

ie

Course

Day 1: Sentiment Analysis - 27 of September from 10:30am to 12:30pm - Online

Day 2: Market Trend Identification - 11th of October from 10:30am to 12:30pm - Online

Manoel Gadi

more than 16 years in Banking and Financial Institutions like Citibank and Santander
most in Analytics functions for Risk Management in Brazil (Sao Paulo), the UK (Milton Keynes) and Spain (Madrid)
Teaching in IE University since 2013 and since 2019 he is fully dedicated to teaching in IE University in courses ranging from Programming, Statistics, Machine Learning, Banking, and Fintech.



Please say your name, department and your highest experience with AI, what you've heard about and what you would like to be exposed to and why in this course:

Fill in here: https://docs.google.com/spreadsheets/d/1EMR9BBLJr-g_B1PnLjSZcSzvuuyHpgg/edit?usp=sharing&oid=100949985522117931296&rtpof=true&sd=true

1 - Prompting involves asking question to AI and receiving the answer, then editing/using the answer in another tool (email, Excel, Word, etc...).

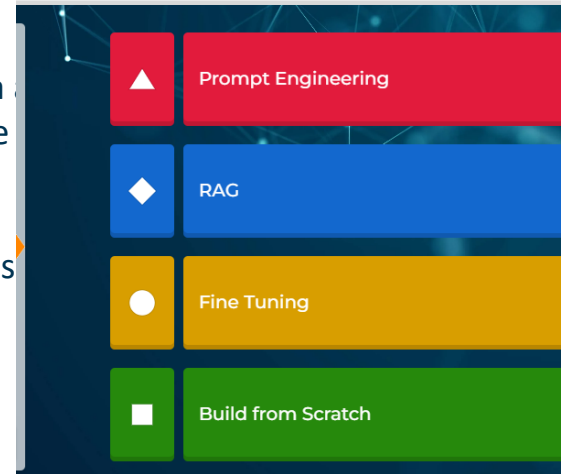
2 - Prompt engineering involves crafting queries providing context and/or data input to elicit better analysis or responses from the model without altering the model.

3 - RAG involves augmenting an LLM with access to a dynamic, curated databases to improve outputs.

4 - Fine-tuning involves training an LLM on a smaller, specialized dataset to adjust its parameters for specific tasks.

5 - Build from Scratch involves learning how to create, train, and tweak large language models (LLMs) by building one from the ground up!

from lowest to highest cost & complexity



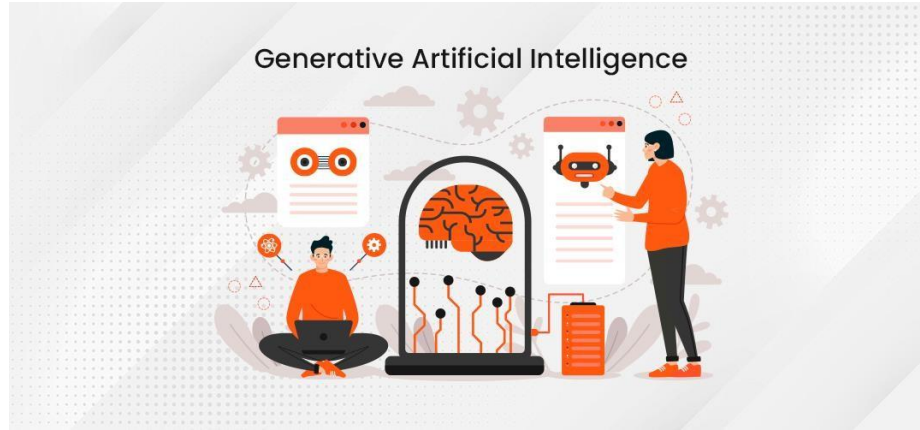
Agenda

Generative AI

- Definitions
- Gen1 types
- Gen2 types
- Applications
- Ecosystem
- Prompt Engineering
- Practice
- Manual Labelling Data
- Discussion
- Wrap up Kahoot

Generative AI - Definition

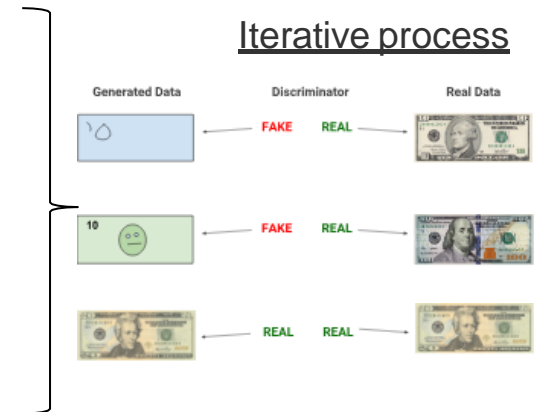
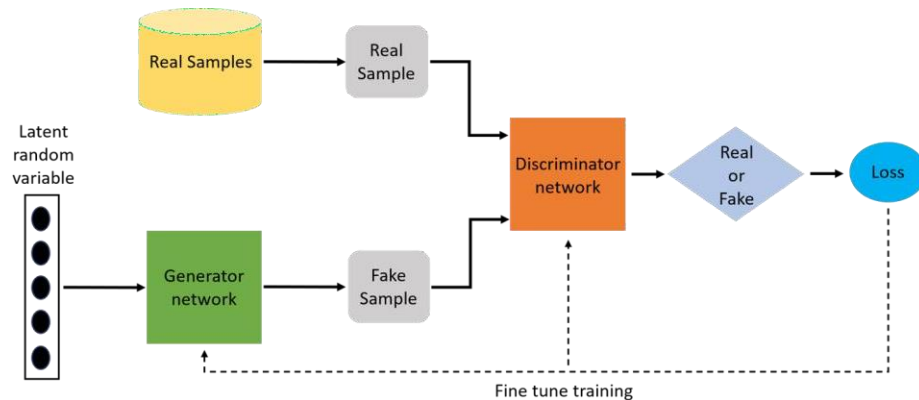
- *“Generative artificial intelligence (AI) describes algorithms that can be used to create new content, including audio, code, images, text, simulations, and videos.”*
– McKinsey, 2023



- *“Generative AI refers to AI techniques that learn a representation of artifacts from data, and use it to generate brand-new, unique artifacts that resemble but don’t repeat the original data.”* – Gartner, 2023

Generative AI – Gen1 types

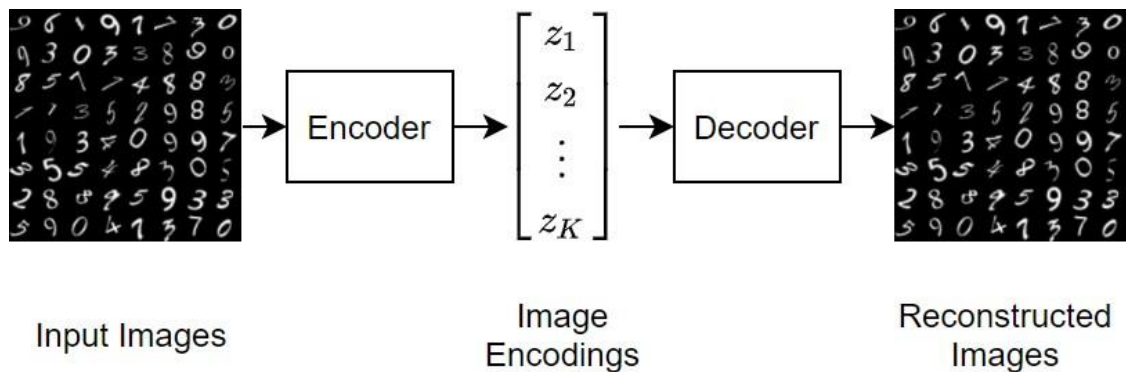
- A Generative adversarial network, or **GAN** for short, is a type of machine learning model that uses deep learning techniques to generate new data based on patterns learned from existing data.



Once the “Generator” is trained, you can use it to create new, synthetic data that is similar to some existing real data.

Generative AI – Gen1 types

- A variational autoencoder (**VAE**) is a type of machine learning model that learns to reproduce its input and map data to latent space*, which contains a compressed representation of the input data.

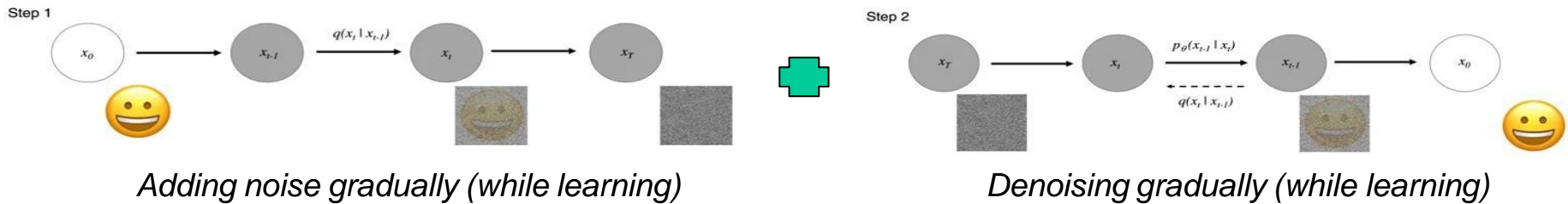


Once trained, VAEs can generate new, synthetic data that is similar to the input data by randomly sampling points from the latent space and decoding them.

** latent space: a condensed or simplified version of the input data.*

Generative AI – Gen2 types

- A **Diffusion model** is a kind of machine learning model that uses a process of gradual change* to turn a simple, known type of random information (like Gaussian noise) into another type of information that we're more interested in (like images or text). This process happens in stages and goes in reverse order, similar to how heat spreads through an object.

