

# Simplifying ChatGPT: Efficient Document Querying with Azure OpenAI

*Luca Zavarella*

**DATA**  
SATURDAYS



# Sponsors



# About me

## Luca Zavarella

Working in Business Intelligence with SQL Server since 2007

Microsoft MVP for Artificial Intelligence & Data Platform

Microsoft Certified: Azure Data Scientist Associate

Author of *"Extending Power BI with Python and R, 2° Ed."* by Packt



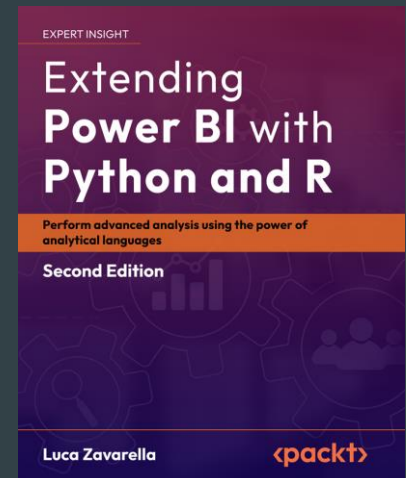
Head of Data & AI @  **icubed**

**Email:** [luca.zavarella@icubed.it](mailto:luca.zavarella@icubed.it)

**Twitter:** @lucazav

**LinkedIn:** <https://it.linkedin.com/in/lucazavarella>

**Blog:** <https://lucazavarella.medium.com>



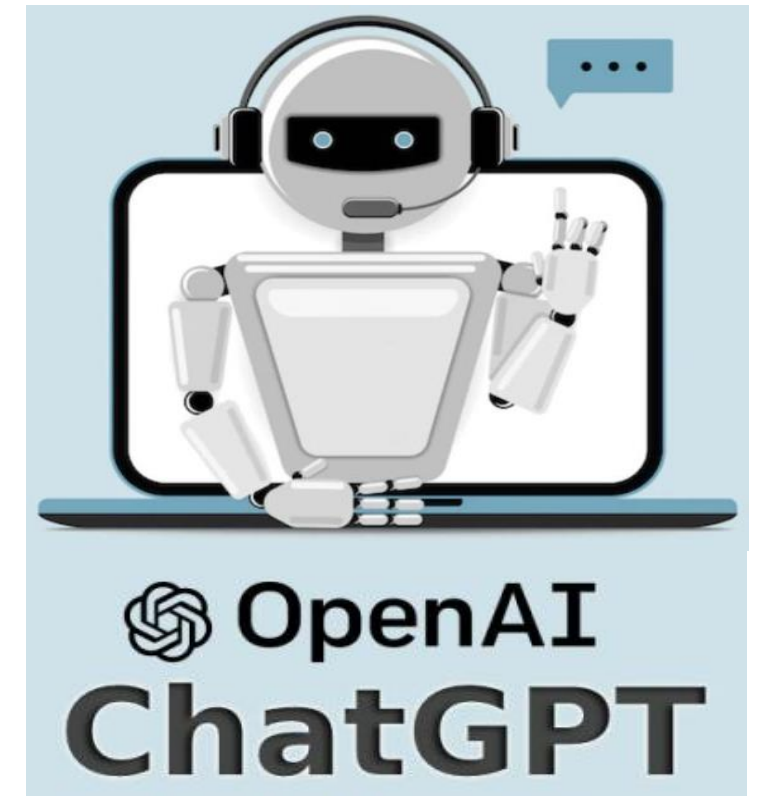
- > What is ChatGPT and how does it work?
- > Why so much success?
- > Emergent Behavior Examples
- > Unlocking Business Potential with Prompt Engineering
- > DEMO 01 – Play with Azure OpenAI on your data
- > DEMO 02 – iCubed AIKE

# What is ChatGPT and How Does it Work?

# What is ChatGPT

AI **chatbot** developed by OpenAI

- > Software application
- > Uses Artificial Intelligence (AI) and Natural Language Processing (NLP) techniques
- > Comprehend customer queries and generate automated responses, creating a conversational experience like that of human interaction

