



Empowering Knowledge Exploration with ChatGPT and Azure OpenAI

Luca Zavarella



Sponsor & Org



UNIVERSITÀ DEGLI STUDI DI PARMA



Bi Factory

DATA KNOWLEDGE ADVISOR



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*" published by Packt

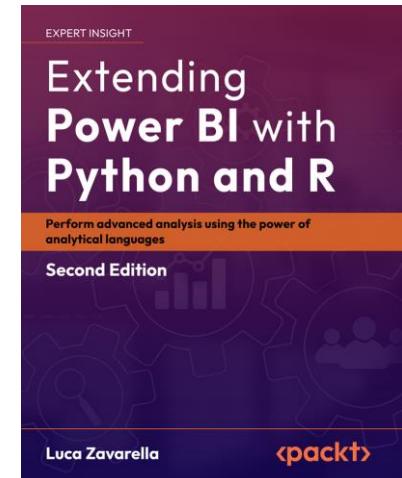
Head of Data & AI @ 

Email: luca.zavarella@icubed.it

Twitter: [@lucazav](https://twitter.com/lucazav)

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

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



Agenda

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

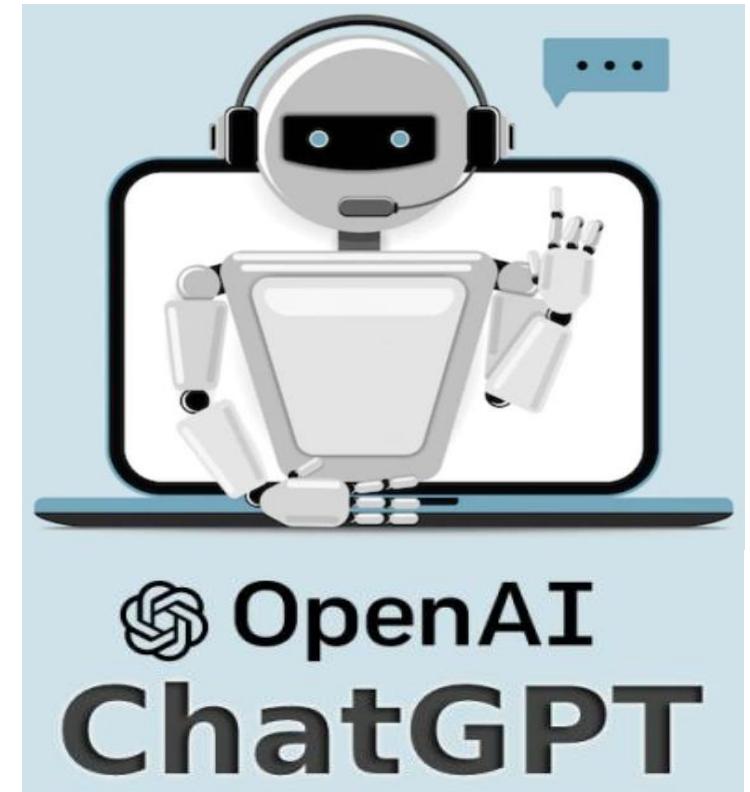
What is ChatGPT and How Does it Work?