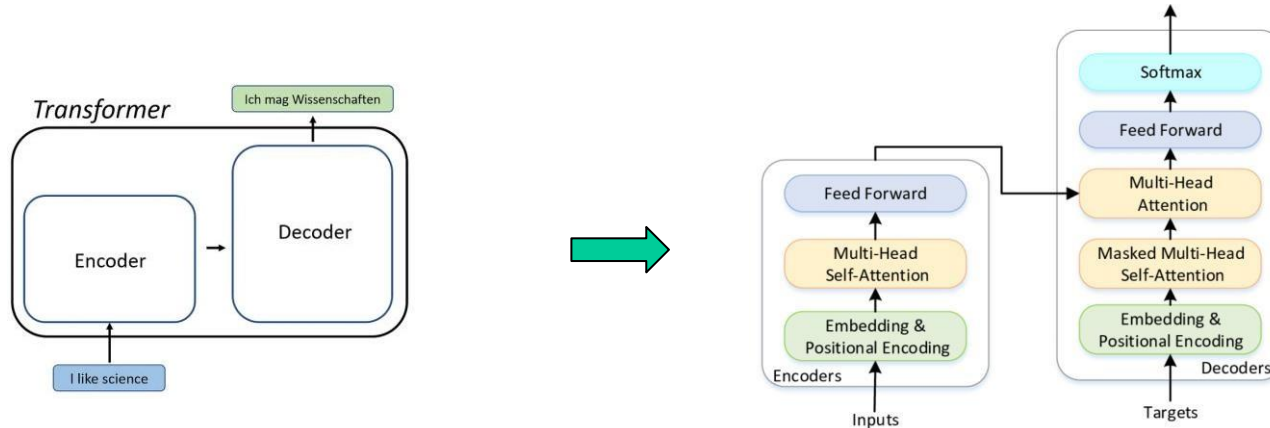


Once the model is trained, it can make new data that looks like the data it learned from quite precisely.

* *GANs use the initial randomness as the seed for a direct transformation (that will be evaluated), while Diffusion models use it as the starting point for a gradual transformation.*

Generative AI – Gen2 types

- A **Transformer**-based model is a type of machine learning model that employs the concepts of self-attention and feedforwarding, meaning they numerically weigh the importance of each element in a sequence in relation to all others.

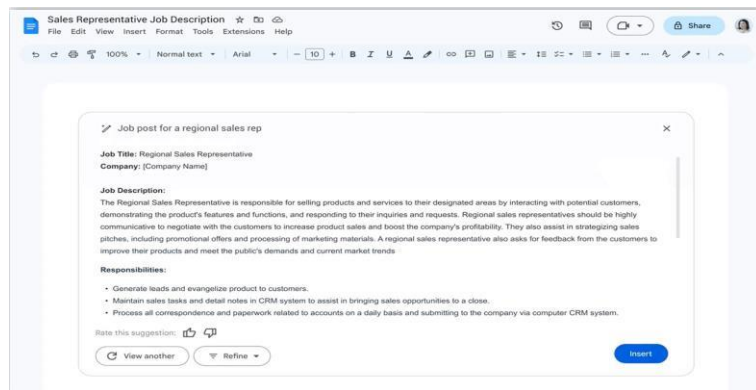
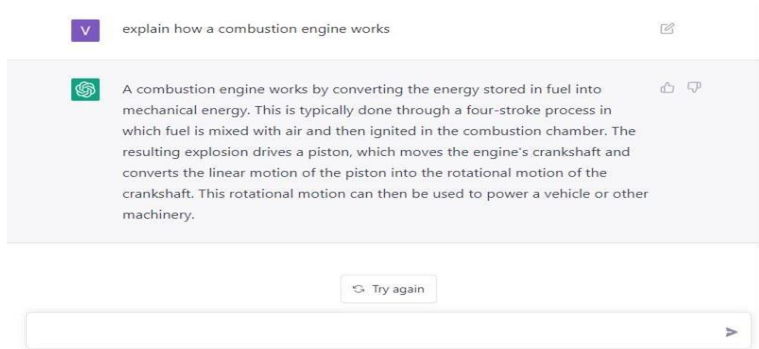


Once a Transformer model is trained, it can understand and generate data that is contextually similar to the training data, maintaining the structural and semantic properties of the input.

Generative AI – Applications

Generative AI, with its ability to create new content based on patterns learned from existing data, has found diverse applications across various domains.

1. Text / LLMs (OpenAI's ChatGPT, Google's Bard, etc.)



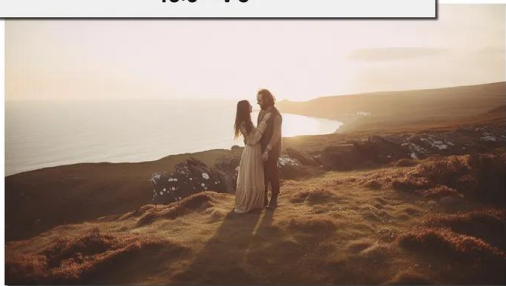
Large Language Models (LLMs) must be capable of capturing the sequential nature of text data. Transformer-based architectures are particularly effective in this regard due to their ability to account for long-range dependencies within the text data.

Generative AI – Applications

2. Images (Midjourney, OpenAI's DALL-E 2, etc.)

Generative Image Models (GIMs) need to be capable of capturing the inherent structure and intricate patterns within image data.

wide angle wedding photography of beautiful couple on the coast of ireland at sunset, aesthetic boho style brown and beige colors, shot on disposable film camera, light leaks --ar 16:9 --v 5



H/T: Christian Maki Grab

colJang

Midjourney ColJang



They have significantly evolved over time, witnessing a major transition from GANs and VAEs to more recent Diffusion models (that offer a blend of the strengths of GANs and VAEs).

Generative AI – Applications

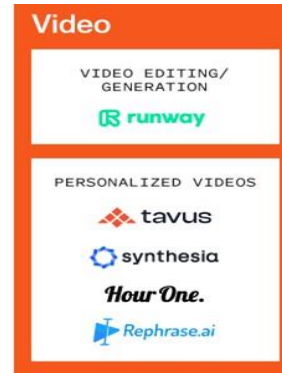
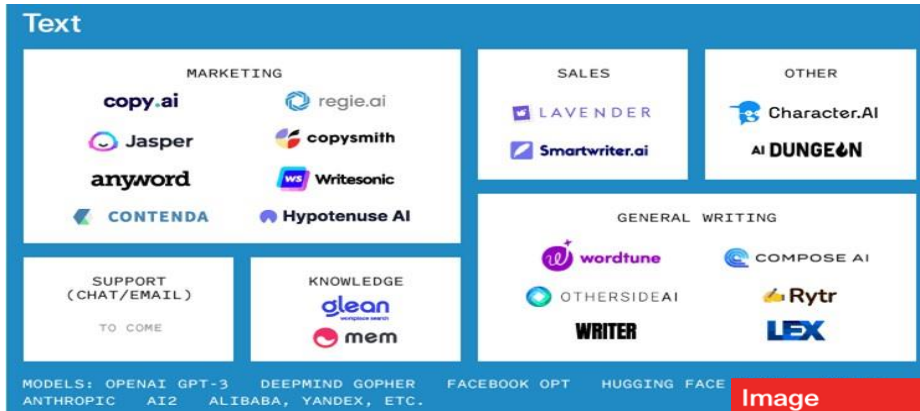
3. Video and Audio (Meta's Make-a-video, Google's AudioLM, etc.)

Generative Video Models (GVMs) need to be able to comprehend the underlying temporal dynamics and spatial relationships within video data. Still young technology, but Transformer-based models are emerging as a promising solution.

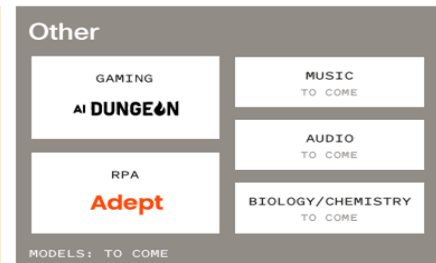
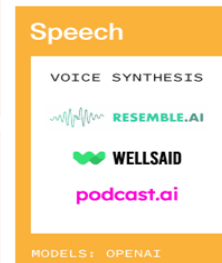
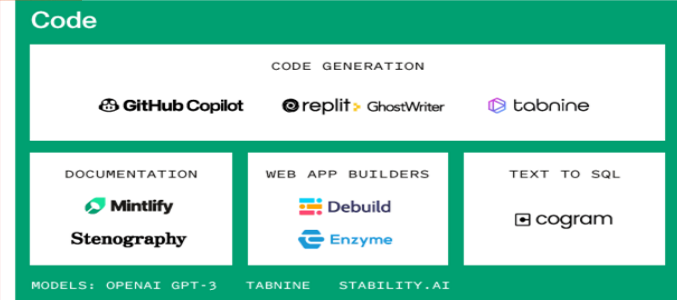
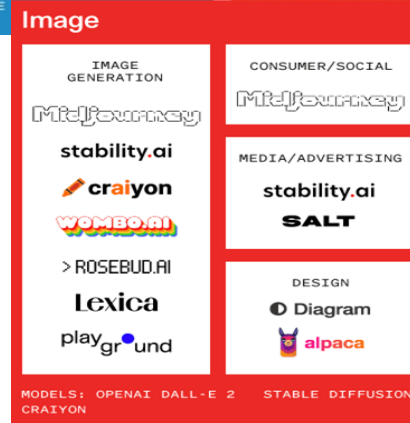
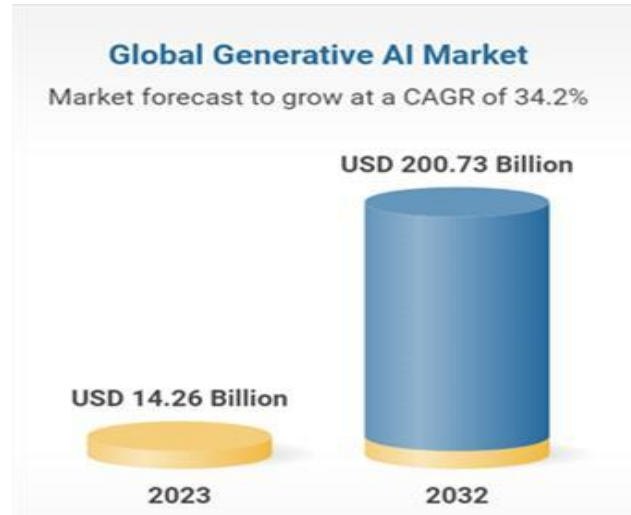


Generative Audio Models (GAMs) are required to capture the time-sequential patterns and rich nuances within audio data. The very last models are using a combination of Transformers-based + Diffusion models with good results.

