Adevinta

Running an open source LLM

Bongani Shongwe

CONF42

Adevinta

France

leboncoin
Groupe Argus
locasun
àvendre àlouer
Truckscorner
MachineryZone
Agriaffaires

Spain

milanuncios
fotocasa
coches.net
InfoJobs
motos.net
habitaclia

Belgium

2ememain
2dehands

Netherlands

Marktplaats

Germany

kleinanzeigen
mobile.de

Italy

subito
InfoJobs
automobile.it

Canada

kijiji

Ireland

daft.ie
ADVERTS.ie
DoneDeal
Gumtree

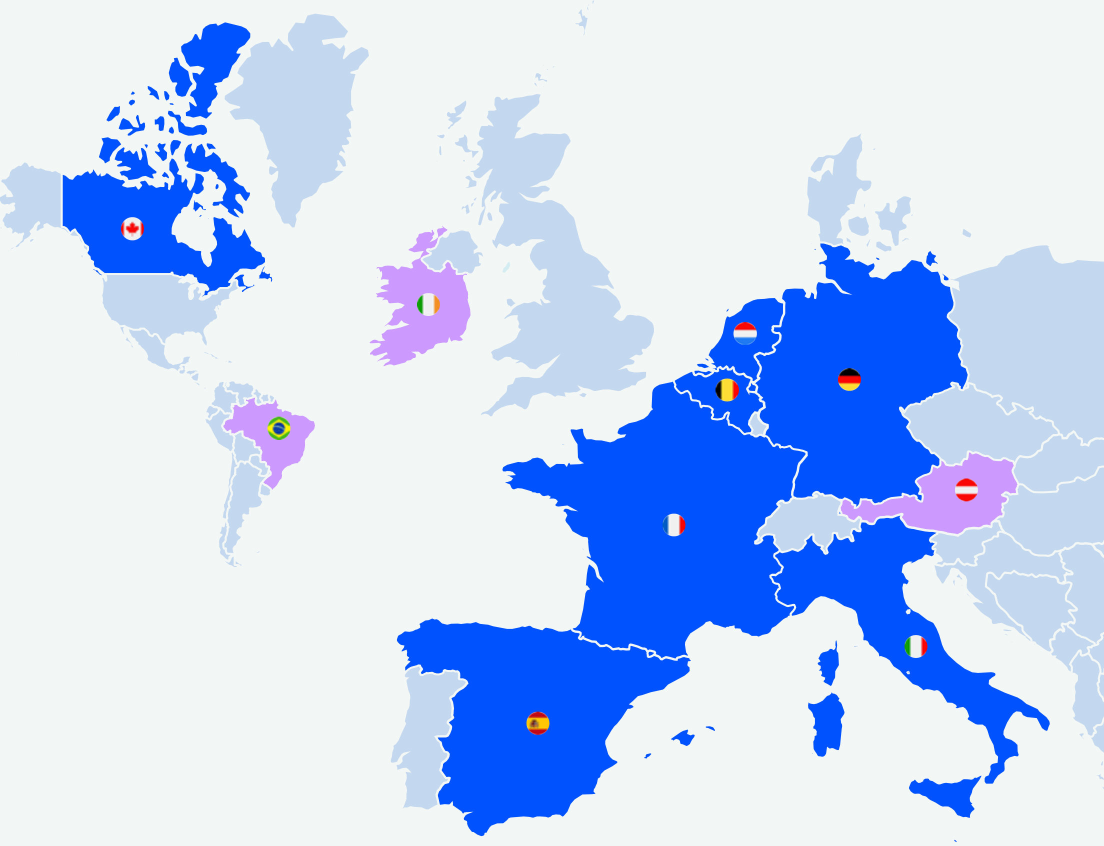
Austria

WILLHABEN

Brazil

olx
zap

Joint ventures



2.5bn

monthly visits worldwide

25+

marketplaces in our digital portfolio