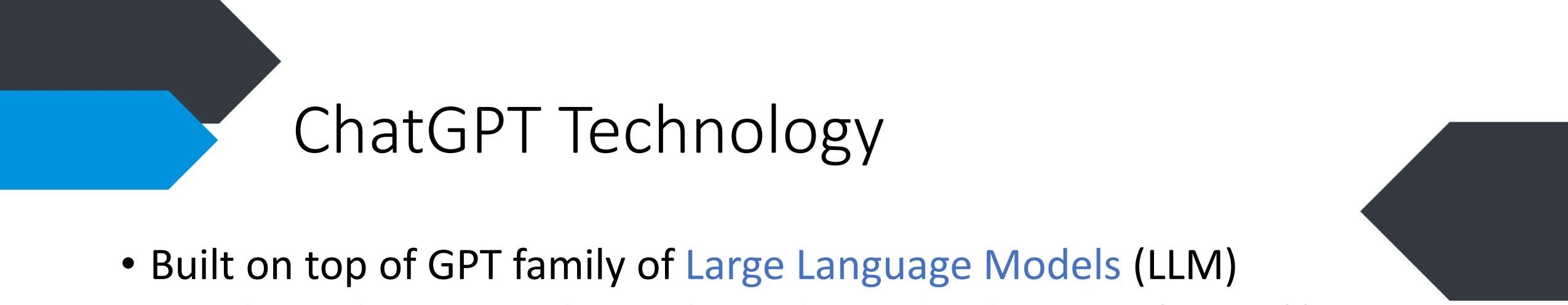


What is ChatGPT

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



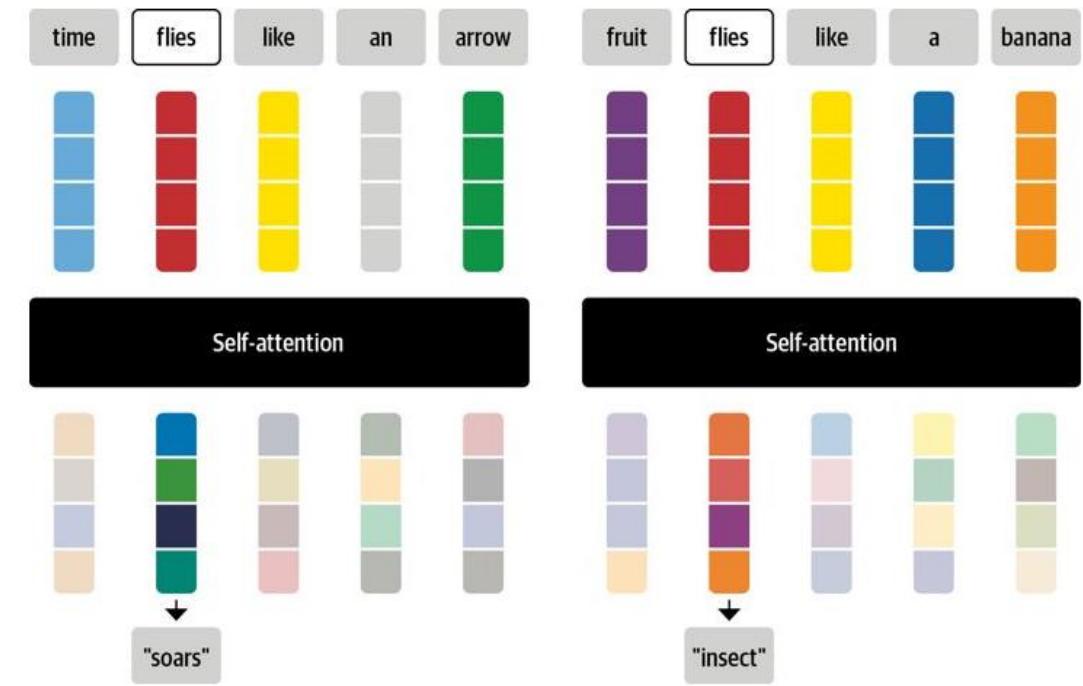


ChatGPT Technology

- 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)**
- A class of **Artificial Intelligence** algorithms that **can generate new and original data** based on patterns it has learned from a training dataset
 - This contrasts with **discriminative AI** algorithms that can classify, or label data based on pre-defined categories

Transformer Foundation Models

- 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



What ChatGPT 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% |

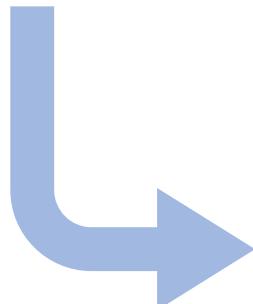


Randomness of ChatGPT Output

- 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

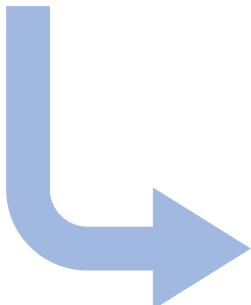
{The best thing about AI is its ability to,
The best thing about AI is its ability to learn,
The best thing about AI is its ability to learn from,
The best thing about AI is its ability to learn from experience,
The best thing about AI is its ability to learn from experience.,
The best thing about AI 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, }