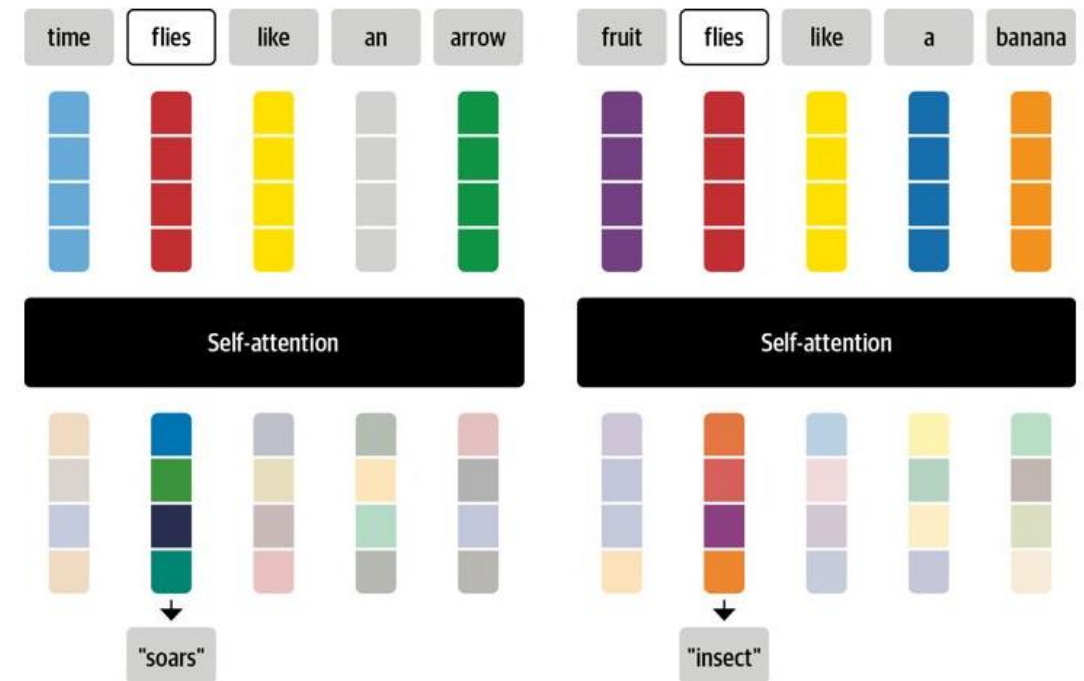


- > Built on top of GPT family of **Large Language Models (LLM)**
  - > Advanced AI systems designed to **understand** and **generate** human-like text based on the input they receive
- > **Generative** Pre-trained **Transformers (GPT)**
  - > Can generate new and original data based on patterns it has learned during training
  - > This contrasts with **discriminative** AI algorithms that can classify, or label data based on pre-defined categories

# Transformer Foundation Model

A neural network that **learns context** and thus meaning by tracking relationships in sequential data like the words in this sentence

- > First described in a 2017 paper from Google (*Attention Is All You Need*)
- > They apply an evolving set of mathematical techniques, called **attention** or self-attention
- > They detect subtle ways even distant data elements in a series influence and depend on each other
- > Called “foundation models” because they drove a paradigm shift in AI



Tunstall et al., Natural Language Processing with Transformers: Building Language Applications with Hugging Face, 2022.



# What ChatGPT Really Does

It generates a coherent and plausible [extension of the given text](#), based on the patterns and relationships it has learned from analyzing vast amounts of text data on the internet and other sources

- > Provide a probability distribution over a sequence of words
- > They estimate the [likelihood of each word](#) or set of words appearing in each sentence or text

*The best thing about AI is its ability to*

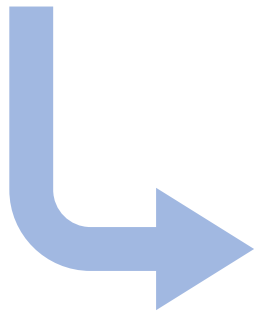
learn	4.5%
predict	3.5%
make	3.2%
understand	3.1%
do	2.9%

# Randomness of ChatGPT Output

- > You may think to pick the highest-ranked word...
  - > You will get a very “flat” essay
  - > Never seems to show any “creativity” (and even sometimes repeats word for word)
- > “Temperature” parameter determines how often lower-ranked words will be used
  - > The term “temperature” stems from its analogy with statistical mechanics and the Boltzmann distribution
  - > A higher temperature ( $\rightarrow 1.0$ ) increases the randomness and diversity of the output
  - > A lower temperature ( $0 \leftarrow$ ) reduces the randomness
  - > For essay generation (“creativity”), it turns out that a “temperature” of 0.8 seems best

# Text Generation Using Temperature 0.0

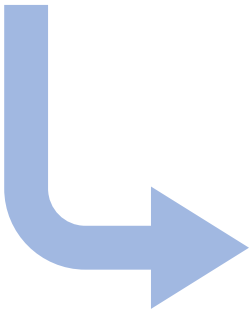
{ The best thing about AI is its ability to,  
The best thing about AI is its ability to learn,  
The best thing about AI is its ability to learn from,  
The best thing about AI is its ability to learn from experience,  
The best thing about AI is its ability to learn from experience.,  
The best thing about AI is its ability to learn from experience. It,  
The best thing about AI is its ability to learn from experience. It's,  
The best thing about AI is its ability to learn from experience. It's not }



The best thing about AI is its ability to learn from experience. It's not just a matter of learning from experience, it's learning from the world around you. The AI is a very good example of this. It's a very good example of how to use AI to improve your life. It's a very good example of how to use AI to improve your life. The AI is a very good example of how to use AI to improve your life. It's a very good example of how to use AI to

# Text Generation Using Temperature 0.8

{ The best thing about AI is its ability to,  
The best thing about AI is its ability to create,  
The best thing about AI is its ability to create worlds,  
The best thing about AI is its ability to create worlds that,  
The best thing about AI is its ability to create worlds that are,  
The best thing about AI is its ability to create worlds that are both,  
The best thing about AI is its ability to create worlds that are both exciting,  
The best thing about AI is its ability to create worlds that are both exciting, }



The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed, taking an infinite number of steps, developing a machine that can be integrated with other systems, or controlling one system that's truly a machine, is one of the most fundamental processes of AI. Aside from the human-machine interaction, AI was also a big part of creativity

# ChatGPT Evolution With Numbers

OpenAI has been developing GPT since 2018

- > GPT-1
  - > Trained with BooksCorpus dataset (5 GB), for language understanding
- > GPT-2
  - > Launched on Valentine's Day 2019 ☺
  - > Trained with Reddit articles with over 3 likes (40 GB)
  - > The training cost was \$43,000
- > GPT-3
  - > Released in June 2020
  - > Trained by a much more comprehensive dataset
- > GPT-4
  - > Released In March 2023
  - > It has **175 billion parameters**
  - > It receives 10 million queries per day
  - > Trained on ~ **570 GB of text datasets**, including web pages, books, and other sources
  - > Costs an estimated **\$12 mln** to train

Why so much success?

