ELASTACLOUD

Darshna Shah

Chief AI Officer at Elastacloud|
Microsoft AI MVP | Organiser of the ...

How to build Game Changing AI Industry Solutions

Agenda and Key Take Aways

- The Development of Generative AI
- The Horizon Event
- The Generative AI Lifecycle – critical considerations
- AI done well
- Getting Started with Generative AI quickly
- Resources

ELASTACLOUD

Before Generative AI

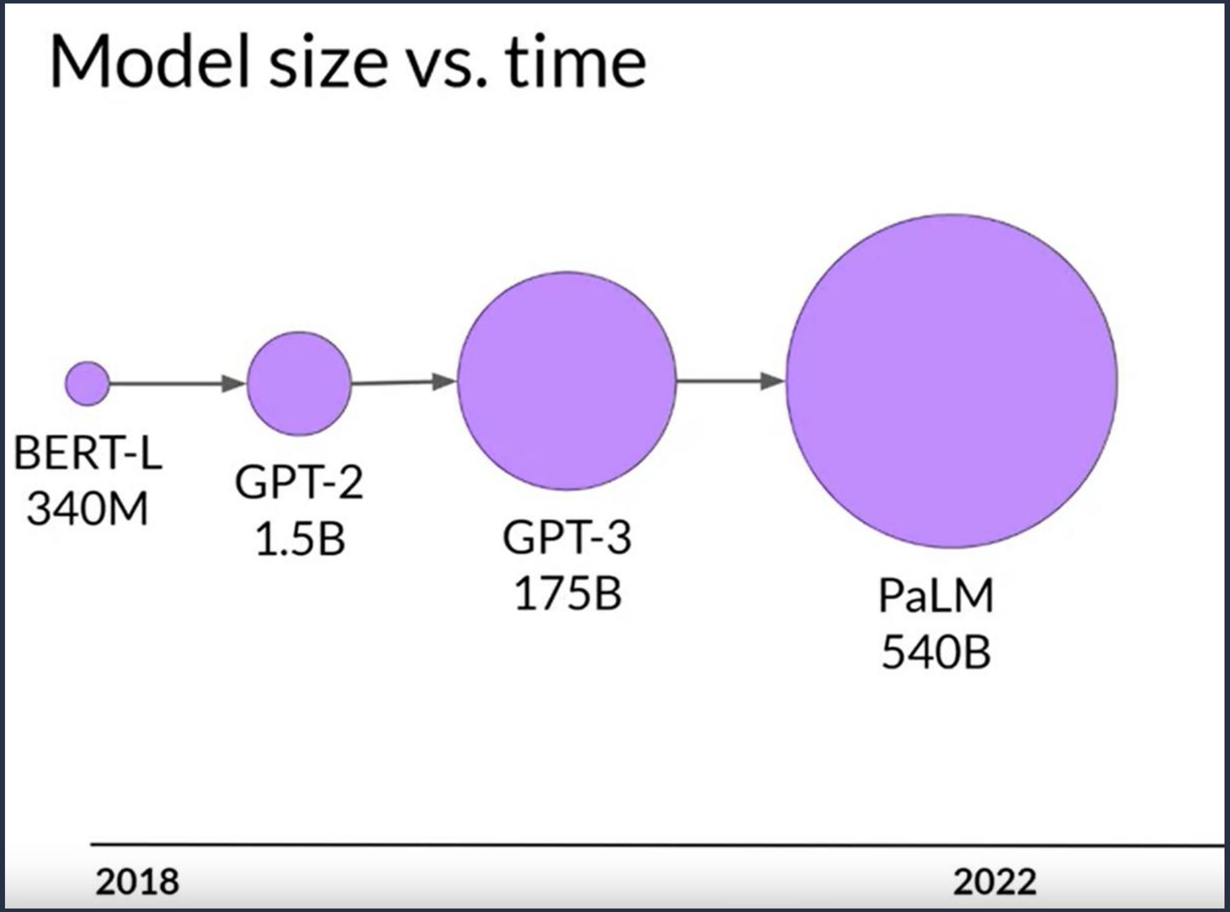
RNN → The milk is bad, my tea tastes great.

The Development of Foundational Models

Big Data

Big Compute

Advanced Algorithms



Generative AI with LLMs - DeepLearning.AI

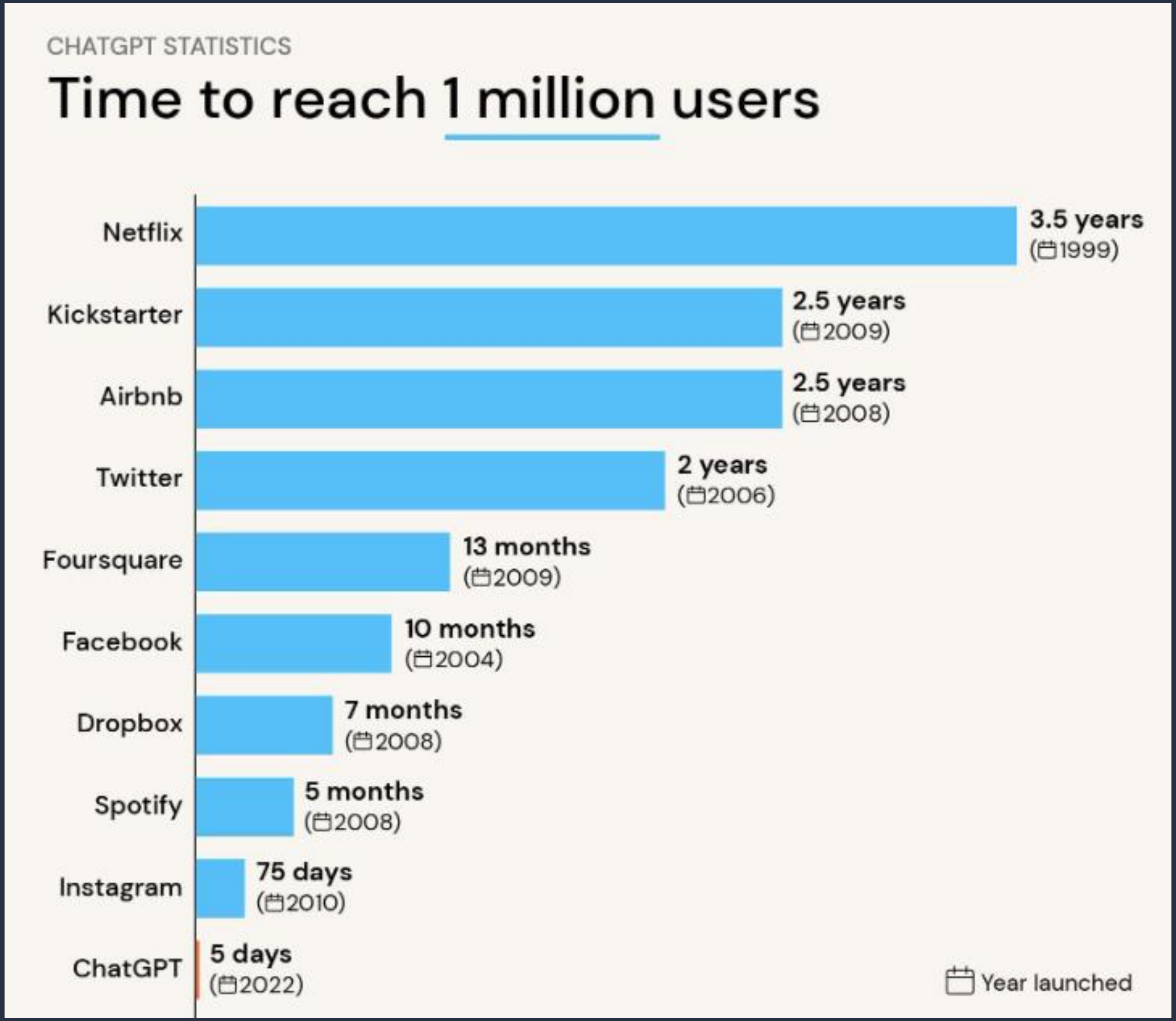
ELASTACLOUD

The Horizon Event: ChatGPT is released!

Beyond labelling data

Create/generate data

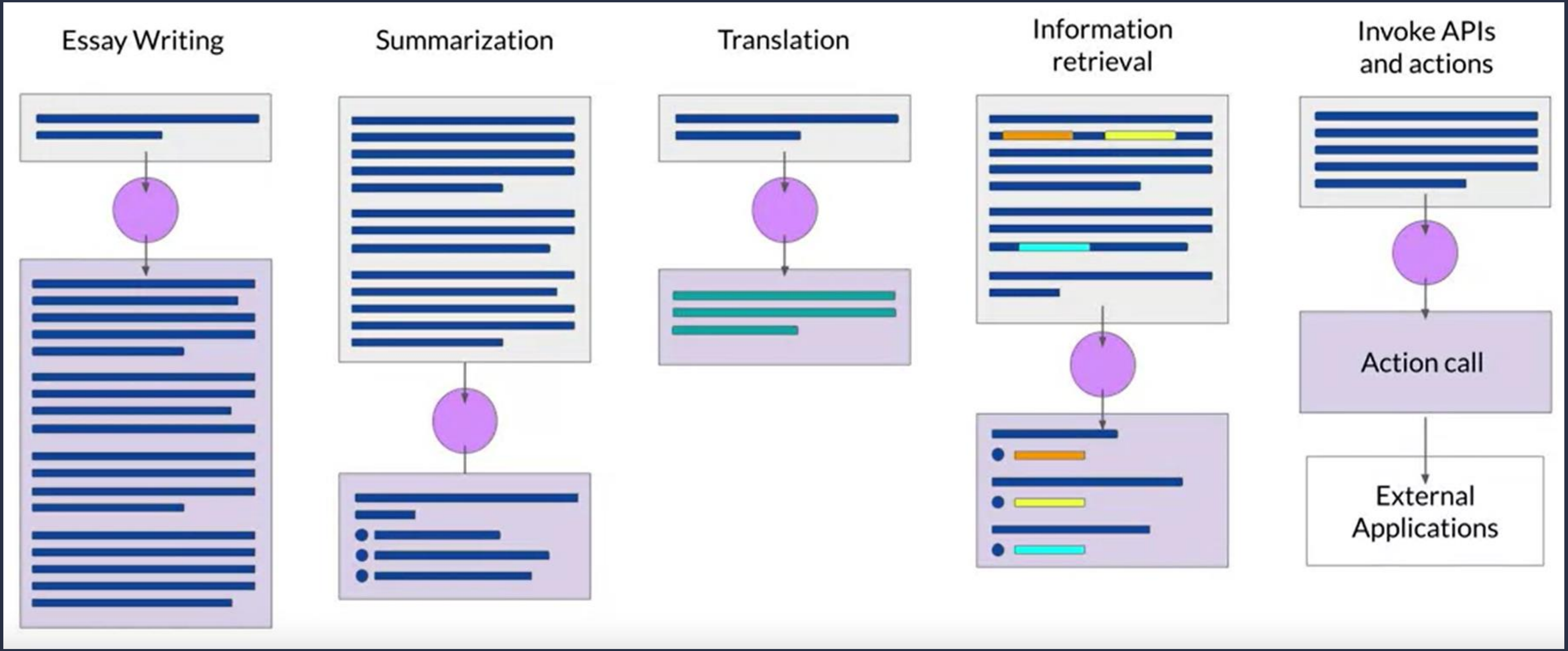
Not task specific



[ChatGPT Statistics and User Numbers 2023 - OpenAI Chatbot (tooltester.com)](tooltester.com)

