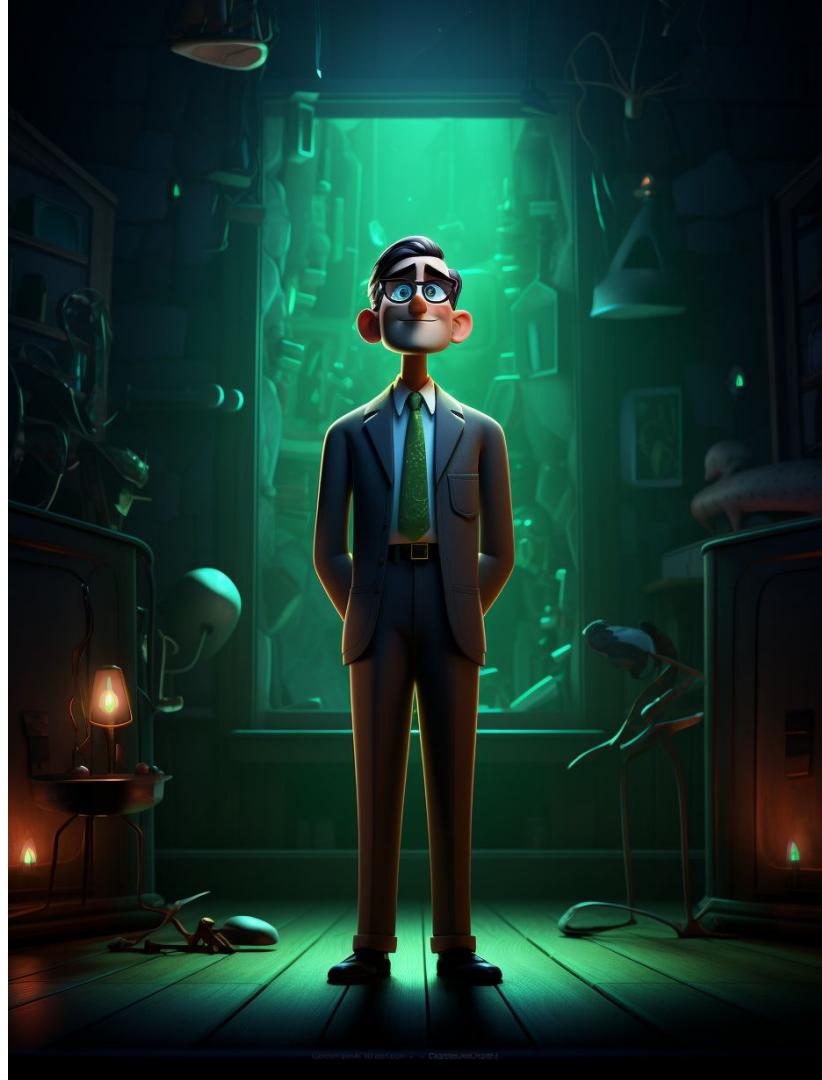


Building LLM Powered Solutions

Module 2:
Getting deeper into
LLMs and ML System
Design

Hamza Farooq



Recap from Module -1

- What are foundational models?
- A larger overview of application of LLMs
- Working Prototypes

Learning outcomes

- What can Gen AI do for you?
- The world of NLP
- Re-Introduction to LLMs
- Transformer & General Architecture
- Introduction to ML System Design
- Future ML Architecture

The main question

What can Gen AI do for you and your organization?

Why
Uber ?



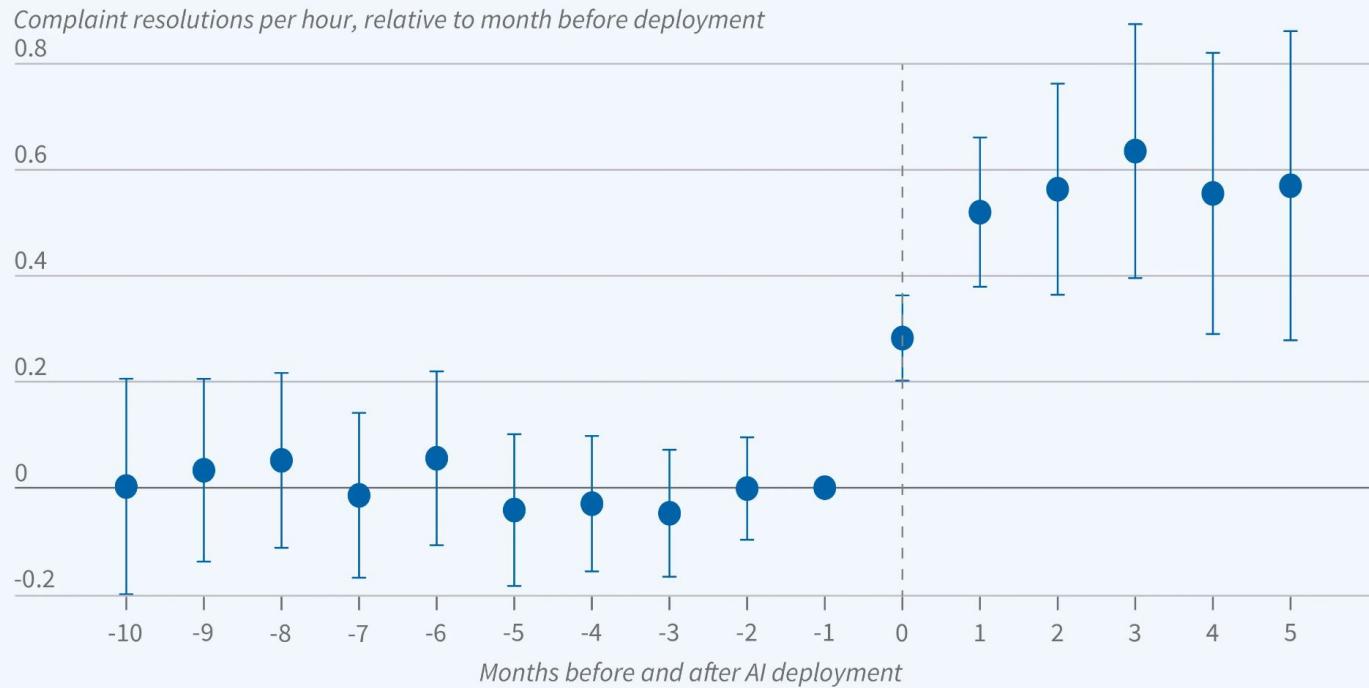


[Source](#)

What can Gen AI do for you and your organization?

One word: Productivity

AI Assistance and Customer Complaint Resolutions



Thin bars represent 95% confidence intervals

Source: Researchers' calculations using data from customer support agents provided by a Fortune 500 enterprise software company



[Source](#)

The value of context in NLP Models

The goal of natural language processing (NLP) is to find answers to four questions:

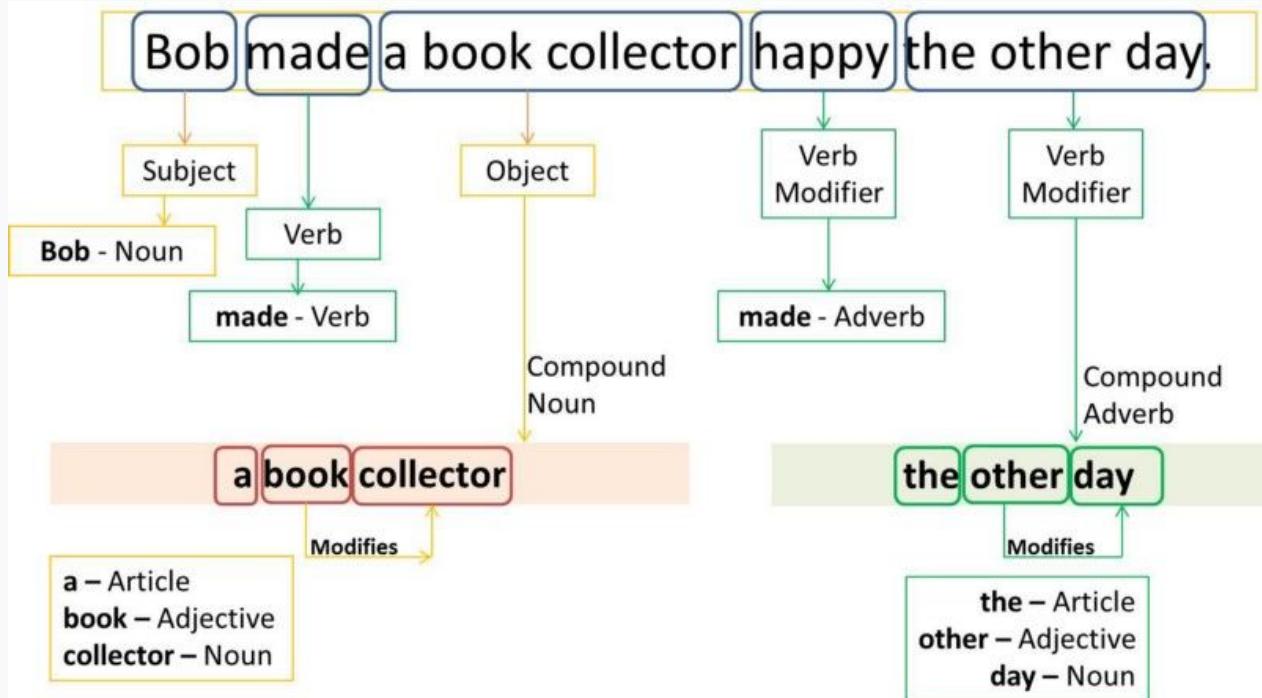
- **Who** is talking?
- **What** are they talking about?
- **How** do they feel?
- **Why** do they feel that way?

