

# Post Read : Gen-Ai Training

## Contents

1.	Introduction to AI .....	3
1.1	What is AI?.....	3
1.2	AI Project Lifecycle .....	5
2.	Exploratory Data Analytics (EDA) .....	8
2.1	EDA Example .....	9
3.	Machine Learning.....	9
3.1	Machine Learning Vs Rule Based Programming .....	9
3.2	Data Science Problem Type with examples .....	10
3.3	Common training methods and algorithms .....	11
3.3.1	Linear Regression .....	12
3.3.2	Logistic Regression .....	13
3.3.3	Decision Tree.....	14
	Random Forest Model .....	15
	Machine Learning Map .....	15
	Model Evaluation.....	17
	Confusion Matrix, F1 score, Precision, Recall.....	17
	Machine Learning With Azure ML Studio And Azure Cloud .....	18
	Data Science and Machine Learning References.....	23
	Learning Python .....	23
	Deep Learning .....	24
	Neural Networks.....	24
	Key Components Of A Neural Network .....	25
	Artificial Neural Network (ANN) .....	26
	Steps In Neural Network Processing.....	26
	Backpropagation .....	27
	Key use cases.....	29
	Computer vision: Convolutional Neural Network (CNN) .....	29
	Generative Adversarial Network (GAN) .....	30

Variable Auto Encoder (VAE) .....	31
Natural Language Processing (NLP) .....	32
Tokenization .....	32
Vectors.....	33
Embeddings .....	33
RNN, LSTM and GRUs.....	34
Challenges With NLP .....	35
Deep Learning References .....	36
The Transformer Model.....	37
GPTs and LLMs .....	39
Foundation Models.....	42
Multi Modal .....	42
Gen AI Core capabilities .....	43
Inferencing, Question & Answer based on pre-trained knowledge .....	43
Summarization.....	44
Reasoning.....	44
ReAct – Reasoning and Action .....	44
Customizing LLM/GPT behaviours .....	45
The need to customize .....	45
Customization options.....	46
Prompt Engineering.....	48
Understanding Tokens .....	50
Prompt Engineering Tools .....	51
Prompt Engineering References .....	52
Retrieval Augmented Generation (RAG) .....	53
LLM Fine Tuning .....	58
Low Rank Adaptation (LoRA).....	58
Quantized Low Rank Adaptation (QLoRA) .....	59
Alignment Fine Tuning (RLHF etc.).....	59
Agents.....	60
Evaluating LLMs .....	63
The RAG Triad .....	64

LLM as a Judge .....	65
Role of Human Reviewers .....	66
Platforms And Tools for Gen AI App Development.....	67
Implementation Challenges.....	71
Costing .....	73
Enterprise Considerations .....	76
Productionalizing LLM Applications .....	76
Securing LLM Applications .....	80
Guardrails .....	80
Risks, Governance And Compliance .....	80
Explainable AI .....	81
Responsible AI Framework .....	83
Gen AI Use cases Across Job Functions .....	84
Marketing and Sales.....	86
Software Development and Quality Assurance .....	87
Human Resources .....	88
Accounting and Finance.....	89
Gen AI Use cases Across Industries.....	89
Finance and BFSI.....	90
Manufacturing.....	92
Media and Entertainment.....	95
Popular Gen-AI tools and applications .....	96
Articles, Blogs, Newsletters, SM Handles.....	100
What does the future hold?.....	105

## 1. Introduction to AI

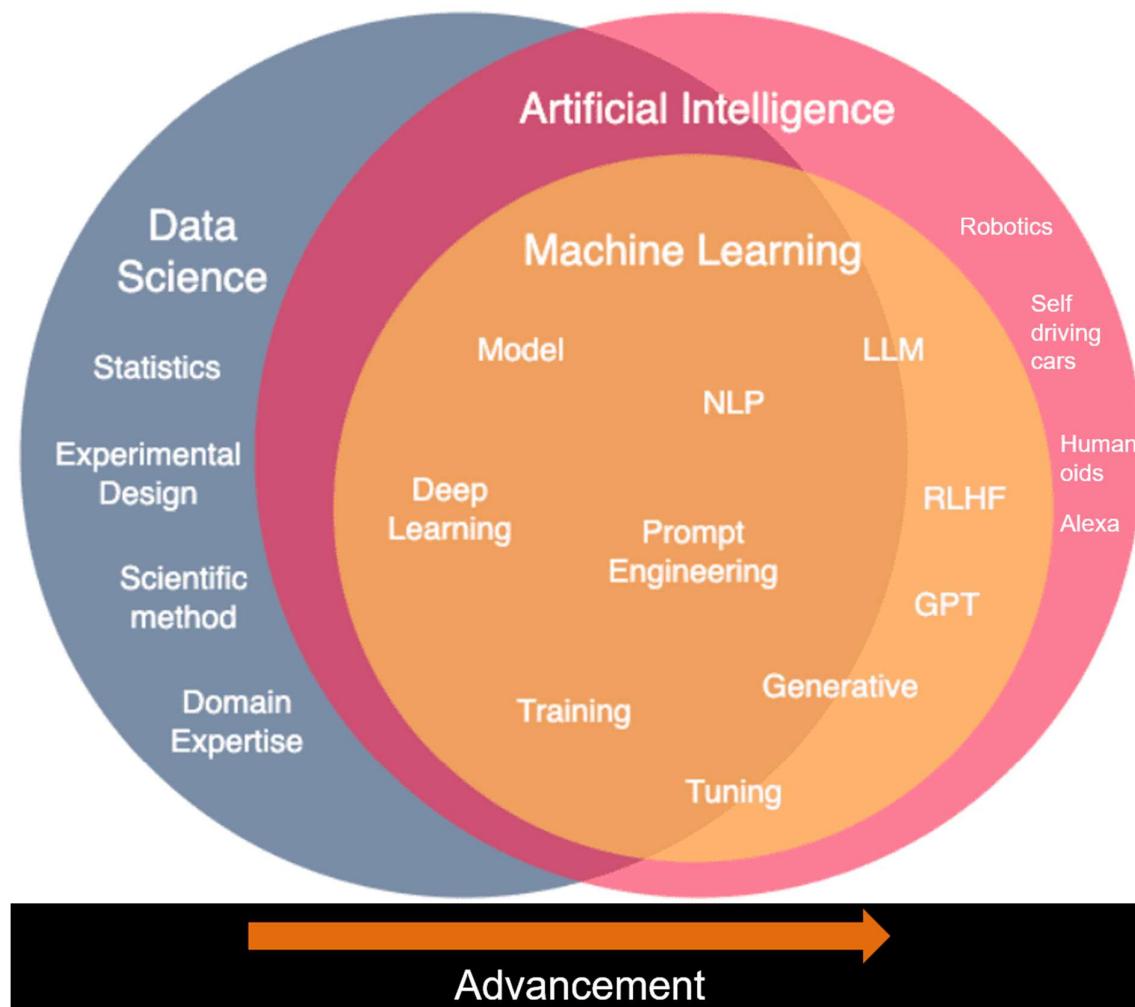
### 1.1 What is AI?

Artificial Intelligence (AI) is the field of computer science focused on creating machines or systems that can perform tasks that would typically require human intelligence.

These tasks often include:

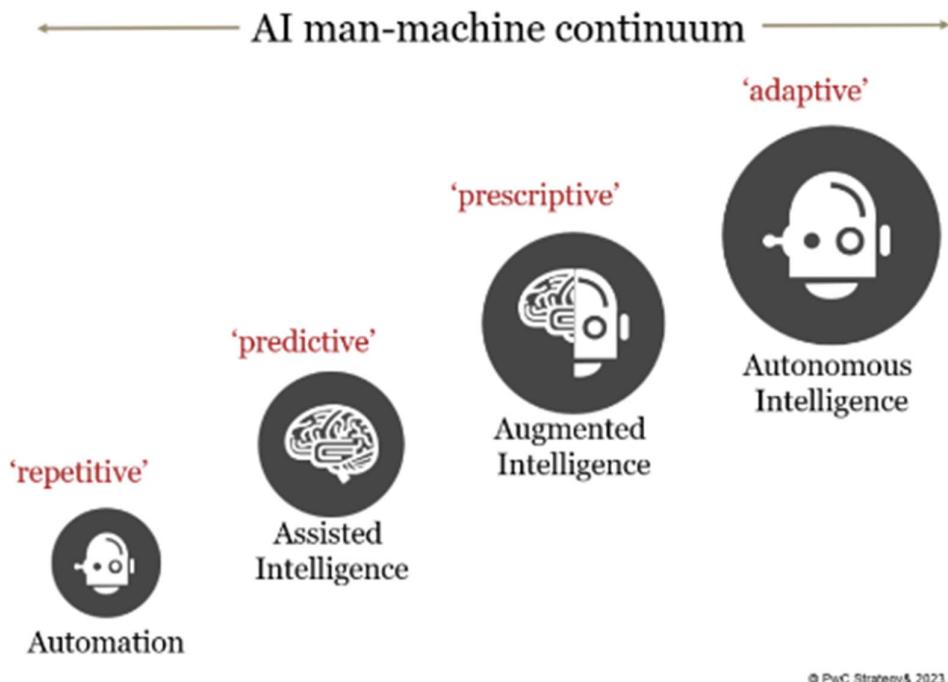
- Reasoning
- Learning
- Problem-solving
- Understanding natural language
- Recognizing patterns
- Making decisions
- Generating new stuff based on it's learning

AI systems can adapt and improve over time, especially those utilizing advanced techniques like machine learning (ML) and deep learning.



**The AI Story:** From Turing to GPT-4 Every great story has a beginning, and AI's starts with the Turing Test in 1950, a brainchild of Alan Turing, to test a machine's ability to imitate human intelligence. Fast forward to 1956, and we have the term "artificial intelligence" coined by John McCarthy. From there, the AI journey takes us to milestones like Shakey (1969), the first-ever general-purpose mobile robot, and Deep Blue's triumph over the world chess champion (1997), which illustrated AI's potential.

As we venture into the 21st century, AI continues to push boundaries. From robotic process automation (RPA) and smart homes to self-driving cars, and more recently, OpenAI's GPT-4 (2022), showcasing high-quality text generation—a leap in natural language processing capabilities.



## 1.2 Roadmap

A good roadmap at:

[h3esam on X: "if you are completely new to LLMs and NLP, here is a solid roadmap to build your knowledge from the foundations: \(+ link to resources\)](#) [1 Foundations ↴ Mathematics for Machine Learning \[Book\] ↴ Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow https://t.co/trhbboVoey" / X](#)

**Foundations ↴ Mathematics for Machine Learning [Book] ↴ Hands-On Machine**

Learning with Scikit-Learn, Keras, and TensorFlow [Book] ↴ (Optional) Introduction to Statistical Learning [Book] — for statistics understanding

**Basic NLP Concepts** ↴ Stopwords, Stemming, Lemmatization ↴ Bag of Words ↴ TF-IDF ↴ BM25 ↴ Word2Vec ↴ N-grams ↴ Named Entity Recognition (NER) ↴ Tokenization ↴ Part-of-Speech (POS) Tagging

**Get Familiar with Frameworks** ↴ NLTK ↴ SpaCy ↴ PyTorch (important, more popular now) ↴ (Optional) Hugging Face basics

**Understand Embeddings** ↴ Word2Vec ↴ GloVe ↴ FastText (Understand pre-transformer embeddings properly)

**Dive into Transformers & Models** ↴ Understand how Transformers work ↴ Explore BERT, ModernBERT, DistilBERT, etc. ↴ Learn Sentence-Transformers & Hugging Face

**Dive into Evaluations** ↴ BLEU Score, ROUGE Score, Perplexity, etc. ↴ Accuracy, Precision, Recall, F1 Score, etc. ↴ Semantic Similarity evaluations (Cosine similarity, embedding-based metrics)

**Explore Fine-tuning of Small Language Models (e.g., BERT)** ↴ Fine-tune for classification ↴ Fine-tune for question-answering ↴ Fine-tune for domain-specific embeddings ↴ Understand task-specific heads (classification/QA) ↴ Learn fine-tuning best practices: LR schedules, early stopping, regularization credits to the amazing Shantanu Ladhwe for his amazing solid roadmaps

## NLP Roadmap (pre-LLM)

### ① Foundations

- ↳ Mathematics for Machine Learning [Book]
- ↳ Hands-on Machine Learning with Scikit-Learn, Keras and TensorFlow [Book]

### ② Basic NLP concepts

- ↳ Bag of words
- ↳ TD-IDF
- ↳ BM25
- ↳ word2vec
- ↳ n-grams
- ↳ Named Entity recognition
- ↳ Tokenization
- ↳ Stop words, Stemming, etc.

### ③ Get familiar with frameworks

- ↳ NLTK
- ↳ Spacy
- ↳ PyTorch → Important {more popular}

### ④ Understand Embeddings

- ↳ word2vec, GloVe, FastText (pre-transformer era)

### ⑤ Dive into Transformers & models

- ↳ Understand Transformer
- ↳ BERT, ModernBERT, DistilBERT, ...
- ↳ Explore sentence-transformer & HuggingFace
- { skipping RNN, seq2seq }

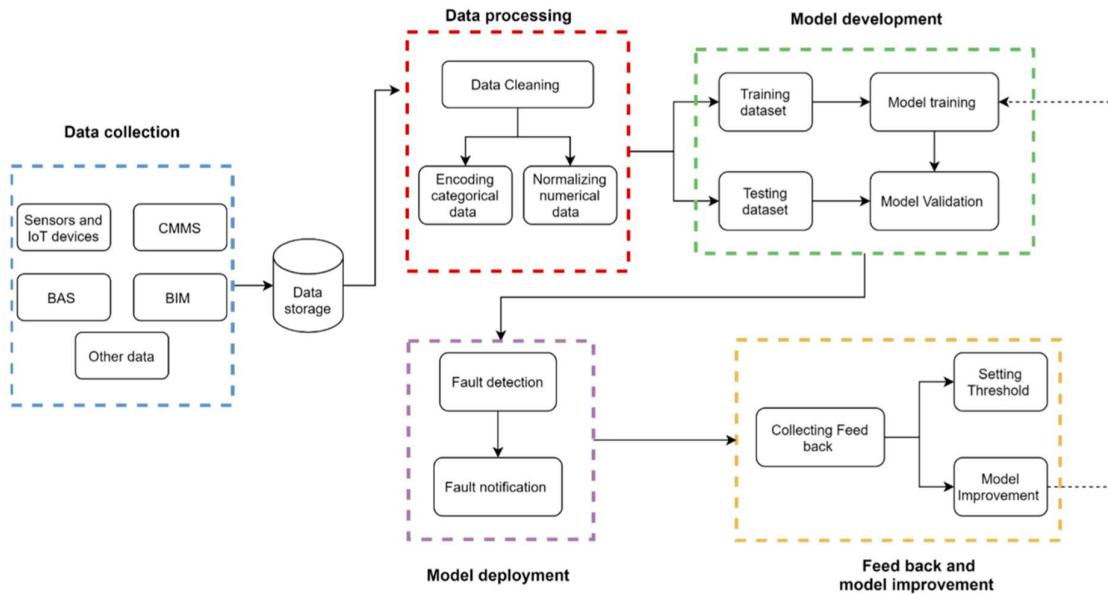
### ⑥ Dive into Eval: →

- ↳ BLEUScore, ROUGE score, Perplexity... (more)

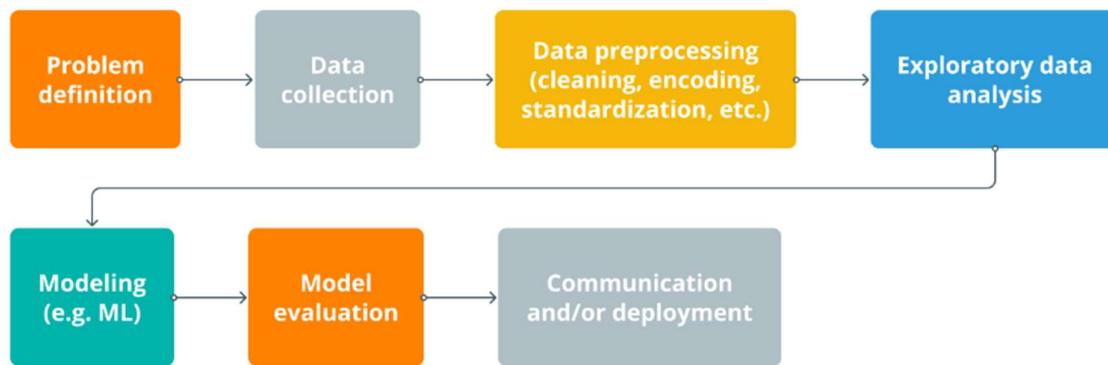
### ⑦ Explore fine-tuning of small language models eg. BERT

- ↳ For classification
- ↳ For question-answering
- ↳ Generate fine-tuned embeddings

## AI Project Lifecycle



## 2. Exploratory Data Analytics (EDA)



## Data Cleaning

Handling missing values, outliers, duplicates, and inconsistencies.

## Descriptive Statistics

Summarizing data using mean, median, mode, standard deviation, etc.

## Data Visualization:

Using plots like histograms, scatter plots, box plots, and correlation matrices to visualize relationships between variables.

## Feature Engineering

Creating new features or transforming existing ones to make data more meaningful.

- Dimensionality Reduction
- Encoding (convert categorical features into numerical, for e.g. One Hot Encoding)
- Normalization (Min-max scaling, Log transformation)
- Binning
- Adding new columns

## 2.1 EDA Example

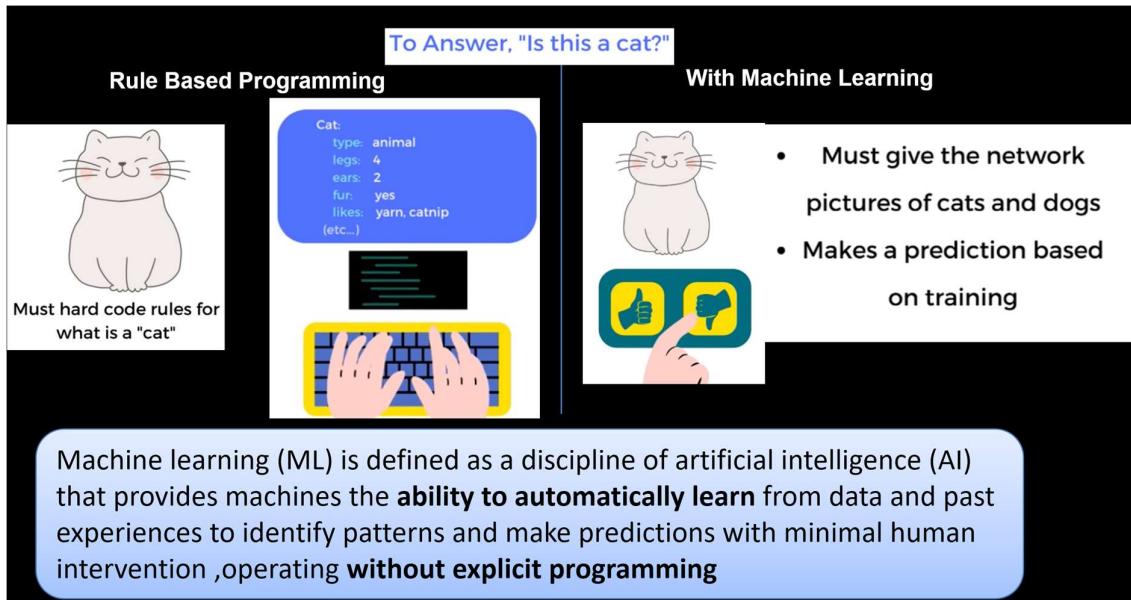
### **Loan Defaulters Prediction | EDA | Lending Club Study**

Python · [Loan Classification Dataset](#)

<https://www.kaggle.com/code/abhishek14398/loan-defaulters-prediction-eda-lending-club-study>

## 3. Machine Learning

### 3.1 Machine Learning Vs Rule Based Programming

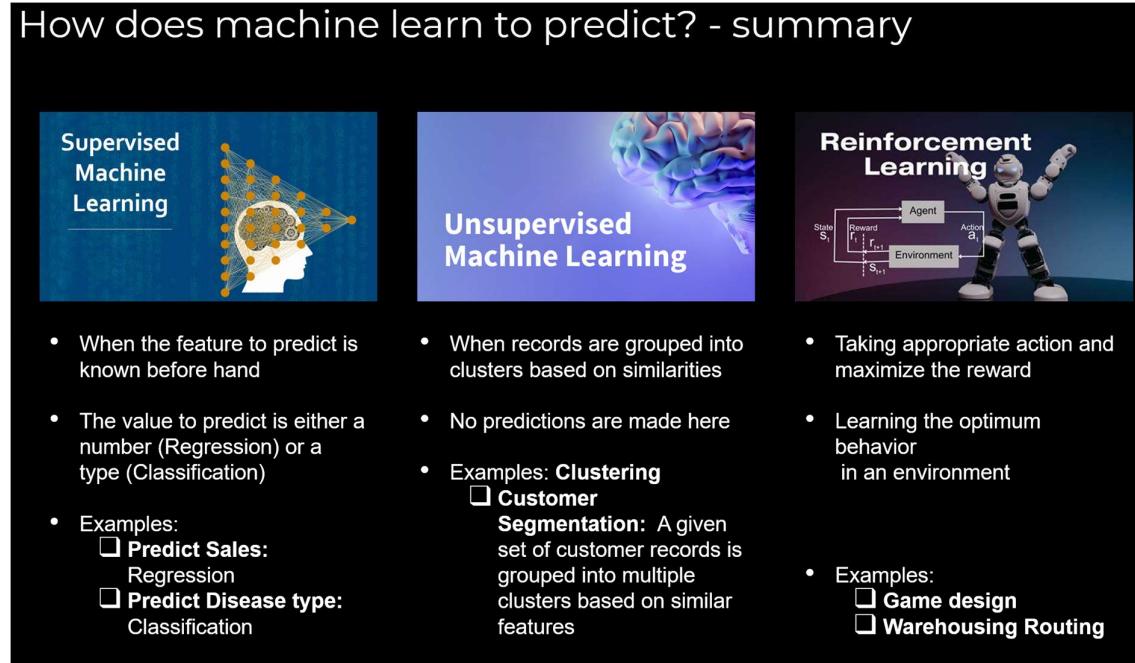


### 3.2 Data Science Problem Type with examples

Problem	Description	Examples
Classification	Predicting whether a data point belongs to one of the predefined classes	Customer Churn Prediction, Spam Detection
Regression	Predicting the numerical value of the target variable	Agricultural Yield, Inflation Rate Prediction
Clustering	Identifying natural groups within the dataset based on similar inherent property of the data points	Social Network Analysis, Crime Incidence Analysis, Search Result Grouping
Anomaly Detection	Predicting whether a data point is an outlier in comparison with other points in the data set	Detecting Network Intrusions, Predicting Machine Failures, Detecting Gene Mutations
Association	Discovering rules that govern frequent simultaneous occurrence of certain items or phenomena	Medical Diagnosis, Protein Sequencing, Building Intelligent Transportation Systems
Recommendation	Suggesting items for users based on the past preferences of theirs and similar users	Recommendation of Movies, Books, Restaurants, Holiday Destinations

### 3.3 Common training methods and algorithms

How does machine learn to predict? - summary



Problem Type	Training Method	Common Algorithms
Classification	Supervised	Logistic Regression, Decision Trees, Random Forest, Support Vector Machines (SVM), K-Nearest Neighbors (KNN), Neural Networks
Regression	Supervised	Linear Regression, Ridge/Lasso Regression, Polynomial Regression, Random Forest, Gradient Boosting Machines (GBMs), Neural Networks
Clustering	Unsupervised	K-Means, Hierarchical Clustering, DBSCAN, Gaussian Mixture Models (GMMs), Spectral Clustering
Anomaly Detection	Semi-supervised, Unsupervised, or Supervised	Isolation Forest, One-Class SVM, Autoencoders, DBSCAN, Gaussian Mixture Models, LOF (Local Outlier Factor)
Association	Unsupervised	Apriori, Eclat, FP-Growth (Frequent Pattern Growth)
Recommendation Engine	Supervised, Unsupervised, or Reinforcement Learning	Collaborative Filtering (Matrix Factorization, Singular Value Decomposition), Content-Based Filtering, Neural Collaborative Filtering, Reinforcement Learning (Bandit algorithms)

Examples:

Bad loan write-offs	→	Classification
Property valuation	→	Regression
Differentiated marketing campaign	→	Clustering
Credit card fraud detection	→	Anomaly detection
Improving cross-sales	→	Association
Personalized online experience	→	Recommendation
Adjust the portfolio optimization model based on market conditions	→	Reinforcement Learning

### 3.3.1 Linear Regression

#### How it works:

Linear regression models the relationship between a dependent variable  $y$  and one or more independent variables  $X$  by fitting a linear equation:

$$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n + \varepsilon$$

where  $\beta_0$  is the intercept,  $\beta_1 \dots \beta_n$  are coefficients, and  $\varepsilon$  is the error term.

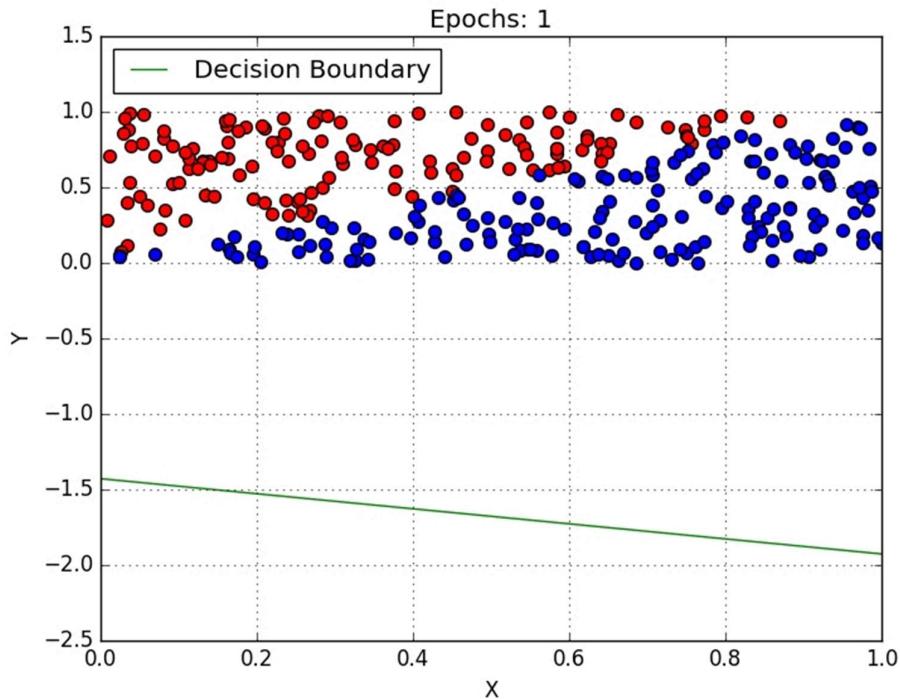
The model estimates these coefficients to minimize the sum of squared residuals (differences between predicted and actual values).

#### Evaluation Metrics:

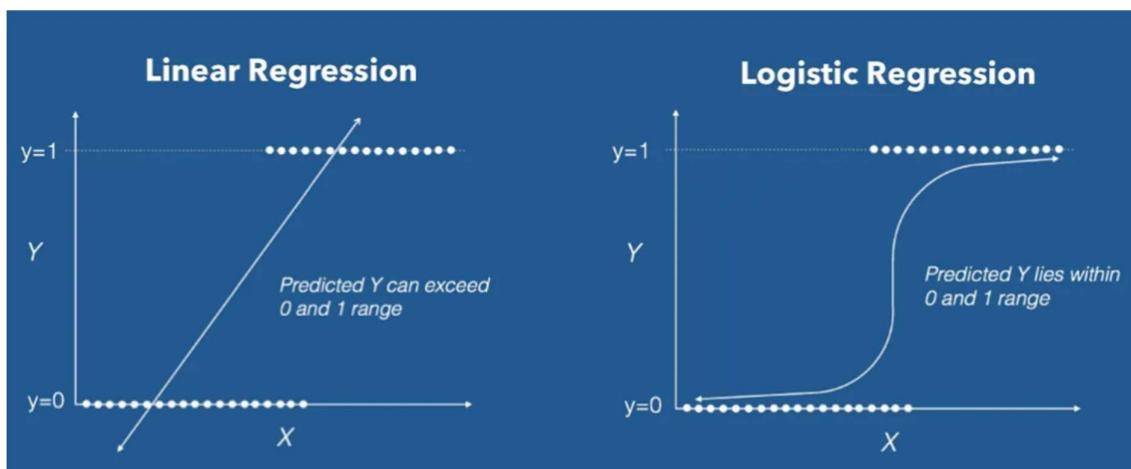
- Mean Squared Error (MSE)
- Root Mean Squared Error (RMSE)
- R-squared (coefficient of determination)
- Adjusted R-squared

### 3.3.2 Logistic Regression

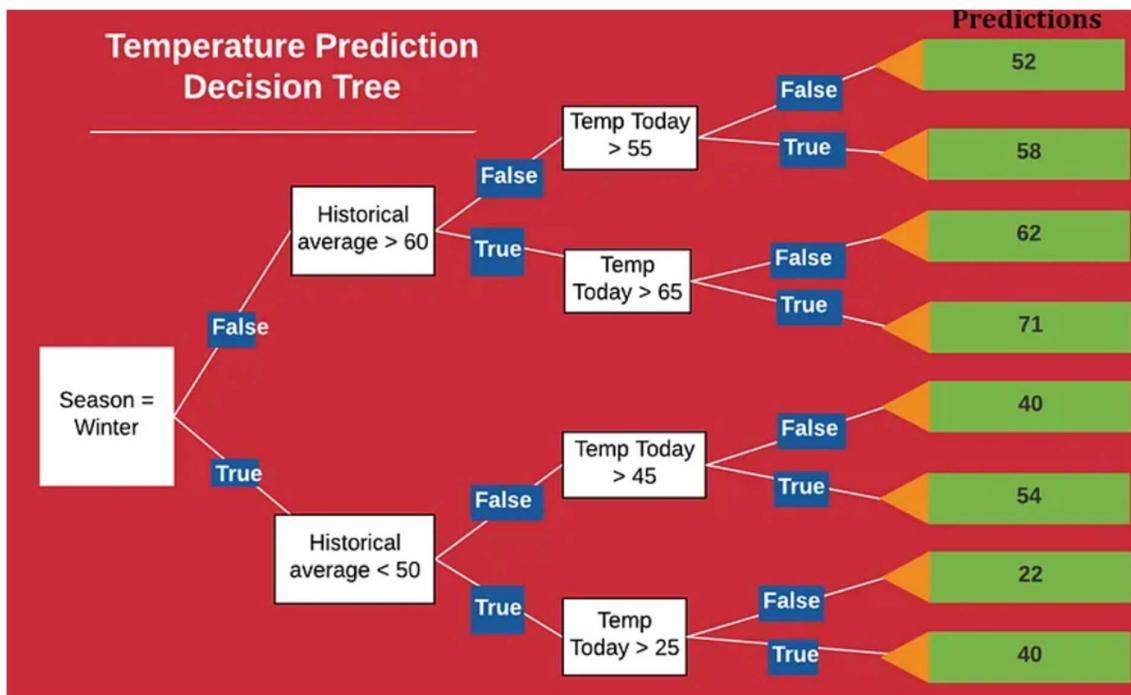
Used for classification tasks:



Logistic regression uses Sigmoid function to suit the classification task:



### 3.3.3 Decision Tree



#### Advantages

- No need of feature scaling
- Handles both categorical and numeric data
- No assumptions about distribution
- Explainability, easy to understand

#### Limitations

- Overfitting
- Biased when data is imbalanced
- Unstable (High variance)
- Inefficient with high dimensional data

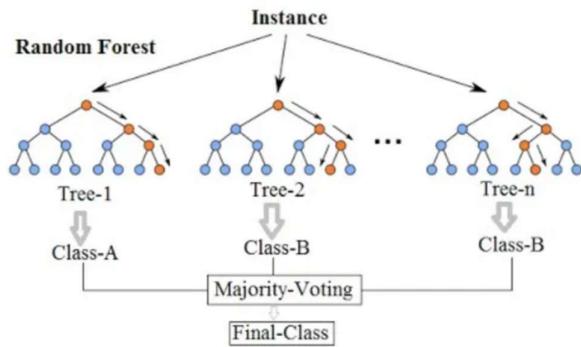
Decision trees have high variance

Models such as Random forest Gradient Boosting helps minimize the variance

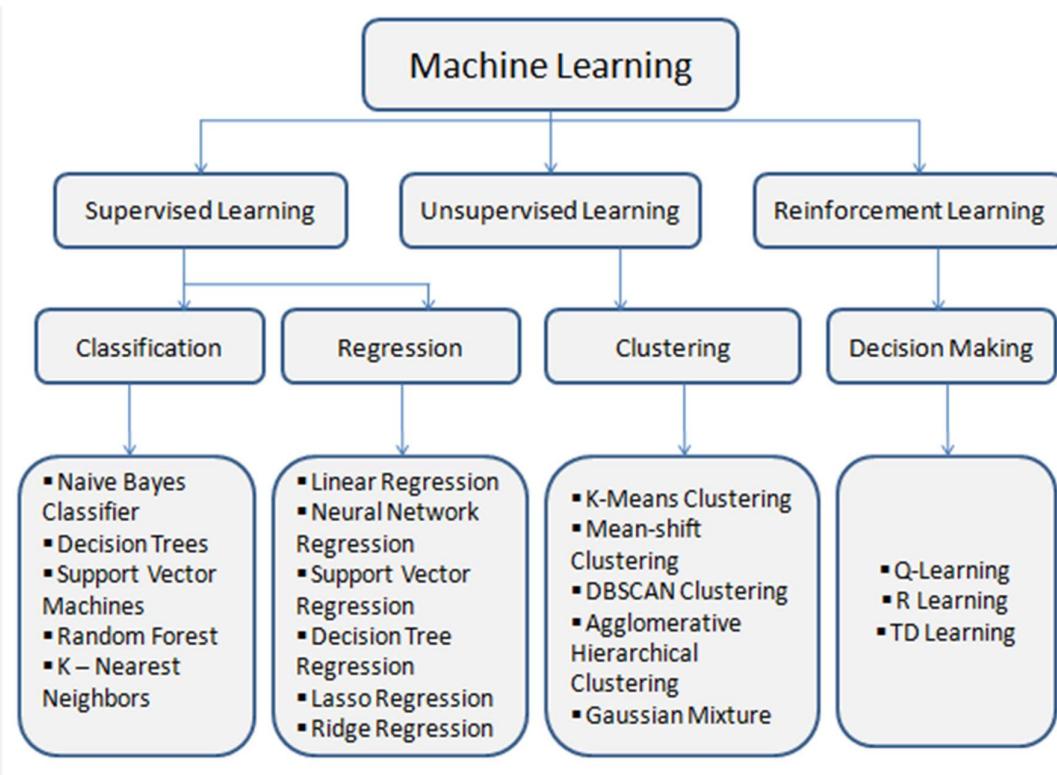
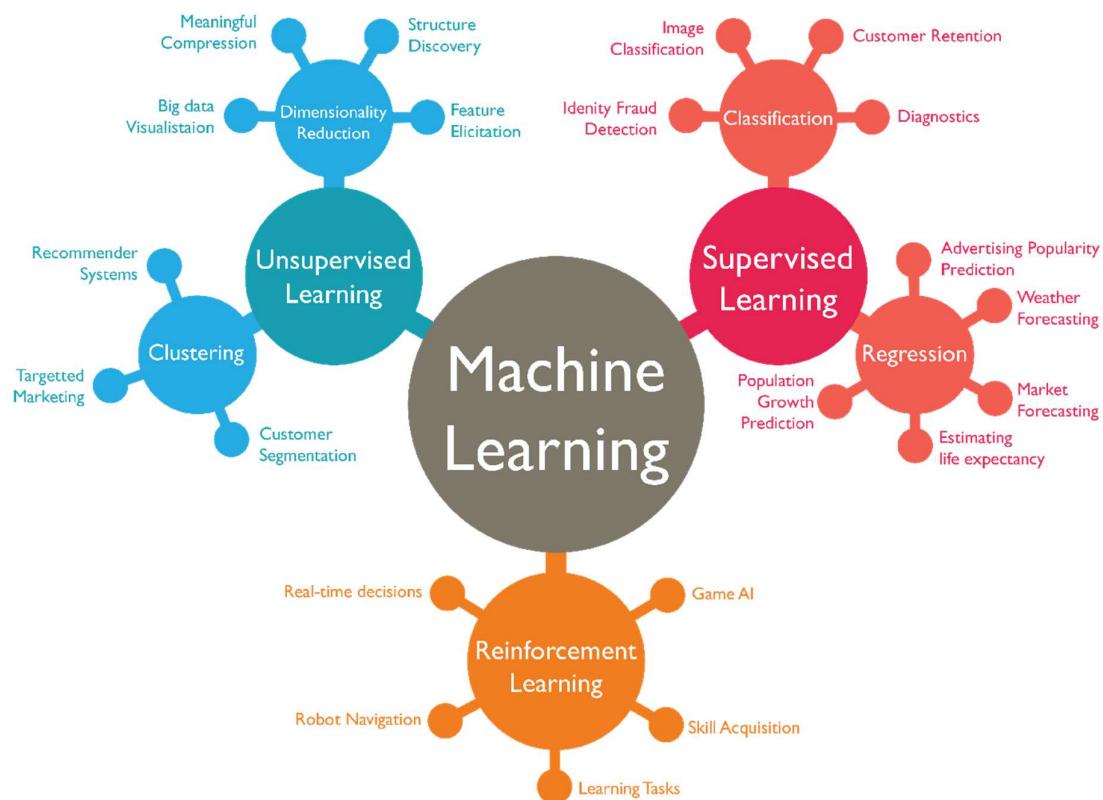
## Random Forest Model

Uses **Bagging** technique to reduce variance.