# ChatGPT Has Gone Viral!

## ChatGPT Sprints to One Million Users

Time it took for selected online services to reach one million users



\* one million backers \*\* one million nights booked \*\*\* one million downloads  
Source: Company announcements via Business Insider/LinkedIn



statista

# Why is it so successful? – 1/2

It is generative! 🤖



**You**

Introduci i presenti alla mia sessione "Simplifying ChatGPT: Efficient Document Querying with Azure OpenAI" utilizzando la metafora simile a quella delle pillole rossa e blu usate da Morpheus in Matrix. Sii breve



**ChatGPT**

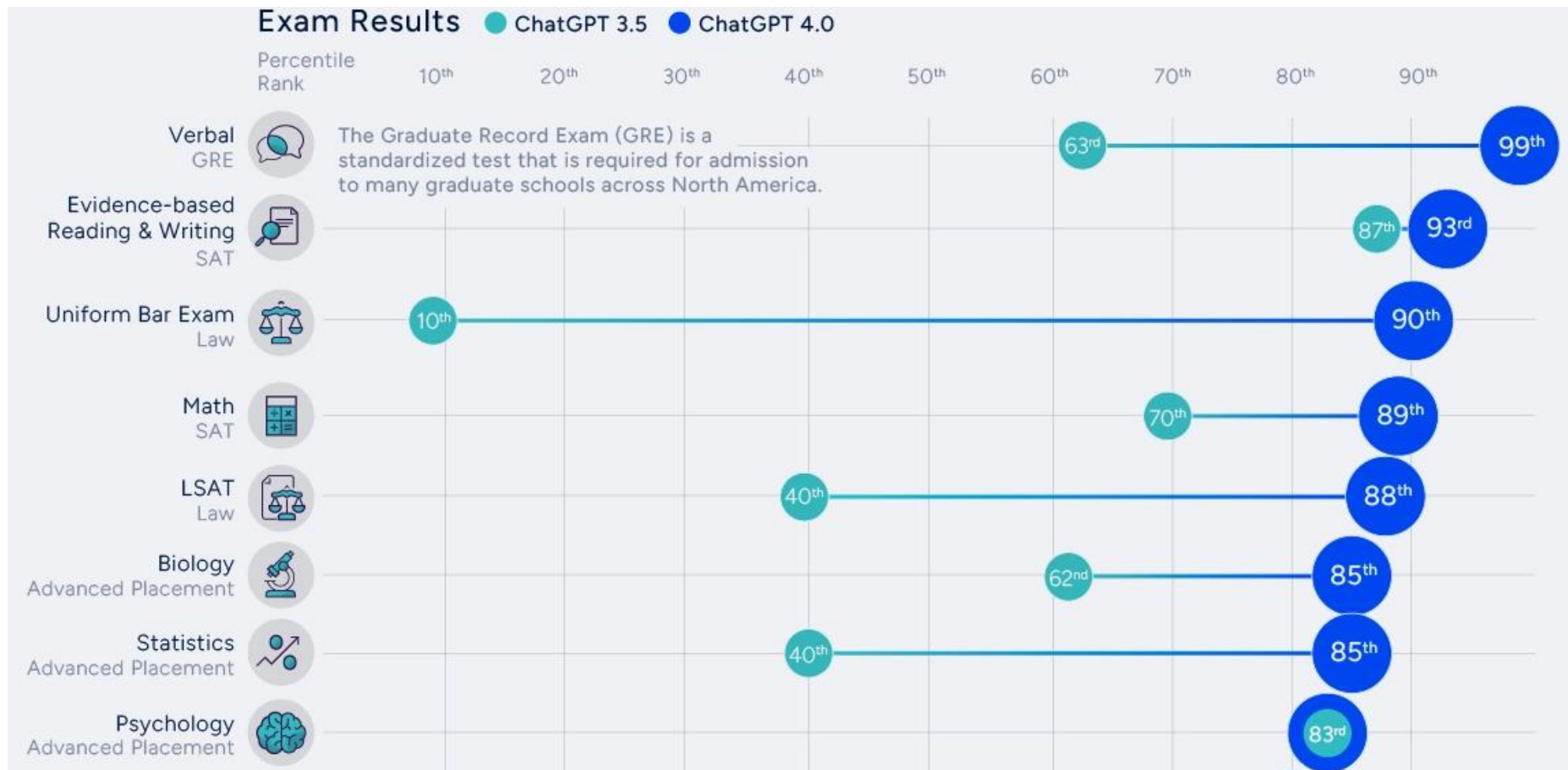
Benvenuti a tutti alla mia sessione "Simplifying ChatGPT: Efficient Document Querying with Azure OpenAI". Oggi vi troverete di fronte a una scelta, simile a quella presentata da Morpheus nel film Matrix con le sue pillole rossa e blu. Se scegliete la pillola blu, tornerete alla vostra routine quotidiana, continuando a interagire con i documenti e i dati nel modo tradizionale, con tutte le sue sfide e limitazioni. Ma se scegliete la pillola rossa, vi porterò in un viaggio attraverso le profondità di ChatGPT e Azure OpenAI, dove scopriremo insieme come semplificare e rivoluzionare il modo in cui interroghiamo i documenti, sbloccando potenzialità nascoste e raggiungendo una nuova efficienza. La scelta è vostra: quale pillola sceglierete?