This last question is a question of **context**

Content vs Context

Content is the material/matter/medium contained within the work that's available for the audience.

Context is the positioning of the content, the storyline or the purpose that provides value to the audience.



01

The role on NLP in our Lives

We live in a
world of NLP

Natural Language Processing

NEW YORK NEWS
Articles for you



What is NLP anyways?

Natural Language Processing (NLP) is defined as the branch of Artificial Intelligence that provides computers with the capability of understanding text and spoken words, in the same way a human being can.

It incorporates machine learning models, statistics, and deep learning models into computational linguistics i.e. rule-based modeling of human language to allow computers to understand text, spoken words and understands human language, intent, and sentiment.

Applications – 1

- Information retrieval
- Information extraction
- Question answering

Google list of good sushi restaurant in nyc

About 505,000,000 results (1.29 seconds)

The search results page shows a map of New York City with several red dots indicating sushi restaurants. Below the map are three cards for Sushi Nakazawa, Sushi Yasuda, and Blue Ribbon Sushi, each with a rating, address, closing time, and a small image of the restaurant.

Sushi Nakazawa
4.7 ★★★★★ (1,038) - \$\$\$ - Sushi
23 Commerce St
Closes soon · 11PM
Dine-in · No takeout · No delivery

Sushi Yasuda
4.4 ★★★★★ (1,119) - \$\$\$ - Japanese
204 E 43rd St
Closes soon · 11PM
Good sushi, but over priced

Blue Ribbon Sushi
4.5 ★★★★★ (1,193) - \$\$ - Sushi
119 Sullivan St
Closes soon · 11PM
Good sushi, extensive menu.

View all

why is the sky blue

All Books Videos Images News More Tools

Child Bill Nye Adult Daddy

The search results page includes a diagram illustrating light dispersion through a prism and a text snippet from Scientific American explaining Rayleigh scattering.

Thus, as sunlight of all colors passes through air, the blue part causes charged particles to oscillate faster than does the red part. ... More of the sunlight entering the atmosphere is blue than violet, however, and our eyes are somewhat more sensitive to blue light than to violet light, so the sky appears blue. Apr 7, 2003

<https://www.scientificamerican.com/article/why-is-the-sky-blue/>

Why is the sky blue? - Scientific American

About featured snippets · Feedback

People also ask

Why is the sky blue short answer?
Is the sky blue because of the ocean?
Why is the sky blue explain to a child?
What is the reason the sky looks blue?

Feedback

Applications –2

- Machine Translation
- Summarization
- Auto Completion
- Spell Correction

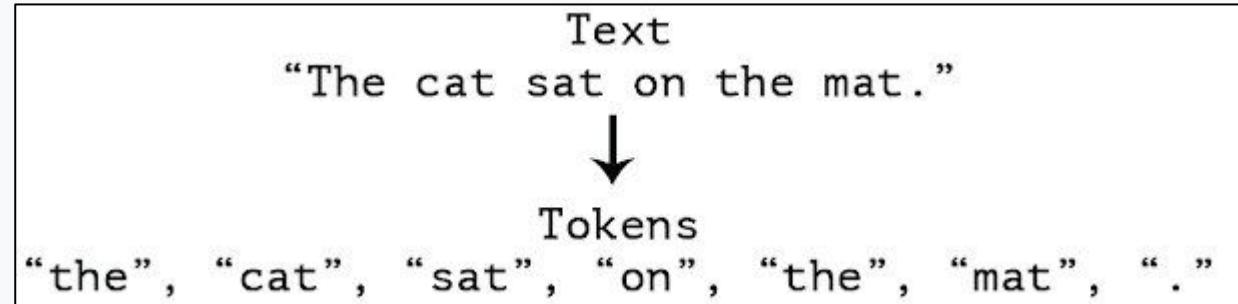
Many More...

The image displays three separate application windows side-by-side.

- Google Translate:** A screenshot of the Google Translate interface. It shows a text input field with placeholder text "Translation" and a character count of "0/5000". Above the input field, language selection tabs are visible: DETECT LANGUAGE, ENGLISH, SPANISH, FRENCH, ENGLISH (selected), SPANISH, ARABIC. Below the input field are three circular icons: History, Saved, and Community.
- Microsoft Word:** A screenshot of the Microsoft Word ribbon. The "AutoCorrect" tab is selected, showing a list of suggestions: "unveiling", "unfeeling", "availing", and "unavailing". There are also options to "Add to dictionary" and "Link...".
- Messaging Application:** A screenshot of a messaging window titled "Taco Tuesday". The message body contains the text: "Hey there. Don't forget about Taco Tuesday! I'll bring the chips and salsa".

How does a machine understand what we are saying?

- Tokenization
 - Named Entity Recognition (NER)
 - Text representation
 - Text classification
 - Natural language generation
 - Multimodal NLP
- Tokenization is the process of breaking down a text into individual units called tokens.
 - Tokens are typically words, but can also be phrases or even individual characters, depending on the application.
 - Tokenization is a crucial step in natural language processing tasks such as machine translation, sentiment analysis, and named entity recognition.



Common NLP tasks

- Tokenization
- **Named Entity Recognition (NER)**
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

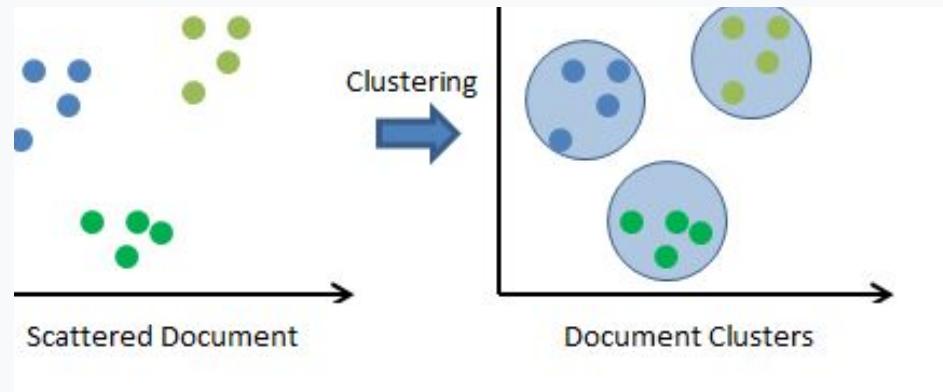
Named entity recognition (NER) is the process of identifying and categorizing named entities in a text, such as people, organizations, locations, and dates.

When Sebastian Thrun PERSON started at Google ORG in 2007 DATE, few people outside of the company took him seriously. "I can tell you very senior CEOs of major American NORP car companies would shake my hand and turn away because I wasn't worth talking to," said Thrun PERSON, now the co-founder and CEO of online higher education startup Udacity, in an interview with Recode ORG earlier this week DATE.

A little less than a decade later DATE, dozens of self-driving startups have cropped up while automakers around the world clamor, wallet in hand, to secure their place in the fast-moving world of fully automated transportation.

Common NLP tasks

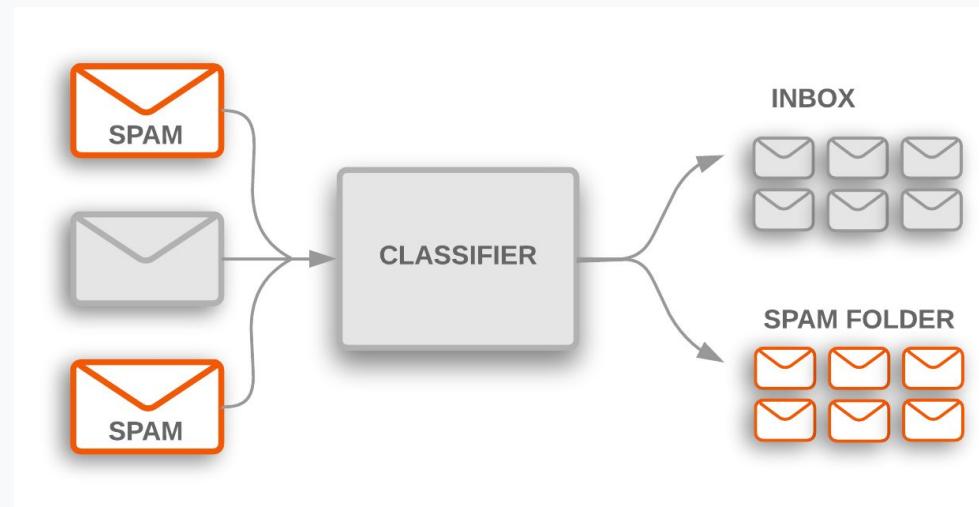
- Tokenization
- Named Entity Recognition (NER)
- **Text representation**
- Text classification
- Natural language generation
- Multimodal NLP