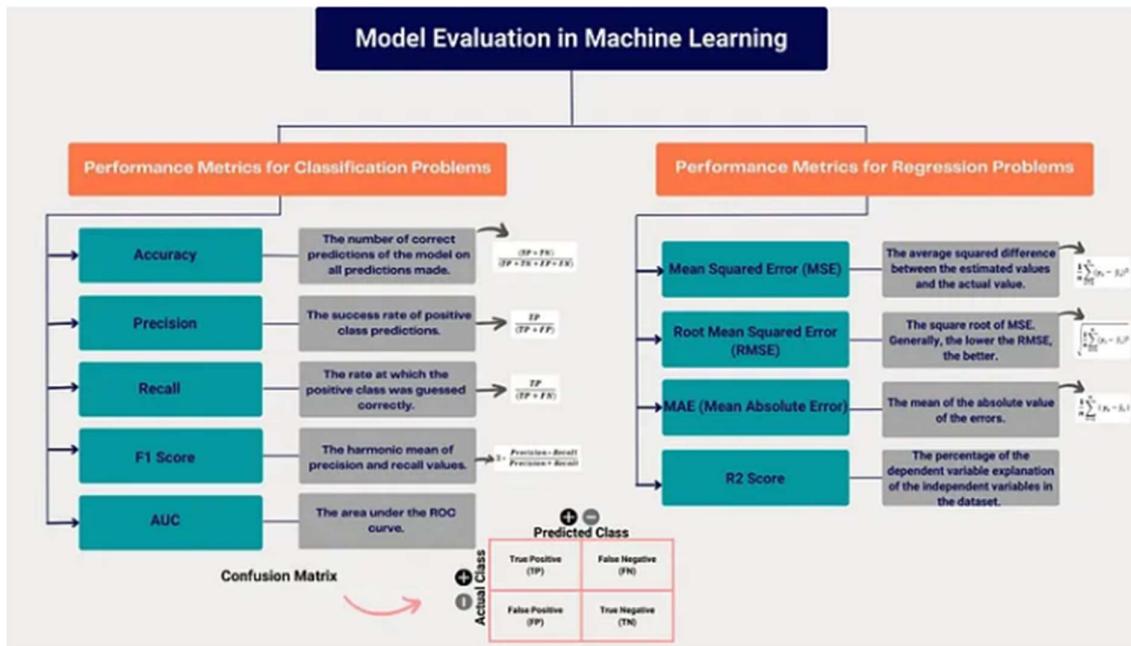
**Bagging** is an ensemble learning technique that builds multiple independent models (in the case of Random Forest, these are decision trees) using different random subsets of the training data.



## Machine Learning Map



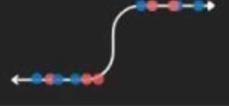
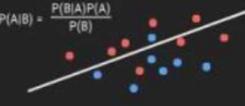
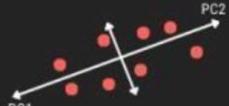
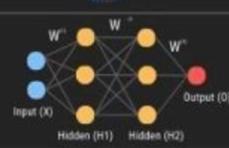
## Model Evaluation



Confusion Matrix, F1 score, Precision, Recall

		Predicted Value	
		No	Yes
Actual Value	No	970 (True Negative)	0 (False Positive)
	Yes	29 (False Negative)	1 (True Positive)

Hyperparameter tuning map

ML Algorithms	Hyperparameters
Linear Regression 	<ul style="list-style-type: none"> <li>L1/L2 Penalty</li> <li>Solver</li> <li>Fit Intercept</li> </ul>
Logistic Regression 	<ul style="list-style-type: none"> <li>L1/L2 Penalty</li> <li>Solver</li> <li>Class Weight</li> </ul>
Naive Bayes 	<ul style="list-style-type: none"> <li>Alpha</li> <li>Fit Prior</li> <li>Binarize</li> </ul>
Decision Tree 	<ul style="list-style-type: none"> <li>Criterion</li> <li>Max Depth</li> <li>Min Sample Split</li> </ul>
Random Forest 	<ul style="list-style-type: none"> <li>Criterion</li> <li>Max Depth</li> <li>N Estimators</li> <li>Max Features</li> </ul>
Gradient Boosted Trees 	<ul style="list-style-type: none"> <li>Criterion</li> <li>Max Depth</li> <li>N Estimators</li> <li>Min Sample Split</li> <li>Learning Rate</li> </ul>
Principal Component 	<ul style="list-style-type: none"> <li>N Component</li> <li>SVD Solver</li> <li>Iterated Power</li> </ul>
K-Nearest Neighbor 	<ul style="list-style-type: none"> <li>N Neighbors</li> <li>Weights</li> <li>Algorithm ('kd_tree', 'brute')</li> </ul>
K-Means 	<ul style="list-style-type: none"> <li>N Clusters</li> <li>Init</li> <li>Max Iter</li> </ul>
Dense Neural Networks 	<ul style="list-style-type: none"> <li>Hidden Layer Sizes</li> <li>Activation</li> <li>Dropout</li> <li>Solver</li> <li>Alpha</li> <li>Learning rate</li> </ul>

Ref: [datainterview.com](http://datainterview.com)

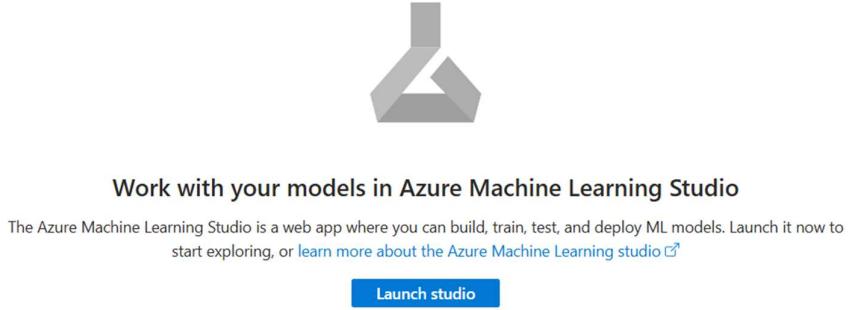
## Machine Learning with Azure ML Studio And Azure Cloud

Azure ML Workspace and ML Studio can be used to model a ML pipeline with no-code capabilities, including Auto-ML

## Steps involved

- Create ML Workspace
- Launch ML Studio from the workspace

└── azmlflow tracking URI : azureml://eastus2.api.azureml.ms/m



- In ML Studio:
  - o Import the dataset

This screenshot shows the "Data" section of the Azure Machine Learning Studio interface. The left sidebar has a "Data" category selected under "Assets". The main area displays a table of data assets:

Name	Source	Version	Created
loandatanew	This workspace	1	Sep
TD-Loan_Approval_ML_Pipeline-	This workspace	2	Sep
MD-Loan_Approval_ML_Pipeline	This workspace	2	Sep
loadata2	This workspace	1	Sep
sentimentdata	This workspace	1	Sep
automl_data1	This workspace	1	Sep
finance_sentiment	This workspace	1	Sep

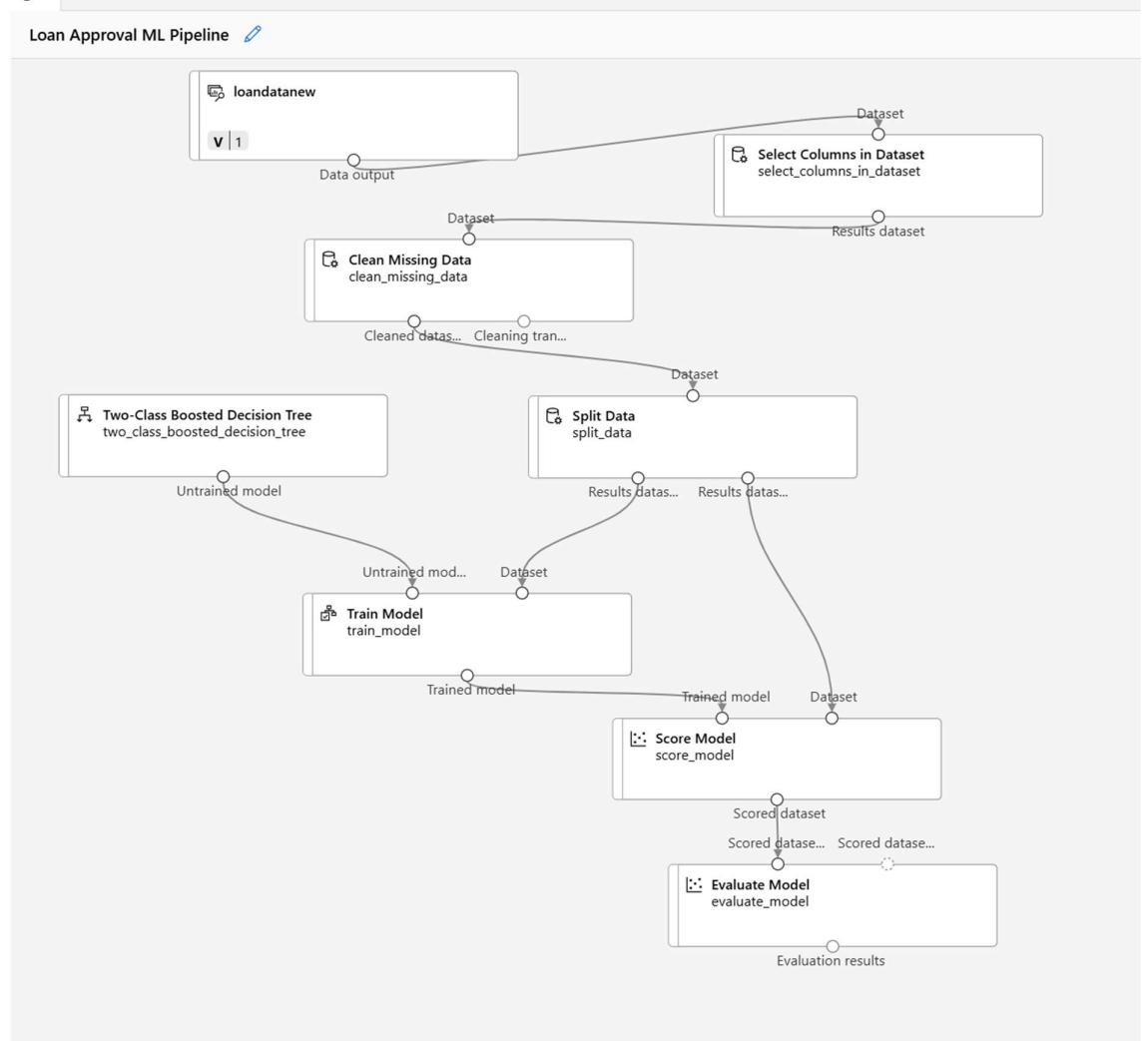
- Create a ML Pipeline

The screenshot shows the Azure AI | Machine Learning Studio Designer interface. The left sidebar navigation includes: All workspaces, Home, Model catalog, Authoring (Notebooks, Automated ML, Designer selected), Prompt flow, Assets (Data, Jobs, Components, Pipelines, Environments, Models, Endpoints), Manage (Compute, Monitoring). The main area is titled 'Designer' and 'New pipeline'. It shows two tabs: 'Classic prebuilt' (selected) and 'Custom'. A note states: 'This low-code option uses existing prebuilt components and earlier dataset types (tab will not have any new components added.)'. Below are three prebuilt pipeline options: 'Create a new pipeline using classic prebuilt components' (with a plus icon), 'Image Classification using DenseNet' (with a neural network icon), and 'Binary Classification using Vowpal Wabbit' (with a user icon). The 'Pipelines' section shows a table with one entry:

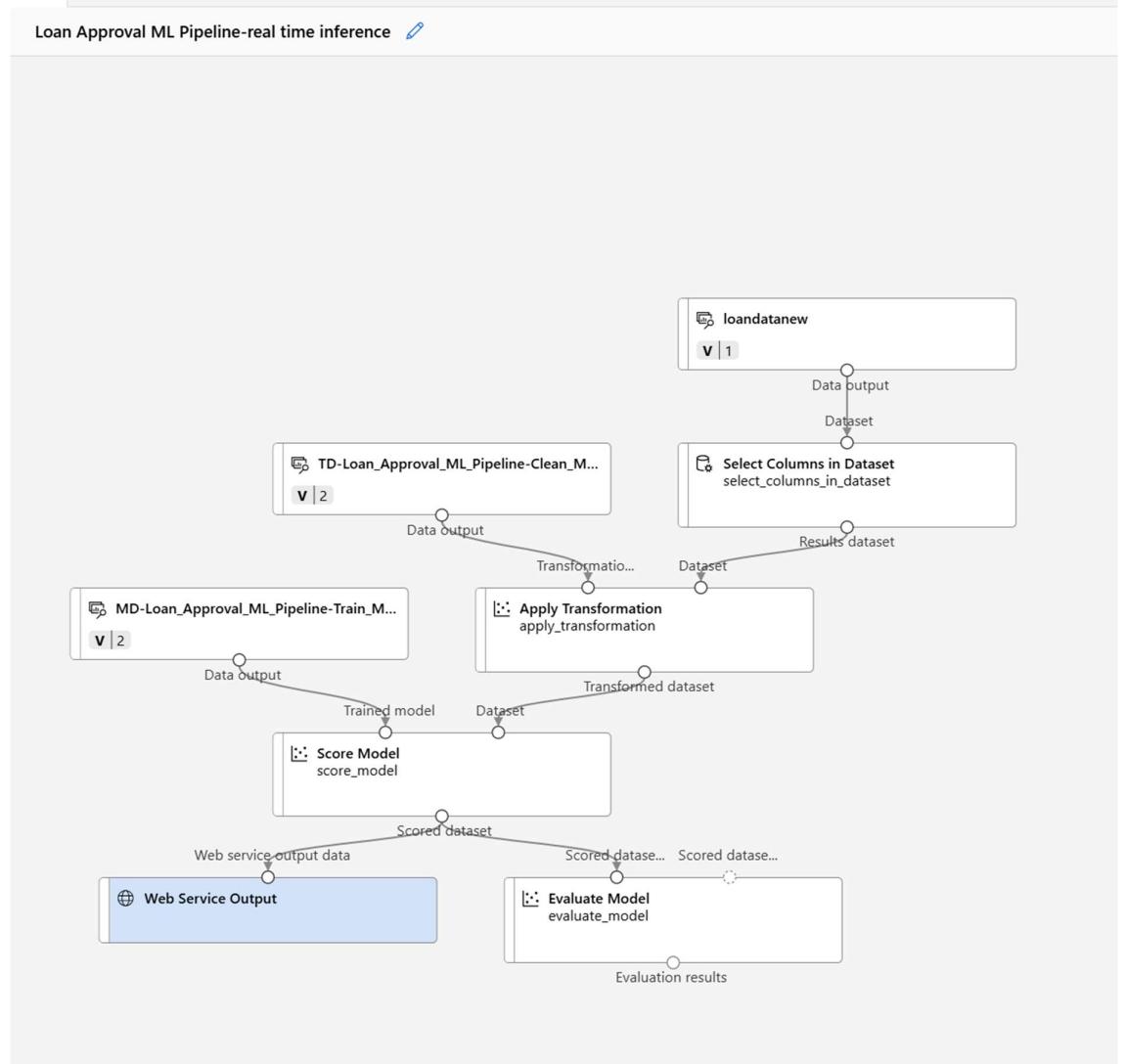
Name	Pipeline type	Last modified
Loan Approval ML Pipeline-real time i...	Real-time inference	Oct 12, 2023, 10:16 AM

Buttons at the top of the pipelines section include Refresh, Delete, and Reset view.

- Design the training pipeline



- Execute it
  - Output will be a trained model
- Create an inference pipeline that will expose the Model as an API web service



- Test the web service with postman or any other API tool

```

1  {
2      "Inputs" : {
3          "input1" : [
4              {
5                  "Age" : 24,
6                  "Experience" : 4,
7                  "Income" : 300000,
8                  "Zipcode" : 400053,
9                  "Family" : 1,
10                 "CCAvg" : 400,
11                 "Education" : 4,
12                 "Mortgage" : 0,
13                 "SecuritiesAccount" : 1,
14                 "CDAccount" : 1,
15                 "Online" : 1,
16                 "CreditCard" : 0
17             }
18         ]
19     }
20 }
21 }
```

## Data Science and Machine Learning References

### Learning Python

<https://github.com/jerry-git/learn-python3/tree/master/notebooks/beginner/notebooks>

<https://www.kaggle.com/code/chats351/introduction-to-numpy-pandas-and-matplotlib>

For deeper understanding of data science and machine learning, recommend following reading:

#### 1) Analytics Vidhya course (free)

<https://www.analyticsvidhya.com/blog/2017/09/common-machine-learning-algorithms/>

### Google Courses (free):

<https://developers.google.com/machine-learning/crash-course/linear-regression>

### Top 4 Linear Regression variations:

<https://towardsdatascience.com/top-machine-learning-algorithms-for-regression-c67258a2c0ac>

### Top 6 ML Algorithms for classification:

<https://towardsdatascience.com/top-machine-learning-algorithms-for-classification-2197870ff501>

ML A-Z

<https://drive.google.com/drive/folders/1OFNnrHRZPZ3unWdErjLHod8lbv2FfG1d>

TensorFlow Bootcamp

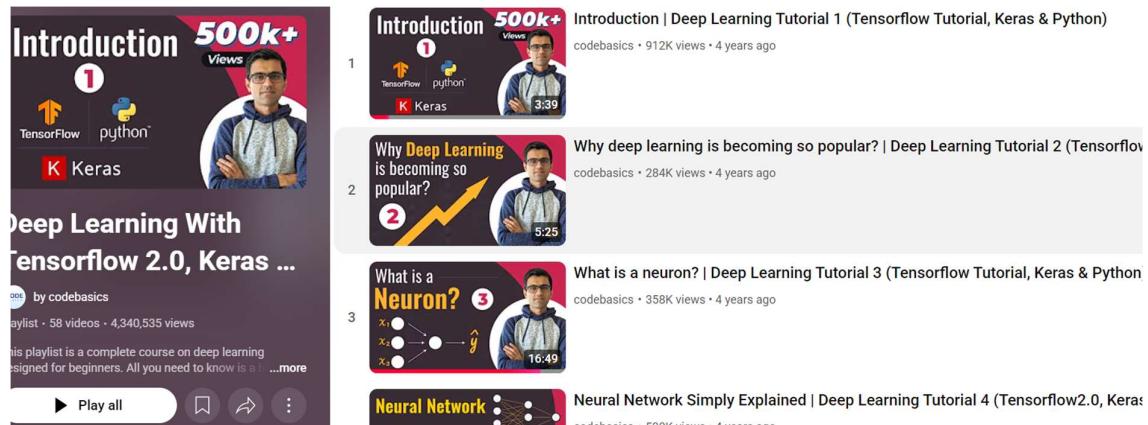
<https://drive.google.com/drive/folders/1rXrgUzzIdsyJ4xp05Suq7ioR5q1tOtFY>

## Deep Learning

Deep learning is a subset of machine learning that uses **neural networks** to model and solve complex problems. It is particularly effective for tasks involving large volumes of data and unstructured data types such as images, audio, and text.

This playlist is good to learn it in depth:

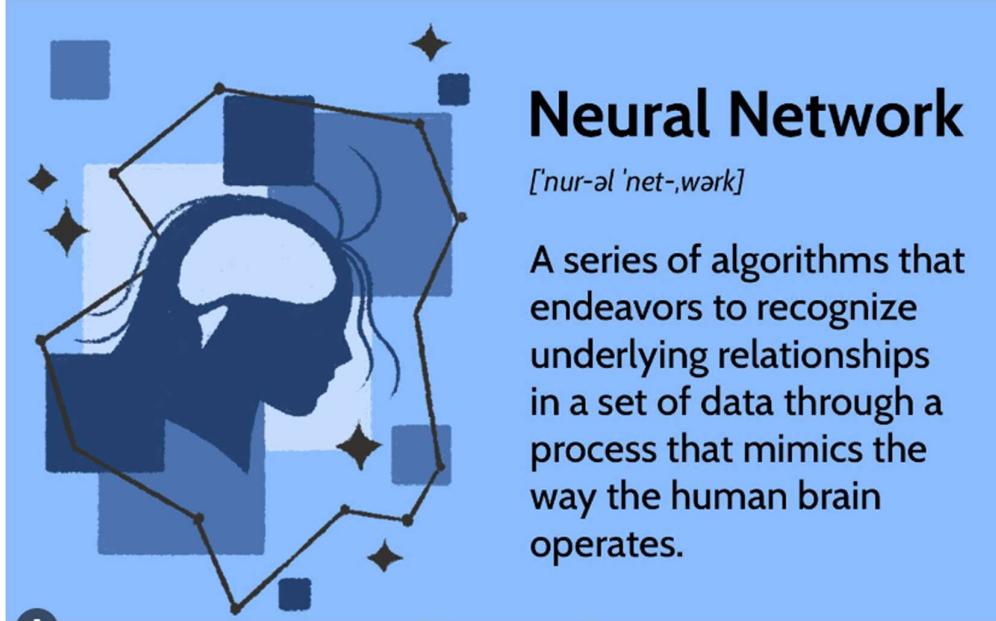
[https://www.youtube.com/playlist?list=PLEo1K3hjS3uu7CxAacxVndl4bE\\_o3BDtO](https://www.youtube.com/playlist?list=PLEo1K3hjS3uu7CxAacxVndl4bE_o3BDtO)



The image shows a YouTube playlist page for "Deep Learning With Tensorflow 2.0, Keras ..." by codebasics. The playlist has over 500k views and 58 videos. The first four videos in the playlist are listed:

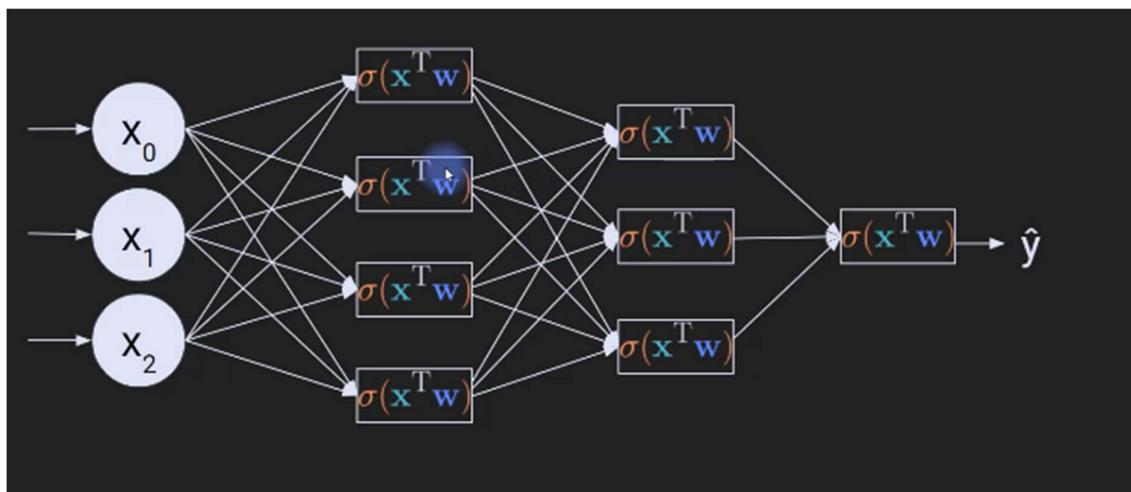
- 1. Introduction | Deep Learning Tutorial 1 (Tensorflow Tutorial, Keras & Python)
- 2. Why Deep Learning is becoming so popular? | Deep Learning Tutorial 2 (Tensorflow)
- 3. What is a Neuron? | Deep Learning Tutorial 3 (Tensorflow Tutorial, Keras & Python)
- 4. Neural Network Simply Explained | Deep Learning Tutorial 4 (Tensorflow2.0, Keras)

## Neural Networks



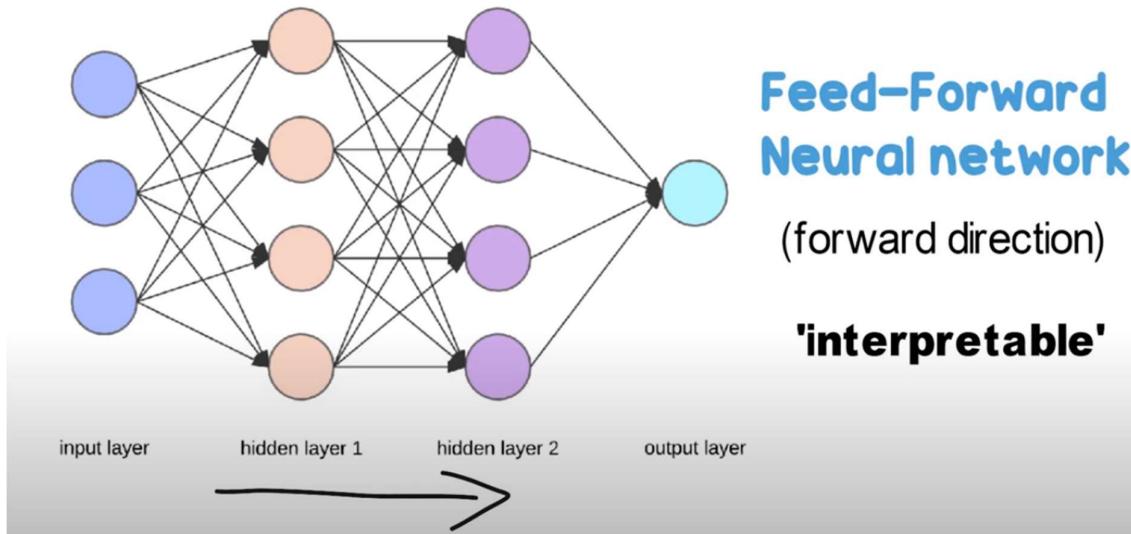
## Key Components Of A Neural Network

Nodes, Layers, Weights

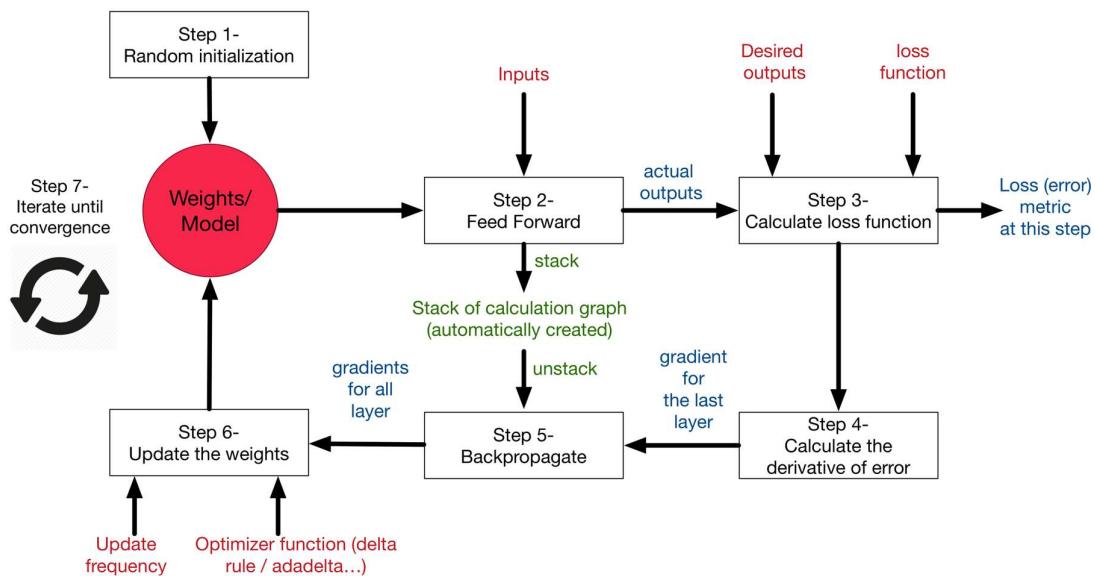


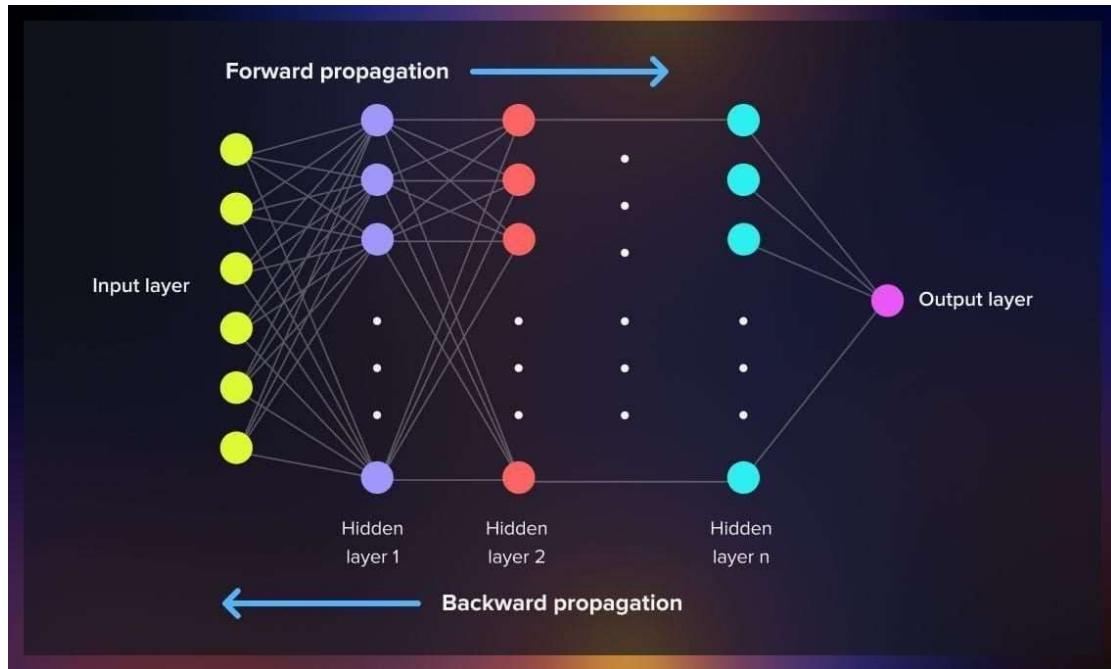
## Artificial Neural Network (ANN)