Generative AI – Ecosystem



Powerful partnerships:



Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

Emerging field in the ecosystem: Prompt Engineering

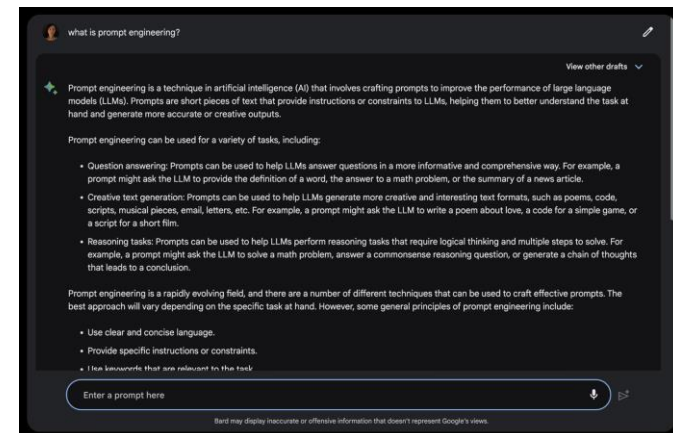
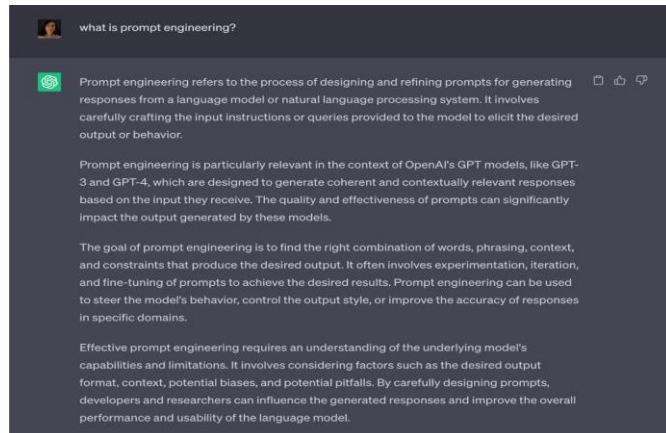
An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

OK, but what is a **prompt**?

Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

OK, but what is a **prompt**?



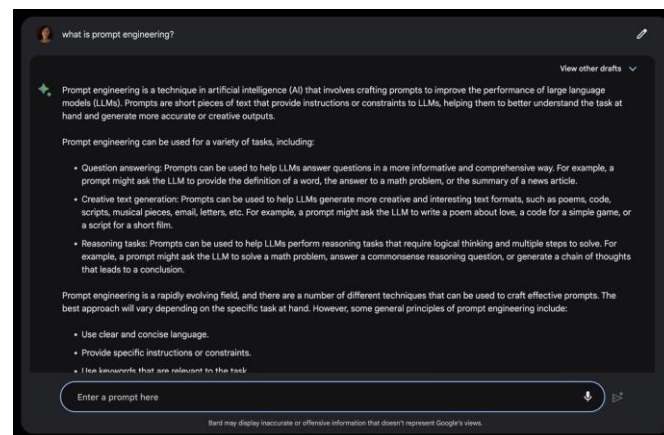
Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

OK, but what is a **prompt**?



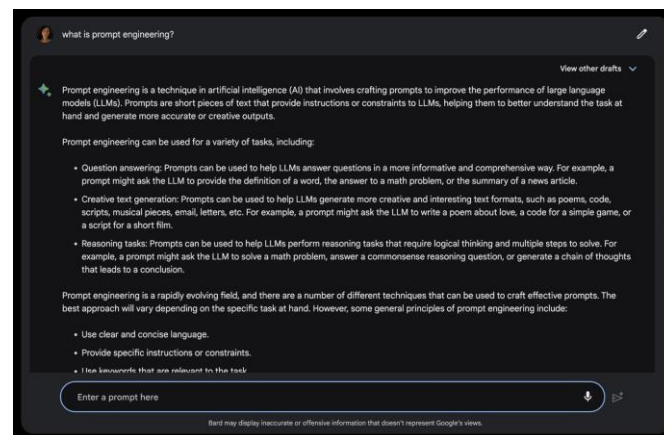
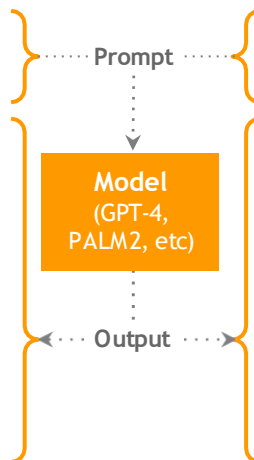
} Prompt {



Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

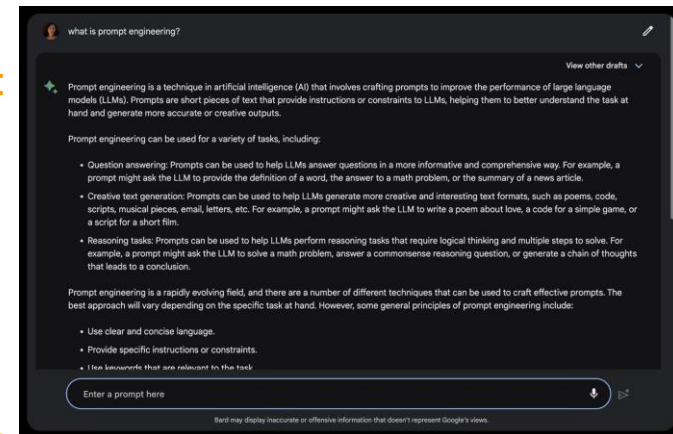
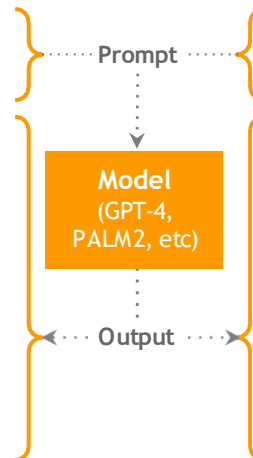
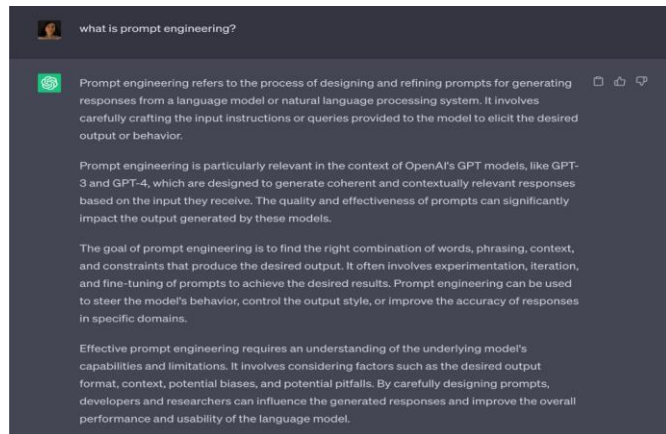
OK, but what is a **prompt**?



Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

OK, but what is a **prompt**?

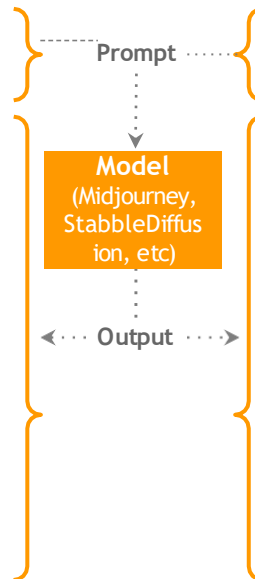


Emerging field in the ecosystem: Prompt Engineering

An AI technique that involves crafting the input instructions to help LLMs understand the task they need to perform and elicit the desired response.

OK, but what is a **prompt**?

watercolor painting of girl. lots of splashes and mistakes

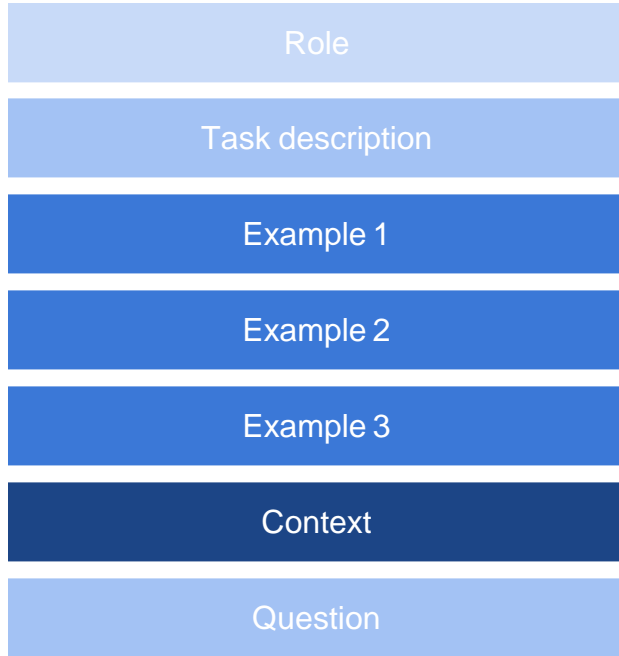


watercolor painting of girl. lots of splashes and mistakes



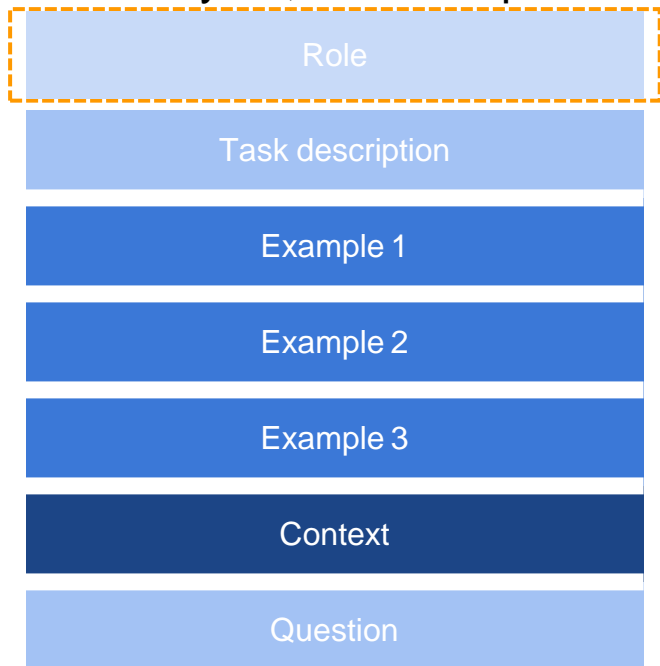
Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



The AI model is given a specific role to guide its responses based on the knowledge, experience or perspective of a given expert or character or persona. This technique can be helpful when you want the model to respond with information from a specific context or domain. For example:

As a financial advisor...

