



Politecnico
di Torino



Introduzione ai Large Language Model

Luca Cagliero

Dipartimento di Automatica e Informatica - Politecnico di Torino

SmartData@PoliTo

Introduzione ai Large Language Model

- L'Intelligenza Artificiale
- I tipi di dato
- Tecniche principali
- Large Language Model: fondamenti
- Il task Text2SQL

Introduzione ai Large Language Model

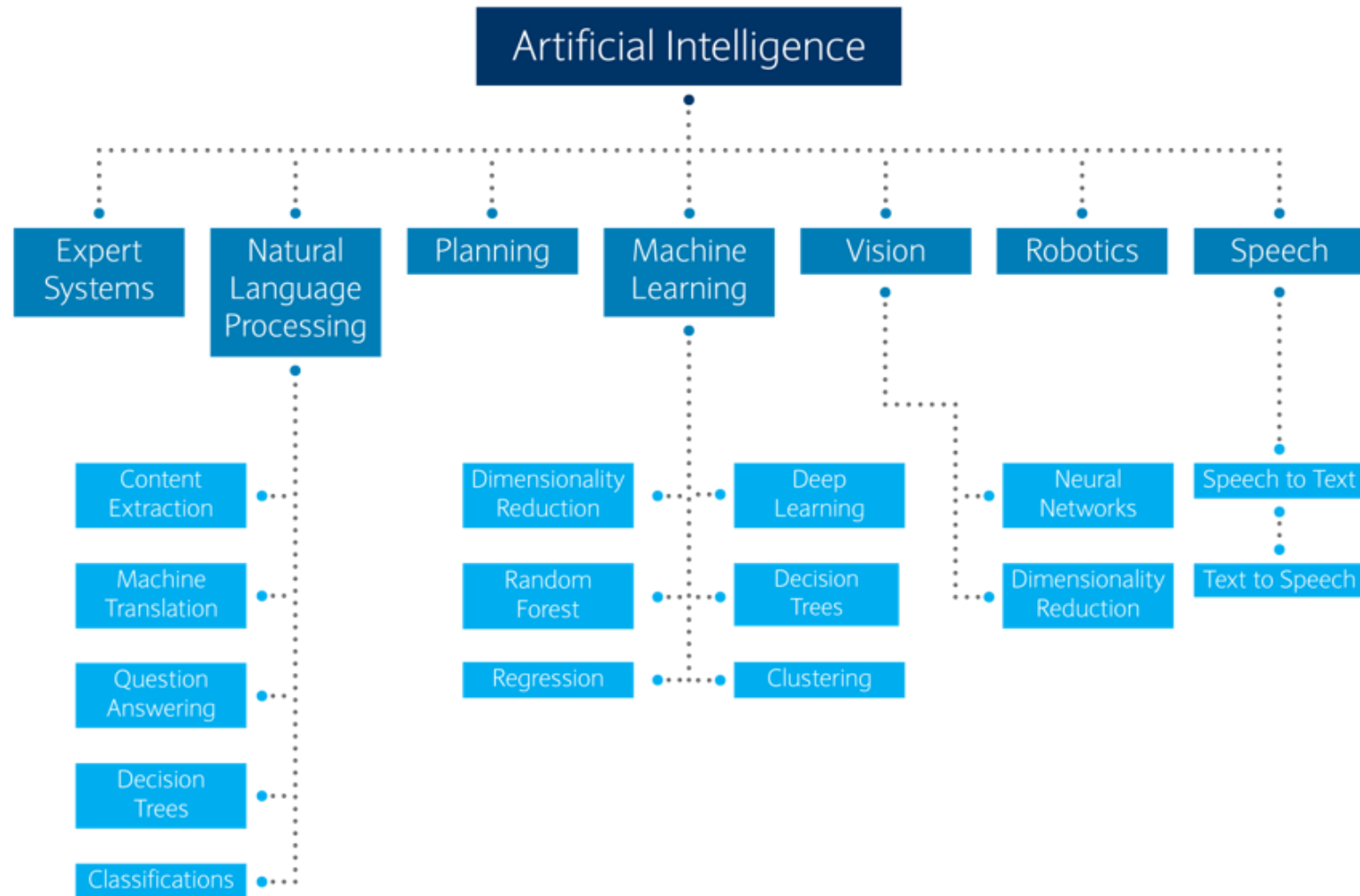
L'intelligenza Artificiale

Definizione di Intelligenza Artificiale (IA)

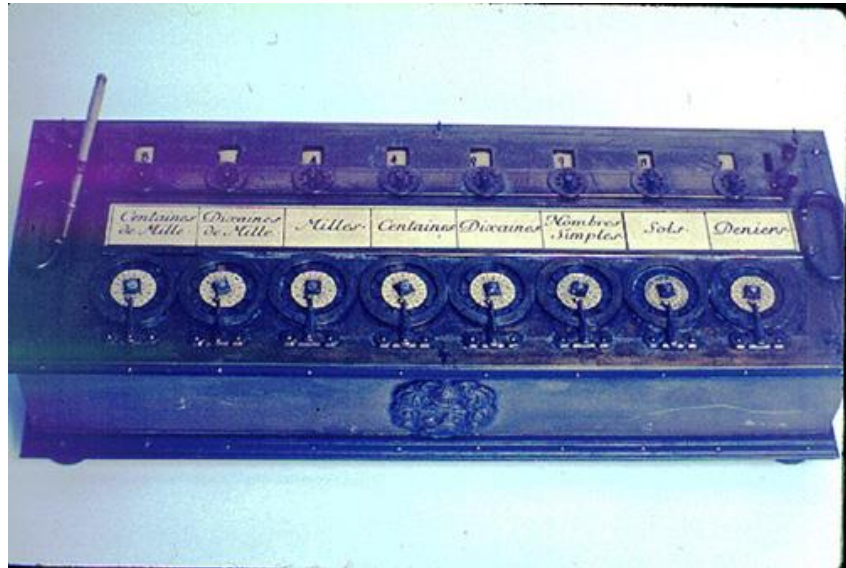
«Disciplina che studia se e in che modo si possano riprodurre i processi mentali più complessi mediante l'uso di un computer. Tale ricerca si sviluppa secondo due percorsi complementari: da un lato l'IA cerca di avvicinare il funzionamento dei computer alle capacità dell'intelligenza umana, dall'altro usa le simulazioni informatiche per fare ipotesi sui meccanismi utilizzati dalla mente umana.»