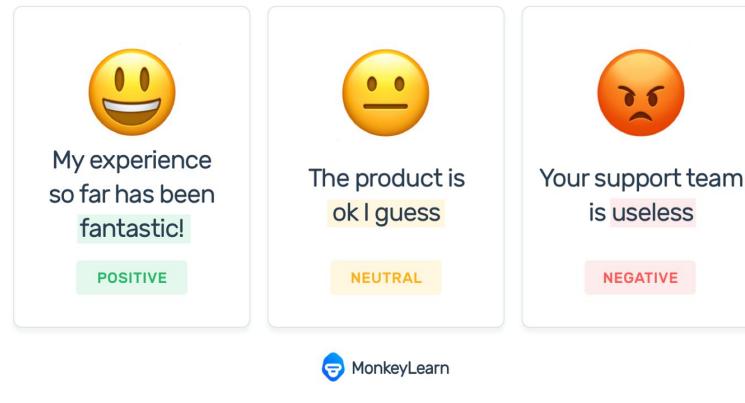
- Text representation is the process of converting unstructured text data into a structured format that can be used for natural language processing tasks.
- It involves selecting a suitable representation scheme, such as bag-of-words, word embeddings, or topic models, to capture the key features and characteristics of the text data in a numerical form that can be processed by machine learning algorithms.

Common NLP tasks

- Tokenization
- Named Entity Recognition (NER)
- Text representation
- **Text classification**
- Natural language generation
- Multimodal NLP



Sentiment Analysis



Common NLP tasks

- Tokenization
- Named Entity Recognition (NER)
- Text representation
- Text classification
- **Natural language generation**
- Multimodal NLP

Input Article

Marseille, France (CNN) The French prosecutor leading an investigation into the crash of Germanwings Flight 9525 insisted Wednesday that he was not aware of any video footage from on board the plane. Marseille prosecutor Brice Robin told CNN that " so far no videos were used in the crash investigation . " He added, " A person who has such a video needs to immediately give it to the investigators ." Robin's comments follow claims by two magazines, German daily Bild and French Paris Match, of a cell phone video showing the harrowing final seconds from on board Germanwings Flight 9525 as it crashed into the French Alps . All 150 on board were killed. Paris Match and Bild reported that the video was recovered from a phone at the wreckage site. ...

Abstractive summarization

Text
Summarization
Models

Extractive summarization

Generated summary

Prosecutor : " So far no videos were used in the crash investigation "

Extractive summary

marseille prosecutor brice robin told cnn that " so far no videos were used in the crash investigation . " robin 's comments follow claims by two magazines , german daily bild and french paris match , of a cell phone video showing the harrowing final seconds from on board germanwings flight 9525 as it crashed into the french alps . paris match and bild reported that the video was recovered from a phone at the wreckage site .

Sentence having the right answer

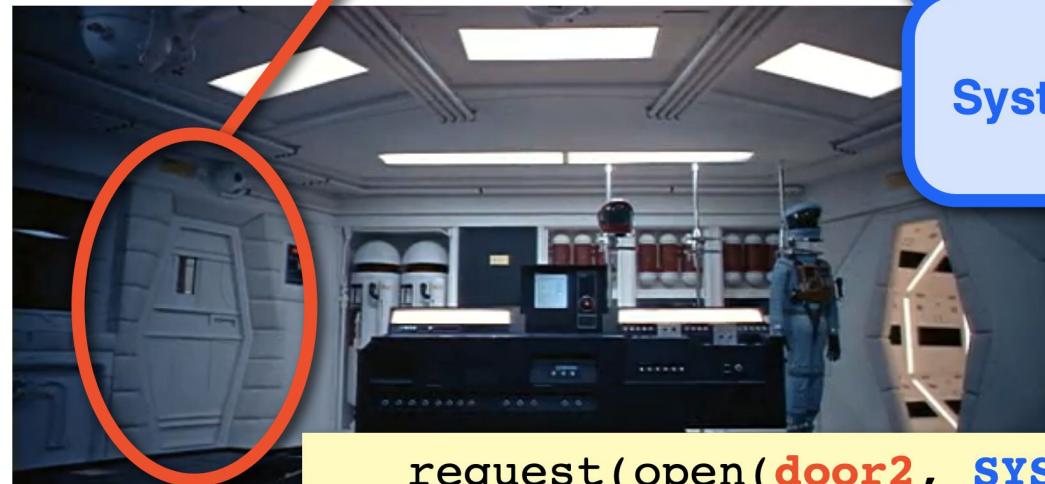
'context': 'Beyoncé Giselle Knowles-Carter (/bi: 'jpnsei/ bee-YON-say) (born September 4, 1981) is an American singer, songwriter, record producer and actress. Born and raised in Houston, Texas, she performed in various singing and dancing competitions as a child, and rose to fame in the late 1990s as lead singer of R&B girl-group Destiny's Child. Managed by her father, Mathew Knowles, the group became one of the world's best-selling girl groups of all time. Their hiatus saw the release of Beyoncé's debut album, Dangerously in Love (2003), which established her as a solo artist worldwide, earned five Grammy Awards and featured the Billboard Hot 100 number-one singles "Crazy in Love" and "Baby Boy".',
'text': 'in the late 1990s'
'question': 'When did Beyonce start becoming popular?'

Exact Answer

Common NLP tasks

- Tokenization
- Named Entity Recognition (NER)
- Text representation
- Text classification
- Natural language generation
- Multimodal NLP

Multimodal NLP: mapping from language to the world

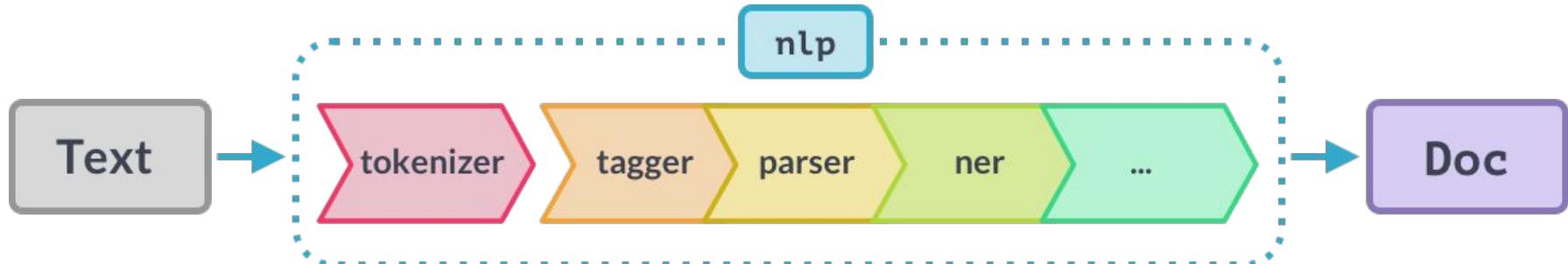
$$\exists x \exists y (\text{pod_door}(x) \And \text{Hal}(y) \\ \And \text{request}(\text{open}(x, y)))$$


`request(open(door2, SYS))`

spaCy Package

spaCy is an open-source library used for natural language processing in python. It is extremely popular for processing a large amount of unstructured data generated at a vast scale in the industry and generate useful and meaningful insights from the data.

spaCy NLP Pipeline



Let's code

Colab

02

Back to LLMs

Coming back to Language models

A Language model refers to a type of model specifically designed to generate human-like text, or predict the probability of a sequence of words. Language models learn patterns and statistics from large amounts of text data, enabling them to generate sensible and contextually appropriate sentences.

The cat

What did we use before Transformers?

RNNs (Recurrent Neural Nets), *predicting* the output by assessing the relationships of a word with its immediate neighbors.

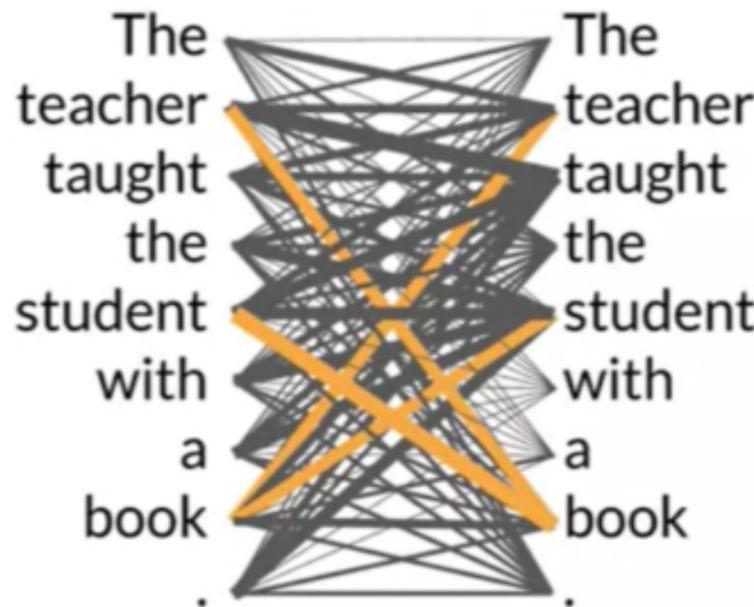
Unfortunately, RNNs had two major flaws, when it comes to NLP tasks:

- **Difficulty in handling long sequences** → Finding it difficult to learn from things that happened long ago in the sequence.
- **Computational Complexity** → Due to the sequential processing of their training, it was quite timely and costly to train them.