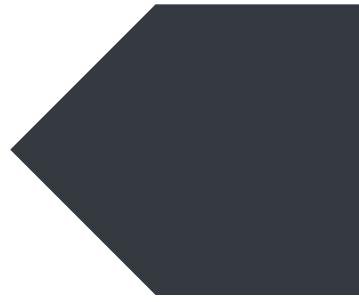


The best thing about AI is its ability to see through, and make sense of, the world around us, rather than panicking and ignoring. This is known as AI "doing its job" or AI "run-of-the-mill." Indeed, taking an infinite number of steps, developing a machine that can be integrated with other systems, or 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
- ChatGPT-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 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



Why is it so successful? – 1/2

It is generative! 🤖



You

Introduci i presenti alla mia sessione "Empowering Knowledge Exploration with ChatGPT and Azure OpenAI" con una frase accattivante che fa riferimento alle pillole rossa e blu usate da Morpheus nel film Matrix



ChatGPT

Certo! Per introdurre la tua sessione "Empowering Knowledge Exploration with ChatGPT and Azure OpenAI", potresti usare una frase come questa:

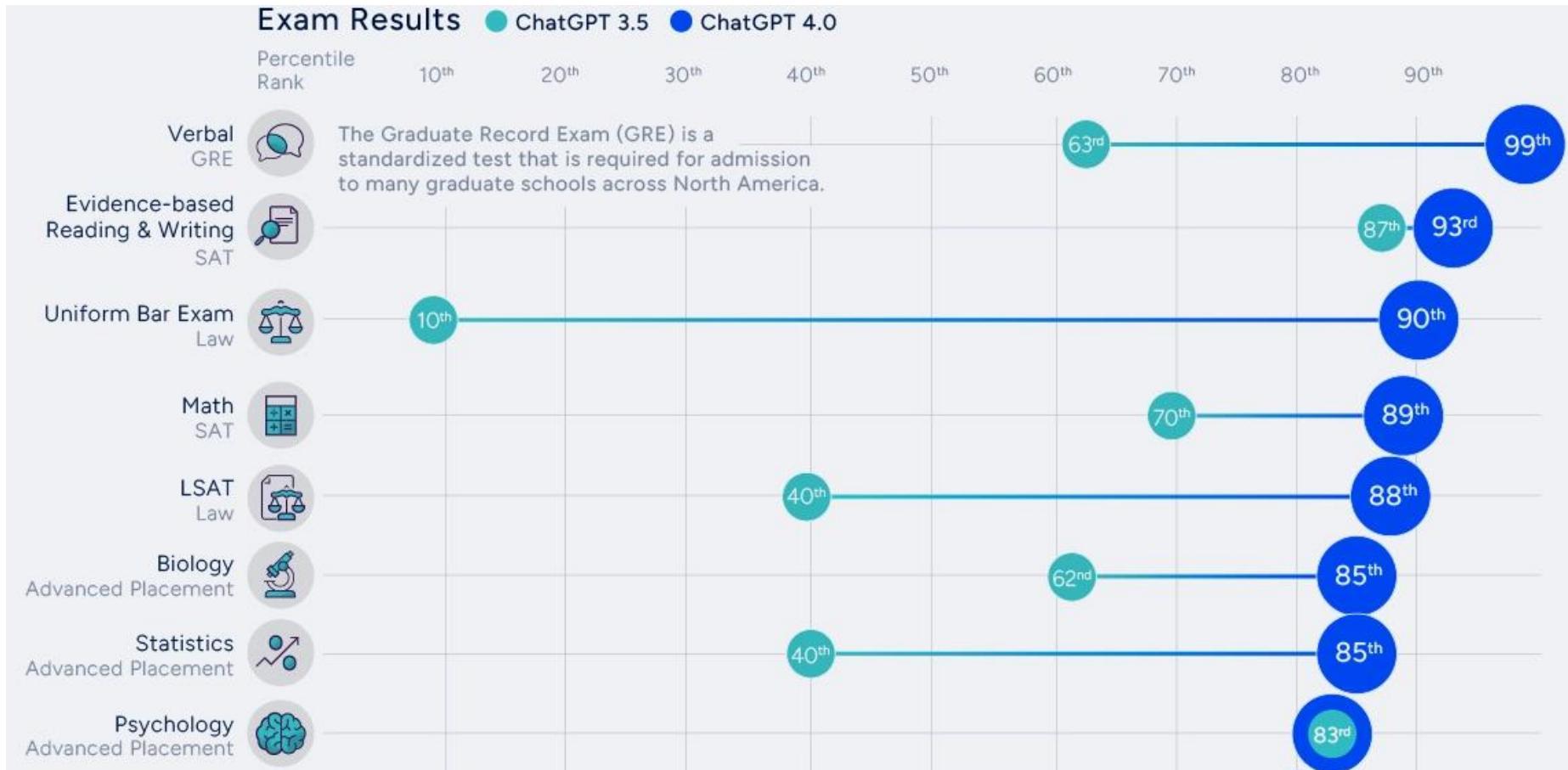
"Benvenuti alla sessione 'Empowering Knowledge Exploration with ChatGPT and Azure OpenAI', dove vi offriamo la scelta tra la pillola rossa e la pillola blu di Matrix: scegliete la pillola rossa per svelare le straordinarie possibilità dell'intelligenza artificiale e del machine learning, o la pillola blu per rimanere nella comfort zone della tecnologia tradizionale. La scelta è vostra, ma ricordate, scegliendo la pillola rossa, vi mostreremo quanto è profondo il coniglio della IA!"