€1.8bn

revenues in FY 2023

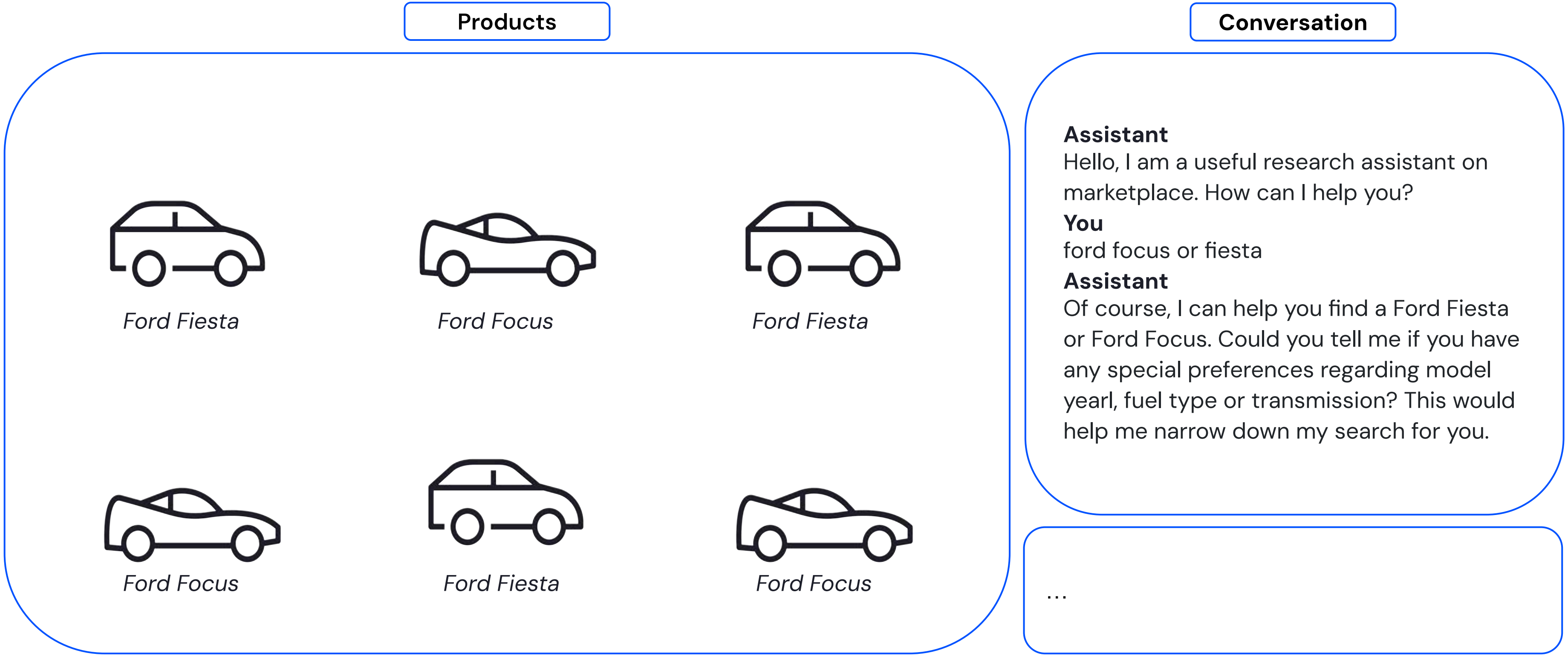
5700+

employees (approx), across **10 countries**

The Product

Conversational search assistant 💬

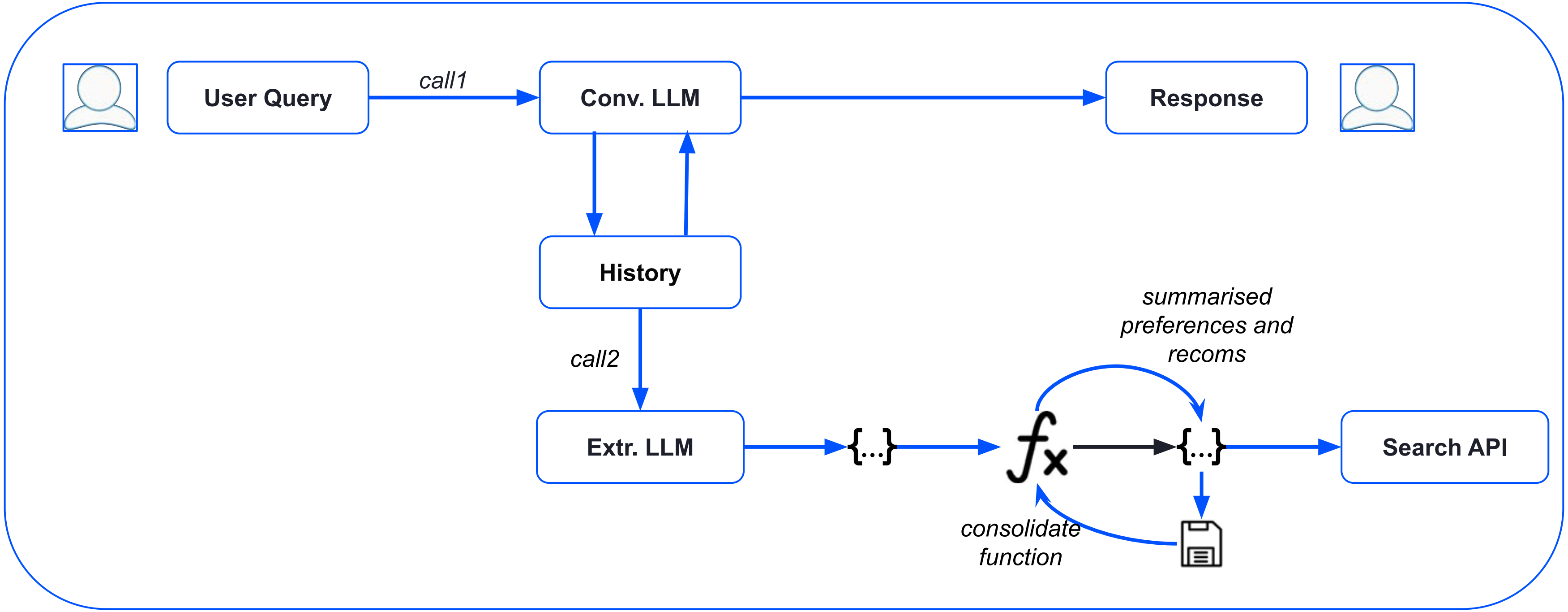
The Product We Envisioned



Assistant
Hello, I am a useful research assistant on marketplace. How can I help you?
You
ford focus or fiesta
Assistant
Of course, I can help you find a Ford Fiesta or Ford Focus. Could you tell me if you have any special preferences regarding model yearl, fuel type or transmission? This would help me narrow down my search for you.