As a travel agent...

As a fun friend...

You are a brilliant mathematician...

I want you to act as a comedian...

You are an art historian specialised in ancient Egyptian art...

Note: in modern LLMs role prompting may not be as effective and could be skipped.

Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Role

Task description

Example 1

Example 2

Example 3

Context

Question

This is where we give instructions to the AI model. For example:

Remove personal identifiable information (PII) from the following text...

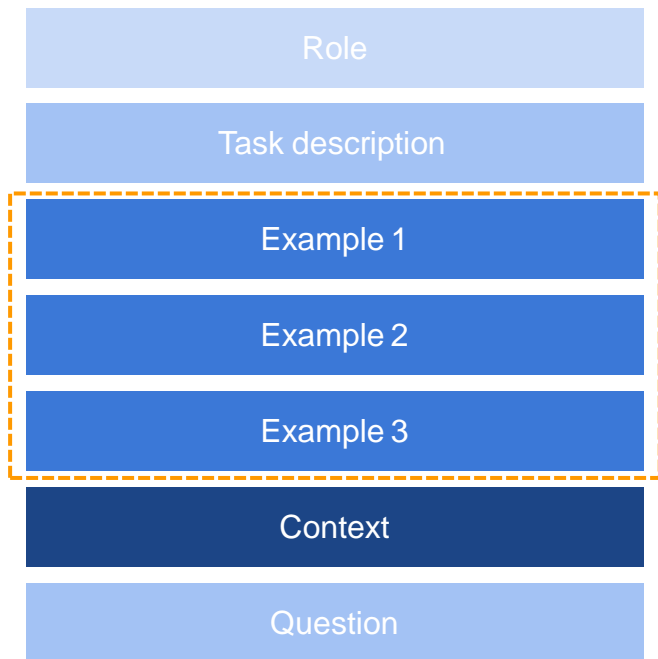
Classify this text...

Extract the data in this text and convert into a table...

Change the register of this text...

Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



For the model to better understand the task that we want them to complete, it's helpful to provide a few examples (called shots) of what we would like to see.

When we show the model a few examples this technique is called few-shot prompting, if we show one is 1-shot prompting, if we don't show any (most basic form of prompting) this is called zero-shot prompting.

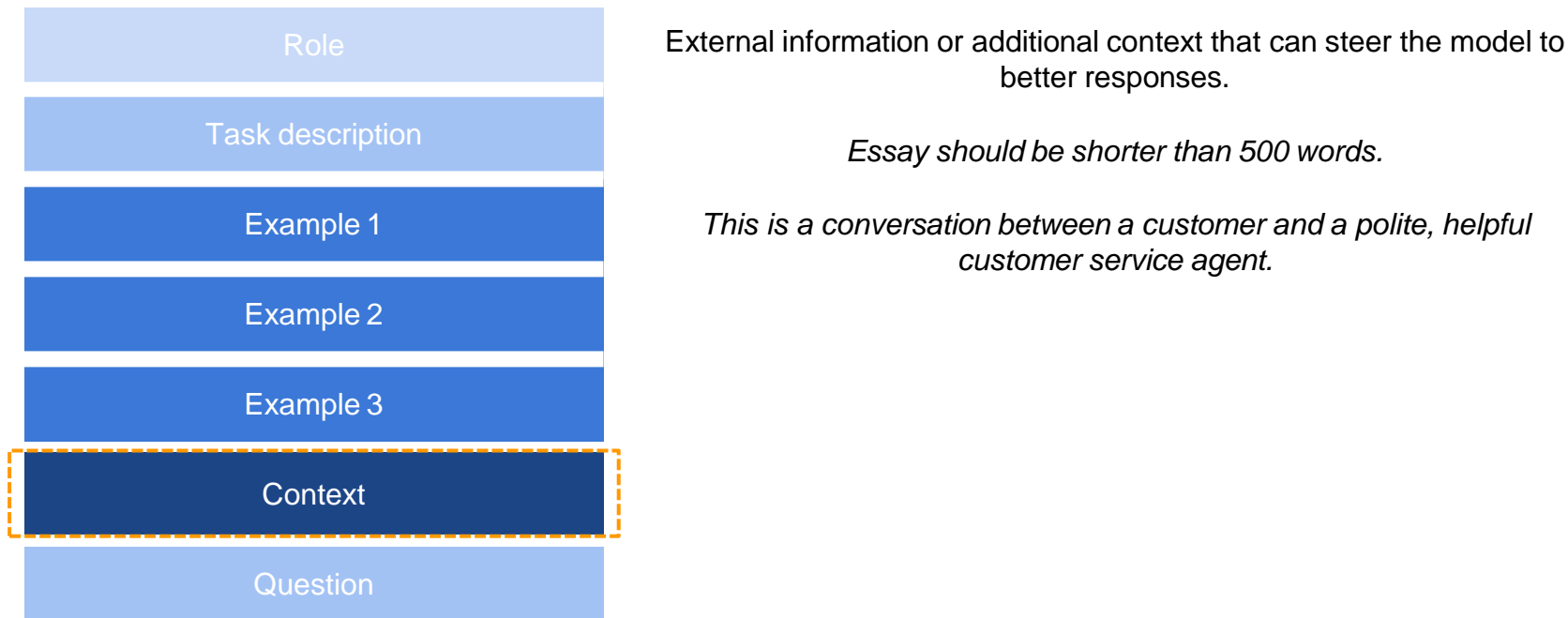
For example, in a sentiment classification task:

Q: Tweet: *"That's what I love about @salesforce. That it's about relationships and about caring about people and it's not only about business and money"*
Is this positive or negative?
A: positive

Q: Tweet: *"The more I use @salesforce the more I dislike it. It's slow and full of bugs. There are elements of the UI that look like they haven't been updated since 2006. Current frustration: app exchange pages won't stop refreshing every 10 seconds"*
Is this positive or negative?
A: negative

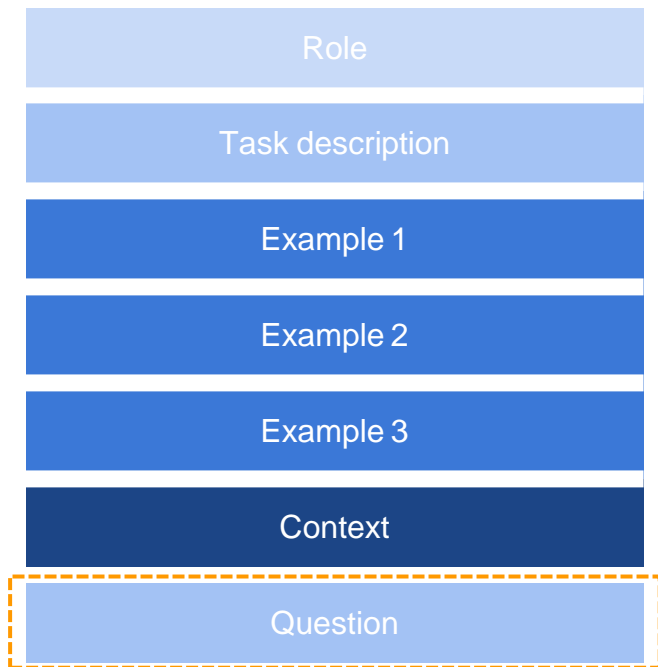
Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



Anatomy of a good prompt (text)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



Self-explanatory. This is just a question!. :)

Model parameters (text)

LLM outputs are affected by the configuration hyperparameters of the model, which control different aspects such as how “random” it is. By adjusting them you can influence the output to be more creative, diverse or interesting which control various aspects of the model, such as how 'random' it is.

Temperature

Temperature controls the randomness of the model output. A high temperature produces more unpredictable and creative results, while a low temperature produces more common and conservative outputs. For example, if you adjust the temperature to 0.5, the model will usually generate text that is more predictable and less creative than if you set the temperature to 1.0.

Top p

Top p (aka nucleus sampling), also controls the randomness of the model output. It sets a threshold probability and selects the top tokens whose cumulative probability exceeds the threshold. The model then randomly samples from this set of tokens to generate an output. For example, if you set top p to 0.9, the model will only consider the most likely words that make up 90% of the probability mass.

Best practices for (text) prompting

Start simple: Iterate adding more elements and context. Try multiple formulations (different words, contexts, examples) until you find what works best for your task.