Intelligenza Artificiale

- Che cos'è l'IA?



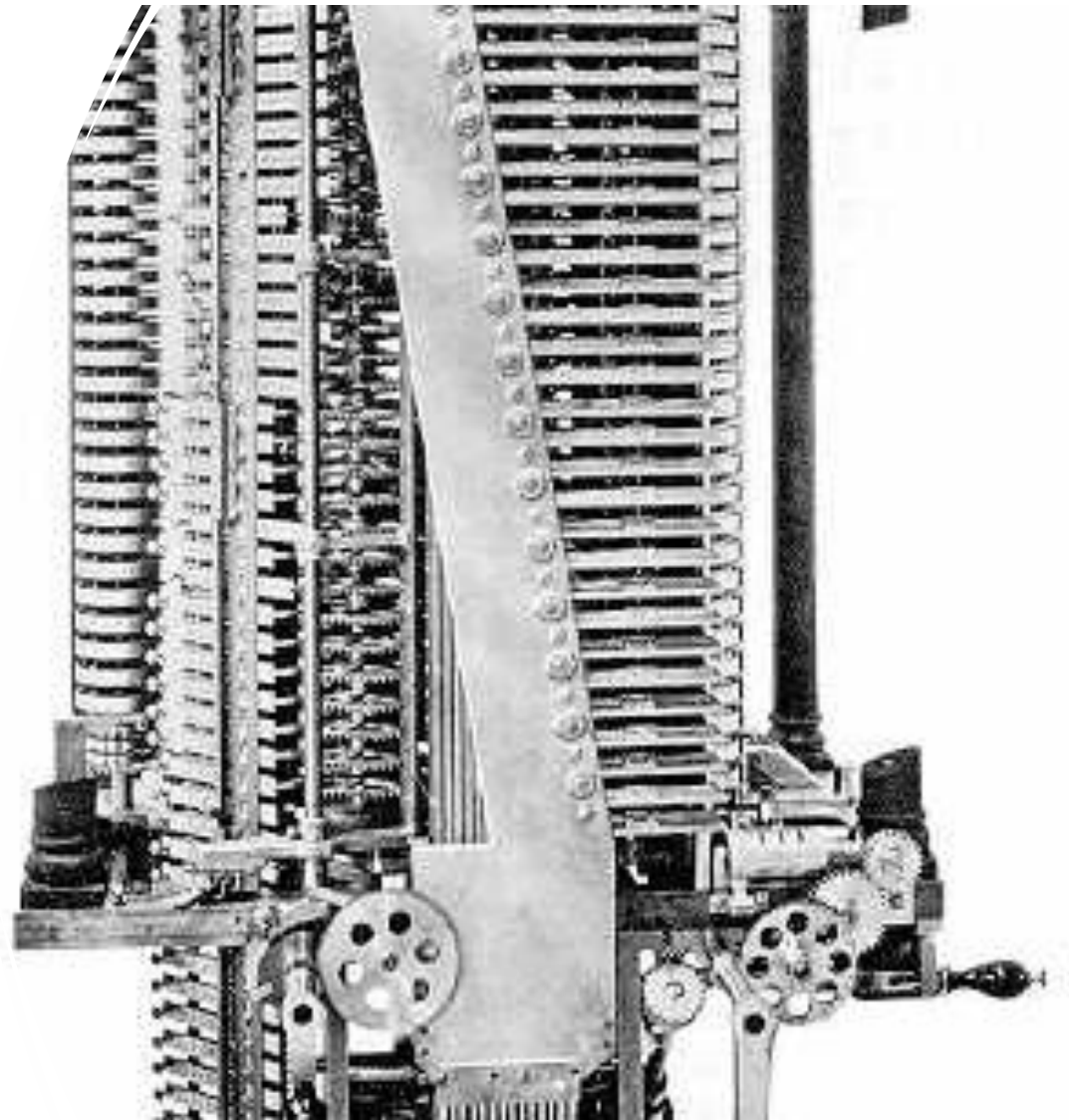
Le radici dell'IA



Pascal



Leibnitz



Charles Babbage

L'evoluzione nel tempo

A.I. TIMELINE

SYZYGY

1950

TURING TEST

Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence



1955

A.I. BORN

Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"



1964

ELIZA

Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans



1966

SHAKY

The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions



1997

DEEP BLUE

Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov



1998

KISMET

Cynthia Breazeal at MIT introduces KISmet, an emotionally intelligent robot insofar as it detects and responds to people's feelings

A.I. WINTER

Many false starts and dead-ends leave A.I. out in the cold



1999

AIBO

Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time



2002

ROOMBA

First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes



2011

SIRI

Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S



2011

WATSON

IBM's question answering computer Watson wins first place on popular \$1M prize television quiz show Jeopardy