What is a transformer?

A transformer is a type of deep learning model architecture that employs self-attention, in order to capture relationships between words in a sequence

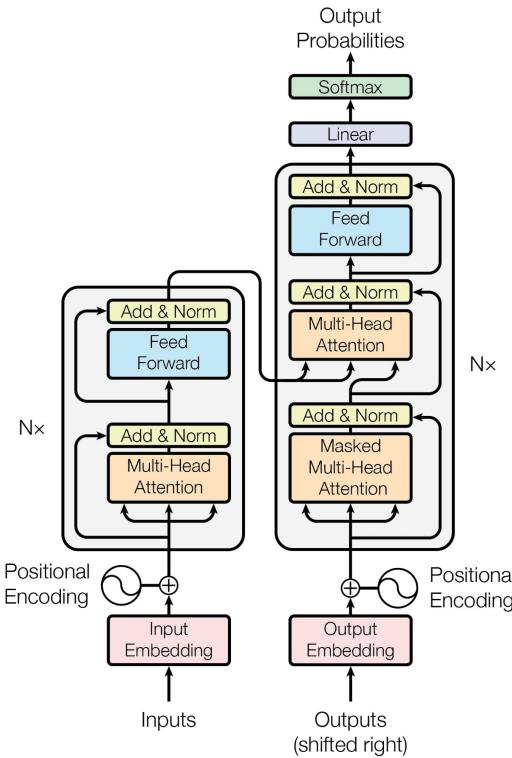
The strength of the transformer model is in its ability to understand the significance and context of every word in a sentence.



The Encoder Decoder Model

Encoder

- \Leftrightarrow Understanding / Classification / Regression / Sequence Labeling
- Given a textual inputs, produce hidden representations
 - o Eg. embedding, topic distribution, class likelihood, etc.



Decoder

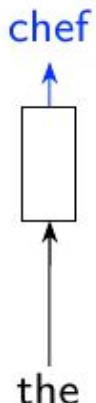
- \Leftrightarrow Generation / Summarization
- Given a hidden representation, produce the textual outputs.
- Note: Understanding tasks can also be modeled as generation task

MLM

Under Masked Language Modelling, we typically mask a certain % of words in a given sentence and the model is expected to predict those masked words based on other words in that sentence.

Such a training scheme makes this model bidirectional in nature, because the representation of the masked word is learnt based on the words that occur it's left as well as right.

Language Modeling



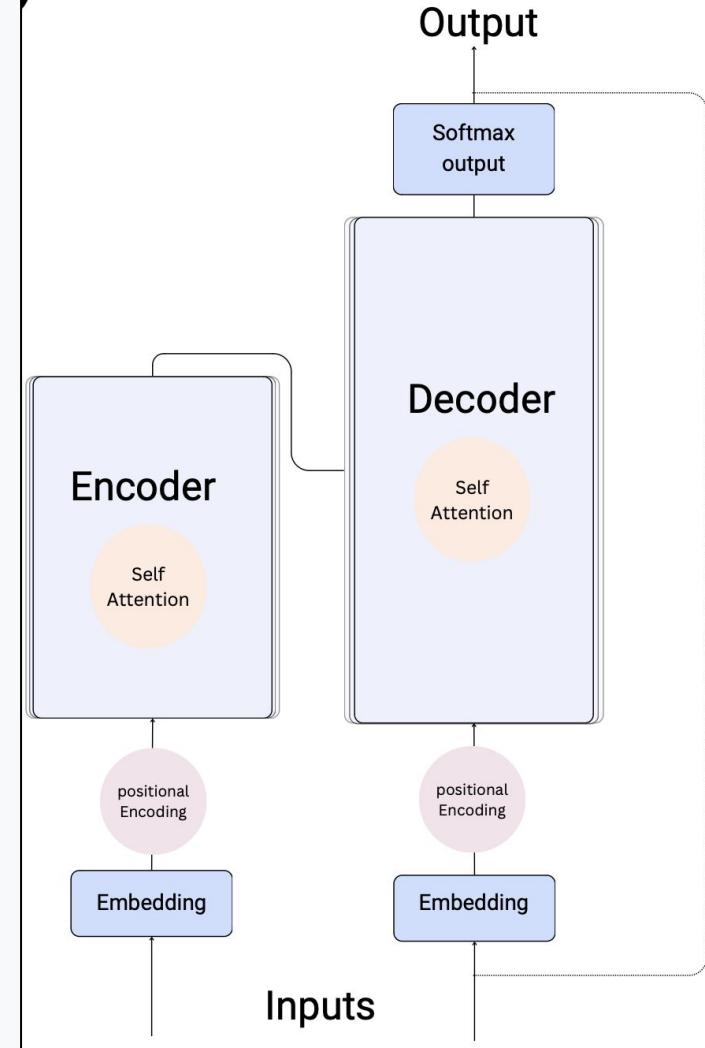
Masked Language Modeling

the chef cooked the meal

How does the transformer model work?

- Tokenization
- Embeddings
- Positional Encodings
- Self attention
- Multi-Headed Attention

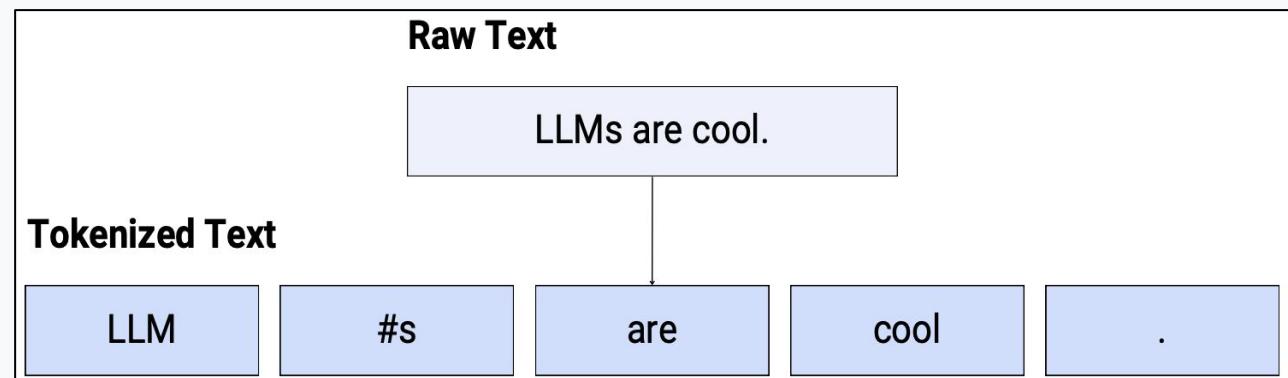
Source: [Data Science Dojo!](#)



How does the transformer model work?

- Tokenization
- Embeddings
- Positional Encodings
- Self attention
- Multi-Headed Attention

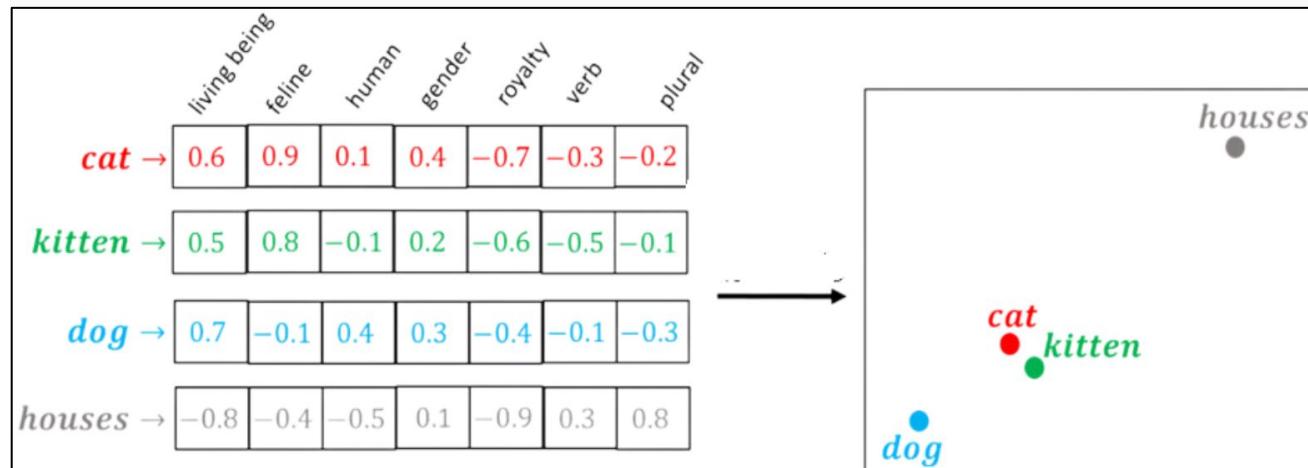
As stated before, Tokenization is the process of breaking down text into smaller units, such as words or phrases, for easier processing and analysis.