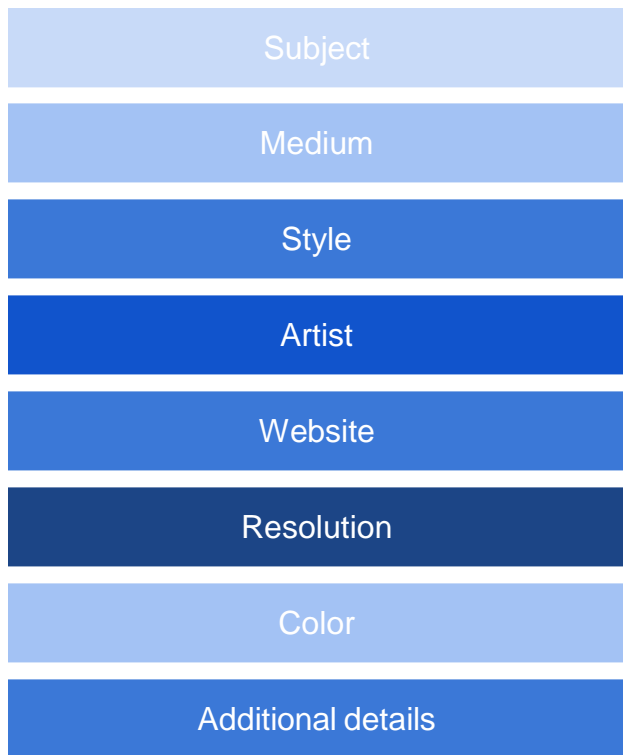
Be specific: the more descriptive and detailed the prompt is, the better the results.

Frame from the positive: avoid saying what not to do and say what to do instead

Be detailed: include specifics about desired response (eg. length, format, style). Include descriptors to tone or refine the output.

Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

If your desired image contains one or more subjects, describe with as much detail as possible. For example:

Medium

Style

Artist

Website

Resolution

Color

Additional details

A young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.



A fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

Resolution

Color

Additional details

Add detail on the medium the image you'd like (eg digital painting, photograph, oil painting).

Pointillism image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.



watercolour image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

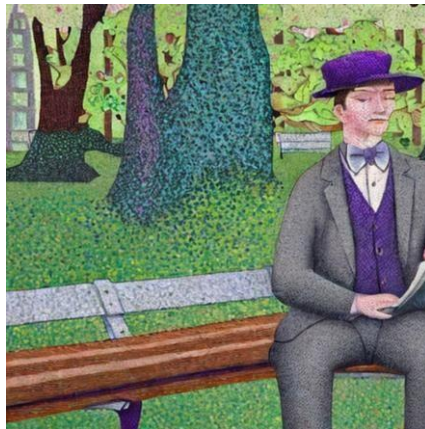
Resolution

Color

Additional details

Describe the style you'd like your image. In the absence of one, the model will choose one that has seen often in related images (pop-art, modernist, hyperrealistic)

art nouveau pointillism image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.



cubist watercolour image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

Resolution

Color

Additional details

You can mention an artist you would like your image to follow the traits of (eg if elegance you could use John Collier, for strong effects in portraits you could use Frida Kahlo). **Use sparingly: has a strong effect**

Monet style, art nouveau pointillism image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.



Picasso style cubist watercolour image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

Resolution

Color

Additional details

For niche images (eg fantasy or anime) there are specialised websites one can include in the prompt. But prompts need to be consistent (eg Picasso + Anime does not go well together)

***artstation** type image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.*



***Pixiv** style anime image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.*



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

Resolution

Color

Additional details

You can further refine the style with keywords related to resolution (eg sharp focus, 8k, vray)

***unreal engine** artstation type image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.*



***3D** Pixiv style anime image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils.*



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

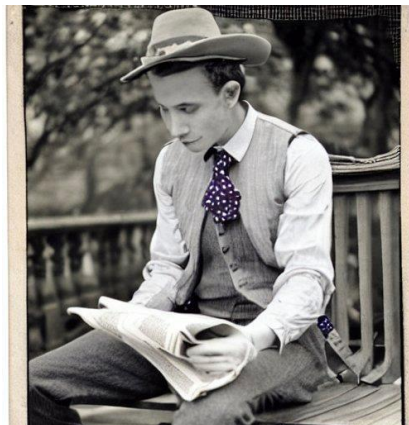
Resolution

Color

Additional details

You can further refine the color scheme of the image with keywords related to color (eg golden, palette of blue, splashes of)

***vintage** image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper.*



*image of a fox pup with bright orange fur and white face and tail tip sleeping curled up on a bed of daffodils under **golden hours iridescent light**.*



Anatomy of a good prompt (image)

A good prompt can contain the following elements. Not all appear in every prompt but if they do, it's not important that they follow a specific order.

Subject

Medium

Style

Artist

Website

Resolution

Color

Additional details

Additional descriptive attributes you'd like your image to be (eg ornate, cinematic lightning, low angle shot)

*unreal engine artstation type image of a young man dressed in a grey suit, white shirt and purple bow tie wearing a hat, sitting on a wooden bench in a highly manicured British-style park reading the newspaper **on a foggy day.***



*3D image of a **smiling** baby fox with bright orange fur **calmly** sleeping on a bed of daffodils, under golden hours iridescent light.*



Best practices for (image) prompting

Start simple: Iterate adding more elements and context. Try multiple formulations (different words, contexts, examples) until you find what works best for your task.

Be detailed: the more descriptive and detailed your descriptions are, the better the results. Use evocative language.

Be consistent: Keywords need to make a coherent matching in the prompt. E.g. photograph should not be used with van Gogh.

Use artist styles sparingly: they have a strong effect. Only choose a given artist if it can give a certain flavour to the image you're after.

PRACTICE TIME!

Exercise #1

1. Engineer prompts to generate images as close to the images below using stable diffusion: <https://www.bing.com/images/create>
1. Save a history will have to upload to [Drive](#) all the iterations of your prompts (and the resulting images) that led you to the image closest to at least 2 of the target images.

IMPORTANT!! You will **not** be able to reproduce the same image twice, so make sure keep a record where you save each prompt - outcome pair and iterate.

TIP: Pay attention to the details

TARGET IMAGES



Fill in here: https://docs.google.com/spreadsheets/d/1EMR9BBLJr-g_B1PnLjSZcSzvuuyHpgg/edit?usp=sharing&ouid=100949985522117931296&rtpof=true&sd=true

RAG vs. fine-tuning vs prompt engineering

Prompt engineering involves crafting queries to elicit better responses from the model without altering the model or its data sources.

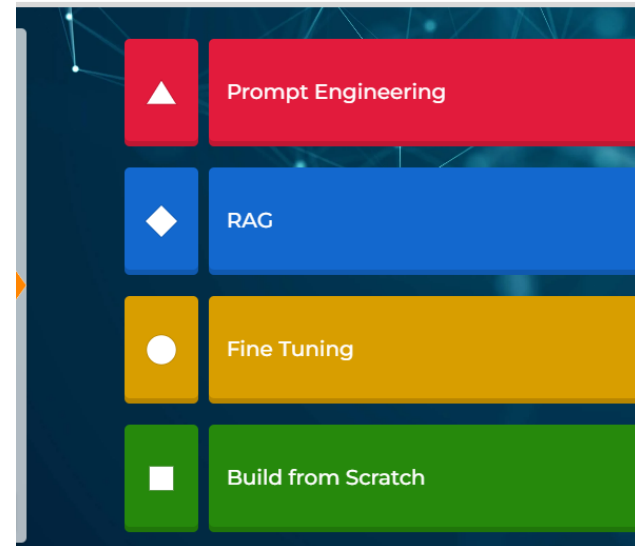
RAG involves augmenting an LLM with access to a dynamic, curated database to improve outputs.

Fine-tuning involves training an LLM on a smaller, specialized dataset to adjust its parameters for specific tasks.

Build from Scratch involves learning how to create, train, and tweak large language models (LLMs) by building one from the ground up!

[source](#)

from lowest to highest cost & complexity



CryptoGdelt -

A systematically create a diverse and relevant
curated cryptocurrency news database

<https://github.com/manoelgadi/CryptoGDelt2022>

CryptoGdelt

Extraction of 243,504 news articles from the Global Database of Events, Language, and Tone (GDELT) between March 31, 2021, and April 30, 2022, with keywords: cryptocurrency, cryptocurrencies, CBDC, Bitcoin, Ethereum, Litecoin, BitcoinCash, BitcoinSV, Polkadot, Chainlink, BinanceCoin, VeChain, Cosmos, Polkadot, NEO, Tezos, Tether, USDCoin, Monero, Dash, Zcash, Ripple, Cardano, Stellar, COUNOSX.

<https://github.com/manoelgadi/CryptoGDelt2022>

FinLin

In 2021, Daudert (2022) points out that two main challenges arise when applying sentiment analysis: relevance and point of view. First, relevance can be summarized as setting up common rules for dropping irrelevant news. Second, point of view refers to finding a common point of view during the annotation process. The second is particularly important for crypto since news regarding crypto mining may be positive for the miners but negative for the environment, generating inconsistent annotation if a point of view is not previously set.

[A multi-source entity-level sentiment corpus for the financial domain: the FinLin corpus | Language Resources and Evaluation \(springer.com\)](#)

GDELT

Extraction of 243504 news from GDELT between the 31st of March 2021 and the 30th of April 2022 containing any of the keywords:
cryptocurrency, cryptocurrencies,CBDC,Bitcoin, Ethereum, Litecoin, BitcoinCash, BitcoinSV, Polkadot, Chainlink, BinanceCoin, VeChain, Cosmos, Polkadot, NEO, Tezos, Tether, USDCoin, Monero, Dash, Zcash, Ripple, Cardano, Stellar, COUNOSX

relevance score

method: pipeline of TfidfVectorizer and MultinomialNB

data set: web scrapped Yahoo Finance
Crypto and non-crypto news / target: 1 news from <https://finance.yahoo.com/topic/crypto> and 0 news from <https://news.yahoo.com>

sentiment score

method: retraining FinBERT

data set: CryptoLinIE /
target: manual labelled by IE Business School students

strength (impact) score

method: pipeline of TfidfVectorizer and MultinomialNB

data set: Fama French /
target: Fama French Three factor

Deployment

Deployment of all three probability and classes to CryptoGDelt2022

Relevance Scores - Methodology

- Data: 1733 news articles over two weeks in June/2022, 1187 were unrelated news and 546 were cryptocurrency-related news.
- Methodology: Supervised model Naive Bayes to classify between cryptocurrency news and non-cryptocurrency news.
- Tokenization techniques to preprocess the headlines resulted in a 3.3 percentage point increase in classification performance in the test sample.

