Artificial intellegence stack

Stack

Required parts for a DL computer

Motherboard: Circuit board that connects all components together such as CPU, memory sticks etc., determines how compatible peripherals will attach themselves according to form factor standards such as ATX, Micro-ATX, Mini ITX etc...

Central processing unit (CPU): The brain of a computer; executes instructions from software programs and processes data; comes with different cores depending on how many threads may be active at once. Modern CPUs have multiple cores enabling multi-tasking capabilities. Must be able to handle tensor operations. **Intel Xeon, AMD Threadripper. 12C/24T

Random Access Memory (RAM): It refers to temporary memory used by a computer while it operates. Typically measured in megabytes or gigabytes (MB/GB). More RAM means faster performance and more programs can be open simultaneously.
**128 Gb

Storage: Storage refers to retaining data on some form of device like SSD drives (Solid State Drives). Unlike traditional spinning disks known as Hard Disk Drives (HDDs), these newer alternative devices effectively provide faster read/write times by utilizing flash technology, often used also when retrieving files quickly. **4Tb SSD

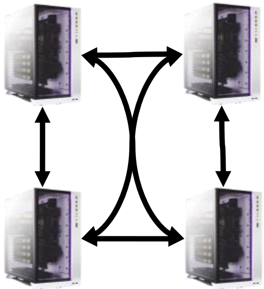
Graphics Processing Unit (GPU)/Tensor Processing Unit (TPU): Processors designed specifically for handling graphics and computations related tasks like rendering images or performing deep learning operations respectively. NVIDIA is known for its powerful GPUs suitable for machine learning applications whereas Google developed TPUs which are optimized for their own AI workloads. *NVIDIA workstation cards (RTX A6000, A100 , or H100)

Power Supply Unit (PSU): Devices responsible for providing power to other hardware components within the system unit. They must match the requirements of the motherboard and other parts they power. Selecting an appropriate power supply unit guarantees smooth functioning if not exceeding wattage requirements could potentially cause instabilities or damage components within the system due to voltage fluctuations. **1100 watt

Cooling System: Important component preventing overheating caused by excessive heat generated during intensive activities – includes heatsinks/fans attached either manually or pre-installed upon purchase.- Considerations before purchasing include compatibility between each piece along with suitability concerning usage demands. For instance if one plans doing heavy artificial intelligence workstation activity then larger RAM and GPU will be used, which generate large amounts of heat. *6 fans @3000 RPM