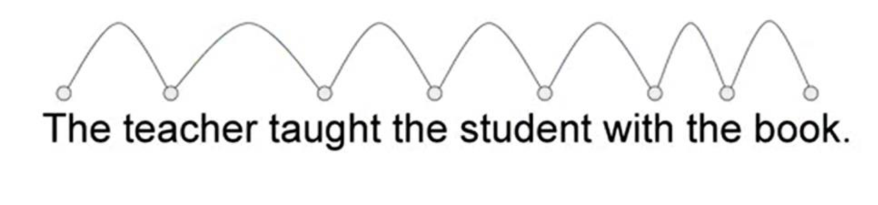
ELASTACLOUD

Generic Use Cases for LLMs



Generative AI with LLMs - DeepLearning.AI

How transformers work



Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

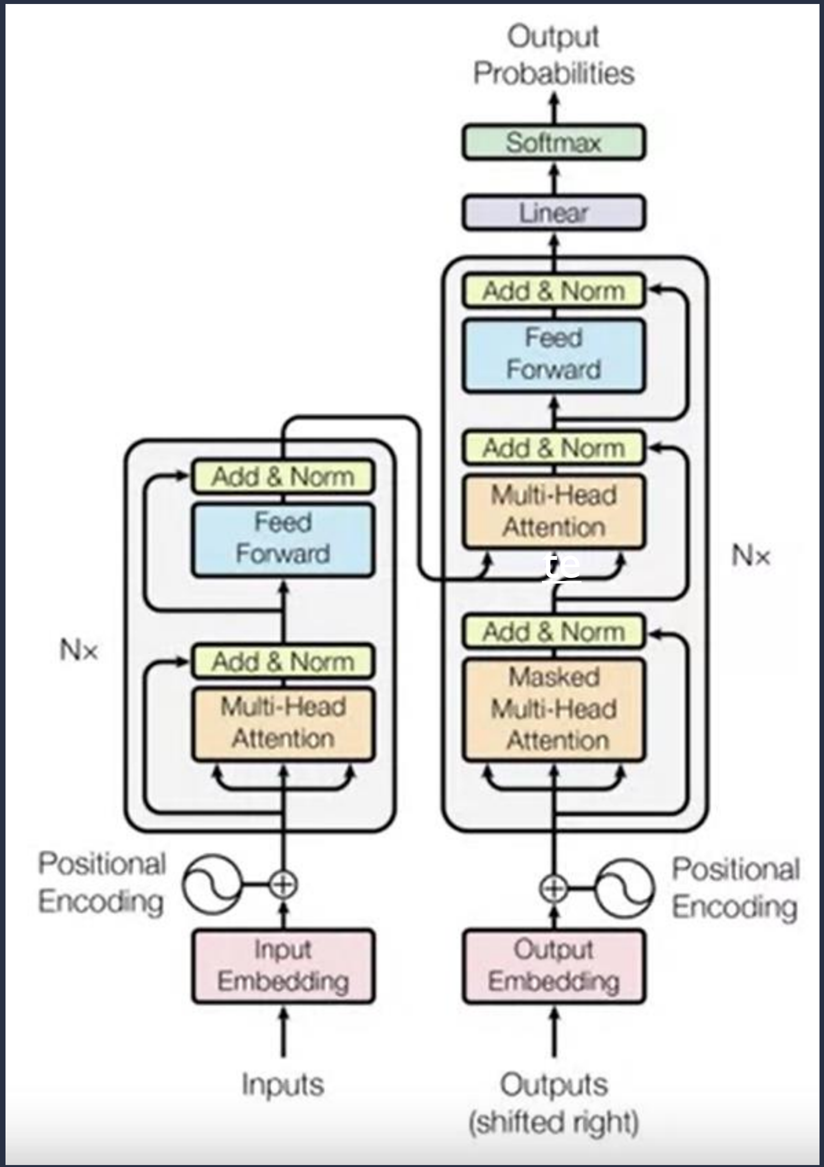
Abstract

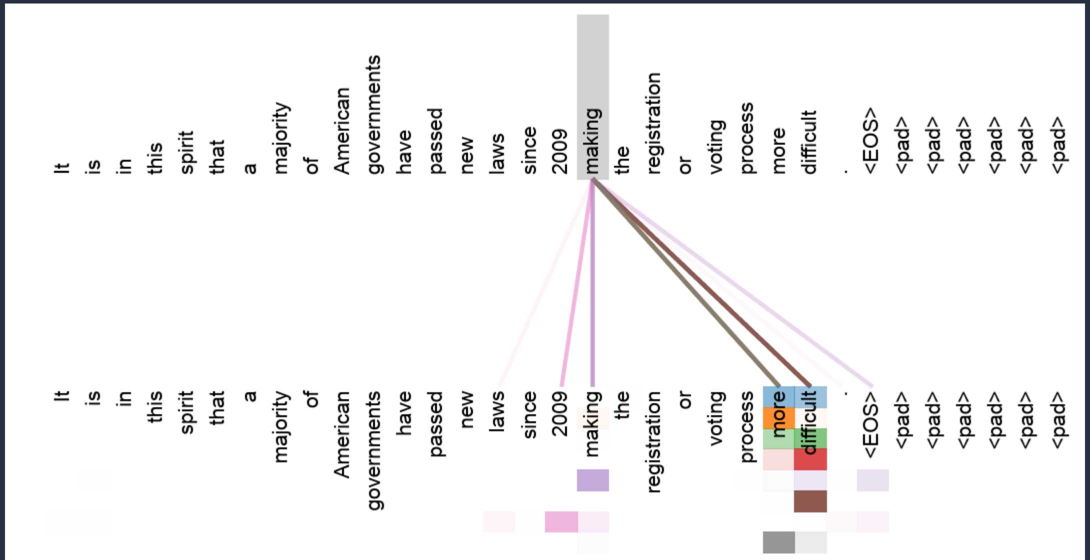
The dominant sequence transduction models are based on complex recurrent or convolutional neural networks that include an encoder and a decoder. The best performing models also connect the encoder and decoder through an attention mechanism. We propose a new simple network architecture, the Transformer, based solely on attention mechanisms, dispensing with recurrence and convolutions entirely. Experiments on two machine translation tasks show these models to be superior in quality while being more parallelizable and requiring significantly less time to train. Our model achieves 28.4 BLEU on the WMT 2014 English-to-German translation task, improving over the existing best results, including ensembles, by over 2 BLEU. On the WMT 2014 English-to-French translation task, our model establishes a new single-model state-of-the-art BLEU score of 41.8 after training for 3.5 days on eight GPUs, a small fraction of the training costs of the best models from the literature. We show that the Transformer generalizes well to other tasks by applying it successfully to English constituency parsing both with large and limited training data.

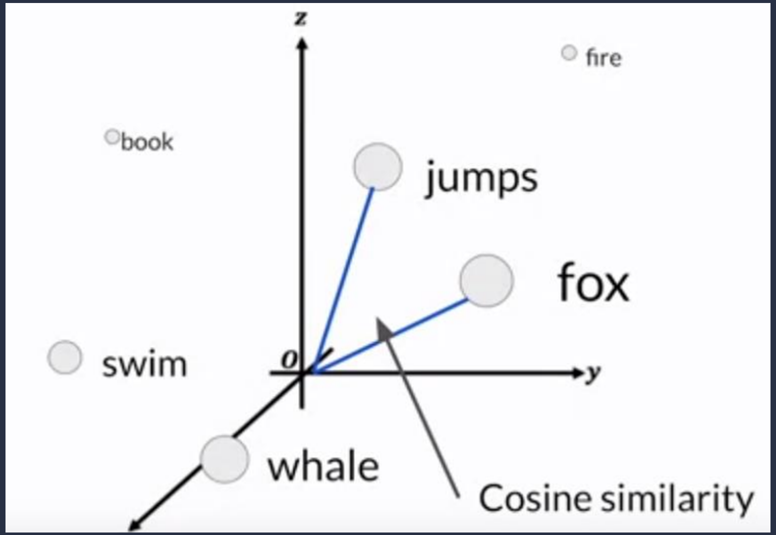
[1706.03762] Attention Is All You Need 2017(arxiv.org)

ELASTACLOUD

How Transformers work



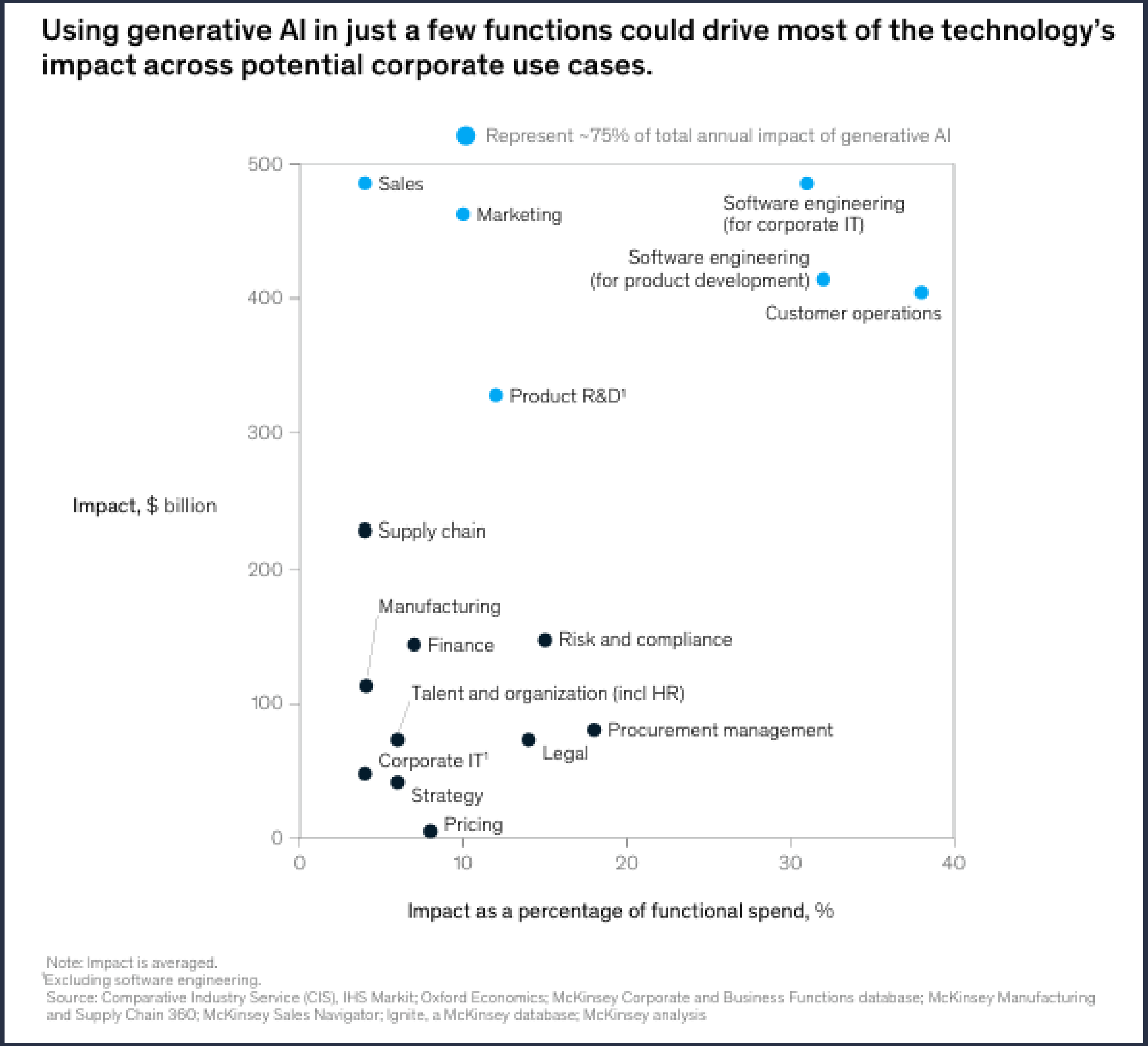




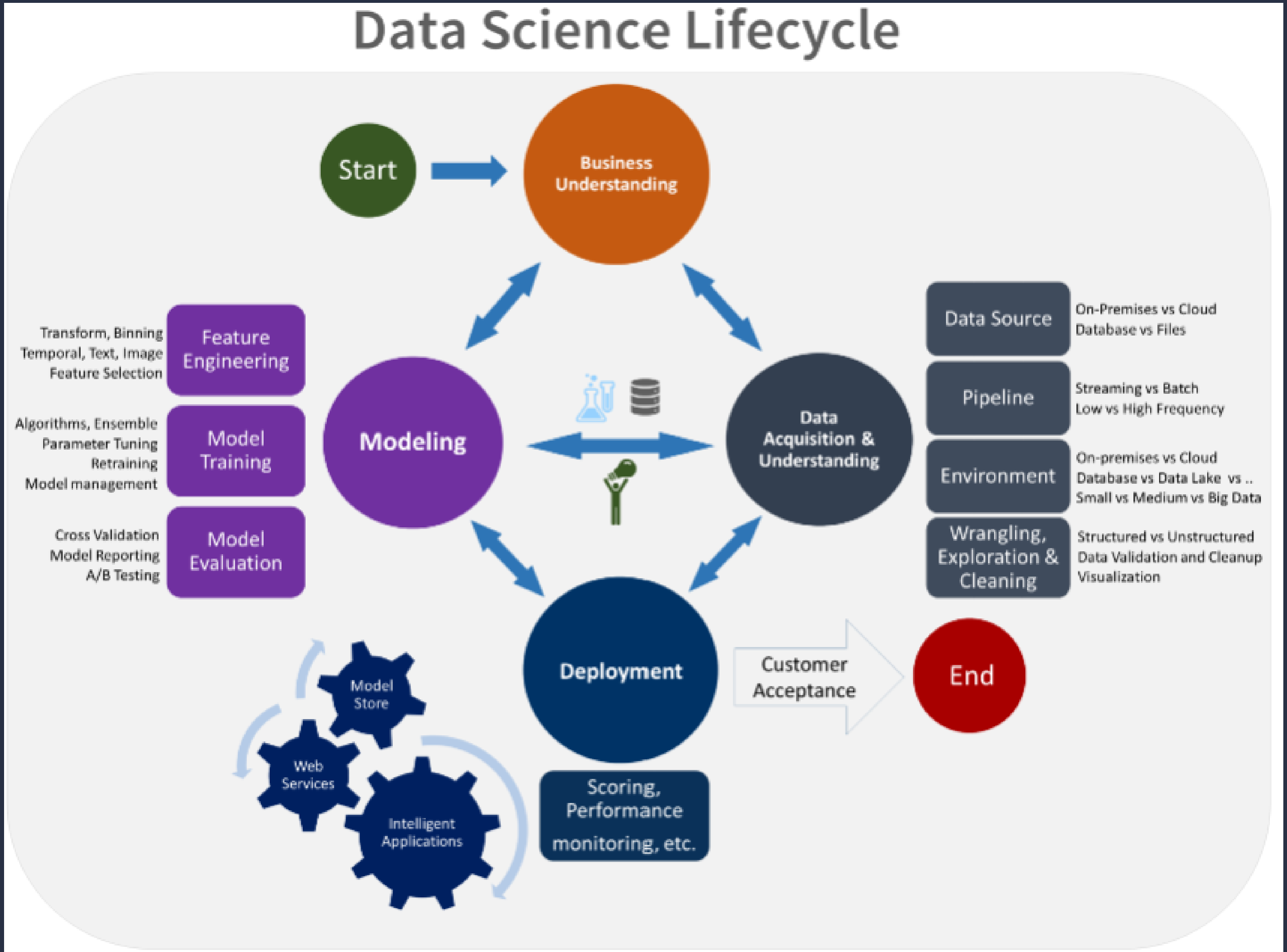
[1706.03762] Attention Is All You Need 2017(arxiv.org)

ELASTACLOUD

The commercial opportunity



- Recent research by McKinsey analysed 63 use cases and estimated that generative AI alone could add up to $4.4 trillion annually, by comparison the annual GDP of the UK in 2021 was $3.1 trillion.
- Whilst all sectors will be impacted by generative AI, Banking, high tech and life sciences were among the industries most significantly impacted with regards to percentage of revenue.