Relevance Scores - Results

- For the training set (1255 news articles), the model showed an accuracy of 97.84%
- For the test set (434 news articles), the model showed an accuracy of 91.70%
- The drop in performance in the test set may indicate overfitting, as a solution in the future we want to try Lasso/Ridge regularization techniques and compare with Transfer Learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) and Generative pre-trained transformer (GPT).
- <https://github.com/manoelgadi/CryptoGDelt2022/tree/master/Relevance>

Sentiment Scores - Methodology

- CryptoLinBERT was developed: a retraining of the Financial BERT algorithm using a Few Shot Learning strategy, Few Shot Learning (FSL), with the dataset created by us called CryptoLin.
- CryptoLin is a dataset composed of cryptocurrency-related news events,
- manually labeled with discrete values -1 (negative), 0 (neutral), and 1 (positive).
- CryptoLin was manually labelled on two occasions, with two groups of annotators, one with 27 Spanish computing students, and another with 113 finance students of various nationalities (3 groups with 26, 27, and 30 respectively).

Few Shot Learning

- **Definition:** Few-Shot Learning (FSL) is a machine learning framework where models learn to make accurate predictions with only a few labelled examples.
- **Applications:** Commonly used in scenarios where obtaining large amounts of labelled data is challenging, such as medical diagnosis, rare species identification, and consulting or regulation recommendations.
- **Techniques:** Utilizes methods like transfer learning, meta-learning, and data augmentation to enhance model performance with limited data.
- **Advantages:** Reduces the need for extensive labelled datasets, lowers training costs, and speeds up the learning process.
- **Challenges:** Ensuring model generalization with minimal data, handling class imbalance, and maintaining accuracy across diverse tasks.

Few Shot Learning – Manual Labelling Exercise

Exercise #2:

- Manual Label the news assigned to your student number - tab: News

Fill in here: https://docs.google.com/spreadsheets/d/1EMR9BBLJr-g_B1PnLjSZcSzvuuyHpgg/edit?usp=sharing&oid=100949985522117931296&rtpof=true&sd=true

Sentiment Scores - Results

- Three pre-trained sentiment analysis models were evaluated: Vader, Textblob, Flair, and our CryptoLinBERT. The results showed that non-financial sentiment algorithms performed poorly on this data.
- Our finetuned CryptoLinBERT achieved an overall accuracy of 92.49 % and 83.33 % on the training and test datasets respectively.

<https://github.com/manoelgadi/CryptoGDelt2022/blob/master/Sentiment/ToReproduceSentiment.ipynb>

Algorithmic Bias and Fairness

We live in a biased world. Training data for LLMs is reflective of our world. You learned in the class pre-work that bias get replicated with serious consequences, from perpetuating harmful stereotypes to reinforcing systemic inequalities and discrimination.





Bias in Prompting

Try it yourself at: <https://huggingface.co/spaces/society-ethics/DiffusionBiasExplorer>

‘ambitious office worker’

Diffusion Bias Explorer

Choose from the prompts below to explore how the text-to-image models like [Stable Diffusion v1.4](#), [Stable Diffusion v2](#) and [DALL-E-2](#) represent different professions and adjectives

Choose a model to compare results	Choose a model to compare results
Stable Diffusion 1.4	Stable Diffusion 1.4
Choose a first adjective (or leave this blank)	Choose a second adjective (or leave this blank)
ambitious	compassionate
Choose a first group	Choose a second group
office worker	office worker
0 Images	0 Images
	

‘compassionate office worker’

Group Discussion

Discussion #1:

- How to agree on point of view for labelling and avoid bias?

Consensus Mechanism and Inter-rate reliability coefficients

Set at least 3 annotator, define consensus table and apply agreement metrics.

annotator 1	annotator 2	annotator 3	decision	reasoning	count
-1	-1	-1	-1	complete majority	380
-1	-1	0	-1	majority with no objection	3
-1	-1	1	0	minority report, majority with one objection	2
-1	0	-1	-1	majority with no objection	0
-1	0	0	0	majority	6
-1	0	1	0	total disagreement	0
-1	1	-1	0	minority report, majority with one objection	1
-1	1	0	0	total disagreement	2
-1	1	1	0	minority report, majority with one objection	2
0	-1	-1	-1	majority with no objection	2
0	-1	0	0	majority	3
0	-1	1	0	total disagreement	3
0	0	-1	0	majority	6
0	0	0	0	complete majority	864
0	0	1	0	majority	13
0	1	-1	0	total disagreement	1
0	1	0	0	majority	13
0	1	1	1	majority with no objection	17
1	-1	-1	0	minority report, majority with one objection	2
1	-1	0	0	total disagreement	3
1	-1	1	0	minority report, majority with one objection	2
1	0	-1	0	total disagreement	1
1	0	0	1	majority	15
1	0	1	1	majority with no objection	21
1	1	-1	0	minority report, majority with one objection	3
1	1	0	1	majority with no objection	13
1	1	1	1	complete majority	1305
					2683

Table 5: Consensus table. Column decision indicates the final manual labelling assigned depending on annotation given by annotator 1, 2 and 3. Column reasoning gives an explanation of the decision. Note minority report to default the labelling to neutral in case of one objection is used. Count represents the number of occurrences in the data set.

Metric	Coeff (1-2)	Coeff (1-3)	Coeff (2-3)
Fleiss' κ	0.942	0.942	0.944
Kappa's StdErr	0.006	0.006	0.006
Kappa's 95% C.I.	(0.958, 0.972)	(0.958, 0.972)	(0.96, 0.973)
Krippendorff's α	0.942	0.942	0.944
Gwet's AC1	0.9499	0.9499	0.952
Landis and Koch [31] benchmark	Almost Perfect	Almost Perfect	Almost Perfect
Fleiss [27] benchmark	Excellent	Excellent	Excellent
Altman [32] benchmark	Very Good	Very Good	Very Good
Cicchetti [33] benchmark	Excellent	Excellent	Excellent

Table 7: Fleiss's Kappa, Krippendorff's Alpha and G-efficients annotators 1 and 2 (Coeff (1-2)), 1 and 3 (Coeff (1-3)) and 2 and 3 (Coeff (2-3)) and its benchmark interpretation according to Landis [31] and Cicchetti [33] benchmarks

Landis and Koch [31] benchmark range	Landis-Koch benchmark interpretation
$0.00 \leq \kappa < 0.2$:	Slight
$0.20 \leq \kappa < 0.4$:	Fair
$0.40 \leq \kappa < 0.6$:	Moderate
$0.60 \leq \kappa < 0.8$:	Substantial
$0.80 \leq \kappa \leq 1.0$:	Almost Perfect
Fleiss [27] benchmark range	Fleiss benchmark interpretation
$\kappa < 0.4$:	Poor
$0.40 \leq \kappa < 0.75$:	Intermediate to Good
$0.75 \leq \kappa \leq 1.0$:	Excellent
Altman [32] benchmark range	Altman benchmark interpretation
$0.00 \leq \kappa < 0.2$:	Poor
$0.20 \leq \kappa < 0.4$:	Fair
$0.40 \leq \kappa < 0.6$:	Moderate
$0.60 \leq \kappa < 0.8$:	Good
$0.80 \leq \kappa \leq 1.0$:	Very Good
Cicchetti [33] benchmark range	Cicchetti benchmark interpretation
$0.00 \leq \kappa < 0.4$:	Poor
$0.40 \leq \kappa < 0.59$:	Fair
$0.59 \leq \kappa < 0.74$:	Good
$0.74 \leq \kappa \leq 1.0$:	Excellent

Table 8: Interpretation of Kappa metric using number with Landis-Koch, Fleiss, Altman, Cicchetti benchmarks

TABLE 7. Fleiss's Kappa, Krippendorff's Alpha, and Gwet's AC1 inter-rater reliability coefficients annotators 1 and 2 (Coeff (1-2)), 1 and 3 (Coeff (1-3)) and 2 and 3(Coeff (2-3)).

Annotators	Metric	Coeff (1-2)	Coeff (1-3)	Coeff (2-3)
UTAD	Fleiss' κ	0.754	0.747	0.775
UTAD	Kappa's StdErr	0.011	0.012	0.011
UTAD	Kappa's 95% C.I.	(0.839, 0.865)	(0.834, 0.861)	(0.852, 0.878)
UTAD	Krippendorff's α	0.753	0.747	0.775
UTAD	Gwet's AC1	0.7886	0.782	0.8064
IE	Fleiss' κ	0.942	0.942	0.944
IE	Kappa's StdErr	0.006	0.006	0.006
IE	Kappa's 95% C.I.	(0.958, 0.972)	(0.958, 0.972)	(0.96, 0.973)
IE	Krippendorff's α	0.942	0.942	0.944
IE	Gwet's AC1	0.9499	0.9499	0.952

TABLE 3. IE University annotators by educational backgrounds. The table has three columns: Row Labels, count, and percentage. The first column lists the education of the people in the group. The second column shows how many people belong to each educational background. The third column shows what percentage of the total number of people in the group each educational background represents.

Row Labels	count	percentage
Accounting	3	3.61%
Aeronautics	1	1.20%
Arts	4	4.82%
Biology	1	1.20%
Business Administration	15	18.07%
Business Engineering	4	4.82%
Chemical Engineering	1	1.20%
Civil Engineering	2	2.41%
Commerce	1	1.20%
Commercial Engineering	1	1.20%
Computer Engineering	8	9.64%
Computer Science	7	8.43%
Economics	1	1.20%
Engineering	3	3.61%
Finance	1	1.20%
Human Resource	3	3.61%
Industrial Engineering	1	1.20%
International Business	1	1.20%
Law	1	1.20%
Marketing	2	2.41%
MBA / Master in Management	10	12.05%
Politics	1	1.20%
Science	7	8.43%
Social Science	1	1.20%
Teaching	2	2.41%
Tourism	1	1.20%

TABLE 2. IE University annotators by nationality. The table has three columns: Row Labels, count, and percentage. The first column lists the nationalities of the people in the group. The second column shows how many people belong to each nationality. The third column shows what percentage of the total number of people in the group each nationality represents.