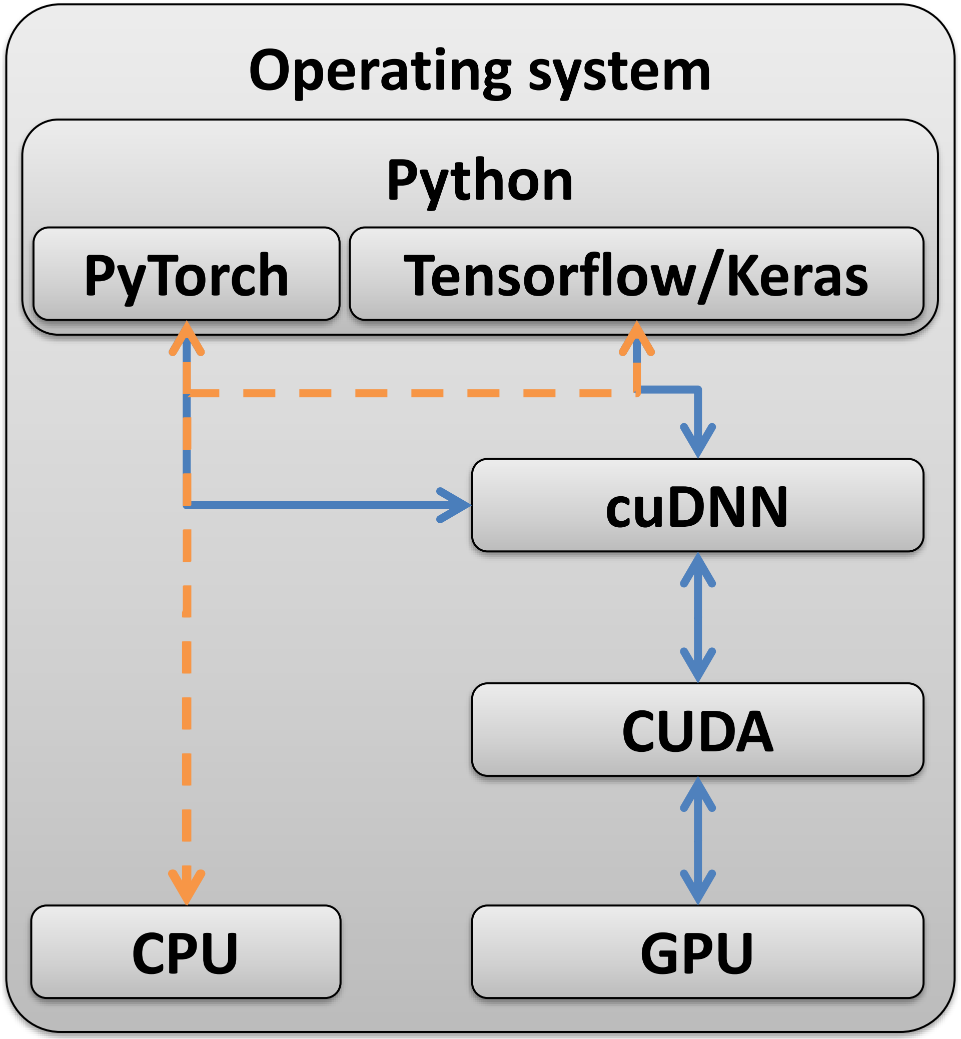
*Summarized with Llama 2 using Faraday

** Suggested minimum for a Ai/ML system



	Laptop	Desktop	Workstations	Virtual Clusters	Centralized Clusters/HPC	Cloud
CPU	4-10 Cores	4-16 Cores	12-64Cores	100s	1000s	“
GPU	2-4 Gb	1x-2x 16Gb	2x-4x 48 Gb	10s	1000s	“
RAM	8-32 Gb	8 – 64	128 - 512	10s Tb	100s Tb	“
Est. Cost	$1K-2K	$1K – 3K	$10K-35K	$50K>	$0.01 - 27.43/hr	“
Other fees/cost	Protection plan	Storage	Storage	Storage Software	Storage Software Infrastructure	Storage I/O Cont. Cost
Scaling options	Virtual cluster	Virtual cluster	Virtual cluster	Add macines	UTH, MDACC, UTMB, BCM, TAMU, Rice, Methodist, UH	$$$
Access	Good	Good	Good	Administrated	Administrated	Open