How does the transformer model work?

- Tokenization
- **Embeddings**
- Positional Encodings
- Self attention
- Multi-Headed Attention

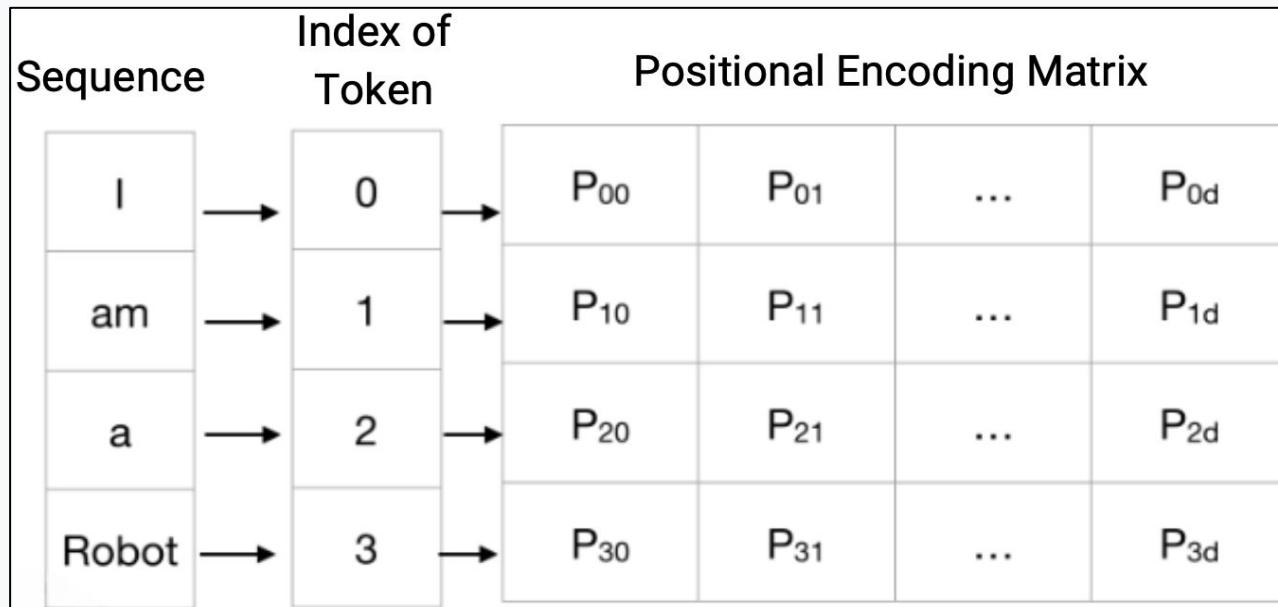
Once again, Embedding is the process where each token is then transformed into a vector in a high-dimensional space. This embedding captures the meaning and context of each word.



How does the transformer model work?

- Tokenization
- Embeddings
- **Positional Encodings**
- Self attention
- Multi-Headed Attention

Positional encoding is the process of adding information to a model about the position of elements in a sequence.

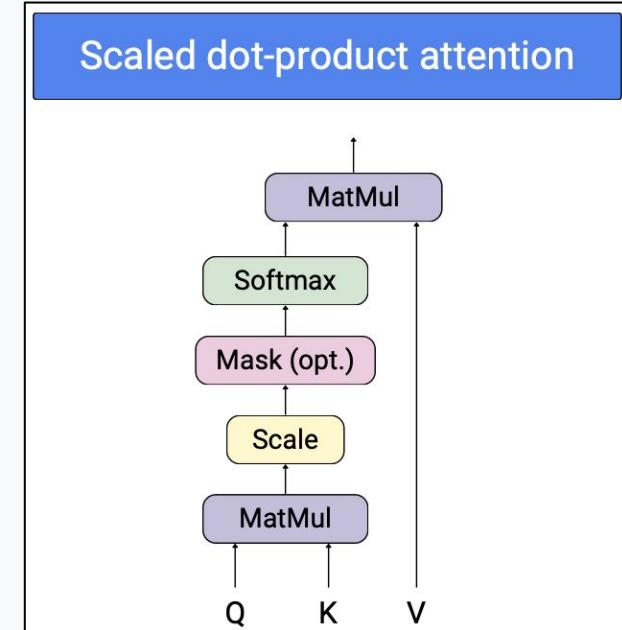


How does the transformer model work?

- Tokenization
- Embeddings
- Positional Encodings
- **Self attention**
- Multi-Headed Attention

The model calculates attention scores for each word, determining how much focus it should put on other words in the sentence when trying to understand a particular word.

This helps the model capture relationships and context within the text.

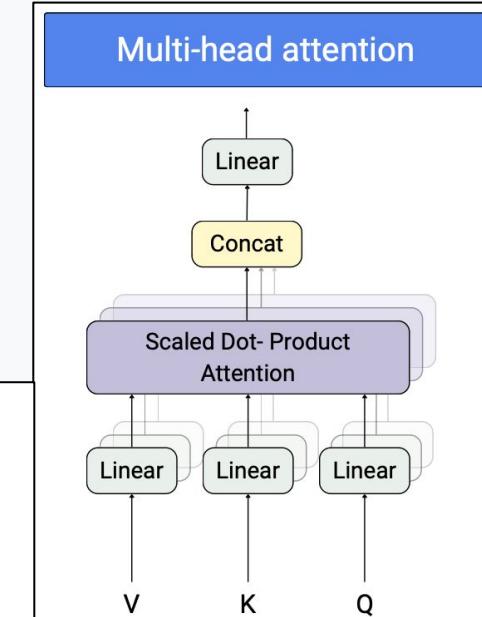


How does the transformer model work?

- Tokenization
- Embeddings
- Positional Encodings
- Self attention
- **Multi-Headed Attention**

Source: [Data Science Dojo!](#)

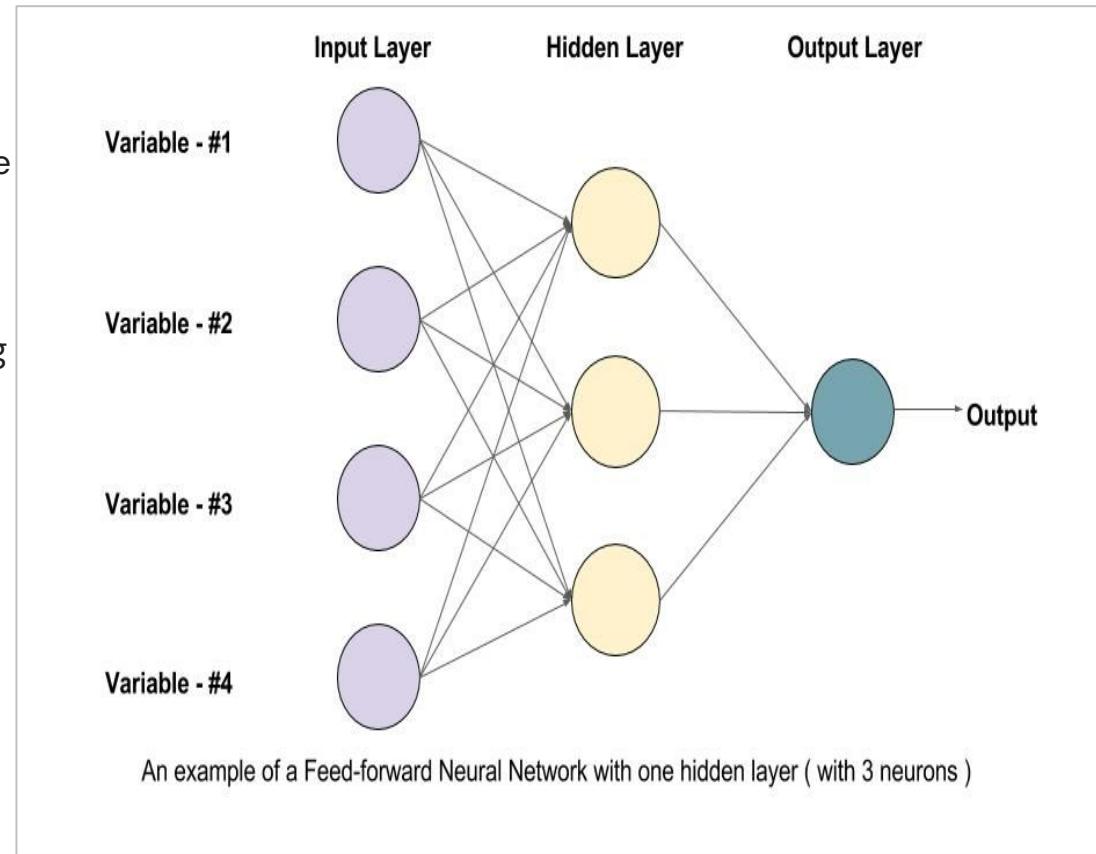
Multi-headed attention is a mechanism in transformers that runs several self-attention processes in parallel, allowing the model to focus on different parts of the input sequence from different perspectives at the same time.