Row Labels	count	percentage
American	2	2.41%
Argentinian	2	2.41%
Belgian	2	2.41%
Canadian	1	1.20%
Chilean	1	1.20%
Chinese	2	2.41%
Colombian	6	7.23%
Costa Rican	1	1.20%
Dominican	3	3.61%
Dutch	1	1.20%
Ecuadorian	3	3.61%
French	4	4.82%
German	10	12.05%
Greek	1	1.20%
Indian	2	2.41%
Italian	5	6.02%
Kazakhstani	1	1.20%
Kenyan	1	1.20%
Lebanese	2	2.41%
Mexican	3	3.61%
Panamanian	1	1.20%
Philippines	1	1.20%
Portuguese	2	2.41%
Saudi Arabian	2	2.41%
South African	1	1.20%
Spanish	18	21.69%
Swiss	3	3.61%
Syrian	1	1.20%
Thai	1	1.20%

Sentiment Scores - Results

- The results indicate that the manual labeling performed by both groups was satisfactory according to the inter-rater reliability coefficients Fleiss's Kappa, Krippendorff's Alpha, and Gwet's AC1. However, the group with greater diversity has better inter-rater reliability coefficients.
- These results have been published in two articles: A sentiment corpus for the cryptocurrency financial domain: the CryptoLin corpus [9] and Annotators' Selection Impact on the Creation of a Sentiment Corpus for the Cryptocurrency Financial Domain [10].
- https://github.com/manoelgadi/CryptoLin/blob/main/CryptoLinIE_analysis_for_reproducibility_purposes.ipynb

Strength (Impact) Scores - Methodology

- Ability of an individual news article or a group of news articles to provoke changes in the cryptocurrency ecosystem, in this study positive or negative abnormal returns were considered.
- Three-factor Fama-French model to evaluate abnormal returns of Bitcoin.
- 7-day time interval, 2 days prior, the day of the news publication, and the 4 days after.
- The target variable was the labeling of positive abnormal, negative abnormal, or normal (neutral) given by Fama-French 3 factors.

Strength (Impact) Scores - Results

- The first attempt with a simple Naïve Bayes classifier with default parameters as the NLP model was trained with daily Bitcoin price data from Yahoo! Finance from March 12, 2022, to April 7, 2022.
- The NLP model accuracy was low, reaching only 64.35 % on the test dataset.
- More research is needed to propose 3 cryptocurrency series to replace the original series of the Fama-French model. The study conducted by Rubén Martínez Conejo during his Master's Thesis at the University of Alcalá confirmed the poor fit of the Fama-French 3-factor model to cryptocurrencies.
- Similarly, more research is needed in the selection of the method comparing Transfer Learning techniques such as Bidirectional Encoder Representations from Transformers (BERT) and Generative pre-trained transformer (GPT).
- <https://github.com/manoelgadi/CryptoGDelt2022/tree/master/Strength>

Relevance Scores - Methodology

- Data: 1733 news articles over two weeks in June/2022, 1187 were unrelated news and 546 were cryptocurrency-related news.
- Methodology: Supervised model Naive Bayes to classify between cryptocurrency news and non-cryptocurrency news.
- Tokenization techniques to preprocess the headlines resulted in a 3.3 percentage point increase in classification performance in the test sample.

Best Practice in Labelling

Define	Define Clear Guidelines: Establish detailed instructions for labelers to ensure consistency and accuracy.
Training	Training and Calibration: Regularly train labelers and conduct calibration sessions to align their understanding and application of labeling criteria.
Implement	Quality Control: Implement a robust quality control process, including consensus table and inter-rater reliability coefficients.
Use	Use of Annotation Tools: Leverage advanced annotation tools that support efficient and accurate labeling, such as those with built-in error-checking features.
Provide	Iterative Feedback: Provide continuous feedback to labelers to help them improve and correct mistakes promptly.
Labeling	Labeling Environment: Create a conducive labeling environment that minimizes distractions and supports focus.
Maintain	Documentation: Maintain comprehensive documentation of the labeling process, including any changes to guidelines or criteria.
Ensure	Ethical Considerations: Ensure that the labeling process respects privacy and ethical standards, especially when dealing with sensitive data.

Best Practice in Labeling

Define

Define Clear Guidelines: Establish detailed instructions for labelers to ensure consistency and accuracy.

Define the point of view.

Training

Training and Calibration: Regularly train labelers and conduct calibration sessions to align their understanding and application of labeling criteria.

In 2021, Daudert [17] points out that two main challenges arise when applying sentiment analysis: relevance and point of view. First, **relevance** can be summarized as setting up common rules for dropping irrelevant news. Second, **point of view** refers to finding a common point of view during the annotation process. The second is particularly important for crypto since news regarding crypto mining may be positive for the miners but negative for the environment, generating inconsistent annotation if a point of view is not previously set.

Best Practice in Labelling

Implement

Quality Control: Implement a robust quality control process, including consensus table and inter-rater reliability coefficients.

Set at least 3 annotator, define consensus table and apply agreement metrics.

annotator 1	annotator 2	annotator 3	decision	reasoning	count
-1	-1	-1	-1	complete majority	380
-1	-1	0	-1	majority with no objection	3
-1	-1	1	0	minority report, majority with one objection	2
-1	0	-1	-1	majority with no objection	0
-1	0	0	0	majority	6
-1	0	1	0	total disagreement	0
-1	1	-1	0	minority report, majority with one objection	1
-1	1	0	0	total disagreement	2
-1	1	1	0	minority report, majority with one objection	2
0	-1	-1	-1	majority with no objection	2
0	-1	0	0	majority	3
0	-1	1	0	total disagreement	3
0	0	-1	0	majority	6
0	0	0	0	complete majority	864
0	0	1	0	majority	13
0	1	-1	0	total disagreement	1
0	1	0	0	majority	13
0	1	1	1	majority with no objection	17
1	-1	-1	0	minority report, majority with one objection	2
1	-1	0	0	total disagreement	3
1	-1	1	0	minority report, majority with one objection	2
1	0	-1	0	total disagreement	1
1	0	0	0	majority	15
1	0	1	1	majority with no objection	21
1	1	-1	0	minority report, majority with one objection	3
1	1	0	1	majority with no objection	13
1	1	1	1	complete majority	1305
					2683

Table 5: Consensus table. Column decision indicates the final manual labelling assigned depending on annotation given by annotator 1, 2 and 3. Column reasoning gives an explanation of the decision. Note minority report to default the labelling to neutral in case of one objection is used. Count represents the number of occurrences in the data set.

Metric	Coeff (1-2)	Coeff (1-3)	Coeff (2-3)
Fleiss' κ	0.942	0.942	0.944
Kappa's StdErr	0.006	0.006	0.006
Kappa's 95% C.I.	(0.958, 0.972)	(0.958, 0.972)	(0.96, 0.973)
Krippendorff's α	0.942	0.942	0.944
Gwet's AC1	0.9499	0.9499	0.952
Landis and Koch [31] benchmark	Almost Perfect	Almost Perfect	Almost Perfect
Fleiss [27] benchmark	Excellent	Excellent	Excellent
Altman [32] benchmark	Very Good	Very Good	Very Good
Cicchetti [33] benchmark	Excellent	Excellent	Excellent

Table 7: Fleiss's Kappa, Krippendorff's Alpha and Gwet's AC1 inter-rater reliability coefficients annotators 1 and 2 (Coeff (1-2)), 1 and 3 (Coeff (1-3)) and 2 and 3 (Coeff (2-3)) and its benchmark interpretation according to Landis and Koch [31], Fleiss [27], Altman [32] and Cicchetti [33] benchmarks

Best Practice in Labelling

Use

Develop an internal Annotation Tool or Use of Annotation Tools: Leverage advanced annotation tools that support efficient and accurate labeling, such as those with built-in error-checking features.

Here are some popular annotation tools available in the market:

1. ClickUp: Great for project management, it offers various annotation features for documents and images.
2. Filestage: Used for reviewing and approving electronic media, it supports collaborative annotations.
3. Prodigy: A tool often used by machine learning professionals for text and image annotation.
4. Annotate: Known for its robust text and image annotation capabilities.
5. Markup.io: Enables real-time annotations on images, websites, and design files.
6. Pastel: Allows users to annotate websites and share feedback easily.
7. Zoho Annotator: Part of the Zoho suite, it offers comprehensive annotation features.
8. Markup Hero: A versatile tool for annotating images, PDFs, and more.
9. GoVisually: Focuses on visual content, allowing for detailed feedback and annotations.
10. Frame.io: Primarily used for video annotation and collaboration.

Top 8 Annotation Tools (Features, Pros, Cons, Pricing). <https://www.spotsaas.com/blog/top-8-annotation-tools-features-pros-cons/>.

Best Practice in Labelling

Provide

Iterative Feedback: Provide continuous feedback to labelers to help them improve and correct mistakes promptly.

Labeling

Labeling Environment: Create a conducive labeling environment that minimizes distractions and supports focus.

Maintain

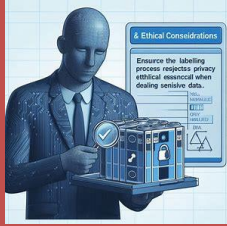
Documentation: Maintain comprehensive documentation of the labeling process, including any changes to guidelines or criteria.