2014

EUGENE

Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human



2014

ALEXA

Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks



2016

TAY

Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments



2017

ALPHAGO

Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

OpenAI



GPT-4



deepseek

2025

Intelligenza artificiale, Trump annuncia investimenti multimiliardari



ECONOMIA | Martedì 28 gennaio 2025

Il crollo in borsa delle aziende di tecnologia a causa di DeepSeek

Nvidia ha perso il 17 per cento e quasi tutto il settore è andato malissimo per le preoccupazioni attorno alla nuova startup cinese di intelligenza artificiale

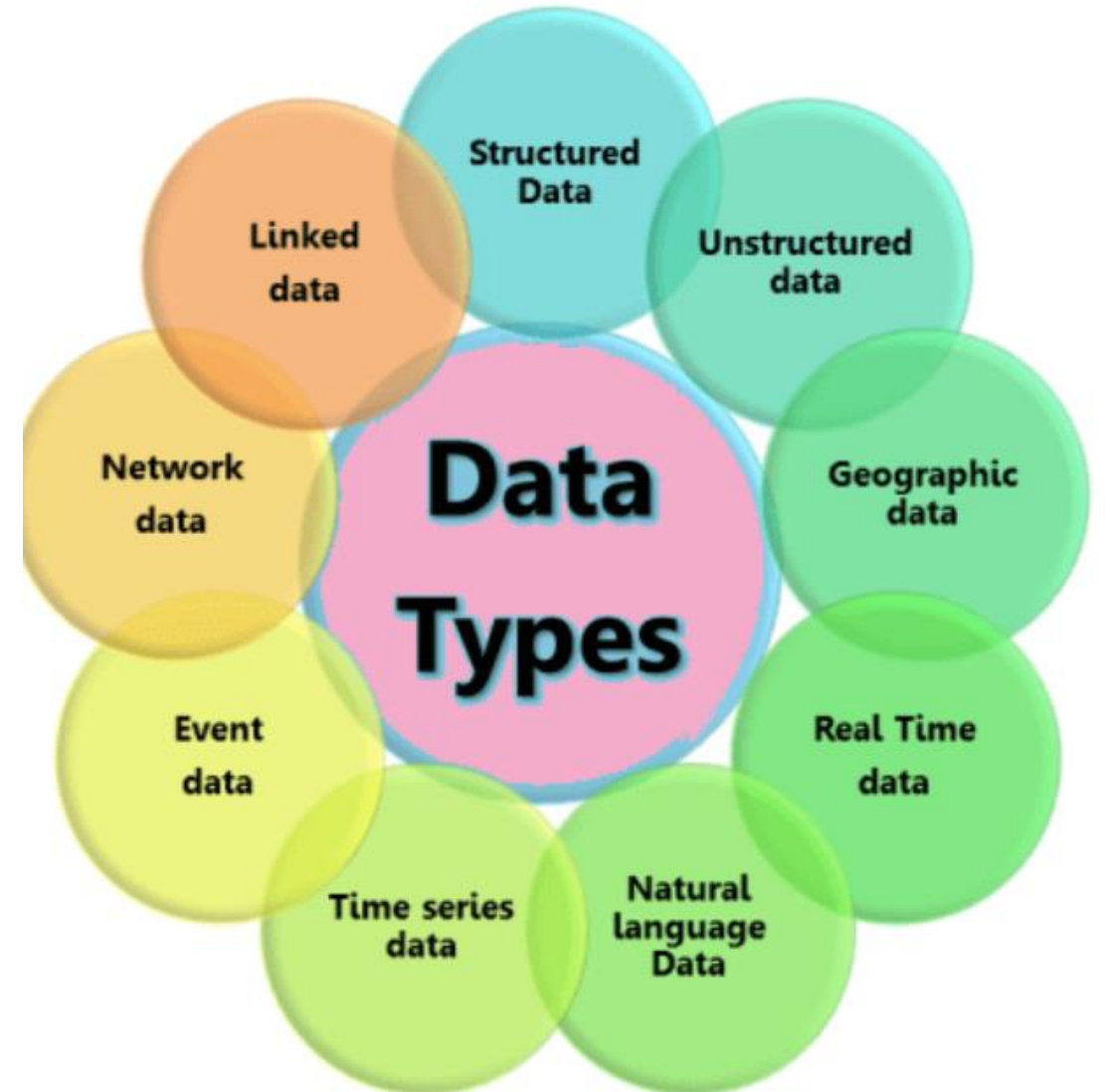


Introduzione ai Large Language Model

I tipi di dato

I tipi di dato

- Tabella
- Documento testuale
- Dato transazionale
- Grafo
- Sequenza
- Serie temporale
- Dato geospaziali
- Immagine
- Video
- Segnale audio
- Ecc.



I tipi di dato

- Tabella
 - “Relazione” secondo il modello relazionale
 - Esempio tipico di **dato strutturato**
 - Collezione di record (sample) descritti da un insieme prefissato di attributi
 - **Metadati**, ovvero “dati che descrivono dati”
 - Gli attributi sono variabili, campi, dimensioni, feature, descrittori, ecc.