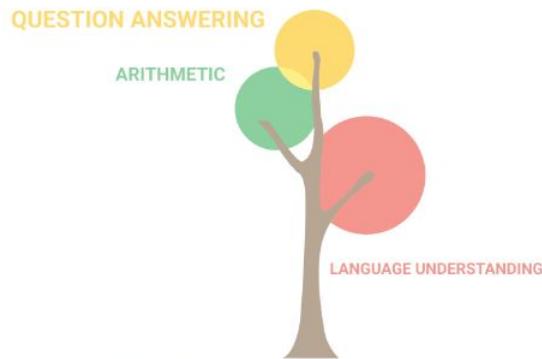


Finally, to generate an Output: The final layers of the transformer convert the processed data into an output format suitable for the task at hand, such as classifying the text or generating new text.

Feedforward Layer

The **feedforward layers** of LLMs have multiple fully connected layers that apply nonlinear transformations to the input embeddings. These layers help the model learn higher-level abstractions from the input text. **Feedforward** is a basic concept used in many areas, including **artificial intelligence** and **neural networks**. In simple words, feedforward refers to a process where *information flows in a single direction, from the input to the output, without any feedback loops*.





8 billion parameters

LLM use case with change in Parameters

Parameters

Parameters refer to the learnable elements or variables within the model.

These parameters capture the knowledge and patterns learned during training, and are used to make predictions or generate text.

They represent the internal representations and connections between the different components of the model, such as the hidden layers and attention mechanisms.

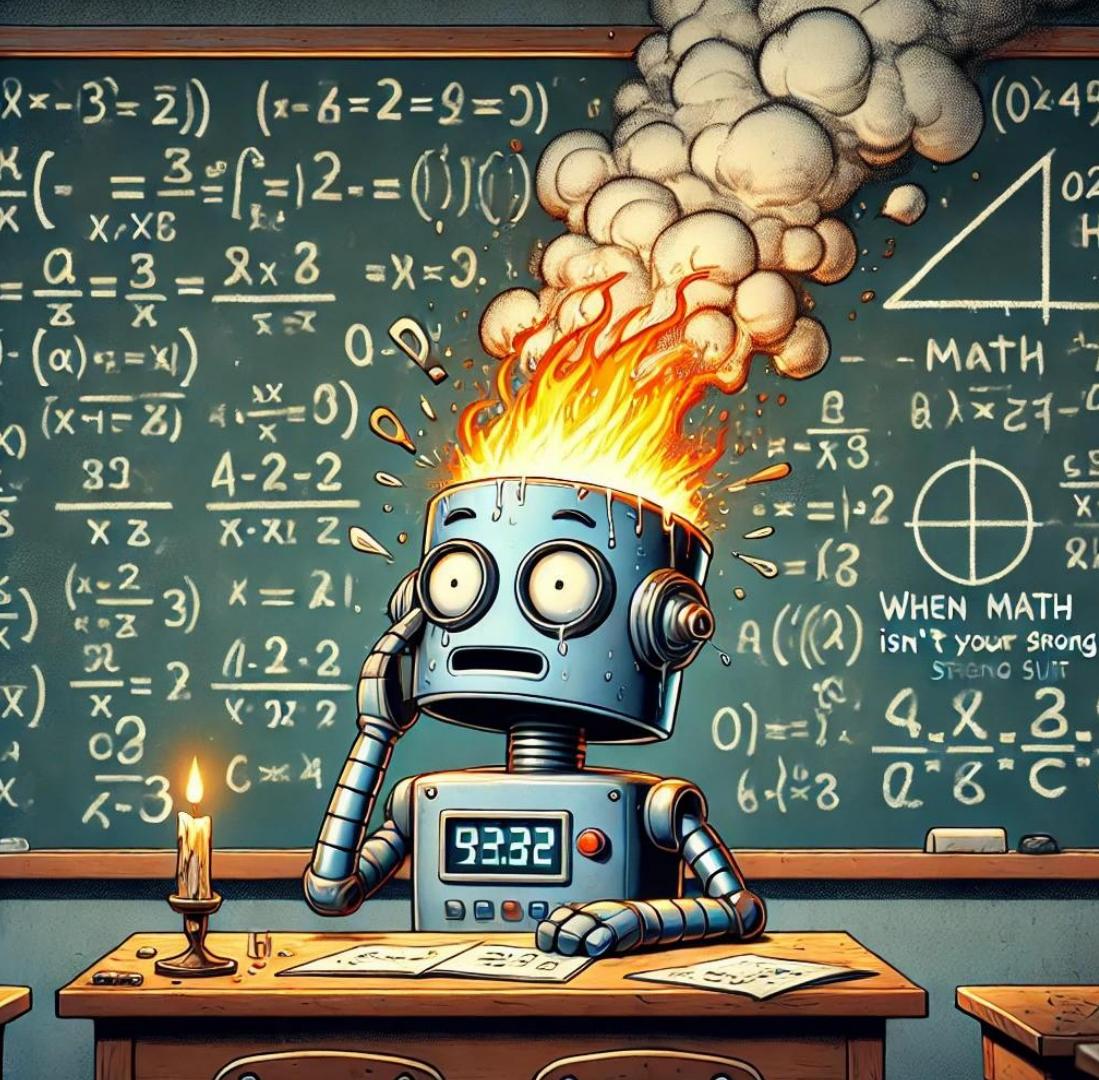
Parameters: Explain like I'm 5

- Imagine you have a very smart robot that can understand and speak in different languages.
- To make this robot understand and respond correctly, we need to give it some instructions and knowledge.
- Parameters in large language models are like little pieces of information that the robot learns and remembers during its training.
- The more parameters the robot has, the more information it can remember and use to understand and speak languages.
- It's like having a bigger memory or a bigger brain!
- With more parameters, the robot can become smarter and perform better at tasks like translating languages, answering questions, or even having a conversation with you.

What about arithmetics?

How do GPT models learn how to calculate numbers and perform arithmetic operations?

What is going on, behind the scenes, with ChatGPT when you ask it to calculate 193893×10342 ?



How GPT models learn arithmetics...

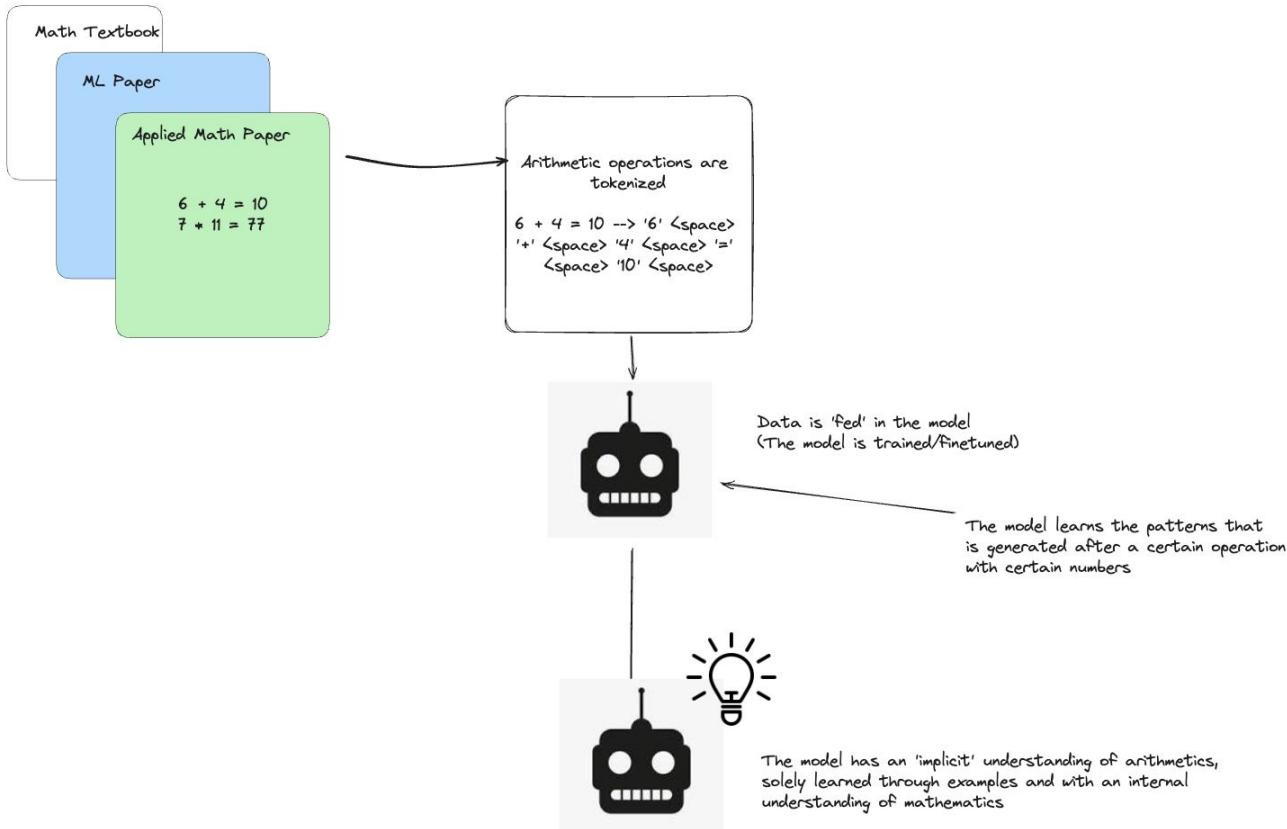
Not all GPT-frameworks have built-in calculators for mathematical queries (ChatGPT is an example), however many GPT models can correctly answer both simple and complex arithmetic questions, from $2+2$ to more complicated multiplications. How can this be?