# 1. Artificial Neural Network (or ANN)



## Steps In Neural Network Processing

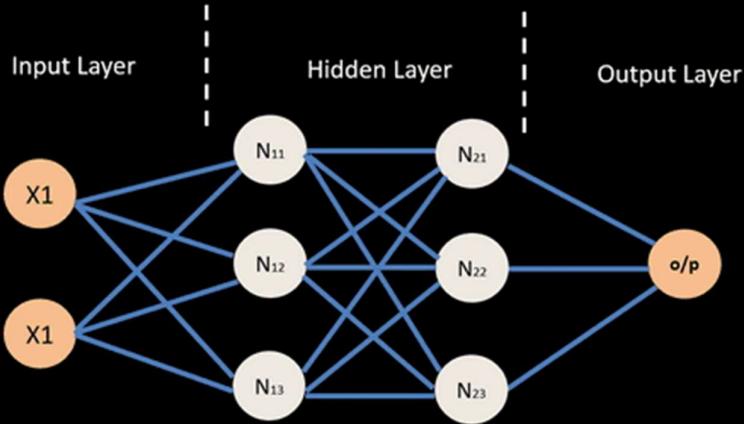




NN Visualization: <https://alexlenail.me/NN-SVG/index.html>

## Backpropagation

# Neural Network – Backpropagation



© machinelearningknowledge.ai

<https://www.youtube.com/watch?v=Ilg3gGewQ5U&t=734s>

Mastering Backpropagation:

<https://www.datacamp.com/tutorial/mastering-backpropagation>

## Gradient Descent

Gradient Descent in 3 minutes

<https://www.youtube.com/watch?v=qg4PchTECck>

Gradient Descent Visualizer

<https://uclaacm.github.io/gradient-descent-visualiser/#playground>

## Key use cases

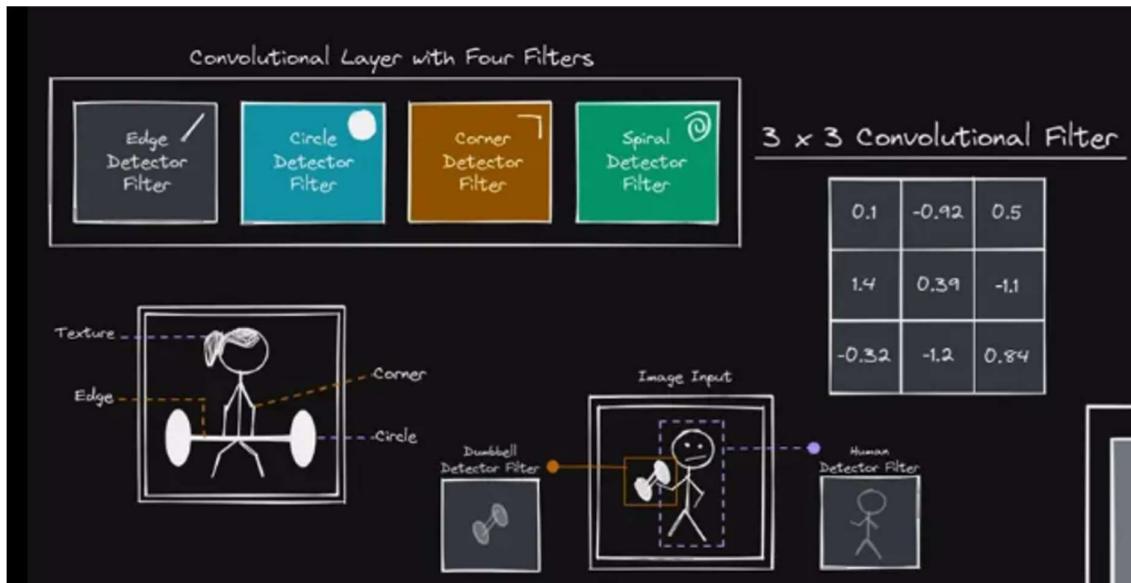
Artificial Neural Networks	Used for Regression & Classification
Convolutional Neural Networks	Used for Computer Vision
Recurrent Neural Networks	Used for Time Series Analysis
Self-Organizing Maps	Used for Feature Detection
Deep Boltzmann Machines	Used for Recommendation Systems
AutoEncoders	Used for Recommendation Systems

- Text classification using NLP (RNN, LSTM)
- Computer Vision (object detection) (Convolutional Neural Networks - CNN)

## Computer vision: Convolutional Neural Network (CNN)

[https://youtu.be/zfiSAzpy9NM?si=1jJF\\_sQJqF9HYdBn](https://youtu.be/zfiSAzpy9NM?si=1jJF_sQJqF9HYdBn)

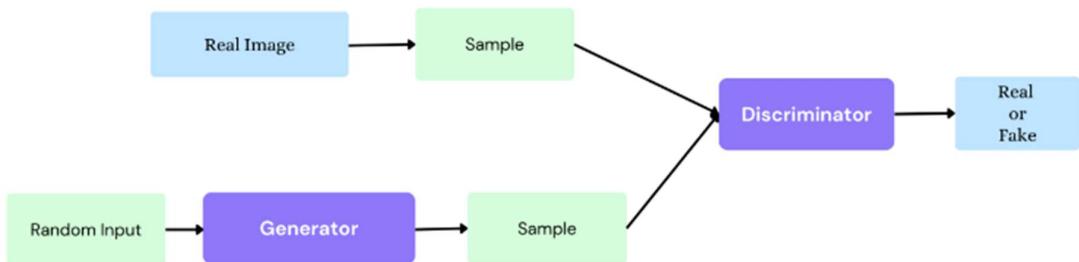
[https://adamharley.com/nn\\_vis/](https://adamharley.com/nn_vis/)



Generative Capabilities:

2014	- VAE - Encoder-decoder - GAN
2015	- Attention - U-NET - Pointer Network - NeuralStyle
2016	- ResNet - Pix2Pix
2017	- WGAN - WGAN-GP - CycleGAN - Transformers - MuseGAN - ProgressiveGAN
2018	- WorldModels - Self Attention GAN - BigGAN - BERT
2019	- StyleGAN - GPT-2 - MuseNET
2020	- GPT-3
2021	- DALL-E - CLIP
.....	
2025	- GPT-7

## Generative Adversarial Network (GAN)

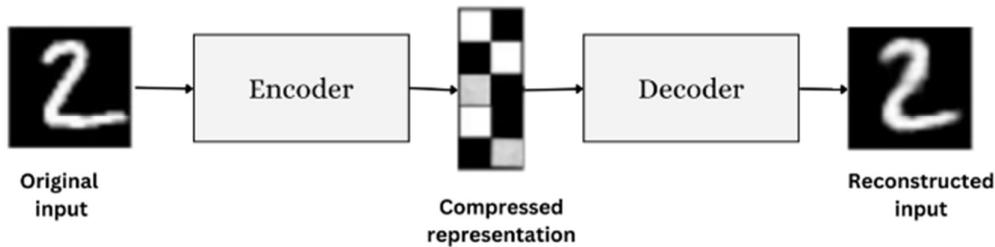


## GAN architecture diagram

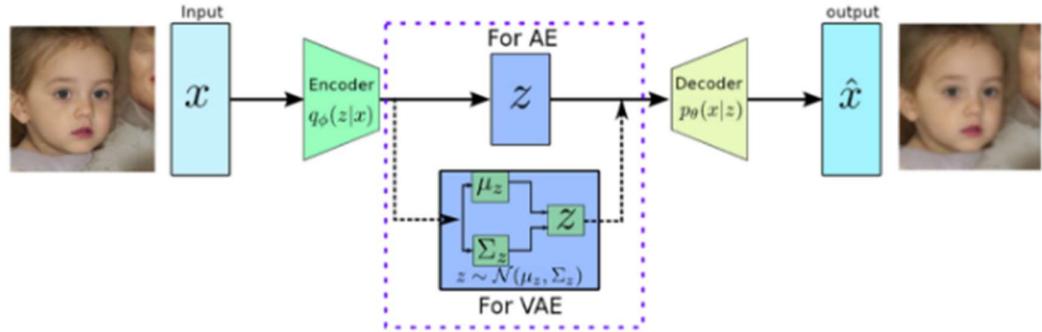
<https://towardsdatascience.com/understanding-generative-adversarial-networks-gans-cd6e4651a29>

## Variable Auto Encoder (VAE)

<https://towardsdatascience.com/understanding-variational-autoencoders-vaes-f70510919f73>

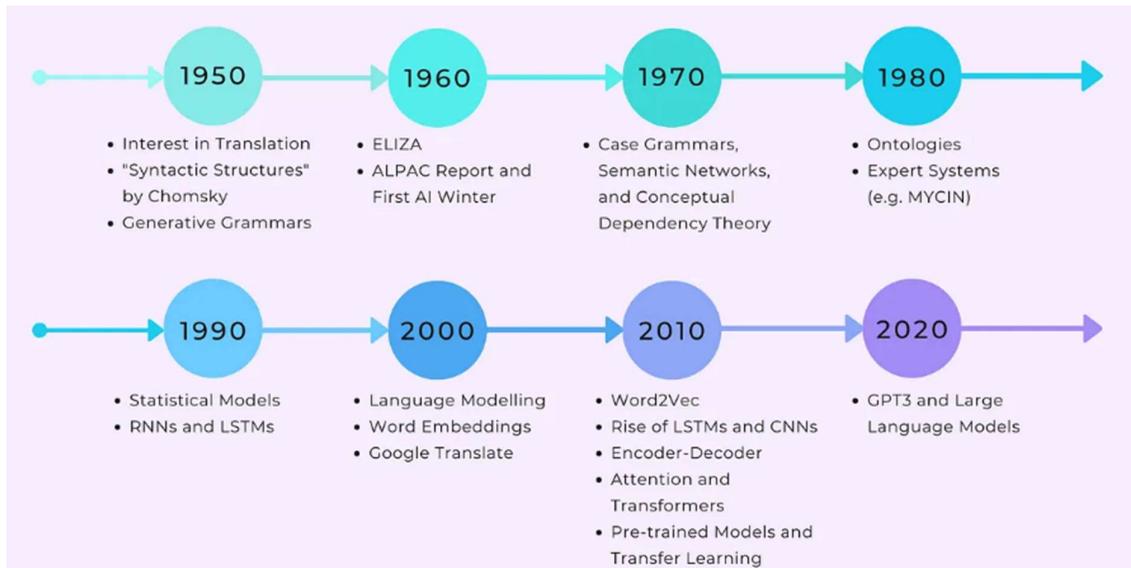


- 
- 
- VAEs aim to learn a probabilistic mapping between the data space and a latent space. This capability allows them to generate new samples closely resembling the patterns observed in the training data.
- 
- GANs are designed to generate realistic data by training a generator network to produce data that is identical to real data.
- GANs are more popular for image generation
- Autoencoders are used for feature learning, compression, reconstruction data and minimize error between input and output.
- Both VAEs and autoencoders use a reconstruction loss function to tune the neural networks using gradient descent.
- 
- The distinction between regular autoencoders (AEs) and variational autoencoders (VAEs) lies in how they handle the latent representation. In conventional autoencoders, the encoder transforms an input into a predetermined and fixed latent vector. On the other hand, in variational autoencoders, the encoder generates not a fixed latent vector but the parameters defining a probability distribution, typically a Gaussian distribution.
- The concept described above is visually depicted in the following figure.



- <https://ubiai.tools/comparing-gan-autoencoders-and-vae/>
- <https://medium.com/@ogre51/difference-between-autoencoder-ae-and-variational-autoencoder-vae-73004cc7e0fb>

## Natural Language Processing (NLP)



## Tokenization

Divide the texts into words or smaller sub-texts, which will enable good generalization of relationship between the texts and the labels.

This determines the “vocabulary” of the dataset (set of unique tokens present in the data).

gpt-3.5 has a vocabulary size of ~50000 tokens

Tokens	Characters
21	105

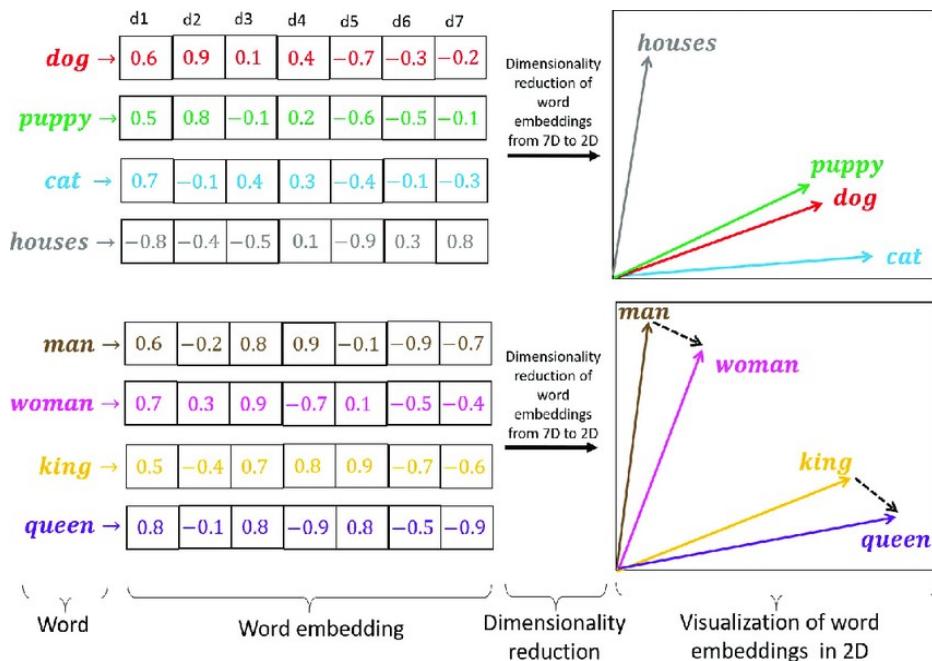
```
Einstein is arguably the greatest scientist of all times. He was born in
germany and lived in switzerland
```

## Vectors

Sequence of numbers - a mathematical representation for an object such as a word.

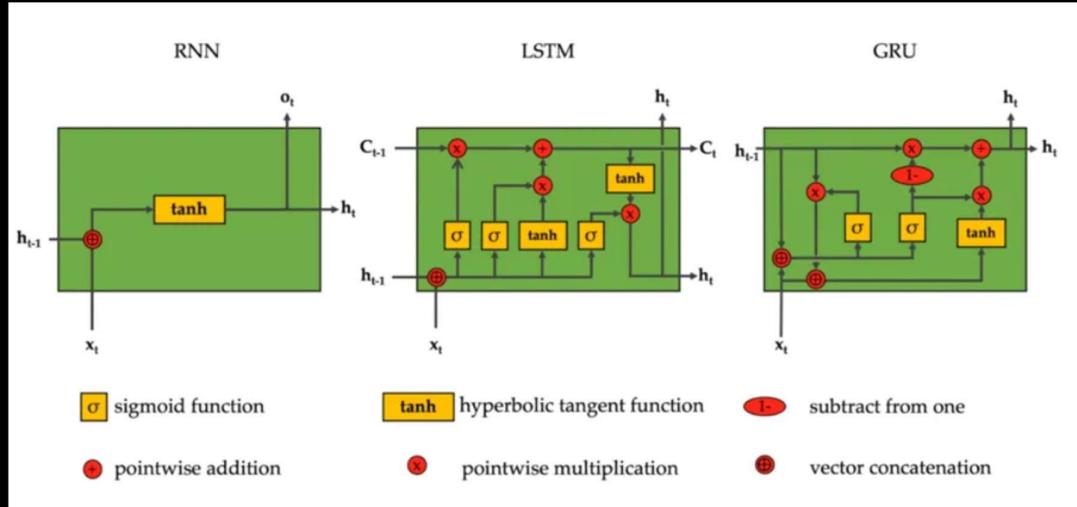
## Embeddings

The sequence of numbers capture the meaning of words, the relationship between words, and the context of different words as they are used naturally. Each number in the sequence represents a dimension, such as gender



## RNN, LSTM and GRUs

### RNN, LSTM and GRU



- Concept of limited memory (state) enabling sequential data processing
- Examples: Timeseries analysis, text processing, Audio processing
- Enabled NLP (Natural Language Processing) use-cases

<b>ANN</b>	<b>CNN</b>	<b>RNN</b>
Tabular or Text Data	Image Data	Sequence data
No Parameter Sharing	Yes	Yes
Operate on Fixed Length input	Operate on Fixed Length input	Don't
No Recurrent Connections	No Recurrent Connections	They are Possible
No Spatial Relationships	They are Possible	No Spatial Relationships
ANN is considered to be less powerful than CNN, RNN	CNN is considered to be more powerful than others	RNN includes less feature compatibility when compared to CNN
Having fault tolerance, Ability to work with incomplete knowledge	High accuracy in image recognition problems, & weight sharing	Remembers each and every information, & offers time series prediction

Ref: <https://arxiv.org/pdf/2401.02843.pdf>

## Challenges With NLP

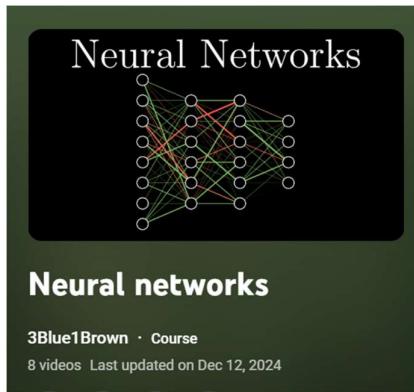
The word "**bank**" in the following two sentences has **different meanings**, but since they are assigned to the **same token id**, their word **embedding** vectors are the **same**.

1. "I went to the **bank** to withdraw some money."
2. "I went to the other side of the river **bank**."

- Earlier neural networks such as RNN, LSTM, GRU had limited memory / capacity to process context
- Suffered from problems such as vanishing gradient, exploding gradient

## Deep Learning References

3Blue1Brown – Neural Networks playlist:



[https://www.youtube.com/playlist?list=PLZHQBObOWTQDNU6R1\\_67000Dx\\_ZCJB-3pi](https://www.youtube.com/playlist?list=PLZHQBObOWTQDNU6R1_67000Dx_ZCJB-3pi)

[https://www.youtube.com/playlist?list=PLqwozWPBo-FvuHWx3\\_aYwG2WVdbb-wC6q](https://www.youtube.com/playlist?list=PLqwozWPBo-FvuHWx3_aYwG2WVdbb-wC6q)



[https://www.youtube.com/playlist?list=PLeo1K3hjS3uu7CxAacxVndl4bE\\_o3BDtO](https://www.youtube.com/playlist?list=PLeo1K3hjS3uu7CxAacxVndl4bE_o3BDtO)

<https://ml-playground.com>

<https://playground.tensorflow.org/>

<https://www.freecodecamp.org/news/deep-learning-neural-networks-explained-in-plain-english/>

Deep Learning book by Ian Goodfellow, creator of G

<https://www.deeplearningbook.org/>

Variable Auto Encoders

<https://github.com/Jackson-Kang/Pytorch-VAE-tutorial>

How Stable Diffusion Works (AI Image Generation)

<https://www.youtube.com/watch?v=sFztPP9qPRc>

Interactive Node link visualization:

[https://adamharley.com/nv\\_vis/](https://adamharley.com/nv_vis/)

Enhanced credit card fraud detection based on attention mechanism and LSTM deep model:

<https://journalofbigdata.springeropen.com/articles/10.1186/s40537-021-00541-8>

Vehicle marketplace recommendation engine with DL:

<https://slideplayer.com/slide/14614243/>

## The Transformer Model

Transformers are a type of neural network that have a **unique ability to recognize long-range connections within sequences**. They are particularly useful for **tasks like generating text**, as the model needs to comprehend the preceding words in order to produce the next one.

**The introduction of transformers in 2018 was a groundbreaking moment for the field of natural language processing.**

Transformers  
are able to  
learn long-  
range  
dependencies  
in sequences.

Transformers  
are able to be  
trained on very  
large datasets.

Transformers  
are able to be  
parallelized

---

## Attention Is All You Need

---

Ashish Vaswani<sup>\*</sup>  
Google Brain  
avaswani@google.com

Noam Shazeer<sup>\*</sup>  
Google Brain  
nseide@google.com

Niki Parmar<sup>\*</sup>  
Google Brain  
nikip@google.com

Jakob Uszkoreit<sup>\*</sup>  
Google Brain  
uszko@google.com

Llion Jones<sup>\*</sup>  
Google Research  
llion@google.com

Aidan N. Gomez<sup>†</sup>  
University of Toronto  
aidan@cs.toronto.edu

Lukasz Kaiser<sup>\*</sup>  
Google Brain  
lukasz.kaiser@google.com

Illia Polosukhin<sup>‡</sup>  
illia.polosukhin@gmail.com

### Abstract

The dominant sequence-to-sequence 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, dispelling the need for recurrent or convolutional units. 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, surpassing the existing best results, including ensemble, by 2.8 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

The boy is holding a blue ball

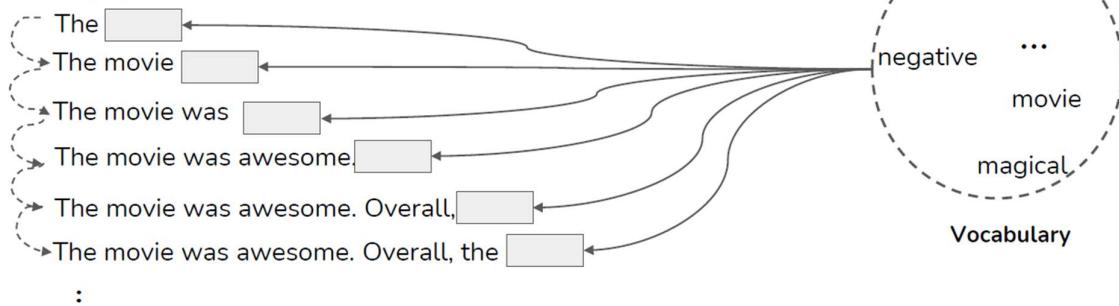
## Transformers: Next token prediction

During inference, the LLM predicts the next word in the input sequence.

Input word = prompt

The

Output, word-by-word



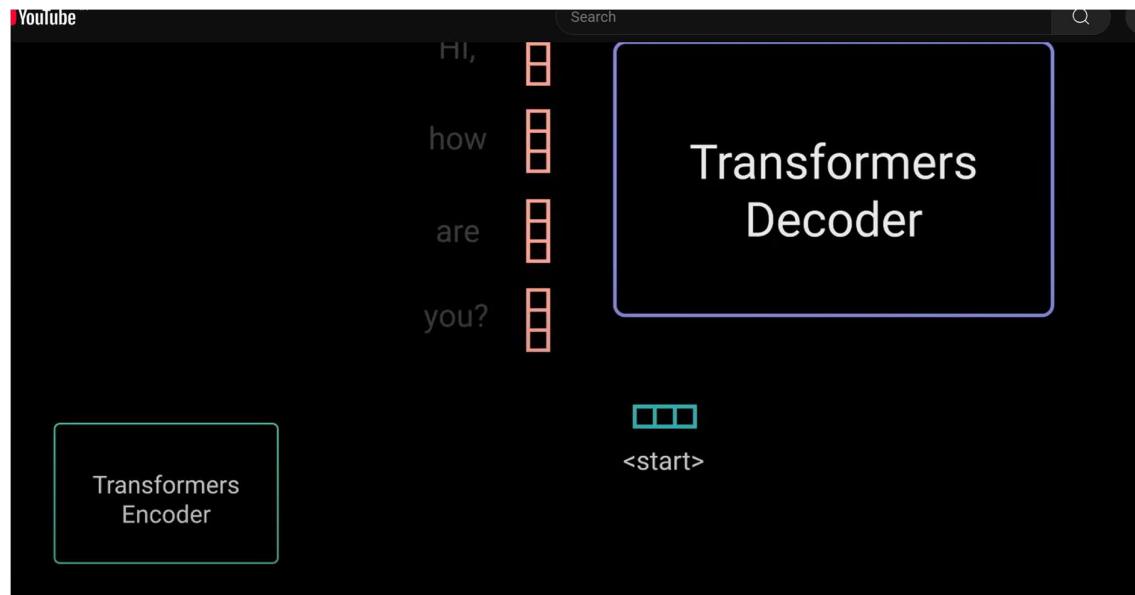
**self attention:** operates on **one set of data** (a word phrase)

**cross attention:** operates on **two sets of data** (word phrase + image)

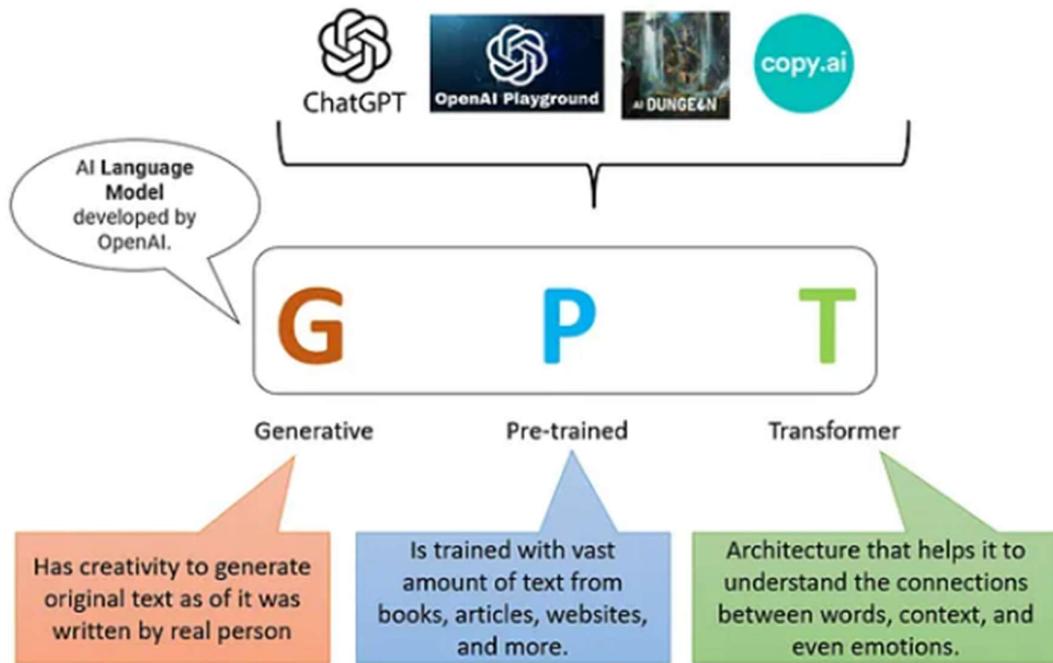
## Transformer Architecture

AI Hacker: Illustrated Guide to Transformers Neural Network: A step by step explanation

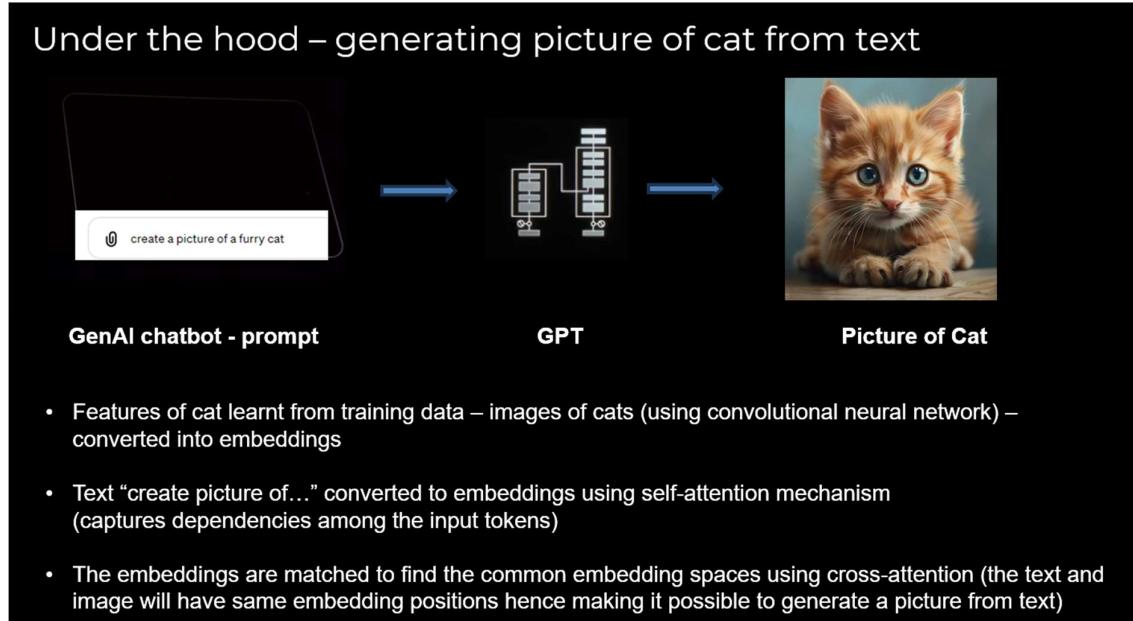
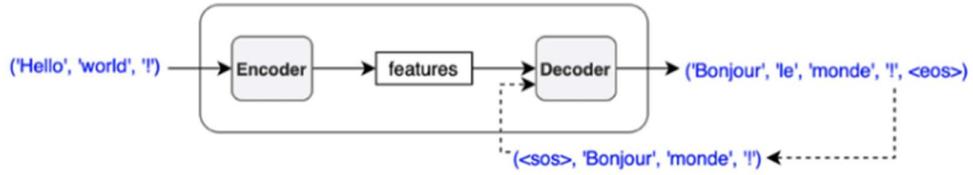
<https://www.youtube.com/watch?v=4Bdc55j80l8>



## GPTs and LLMs



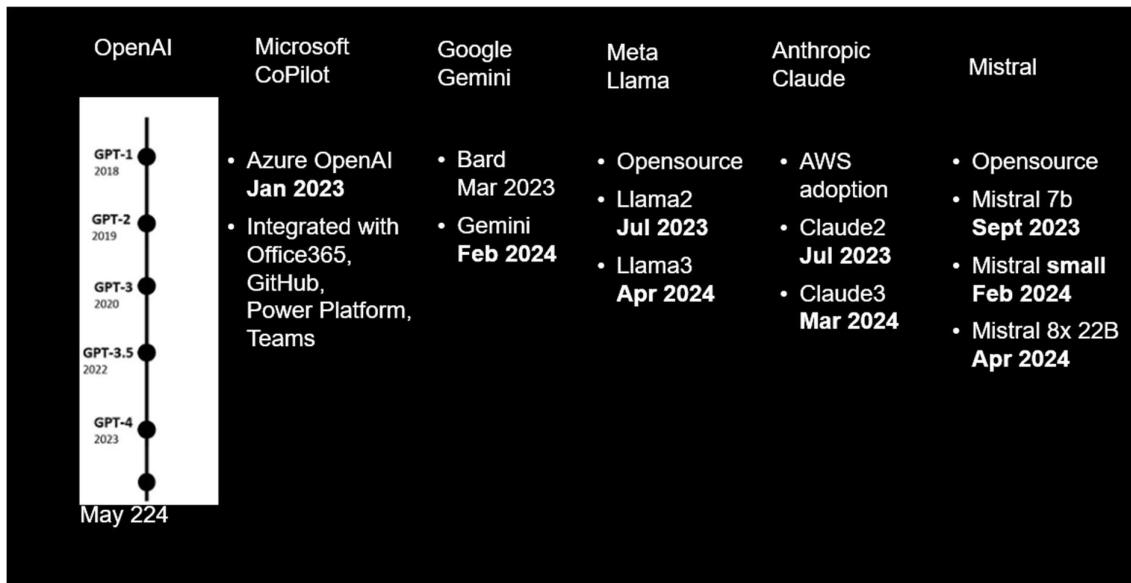
### Transformers: Encoders and Decoders



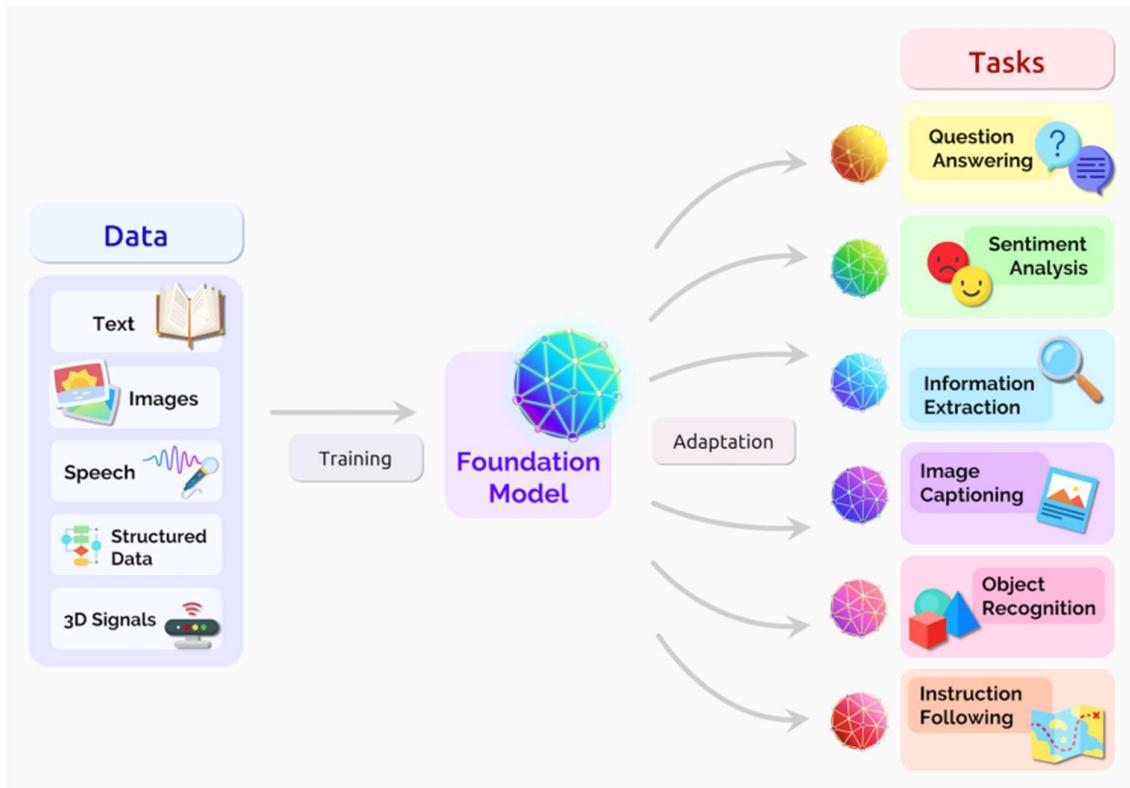
## Large Language Models (LLMs)

- Pre-trained Deep Learning models
- Trained on massive textual data (tokens)
- Data source
  - Online contents and literature
  - News and current affairs
  - Social media
  - Other sources like blogs, comments, feedback, reviews etc.
- Millions of parameters (weights, neurons)
- LLMs can be fine-tuned

- for specific problems
- for specific domains
- Using smaller datasets



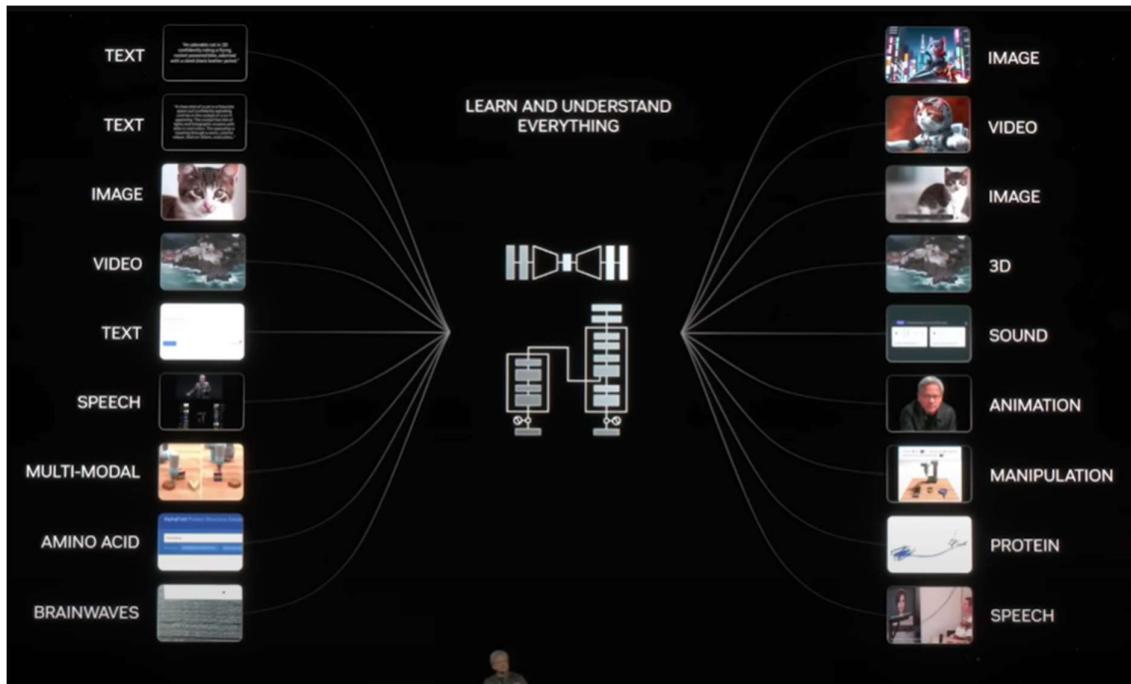
## Foundation Models



## Multi Modal

**multi-modal** refers to AI systems that can process, generate, or interact with multiple types of data modalities, such as text, images, audio, video, and more.