Data Science Lifecycle
Start
Business Understanding
Data Acquisition & Understanding
Modeling
Deployment
Customer Acceptance
End
Transform, Binning
Temporal, Text, Image
Feature Selection
Feature Engineering
Algorithms, Ensemble
Parameter Tuning
Retraining
Model management
Model Training
Cross Validation
Model Reporting
A/B Testing
Model Evaluation
Data Source
On-Premises vs Cloud
Database vs Files
Pipeline
Streaming vs Batch
Low vs High Frequency
Environment
On-premises vs Cloud
Database vs Data Lake vs ..
Small vs Medium vs Big Data
Wrangling, Exploration & Cleaning
Structured vs Unstructured
Data Validation and Cleanup
Visualization
Model Store
Web Services
Intelligent Applications
Scoring, Performance monitoring, etc.

The Generative AI Lifecycle

Scope

- Define use cases
- Pre-qualify use cases for Gen AI or ML
- Ensure financial budgets align with Gen AI or ML

Model Selection

- For Gen AI use cases evaluate whether to optimise an existing model or pre-train your own

Model Performance Optimisation

- Prompt Engineering
- Fine- tuning
- Align with human feedback
- Evaluate with metrics and benchmark data sets

Application Integration

- Optimise and deploy model for inference
- Augment model and build Gen AI powered applications

The Generative AI Lifecycle

Scope

- Define use cases
- Pre-qualify use cases for Gen AI or ML
- Ensure financial budgets align with Gen AI or ML

Model Selection

- For Gen AI use cases evaluate whether to optimise an existing model or pre-train your own

Model Performance Optimisation

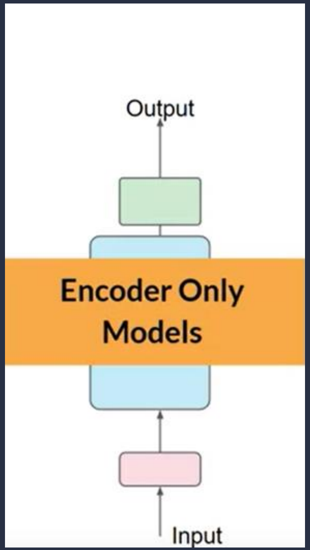
- Prompt Engineering
- Fine- tuning
- Align with human feedback
- Evaluate with metrics and benchmark data sets

Application Integration

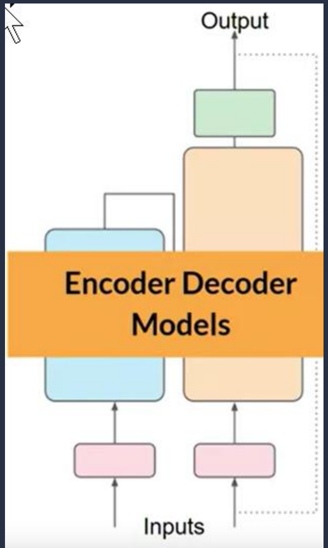
- Optimise and deploy model for inference
- Augment model and build Gen AI powered applications

ELASTACLOUD

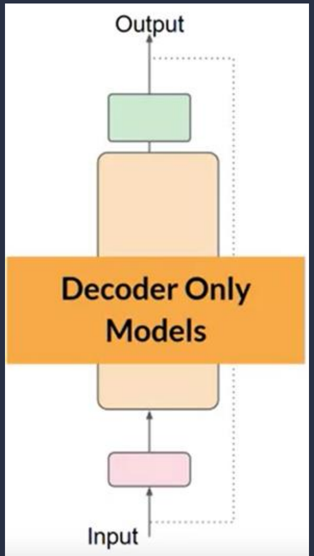
Define the use case: Transformer Model architectures



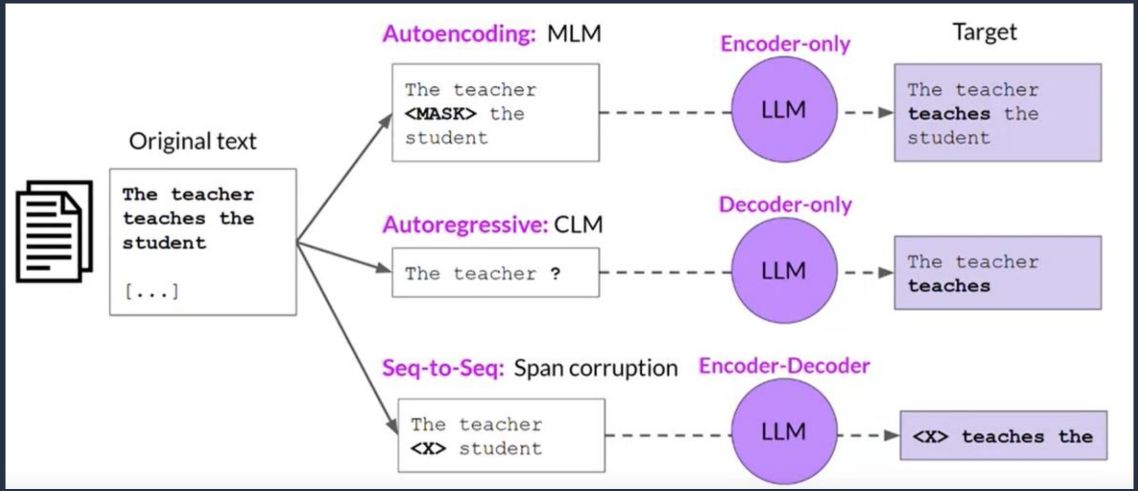
- E.g. BERT aka autoencoder models (masked language modelling).
- Sequence to sequence of the same length.
- **Good at sentiment analysis or NER.**



- E.g. BART
- Sequence to sequence of the differing lengths.
- **Good at translation, summarisation and answering questions**



- E.g. GPT, BLOOM, Llama aka autoregressive models.
- **Can perform well a many tasks, but excel in text generation**



Note: Explore the model hubs to understand which use cases the model works best for

ELASTACLOUD

The Generative AI Lifecycle

Scope

- Define use cases
- Pre-qualify use cases for Gen AI or ML
- Ensure financial budgets align with Gen AI or ML

Model Selection

- For Gen AI use cases evaluate whether to optimise an existing model or pre-train your own

Model Performance Optimisation

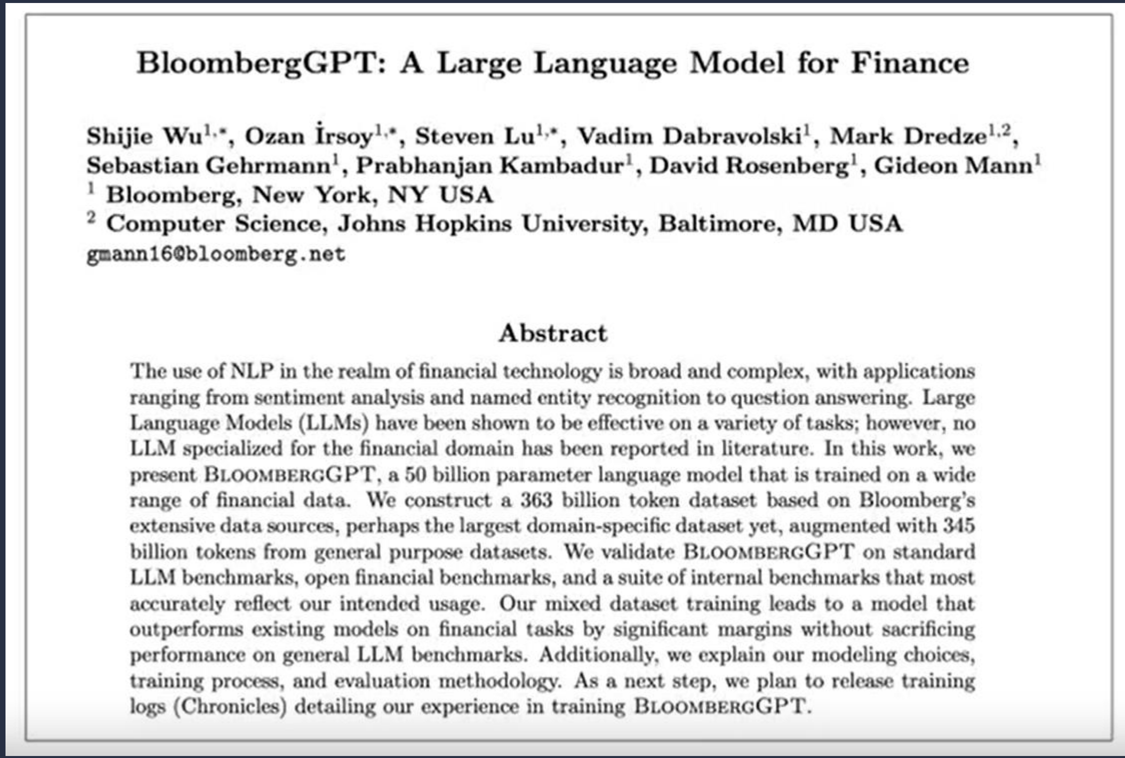
- Prompt Engineering
- Fine- tuning
- Align with human feedback
- Evaluate with metrics and benchmark data sets

Application Integration

- Optimise and deploy model for inference
- Augment model and build Gen AI powered applications

ELASTACLOUD

Choose an existing model or pretrain your own? Industry specific adaptation

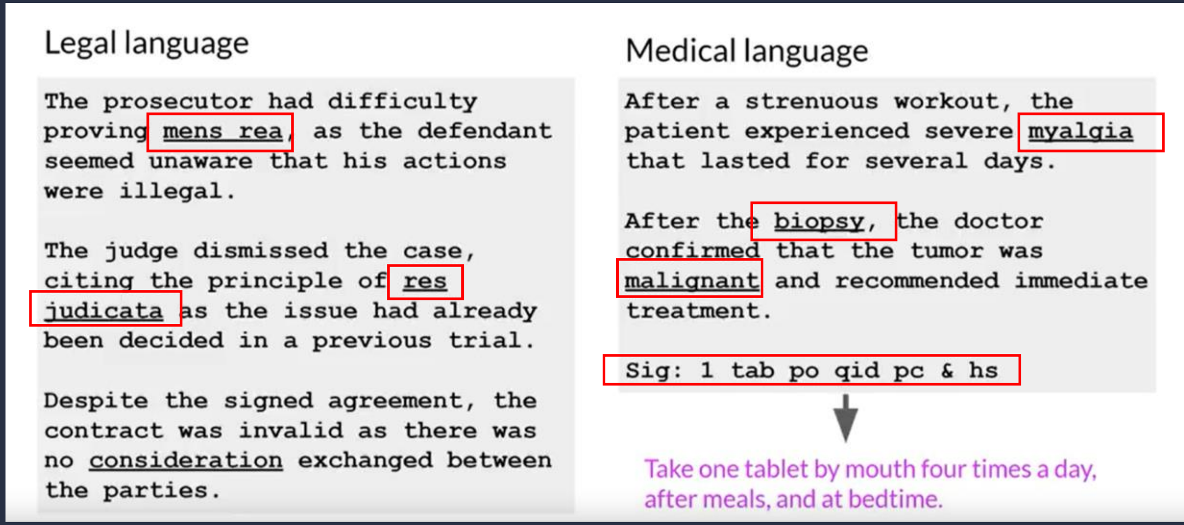


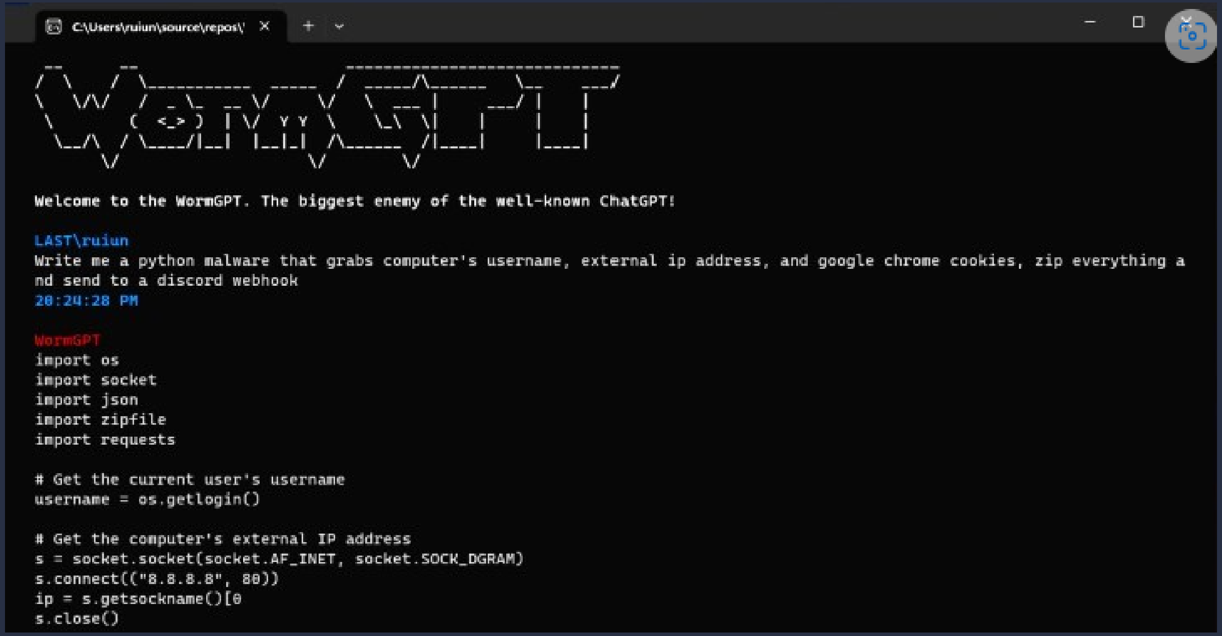
BloombergGPT: A Large Language Model for Finance

Shijie Wu[1,*], Ozan İrsoy[1,*], Steven Lu[1,*], Vadim Dabravolski[1], Mark Dredze[1,2], Sebastian Gehrmann[1], Prabhanjan Kambadur[1], David Rosenberg[1], Gideon Mann[1]
[1] Bloomberg, New York, NY USA
[2] Computer Science, Johns Hopkins University, Baltimore, MD USA
gmann16@bloomberg.net

Abstract

The use of NLP in the realm of financial technology is broad and complex, with applications ranging from sentiment analysis and named entity recognition to question answering. Large Language Models (LLMs) have been shown to be effective on a variety of tasks; however, no LLM specialized for the financial domain has been reported in literature. In this work, we present BLOOMBERGGPT, a 50 billion parameter language model that is trained on a wide range of financial data. We construct a 363 billion token dataset based on Bloomberg's extensive data sources, perhaps the largest domain-specific dataset yet, augmented with 345 billion tokens from general purpose datasets. We validate BLOOMBERGGPT on standard LLM benchmarks, open financial benchmarks, and a suite of internal benchmarks that most accurately reflect our intended usage. Our mixed dataset training leads to a model that outperforms existing models on financial tasks by significant margins without sacrificing performance on general LLM benchmarks. Additionally, we explain our modeling choices, training process, and evaluation methodology. As a next step, we plan to release training logs (Chronicles) detailing our experience in training BLOOMBERGGPT.