Attributes

<i>Tid</i>	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

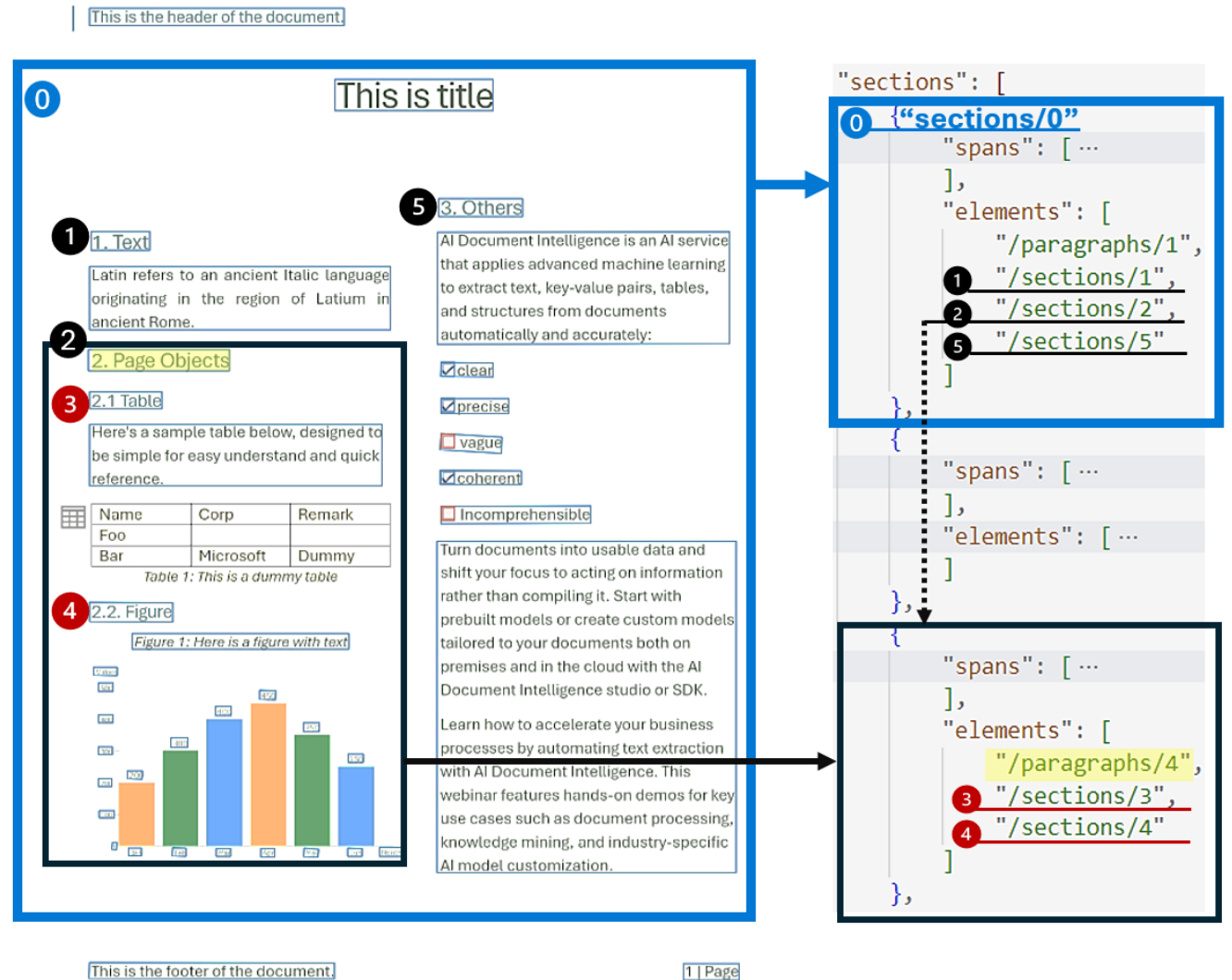
Samples

I tipi di dato

- Documento testuale
 - Semi-strutturato o destrutturato



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



<https://learn.microsoft.com/it-it/azure/ai-services/>

I tipi di dato

- **Documento testuale**

- Posso essere rappresentati in formato tabulare
- Ogni record è un documento
- Ogni attributo è un'unità di testo
 - Parola, frase, ecc.
- I valori indicano l'importanza dell'unità di testo nel documento
 - Frequenza, presenza/assenza, ecc.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

I tipi di dato

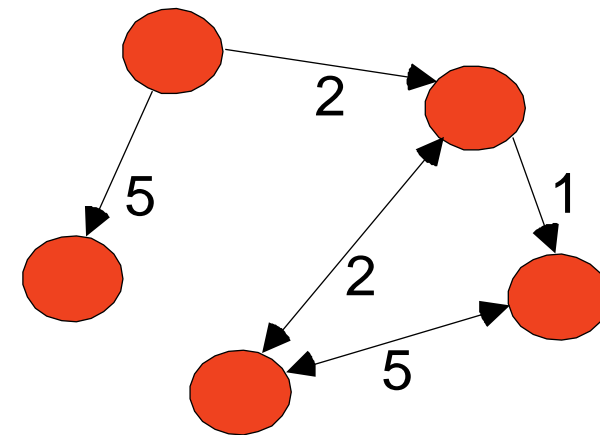
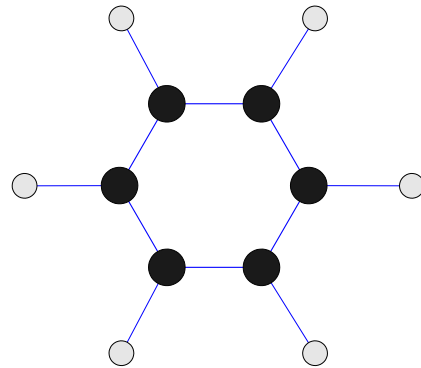
- **Dataset** (set di dati) **transazionale**
 - Ogni transazione è un insieme di oggetti, definiti “item”, all’interno di un vocabolario
 - Le transazioni possono contenere un numero arbitrario di item