Multi-modal AI models are designed to integrate and understand relationships across these different modalities, enabling them to perform tasks that involve combinations of data types.



## Gen AI Core capabilities

Inferencing, Question & Answer based on pre-trained knowledge

Prompt: "What is the capital of France?"  
 Response: "Paris."

Prompt: A detailed article about climate change.  
 Response: "Climate change is a significant global issue caused by human activities leading to rising temperatures and environmental impacts."

Translation: Translating text from one language to another.  
 Prompt: "Hello, how are you?" (English)  
 Response: "Hola, ¿cómo estás?" (Spanish)

Text Generation: Producing creative content such as stories, poems, or essays.  
 Prompt: "Once upon a time in a faraway land,"  
 Response: "there was a small village surrounded by towering mountains and lush forests."

Sentiment Analysis: Determining the sentiment or emotional tone of a text.

Example: Input: "I love this product!"  
 Output: "Positive sentiment."

## Summarization

It can summarize large documents, videos, images, flowcharts, audio – any modality, multiple languages

### Extractive Summarization:

The LLM selects and concatenates key sentences or phrases from the original text

### Abstractive Summarization:

The LLM generates new sentences that paraphrase the original content.

### Practical applications

News summarization, scientific papers summarization, legal documents, product reviews

## Reasoning

Reasoning in AI involves higher-level cognitive processes (compared to inferencing) where the AI system makes decisions based on logic, understanding relationships, and drawing conclusions from given information.

### Deductive reasoning:

All humans are mortal. Socrates is a human. Therefore, Socrates is mortal.

### Inductive reasoning

Every time I water my plant, it grows taller. Watering plants regularly helps them grow taller.

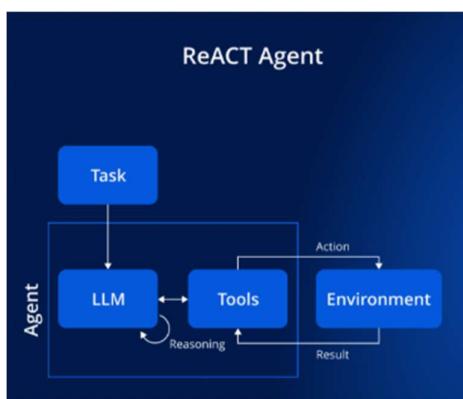
### Abductive reasoning

Ability to choose best possible inference. Lawn is wet – rain? Or sprinkler?

### Practical applications

Analyse market data to analyse patterns, customer behaviour analysis, drug discovery

## ReAct – Reasoning and Action



1. **Input:** The agent receives a task description in natural language, which the core LLM processes.
2. **Reasoning:** The LLM breaks down the task into smaller steps, analyzes the situation, considers available information, and plans the necessary actions.
3. **Action:** Based on this reasoning, the LLM selects the appropriate tool (e.g., search engine, database, API) and performs actions to gather information or interact with the environment. This could involve querying Wikipedia for facts or retrieving data from a company database.
4. **Observation:** The agent observes the results of its actions and updates its knowledge. This new information is used to refine its reasoning in the next cycle.
5. **Response:** Finally, the agent generates a response based on its reasoning and the information gathered.

<https://www.linkedin.com/pulse/react-agents-revolutionizing-ai-reasoning-action-allen-adams-fxqrc/>

## Customizing LLM/GPT behaviours

The need to customize

## Why customize?

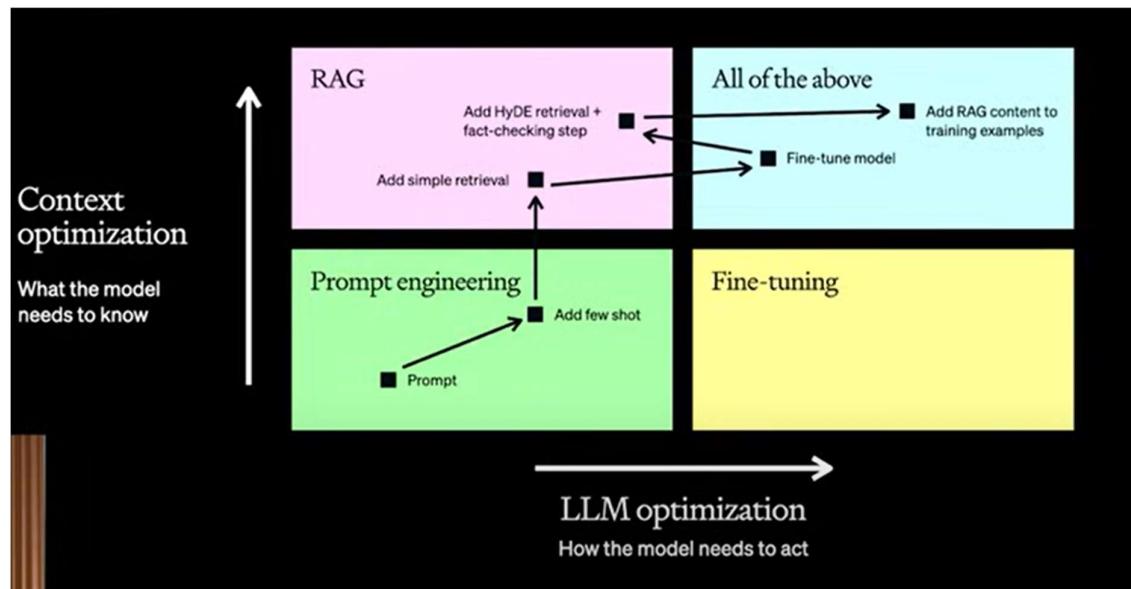
- LLMs lack up to date / real time data  
Their knowledge is restricted till the training date cut-off of the model
- Some use-cases demand models work with proprietary / IP protected / subscription based datasets / documents
- Reduce hallucinations, improve accuracy and response, change the response tone, behaviour

## Customization options

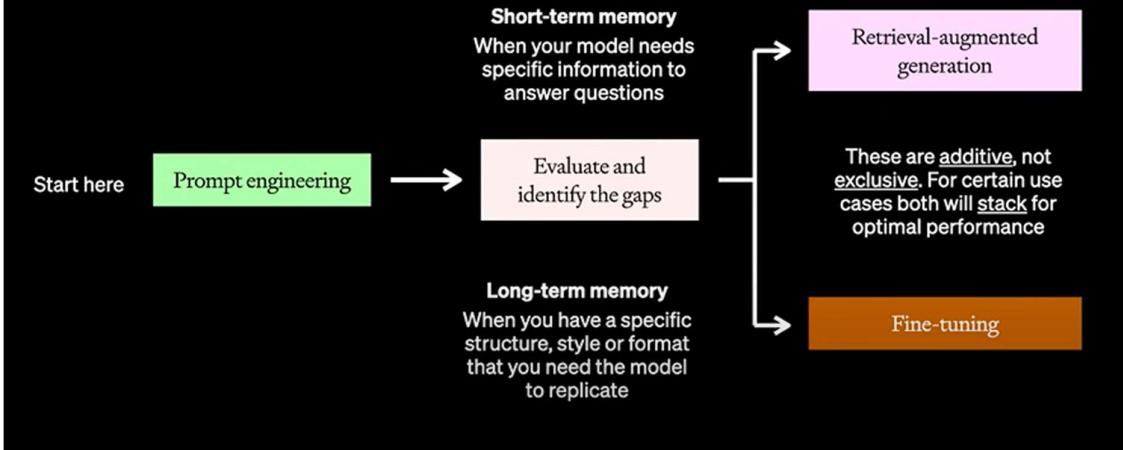


- **Prompt Engineering**  
Refines model input to guide its output.
- **Retrieval Augmented Generation (RAG)**  
Augments LLMs knowledge with custom data.  
Merges prompt engineering with index & retrieval mechanism to create a customized context for the LLM
- **Full Fine-tuning**  
Adjusts all parameters of the LLM using task-specific data.
- **Parameter-efficient Fine-tuning (PEFT)**  
Modifies select parameters for more efficient adaptation.
- **Agents**  
Agents enable tools and functions that can be invoked by the LLM to fetch specific data. LLM acts as a ReAct (Reasoning and Action) engine that can invoke any tool on demand to complete a certain assigned task (prompt)

A combination can be leveraged:



## RAG Vs Fine tuning



## Prompt Engineering

Prompt engineering is the process of crafting a high-quality prompt for a GenAI model to generate response

### Why are Prompts Important ?



**Provide clear direction:** Prompts provide clear direction to AI, ensuring it understands your needs.

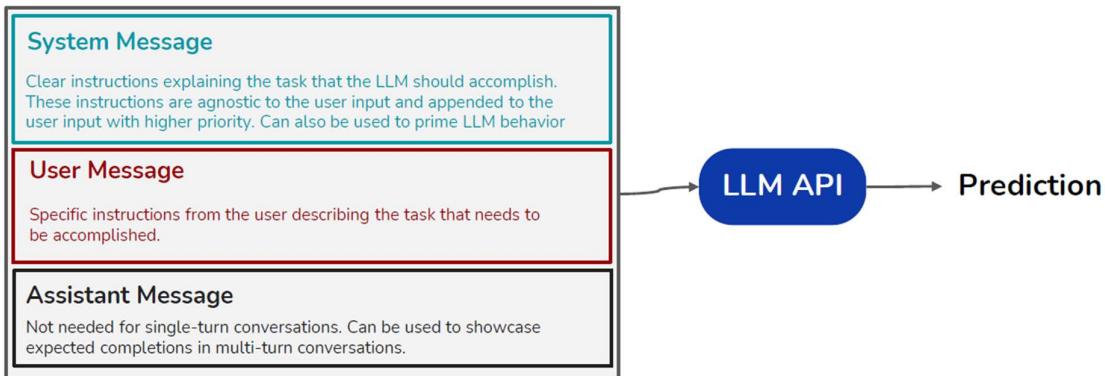


**Improve accuracy:** Well-crafted prompts enhance the accuracy of AI responses.



**Save time:** Specific prompts save time by reducing back-and-forth clarifications.

Azure Open AI APIs are compatible with the Open AI APIs and have the following three components.



## Zero, One, and Few-shot Prompting



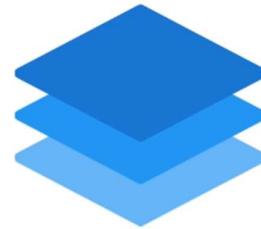
Gives task directly

e.g. translate "this cat is sleeping" into French



Provides one example

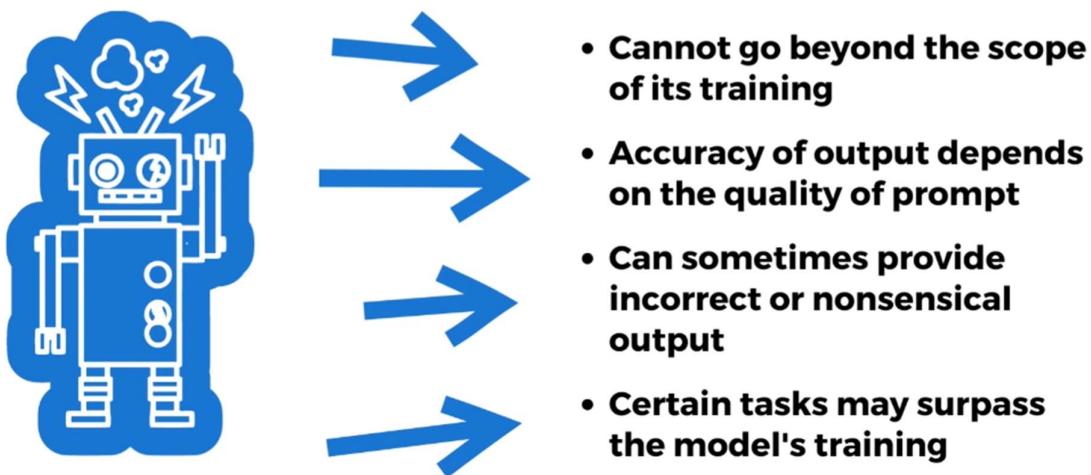
e.g. continue this story, here is an example



Provides multiple examples

e.g. write a review based on these samples

# Understanding the Limitations



## Understanding Tokens

### Understanding tokens

**Token:**

- Smallest unit into which text data can be broken down for an AI model to process (Think 1,000 tokens to be about 750 words)
- Exact tokenization process varies between models.
- Request Tokens are generated from the input prompt. Response Tokens are generated by the model from the output.

**Token Limit:**

- Total tokens (input + output) a model can process in one request

**Context Window:**

- The model's capacity to retain and consider tokens for generating responses.

Tokenizer tool:

<https://platform.openai.com/tokenizer>

# Prompt Engineering Tools

## 1. PromptPerfect: Create a perfect prompt by describing the use-case:

 Vijay Agrawal

create a prompt for generating a loan approval document. Ask for customer name, loan tenure, interest rate, address and generate a nicely formatted loan approval letter

 Optimizer

Understood, let's create a prompt to generate a loan approval document. Here is my plan:

1. The model will be tasked to generate a loan approval letter.
2. The prompt will include placeholders for customer details such as name, loan tenure, interest rate, and address.
3. The generated letter should be professionally formatted and articulate the key points clearly.

Let's write the prompt:

**PROMPT**

You are tasked with generating a professionally formatted loan approval letter.  
To do so, you will need the following details: customer name, loan tenure, interest rate, and address.

The letter should include the following elements:

- A clear and formal salutation

Tool URL:

<https://promptperfect.jina.ai/>

## 2. AI For Work

<https://www.aiforwork.co/>

Google AI Studio – Prompt Gallery

<https://aistudio.google.com/app/gallery>

30,000+ PROFESSIONALS USE AI FOR WORK

# ChatGPT Prompts to Get Work Done ✓

[Your Account](#)

Select Your Department 👇



Legal



Personal Development



Human Resources



Executive Management



Media And Communication



Entrepreneur



Finance



Retail



Customer Service



Marketing

## Prompt Engineering References

### Prompt Engineering 101 - Crash Course & Tips



AssemblyAI  
154K subscribers

[Subscribe](#)

167K views 1 year ago #promptengineering #chatgpt #MachineLearning

<https://www.youtube.com/watch?v=aOm75o2Z5-o>

## Retrieval Augmented Generation (RAG)

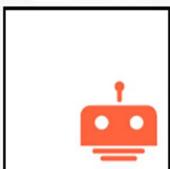
# RAG Use Cases

The development of RAG technique is rooted in use cases that were limited by the inherent weaknesses of the LLMs. As of today some commercial applications of RAG are in -



### Document Question Answering Systems

By providing access to proprietary enterprise document to an LLM, the responses are limited to what is provided within them. A retriever can search for the most relevant documents and provide the information to the LLM. Check out [this blog](#) for an example



### Conversational agents

LLMs can be customised to product/service manuals, domain knowledge, guidelines, etc. using RAG. The agent can also route users to more specialised agents depending on their query. [SearchUnify has an LLM+RAG powered conversational agent](#) for their users.



### Real-time Event Commentary

Imagine an event like a sports or a new event. A retriever can connect to real-time updates/data via APIs and pass this information to the LLM to create a virtual commentator. These can further be augmented with Text To Speech models. [IBM leveraged the technology for commentary during the 2023 US Open](#)



### Content Generation

The widest use of LLMs has probably been in content generation. Using RAG, the generation can be personalised to readers, incorporate real-time trends and be contextually appropriate. [Yarnit is an AI based content marketing platform that uses RAG for multiple tasks.](#)



### Personalised Recommendation

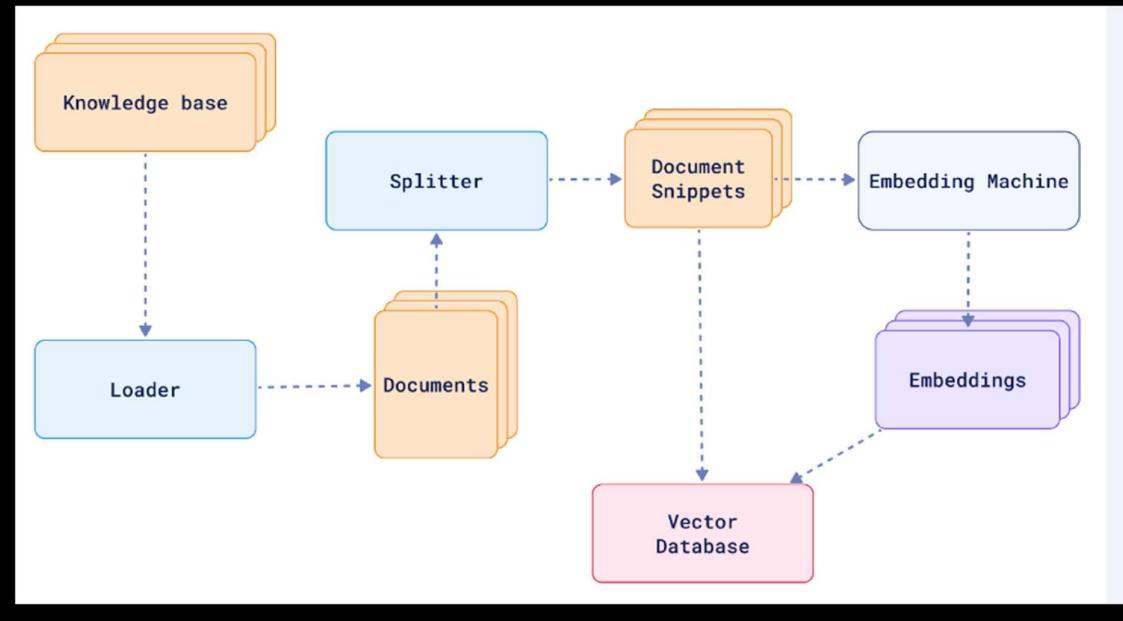
Recommendation engines have been a game changes in the digital economy. LLMs are capable of powering the next evolution in content recommendations. Check out [Aman's blog](#) on the utility of LLMs in recommendation systems.



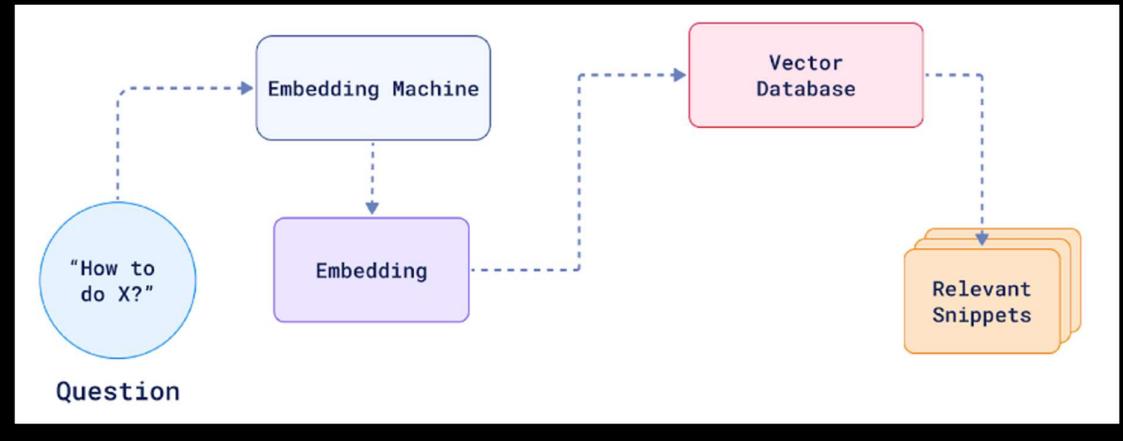
### Virtual Assistants

Virtual personal assistants like Siri, Alexa and others are in plans to use LLMs to enhance the experience. Coupled with more context on user behaviour, these assistants can become highly personalised.

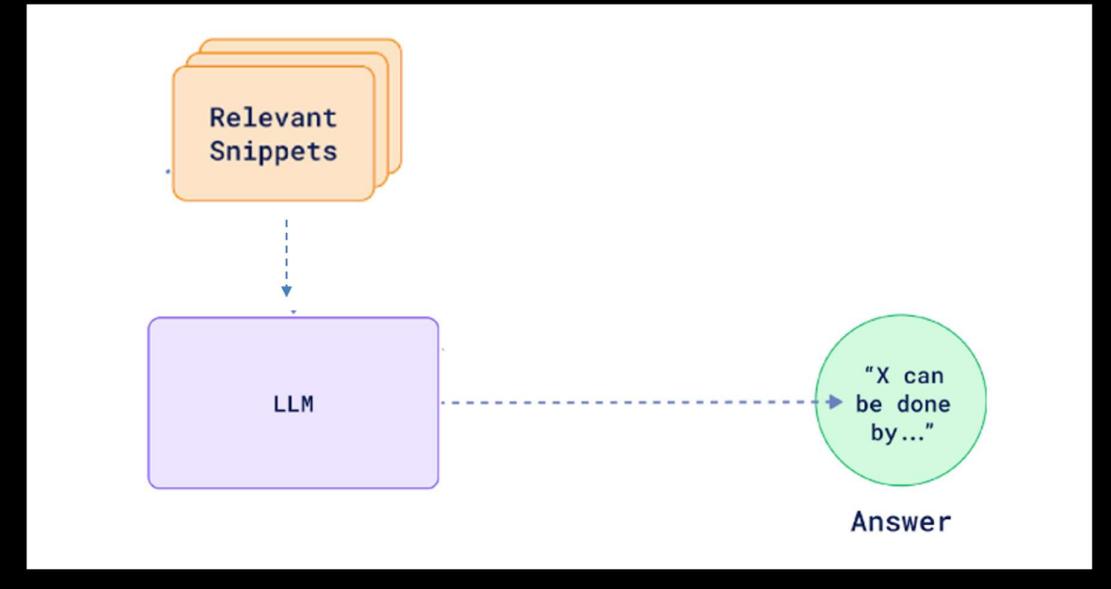
## RAG step 1 – Indexing (of new knowledge)



## RAG step 2: Retrieval



## RAG step 3: Generation



## Popular Embedding Models

**word2vec** Google's Word2Vec is one of the most popular pre-trained word embeddings. The official paper -  
<https://arxiv.org/pdf/1301.3781.pdf>

**GLOVE** The 'Global Vectors' model is so termed because it captures statistics directly at a global level. The official paper -  
<https://nlp.stanford.edu/pubs/glove.pdf>

**fastText** Facebook's AI research, fastText builds embeddings composed of characters instead of words. The official paper -  
<https://arxiv.org/pdf/1607.04606.pdf>

**Elmo** Embeddings from Language Models, are learnt from the internal state of a bidirectional LSTM. The official paper -  
<https://arxiv.org/pdf/1802.05365.pdf>

**BERT** Bidirectional Encoder Representations from Transformers is a transformer bases approach. The official paper -  
<https://arxiv.org/pdf/1810.04805.pdf>

# Top 10 Retrievers / Vector Stores (2024)

1	 chroma	6	 Milvus
2	 Ecos	7	 learn
3	 Pinecone	8	 elasticsearch
4	 qdrant	9	 MongoDB Atlas
5	 neo4j	10	 neo4j

Source: LangChain: <https://x.com/LangChainAI/status/1869812624998969836>

## RAG Optimization knobs

### Chunking

Semantic chunking  
Agentic chunking

### Indexing

vector compression  
Metadata  
Hyperparameter tuning  
Choose right ANN algo

### Query optimization

Prompt engineering  
Query transformation, splitting  
Query expansion, compression  
Query routing, augmentation  
*Lost in the middle problem*

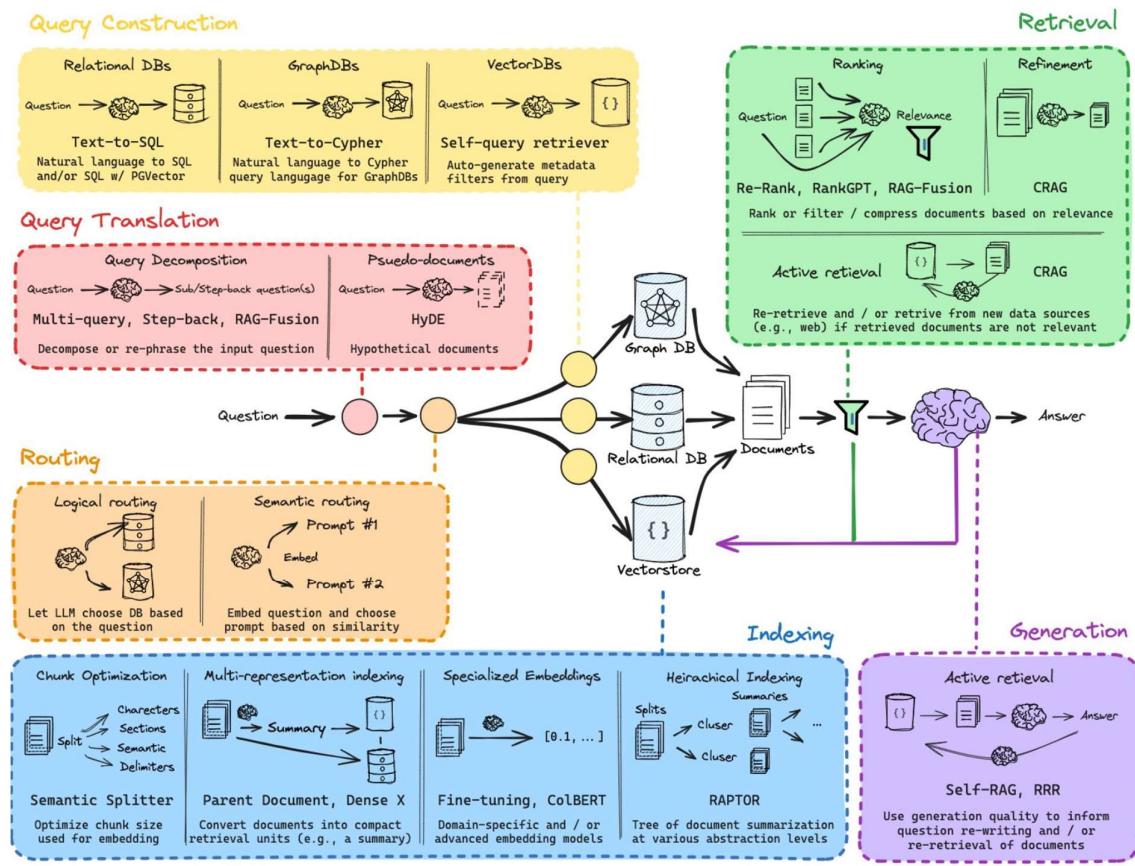
### Retrieval knobs

Embedding models  
hybrid search weighting  
re-ranker models  
Multi-index search configurations  
Metadata filtering.  
Time-weighted retrieving  
Summary based retrieval  
Ensemble retriever

### Generation knobs

Choice of LLM  
LLM configuration & Tuning  
(PEFT, FFT)

## Advanced RAG Implementation:



Source: <https://x.com/LangChainAI/status/1754915914796216654>

## Quantization to improve RAG performance:

<https://qdrant.tech/articles/what-is-vector-quantization/>

## RAG References

Top 9 Vector databases you should know:

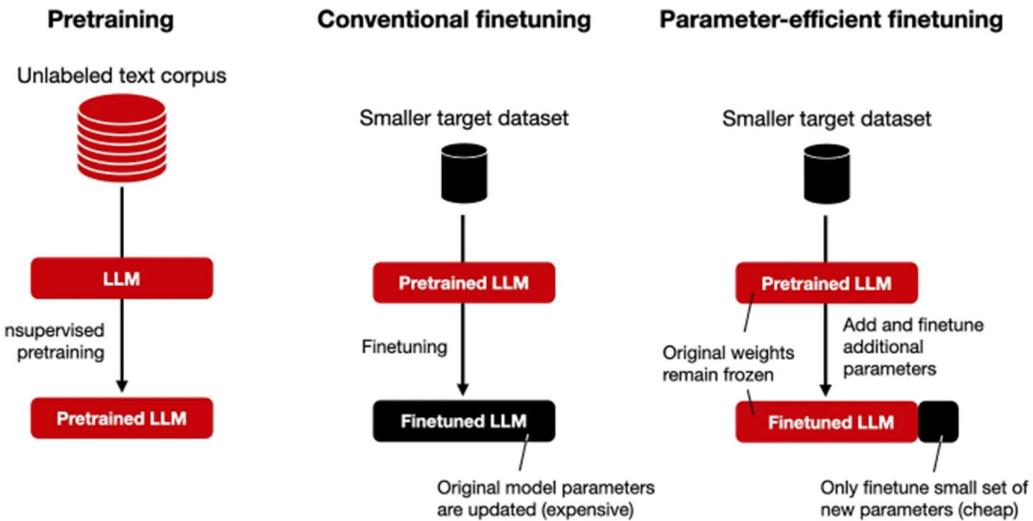
[https://machinelearningknowledge.ai/top-vector-databases-you-should-know/#google\\_vignette](https://machinelearningknowledge.ai/top-vector-databases-you-should-know/#google_vignette)

How does similarity search work?