Why is it so successful? – 2/2

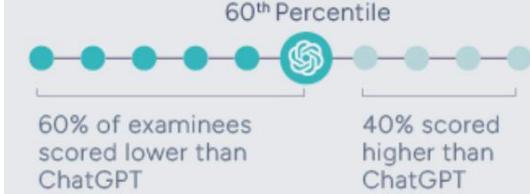
- These models supposedly have one directive
 - To accept a string of text as input and **predict what comes next**, over and over, based purely on statistics
- Computer scientists anticipated that scaling up would boost performance on known tasks
- They didn't expect the models to suddenly handle so 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?



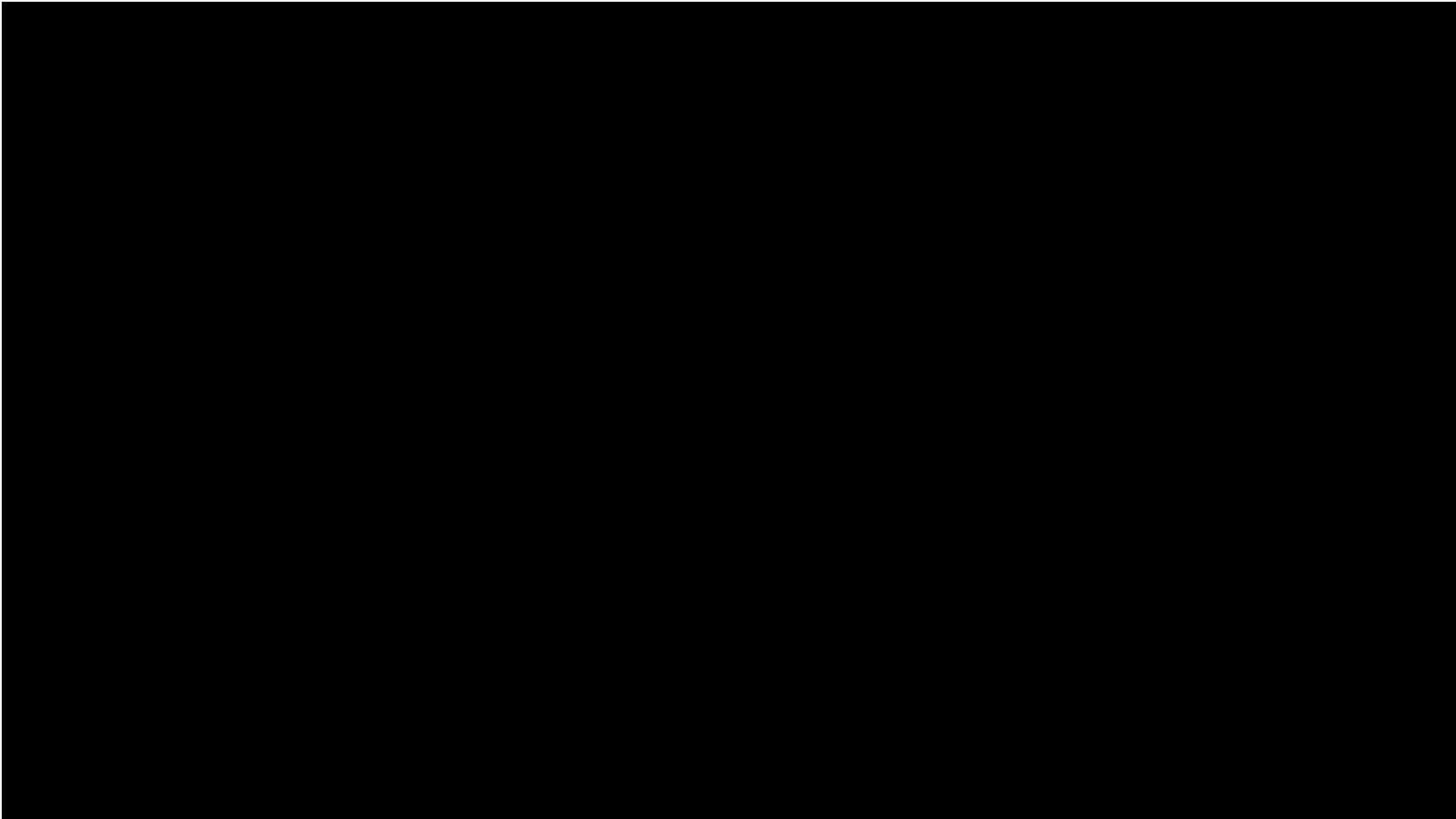
A percentile describes how an examinee's score ranks in comparison to others.