...

Basic Overview



Conversation LMM: GPT-3.5-turbo
Extraction LLM: GPT-3.5-turbo

Host own LLM

Great privacy and compliance

Customisation

Avoid vendor lock-in

Control of quality of service

????

Enterprise LLM Access

Easy entry point to LLMs

Minimal infrastructure

High speed implementation

Slow API response time

Slow onboarding

Costs of tokens high

Privacy of data

External dependency on service

Host own LLM

Great privacy and compliance

Customisation

Avoid vendor lock-in

Control of quality of service

????

Enterprise LLM Access

Easy entry point to LLMs

Minimal infrastructure

High speed implementation

Slow API response time

Slow onboarding

Costs of tokens high

Privacy of data

External dependency on service

Hugging Face

Falcon 40B

Chat tuned

Multiple language support

140B parameter model also available

Llama 2 70B

Chat tuned

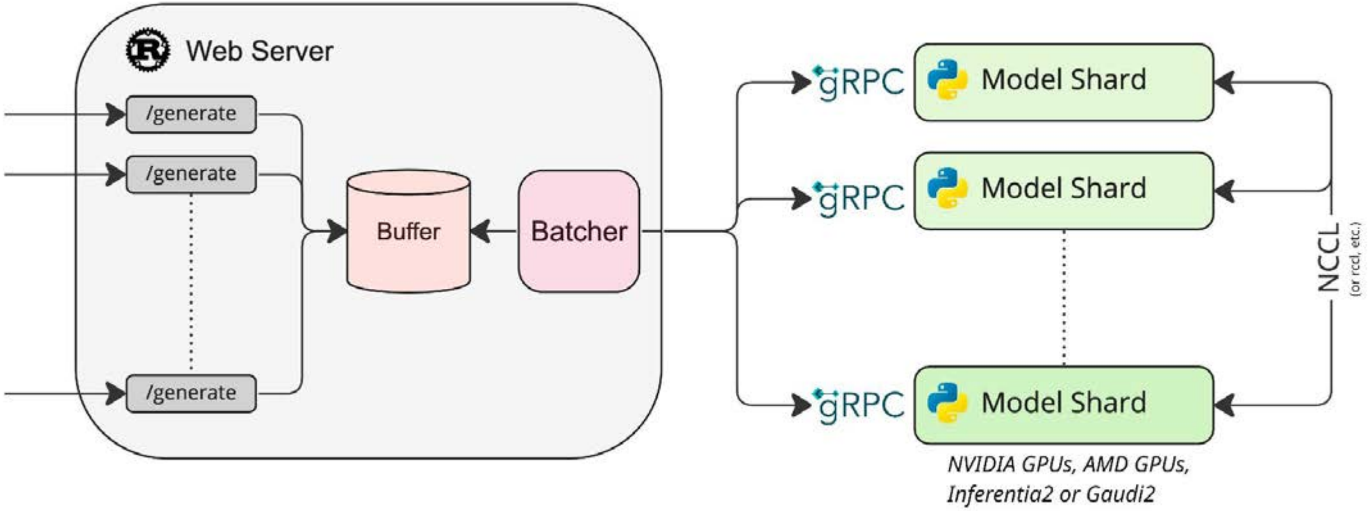
Multiple language support

Gated model

*Falcon 7B for initial setup

Hosting LLM

Text-Generation-Inference (TGI)



https://github.com/huggingface/text-generation-inference/pkgs/container/text-generation-inference

Hosting LLM

Text-Generation-Inference (TGI)

```
model=HuggingFaceH4/zephyr-7b-beta
volume=$PWD/data # share a volume with the Docker container to avoid downloading weights every run

docker run --gpus all --shm-size 1g -p 8080:80 -v $volume:/data
ghcr.io/huggingface/text-generation-inference:1.3 --model-id $model
```

```
curl 127.0.0.1:8080/generate \
    -X POST \
    -d '{"inputs":"What is Deep Learning?","parameters":{"max_new_tokens":20}}' \
    -H 'Content-Type: application/json'
```

https://github.com/huggingface/text-generation-inference/pkgs/container/text-generation-inference