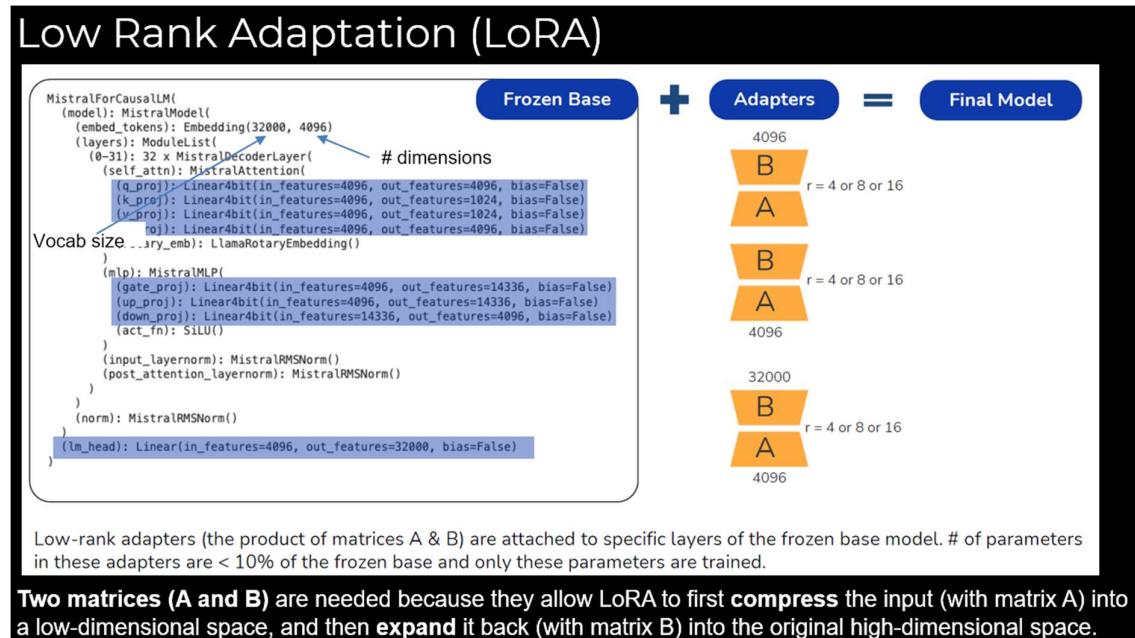
<https://towardsdatascience.com/similarity-search-part-4-hierarchical-navigable-small-world-hnsw-2aad4fe87d37/>

<https://www.youtube.com/watch?v=QvKMwLjdK-s>

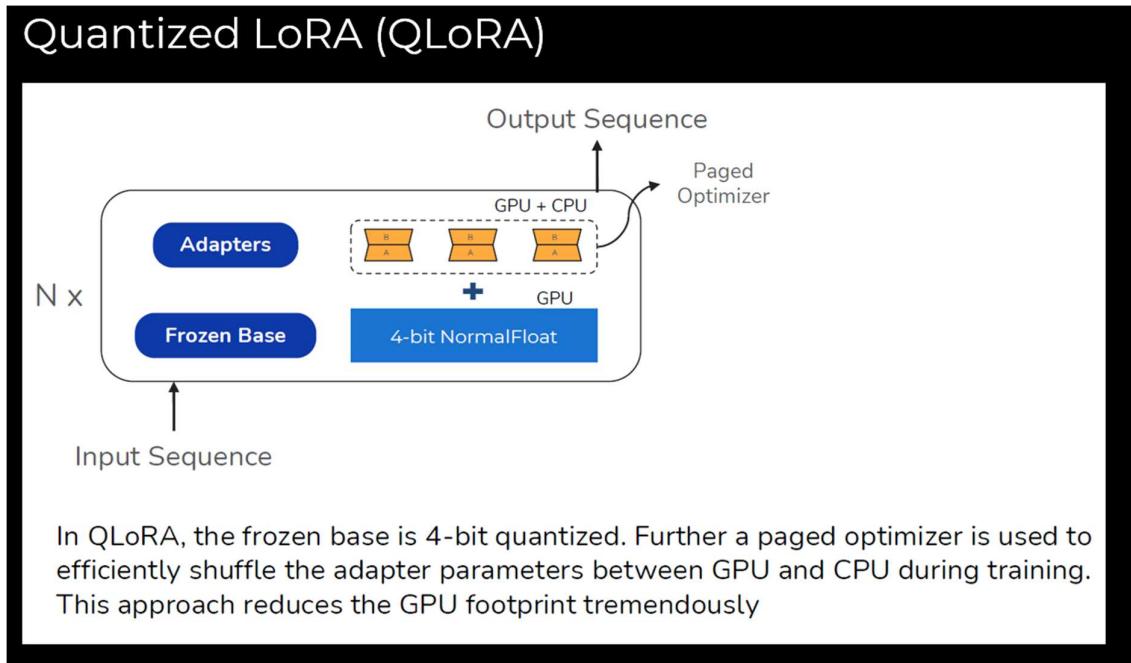
# LLM Fine Tuning



## Low Rank Adaptation (LoRA)



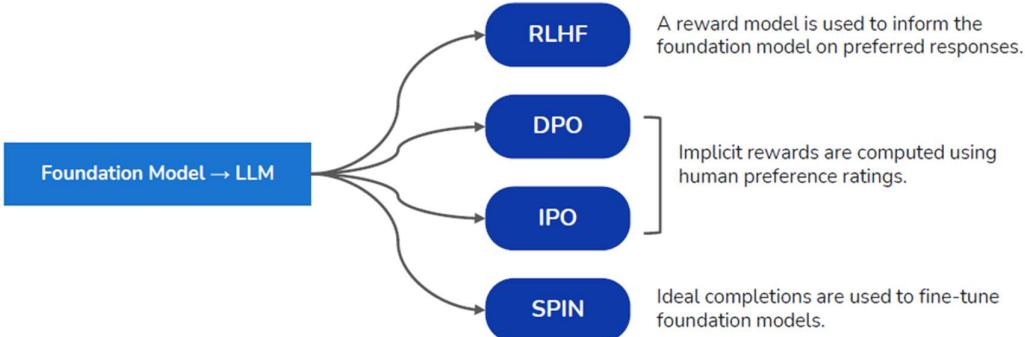
## Quantized Low Rank Adaptation (QLoRA)



## Alignment Fine Tuning (RLHF etc.)

## Alignment fine tuning techniques

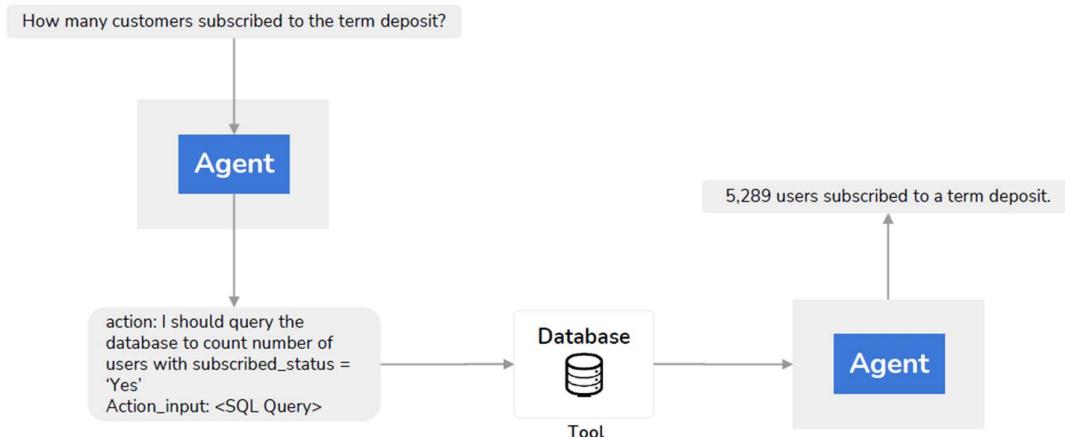
Foundation models are aligned with human preferences by rewarding high human ratings of model responses.



PPO: Proximal Policy Optimization

DPO: Direct Preference Optimization

## Agents



## NVIDIA CEO on Agents Being the Future of AI



Matthew Berman  
369K subscribers

Join



Subscribed



93K views 3 months ago • Members first

Join My Newsletter for Regular AI Updates

<https://forwardfuture.ai>

more



<https://www.youtube.com/watch?v=SMnBnGQqlR4>

zbrain.ai/agents/

we-eax Webmail Bookmarks Travel Local Invest Bookmarks social New Tab stockholm Kick off dry January... Slideshow Maker Untitled - FlexClip Untitled - til Adobe Acrobat

**ZBrain**

PLATFORM INDUSTRIES AGENTS BETA RESOURCES

All Agents Legal Human Resources Sales Finance Marketing Customer Service Procurement Informa

**ZBrain AI Agents: Streamlining Enterprise Operations**

ZBrain AI agents are designed to automate specific tasks within enterprise processes using GenAI. By deploying these agents, organizations can reduce manual workload and enhance operational productivity.



**Utilities**  
**Meeting Research Agent**  
Provides meeting preparation reports with details about external attendees, enhancing meeting effectiveness.



**Customer Service**  
**Customer Support Email Responder Agent**  
Monitors the email inbox for customer queries, retrieves answers from the knowledge base, sends replies, or creates tickets for unresolved queries.



**Utilities**  
**Document Comparison Agent**  
Compares documents to previous versions, ensuring consistency, accuracy, and compliance with predefined standards.



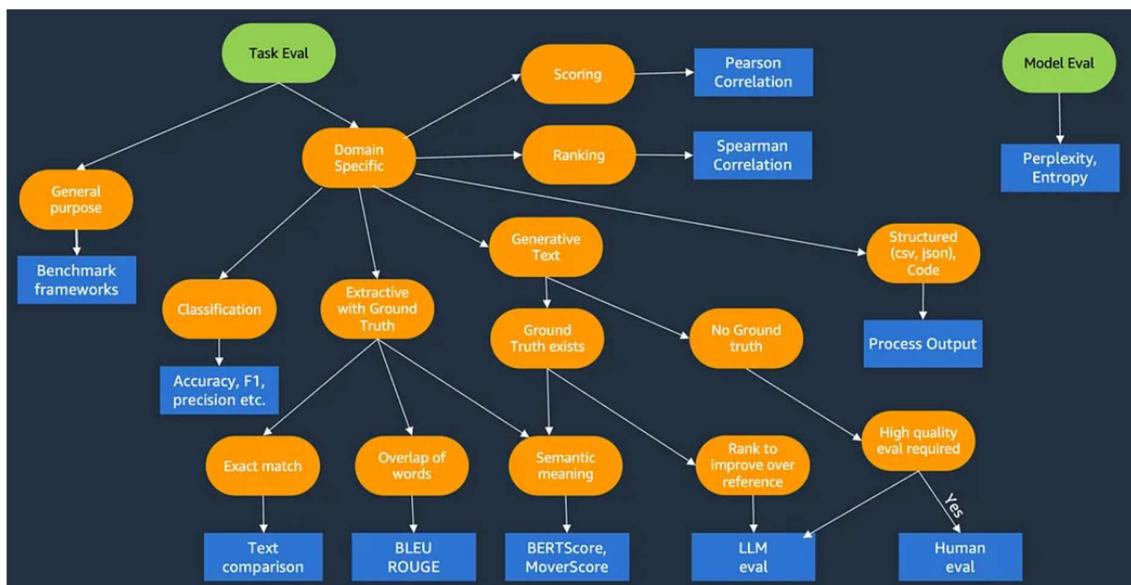
**Human Resources**  
**Acknowledgment Email Sender Agent**  
Automatically sends acknowledgment emails based on predefined criteria, ensuring timely and consistent communication with employees and candidates.

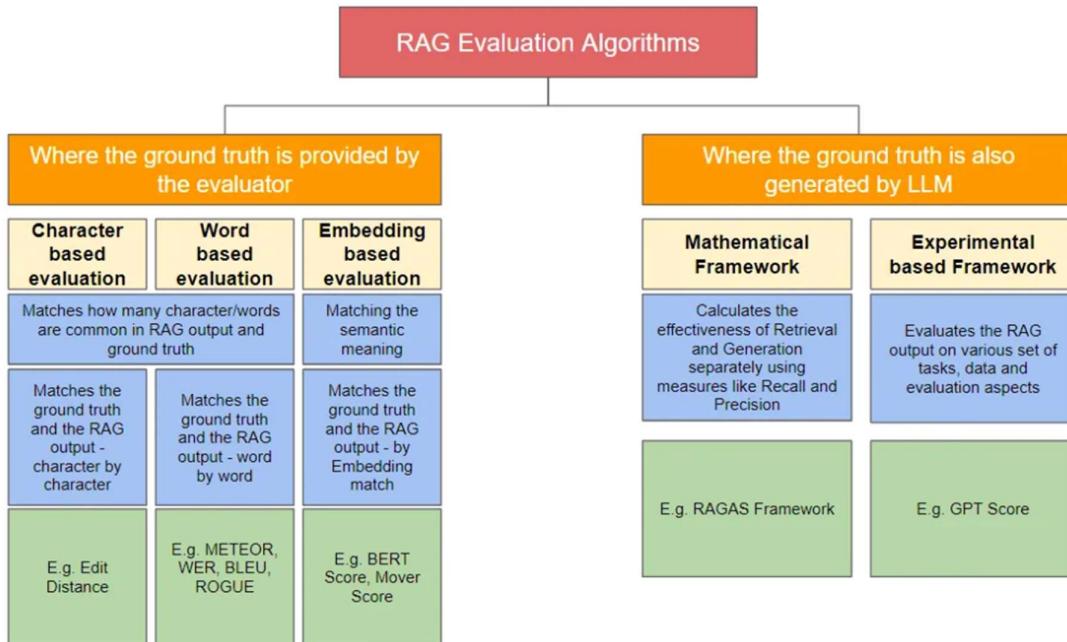
## Andrew Ng Explores The Rise Of AI Agents And Agentic Reasoning | BUILD 2024 Keynote, Snowflakes Inc



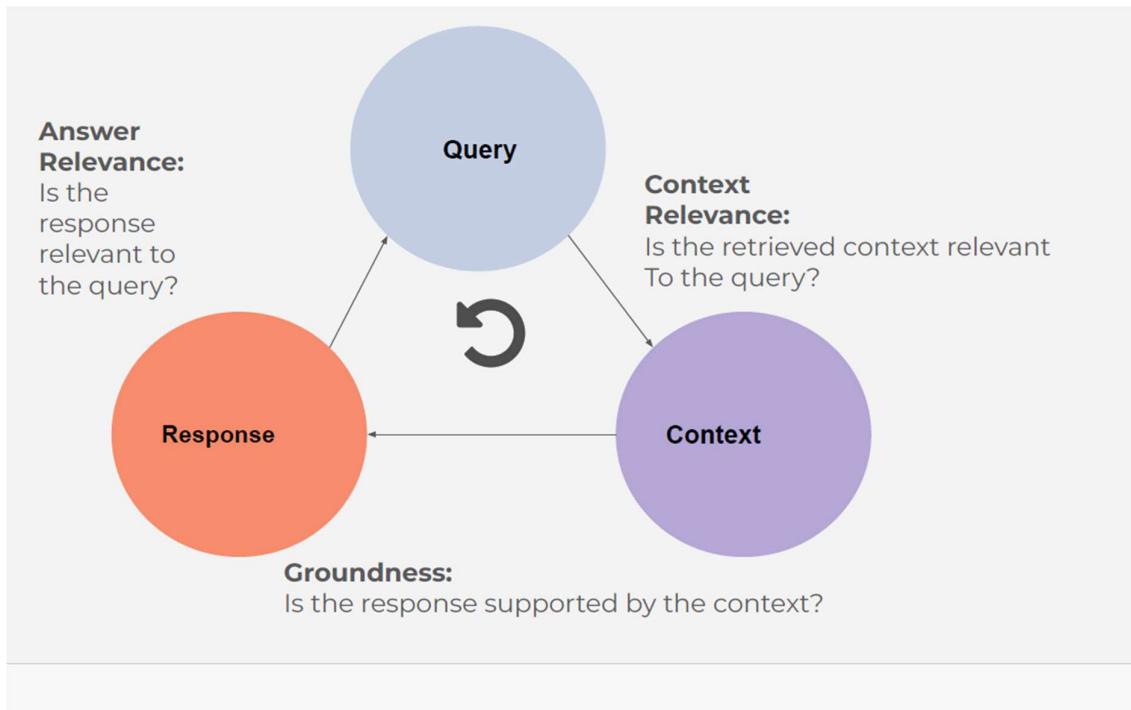
<https://www.youtube.com/watch?v=KrRD7r7y7NY>

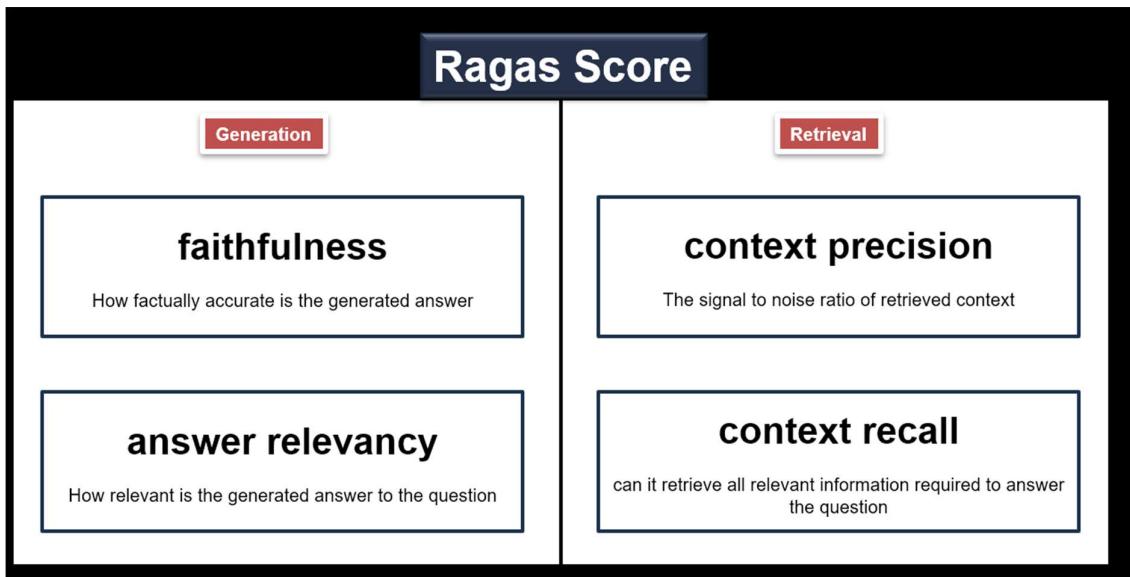
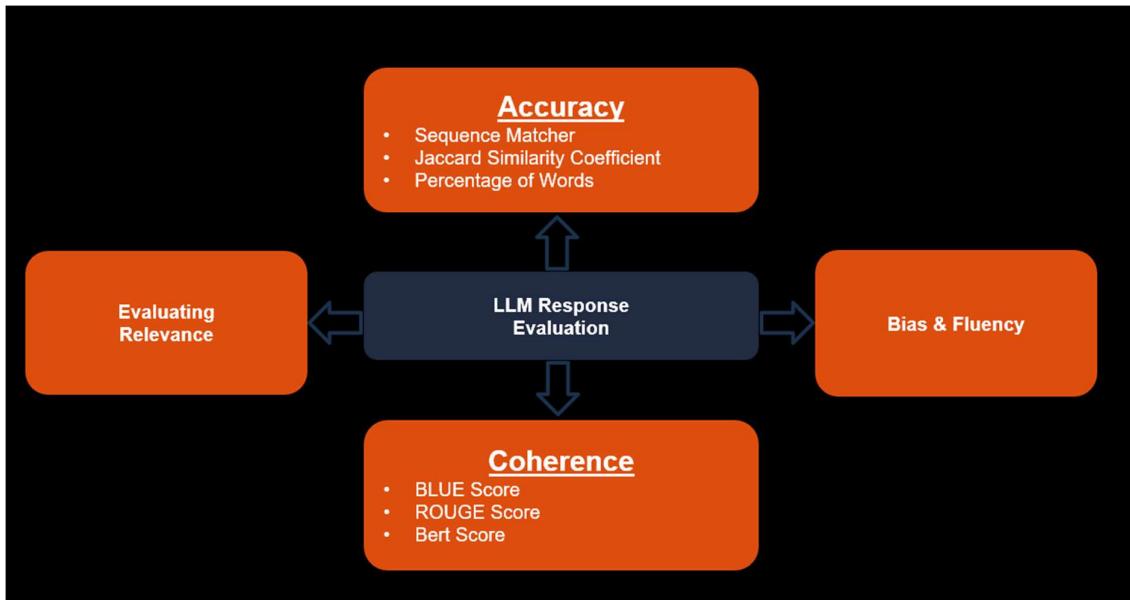
# Evaluating LLMs





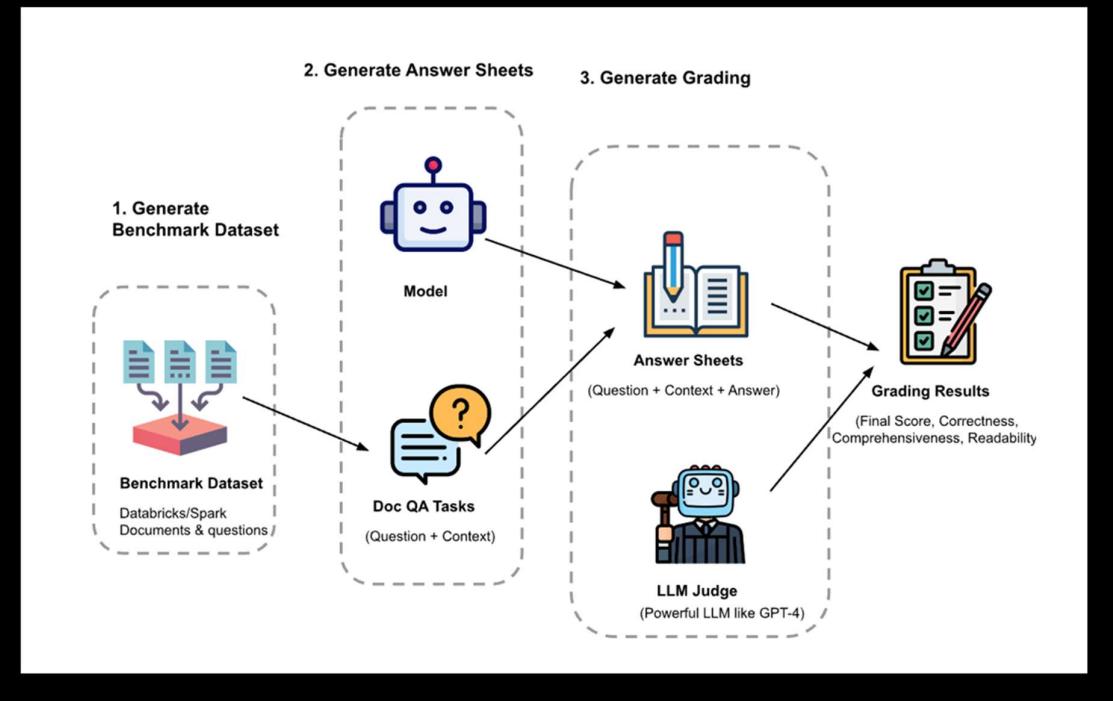
## The RAG Triad





LLM as a Judge

## LLM as a judge



## Role of Human Reviewers

### Role of Human reviewers in assessing response quality

- Human reviewers play a crucial role in evaluating how well the responses from LLM align with the intended goals and expectations
- It makes sure that LLM is not only informative but also produces human-like text
- Researchers assess LLM for empathetic responses and effective addressing of user concerns

# Platforms And Tools for Gen AI App Development

Langchain

Langflow

Llamaindex

LmStudio

Replit

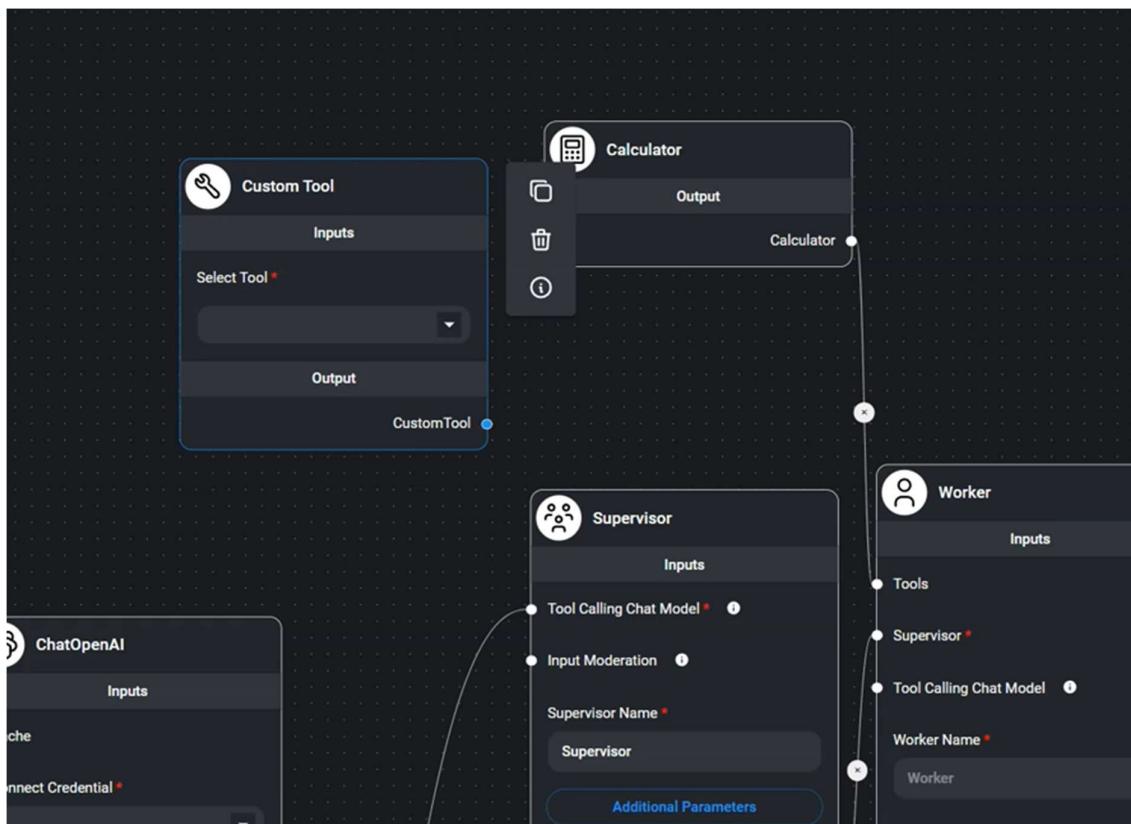
Anything LLM

Make.com

Firecrawl

Autotrain

Flowise



LlamaCloud Default Org / Default

**TOOLS**

- Parse
- Extraction (beta)

**RESOURCES**

- Settings
- API Key
- Documentation
- Support

**PARTNER INTEGRATIONS**

- LlamaTrace with Arize

**YOUR INFO**

- Free Plan

0 / 1000 pages per day

## Extraction (beta)

Infer the underlying schema in your document(s) to convert unstructured docs to structured data.

Extraction is an experimental feature that we're actively working to improve. Please, report any issues to our [Github](#).

Name ↑↓	Last Updated ↑↓
No results.	

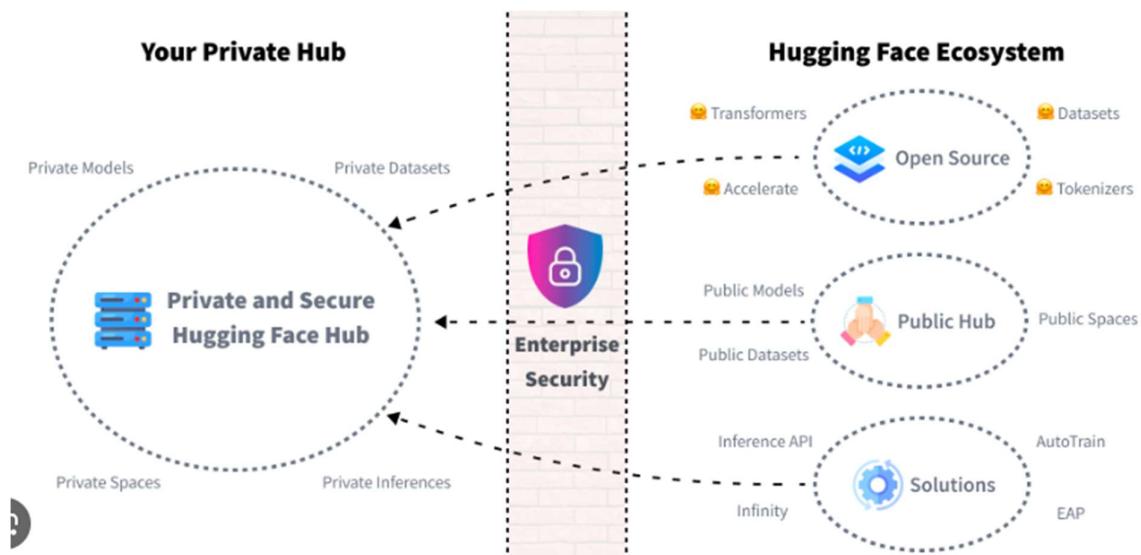
Huggingface

<https://huggingface.co>

HF cookbooks:

[https://huggingface.co/learn/cookbook/mlflow\\_ray\\_serve](https://huggingface.co/learn/cookbook/mlflow_ray_serve)

HF Hub:



LLM Finetuning:

<https://github.com/ashishpatel26/LLM-Finetuning>

Kore.ai

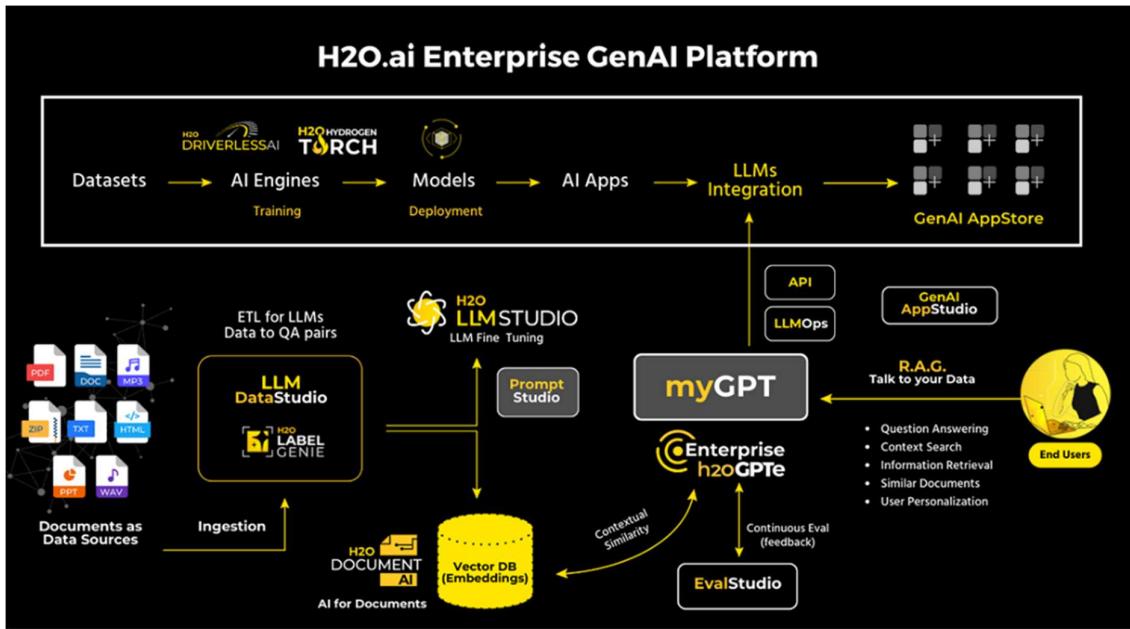
<https://www.youtube.com/watch?v=p06tMqyohBI>

<https://kore.ai>



H2O.AI

<https://h2o.ai/>



<https://dynamo.ai/>



Serper.dev for google search API

The World's Fastest & Cheapest Google Search API

Experience unparalleled speed with our industry-leading SERP API, delivering lightning-fast Google search results in 1-2 seconds, at an unbeatable price.

Get 2,500 free queries

No credit card required

## Implementation Challenges

- 1) Lack of good training data

Mitigation: Synthetic data – use LLMs to generate data

- 2) Quality of output:

<b>Accuracy</b> May occasionally miss critical details or emphasize less important information, affecting the accuracy of the output	<b>Bias and fairness</b> Inherits potential biases in training data	<b>Lack of deep understanding and coherence</b> LLMs do not understand text like humans which can sometimes result in incoherent or logically flawed responses
<b>Consistency and reliability</b> Can sometimes produce inconsistent or unreliable reasoning outputs, particularly if the prompt is ambiguous or if the task requires multi-step reasoning	<b>Errors / Hallucinations</b> Incorrect information presented as factually correct (misleading)	

## Problems

- 1 Controllability of AI output
- 2 Accuracy and reliability
- 3 IP protection
- 4 Latency
- 5 Governance challenges
- 6 Costs
- 7 Privacy concerns

Mitigation: Prompt Engineering, RAG, Fine Tuning

## Hallucination



Hallucinations occur when AI model generates incorrect or misleading information but presents it as if it were a fact.

**Example:** LLM designed to generate summaries of news articles may produce a summary that includes details not present in the original article, or even fabricates information entirely.

### Reasons

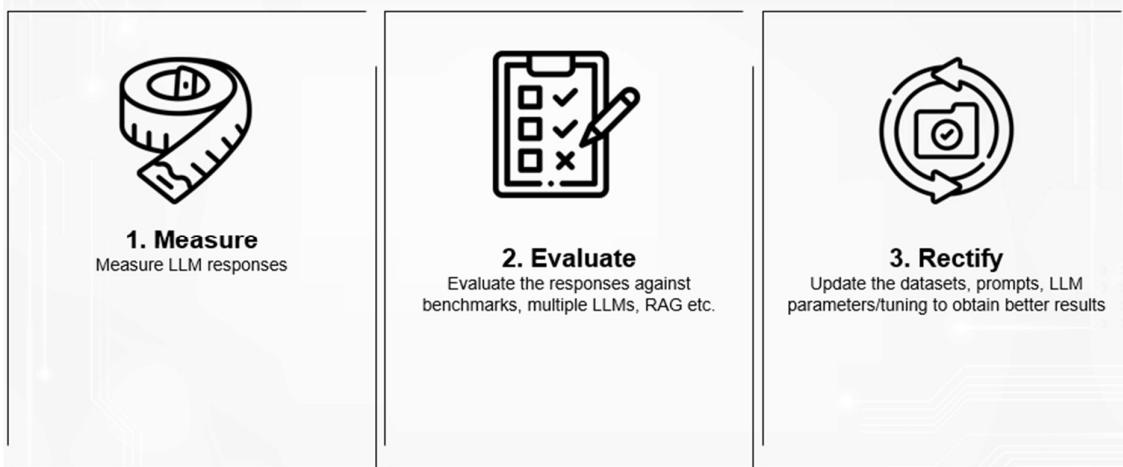
Flawed training data – Inaccurate, Incomplete.

Overfitting to training data

Extrapolation beyond training.

Insufficient context or context switching

Ambiguous prompt



[AI, Investment Decisions, and Inequality by Alex G. Kim, David S. Kim, Maximilian Muhn, Valeri V. Nikolaev , Eric C. So :: SSRN](#)

## Costing

### Enterprise Gen AI Bill Of Materials

#### LLM

- Open source or closed source?
- Pre-trained LLM or customized?
- As-an API service from cloud?
- As managed service?
- deploy and manage on our own
  - on cloud?
  - on-prem?

#### Customizations

- RAG
- Model Fine Tuning (PEFT/LoRA/QLoRA, DPO, PPO, RLHF)
- Agents

Recent costing (check latest from vendors)

Model	Quality	Resolution	Price	
DALL-E 3	Standard	1024×1024	\$0.040 / image	
	Standard	1024×1792, 1792×1024	\$0.080 / image	
DALL-E 3	HD	1024×1024	\$0.080 / image	
	HD	1024×1792, 1792×1024	\$0.120 / image	
DALL-E 2		1024×1024	\$0.020 / image	
		512×512	\$0.018 / image	
		256×256	\$0.016 / image	

GPT-4o mini is our most cost-efficient small model that's smarter and more efficient than GPT-3.5 Turbo, and has vision capabilities. The model has 128K context tokens and an October 2023 knowledge cutoff.

