llama-30b XORs
special_tokens_map.json	213 Bytes	LFS	↓	Add oasst-sft-6-llama-30b XORs
tokenizer.model	500 kB	LFS	↓	Add oasst-sft-6-llama-30b XORs
tokenizer_config.json	277 Bytes	LFS	↓	Add oasst-sft-6-llama-30b XORs

KNOW THE HARDWARE & ITS COST



Display price: per hour ● per month

CPU basic ✓
2 vCPU · 16 GB RAM
Current · Free

CPU upgrade
8 vCPU · 32 GB RAM
$0.03/hour

Nvidia T4 small
4 vCPU · 15 GB RAM · 16GB VRAM
$0.60/hour

Nvidia T4 medium
8 vCPU · 30 GB RAM · 16GB VRAM
$0.90/hour

Nvidia A10G small
4 vCPU · 15 GB RAM · 24GB VRAM
$1.05/hour

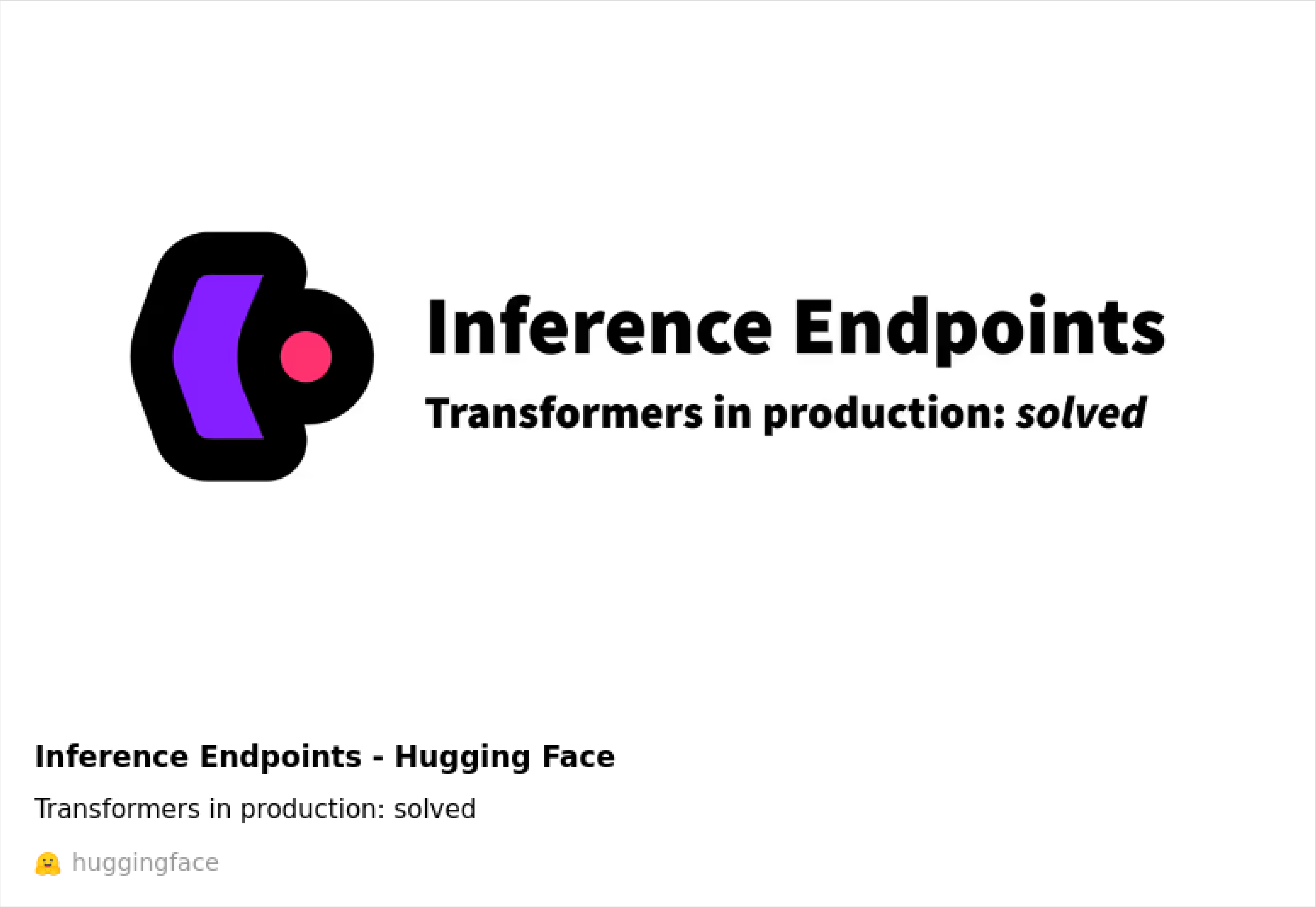
Nvidia A10G large
12 vCPU · 46 GB RAM · 24GB VRAM
$3.15/hour

Nvidia A100 large
12 vCPU · 142 GB RAM · 40GB VRAM
$4.13/hour

AI Accelerator
HPU · IPU · ...
Coming soon

LETS HEAD TO HUGGING FACE

NOW WE ARE TALKING. PRACTICE???



Inference Endpoints

Transformers in production: *solved*

Inference Endpoints - Hugging Face

Transformers in production: solved

🤗 huggingface

Spaces

hf.co/spaces

Spaces - Hugging Face

Discover amazing ML apps made by the community

🤗 huggingface

THANKS FOR WATCHING

LIKE

SHARE

SUBSCRIBE

HTTPS://GITHUB.COM/INSIGHTBUILDER