For example





The Magic of Complex Systems



Emergent Behavior Examples

Advanced Problem Solving 1/2



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 osto disonesto e ubriacone beve 6 litri di vino da un barile che ne contiene 360 litri 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'osto disonesto?



ChatGPT

Per calcolare la quantità di vino bevuto dall'osto disonesto, dobbiamo considerare che ogni volta che preleva vino dal barile e lo sostituisce con acqua, la concentrazione di vino nel barile diminuisce.

All'inizio, il barile contiene 360 litri di vino. Ogni volta che l'osto preleva 6 litri di vino e li sostituisce con acqua, il contenuto totale del barile rimane 360 litri, ma la quantità di vino puro diminuisce.

Calcoliamo la quantità di vino puro rimasta nel barile dopo ogni prelievo e sostituzione:

1. Dopo il primo prelievo, l'osto ha bevuto 6 litri di vino puro. Quindi, nel barile rimangono $360 - 6 = 354$ litri di vino puro.
2. Quando l'osto preleva la seconda volta, non sta prelevando 6 litri di vino puro, ma una miscela di vino e acqua. La concentrazione di vino in questa miscela è $\frac{354}{360}$. Quindi, l'osto beve $6 \times \frac{354}{360}$ litri di vino puro.
3. La terza volta che l'osto preleva, la concentrazione di vino nel barile è ulteriormente diminuita.

Dopo il secondo prelievo, nel barile rimangono $354 - 6 \times \frac{354}{360}$ litri di vino puro. La concentrazione di vino nella miscela durante il terzo prelievo è $\frac{354 - 6 \times \frac{354}{360}}{360}$, e l'osto beve $6 \times \frac{354 - 6 \times \frac{354}{360}}{360}$ litri di vino puro.

Sommendo queste quantità otteniamo la quantità totale di vino puro bevuto dall'osto. Calcoliamola.

L'osto disonesto ha bevuto complessivamente circa 17,70 litri di vino puro.[\[→\]](#)

Advanced Problem Solving 2/2

```
[{  
  "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"  
      }  
    }  
  }]
```



You

Given [{"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": 1000000.0, "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"}}]

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



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

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

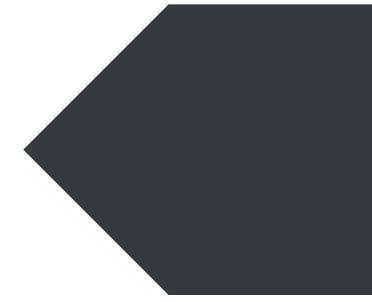
- GPT-3 solves around 40% of false-belief tasks
 - Comparable to 3.5-year-old children
- GPT-3.2 achieves 70% accuracy
 - Comparable to six-year-olds
- GPT-3.5 astounding 90% accuracy
 - Comparable to seven-year-olds

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 is a generative model
- It is not always able to provide accurate answers to specific questions
- In some cases, the generated responses may be irrelevant or nonsensical (*hallucinations*)