[Learn about GPT-4o mini ↗](#)

Model	Pricing	Pricing with Batch API*
gpt-4o-mini	0,150 \$ / 1M input tokens	0,075 \$ / 1M input tokens
	0,600 \$ / 1M output tokens	0,300 \$ / 1M output tokens
gpt-4o-mini-2024-07-18	0,150 \$ / 1M input tokens	0,075 \$ / 1M input tokens
	0,600 \$ / 1M output tokens	0,300 \$ / 1M output tokens

- Use case : Text generation, Code generation, image generation(Media, marketing, design), voice synthesis, embeddings, etc.
- Evaluate data- What type of datasets your use case requires? (General purpose or Domain specific)?
- Performance - Quality of the response and supported latency
- Context window size
- Fine tuning & customization support
- Required Modality support- Single, multiple
- Type of model- General purpose model (Pre trained model), instruction tuned for your domain specific tasks & RL tuned models
- Hosting type - Self hosted or fully managed with model as a service
- Training Data – Type of data used to train the model- internet data, code
- License type- Open source, Open model or Proprietary
- Licensing conditions
- Data Privacy
- Ethical & Responsible AI considerations
- Language support – Most models are trained on English
- Cost- Infrastructure, software requirement to host the model
- Pricing- Hosted models are typically priced based on input tokens and completions.

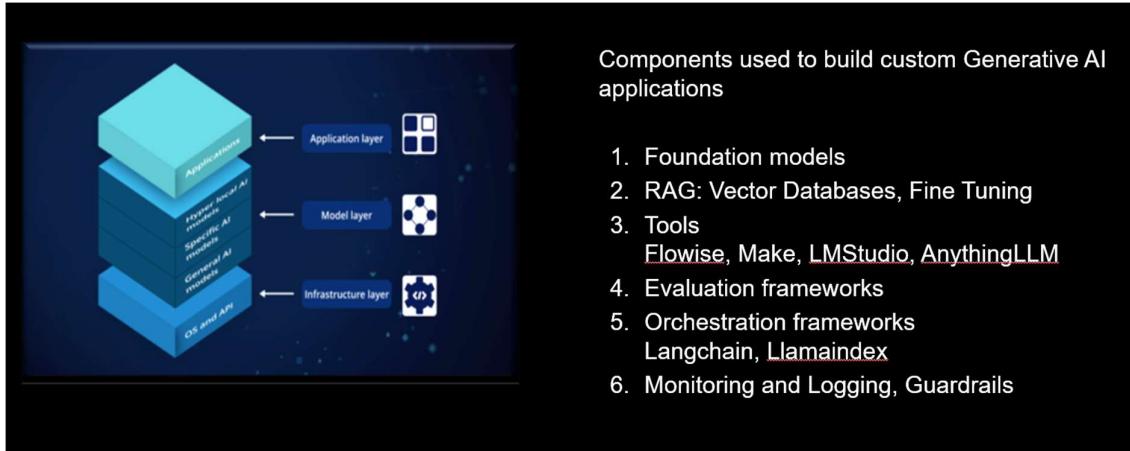
## Considerations for using LLMs as API service Vs Hosted Vs on-prem

- **Data privacy and security**
- **Time-to-market**
- **Usage (Application characteristics)**
- **Cost**
- **Skills**
- **Performance (Speed and accuracy/precision)**
- **Intellectual property ownership**
- **Available data / volume of proprietary or RAG data**

<https://medium.com/@gopikwork/comprehensive-guide-for-model-selection-and-evaluation-fcd7fe299a50>

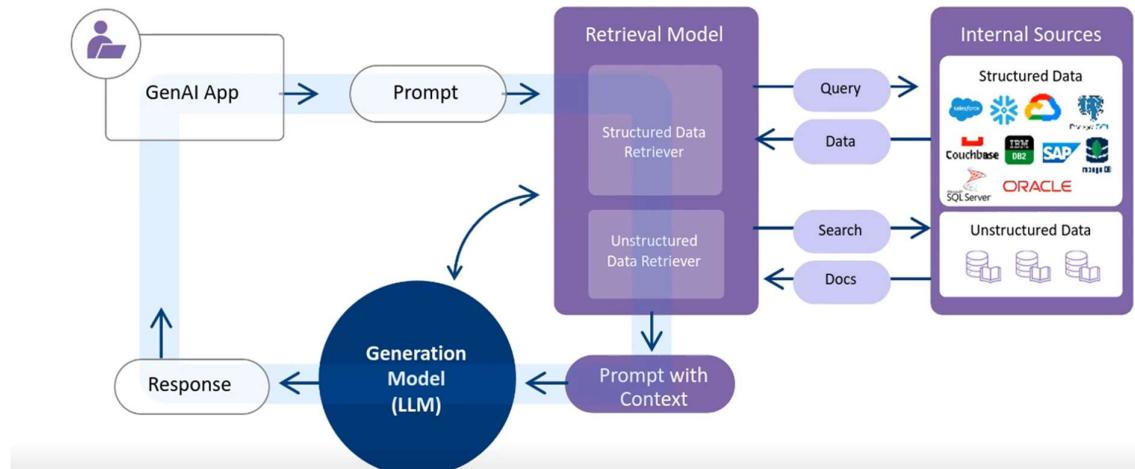
<https://medium.com/emalpha/the-economics-of-large-language-models-2671985b621c>

# Enterprise Considerations



k2view

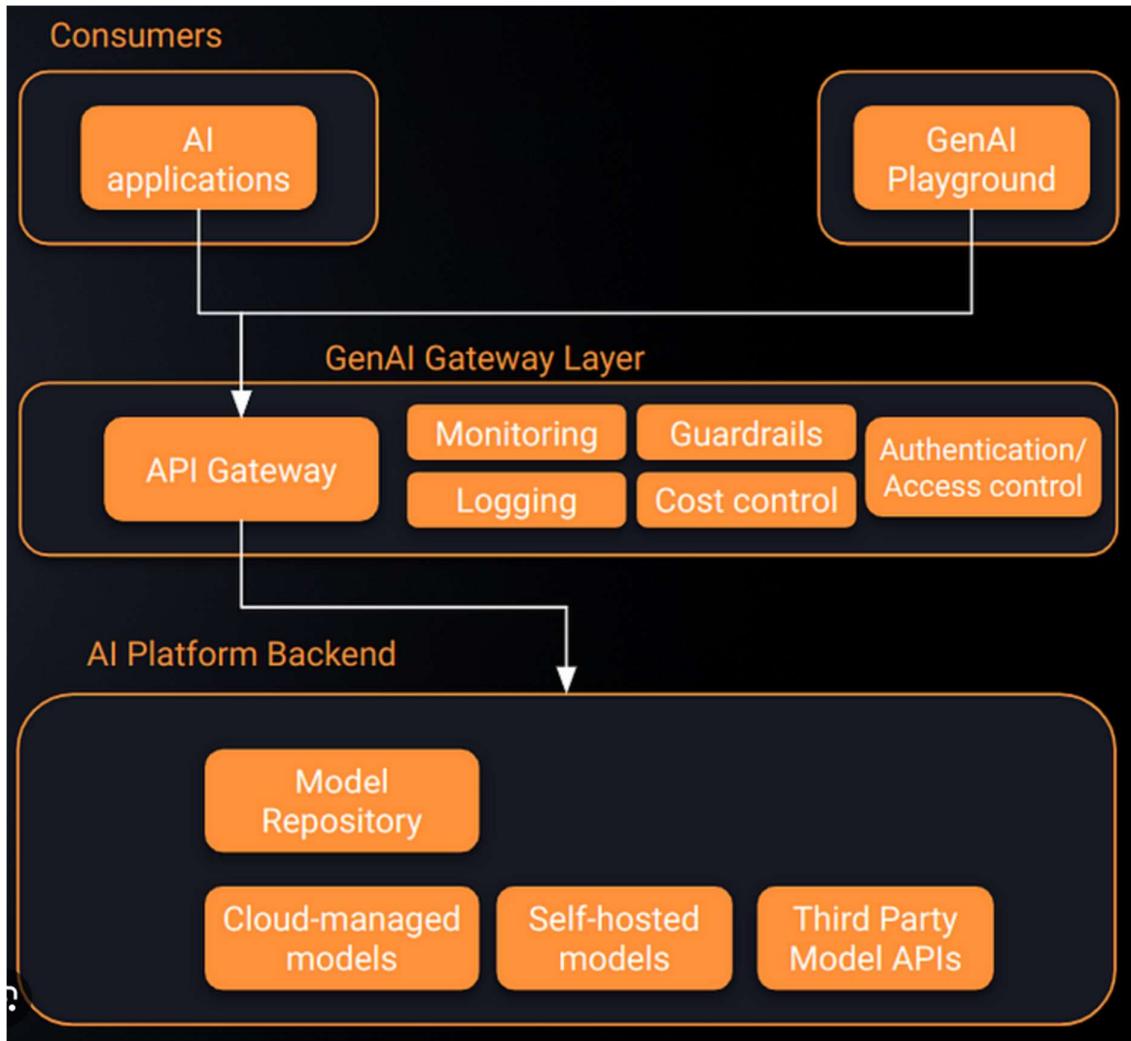
## RAG turns generic LLMs into business-specific LLMs



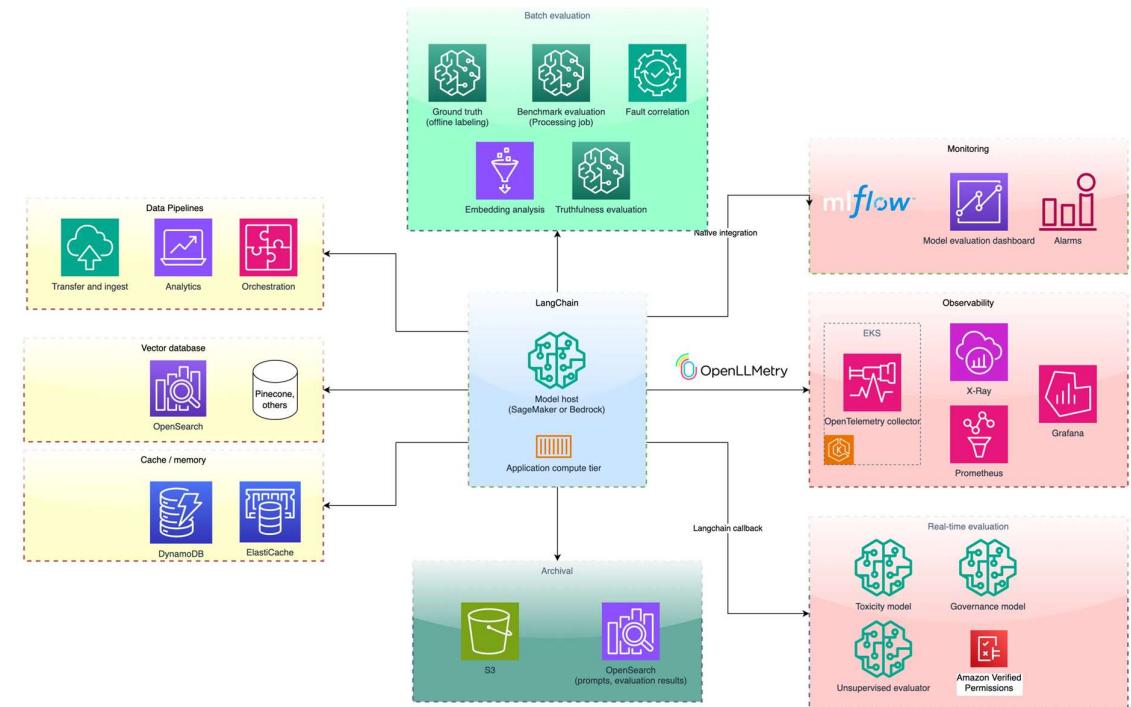
## Productionalizing LLM Applications

[What We've Learned From A Year of Building with LLMs – Applied LLMs](#)

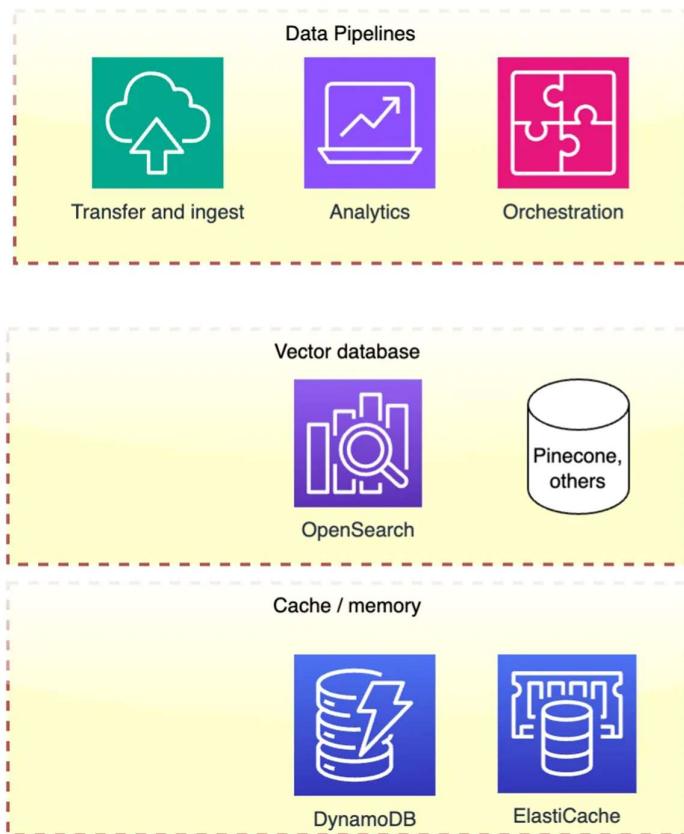
<https://community.aws/content/2i7BzRVM4ppZSGaZKoEHdpNwick/observability-monitoring-and-layered-guardrails-for-genai>

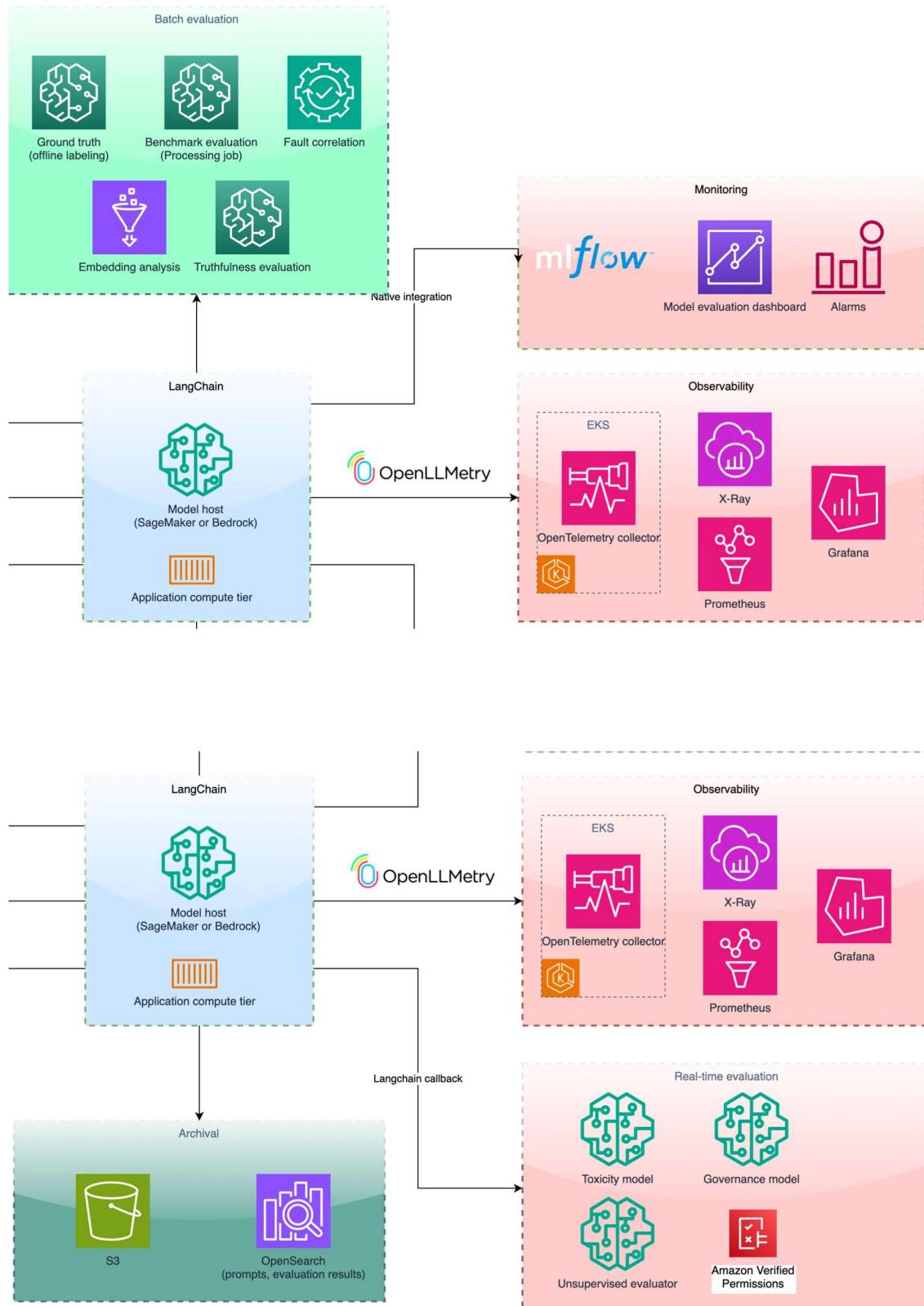


<https://www.ml6.eu/blogpost/why-you-need-a-genai-gateway>

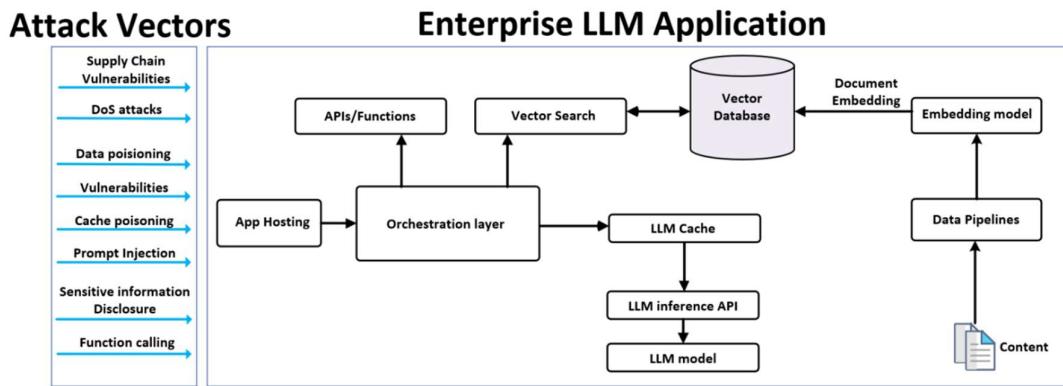


<https://community.aws/content/2i7BzRVM4ppZSGaZKoEHdpNwick/observability-monitoring-and-layered-guardrails-for-genai>





## Securing LLM Applications



<https://msandbu.org/how-do-you-secure-an-genai-application-or-service/>

## Guardrails

**Guardrails**

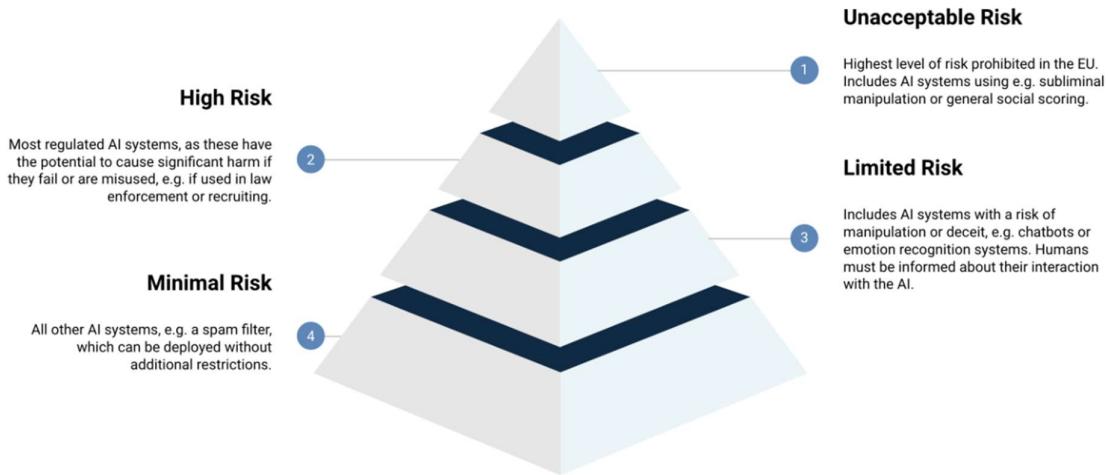
• Mechanisms to ensure AI systems operate safely, ethically, and in compliance with regulations

**Examples**

- **LLMGuard:** Provides tools to ensure the outputs of large language models are safe and aligned with desired guidelines.
- **NeMo Guardrails:** NVIDIA's toolkit for ensuring AI models produce outputs that meet specific safety and ethical standards.
- **AI Fairness 360:** An extensible open-source toolkit that can help users examine, report, and mitigate discrimination and bias in machine learning models.

## Risks, Governance And Compliance

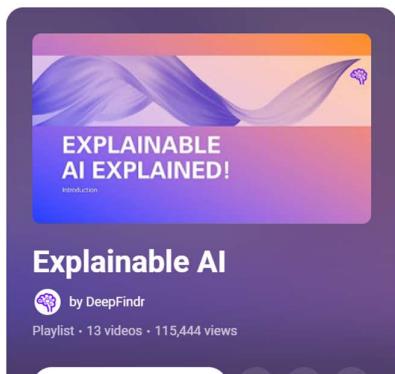
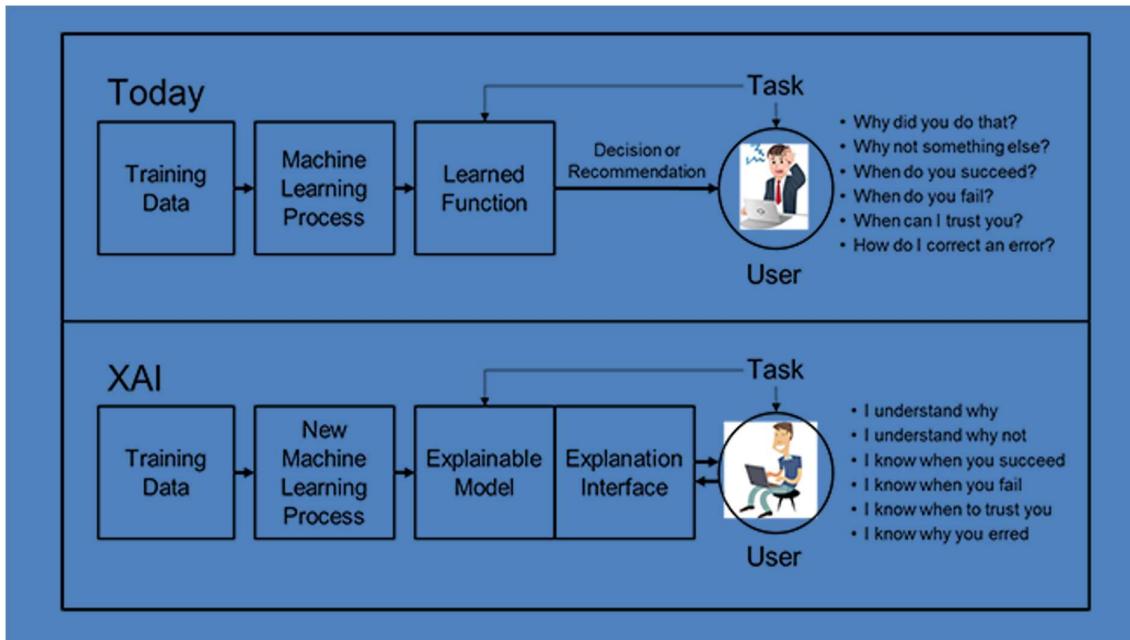
EU AI ACT



## ISO 42001



## Explainable AI



1 **Explainable AI explained! | #1 Introduction**  
DeepFindr • 72K views • 3 years ago

2 **Explainable AI explained! | #2 By-design interpretable models**  
DeepFindr • 36K views • 3 years ago

3 **Explainable AI explained! | #3 LIME**  
DeepFindr • 64K views • 3 years ago

<https://www.youtube.com/playlist?list=PLV8yxwGOxvvovp-j6ztxhF3QcKXT6vORU>

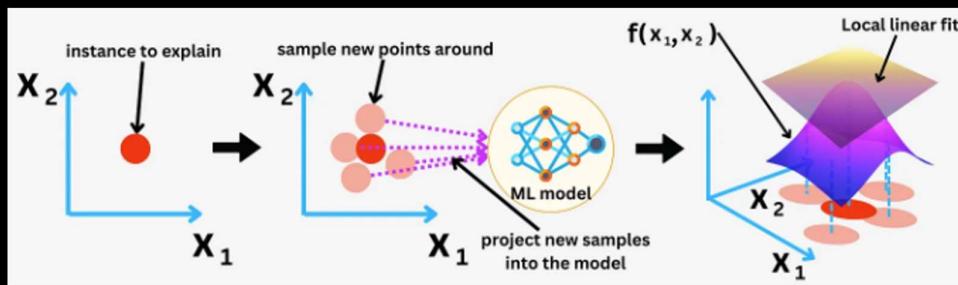
## Explainable AI (XAI) frameworks – LIME and SHAP

### LIME: Local Interpretable Model-agnostic Explanations

- Perturb input data
- Model predictions on perturbed data
- Train a simple interpretable model
- Local explanation

### SHAP (Shapley Additive exPlanations)

- Shapley Values (Game theory concept)
- Evaluate feature contributions
- Build a SHAP explanation



LIME

## Responsible AI Framework



# Gen AI Use cases Across Job Functions

## Key capabilities enabled by Generative AI

### Vast, accessible knowledgebase – democratization of AI

Pre-trained models trained on vast knowledge base, accessible via a context aware chat interface

### Natural Language Processing (NLP)

Transcribe, translate, summarize, classify, question & answer, conversational AI, chatbots any text

### Computer vision

Segmentation, Object detection, Scene/image understanding and processing

### Multimodal

Process from multiple modalities (text, image, audio, video etc.), generate multi-modal outputs

### Generative capabilities

Intelligently generate summaries, reports, presentations, stories, blogs, content, documents, code, designs, images, audio, video

### Decision making - Reasoning and Action

LLMs can reason and create an execution plan, invoke right tools with Agents and Agentic workflows

### Intelligence from unstructured data

Automatically organizes and bring out patterns and structure from unstructured data such as text, audio, video, images. Interpret and understand text, images, audio, video

### Scenario simulation

For example, factory floor, transportation routes

### Extensible and adaptable

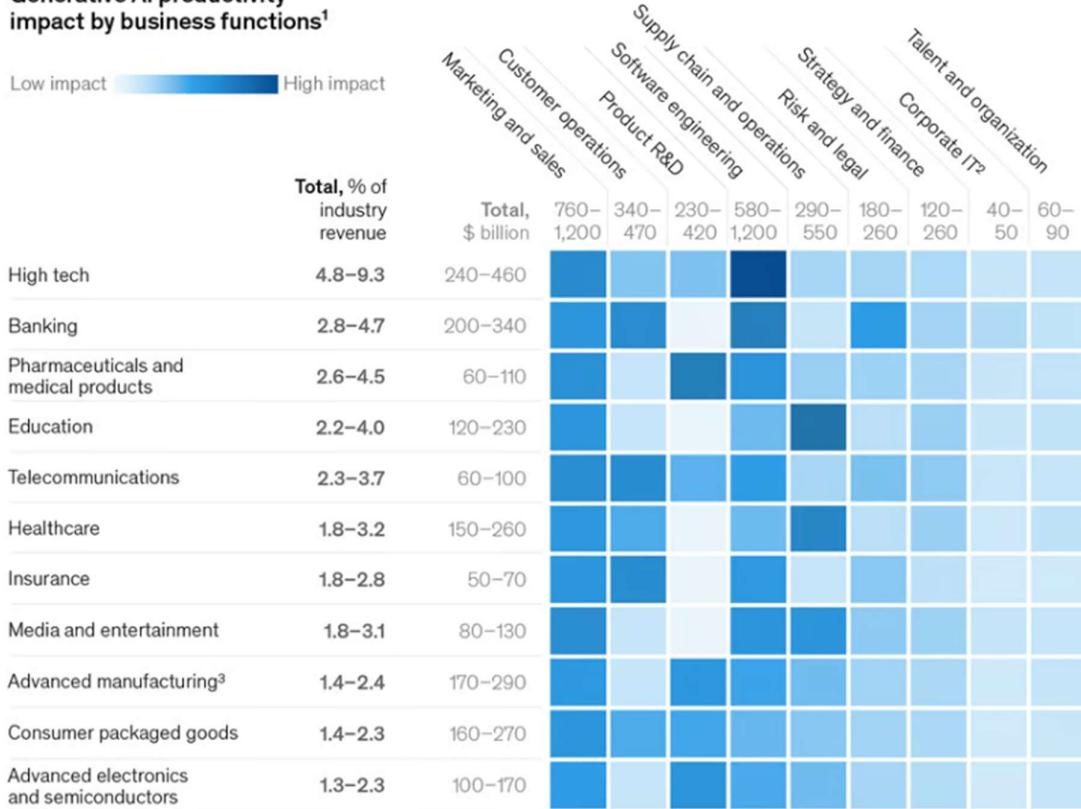
Augment knowledgebase (RAG)

Fine Tuning (Models can be fine tuned for specific purposes)

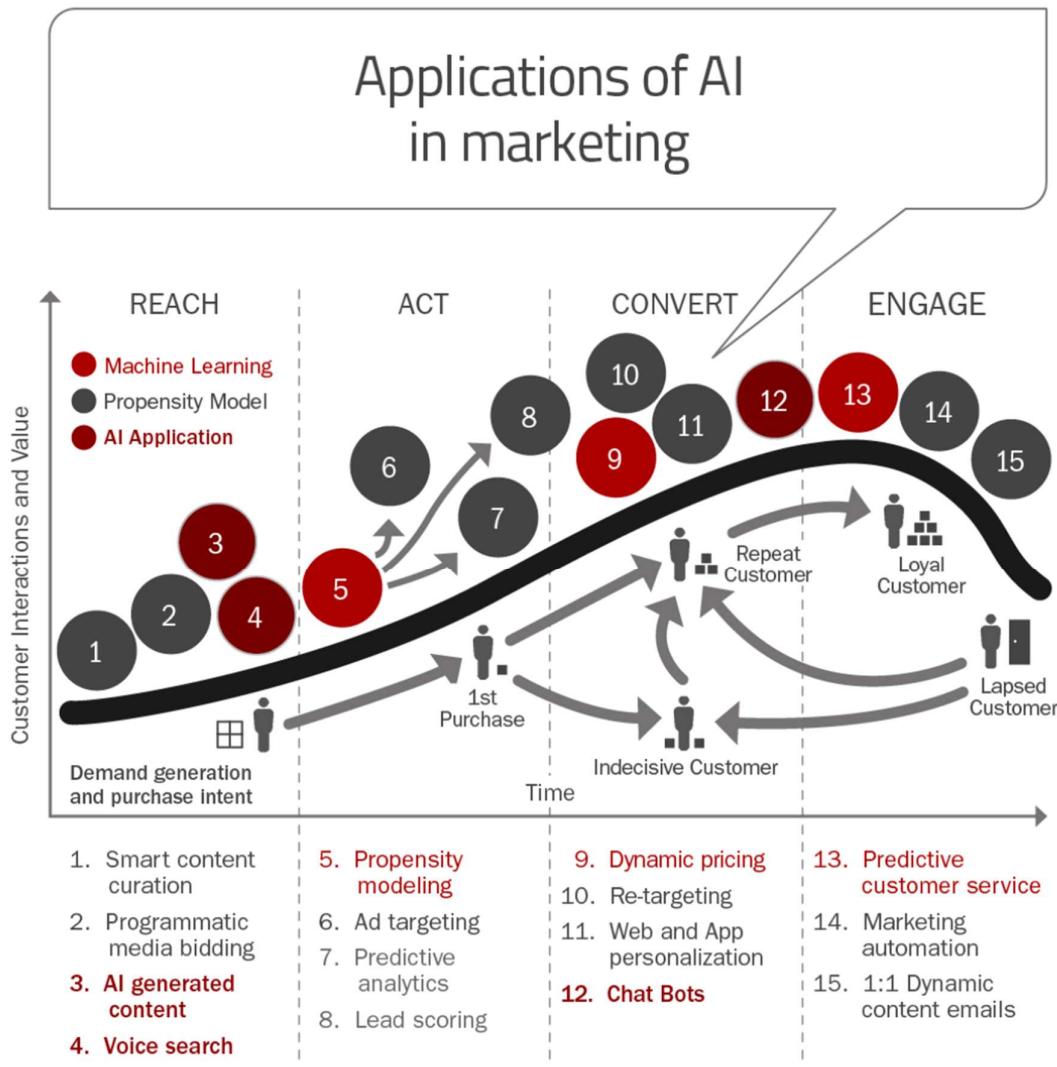
Automation, workflows, real-time data fetching (Agents, ReAct)

Gen AI – Impact by business functions, across industries

## Generative AI productivity impact by business functions<sup>1</sup>



## Marketing and Sales



SOURCE: Smart Insights © March 2018 The Financial Brand

## Sample Areas of AI's Impact on Sales



Source: Gartner  
© 2023 Gartner, Inc. and/or its affiliates. All rights reserved. 2473717