## Why is it so successful? – 2/2

- > These models supposedly have one directive
  - > To accept a string of text as input and **predict what comes next**, over and over, based purely on statistics
- > Computer scientists anticipated that scaling up would boost performance on known tasks
- > They didn't expect the models to suddenly handle so many new unpredictable tasks (**emergent behavior**)! 🤖
  - > Abilities not present in smaller-scale models but are present in large-scale models
  - > Thus, they cannot be predicted by simply extrapolating the performance improvements on smaller-scale models

# How Smart is ChatGPT?



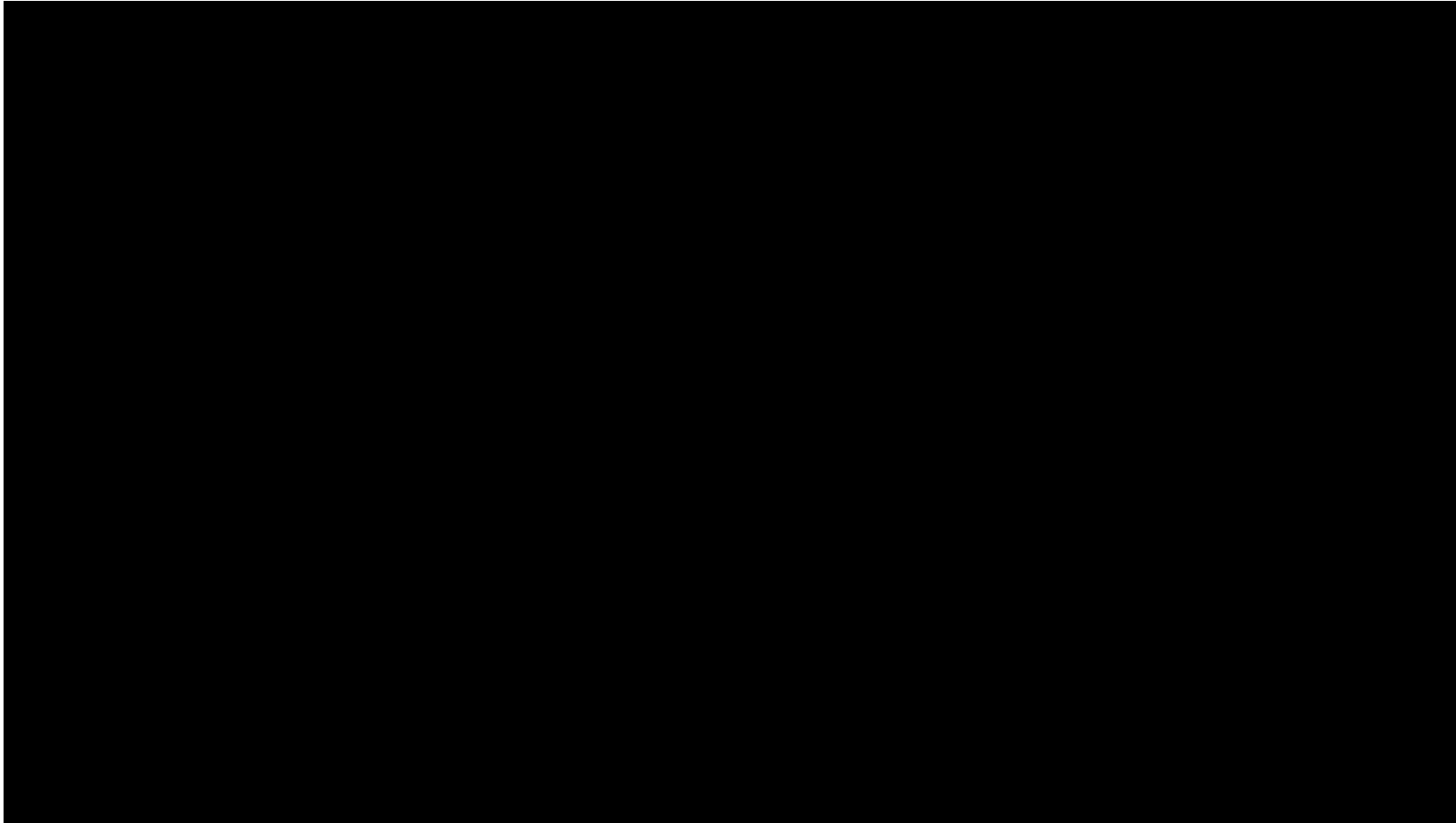
A percentile describes how an examinee's score ranks in comparison to others.

For example

60<sup>th</sup> Percentile



<https://www.visualcapitalist.com/how-smart-is-chatgpt/>



# Emergent Behavior Examples

# Advanced Problem Solving: Logical Riddles 1



**You**

In un pianeta extrasolare ci sono degli organismi viventi che ogni giorno raddoppiano la loro estensione sulla superficie in cui giacciono. Se impiegano 48 giorni per ricoprire l'intera superficie di un territorio, quanto impiegano per coprirne la metà?



**ChatGPT**

Gli organismi viventi in questo scenario seguono una crescita esponenziale, dove raddoppiano la loro estensione ogni giorno. Se impiegano 48 giorni per ricoprire l'intera superficie di un territorio, per coprirne la metà impiegheranno 47 giorni.

Questo perché il giorno successivo, il 48°, raddoppieranno la loro estensione da metà a tutta la superficie. Quindi, il giorno prima dell'ultimo, avranno coperto esattamente la metà del territorio.





**You**

Un oste disonesto e ubriaccone beve 6 litri di vino da un barile che ne contiene 360, e li sostituisce con acqua, in modo che nessuno si accorga del prelievo. Dopo una settimana ripete la malefatta. Dopo un'altra settimana la ripete di nuovo. Quanto vino ha bevuto l'oste?



## ChatGPT

Per risolvere questo problema, è necessario considerare che ogni volta che l'oste sostituisce il vino prelevato con acqua, la concentrazione di vino nel barile diminuisce. Dobbiamo calcolare la quantità di vino effettivamente bevuta nelle tre occasioni.