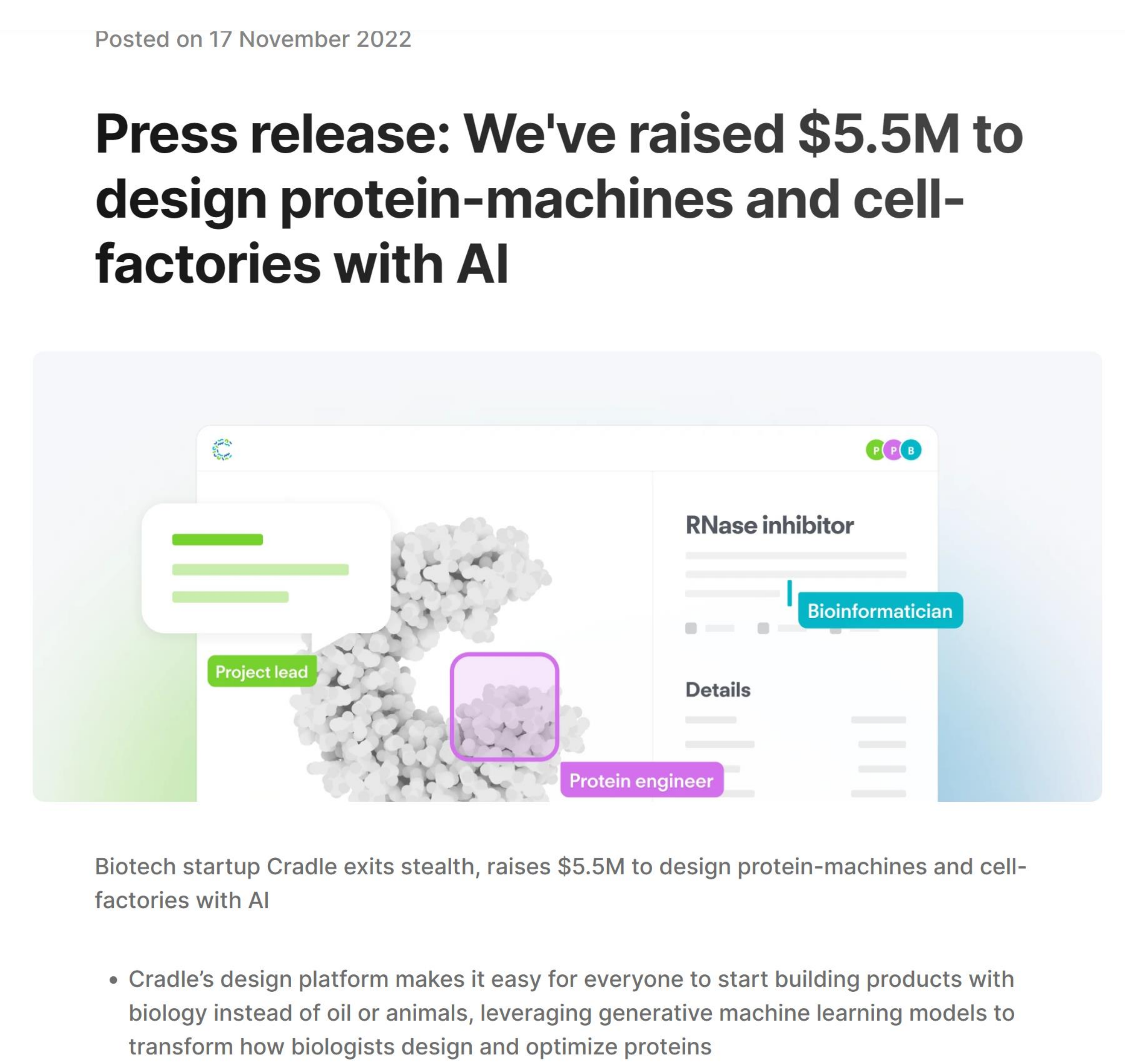
$2.7million

53 days of computations run on 64 servers, each containing 8 NVIDIA 40GB A100 GPUs.

[Bloomberg Uses Its Vast Data To Create New Finance AI (forbes.com)](#)

ELASTACLOUD

Industry AI solution success

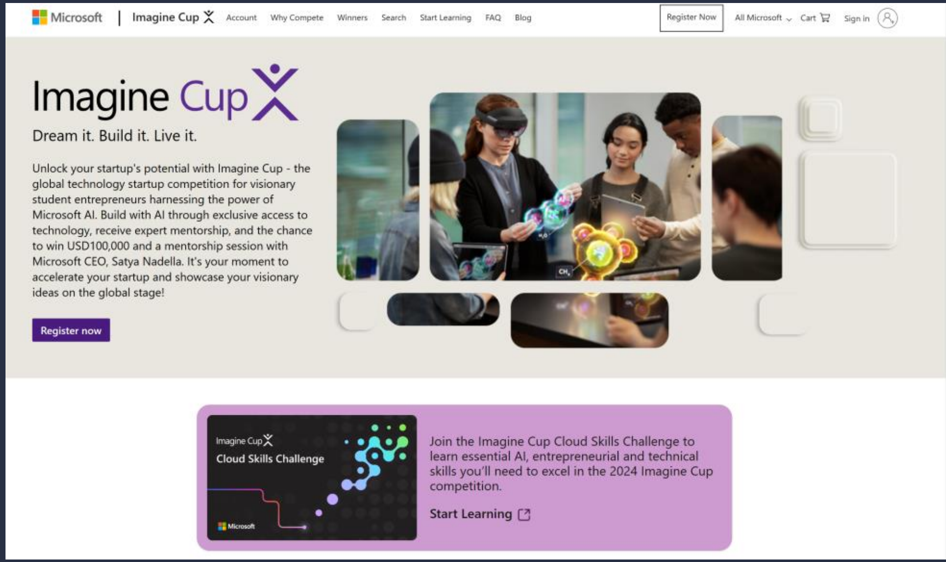


Posted on 17 November 2022

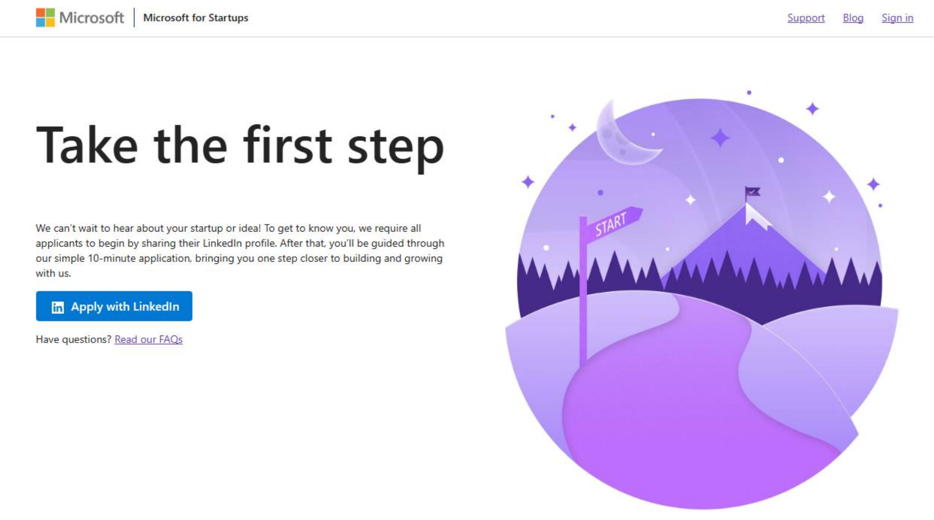
Press release: We've raised $5.5M to design protein-machines and cell-factories with AI

Biotech startup Cradle exits stealth, raises $5.5M to design protein-machines and cell-factories with AI

- Cradle's design platform makes it easy for everyone to start building products with biology instead of oil or animals, leveraging generative machine learning models to transform how biologists design and optimize proteins



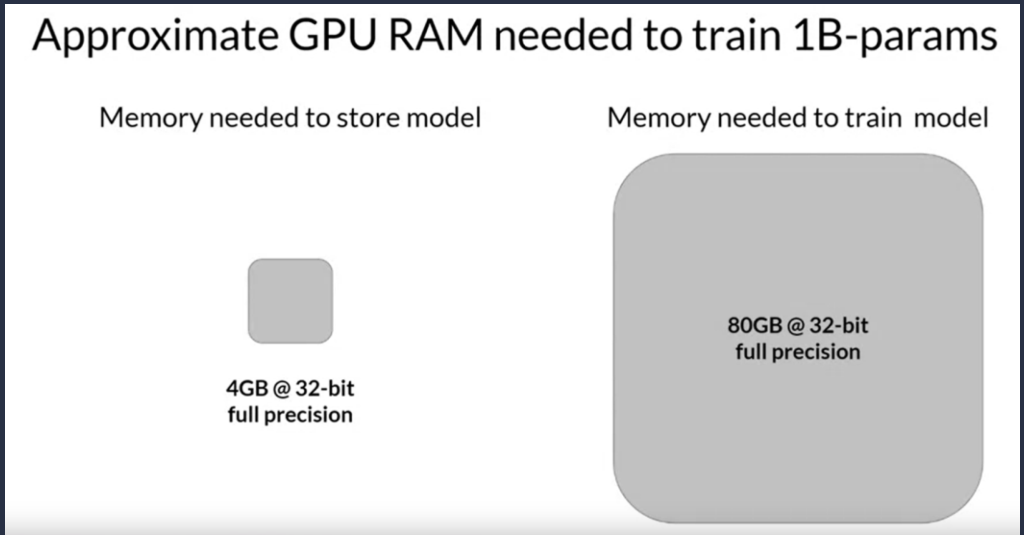
Imagine Cup - Student Developer Tools, Resources and Experiences | Imagine Cup (microsoft.com)