<i>TID</i>	<i>Items</i>
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

I tipi di dato

- Grafo

- Reti di nodi connessi da archi
- Tipologie
 - Grafo orientato /non orientato: ogni arco possiede/non possiede un verso
 - Grafo pesato/non pesato: ogni arco o nodo ha/non ha un peso (rilevanza)

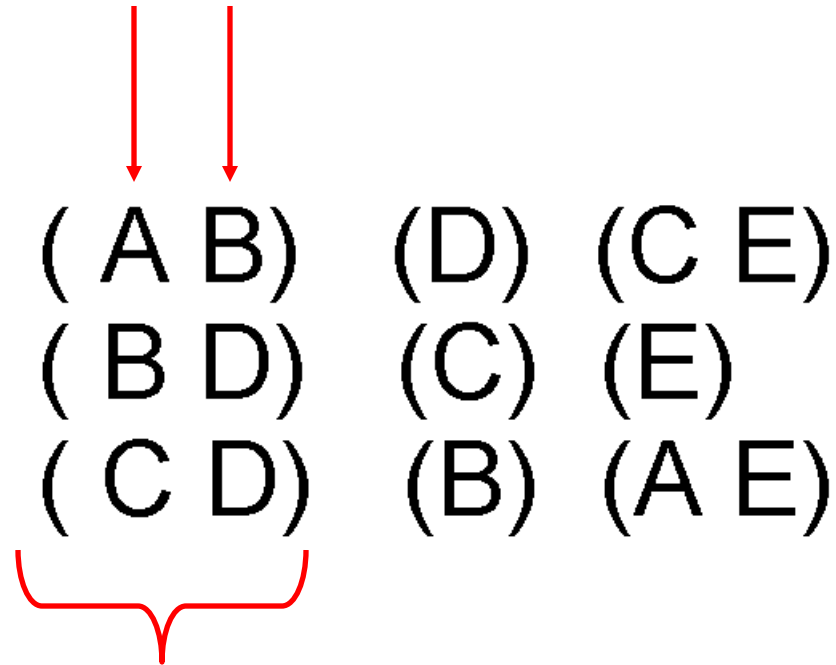


I tipi di dato

- Sequenza

- Sequenza **ordinata** di transazioni/item/eventi

Items/Events



**An element of
the sequence**

I tipi di dato

- Sequenza

- Sequenza ordinata di transazioni/item/eventi

**GGTTCGCTTCAGCCCCGCGCC
CGCAGGGCCCCGCCCCGCGCCGTC
GAGAAGGGCCCGCCTGGCGGGCG
GGGGGAGGCGGGGCCGCCCGAGC
CCAACCGAGTCCGACCAGGTGCC
CCCTCTGCTCGGCCTAGACCTGA
GCTCATTAGGCGGCAGCGGACAG
GCCAAGTAGAACACGCGAAGCGC
TGGGCTGCCTGCTGCGACCAGGG**

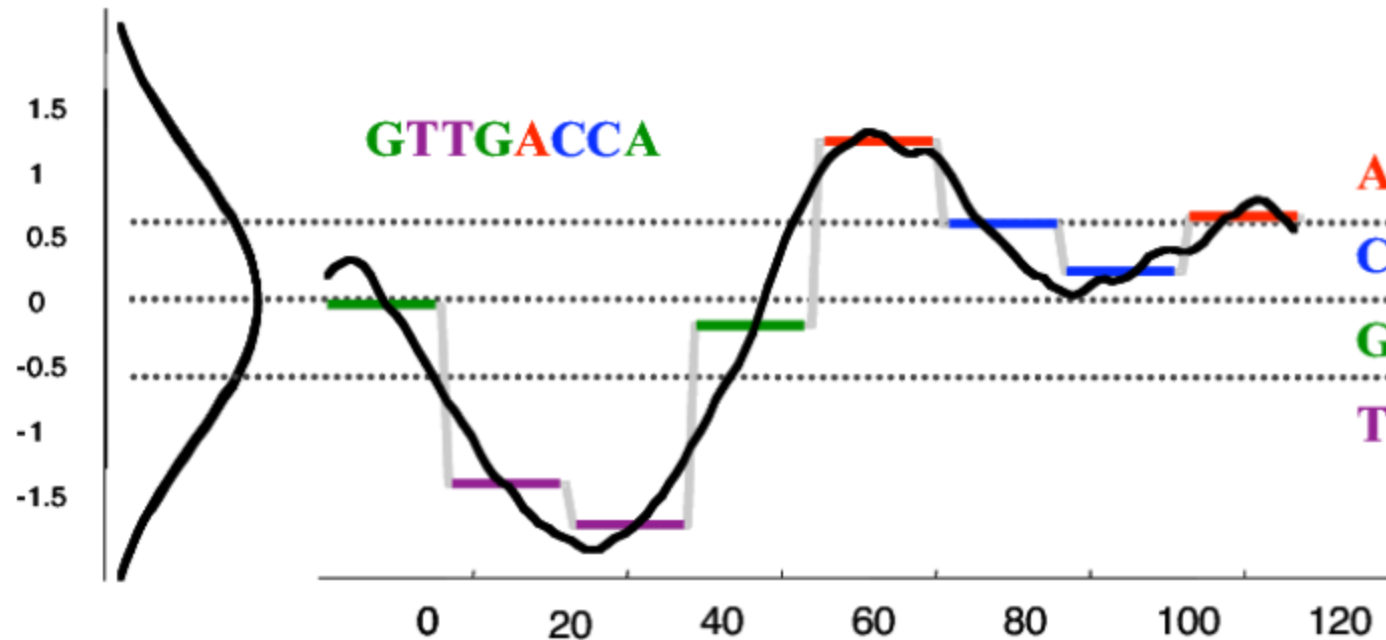
I tipi di dato

- **Serie temporale** (time series)
 - Sequenza di elementi ordinati temporalmente
 - Granularità temporale
 - Intervallo temporale tra un sample e il successivo