1. **Primo prelievo:** l'oste preleva 6 litri di vino da un barile che ne contiene 360 litri. Dopo aver sostituito i 6 litri di vino con acqua, nel barile rimangono 354 litri di vino e 6 litri di acqua.
2. **Secondo prelievo:** una settimana dopo, quando l'oste preleva altri 6 litri, il contenuto del barile non è più puro vino, ma una miscela di vino e acqua. Dobbiamo calcolare quale frazione del contenuto prelevato è vino. Dopo il primo prelievo, la concentrazione di vino nel barile è di  $\frac{354}{360}$ . Dunque, quando preleva 6 litri di questa miscela, preleva  $6 \times \frac{354}{360}$  litri di vino e il resto acqua. Sostituisce di nuovo i 6 litri prelevati con acqua.
3. **Terzo prelievo:** analogamente al secondo prelievo, l'oste preleva altri 6 litri da una miscela ancora più diluita. Dobbiamo calcolare la nuova concentrazione di vino e quindi quanto vino effettivamente preleva in questa terza occasione.

Calcoliamo ora la quantità esatta di vino bevuta in ciascuna delle tre occasioni.

L'oste ha bevuto in totale circa 17.7 litri di vino dopo aver ripetuto la sua malefatta per tre volte, sostituendo ogni volta il vino prelevato con acqua per non far notare il prelievo. [-]

# Advanced Problem Solving: Structured Data 1



You

Given [{"id": "TRANS01", "transaction": {"num\_shares": 50000, "amount": 50044568.89, "txn\_ts": "2020-11-18

How much is the total amount spent by Jill Smith as of January 1 2021?

```
[{
  "id": "TRANS01",
  "transaction": {
    "num_shares": 50000,
    "amount": 50044568.89,
    "txn_ts": "2020-11-18 02:31:43",
    "customer": {
      "first_name": "Jill",
      "last_name": "Smith",
      "id": 1234567,
      "email": "jsmith@gmail.com"
    },
    "company": {
      "name": "ACME Corp",
      "ticker": "ACMC",
      "id": "ACME837275222752952",
      "address": "Anytown USA, 333333"
    }
  }
},
]
```

```
{
  "id": "TRANS02",
  "transaction": {
    "num_shares": 60000,
    "amount": 80044568.89,
    "txn_ts": "2021-09-01 19:30:00",
    "customer": {
      "first_name": "Jill",
      "last_name": "Smith",
      "id": 1234567,
      "email": "jsmith@gmail.com"
    },
    "company": {
      "name": "iCubed srl",
      "ticker": "ICUB",
      "id": "ICUB837275222754789",
      "address": "Milan Italy, 20100"
    }
  }
},
]
```

```
{
  "id": "TRANS03",
  "transaction": {
    "num_shares": 10000,
    "amount": 1000000.00,
    "txn_ts": "2021-12-13 12:30:00",
    "customer": {
      "first_name": "Luca",
      "last_name": "Zavarella",
      "id": 654321,
      "email": "lucazavarella@yahoo.com"
    },
    "company": {
      "name": "iCubed srl",
      "ticker": "ICUB",
      "id": "ICUB837275222754789",
      "address": "Milan Italy, 20100"
    }
  }
},
]
```





## ChatGPT

To calculate the total amount spent by Jill Smith as of January 1, 2021, we need to sum up the amounts from her transactions that occurred before this date. Based on the given data, Jill Smith has made two transactions:

1. Transaction with ID "TRANS01" on November 18, 2020, for an amount of \$50,044,568.89.
2. Transaction with ID "TRANS02" on September 1, 2021, for an amount of \$80,044,568.89.

Since the second transaction occurred after January 1, 2021, we only consider the first transaction for this calculation. Therefore, the total amount spent by Jill Smith as of January 1, 2021, is \$50,044,568.89.

**Theory of Mind (ToM):** The ability to **understand** and **attribute mental states to others** (crucial role in human social interactions, communication, empathy, self-consciousness, and morality)

## GPT-3 Milestones

- > GPT-3 solves around 40% of false-belief tasks
  - > Comparable to 3.5-year-old children
- > GPT-3.2 achieves 70% accuracy
  - > Comparable to six-year-olds
- > GPT-3.5 astounding 90% accuracy
  - > Comparable to seven-year-olds

## Breakthrough GPT-4

- > GPT-4 achieves a remarkable milestone by solving nearly all false-belief tasks (95%) 🧠
- > Suggests a potential emergence of ToM, previously considered uniquely human, in language models
- > It usually takes **9 years for a human being** to achieve this state of social comprehension. **AI has achieved it in just 3 years!** 🧠

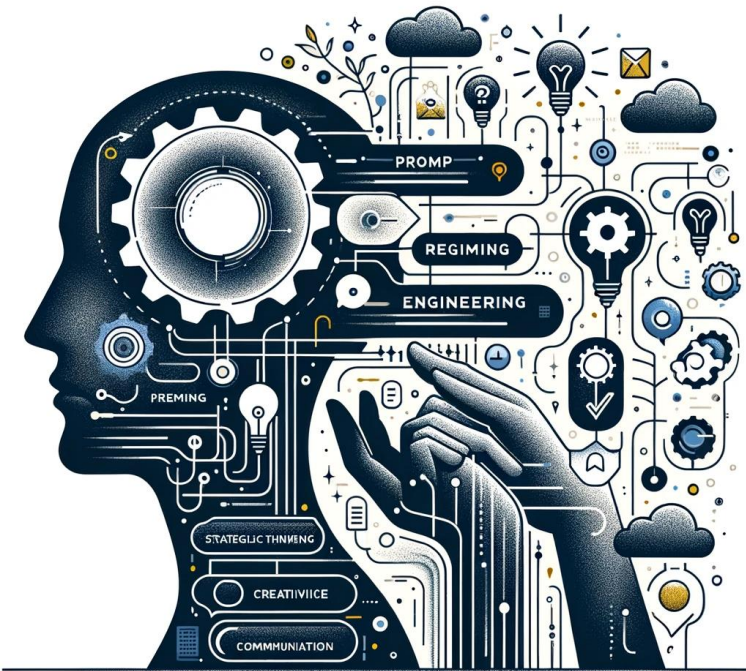
# Be Aware To Hallucinations!