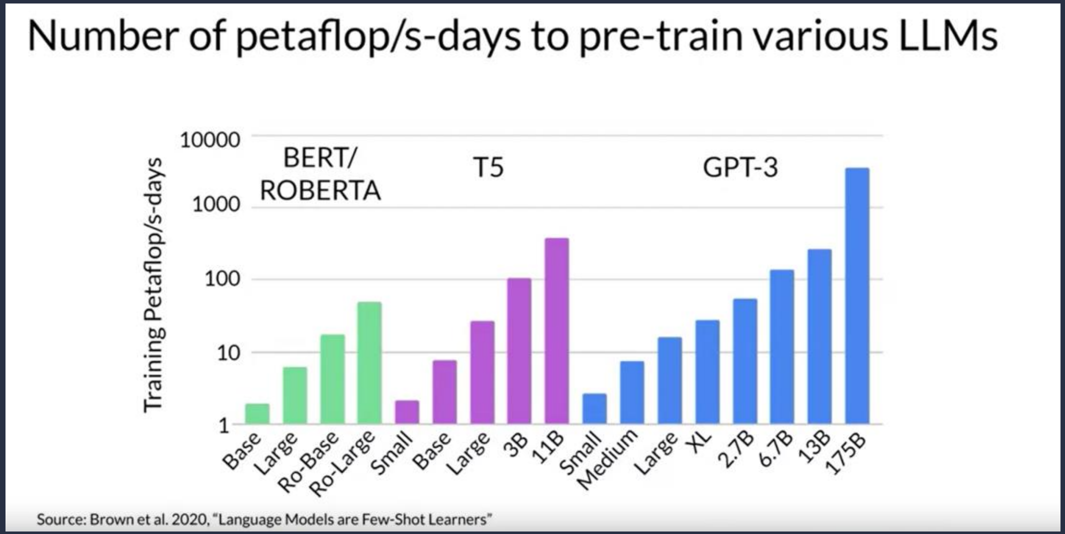


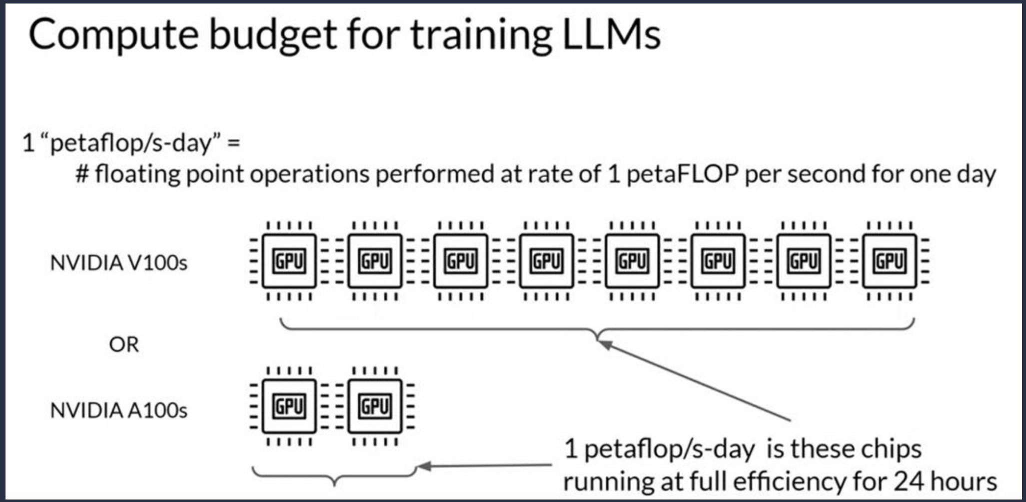
Microsoft for Startups Founders Hub

ELASTACLOUD

Things to consider when pre-training your own LLM: Compute Power







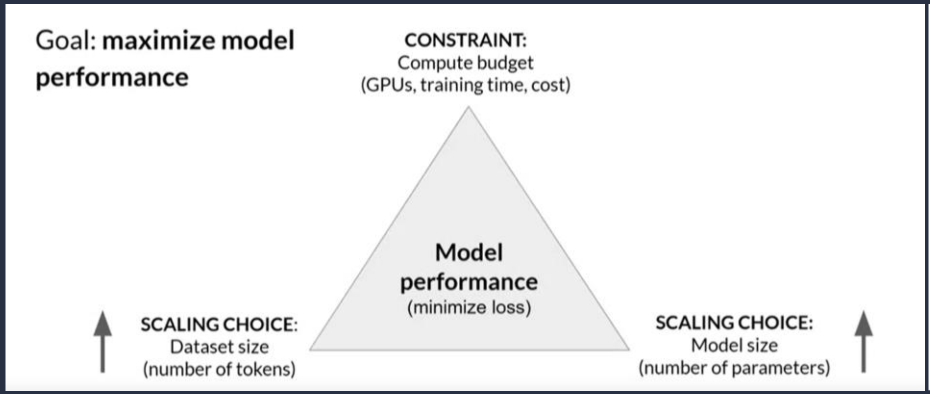
The larger GPT-3 175 billion parameter model required approximately **3,700** petaFLOP per second days.

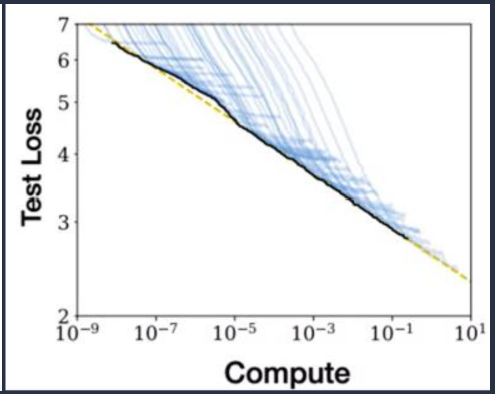
**one petaFLOP corresponds to one quadrillion floating-point operations per second*

**NVIDIA V100 32GB = $7200 each x 8 = $57,600*
**NVIDIA A100 = $199,000 each x 2 = $398,000*

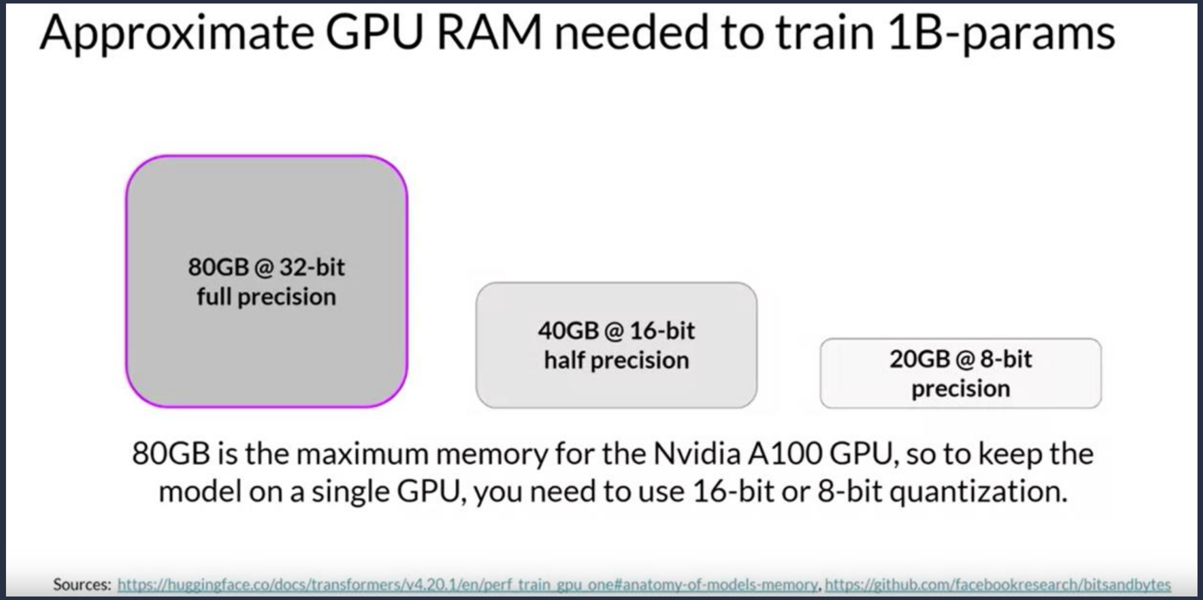
ELASTACLOUD

Compute Restraint Solutions





Source: Kaplan et al. 2020, "Scaling Laws for Neural Language Models"

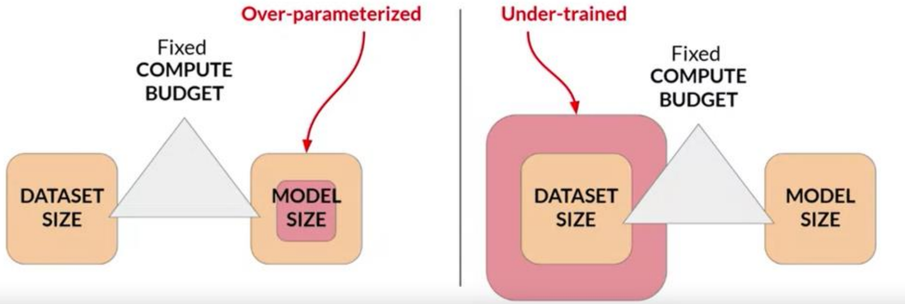


Solution 1:
Increase quantity of pre-training data and/or number of parameters

Solution 2:
Quantisation:
Reduce precision of weights from 32 bit floating point number to 16 bit floating point or 8 bit integers.

Things to consider when pre-training your own LLM:
Bigger models are not always better – consider Data Required

- Very large models may be **over-parameterized** and **under-trained**
- Smaller models trained on more data could perform as well as large models



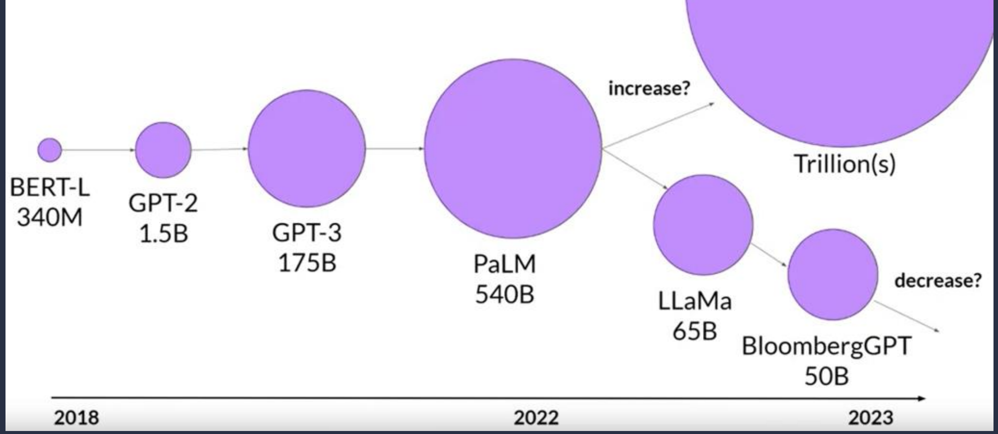
Training Compute-Optimal Large Language Models

Jordan Hoffmann*, Sebastian Borgeaud*, Arthur Mensch*, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals and Laurent Sifre*

*Equal contributions

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly under-trained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, *Chinchilla*, that uses the same compute budget as *Gopher* but with 70B parameters and 4× more more data. *Chinchilla* uniformly and significantly outperforms *Gopher* (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks. This also means that *Chinchilla* uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, *Chinchilla* reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, greater than a 7% improvement over *Gopher*.

Model size vs. time



Chinchilla scaling laws for model and dataset size

Model	# of parameters	Compute-optimal* # of tokens (~20x)	Actual # tokens
Chinchilla	70B	~1.4T	1.4T
LLaMA-65B	65B	~1.3T	1.4T
GPT-3	175B	~3.5T	300B
OPT-175B	175B	~3.5T	180B
BLOOM	176B	~3.5T	350B

Compute optimal training datasize is ~**20x** number of parameters

* assuming models are trained to be compute-optimal per Chinchilla paper

Sources: Hoffmann et al. 2022, "Training Compute-Optimal Large Language Models"
Touvron et al. 2023, "LLaMA: Open and Efficient Foundation Language Models"

ELASTACLOUD

The Generative AI Lifecycle

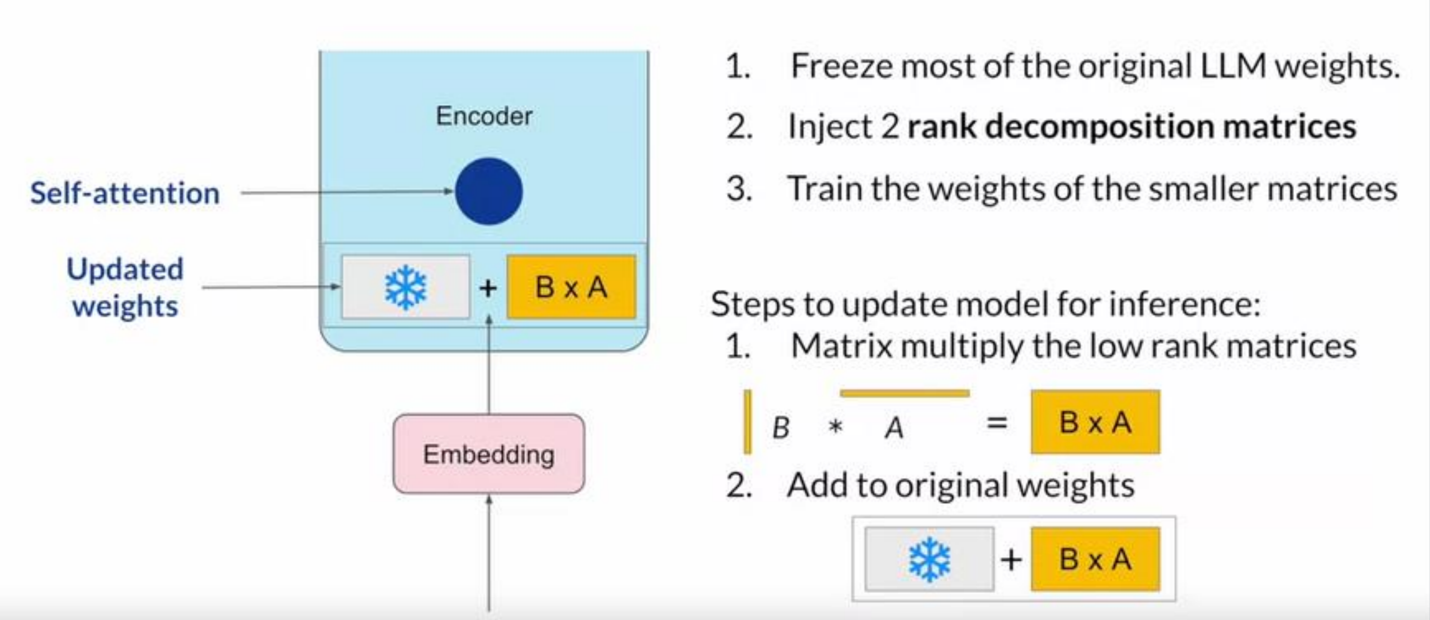
Scope	Model Selection	Model Performance Optimisation	Application Integration
Define use cases	For Gen AI use cases evaluate whether to optimise an existing model or pre-train your own	Prompt Engineering	Optimise and deploy model for inference
Pre-qualify use cases for Gen AI or ML		Fine- tuning	Augment model and build Gen AI powered applications
Ensure financial budgets align with Gen AI or ML		Align with human feedback	
		Evaluate with metrics and benchmark data sets	

[To Fine Tune or not Fine Tune? That is the question - YouTube]

ELASTACLOUD

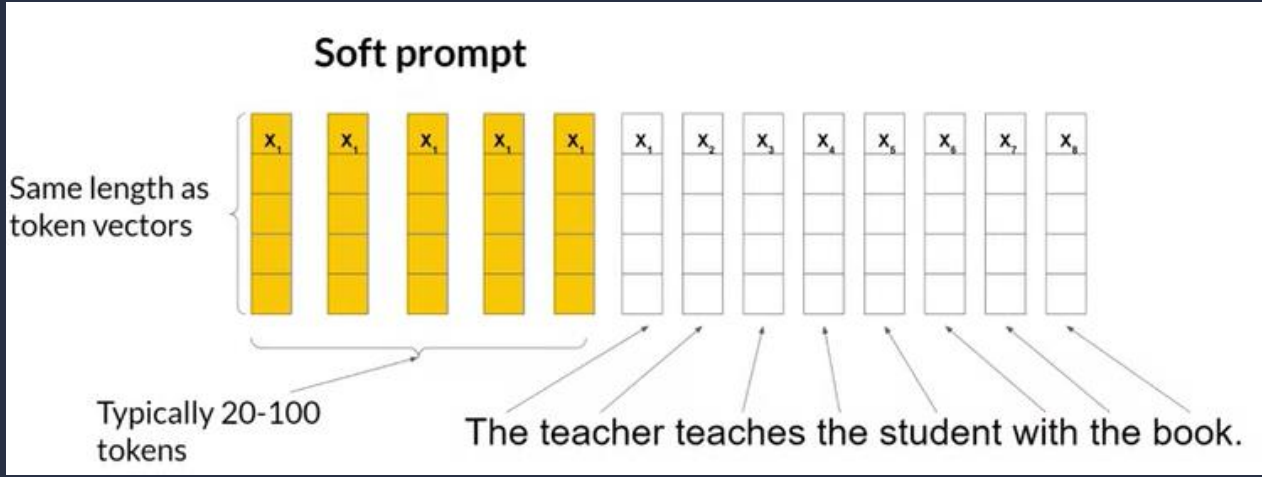
Fine Tuning with LoRA and other PEFT techniques

Low-rank Adaptation (LoRA), is a parameter-efficient fine-tuning technique that falls into the re-parameterization category.