Configuring a deep learning machine



Operating system

Python

PyTorch

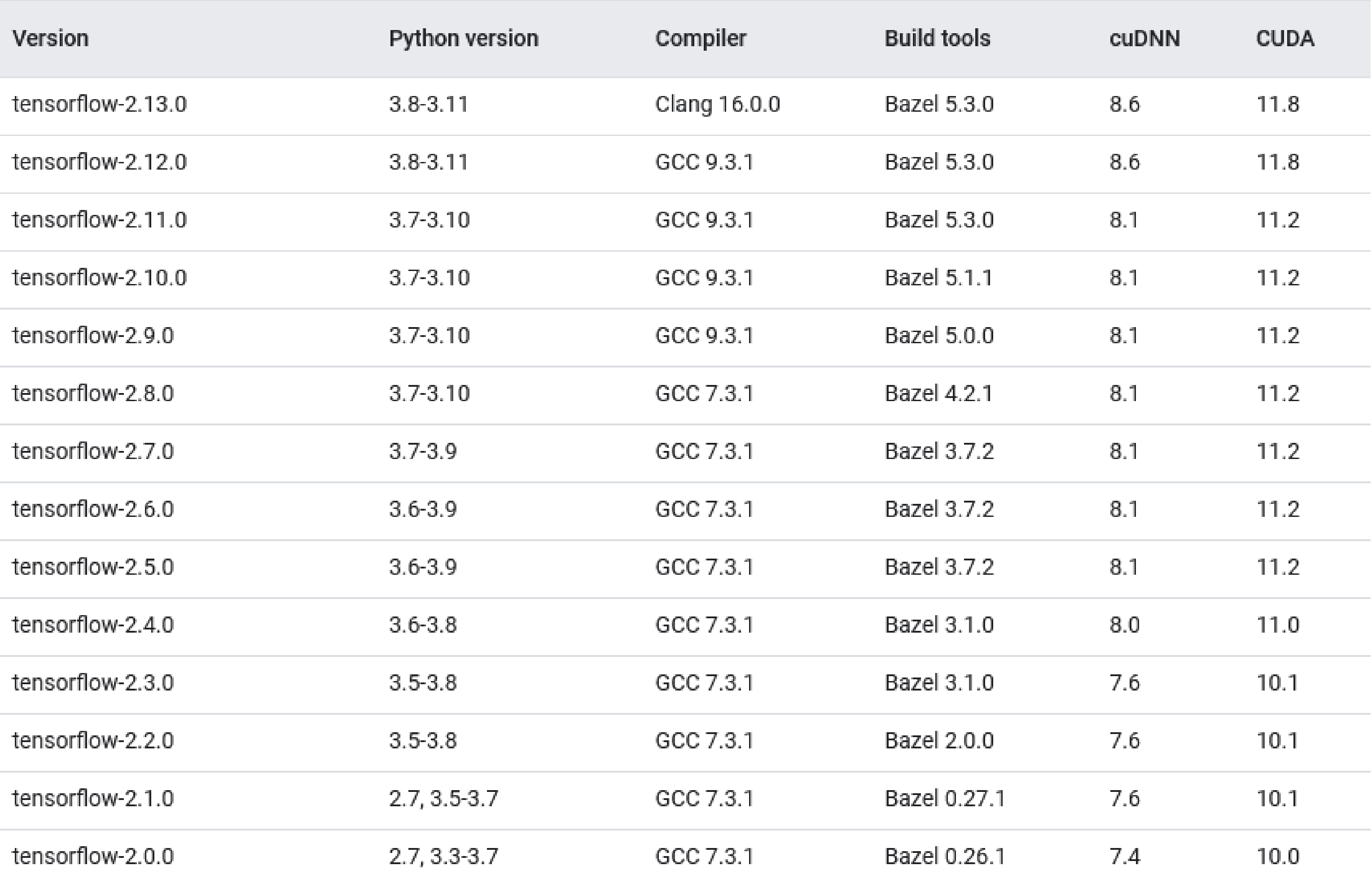
Tensorflow/Keras

cuDNN

CUDA

CPU

GPU



Version	Python version	Compiler	Build tools	cuDNN	CUDA
tensorflow-2.13.0	3.8-3.11	Clang 16.0.0	Bazel 5.3.0	8.6	11.8
tensorflow-2.12.0	3.8-3.11	GCC 9.3.1	Bazel 5.3.0	8.6	11.8
tensorflow-2.11.0	3.7-3.10	GCC 9.3.1	Bazel 5.3.0	8.1	11.2
tensorflow-2.10.0	3.7-3.10	GCC 9.3.1	Bazel 5.1.1	8.1	11.2
tensorflow-2.9.0	3.7-3.10	GCC 9.3.1	Bazel 5.0.0	8.1	11.2
tensorflow-2.8.0	3.7-3.10	GCC 7.3.1	Bazel 4.2.1	8.1	11.2
tensorflow-2.7.0	3.7-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.6.0	3.6-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.5.0	3.6-3.9	GCC 7.3.1	Bazel 3.7.2	8.1	11.2
tensorflow-2.4.0	3.6-3.8	GCC 7.3.1	Bazel 3.1.0	8.0	11.0
tensorflow-2.3.0	3.5-3.8	GCC 7.3.1	Bazel 3.1.0	7.6	10.1
tensorflow-2.2.0	3.5-3.8	GCC 7.3.1	Bazel 2.0.0	7.6	10.1
tensorflow-2.1.0	2.7, 3.5-3.7	GCC 7.3.1	Bazel 0.27.1	7.6	10.1
tensorflow-2.0.0	2.7, 3.3-3.7	GCC 7.3.1	Bazel 0.26.1	7.4	10.0

tf: https://www.tensorflow.org/install/source#gpu

torch: https://pytorch.org/get-started/previous-versions/

Choice of operating system

OS:	Linux	Windows	WSL	MacOS
Common:	Ubuntu CentOS Red Hat	Windows 10 Windows 11	WSL1 WSL2*	Ventura
Command interface:	Terminal (Bash)	Cmd, Powershell	Terminal (Bash)	Terminal (Bash)