Based on Google Cloud

Google Cloud

GPU Requirements

High GPU memory required

Nvidia A100 80GB
most sort for

The higher the
parameters for
model more
memory required

Models can run in 4bit mode
to decrease memory footprint

Deployment Infrastructure

Kubernetes

Using a deployment yaml

Autoscale VMs

CUDA ready machines

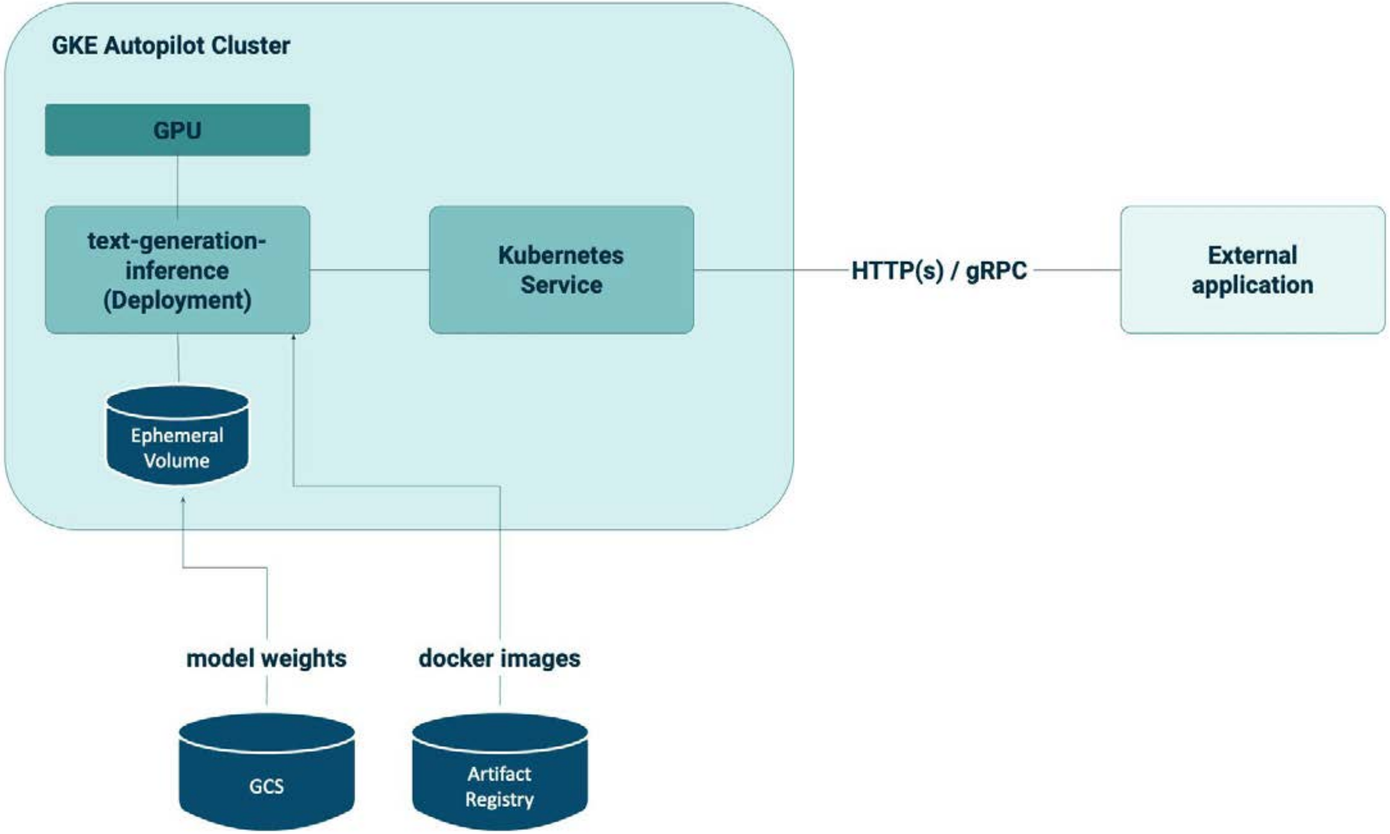
VM

Simple docker command

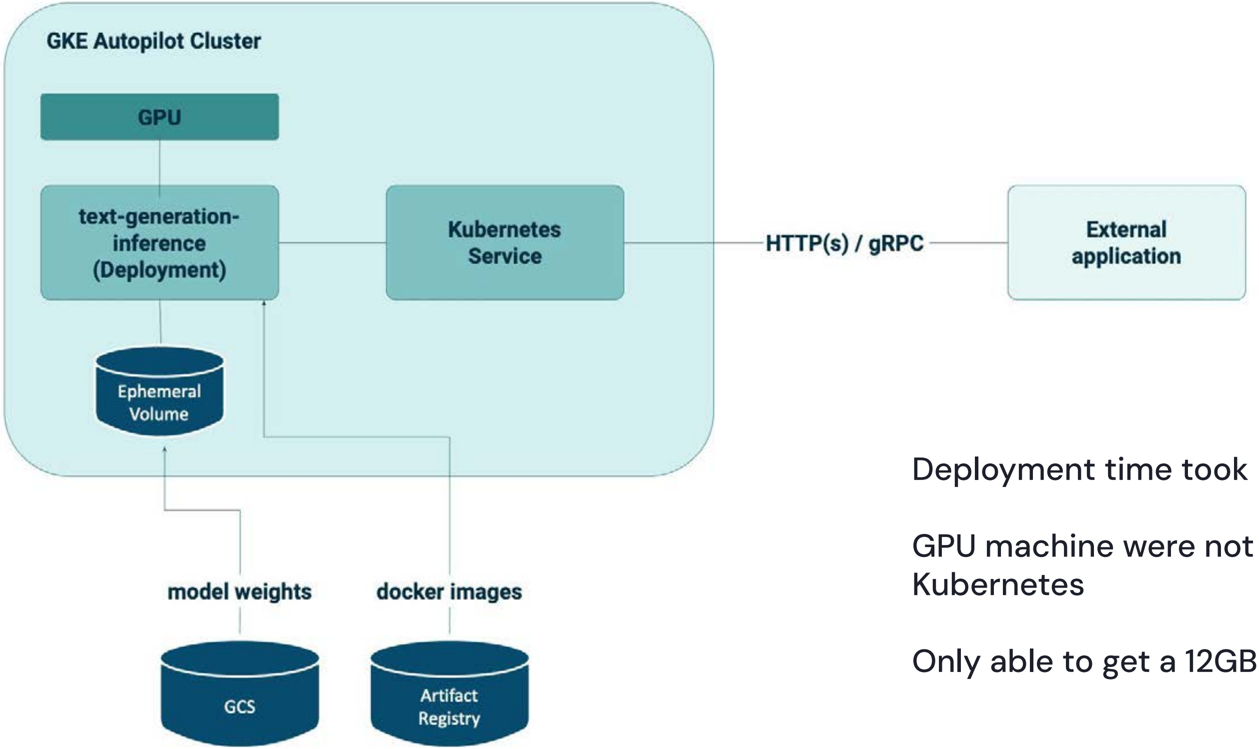
Self managed machines

CUDA ready machines

Kubernetes


GKE Autopilot Cluster
GPU
text-generation-inference (Deployment)
Kubernetes Service
HTTP(s) / gRPC
External application
Ephemeral Volume
model weights
docker images
GCS
Artifact Registry

Kubernetes ✕



Deployment time took long or not at all