Gartner®

## HubSpot AI in Sales 2024 trends report

[https://www.hubspot.com/hubfs/Content%20Offers/HubSpot%20AI%20Sales%20Trends%20Report%202024.pdf?hubs\\_signup-url=offers.hubspot.com%2Fai-sales&hubs\\_signup-cta=Submit&hubs\\_offer=offers.hubspot.com%2Fai-sales&\\_gl=1\\*15wcawt\\*\\_gcl\\_au\\*MjExMDM1NjAzOS4xNzM0MzY1MjUx&\\_ga=2.267153549.261305445.1734406649-259091869.1734406649](https://www.hubspot.com/hubfs/Content%20Offers/HubSpot%20AI%20Sales%20Trends%20Report%202024.pdf?hubs_signup-url=offers.hubspot.com%2Fai-sales&hubs_signup-cta=Submit&hubs_offer=offers.hubspot.com%2Fai-sales&_gl=1*15wcawt*_gcl_au*MjExMDM1NjAzOS4xNzM0MzY1MjUx&_ga=2.267153549.261305445.1734406649-259091869.1734406649)

### AI Sales Tools list:

#### Hubspot: 10 best AI tools in 2024:

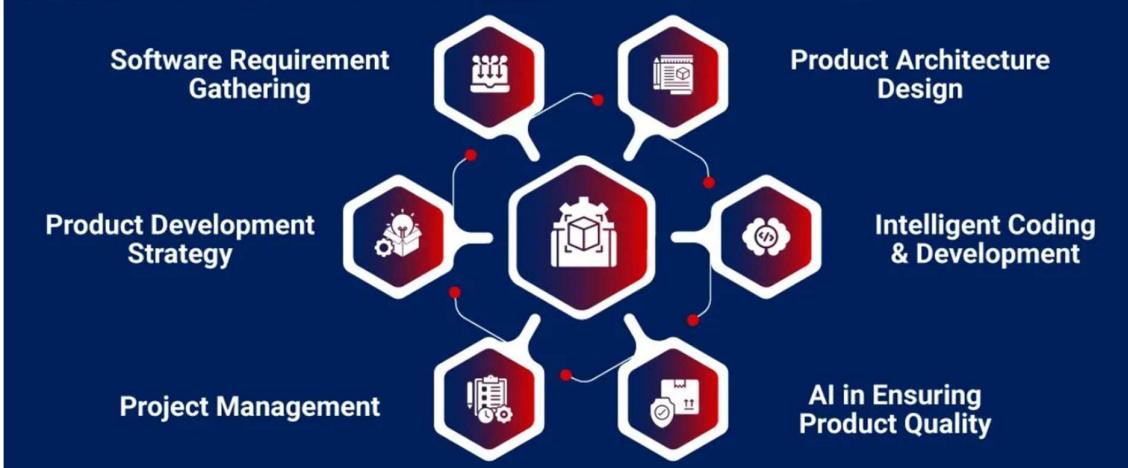
<https://blog.hubspot.com/sales/ai-sales-tools>

A long list of AI tools for Sales:

<https://www.coldiq.com/ai-sales-tools>

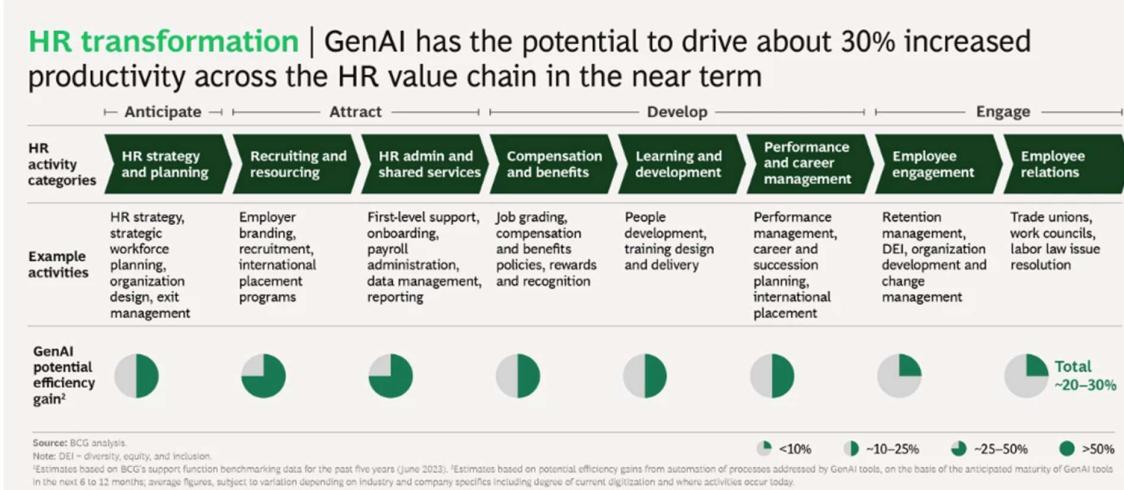
Software Development and Quality Assurance

## AI Across Stages of Product Development



<https://www.rishabhsoft.com/blog/ai-in-product-development>

## Human Resources



## AI in HR by 2025

- Smart hiring decisions
- AI-powered resume screening
- Video interview analysis
- Improved candidate experience
- AI for compliance management
- AI for data management
- Personalized training programs
- NLP for multinational training
- Enhanced training insights
- Real-time performance reviews
- SMART goal-tracking
- Sentiment analysis
- Automated recognition

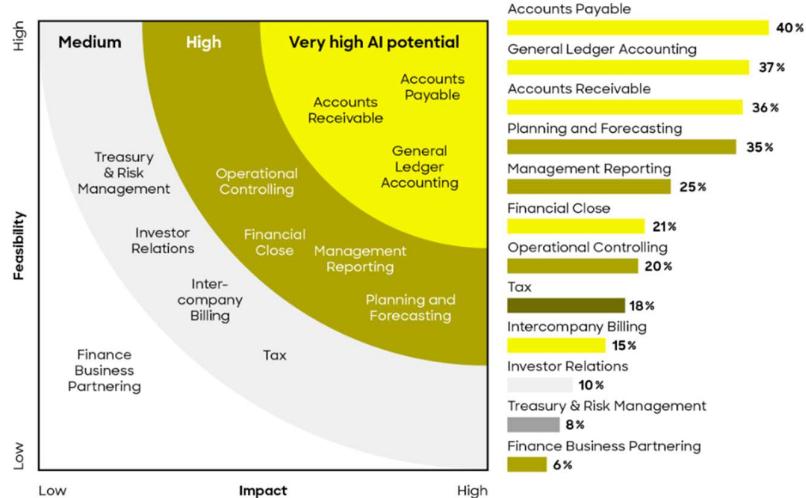


## Accounting and Finance

### Top use cases

Accounting processes, such as accounts receivable and accounts payable, have the potential for up to 40% AI efficiency gains<sup>1</sup>

- Accounting
- Controlling
- Treasury & Risk Management
- Tax
- Investor Relations



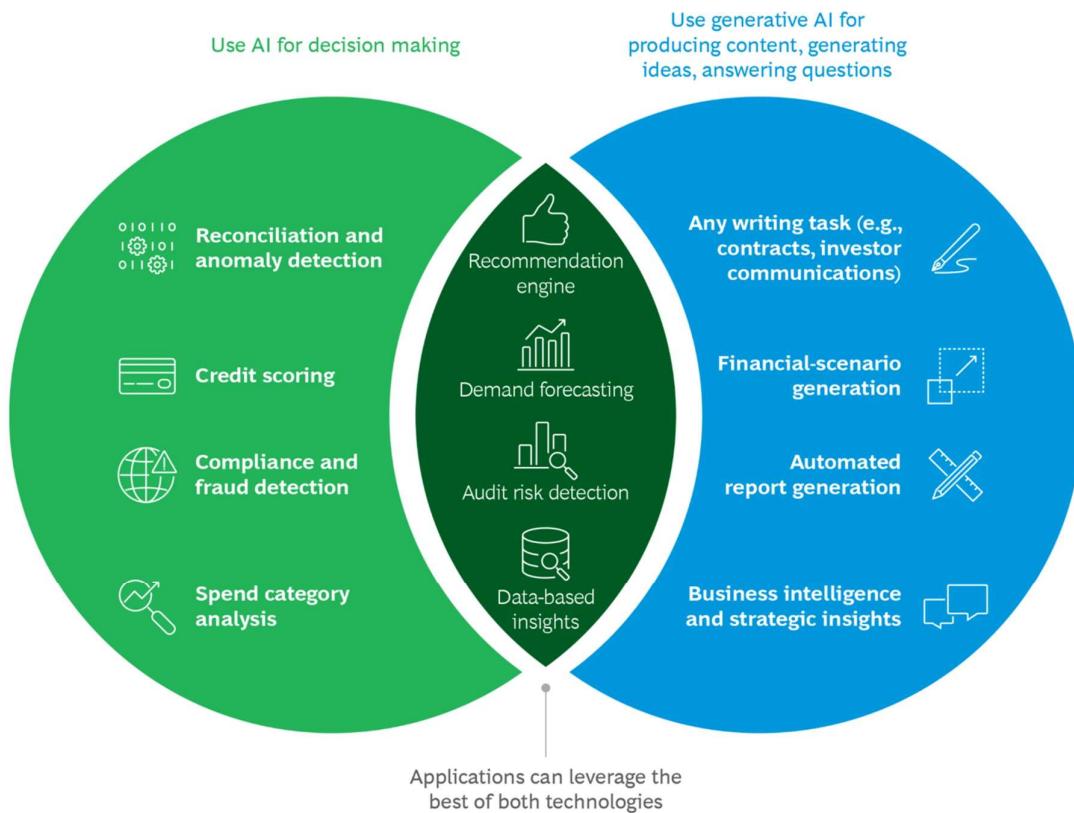
1 Efficiency gains (time, cost) in % estimated based on Roland Berger analysis of >100 companies already using AI  
Source Gartner, The Hackett Group, Roland Berger

Roland Berger

<https://www.rolandberger.com/en/Insights/Publications/Mastering-AI-in-the-finance-function.html>

## Gen AI Use cases Across Industries

## Finance and BFSI



Source: BCG analysis.

<https://www.bcg.com/publications/2023/generative-ai-in-finance-and-accounting>

Moody's copilot:

<https://www.youtube.com/watch?v=7O8jtu3bpS8>

Morgan Stanley uses AI evals to shape the future of financial services

<https://openai.com/index/morgan-stanley/>



Exploring the Impact of AI in Indian Banking: Classical vs. Generative AI

AI Dialogues | Combination Of Classical AI & Gen AI To Help Collate Data From Public Infrastructure Records For Efficient Credit Administration

7:51 / 8:46

Subscribed 3.17M subscribers

Like 19 Share Download Thanks Save ...

All From CNBC-TV18 AI Stock market

<https://www.youtube.com/watch?v=BO1n3NdUWPs>

Blog > Reports And Research

## Generative AI in Finance: Real-World Examples of LLMs in Banking in 2023

August 04, 2023 5 min read Kate Kopyl

<https://tovie.ai/blog/generative-ai-in-finance-real-world-examples-of-llms-in-banking-in-2023>

<https://www.mckinsey.com/industries/financial-services/our-insights/capturing-the-full-value-of-generative-ai-in-banking>

## INSURANCE

**A** Your proprietary data could be in any type and storage facility e.g.:

- File System (Cloud/On-prem)
- DW/OL/ Data Lakehouse
- NoSQL Databases

1 And of any type...  
2 Structured Data  
Unstructured Data

**B** Choice of LLMs & Providers

- Hugging Face
- Qwen
- Codex
- GPT
- QwenP
- QwenP-4
- QwenP-7
- QwenP-12
- QwenP-16
- QwenP-24
- QwenP-32
- QwenP-48
- QwenP-64
- QwenP-96
- QwenP-128
- QwenP-160
- QwenP-256
- QwenP-384
- QwenP-512
- QwenP-768
- QwenP-1024
- QwenP-1280
- QwenP-1536
- QwenP-1920
- QwenP-2448
- QwenP-3072
- QwenP-3840
- QwenP-4864
- QwenP-6400
- QwenP-7680
- QwenP-9600
- QwenP-11520
- QwenP-14080
- QwenP-17280
- QwenP-20480
- QwenP-24192
- QwenP-28160
- QwenP-32256
- QwenP-38400
- QwenP-44640
- QwenP-51200
- QwenP-61440
- QwenP-71680
- QwenP-81920
- QwenP-92160
- QwenP-102400
- QwenP-113760
- QwenP-124480
- QwenP-135200
- QwenP-145920
- QwenP-156640
- QwenP-167360
- QwenP-178080
- QwenP-188800
- QwenP-200000
- QwenP-211760
- QwenP-222400
- QwenP-233120
- QwenP-243840
- QwenP-254560
- QwenP-265280
- QwenP-275920
- QwenP-286640
- QwenP-297360
- QwenP-308080
- QwenP-318800
- QwenP-329520
- QwenP-340240
- QwenP-350960
- QwenP-361680
- QwenP-372400
- QwenP-383120
- QwenP-393840
- QwenP-404560
- QwenP-415280
- QwenP-426000
- QwenP-436720
- QwenP-447440
- QwenP-458160
- QwenP-468880
- QwenP-479600
- QwenP-480320
- QwenP-491040
- QwenP-501760
- QwenP-512480
- QwenP-523200
- QwenP-533920
- QwenP-544640
- QwenP-555360
- QwenP-566080
- QwenP-576800
- QwenP-587520
- QwenP-598240
- QwenP-609000
- QwenP-619760
- QwenP-630520
- QwenP-641280
- QwenP-652040
- QwenP-662800
- QwenP-673560
- QwenP-684320
- QwenP-695080
- QwenP-705840
- QwenP-716600
- QwenP-727360
- QwenP-738120
- QwenP-748880
- QwenP-759640
- QwenP-770400
- QwenP-781160
- QwenP-791920
- QwenP-802680
- QwenP-813440
- QwenP-824200
- QwenP-834960
- QwenP-845720
- QwenP-856480
- QwenP-867240
- QwenP-878000
- QwenP-888760
- QwenP-899520
- QwenP-910280
- QwenP-921040
- QwenP-931800
- QwenP-942560
- QwenP-953320
- QwenP-964080
- QwenP-974840
- QwenP-985600
- QwenP-996360
- QwenP-1007120
- QwenP-1017880
- QwenP-1028640
- QwenP-1039400
- QwenP-1041000
- QwenP-1051760
- QwenP-1062520
- QwenP-1073280
- QwenP-1084040
- QwenP-1094800
- QwenP-1105560
- QwenP-1116320
- QwenP-1127080
- QwenP-1137840
- QwenP-1148600
- QwenP-1159360
- QwenP-1170120
- QwenP-1180880
- QwenP-1191640
- QwenP-1202400
- QwenP-1213160
- QwenP-1223920
- QwenP-1234680
- QwenP-1245440
- QwenP-1256200
- QwenP-1266960
- QwenP-1277720
- QwenP-1288480
- QwenP-1299240
- QwenP-1310000
- QwenP-1320760
- QwenP-1331520
- QwenP-1342280
- QwenP-1353040
- QwenP-1363800
- QwenP-1374560
- QwenP-1385320
- QwenP-1396080
- QwenP-1406840
- QwenP-1417600
- QwenP-1428360
- QwenP-1439120
- QwenP-1450000
- QwenP-1460760
- QwenP-1471520
- QwenP-1482280
- QwenP-1493040
- QwenP-1503800
- QwenP-1514560
- QwenP-1525320
- QwenP-1536080
- QwenP-1546840
- QwenP-1557600
- QwenP-1568360
- QwenP-1579120
- QwenP-1589880
- QwenP-1590640
- QwenP-1601400
- QwenP-1612160
- QwenP-1622920
- QwenP-1633680
- QwenP-1644440
- QwenP-1655200
- QwenP-1665960
- QwenP-1676720
- QwenP-1687480
- QwenP-1698240
- QwenP-1709000
- QwenP-1719760
- QwenP-1730520
- QwenP-1741280
- QwenP-1752040
- QwenP-1762800
- QwenP-1773560
- QwenP-1784320
- QwenP-1795080
- QwenP-1805840
- QwenP-1816600
- QwenP-1827360
- QwenP-1838120
- QwenP-1848880
- QwenP-1859640
- QwenP-1870400
- QwenP-1881160
- QwenP-1891920
- QwenP-1902680
- QwenP-1913440
- QwenP-1924200
- QwenP-1935000
- QwenP-1945760
- QwenP-1956520
- QwenP-1967280
- QwenP-1978040
- QwenP-1988800
- QwenP-1999560
- QwenP-2010320
- QwenP-2021080
- QwenP-2031840
- QwenP-2042600
- QwenP-2053360
- QwenP-2064120
- QwenP-2074880
- QwenP-2085640
- QwenP-2096400
- QwenP-2107160
- QwenP-2117920
- QwenP-2128680
- QwenP-2139440
- QwenP-2150200
- QwenP-2160960
- QwenP-2171720
- QwenP-2182480
- QwenP-2193240
- QwenP-2204000
- QwenP-2214760
- QwenP-2225520
- QwenP-2236280
- QwenP-2247040
- QwenP-2257800
- QwenP-2268560
- QwenP-2279320
- QwenP-2281080
- QwenP-2291840
- QwenP-2302600
- QwenP-2313360
- QwenP-2324120
- QwenP-2334880
- QwenP-2345640
- QwenP-2356400
- QwenP-2367160
- QwenP-2377920
- QwenP-2388680
- QwenP-2399440
- QwenP-2410200
- QwenP-2420960
- QwenP-2431720
- QwenP-2442480
- QwenP-2453240
- QwenP-2464000
- QwenP-2474760
- QwenP-2485520
- QwenP-2496280
- QwenP-2507040
- QwenP-2517800
- QwenP-2528560
- QwenP-2539320
- QwenP-2550000
- QwenP-2560760
- QwenP-2571520
- QwenP-2582280
- QwenP-2593040
- QwenP-2603800
- QwenP-2614560
- QwenP-2625320
- QwenP-2636080
- QwenP-2646840
- QwenP-2657600
- QwenP-2668360
- QwenP-2679120
- QwenP-2689880
- QwenP-2699640
- QwenP-2710400
- QwenP-2721160
- QwenP-2731920
- QwenP-2742680
- QwenP-2753440
- QwenP-2764200
- QwenP-2774960
- QwenP-2785720
- QwenP-2796480
- QwenP-2807240
- QwenP-2818000
- QwenP-2828760
- QwenP-2839520
- QwenP-2850280
- QwenP-2861040
- QwenP-2871800
- QwenP-2882560
- QwenP-2893320
- QwenP-2904080
- QwenP-2914840
- QwenP-2925600
- QwenP-2936360
- QwenP-2947120
- QwenP-2957880
- QwenP-2968640
- QwenP-2979400
- QwenP-2989160
- QwenP-2999920
- QwenP-3010680
- QwenP-3021440
- QwenP-3032200
- QwenP-3042960
- QwenP-3053720
- QwenP-3064480
- QwenP-3075240
- QwenP-3086000
- QwenP-3096760
- QwenP-3107520
- QwenP-3118280
- QwenP-3129040
- QwenP-3139800
- QwenP-3150560
- QwenP-3161320
- QwenP-3172080
- QwenP-3182840
- QwenP-3193600
- QwenP-3204360
- QwenP-3215120
- QwenP-3225880
- QwenP-3236640
- QwenP-3247400
- QwenP-3258160
- QwenP-3268920
- QwenP-3279680
- QwenP-3289440
- QwenP-3299200
- QwenP-3310960
- QwenP-3321720
- QwenP-3332480
- QwenP-3343240
- QwenP-3354000
- QwenP-3364760
- QwenP-3375520
- QwenP-3386280
- QwenP-3397040
- QwenP-3407800
- QwenP-3418560
- QwenP-3429320
- QwenP-3439080
- QwenP-3459840
- QwenP-3469600
- QwenP-3479360
- QwenP-3489120
- QwenP-3499880
- QwenP-3510640
- QwenP-3521400
- QwenP-3532160
- QwenP-3542920
- QwenP-3553680
- QwenP-3564440
- QwenP-3575200
- QwenP-3585960
- QwenP-3596720
- QwenP-3607480
- QwenP-3618240
- QwenP-3629000
- QwenP-3649760
- QwenP-3659520
- QwenP-3670280
- QwenP-3681040
- QwenP-3691800
- QwenP-3702560
- QwenP-3713320
- QwenP-3724080
- QwenP-3734840
- QwenP-3745600
- QwenP-3756360
- QwenP-3767120
- QwenP-3777880
- QwenP-3788640
- QwenP-3799400
- QwenP-3810160
- QwenP-3820920
- QwenP-3831680
- QwenP-3842440
- QwenP-3853200
- QwenP-3863960
- QwenP-3874720
- QwenP-3885480
- QwenP-3896240
- QwenP-3907000
- QwenP-3917760
- QwenP-3928520
- QwenP-3939280
- QwenP-3959040
- QwenP-3969800
- QwenP-3979560
- QwenP-3989320
- QwenP-3999080
- QwenP-4019840
- QwenP-4029600
- QwenP-4039360
- QwenP-4049120
- QwenP-4059880
- QwenP-4069640
- QwenP-4079400
- QwenP-4089160
- QwenP-4099920
- QwenP-4119680
- QwenP-4129440
- QwenP-4139200
- QwenP-4149960
- QwenP-4159720
- QwenP-4169480
- QwenP-4179240
- QwenP-4189000
- QwenP-4199760
- QwenP-4209520
- QwenP-4219280
- QwenP-4229040
- QwenP-4239800
- QwenP-4249560
- QwenP-4259320
- QwenP-4269080
- QwenP-4279840
- QwenP-4289600
- QwenP-4299360
- QwenP-4309120
- QwenP-4319880
- QwenP-4329640
- QwenP-4339400
- QwenP-4349160
- QwenP-4359920
- QwenP-4369680
- QwenP-4379440
- QwenP-4389200
- QwenP-4399960
- QwenP-4409720
- QwenP-4419480
- QwenP-4429240
- QwenP-4439000
- QwenP-4449760
- QwenP-4459520
- QwenP-4469280
- QwenP-4479040
- QwenP-4489800
- QwenP-4499560
- QwenP-4509320
- QwenP-4519080
- QwenP-4529840
- QwenP-4539600
- QwenP-4549360
- QwenP-4559120
- QwenP-4569880
- QwenP-4579640
- QwenP-4589400
- QwenP-4599160
- QwenP-4609920
- QwenP-4619680
- QwenP-4629440
- QwenP-4639200
- QwenP-4649960
- QwenP-4659720
- QwenP-4669480
- QwenP-4679240
- QwenP-4689000
- QwenP-4699760
- QwenP-4709520
- QwenP-4719280
- QwenP-4729040
- QwenP-4739800
- QwenP-4749560
- QwenP-4759320
- QwenP-4769080
- QwenP-4779840
- QwenP-4789600
- QwenP-4799360
- QwenP-4809120
- QwenP-4819880
- QwenP-4829640
- QwenP-4839400
- QwenP-4849160
- QwenP-4859920
- QwenP-4869680
- QwenP-4879440
- QwenP-4889200
- QwenP-4899960
- QwenP-4909720
- QwenP-4919480
- QwenP-4929240
- QwenP-4939000
- QwenP-4949760
- QwenP-4959520
- QwenP-4969280
- QwenP-4979040
- QwenP-4989800
- QwenP-4999560
- QwenP-5009320
- QwenP-5019080
- QwenP-5029840
- QwenP-5039600
- QwenP-5049360
- QwenP-5059120
- QwenP-5069880
- QwenP-5079640
- QwenP-5089400
- QwenP-5099160
- QwenP-5109920
- QwenP-5119680
- QwenP-5129440
- QwenP-5139200
- QwenP-5149960
- QwenP-5159720
- QwenP-5169480
- QwenP-5179240
- QwenP-5189000
- QwenP-5199760
- QwenP-5209520
- QwenP-5219280
- QwenP-5229040
- QwenP-5239800
- QwenP-5249560
- QwenP-5259320
- QwenP-5269080
- QwenP-5279840
- QwenP-5289600
- QwenP-5299360
- QwenP-5309120
- QwenP-5319880
- QwenP-5329640
- QwenP-5339400
- QwenP-5349160
- QwenP-5359920
- QwenP-5369680
- QwenP-5379440
- QwenP-5389200
- QwenP-5399960
- QwenP-5409720
- QwenP-5419480
- QwenP-5429240
- QwenP-5439000
- QwenP-5449760
- QwenP-5459520
- QwenP-5469280
- QwenP-5479040
- QwenP-5489800
- QwenP-5499560
- QwenP-5509320
- QwenP-5519080
- QwenP-5529840
- QwenP-5539600
- QwenP-5549360
- QwenP-5559120
- QwenP-5569880
- QwenP-5579640
- QwenP-5589400
- QwenP-5599160
- QwenP-5609920
- QwenP-5619680
- QwenP-5629440
- QwenP-5639200
- QwenP-5649960
- QwenP-5659720
- QwenP-5669480
- QwenP-5679240
- QwenP-5689000
- QwenP-5699760
- QwenP-5709520
- QwenP-5719280
- QwenP-5729040
- QwenP-5739800
- QwenP-5749560
- QwenP-5759320
- QwenP-5769080
- QwenP-5779840
- QwenP-5789600
- QwenP-5799360
- QwenP-5809120
- QwenP-5819880
- QwenP-5829640
- QwenP-5839400
- QwenP-5849160
- QwenP-5859920
- QwenP-5869680
- QwenP-5879440
- QwenP-5889200
- QwenP-5899960
- QwenP-5909720
- QwenP-5919480
- QwenP-5929240
- QwenP-5939000
- QwenP-5949760
- QwenP-5959520
- QwenP-5969280
- QwenP-5979040
- QwenP-5989800
- QwenP-5999560
- QwenP-6009320
- QwenP-6019080
- QwenP-6029840
- QwenP-6039600
- QwenP-6049360
- QwenP-6059120
- QwenP-6069880
- QwenP-6079640
- QwenP-6089400
- QwenP-6099160
- QwenP-6109920
- QwenP-6119680
- QwenP-6129440
- QwenP-6139200
- QwenP-6149960
- QwenP-6159720
- QwenP-6169480
- QwenP-6179240
- QwenP-6189000
- QwenP-6199760
- QwenP-6209520
- QwenP-6219280
- QwenP-6229040
- QwenP-6239800
- QwenP-6249560
- QwenP-6259320
- QwenP-6269080
- QwenP-6279840
- QwenP-6289600
- QwenP-6299360
- QwenP-6309120
- QwenP-6319880
- QwenP-6329640
- QwenP-6339400
- QwenP-6349160
- QwenP-6359920
- QwenP-6369680
- QwenP-6379440
- QwenP-6389200
- QwenP-6399960
- QwenP-6409720
- QwenP-6419480
- QwenP-6429240
- QwenP-6439000
- QwenP-6449760
- QwenP-6459520
- QwenP-6469280
- QwenP-6479040
- QwenP-6489800
- QwenP-6499560
- QwenP-6509320
- QwenP-6519080
- QwenP-6529840
- QwenP-6539600
- QwenP-6549360
- QwenP-6559120
- QwenP-6569880
- QwenP-6579640
- QwenP-6589400
- QwenP-6599160
- QwenP-6609920
- QwenP-6619680
- QwenP-6629440
- QwenP-6639200
- QwenP-6649960
- QwenP-6659720
- QwenP-6669480
- QwenP-6679240
- QwenP-6689000
- QwenP-6699760
- QwenP-6709520
- QwenP-6719280
- QwenP-6729040
- QwenP-6739800
- QwenP-6749560
- QwenP-6759320
- QwenP-6769080
- QwenP-6779840
- QwenP-6789600
- QwenP-6799360
- QwenP-6809120
- QwenP-6819880
- QwenP-6829640
- QwenP-6839400
- QwenP-6849160
- QwenP-6859920
- QwenP-6869680
- QwenP-6879440
- QwenP-6889200
- QwenP-6899960
- QwenP-6909720
- QwenP-6919480
- QwenP-6929240
- QwenP-6939000
- QwenP-6949760
- QwenP-6959520
- QwenP-6969280
- QwenP-6979040
- QwenP-6989800
- QwenP-6999560
- QwenP-7009320
- QwenP-7019080
- QwenP-7029840
- QwenP-7039600
- QwenP-7049360
- QwenP-7059120
- QwenP-7069880
- QwenP-7079640
- QwenP-7089400
- QwenP-7099160
- QwenP-7109920
- QwenP-7119680
- QwenP-7129440
- QwenP-7139200
- QwenP-7149960
- QwenP-7159720
- QwenP-7169480
- QwenP-7179240
- QwenP-7189000
- QwenP-7199760
- QwenP-7209520
- QwenP-7219280
- QwenP-7229040
- QwenP-7239800
- QwenP-7249560
- QwenP-7259320
- QwenP-7269080
- QwenP-7279840
- QwenP-7289600
- QwenP-7299360
- QwenP-7309120
- QwenP-7319880
- QwenP-7329640
- QwenP-7339400
- QwenP-7349160
- QwenP-7359920
- QwenP-7369680
- QwenP-7379440
- QwenP-7389200
- QwenP-7399960
- QwenP-7409720
- QwenP-7419480
- QwenP-7429240
- QwenP-7439000
- QwenP-7449760
- QwenP-7459520
- QwenP-7469280
- QwenP-7479040
- QwenP-7489800
- QwenP-7499560
- QwenP-7509320
- QwenP-7519080
- QwenP-7529840
- QwenP-7539600
- QwenP-7549360
- QwenP-7559120
- QwenP-7569880
- QwenP-7579640
- QwenP-7589400
- QwenP-7599160
- QwenP-7609920
- QwenP-7619680
- QwenP-7629440
- QwenP-7639200
- QwenP-7649960
- QwenP-7659720
- QwenP-7669480
- QwenP-7679240
- QwenP-7689000
- QwenP-7699760
- QwenP-7709

<https://www.royalcyber.com/resources/videos/webcasts/transforming-insurance-operations-through-generative-ai/>

## Company financial analysis:

<https://chat.financialdatasets.ai/>

<https://github.com/virattt/ai-financial-agent>

## Stocks sentiment analysis:

<https://huggingface.co/spaces/Karthikeyen92/Stock-Sentiment-Analysis>

LLM in fraud detection

<https://www.youtube.com/watch?v=lRRtDrwd2z8>

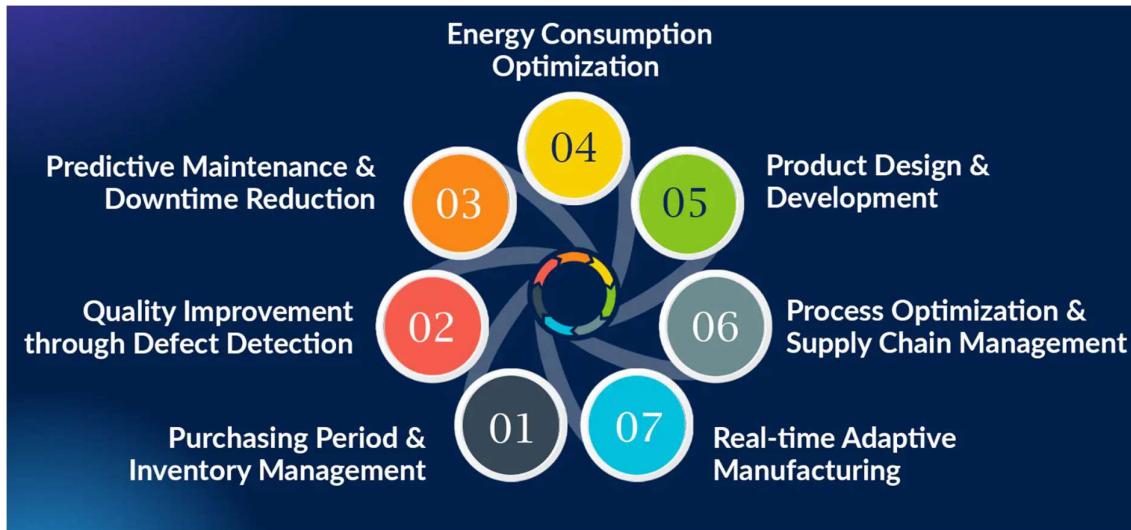
<https://github.com/amitkedia007/Financial-Fraud-Detection-Using-LLMs>

<https://www.kaggle.com/datasets/amitkedia/financial-statement-fraud-data>

Fine tuning llm for fraud detection using Azure OpenAI

<https://mkonda007.medium.com/fine-tuning-ai-models-creating-a-financial-fraud-detection-assistant-c23505c8d669>