Pros:	1. Ubuntu is free 2. Open-source 3. Customizable 4. Less bloated 5. Common on HPRC	1. Most widely used OS for professional systems 2. Large user base	1. Lightweight Linux machine running on Windows 11 2. Best of both worlds 3. Supports Dockers	User interface Linux-based
Cons:	1. Step learning curve 2. Some compatibility issues 3. Commercial support issues	1. Larger computational overhead 2. Less customizable/flexable	1. Slightly reduced performance	Increased cost Limited customizability

**Pro’s and con’s generated by Wizard v1.1 model with GPT4ALL

Reputable cloud providers

Host	Azure	AWS	Colab	OCI	DigitalOcean	IBM	Lambda
Parent Co.	Microsoft	Amazon	Google	Oracle	DigitalOcean, Inc.	IBM	Lambda Labs
Interface	Virtual compute Database Storage	Virtual compute Database Storage	Virtual Notebook	Virtual compute Database Storage	App hosting	Virtual compute Database Storage	Jupyter

Common code/scripting langs.

Software:	**Matlab**	**R statistics**	**Python**
Interface:	Programing Scripting	Programing Scripting	Programing Scripting
Features:	Machine Learning Toolbox	caret e1071 randomForest xgBoost nueralnet	PIL scikit-image scikit-learn Pytorch Tensorflow
Licensing:	Commercial	Open Source	Open Source

Interactive Developer Environments (IDE)

	R	PC	jupyter
Software:	**R studios**	**Jetbeans/PyCharm**	**Jupyter Labs/notebook**
Languages:	R stats Python	R stats Python	R stats Python
Features:	Scripting Terminal Notebooks Env editor/preview	Scripting Terminal Env editor/preview	Scripting Terminal Notebooks Env editor/preview
Licensing:	Free Commercial	Free Commercial	Free Commercial