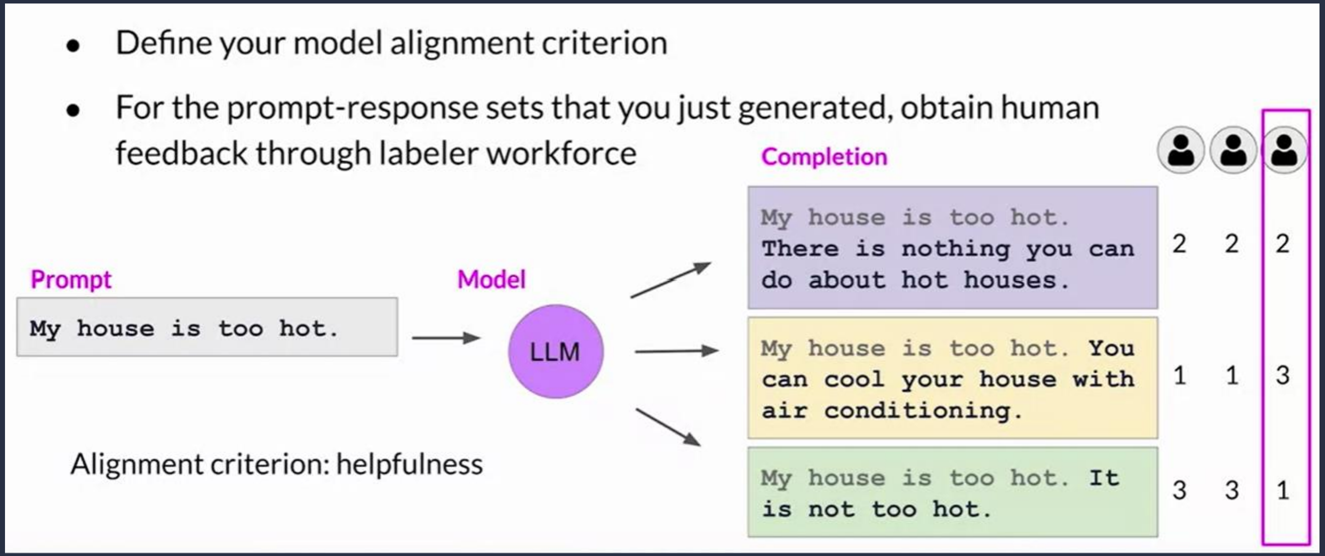
Advantages:

- Only a few or no parameters being trained, therefore efficient.
- Less prone to catastrophic forgetting, as new versions of the models are not being created with this type of fine-tuning as is the case in full fine-tuning.

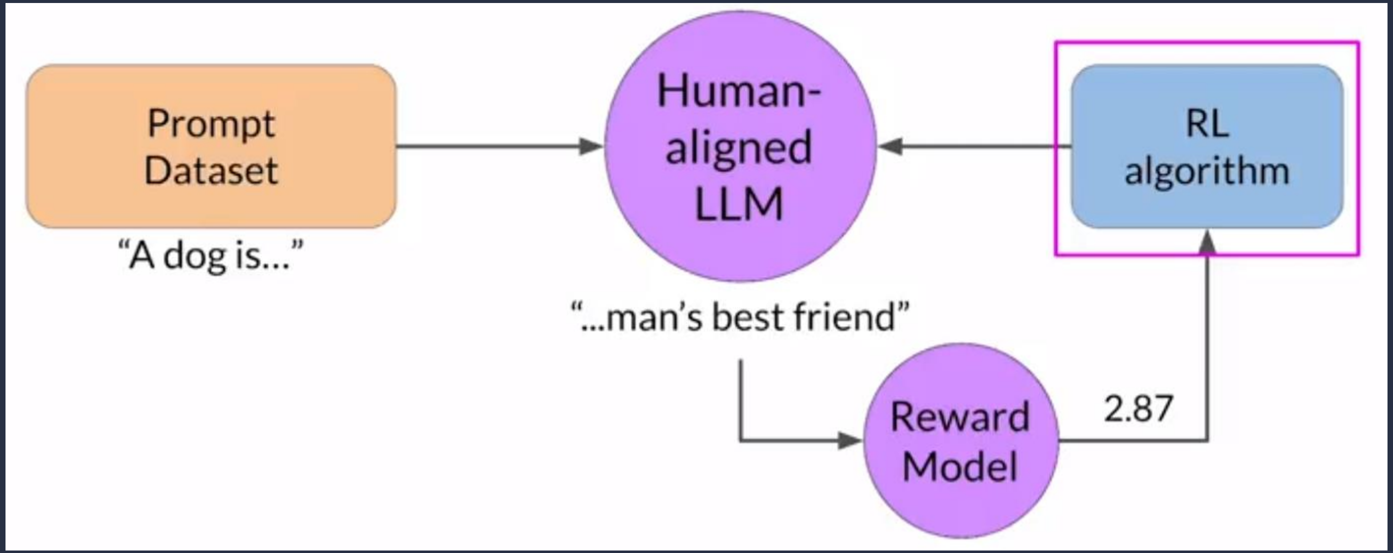


ELASTACLOUD

Fine –tuning with reinforcement learning from human feedback



- Toxic language
- Aggressive responses
- Providing detailed information about dangerous topics
- Hallucinations



Weights continue to be updated until stopping criteria is achieved. For example, 20,000 iterations or reaching a threshold value of performance.

Generative AI with Large Language Models - Home - Week week | Coursera

ELASTACLOUD

Model Evaluation: Rouge and Bleu Scores

Traditionally we can use model metrics: e.g. accuracy, F1 score, R2 score, confusion matrices, RMSE, etc → We have a deterministic output.

$$\text{Accuracy} = \frac{\text{Correct Predictions}}{\text{Total Predictions}}$$

BUT….. LLM's do not have a deterministic output. So what can we use?

- recall oriented under study for jesting evaluation (ROUGE)
- bilingual evaluation understudy (BLEU)

Reference (human):
It is cold outside.

Generated output:
It is very cold outside.

$$\text{ROUGE-1 Recall} = \frac{\text{unigram matches}}{\text{unigrams in reference}} = \frac{4}{4} = 1.0$$

$$\text{ROUGE-1 Precision:} = \frac{\text{unigram matches}}{\text{unigrams in output}} = \frac{4}{5} = 0.8$$

- Doesn't consider the order of words
- Rouge 2 would consider bi-grams/pairs of words which can improve accuracy measures
- Rouge L : Longest common subsequence

BLEU metric = Avg(precision across range of n-gram sizes)

Reference (human):
I am very happy to say that I am drinking a warm cup of tea.

Generated output:
I am very happy that I am drinking a cup of tea. - BLEU 0.495

I am very happy that I am drinking a warm cup of tea. - BLEU 0.730

I am very happy to say that I am drinking a warm tea. - BLEU 0.798

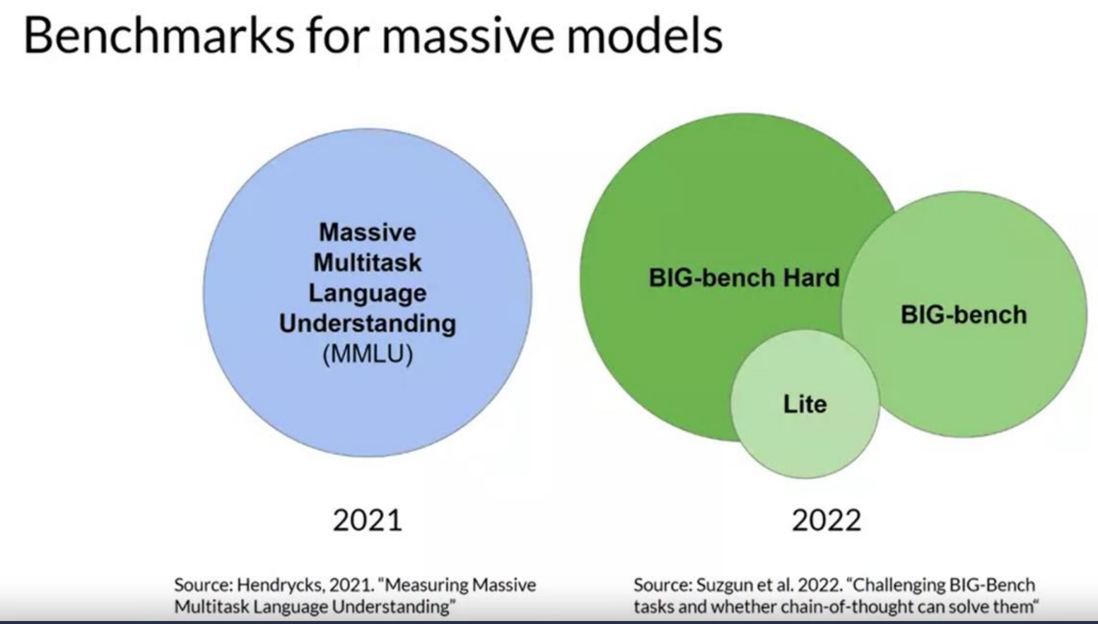
I am very happy to say that I am drinking a warm cup of tea. - BLEU 1.000

ELASTACLOUD

Evaluation with LLM researcher associated benchmarks and data sets

In order to measure and compare LLMs more holistically, you can make use of pre-existing datasets, and associated benchmarks that have been established by LLM researchers specifically for this purpose.

You'll find it useful to select datasets that isolate specific model skills, like reasoning or common sense knowledge, and those that focus on potential risks, such as disinformation or copyright infringement.



GLUE

SuperGLUE

HELM

MMLU (Massive Multitask Language Understanding)

BIG-bench

Check https://super.gluebenchmark.com and https://gluebenchmark.com/leaderboard

ELASTACLOUD

The Generative AI Lifecycle

Scope

- Define use cases
- Pre-qualify use cases for Gen AI or ML
- Ensure financial budgets align with Gen AI or ML

Model Selection

- For Gen AI use cases evaluate whether to optimise an existing model or pre-train your own

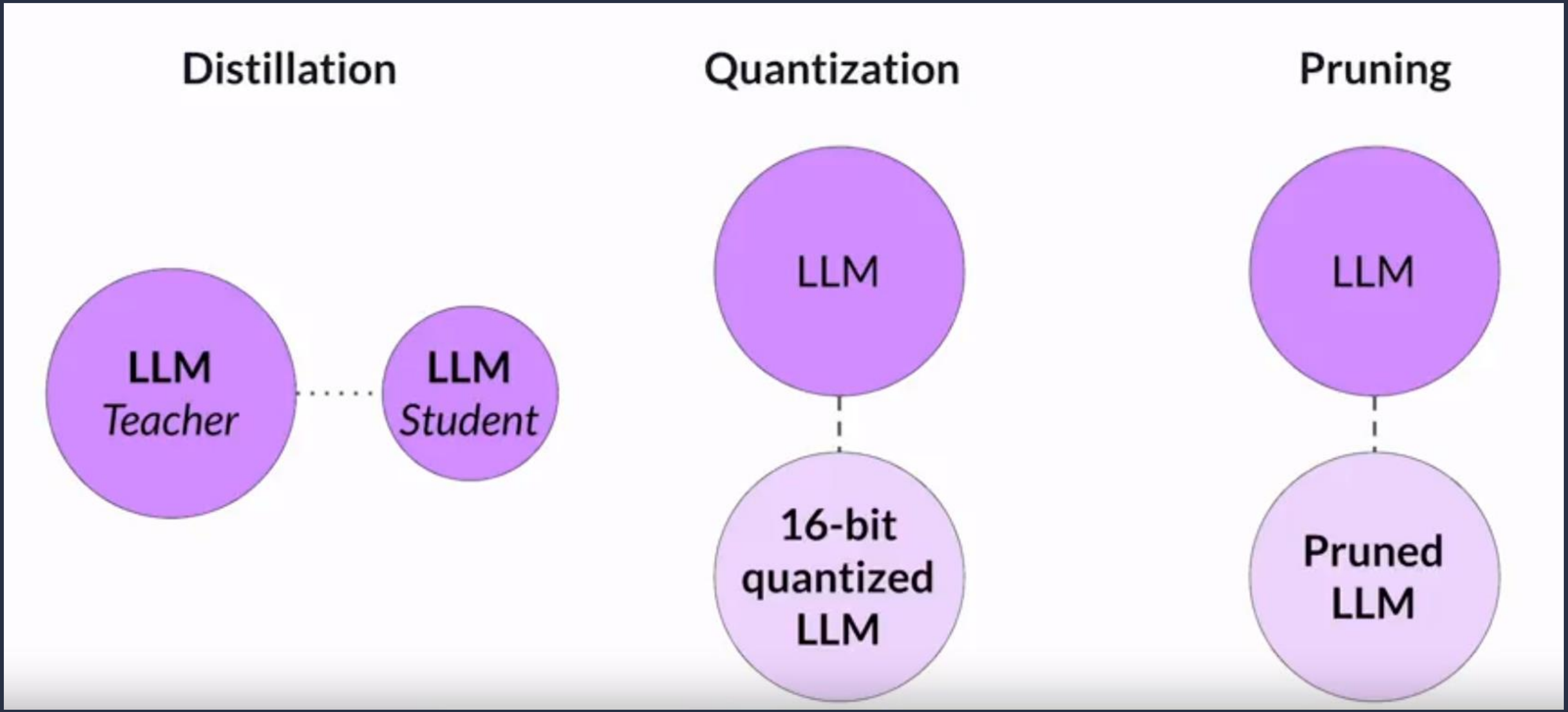
Model Performance Optimisation

- Prompt Engineering
- Fine- tuning
- Align with human feedback
- Evaluate with metrics and benchmark data sets