Unlocking Business Potential with Prompt Engineering



Prompt Engineering for LLMs

Optimizing Prompts

Refining prompt design to enhance LLM effectiveness

Skillful Guidance

Utilizing techniques to direct LLMs in varied tasks

Capability Enhancement

Boosting LLM performance and ensuring safe operation

Knowledge Integration

Merging LLMs with specialized knowledge and tools

Injecting Information into Prompts for Precise Query Responses



You

Given the following context:

==== CONTEXT ===

In a remarkable turn of events at the ATP Finals in Turin, Italy, Jannik Sinner achieved his maiden victory over the world No. 1, Novak Djokovic, igniting the passionate home crowd. This electrifying contest endured for more than three hours, ultimately resulting in Sinner's 7-5, 6-7(5), 7-6(2) triumph. The 22-year-old Italian, a local favorite, clinched his first win against Djokovic in their four previous encounters, and in doing so, halted Djokovic's impressive 19-match winning streak. Sinner's victory propelled him to the top of the Green Group standings.

=====

How did the match between Sinner and Djokovic end up in the 2023 ATP Finals?



ChatGPT

The match between Jannik Sinner and Novak Djokovic at the 2023 ATP Finals ended with Jannik Sinner achieving a thrilling victory. Sinner won the match with a scoreline of 7-5, 6-7(5), 7-6(2) after more than three hours of intense play. This victory was particularly significant for Sinner as it was his first win against Djokovic in their four previous encounters. Additionally, Sinner's win halted Djokovic's impressive 19-match winning streak. This victory also helped Sinner secure a top position in the Green Group standings at the ATP Finals.

|||||

Your enterprise
data



You

How did the match between Sinner and Djokovic end up in the 2023 ATP Finals?



ChatGPT

I'm sorry, but I do not have access to real-time information or events that have occurred after my last knowledge update in January 2022. Therefore, I cannot provide you with the specific details of the match between Jannik Sinner and Novak Djokovic in the 2023 ATP Finals. To find out the result of that match, I recommend checking a reliable sports news website, the official ATP website, or a tennis news source for the most up-to-date information. Is there anything else I can help you with today?

|||||

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 question text and solve it!



The screenshot shows the Chat session and Configuration panels of the ChatGPT interface.

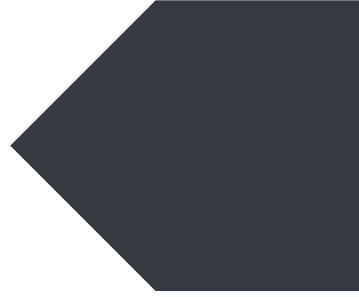
Chat session: Displays a message about a "Token limit error".

Configuration:

- Deployment:** Set to "gpt-35-turbo-0301".
- Session settings:** Shows "Past messages included" set to 10 and "Current token count" showing "Token limit exceeded".
- Input tokens progress indicator:** Shows "9476/4000".