No-code & Low-code ML

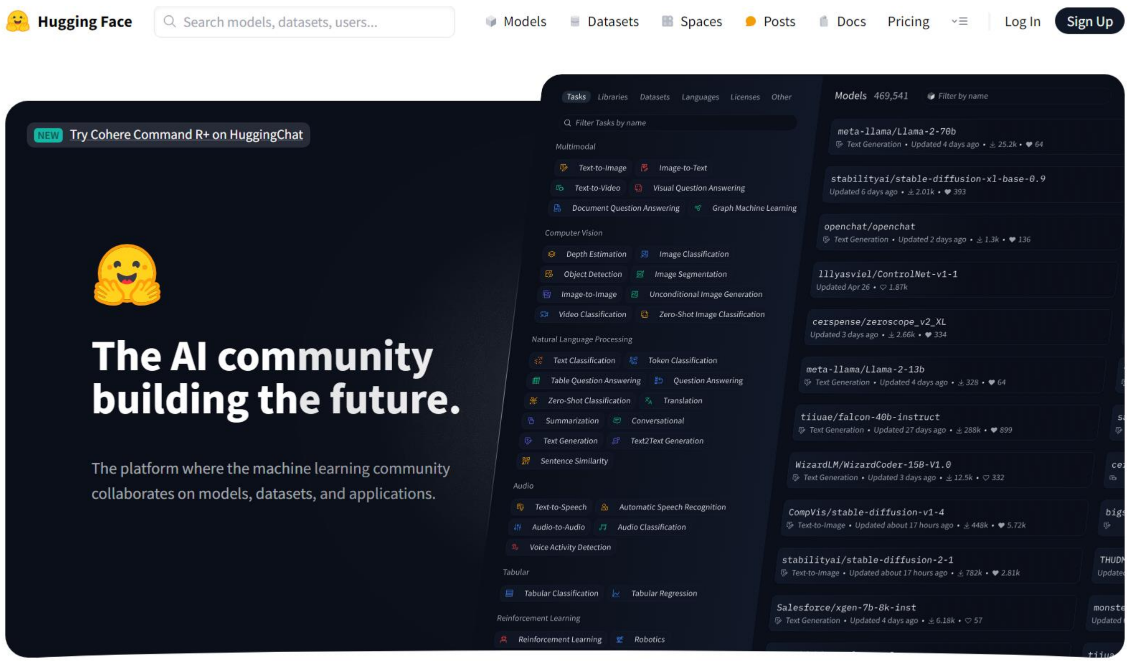
Software:	H2O Flow	Orange	Knime	Rapidminer	Pipeline Pilot
Interface:	Low-code	No-code	No-code Low-code	No-code Low-code	No-code Low-code
Features:	AutoML Notebook GPT	Python with visual programing	Integration Automation Visualization	Automation Visualization	Integration Automation Visualization
Licensing:	Open Source	Open Source	Open source	Commercial	Commercial

Chat-based AI interfaces

Name	ChatGPT	GPT4ALL	Faraday.dev	H2Ogpt LLMstudios	Text-generation-webui	Jupyter_AI	Co-Pilot
Parent Company	OpenAI	GPT4ALL	Faraday	H2O.ai	oobabooga	Jupyter	Microsoft
Release	Alpha	Alpha	Alpha	Alpha	Alpha	Beta	Beta
Base models	GPT3.5 turbo GPT4	HF	HF	HF	HF	GPT3.5 turbo	GPT3.5 turbo
Instance	Cloud	CPU_Local	CPU_Local GPU_Local	GPU_Local	GPU_Local	Cloud_API	Cloud
Pricing model	Free Subscription ($20/mo)	Free	Free	Free	Free	Free Requires API Keys	Subscription ($30/mo)

The new source of knowledge

HuggingFace



The AI community building the future.

The platform where the machine learning community collaborates on models, datasets, and applications.