- **Learning Through Patterns:** Like the rest of their knowledge, the models learn patterns by observing it in the data, in a tokenized form. Numbers themselves are tokenized, helping the models process them similarly to how they processes textual/linguistic data (i.e. the number 634 is interpreted as '6' '3' '4' <space>, not as a single entity)
- **Training Examples:** The models are exposed to countless arithmetic examples in their training data (math textbooks, scientific papers...).
- **Tokenization and Statistics:** Between the tokenization, and the presence of examples, the models learn digit relationships statistically, not by traditional computing methods.

How GPT models learn arithmetics...

- **Carry Overs:** GPTs don't "understand" carryover as a computational necessity, but predict it through learned patterns in text
- **Power and Complexity of Neural Nets:** The models demonstrate the capacity to learn arithmetic patterns through example-based training, without being specifically programmed for these tasks.
- **Future and Theoretical Implications:** The research in this area is fascinating; whether the models' internal methods for arithmetic computing are akin to traditional algorithms or unique neural network solutions is still speculative, but open fascinating possibilities for understanding neural network operations.

How GPT models learn arithmetics...



Here's a few examples



241 - (-241) + 1



241 - (-241) + 1 is equivalent to 241 + 241 + 1, which simplifies to 483 + 1. So
241 - (-241) + 1 is equal to 484.

Like Dislike



If it takes 5 machines 5 minutes to make 5 devices, how long would it take 100 machines to make 100 devices?



If it takes 5 machines 5 minutes to make 5 devices, then it would take 100 machines 100 minutes to make 100 devices.

Like Dislike



That's not right. Check carefully and try again.



I apologize if my response was not correct. Can you please provide more information or clarify your question? I may have misunderstood the scenario you described.

Like Dislike



It takes a machine 5 minutes to make a device. How long would it take 100 machines to make 100 devices?



If it takes a single machine 5 minutes to make a single device, then it would take 100 machines 5 minutes to make 100 devices. Therefore, it would take a total of 500 minutes for 100 machines to make 100 devices.

Like Dislike

Try again



[Source](#)

Resources for understanding Transformer Architecture

Wordy, but awesome

Awesome and easy to watch

Should have been the first one

Foundational Models

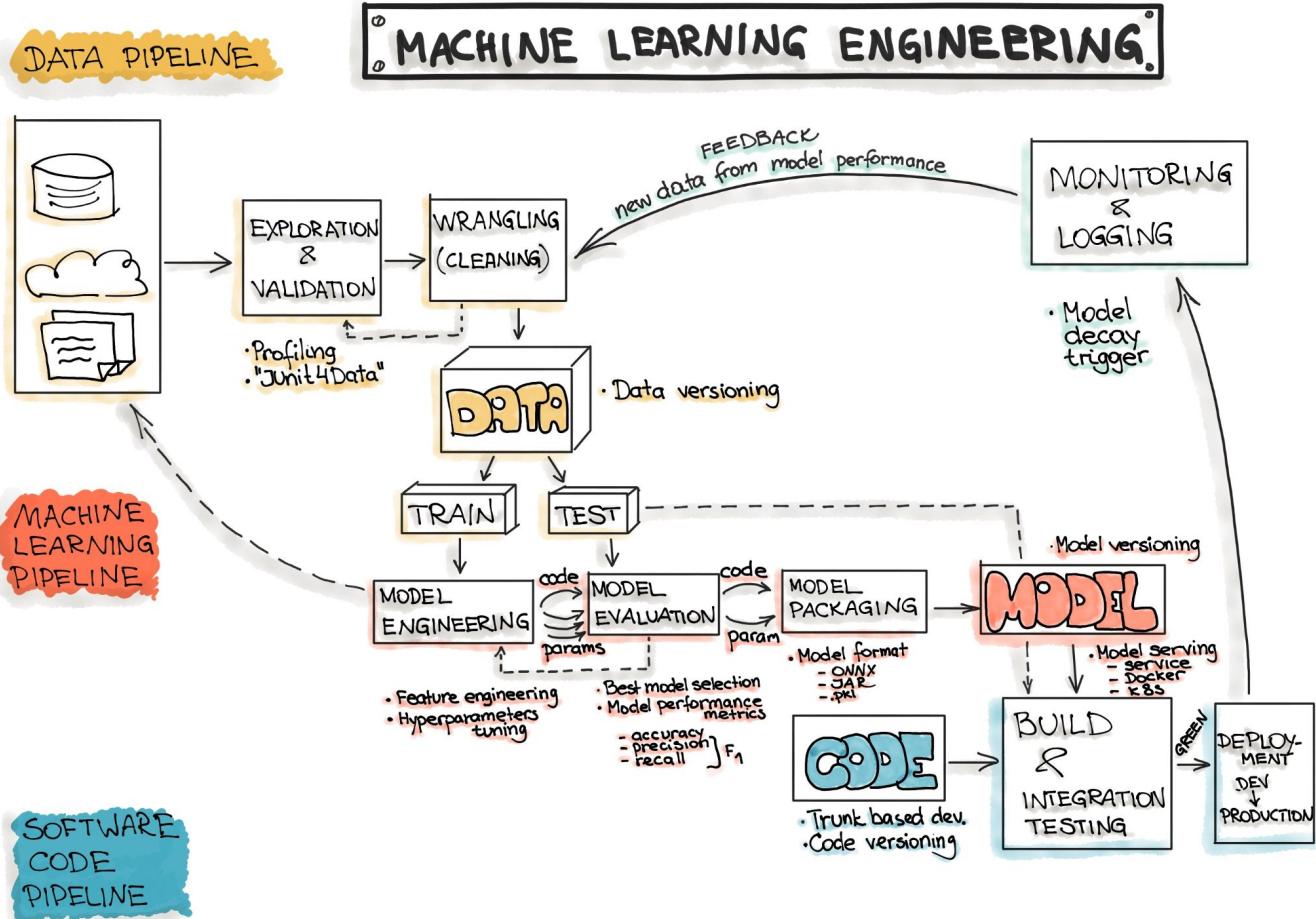
Let's shift gears: Introduction to ML System Design



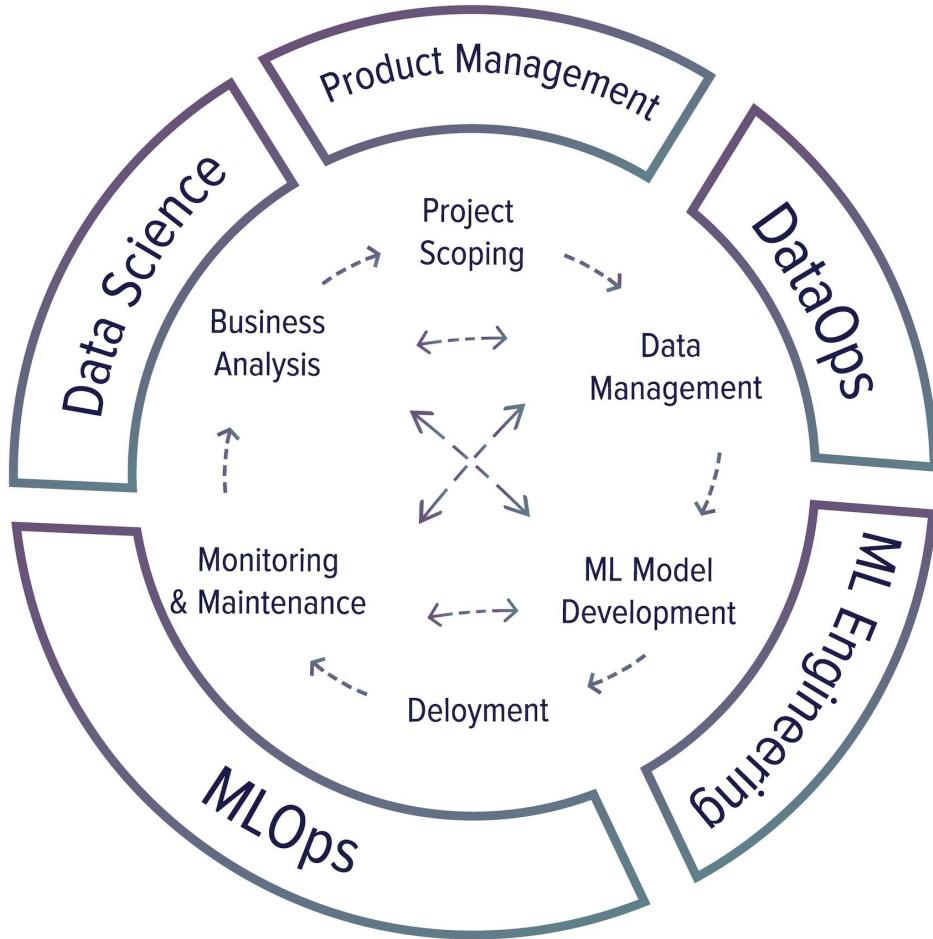
What does it really mean?