- > ChatGPT operates as a **stochastic** generative AI model
  - > Its responses are probabilistic in nature
  - > It may not always provide accurate responses to specific queries
  - > It **does not understand the meaning** behind the content it generates, so responses may occasionally be irrelevant or nonsensical → **hallucinations**

# Unlocking Business Potential with Prompt Engineering

# Prompt Engineering: Enhancing AI Interactions



- > **Essence:** Tailoring prompts to improve AI's response accuracy and relevance
- > **Goal:** Facilitate precise human-AI communication
- > **Key Strategies**
  - > **Clarity:** Use direct, specific prompts
  - > **Context:** Include relevant background information
  - > **Iteration:** Refine prompts based on AI feedback
  - > **Experimentation:** Test various prompt styles



# Asking For Recent Information



**You**

How did the match between Sinner and Djokovic end up in the 2024 Australian Open?



**ChatGPT**

I'm sorry, but as of my last update in January 2022, I don't have information on specific matches or events that occurred after that date, including the 2024 Australian Open. For the most accurate and up-to-date information on the match between Jannik Sinner and Novak Djokovic in the 2024 Australian Open, I recommend checking the latest sports news or official tournament updates.



# Injecting Context Into Prompts 1



**You**

Given the following context:

<CONTEXT>

Italy's Jannik Sinner has become the first man to defeat Novak Djokovic in an Australian Open semi-final after a stunning 6-1, 6-2, 6-7 (6-8), 6-3 win in Melbourne on Friday (26 January).

The 22-year-old fourth seed – who missed a match point in the third-set tiebreak for a straight-sets victory – was dominant throughout, with his Serbian opponent committing 54 unforced errors in an uncharacteristic performance.

</CONTEXT>

How did the match between Sinner and Djokovic end up in the 2024 Australian Open?



## ChatGPT

Based on the provided context, the match between Jannik Sinner and Novak Djokovic in the 2024 Australian Open ended with Sinner defeating Djokovic. Sinner won the match with a score of 6-1, 6-2, 6-7 (6-8), 6-3 in the semi-final held in Melbourne on Friday, January 26th. It was a notable victory for Sinner as he became the first man to defeat Djokovic in an Australian Open semi-final. Despite missing a match point in the third-set tiebreak, Sinner's dominant performance prevailed, with Djokovic committing 54 unforced errors in the match.



# Injecting Context About Enterprise



You

Given the following context:

<CONTEXT>

Italy's Jannik Sinner has become the first man to defeat Novak Djokovic in an Australian Open semi-final after a stunning 6-1, 6-2, 6-7 (6-8), 6-3 win in Melbourne on Friday (26 January).

## ENTERPRISE DATA

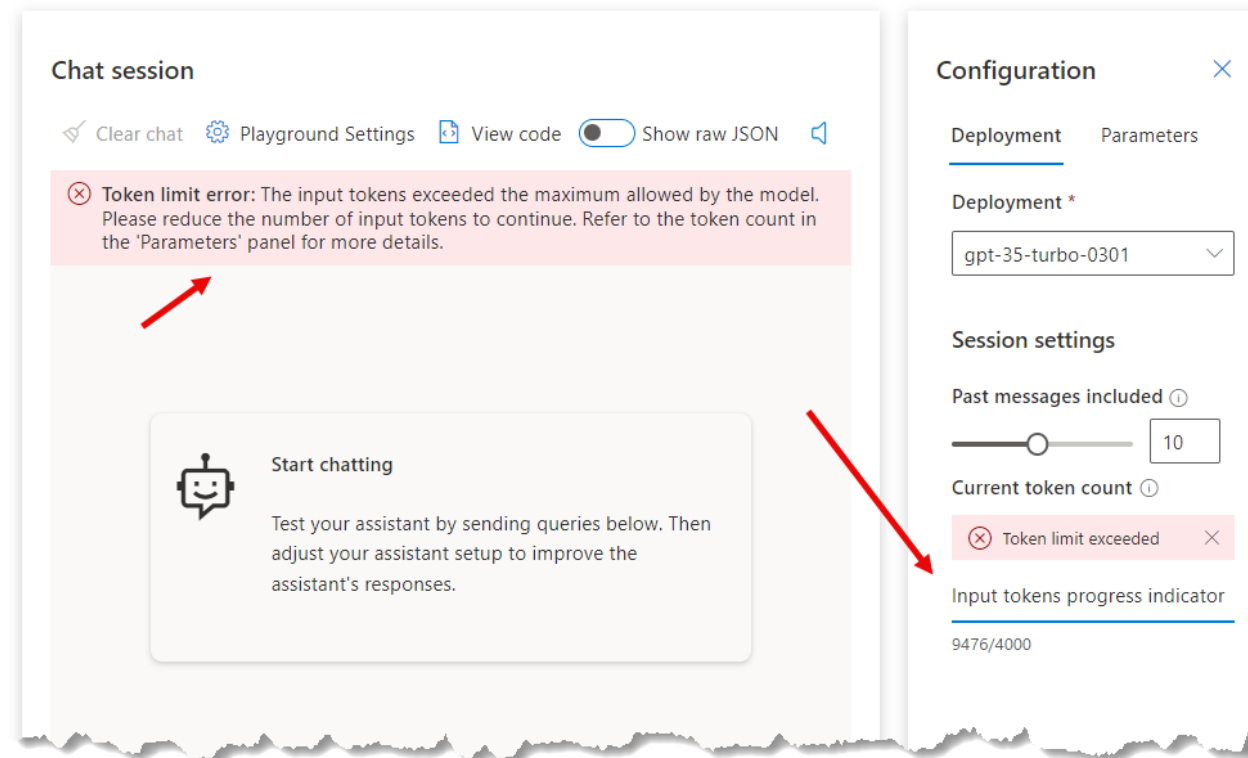
The 22-year-old fourth seed – who missed a match point in the third-set tiebreak for a straight-sets victory – was dominant throughout, with his Serbian opponent committing 54 unforced errors in an uncharacteristic performance.