GPU machine were not readily available for use with Kubernetes

Only able to get a 12GB GPU

Adevinta

* issues can be different to Kubeflow versions

VM Deployment

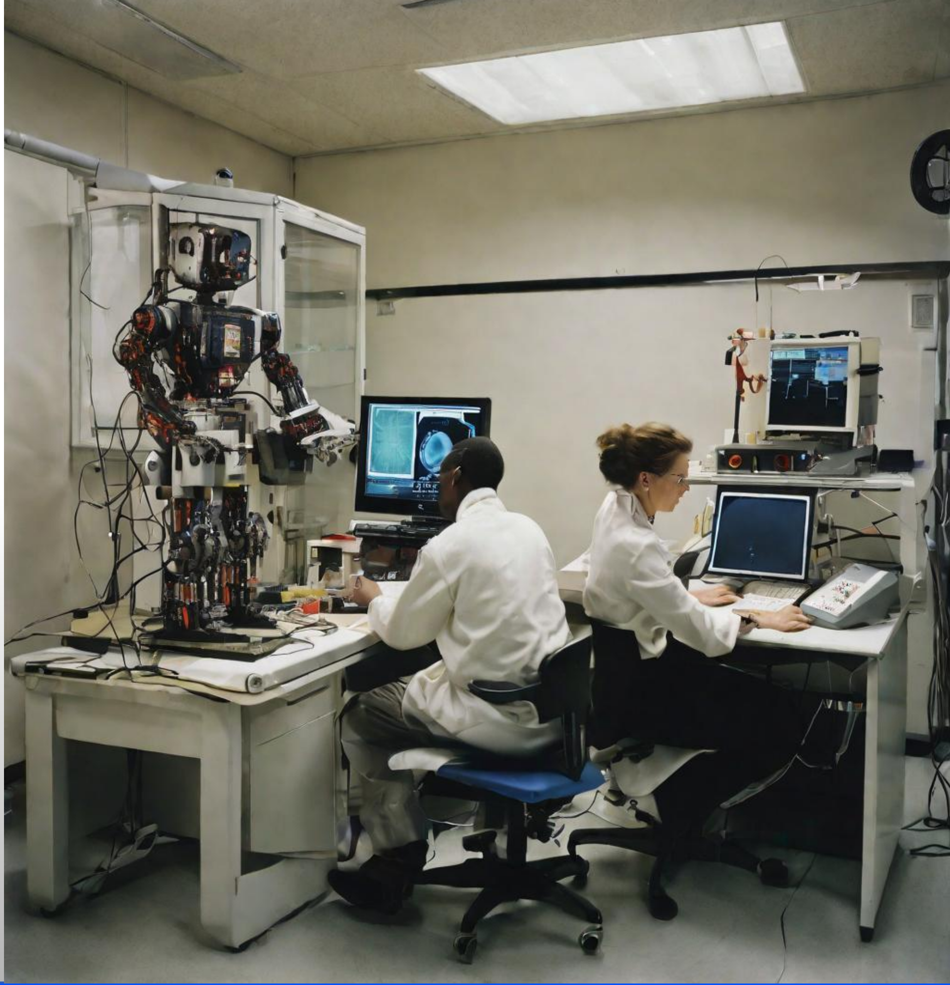
VM with single GPU

Use a single high memory GPU

VM with multiple GPUs

Use multiple lower spec GPU that can meet the requirements

Experimentation and Results



VM Deployment

VM with single GPU

Short response latency

Quick deployment

Max token processing limited to GPU memory

VM with multiple GPUs

Increased GPU memory footprint

Increased max token processing

Model readiness time increase

Response latency increase

Open LLMs - Tokens

Falcon 40B

256 tokens, the model seemed to perform adequately

Increasing of max token size:

Observed hallucinations

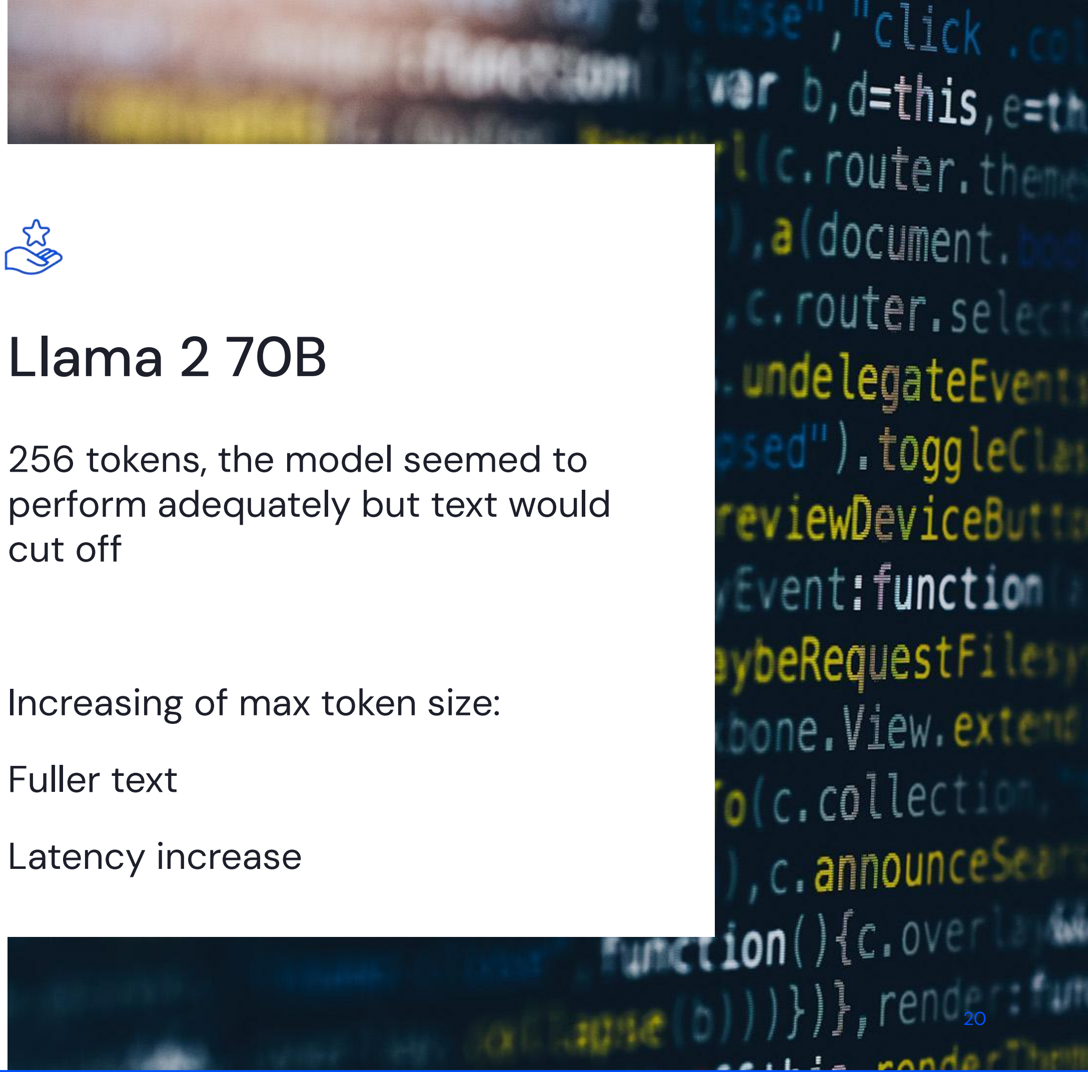
Llama 2 70B

256 tokens, the model seemed to perform adequately but text would cut off

Increasing of max token size:

Fuller text

Latency increase



Open LLMs – Languages

Falcon 40B

Spoke English -> Response English

Spoke French -> Response French

Mix of French/English -> Stuck to French

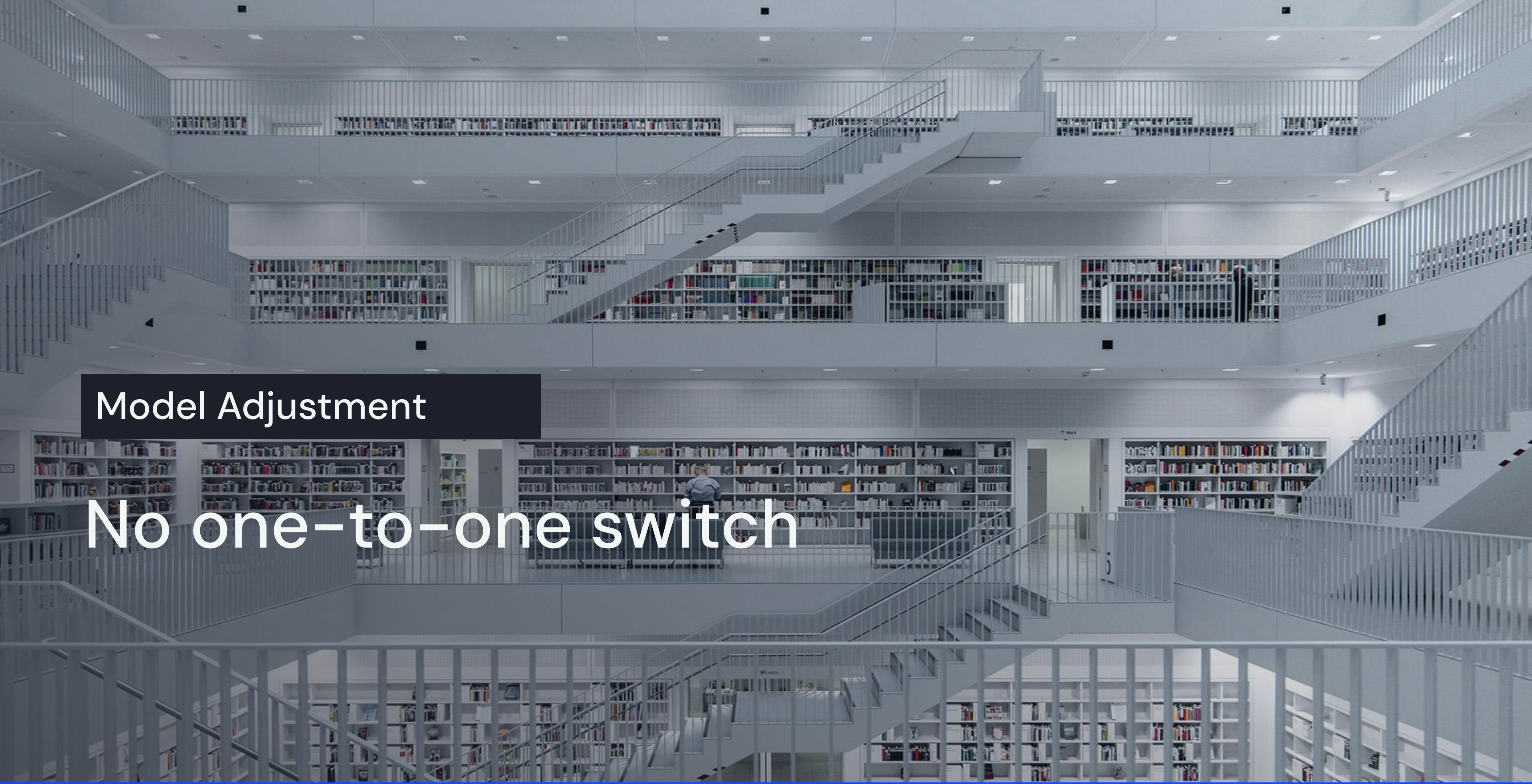
Llama 2 70B

Spoke English -> Response English

Spoke French -> Response French

Mix of French/English -> Stuck to French

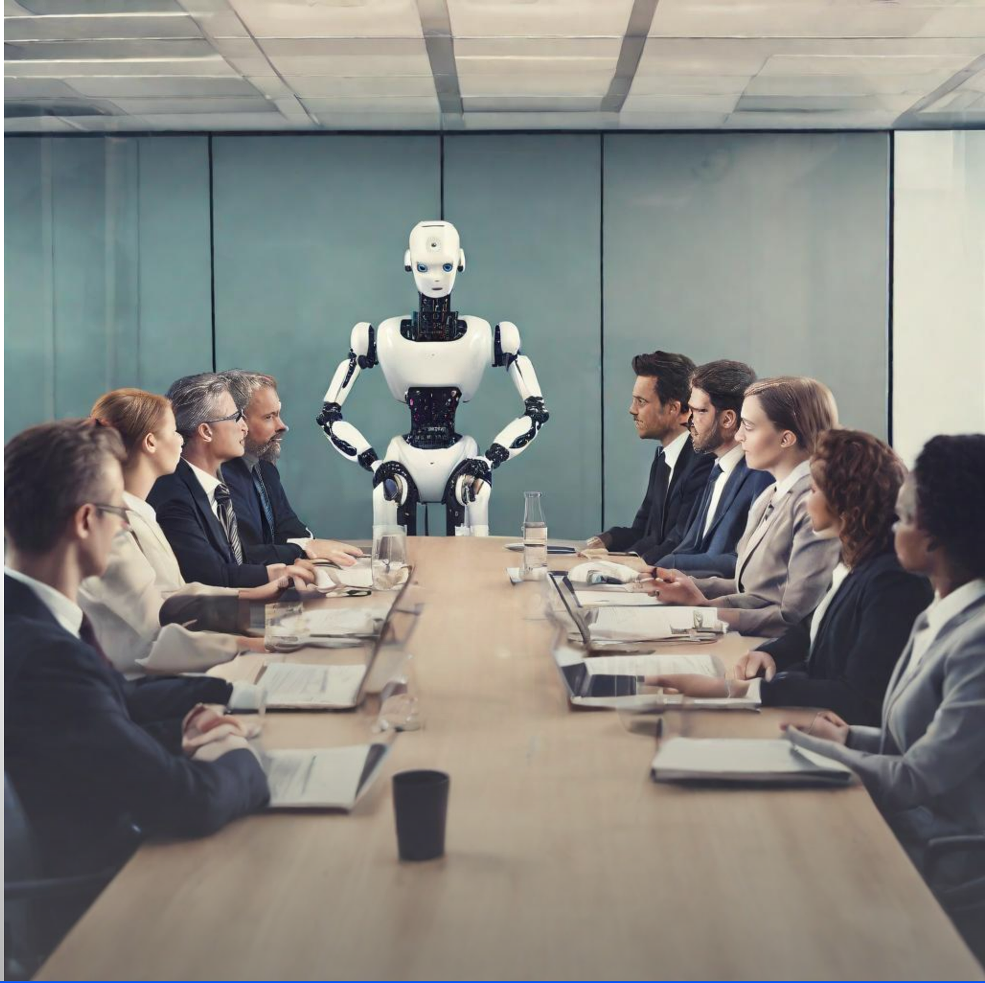
Observed hallucinations in French language