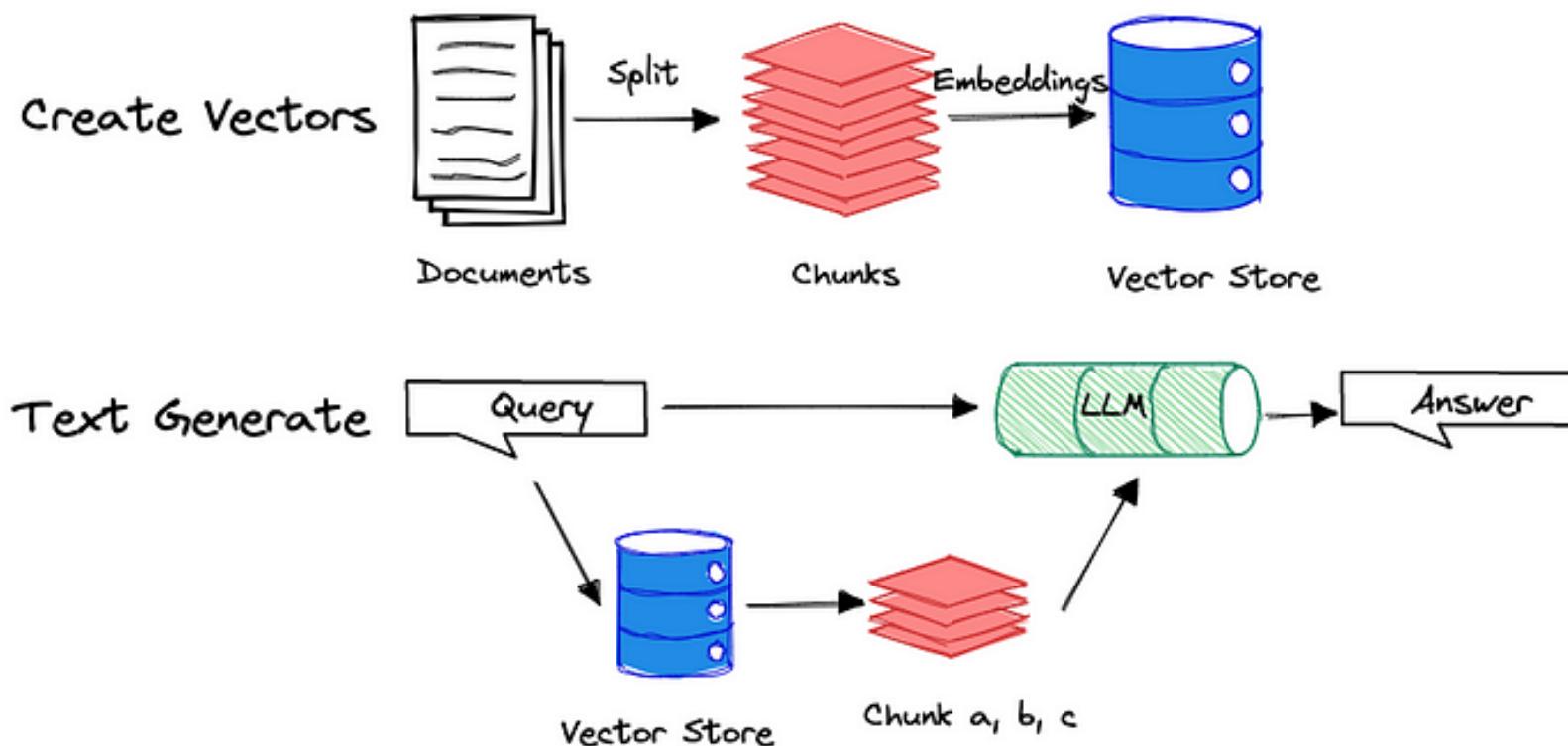
Two red arrows point from the text in the Chat session panel to the "Token limit exceeded" message in the Configuration panel.