Landis and Koch [31] benchmark range	Landis-Koch benchmark interpretation
$0.00 \leq \kappa < 0.2$:	Slight
$0.20 \leq \kappa < 0.4$:	Fair
$0.40 \leq \kappa < 0.6$:	Moderate
$0.60 \leq \kappa < 0.8$:	Substantial
$0.80 \leq \kappa \leq 1.0$:	Almost Perfect
Fleiss [27] benchmark range	Fleiss benchmark interpretation
$\kappa < 0.4$:	Poor
$0.40 \leq \kappa < 0.75$:	Intermediate to Good
$0.75 \leq \kappa \leq 1.0$:	Excellent
Altman [32] benchmark range	Altman benchmark interpretation
$0.00 \leq \kappa < 0.2$:	Poor
$0.20 \leq \kappa < 0.4$:	Fair
$0.40 \leq \kappa < 0.6$:	Moderate
$0.60 \leq \kappa < 0.8$:	Good
$0.80 \leq \kappa \leq 1.0$:	Very Good
Cicchetti [33] benchmark range	Cicchetti benchmark interpretation
$0.00 \leq \kappa < 0.4$:	Poor
$0.40 \leq \kappa < 0.59$:	Fair
$0.59 \leq \kappa < 0.74$:	Good
$0.74 \leq \kappa \leq 1.0$:	Excellent

Table 8: Interpretation of Kappa metric using number with Landis-Koch, Fleiss, Altman, Cicchetti benchmarks

Best Practice in Labelling

Ensure



Ethical Considerations: Ensure that the labeling process **avoid biases**, respects privacy and ethical standards, especially when dealing with sensitive data. Narrow annotators -> risk of bias!

TABLE 7. Fleiss's Kappa, Krippendorff's Alpha, and Gwet's AC1 inter-rater reliability coefficients annotators 1 and 2 (Coeff (1-2)), 1 and 3 (Coeff (1-3)) and 2 and 3 (Coeff (2-3)).

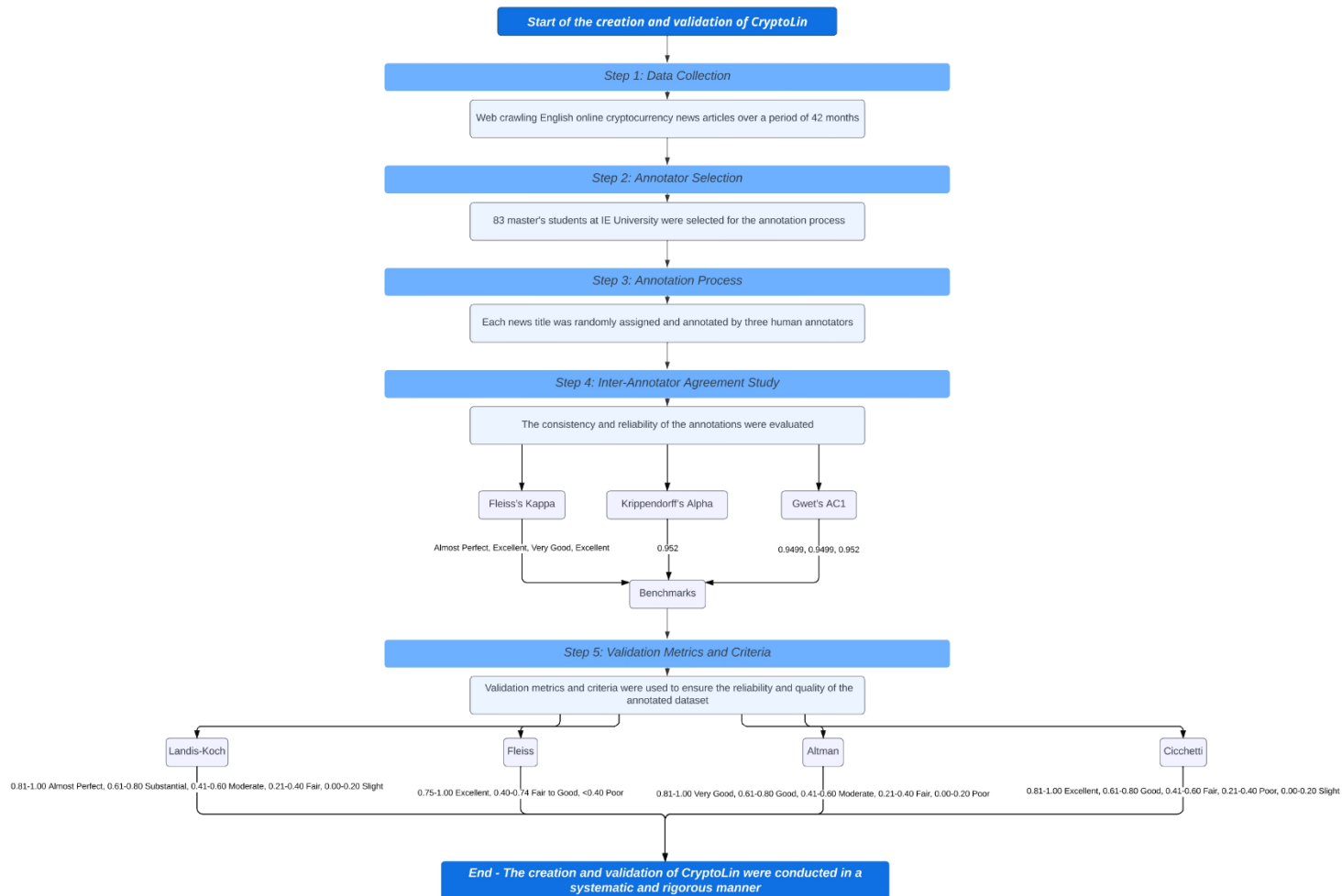
Annotators	Metric	Coeff (1-2)	Coeff (1-3)	Coeff (2-3)
UTAD	Fleiss' κ	0.754	0.747	0.775
UTAD	Kappa's StdErr	0.011	0.012	0.011
UTAD	Kappa's 95% C.I.	(0.839, 0.865)	(0.834, 0.861)	(0.852, 0.878)
UTAD	Krippendorff's α	0.753	0.747	0.775
UTAD	Gwet's AC1	0.7886	0.782	0.8064
IE	Fleiss' κ	0.942	0.942	0.944
IE	Kappa's StdErr	0.006	0.006	0.006
IE	Kappa's 95% C.I.	(0.958, 0.972)	(0.958, 0.972)	(0.96, 0.973)
IE	Krippendorff's α	0.942	0.942	0.944
IE	Gwet's AC1	0.9499	0.9499	0.952

TABLE 3. IE University annotators by educational backgrounds. The table has three columns: Row Labels, count, and percentage. The first column lists the education of the people in the group. The second column shows how many people belong to each educational background. The third column shows what percentage of the total number of people in the group each educational background represents.

Row Labels	count	percentage
Accounting	3	3.61%
Aeronautics	1	1.20%
Arts	4	4.82%
Biology	1	1.20%
Business Administration	15	18.07%
Business Engineering	4	4.82%
Chemical Engineering	1	1.20%
Civil Engineering	2	2.41%
Commerce	1	1.20%
Commercial Engineering	1	1.20%
Computer Engineering	1	1.20%
Computer Science	8	9.64%
Economics	7	8.43%
Engineering	1	1.20%
Finance	3	3.61%
Human Resource	1	1.20%
Industrial Engineering	3	3.61%
International Business	1	1.20%
Law	1	1.20%
Marketing	2	2.41%
MBA / Master in Management	10	12.05%
Politics	1	1.20%
Science	7	8.43%
Social Science	1	1.20%
Teaching	2	2.41%
Tourism	1	1.20%

TABLE 2. IE University annotators by nationality. The table has three columns: Row Labels, count, and percentage. The first column lists the nationalities of the people in the group. The second column shows how many people belong to each nationality. The third column shows what percentage of the total number of people in the group each nationality represents.

Row Labels	count	percentage
American	2	2.41%
Argentinian	2	2.41%
Belgian	2	2.41%
Canadian	1	1.20%
Chilean	1	1.20%
Chinese	2	2.41%
Colombian	6	7.23%
Costa Rican	1	1.20%
Dominican	3	3.61%
Dutch	1	1.20%
Ecuadorian	3	3.61%
French	4	4.82%
German	10	12.05%
Greek	1	1.20%
Indian	2	2.41%
Italian	5	6.02%
Kazakhstan	1	1.20%
Kenyan	1	1.20%
Lebanese	2	2.41%
Mexican	3	3.61%
Panamanian	1	1.20%
Philippines	1	1.20%
Portuguese	2	2.41%
Saudi Arabian	2	2.41%
South African	1	1.20%
Spanish	18	21.69%
Swiss	3	3.61%
Syrian	1	1.20%
Thai	1	1.20%



Best Practice in Labeling

Based on articles:

“Annotators’ Selection Impact on the Creation of a Sentiment Corpus for the Cryptocurrency Financial Domain”

<https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=10322757>

And

“A sentiment corpus for the cryptocurrency financial domain: the CryptoLin corpus”

<https://link.springer.com/article/10.1007/s10579-024-09743-x>

Kahoot



**IE University - ESMA -
Generative AI Data
Foundation**

<https://create.kahoot.it/details/39eb5788-4e56-470d-9897-2cc87449dd4f>

Manoel Gadi

Professor of the School of Science and Technologies

mfalonso@faculty.ie.edu

Madrid

Paseo de la Castellana, 259.
28046 Madrid, Spain

+34 915 689 600

Segovia

Cardenal Zúñiga, 12
40003 Segovia, Spain

+34 921 412 410

university@ie.edu
www.ie.edu/university

Contact us



@ieuniversity

Group Discussion

Discussion #2:

- How to identify a chatbot is hallucinating, making up stats, or lying about something?

Group Discussion

Discussion #3:

- If the chatbot provides references, how to verify it is not creating or spreading miss information?

Group Discussion

Discussion #4:

Watch the video:

<https://www.youtube.com/watch?v=oXRFOxMq7e4>

- In groups, design a framework of validation and auditing that would make “Artificial Justice” reliable.

How to train the AI? How to validate? How to Audit it? How to make sure the new system can deal with new jurisdictions?