I tipi di dato

- Serie temporale (time series)
 - **Symbolic Aggregate Approximation (SAX)**
 - Trasformazione da valori continui a valori discreti

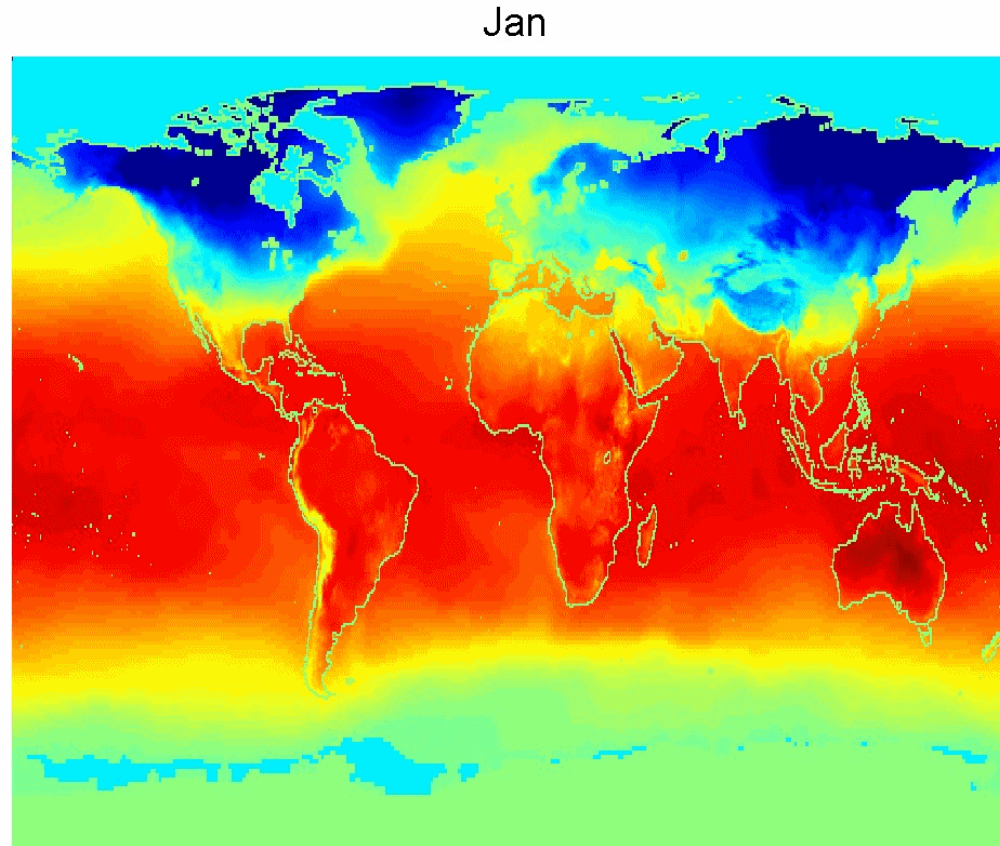


Keogh et al. (2006) Intelligent Icons: Integrating intelligent data mining and visualization into GUI operating systems

I tipi di dato

- Dataset **spazio-temporale**

**Average Monthly
Temperature of
land and ocean**



I tipi di dato

- Media digitali

- Audio, immagini, video

- Le immagini sono matrici di pixel
 - Audio e video sono sequenze temporali di frame, fonemi, ecc.



Introduzione ai Large Language Model

Tecniche principali

Intelligenza Artificiale

A.I. TIMELINE

SYZYGY

1950

TURING TEST

Computer scientist Alan Turing proposes a test for machine intelligence. If a machine can trick humans into thinking it is human, then it has intelligence

1955

A.I. BORN

Term 'artificial intelligence' is coined by computer scientist, John McCarthy to describe "the science and engineering of making intelligent machines"

1961

UNIMATE

First industrial robot, Unimate, goes to work at GM replacing humans on the assembly line

1964

ELIZA

Pioneering chatbot developed by Joseph Weizenbaum at MIT holds conversations with humans

1966

SHAKY

The 'first electronic person' from Stanford, Shakey is a general-purpose mobile robot that reasons about its own actions

A.I. WINTER

Many false starts and dead-ends leave A.I. out in the cold

1997

DEEP BLUE

Deep Blue, a chess-playing computer from IBM defeats world chess champion Garry Kasparov

1998

KISMET

Cynthia Breazeal at MIT introduces Kismet, an emotionally intelligent robot insofar as it detects and responds to people's feelings



1999

AIBO

Sony launches first consumer robot pet dog AiBO (AI robot) with skills and personality that develop over time



2002

ROOMBA

First mass produced autonomous robotic vacuum cleaner from iRobot learns to navigate and clean homes



2011

SIRI

Apple integrates Siri, an intelligent virtual assistant with a voice interface, into the iPhone 4S



2011

WATSON

IBM's question answering computer Watson wins first place on popular \$1M prize television quiz show Jeopardy