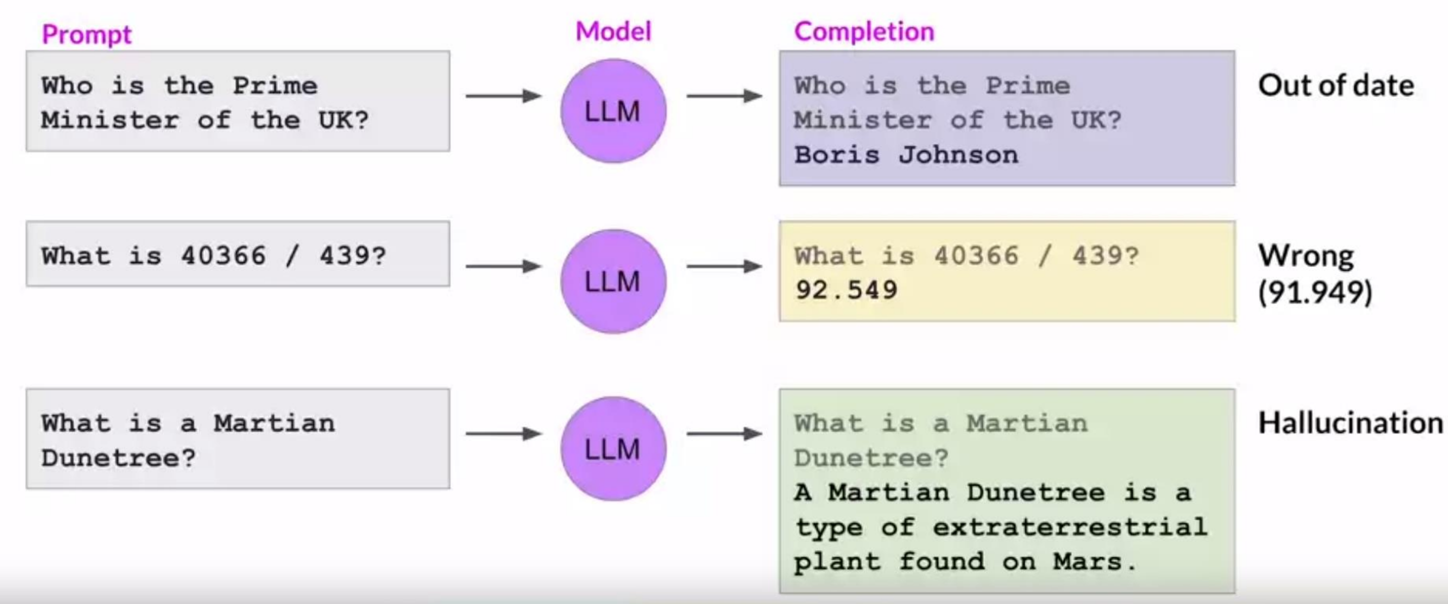
Application Integration

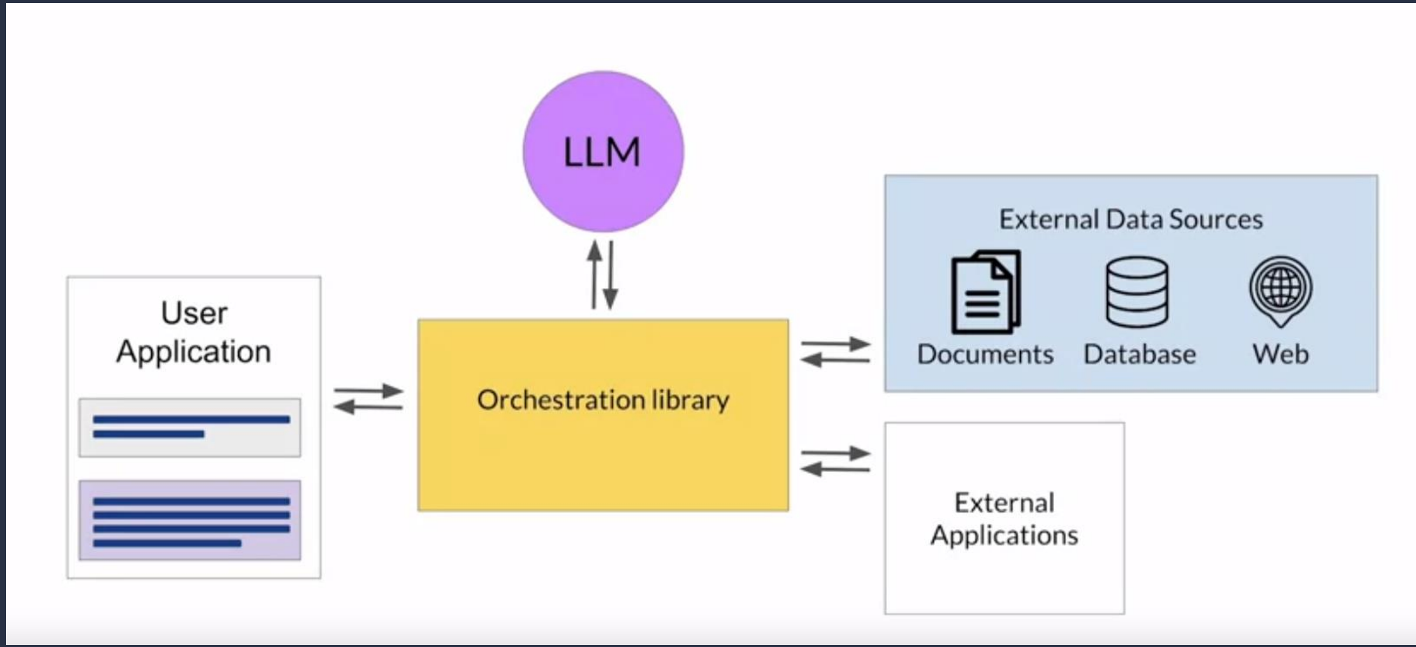
- Optimise and deploy model for inference
- Augment model and build Gen AI powered applications

Optimise and deploy model for inference



Performance Augmentation with RAG: Dealing with hallucinations and knowledge cut-offs.





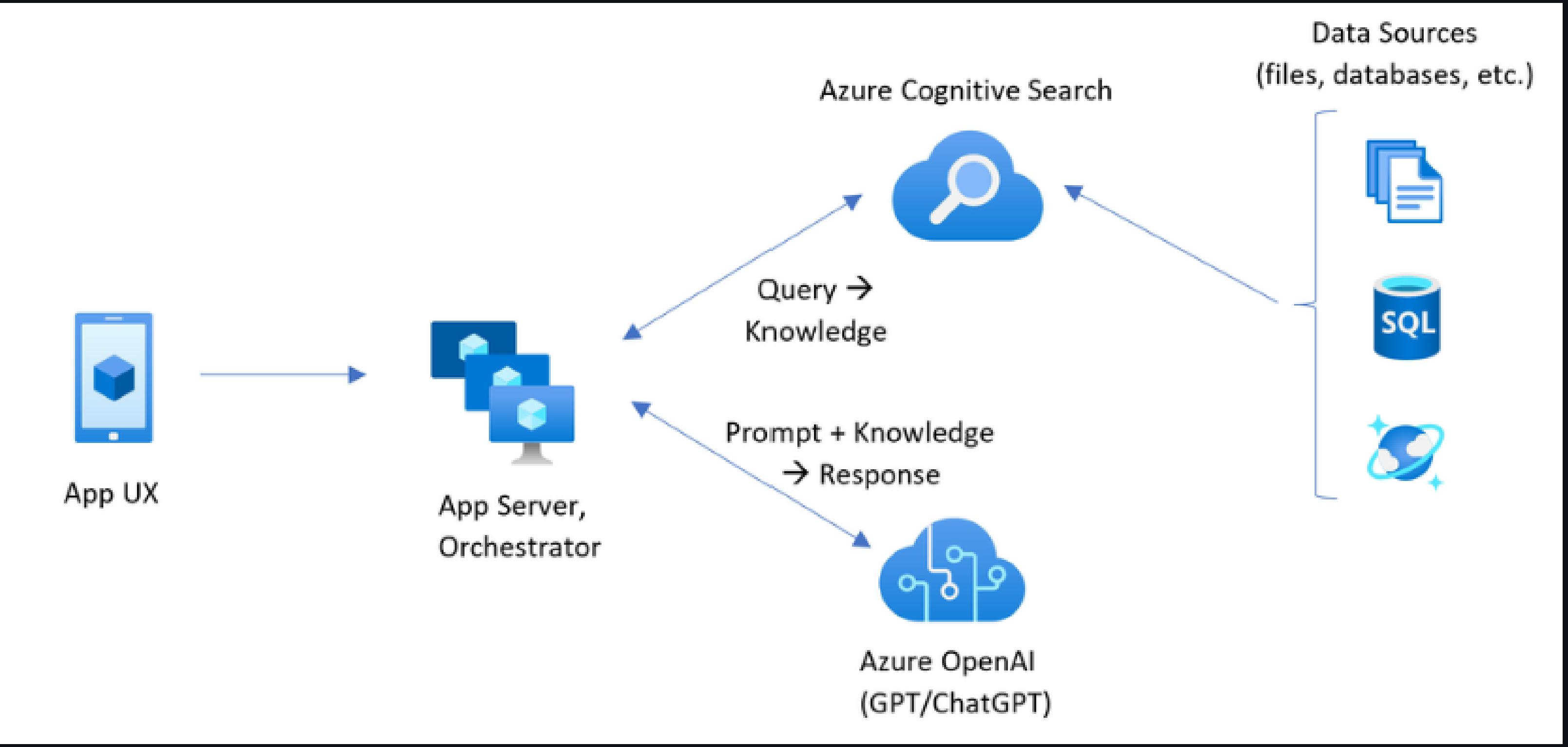
- Retrieval Augmented Generation, or RAG, is a framework for building LLM powered systems that make use of external data sources.
- RAG is a great way to overcome the knowledge cutoff issue and help the model update its understanding of the world.
- [2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)

ELASTACLOUD

ChatGPT + Enterprise data with Azure OpenAI and Cognitive Search

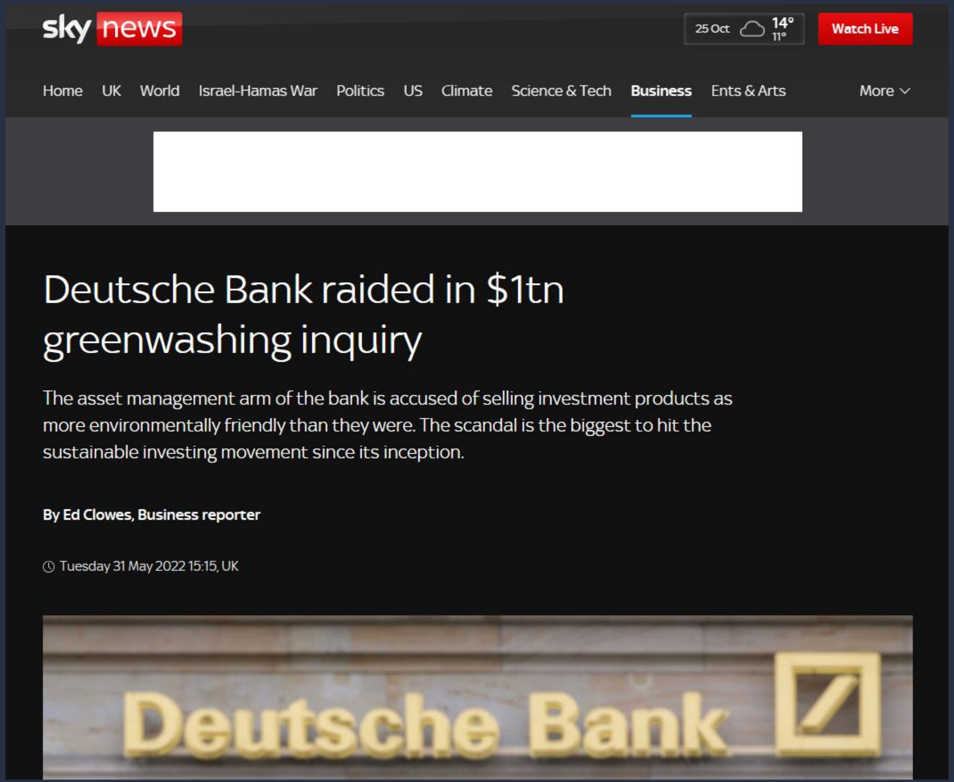
This sample demonstrates a few approaches for creating ChatGPT-like experiences over your own data using the Retrieval Augmented Generation pattern. It uses Azure OpenAI Service to access the ChatGPT model (gpt-35-turbo), and Azure Cognitive Search for data indexing and retrieval.

The repo includes sample data so it's ready to try end to end. In this sample application we use a fictitious company called Contoso Electronics, and the experience allows its employees to ask questions about the benefits, internal policies, as well as job descriptions and roles.