Document Knowledge Exploration with ChatGPT



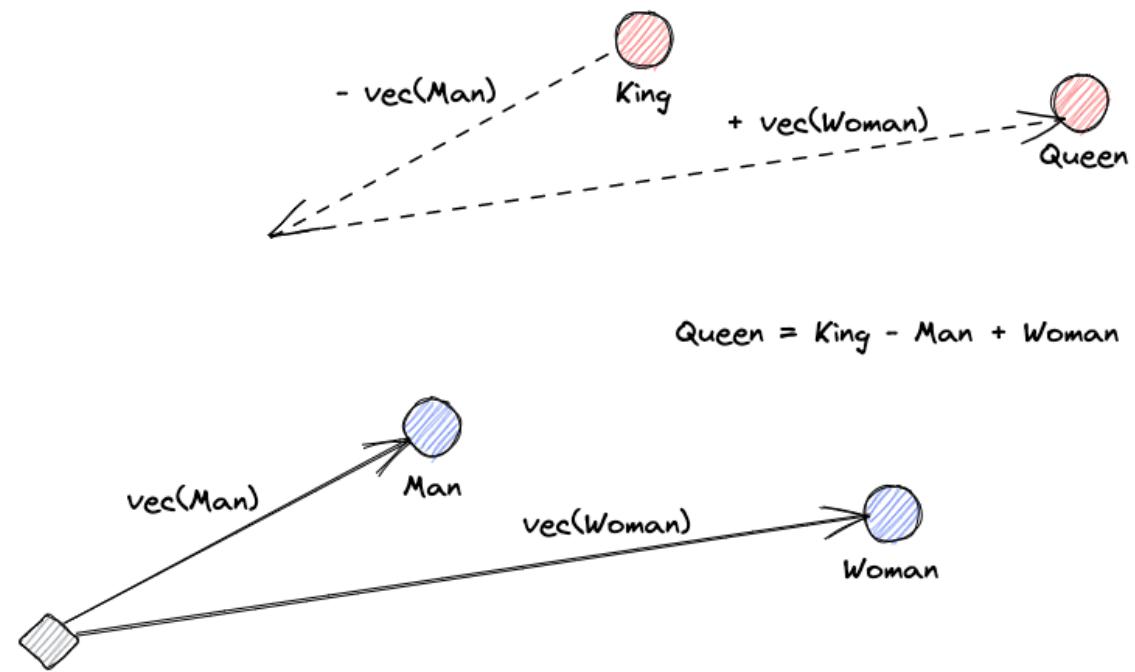
RAG Framework Flow

The **Retrieval-Augmented Generation (RAG)** framework is an approach that combines the strengths of both retrieval-based and generative systems

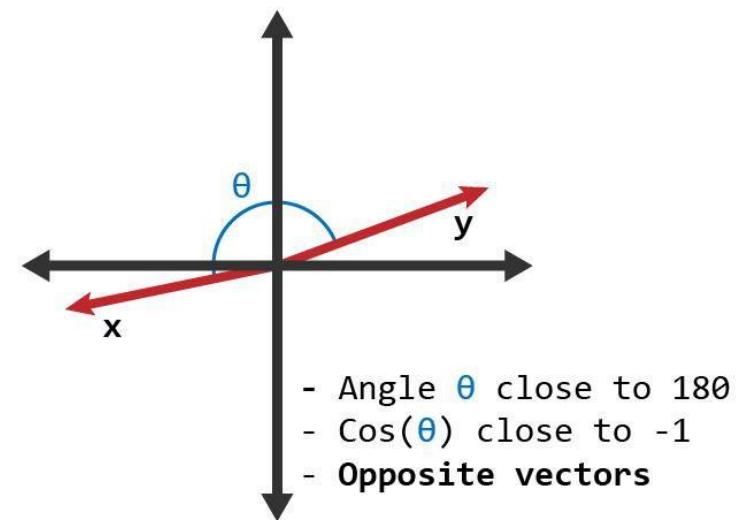
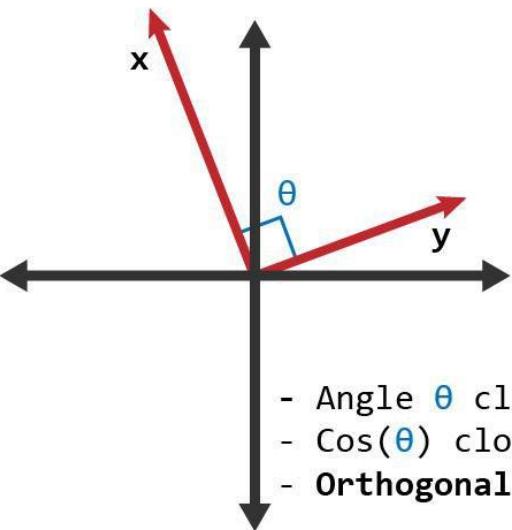
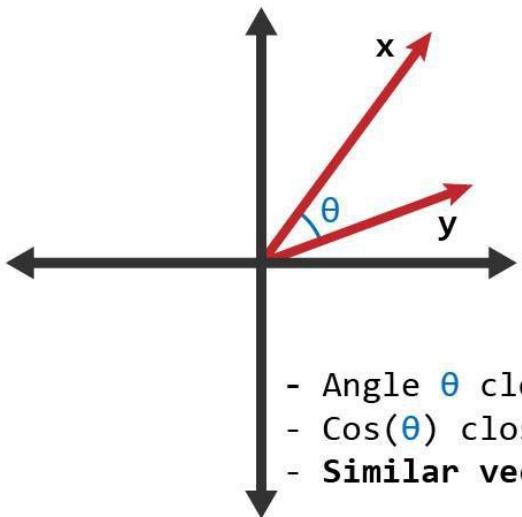


What are Vector Embeddings

- 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 **syntactical and semantical relationships** between words



Cosine-Similarity Search



DEMO 01

Play with Azure OpenAI on your data

References

- 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>)

Waiting for Your Feedback!

Data Saturday #37 Feedback Form





#37 PARMA 2023

Grazie!!!