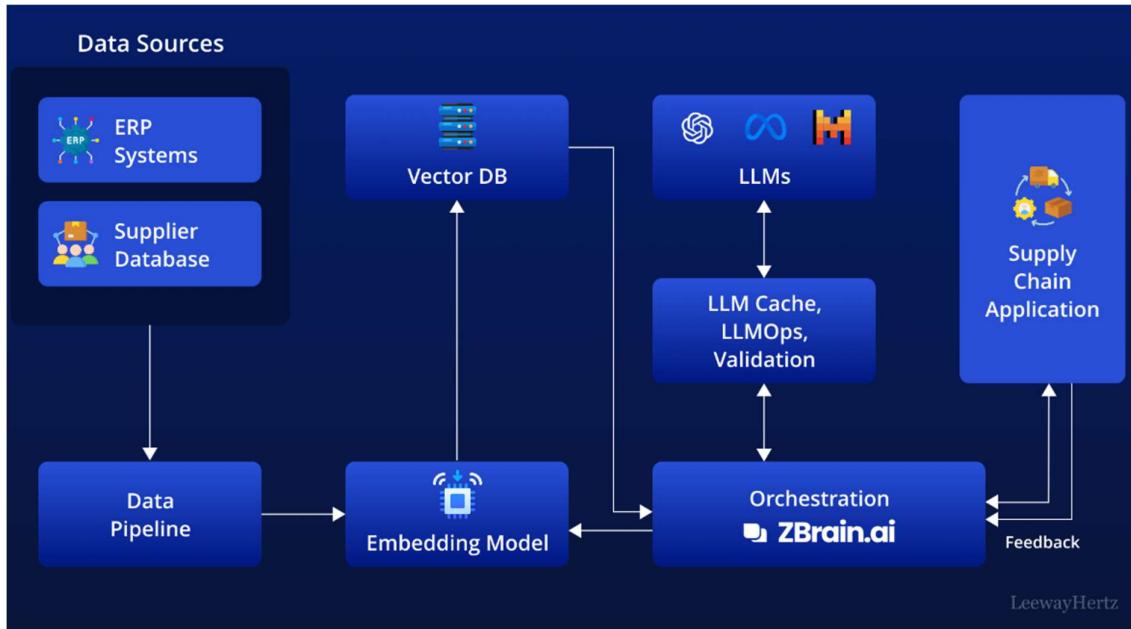
## Manufacturing



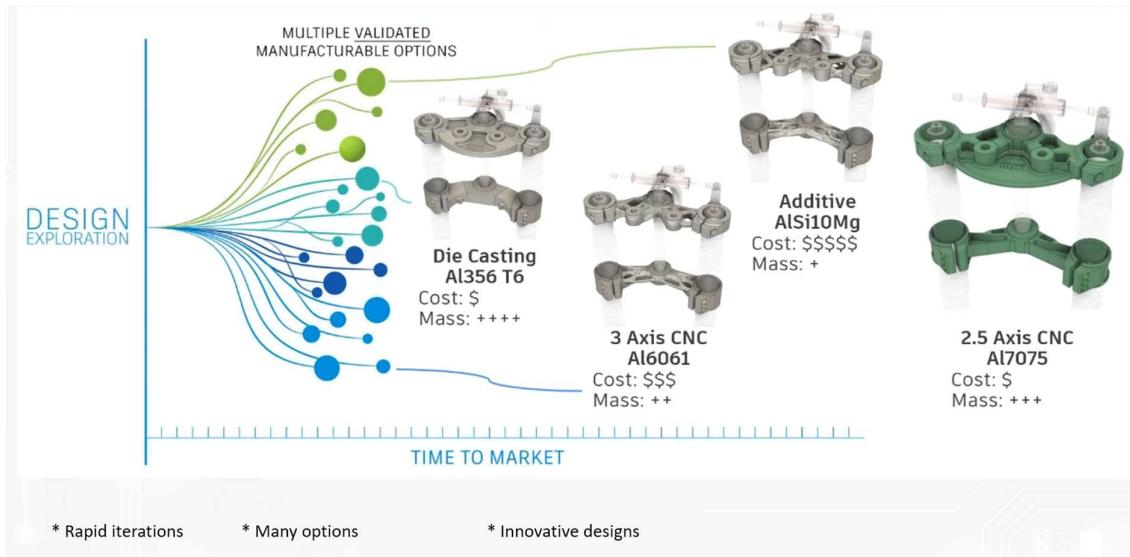
<https://www.matellio.com/blog/generative-ai-in-manufacturing-use-cases/>

<https://nextgeninvent.com/blogs/generative-ai-in-manufacturing-use-cases/>

<https://www.leewayhertz.com/generative-ai-in-supply-chain/>



**Generative AI in CAD/CAM, Product Design:**



<https://www.youtube.com/watch?v=PSSt8wsNQ>

<https://www.youtube.com/watch?v=a7nRzOOVgB8>

**AI and Digital Twin:**



<https://www.youtube.com/watch?v=g78YHYXXils>



<https://aichat.com/mitsubishi-motors-singapore-chatbot/>

05.10.2023 Press Release

## Hyundai Motor's AI-Generated Campaign Invites Instagram Users to Visualize Dream Destinations with All-New SANTA FE



<https://www.hyundai.news/eu/articles/press-releases/santa-fe-ai-generated-campaign-invites-instagram-users-to-dream-destinations.html>

### Media and Entertainment

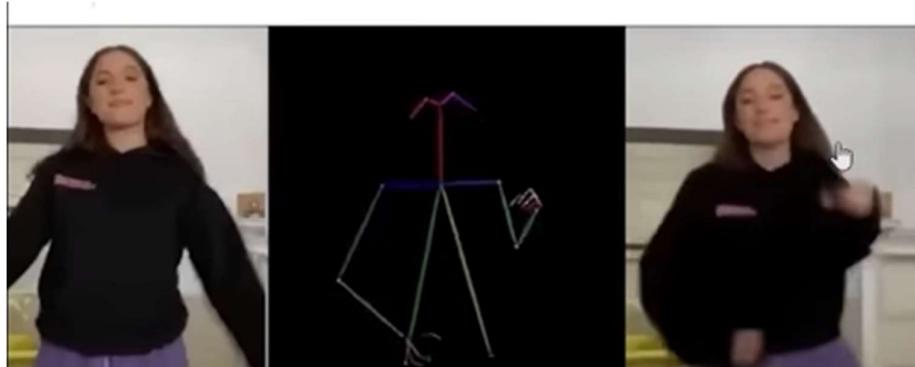
#### AI Influencers + Verification

'Spanish influencer' created entirely by AI generates its modelling agency £9,000 a month with 200,000 followers

4 December 2023, 18:02 | Updated: 4 December 2023, 19:03



AI can make you dance

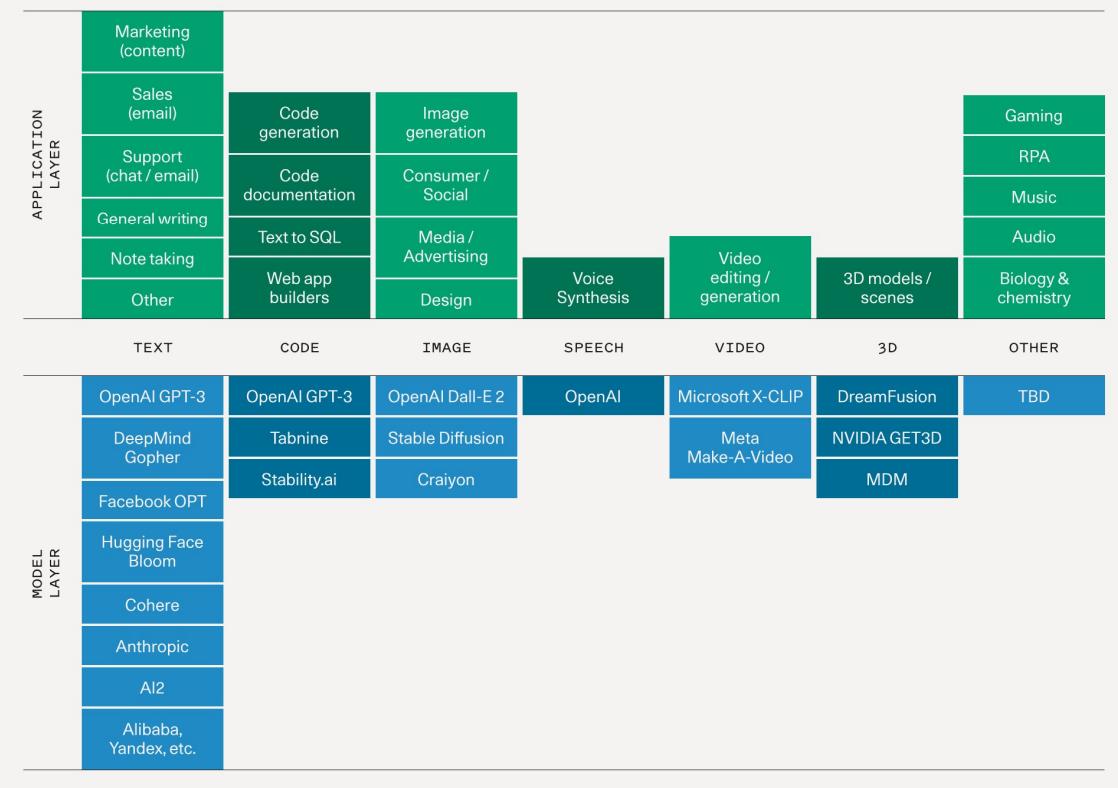


Top AI tools for content creation

[https://www.instagram.com/awa\\_k\\_penn/p/DESWVNSNsZ/](https://www.instagram.com/awa_k_penn/p/DESWVNSNsZ/)

Popular Gen-AI tools and applications

## The Generative AI Application Landscape



A good curated list here:

<https://hashcollision.substack.com/p/a-primer-on-generative-ai-genai>

### AI Chatbots:

OpenAI ChatGPT: <https://chatgpt.com>

Microsoft copilot: <https://copilot.microsoft.com>

Google Gemini: <https://gemini.google.com>

Grok from X: <https://x.com/i/grok?focus=1>

Anthropic Claude (Amazon backed): <https://claude.ai>

Perplexity AI: <https://www.perplexity.ai/>

Meta AI: <https://www.meta.ai/>

Cohere: <https://dashboard.cohere.com/>

### Presentation Slides Generator:

<https://gamma.app/>

<https://www.slidesai.io/>

<https://pptgenai.com/>

**Text to video:**

<https://invideo.io>

<https://sora.com/login?next=%2Flibrary>

<https://help.openai.com/en/articles/9957612-generating-videos-on-sora>

Sora demo:

<https://openai.com/index/sora/>

AI Studios:

<https://app.aistudios.com/dashboard>

**Summarize PDF , chat with pdf, generate slide deck from pdf**

<https://chatpdf.com>

<https://popai.pro>

<https://www.presentations.ai/tools/pdf-to-ppt>

<https://getcoralai.com>

**Excel AI, Dashboards with chat**

Arcwise:

<https://arcwise.app/>

Google AI Studio:

<https://aistudio.google.com/live>

See demo at:

<https://www.youtube.com/watch?v=3cvczHJSRNs>

Excel Dashboard AI:

<https://www.exceldashboard.ai/>

Bricks:

[https://www.thebricks.com/?utm\\_source=bricks&utm\\_medium=resources&utm\\_campaign=how-to-create-a-sales-dashboard-with-ai](https://www.thebricks.com/?utm_source=bricks&utm_medium=resources&utm_campaign=how-to-create-a-sales-dashboard-with-ai)

<https://www.thebricks.com/resources/how-to-create-a-sales-dashboard-with-ai>

Demo:

<https://www.youtube.com/watch?v=-tl-oxY-DAE>

<https://www.youtube.com/watch?v=1vz6EkUhbOg>

### **Marketing and Sales**

Apollo: <https://www.apollo.io/>

Hubspot: 10 best AI tools in 2024:

<https://blog.hubspot.com/sales/ai-sales-tools>

A long list of AI tools for Sales:

<https://www.coldiq.com/ai-sales-tools>

### **Image Generation:**

Midjourney: <https://www.midjourney.com/explore?tab=top>

### **Conversational AI and Agents/Audio:**

Eleven Labs:

<https://elevenlabs.io/app/conversational-ai>

Murf:

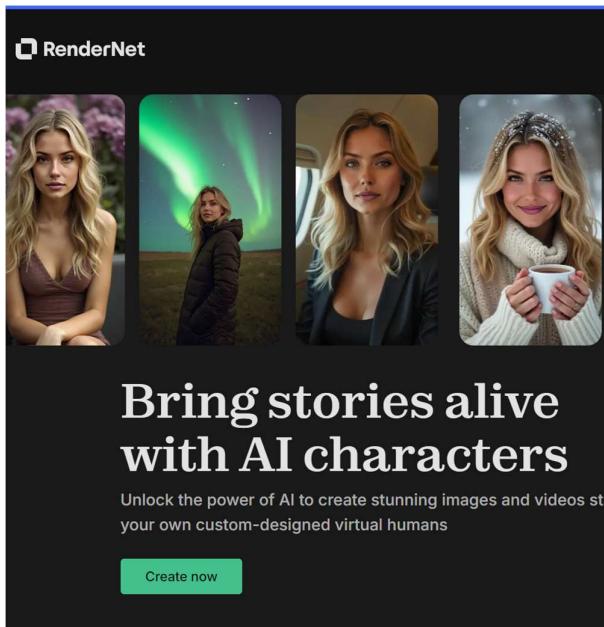
<https://murf.ai>

### **Vision Agent**

<https://va.landing.ai/agent>

RenderNet - create stunning images and videos starring your own custom-designed virtual humans, including voice generation

<https://rendernet.ai/>



## Articles, Blogs, Newsletters, SM Handles

Good sites and blogs for latest in AI:

<https://lablab.ai/blog>

<https://www.thedepview.co/>

<https://arxiv.org/pdf/2401.02843>

<https://medium.com/@akash.kesrwani99/understanding-next-token-prediction-concept-to-code-1st-part-7054dabda347>

Image embeddings:

<https://encord.com/blog/image-embeddings-to-improve-model-performance/>

---

# THOUSANDS OF AI AUTHORS ON THE FUTURE OF AI

---

PREPRINT

<b>Katja Grace<sup>*†</sup></b> AI Impacts Berkeley, California United States katja@aiimpacts.org	<b>Harlan Stewart<sup>†</sup></b> AI Impacts Berkeley, California United States	<b>Julia Fabienne Sandkühler<sup>†</sup></b> Department of Psychology University of Bonn Germany	<b>Stephen Thomas<sup>†</sup></b> AI Impacts Berkeley, California United States
<b>Ben Weinstein-Raun</b> Independent Berkeley, California United States	<b>Jan Brauner</b> Department of Computer Science University of Oxford United Kingdom		

January 2024

## ABSTRACT

In the largest survey of its kind, we surveyed 2,778 researchers who had published in top-tier artificial intelligence (AI) venues, asking for their predictions on the pace of AI progress and the nature and impacts of advanced AI systems. The aggregate forecasts give at least a 50% chance of AI systems achieving several milestones by 2028, including autonomously constructing a payment processing site from scratch, creating a song indistinguishable from a new song by a popular musician, and

<https://hashcollision.substack.com/p/a-primer-on-generative-ai-general>

<https://www.mckinsey.com/capabilities/mckinsey-digital/our-insights/the-economic-potential-of-generative-ai-the-next-productivity-frontier#industry-impacts>

<https://www.deeplearning.ai/>

## Measuring ROI of Generative AI

<https://www.gartner.com/en/articles/take-this-view-to-assess-roi-for-generative-ai#:~:text=Generative%20AI%20promises%20unprecedented%20productivity,Quick%20wins>

<https://embracingenigmas.substack.com/p/next-token-prediction-is-a-fundamental>

## Notable websites on AI:

datacamp.com  
deeplearning.ai  
towardsai.net

### Notable Instagram Handles on AI:

Name	URL	Description
Aiverse	<a href="https://www.instagram.com/aiversepage/">https://www.instagram.com/aiversepage/</a>	Shares insights and updates on artificial intelligence and its applications.
Artificial Intelligence News	<a href="https://www.instagram.com/artificialintelligencenews.in/">https://www.instagram.com/artificialintelligencenews.in/</a>	Provides the latest news and developments in AI and machine learning.
AI Explaining	<a href="https://www.instagram.com/aiexplaining/">https://www.instagram.com/aiexplaining/</a>	Breaks down complex AI concepts into understandable content for enthusiasts.
NVIDIA AI	<a href="https://www.instagram.com/nvidiaai/">https://www.instagram.com/nvidiaai/</a>	Showcases advancements and applications of AI technology by NVIDIA.
Microsoft Research	<a href="https://www.instagram.com/microsoftresearch/">https://www.instagram.com/microsoftresearch/</a>	Highlights research and innovations in AI and other computing fields by Microsoft.
Amazon Science	<a href="https://www.instagram.com/amazon.science/">https://www.instagram.com/amazon.science/</a>	Shares Amazon's scientific research and developments in AI and related areas.
MIT Technology Review	<a href="https://www.instagram.com/technologyreview/">https://www.instagram.com/technologyreview/</a>	Covers the latest in technology, including AI breakthroughs and analyses.
Deep Learning Nerds	<a href="https://www.instagram.com/deeplearningnerds/">https://www.instagram.com/deeplearningnerds/</a>	Focuses on AI, data science, and machine learning content for enthusiasts and professionals.
Best of AI	<a href="https://www.instagram.com/bestofaidaily/">https://www.instagram.com/bestofaidaily/</a>	Curates top AI-related content, news, and innovations from around the world.
AI Art Universe	<a href="https://www.instagram.com/aiartuniverse/">https://www.instagram.com/aiartuniverse/</a>	Showcases AI-generated art and explores the intersection of creativity and technology.
AI Trends	<a href="https://www.instagram.com/aitrends/">https://www.instagram.com/aitrends/</a>	Provides updates on the latest trends and developments in artificial intelligence.
AI Daily	<a href="https://www.instagram.com/aidaily/">https://www.instagram.com/aidaily/</a>	Shares daily news, insights, and articles related to AI and machine learning.
AI Innovations	<a href="https://www.instagram.com/ai_innovations/">https://www.instagram.com/ai_innovations/</a>	Focuses on innovative applications and breakthroughs in the field of AI.
AI Hub	<a href="https://www.instagram.com/aihub/">https://www.instagram.com/aihub/</a>	A community page sharing AI news, research, and technological advancements.

### Notable Twitter (X) handles on AI:

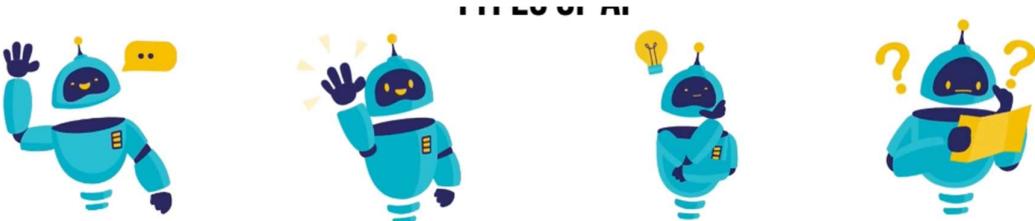
X Handle	URL	Description
Tut_ml	<a href="https://x.com/tut_ml">https://x.com/tut_ml</a>	ML Blogs
LangChainAI	<a href="https://x.com/LangChainAI">https://x.com/LangChainAI</a>	Building applications with LLMs through composability.
Ethan Mollick	<a href="https://x.com/emollick">https://x.com/emollick</a>	Professor at Wharton, sharing insights on AI, innovation, and education.
Omar Sar0	<a href="https://x.com/omarsar0">https://x.com/omarsar0</a>	AI researcher and enthusiast, focusing on machine learning advancements.
Rob Feers	<a href="https://x.com/rfeers">https://x.com/rfeers</a>	Data scientist sharing thoughts on AI, data analysis, and technology trends.
DataCamp	<a href="https://x.com/DataCamp">https://x.com/DataCamp</a>	Learn data skills online at your own pace.
KDnuggets	<a href="https://x.com/KDnuggets">https://x.com/KDnuggets</a>	Leading site on AI, Analytics, Big Data, Data Mining, and Data Science.
iScienceLuvr	<a href="https://x.com/iScienceLuvr">https://x.com/iScienceLuvr</a>	Enthusiast sharing the latest in AI research and machine learning developments.
Shubham Saboo	<a href="https://x.com/Saboo_Shubham">https://x.com/Saboo_Shubham</a>	AI practitioner and educator, providing tutorials and insights on machine learning.
Victor Dey	<a href="https://x.com/victor_explor_e">https://x.com/victor_explor_e</a>	AI journalist exploring the latest trends and breakthroughs in artificial intelligence.
Sam Altman	<a href="https://x.com/sama">https://x.com/sama</a>	CEO of OpenAI, sharing insights on AI development and its implications.
Greg Brockman	<a href="https://x.com/gdb">https://x.com/gdb</a>	Co-founder and Chairman of OpenAI, discussing AI advancements and research.
Demis Hassabis	<a href="https://x.com/demishassabis">https://x.com/demishassabis</a>	CEO of DeepMind, focusing on AI research and its applications.
Andrew Ng	<a href="https://x.com/AndrewYNg">https://x.com/AndrewYNg</a>	Co-founder of Coursera and AI researcher, offering insights into machine learning and education.
Allie K. Miller	<a href="https://x.com/alliekmiller">https://x.com/alliekmiller</a>	AI leader and investor, sharing knowledge on machine learning and AI startups.

### Notable YouTube channels on AI:

Name	Description	URL
Infinite Codes	A channel dedicated to coding tutorials and software development insights.	<a href="https://www.youtube.com/@InfiniteCodes">https://www.youtube.com/@InfiniteCodes</a>
AI Papers Academy	Focuses on discussing and analyzing academic papers in the field of artificial intelligence.	<a href="https://www.youtube.com/@aipapersacademy">https://www.youtube.com/@aipapersacademy</a>
GPT for Work	Provides tutorials on integrating GPT models into Excel, Word, Sheets, and Docs.	<a href="https://www.youtube.com/@gptforwork">https://www.youtube.com/@gptforwork</a>
StatQuest	Simplifies complex statistical concepts with clear and engaging explanations.	<a href="https://www.youtube.com/@statquest">https://www.youtube.com/@statquest</a>
Grant Sanderson	Creator of 3Blue1Brown, offering deep dives into mathematical	<a href="https://www.youtube.com/@GrantSanderson">https://www.youtube.com/@GrantSanderson</a>

	concepts with visual explanations.	
AI Revolution	Explores the latest advancements and trends in artificial intelligence and machine learning.	<a href="https://www.youtube.com/@airevolutionx">https://www.youtube.com/@airevolutionx</a>
Vector Institute	Shares research and insights from the Vector Institute, focusing on AI and machine learning.	<a href="https://www.youtube.com/@vectorinstituteari">https://www.youtube.com/@vectorinstituteari</a>
The AI Nexus	Discusses AI developments, tools, and their applications across various industries.	<a href="https://www.youtube.com/@TheAINexusOfficial">https://www.youtube.com/@TheAINexusOfficial</a>
Matthew Berman	Offers content on AI, technology, and their intersection with society.	<a href="https://www.youtube.com/@matthew_berman">https://www.youtube.com/@matthew_berman</a>
AI Thought Leaders	Features interviews and discussions with leading experts in the field of artificial intelligence.	<a href="https://www.youtube.com/@AIThoughtLeaders">https://www.youtube.com/@AIThoughtLeaders</a>
3Blue1Brown	Visualizes mathematical concepts, making them accessible and engaging for a broad audience.	<a href="https://www.youtube.com/@3blue1brown">https://www.youtube.com/@3blue1brown</a>
TheAIGRID	Focuses on AI research, tutorials, and the latest news in the AI community.	<a href="https://www.youtube.com/@TheAiGrid">https://www.youtube.com/@TheAiGrid</a>
Simplilearn	Provides online training videos on various topics, including AI, data science, and programming.	<a href="https://www.youtube.com/@SimplilearnOfficial">https://www.youtube.com/@SimplilearnOfficial</a>
Kore.ai	Specializes in AI-powered chatbots and conversational AI platform insights.	<a href="https://www.youtube.com/@Koreai">https://www.youtube.com/@Koreai</a>
AI Uncovered	Delves into AI technologies, their applications, and implications in the modern world.	<a href="https://www.youtube.com/@AI.Uncovered">https://www.youtube.com/@AI.Uncovered</a>
Hume AI	Explores the intersection of AI and human emotions, focusing on empathetic AI development.	<a href="https://www.youtube.com/@hume_ai">https://www.youtube.com/@hume_ai</a>
The AI Hacker	Provides tutorials and insights on AI programming, hacking, and development techniques.	<a href="https://www.youtube.com/@theaihacker77">https://www.youtube.com/@theaihacker77</a>
Arseny Shatokhin	Shares content on AI, machine learning, and related technological advancements.	<a href="https://www.youtube.com/@vrseen">https://www.youtube.com/@vrseen</a>
AI News	Offers the latest news and updates in the world of artificial intelligence.	<a href="https://www.youtube.com/@AINewsOfficial">https://www.youtube.com/@AINewsOfficial</a>
David Shapiro	Discusses AI development, programming, and the ethical considerations surrounding AI.	<a href="https://www.youtube.com/@DaveShap">https://www.youtube.com/@DaveShap</a>

# What does the future hold?



## Reactive AI

Can only respond to the current state of the world. They do not have any memory of past events, and they cannot plan for the future.

## Limited-memory

Can remember past events and use this information to make decisions. Cannot reason about the future or understand the intentions of other agents.

## Theory of Mind

Can understand the thoughts and intentions of other agents. Cooperate with other agents and to achieve goals that would be impossible for a single agent.

## Self Aware

A hypothetical type of AI that would be conscious and have its own subjective experiences. Does not yet exist, but it is a possibility that some researchers believe is worth pursuing.

## What is AI?

### ANI vs. AGI vs. ASI



#### Artificial narrow intelligence (ANI)

Designed to perform specific tasks



#### Artificial general intelligence (AGI)

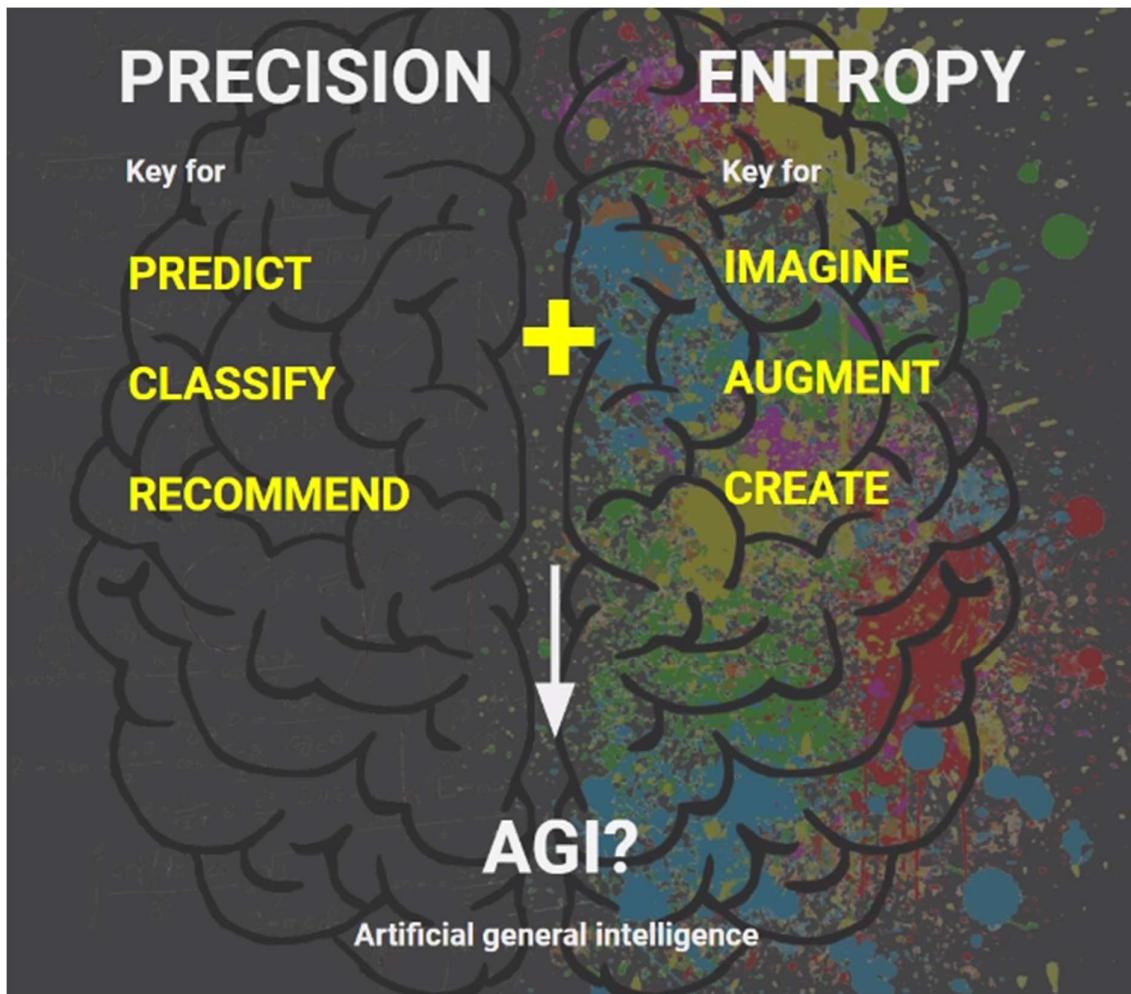
Can behave in a human-like way across all tasks



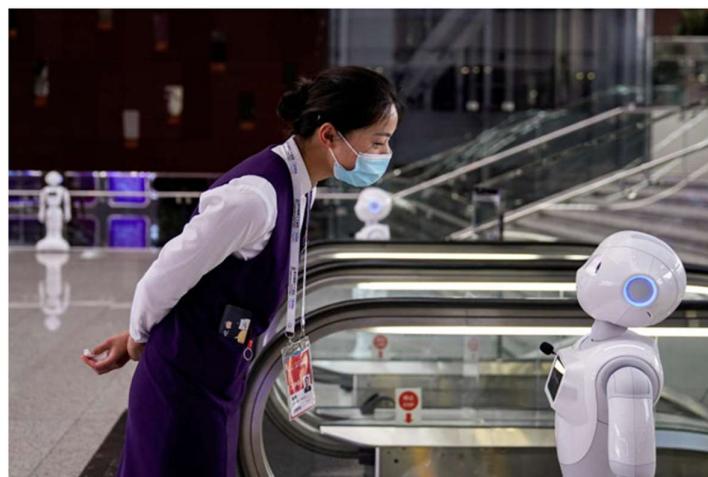
#### Artificial super intelligence (ASI)

Smarter than humans—the stuff of sci-fi

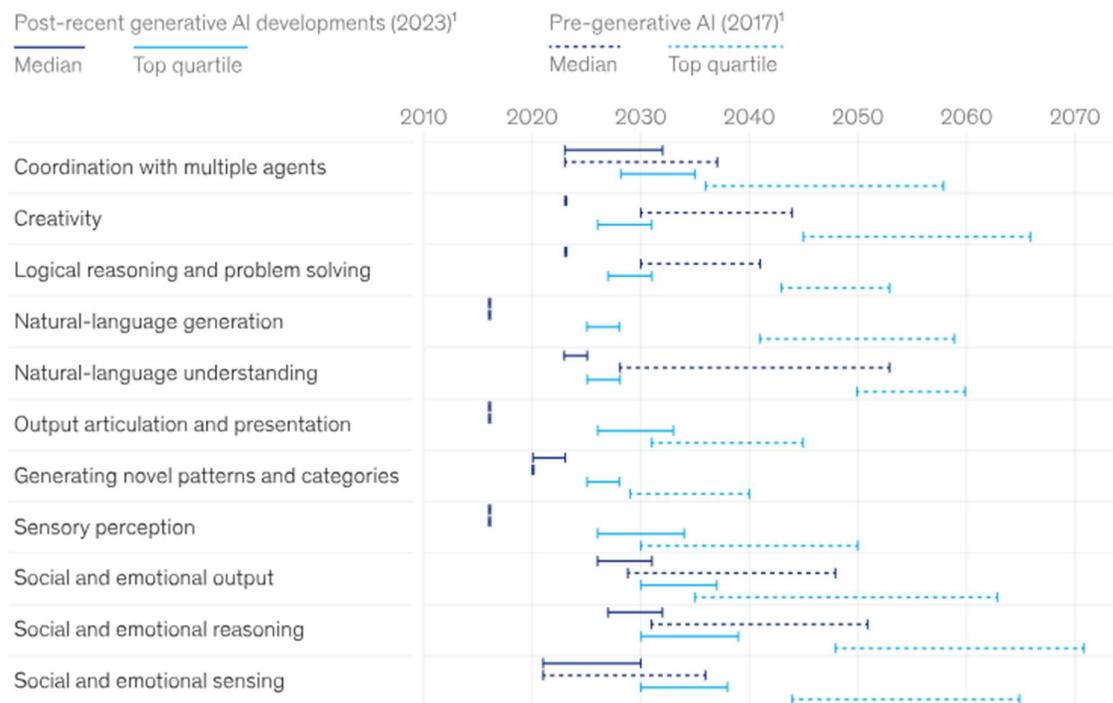
 zapier



Sentience refers to the capacity to experience feelings and sensations, a concept traditionally associated with living beings.



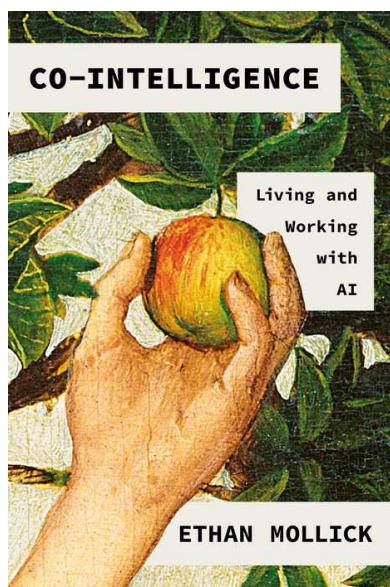
## Estimated range for technology to achieve human-level performance, by technical capability



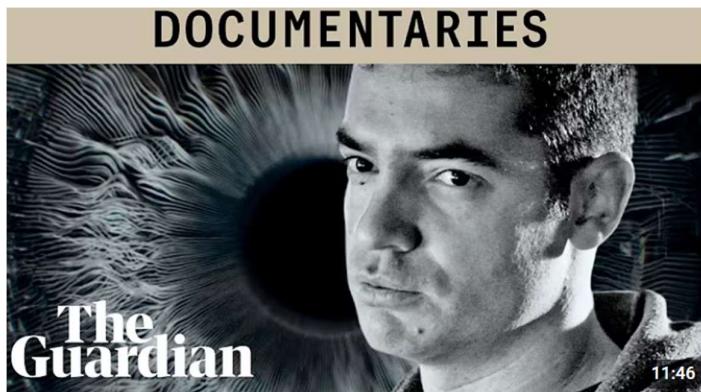
<sup>1</sup>Comparison made on the business-related tasks required from human workers.  
Source: McKinsey Global Institute occupation database; McKinsey analysis

McKinsey & Company

<https://www.penguinrandomhouse.com/books/741805/co-intelligence-by-ethan-mollick/?ref=PRH410E2C567AF>



Ilya interview with the GUARDIAN:



<https://www.youtube.com/watch?v=9iqn1HhFJ6c>

THANK YOU!!!

**Prepared By: Vijay Agrawal**

Connect on LinkedIn: <https://www.linkedin.com/in/agrawalvijay/>

Disclaimers:

- 1) These notes are meant as supplemental reading for participants in the author's Gen AI training program.
- 2) They are compendium of author's own notes and snippets from a variety of publicly available sources. Citations and source credits provided where applicable.
- 3) Author does not endorse any of the mentioned tools nor does he have any commercial interests with any of the vendors mentioned
- 4) There are no explicit or implied warranties – the materials are provided as-is for reference and learning
- 5) Highly recommended to do own research when applying Gen AI to real life problems as this space is rapidly evolving and new tools and techniques are coming up every day.