Machine learning system design refers to the process of creating and developing a system that utilizes machine learning algorithms to solve a specific problem or task.

The horror



Simplified Version



In summary...

machine learning system design involves

- Problem identification,
- Data preparation,
- Model selection and
- Training, evaluation and optimization,
- Deployment, and
- Ongoing monitoring.

It is a cyclical process aimed at creating effective and efficient systems, that leverage machine learning techniques to solve real-world problems.

Word of caution

We want to stay **away** from software system design because that is a whole new paradigm

How do LLMs change all this?

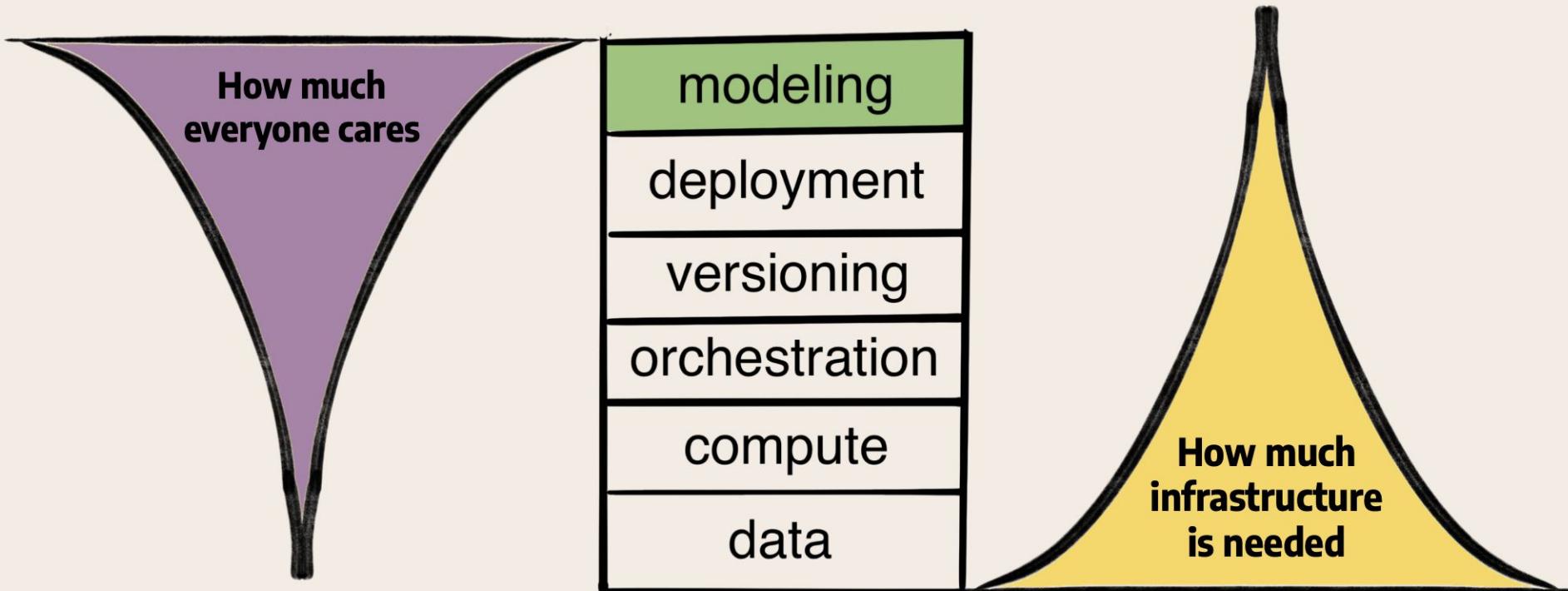
Adoption of LLMs and Foundation Models:

LLMs and other foundation models are expected to be embraced as a potent addition to the machine learning (ML) stack, complementing existing models rather than replacing them.

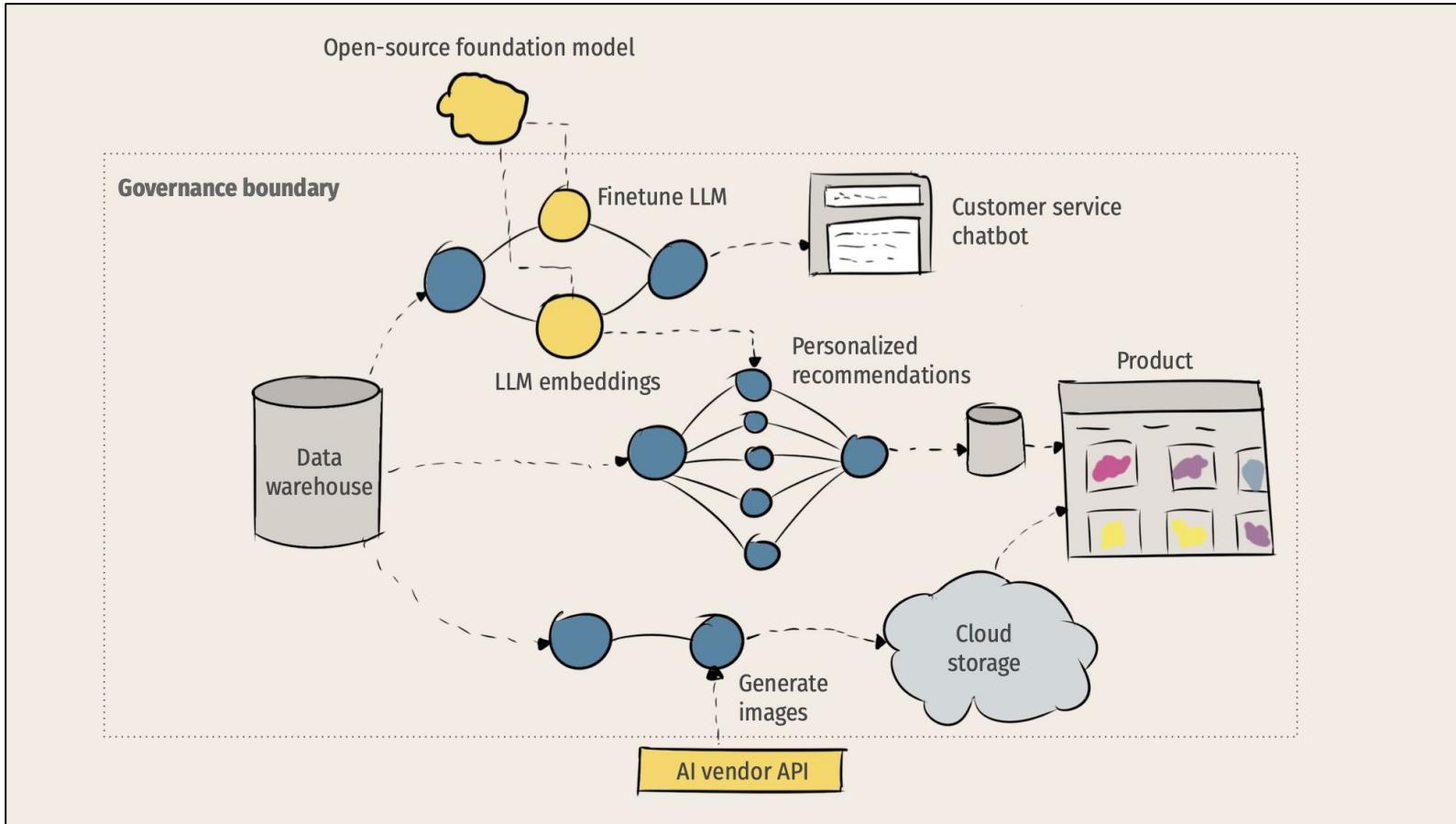
Consistency in Technology Adoption:

Companies are likely to maintain their existing approaches to technology adoption when incorporating LLMs and foundation models, as there is no compelling reason to suggest otherwise.

LLMs are just one part of the larger system



What should LLM Powered ML Systems look like...



[Source](#)



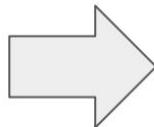
BullsAI

- Default
- Unknown
- Unknown
- Unknown



Angel Midi Dress

Embrace sophistication with the Angel Midi Dress Black, crafted from 100% reclaimed polyester for eco-conscious elegance. Featuring a silky feel, a stunning plunging neckline, and a chic high leg slit, this dress combines sustainable fashion with a touch of allure. Perfect for any occasion that calls for a sleek and stylish statement.

[Add to Cart](#)

Olivia

New York

Summer dress (size XL), jeans (size XL)

34 years old



Angel Midi Broadway Dress

The Angel Midi Dress Black, combines high-end urban elegance with comfort. Its smooth fabric and graceful cut make it a versatile piece for any wardrobe, perfect for transitioning from work to social events.

[Add to Cart](#)

Use case – Virtual Photoshoots



Thank you.

Appendix