</CONTEXT>

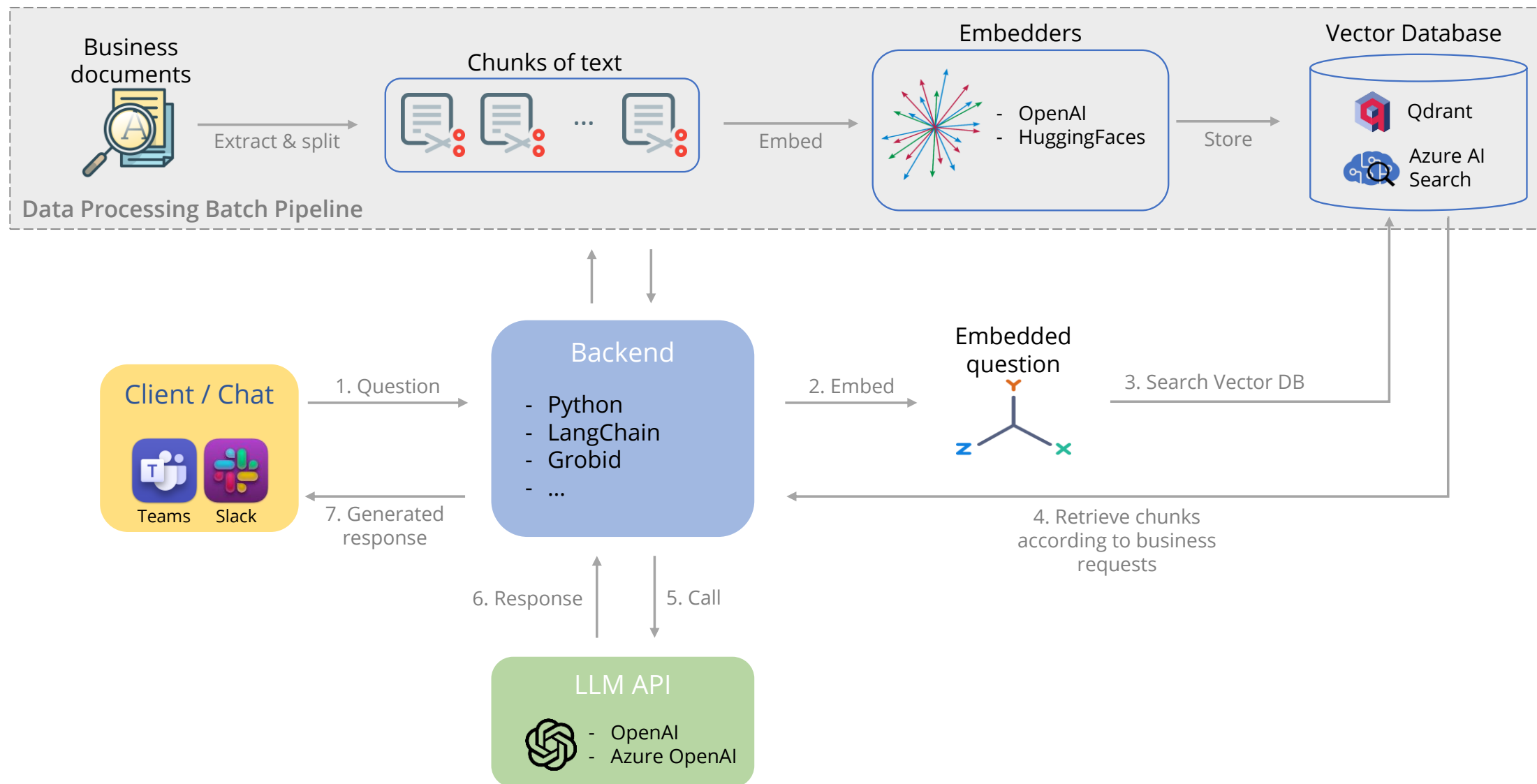
How did the match between Sinner and Djokovic end in the Australian Open?

# Why Not Drop Everything Into a Prompt?

I have a knowledge base of several GB that I can query with ChatGPT.  
I **put everything** in the prompt as context and that's it!