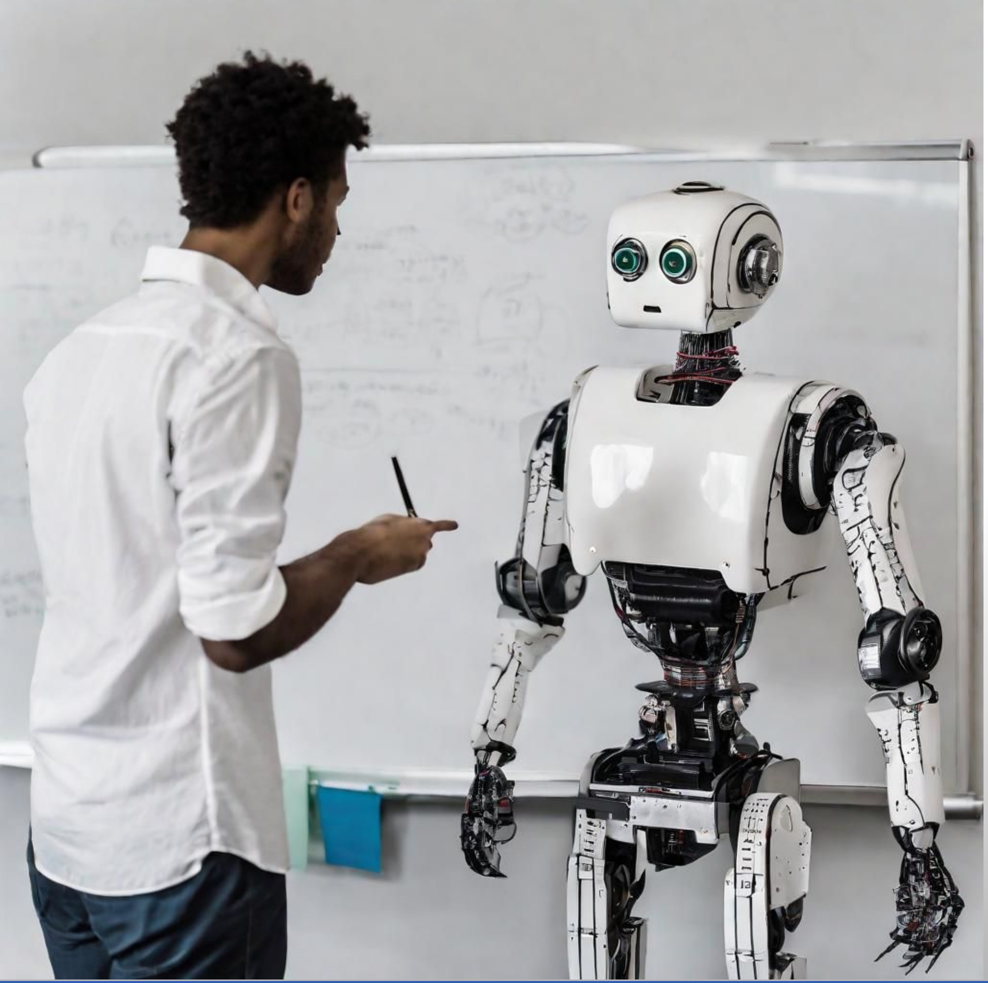
Model Adjustment

No one-to-one switch

How much is this going to cost?



Item	***Monthly estimate***
12 vCPU + 85 GB memory	US$594.27
1 NVIDIA A100 40GB	US$2,141.75
500 GB balanced persistent disk	US$55.00
Total	US$2,791.02



But tests show we need the highest GPU which costs much more…

Item	*Monthly estimate*
12 vCPU + 170 GB memory	US$883.67
1 NVIDIA A100 80GB	US$3,157.40
375 GB local SSD scratch disk	US$33.00
500 GB balanced persistent disk	US$55.00
Total	US$4,129.07

* The cost estimation does not include networking charges, human resources

Host on LLM

Great privacy and compliance

Customisation

Avoid vendor lock-in

Difficult to get adequate GPU

High hosting costs

Internal support costs

Security maintenance

Expertise needed in hosting model

Enterprise LLM Access

Easy entry point to LLMs

Minimal infrastructure

High speed implementation

ChatGPT 3.5 Turbo costs lowered

~~Slow API response time~~

Contract negotiations slow

Sensitive to privacy

https://openai.com/blog/function-calling-and-other-api-updates

Learnings

Explored deploying LLM

Based on use case, the lightweight Falcon 7B is possible to use internally

Use enterprise LLM API,and gradually seek if replacing functionality is possible

Thank you / Gracias / Merci / Danke
Grazie / Dank u wel / Obrigado