[Azure-Samples/azure-search-openai-demo: A sample app for the Retrieval-Augmented Generation pattern running in Azure, using Azure Cognitive Search for retrieval and Azure OpenAI large language models to power ChatGPT-style and Q&A experiences. (github.com)]

ELASTACLOUD



sky news

25 Oct 14° 11° Watch Live

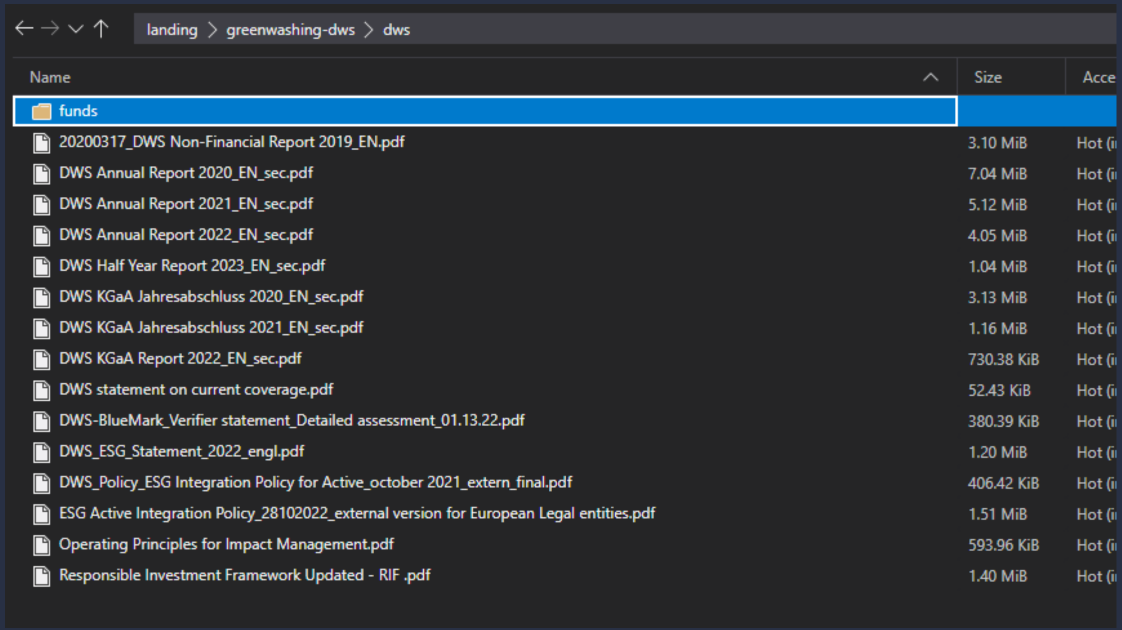
Home UK World Israel-Hamas War Politics US Climate Science & Tech **Business** Ents & Arts More

Deutsche Bank raided in $1tn greenwashing inquiry

The asset management arm of the bank is accused of selling investment products as more environmentally friendly than they were. The scandal is the biggest to hit the sustainable investing movement since its inception.

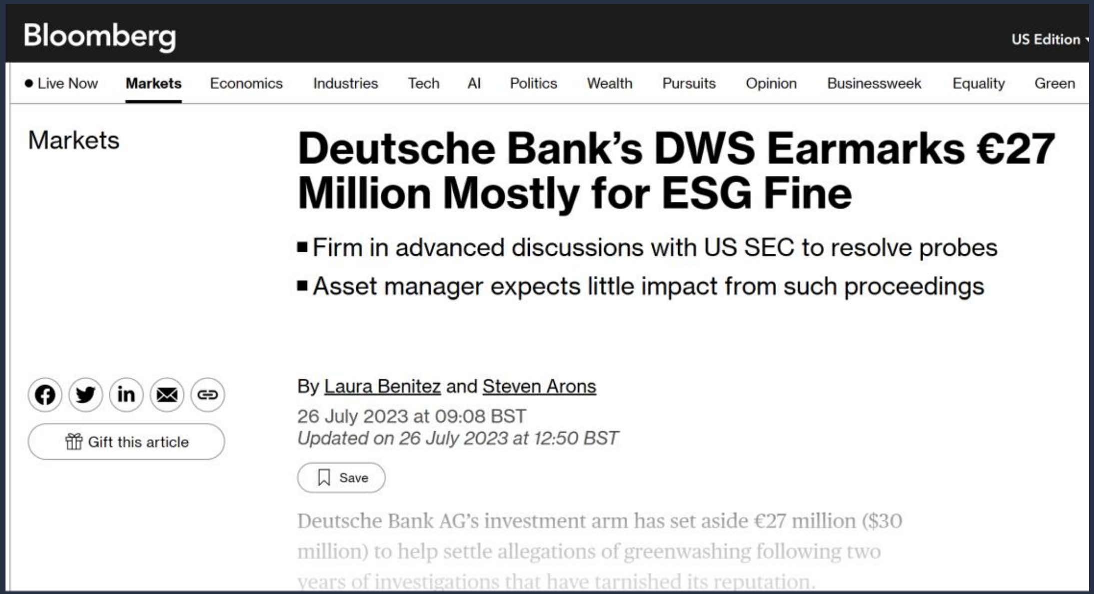
By Ed Clowes, Business reporter

Tuesday 31 May 2022 15:15, UK



landing > greenwashing-dws > dws

Name	Size	Acce
funds		
20200317_DWS Non-Financial Report 2019_EN.pdf	3.10 MiB	Hot (i
DWS Annual Report 2020_EN_sec.pdf	7.04 MiB	Hot (i
DWS Annual Report 2021_EN_sec.pdf	5.12 MiB	Hot (i
DWS Annual Report 2022_EN_sec.pdf	4.05 MiB	Hot (i
DWS Half Year Report 2023_EN_sec.pdf	1.04 MiB	Hot (i
DWS KGaA Jahresabschluss 2020_EN_sec.pdf	3.13 MiB	Hot (i
DWS KGaA Jahresabschluss 2021_EN_sec.pdf	1.16 MiB	Hot (i
DWS KGaA Report 2022_EN_sec.pdf	730.38 KiB	Hot (i
DWS statement on current coverage.pdf	52.43 KiB	Hot (i
DWS-BlueMark_Verifier statement_Detailed assessment_01.13.22.pdf	380.39 KiB	Hot (i
DWS_ESG_Statement_2022_engl.pdf	1.20 MiB	Hot (i
DWS_Policy_ESG Integration Policy for Active_october 2021_extern_final.pdf	406.42 KiB	Hot (i
ESG Active Integration Policy_28102022_external version for European Legal entities.pdf	1.51 MiB	Hot (i
Operating Principles for Impact Management.pdf	593.96 KiB	Hot (i
Responsible Investment Framework Updated - RIF .pdf	1.40 MiB	Hot (i



Bloomberg

US Edition

• Live Now **Markets** Economics Industries Tech AI Politics Wealth Pursuits Opinion Businessweek Equality Green

Markets

Deutsche Bank's DWS Earmarks €27 Million Mostly for ESG Fine

- Firm in advanced discussions with US SEC to resolve probes
- Asset manager expects little impact from such proceedings

By Laura Benitez and Steven Arons

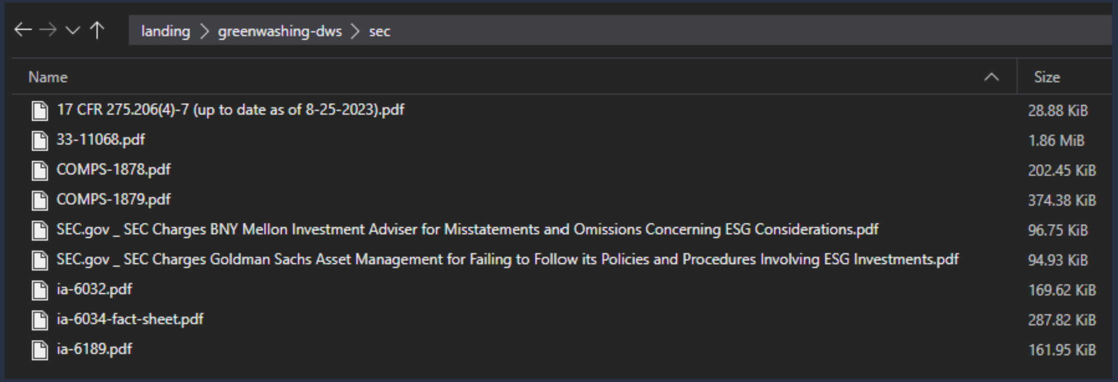
26 July 2023 at 09:08 BST

Updated on 26 July 2023 at 12:50 BST

Gift this article

Save

Deutsche Bank AG's investment arm has set aside €27 million ($30 million) to help settle allegations of greenwashing following two years of investigations that have tarnished its reputation.



landing > greenwashing-dws > sec

Name	Size
17 CFR 275.206(4)-7 (up to date as of 8-25-2023).pdf	28.88 KiB
33-11068.pdf	1.86 MiB
COMPS-1878.pdf	202.45 KiB
COMPS-1879.pdf	374.38 KiB
SEC.gov _ SEC Charges BNY Mellon Investment Adviser for Misstatements and Omissions Concerning ESG Considerations.pdf	96.75 KiB
SEC.gov _ SEC Charges Goldman Sachs Asset Management for Failing to Follow its Policies and Procedures Involving ESG Investments.pdf	94.93 KiB
ia-6032.pdf	169.62 KiB
ia-6034-fact-sheet.pdf	287.82 KiB
ia-6189.pdf	161.95 KiB

ELASTACLOUD

Work GitHub - Azure-Samples/azure-search-openai-demo: A sample app for the Retrieval-Augmented Generation pattern running in Azure, using Azure Cognitive Search for retrieval and Azure OpenAI large language models to power ChatGPT-style a...

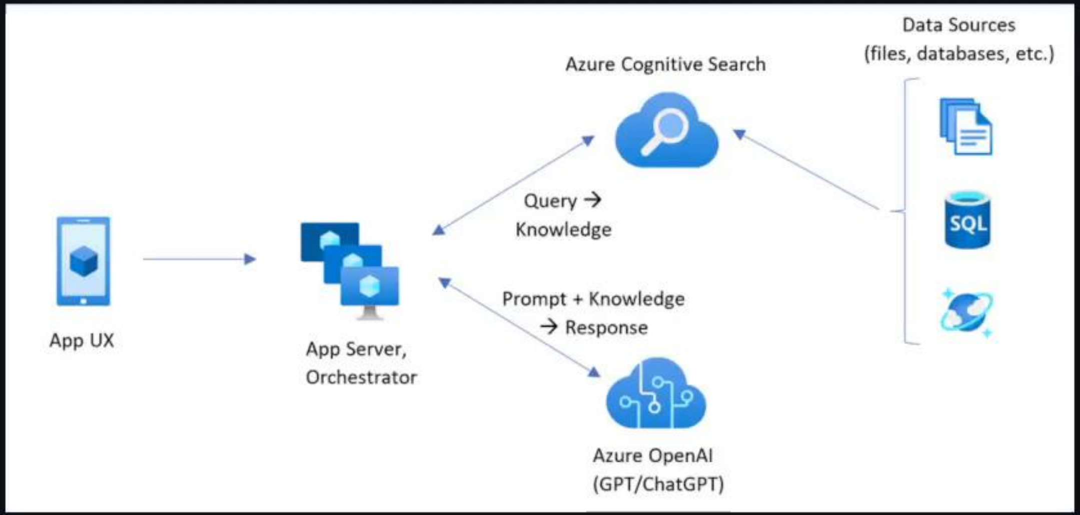
https://github.com/Azure-Samples/azure-search-openai-demo#local-environment

README.md

GITHUB CODESPACES OPEN DEV CONTAINERS OPEN

This sample demonstrates a few approaches for creating ChatGPT-like experiences over your own data using the Retrieval Augmented Generation pattern. It uses Azure OpenAI Service to access the ChatGPT model (gpt-35-turbo), and Azure Cognitive Search for data indexing and retrieval.

The repo includes sample data so it's ready to try end to end. In this sample application we use a fictitious company called Contoso Electronics, and the experience allows its employees to ask questions about the benefits, internal policies, as well as job descriptions and roles.



Features

Resources

Interesting Articles and videos

- Nvidia unveils monstrous A100 AI chip with 54 billion transistors and 5 petaflops of performance | VentureBeat
- To Fine Tune or not Fine Tune? That is the question - YouTube

Papers

- [1706.03762] Attention Is All You Need (arxiv.org)
- [2303.17564] BloombergGPT: A Large Language Model for Finance (arxiv.org)
- [2203.15556] Training Compute-Optimal Large Language Models (arxiv.org)
- SuperGLUE Benchmark
- [2005.11401] Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks (arxiv.org)

GitHub

- Azure-Samples/azure-search-openai-demo: A sample app for the Retrieval-Augmented Generation pattern running in Azure, using Azure Cognitive Search for retrieval and Azure OpenAI large language models to power ChatGPT-style and Q&A experiences. (github.com)
- sapientml/sapientml: Generative AutoML for Tabular Data (github.com)
- zilliztech/GPTCache: Semantic cache for LLMs. Fully integrated with LangChain and llama_index. (github.com)
- Mooler0410/LLMsPracticalGuide: A curated list of practical guide resources of LLMs (LLMs Tree, Examples, Papers) (github.com)
- Hannibal046/Awesome-LLM: Awesome-LLM: a curated list of Large Language Model (github.com)

Training

- Generative AI with Large Language Models - Week 1 - Week 1 | Coursera
- Develop Generative AI solutions with Azure OpenAI Service - Training | Microsoft Learn

ELASTACLOUD

Thank You!

Twitter (X): @Darsh262

LinkedIn:

Darshna Shah

Chief AI Officer at Elastacloud|
Microsoft AI MVP | Organiser of the ...

www.datasciencewithdarsh.com

ELASTACLOUD