Models 469,541

meta-llama/Llama-2-70b
Text Generation • Updated 4 days ago • 25.2k • 64

stabilityai/stable-diffusion-xl-base-0.9
Updated 6 days ago • 2.01k • 393

openchat/openchat
Text Generation • Updated 2 days ago • 1.3k • 136

lllyasviel/ControlNet-v1-1
Updated Apr 26 • 1.87k

cerspense/zeroscope_v2_XL
Updated 3 days ago • 2.66k • 334

meta-llama/Llama-2-13b
Text Generation • Updated 4 days ago • 328 • 64

tiiuae/falcon-40b-instruct
Text Generation • Updated 27 days ago • 288k • 899

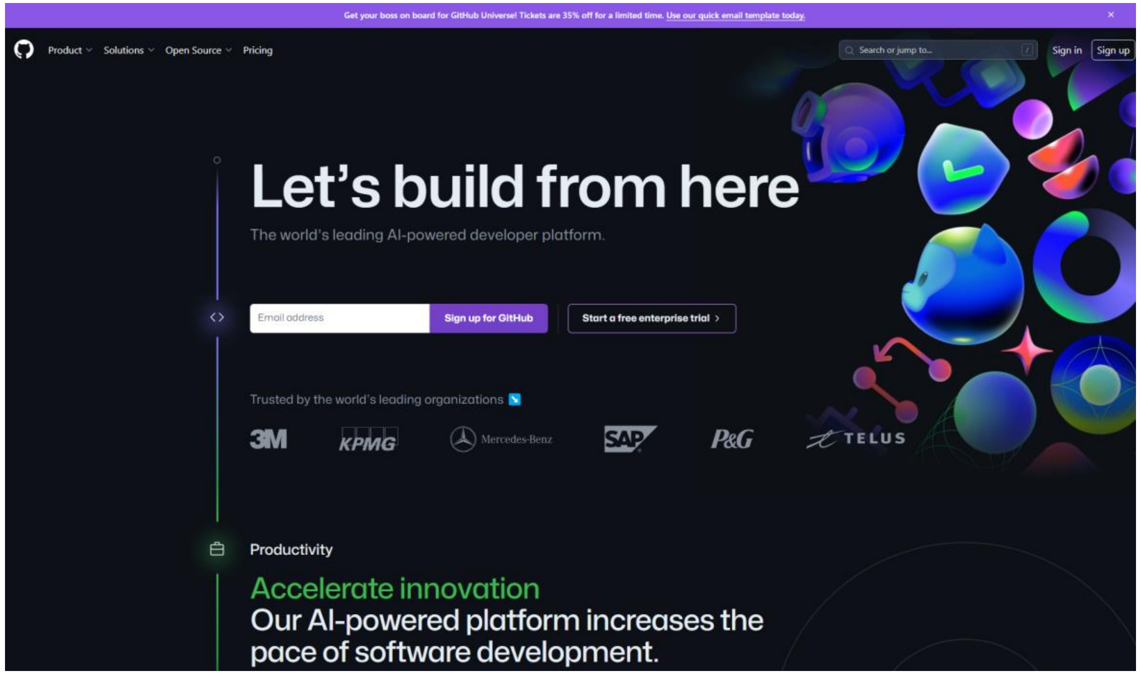
WizardLM/WizardCoder-15B-V1.0
Text Generation • Updated 3 days ago • 12.5k • 332

CompVis/stable-diffusion-v1-4
Text-to-Image • Updated about 17 hours ago • 448k • 5.72k

stabilityai/stable-diffusion-2-1
Text-to-Image • Updated about 17 hours ago • 782k • 2.81k

Salesforce/xgen-7b-8k-inst
Text Generation • Updated 4 days ago • 6.18k • 57

Github



Get your boss on board for GitHub Universe! Tickets are 35% off for a limited time. Use our quick email template today.

Let's build from here

The world's leading AI-powered developer platform.

Trusted by the world's leading organizations

Productivity

Accelerate innovation

Our AI-powered platform increases the pace of software development.

Huggingface models 🤗

Name	GPT	Llama(2, 3)	Orca(2)	Wizard(Coder)	Falcon	T5	BERT
Parent Co.	OpenAi	Meta	Microsoft	WizardLM	TII	Google	Google
Base Tech	GPT	Llama	Llama	Llama	Llama	t5	BERT
Scope	Assistant Coding Text generation	Assistant Coding Text generation	Instruction Task Completion	Instruction Coding	Text generation	Seq2Seq	Text generation Text Classification NER
Sizes		7B 13B 70B	7B 13B	7B 13B 70B	7B 40B 180B		