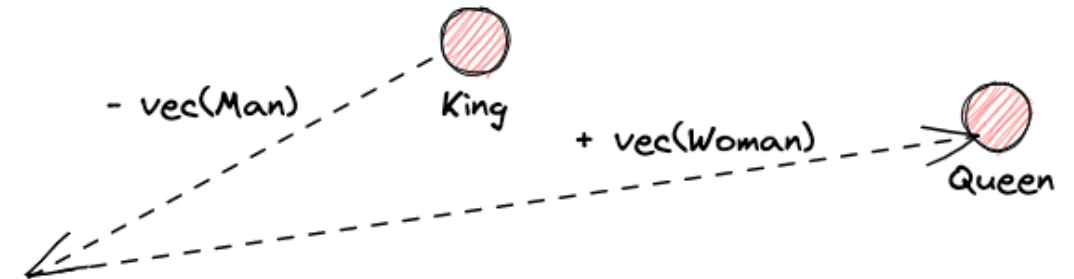


# Document Knowledge Exploration with GPTs

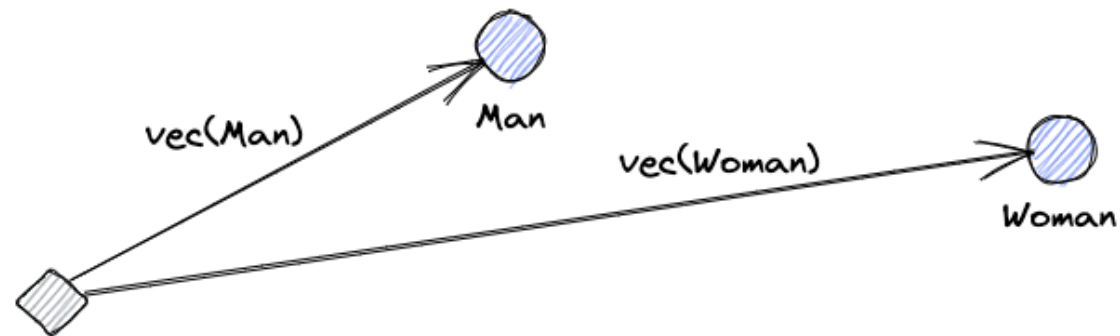


# What Are Vector Embeddings?

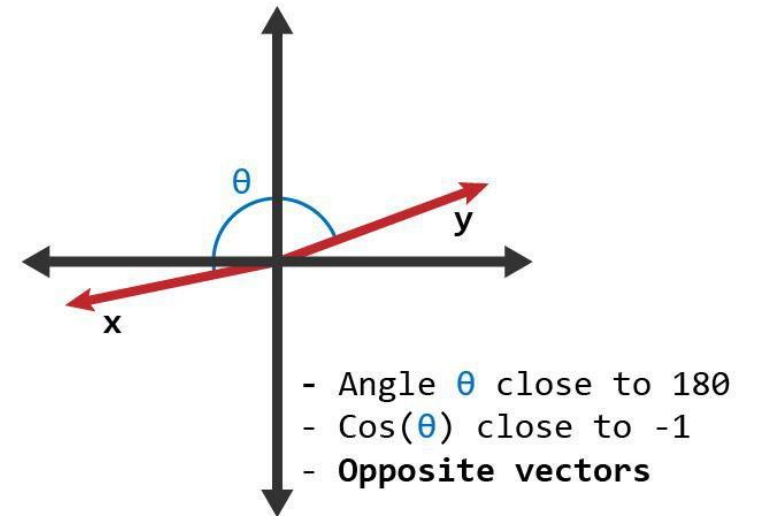
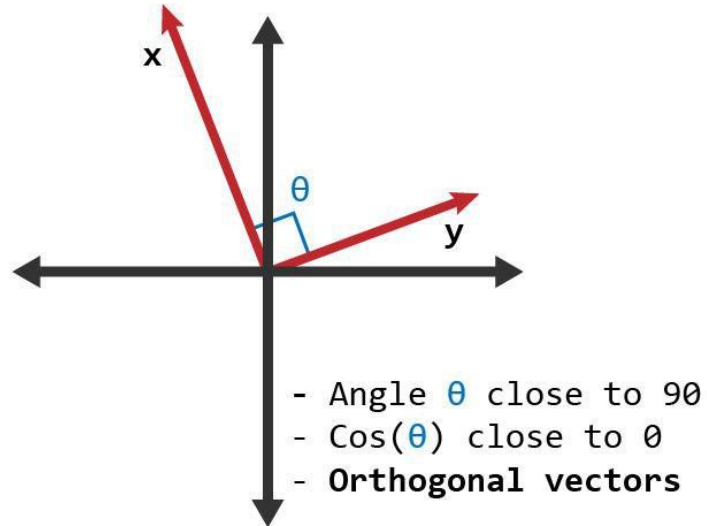
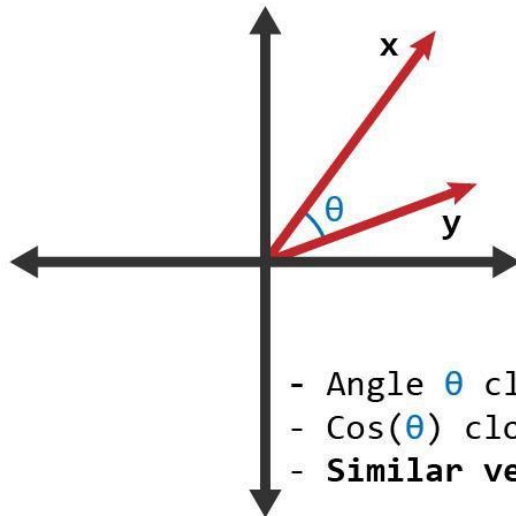
- > A neural network learns to transform text to a **vector space** that contains their actual meaning
- > This is more useful because it can find synonyms and the **syntactic and semantic relationships** between words.



$$\text{Queen} = \text{King} - \text{Man} + \text{Woman}$$



# Cosine-Similarity Search



# DEMO 1

Play with Azure OpenAI  
on your data

- > Theory of Mind Might Have Spontaneously Emerged in Large Language Models (<https://arxiv.org/abs/2302.02083>)
- > Prompt Engineering Guide (<https://www.promptingguide.ai/>)
- > Azure OpenAI on your data (<https://learn.microsoft.com/en-us/azure/ai-services/openai/concepts/use-your-data>)



THANK YOU!! 😊