2014

EUGENE

Eugene Goostman, a chatbot passes the Turing Test with a third of judges believing Eugene is human



2014

ALEXA

Amazon launches Alexa, an intelligent virtual assistant with a voice interface that completes shopping tasks



2016

TAY

Microsoft's chatbot Tay goes rogue on social media making inflammatory and offensive racist comments



2017

ALPHAGO

Google's A.I. AlphaGo beats world champion Ke Jie in the complex board game of Go, notable for its vast number (2^{170}) of possible positions

OpenAI



GPT-4

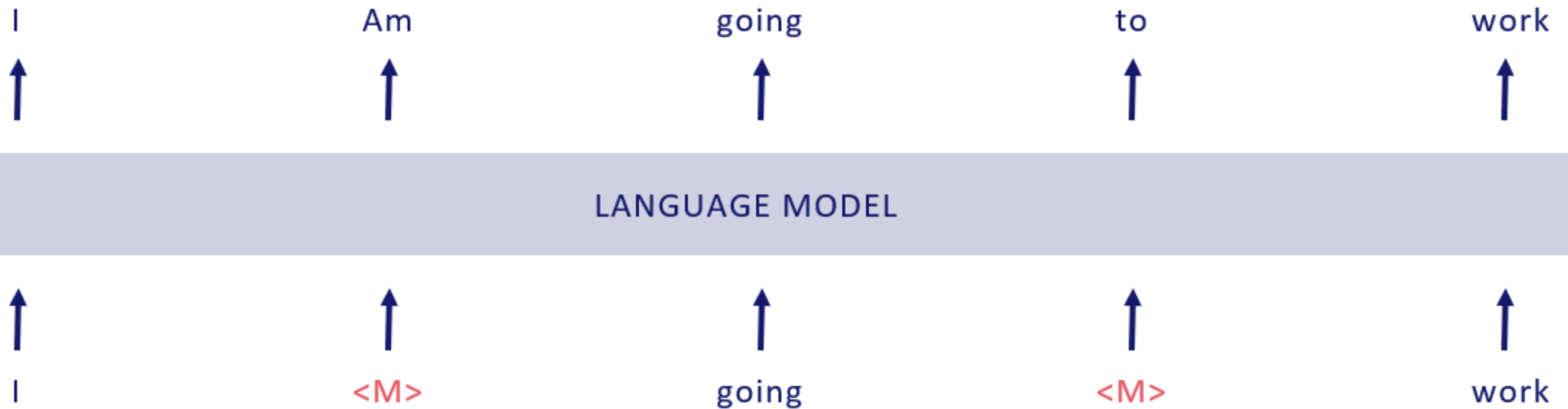
2024

Obiettivo dell'analisi

- Analisi descrittiva
- Analisi predittiva
- Analisi prescrittiva



Tecniche predittive per l'apprendimento del linguaggio



Tecniche predittive per l'apprendimento del linguaggio

- **Next Token Prediction:** Data una sequenza di token (parole o altre unità di testo) predire il token successivo nella sequenza
- Imparo da enormi collezioni di documenti non necessariamente annotati
- L'annotazione la genero automaticamente nascondendo il next token ("three") e allenando il modello di IA a predirlo correttamente

Enter text:

One, two,



3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Classificazione delle tecniche

- **Next Token Prediction** è il modello alla base del funzionamento dei principali Large Language Model



Enter text:
One, two,



3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Introduzione ai Large Language Model

Large Language Model: fondamenti

Large Language Model

- Modello di AI specializzato nella comprensione e generazione di testo in linguaggio naturale
- Addestrato su enormi moli di dati (TeraByte)
 - Necessita ingenti quantità di GPU, memoria e tempo macchina per l'addestramento
- Addestramento solitamente su un mix di tecniche
 - Modelli predittivi, Reinforcement Learning, ecc.

Large Language Model

- Come generare una risposta a una domanda formulata in linguaggio naturale?
- Ad ogni step il LM
 - stima la probabilità di occorrenza di ciascun token presente in un vocabolario in base a
 - Sequenza di token presenti nella domanda
 - Token già inclusi nella risposta agli step precedenti
 - Restituisce il token più probabile tra quelli considerati

Enter text:

One, two,



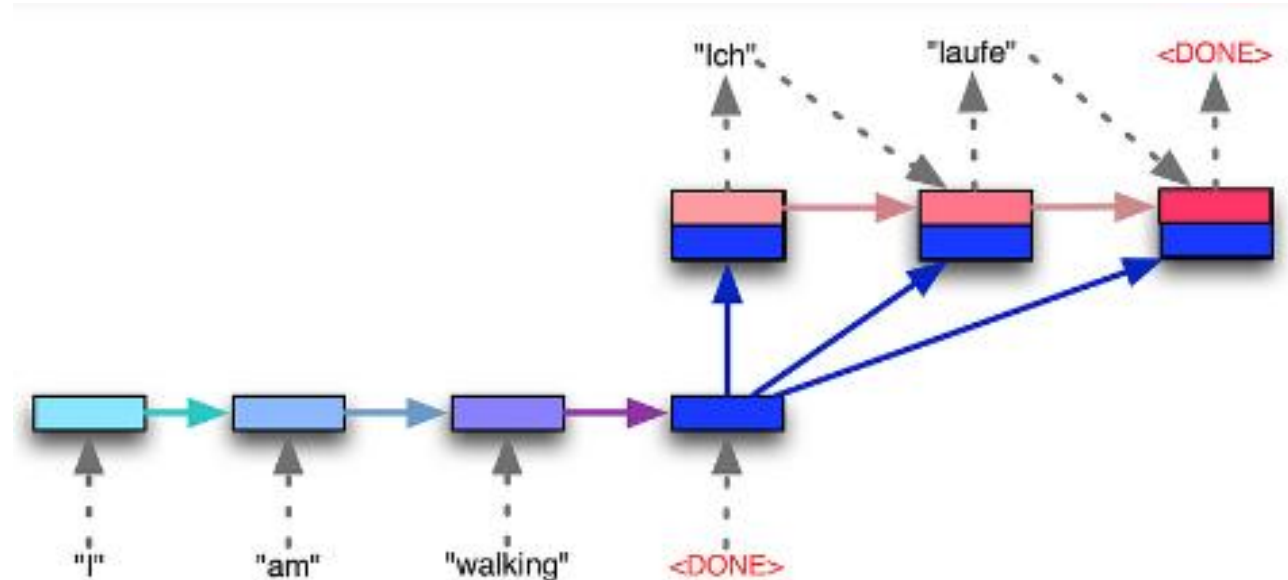
3198 11 734 11

Prediction

#	probs	next token ID	predicted next token
0	39.71%	1115	three
1	16.97%	290	and
2	7.55%	734	two
3	3.76%	1440	four
4	2.76%	393	or
5	2.18%	1936	five
6	1.57%	530	one
7	1.43%	345	you
8	1.15%	257	a
9	0.84%	3598	seven

Large Language Model

- Il testo generato può essere
 - Una risposta a una domanda
 - Una traduzione di un testo
 - Vedi figura
 - Una sintesi di un documento
 - Una parafrasi di un testo
 - ...
 - Una query SQL
 - Laboratorio su LLM #1
 - Un descrizione di un ragionamento per arrivare a una soluzione
 - Laboratorio su LLM #2

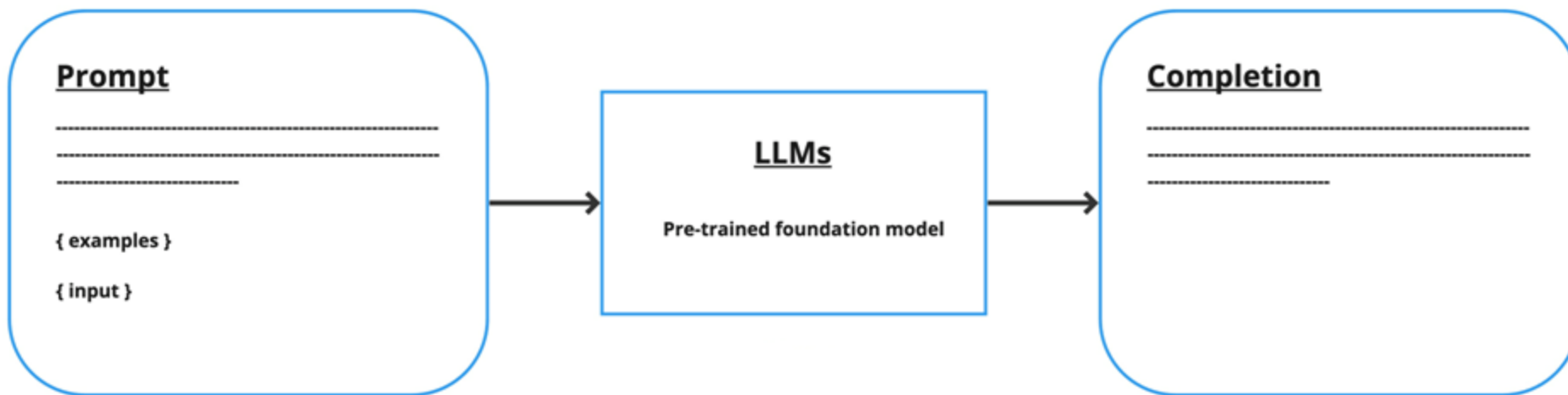


Prompt tuning

- La qualità del testo generato dipende, tra gli altri fattori, da
 - Modello LLM utilizzato
 - Configurazione del modello
 - Stile del prompt
 - Presenza/assenza/tipologia di un template di risposta
 - Eventuali esempi forniti
 - Lingua utilizzata
 - L'inglese è quello su cui la maggior parte degli LLM si è addestrata in modo preponderante
 - Complessità della query richiesta
- E' necessario testare varie combinazioni dei parametri sopra citati per trovare la configurazione ottimale

Large Language Model

- Generazione di testo
 - Input: **prompt**
 - testo in linguaggio naturale di una lunghezza massima di token
 - **Richiesta** (obbligatoria)
 - **Template**: formato atteso della risposta (opzionale)
 - **Esempi**: uno o più esempi di input e output (opzionali)
 - Output: **completion**
 - testo in linguaggio naturale generato dal LLM



Large Language Model

- **In-context learning**

- Fornire esempi di input e output nel prompt per adattare la risposta del modello

- **Zero-Shot Learning (ZSL)**: Nessun esempio

- la risposta dipende esclusivamente dal modello pre-addestrato e dal formato della richiesta (più dall'eventuale template)

- **Few-Shot Learning (FSL)**: K esempi di domande e risposte forniti dall'utente

- La risposta dipende anche dagli esempi forniti

- Il modello LLM può essere istruito con esempi di uno o più task nel medesimo prompt

- Ad es. Traduzioni da inglese a italiano e da inglese a spagnolo

In-context learning: limitazioni

- Con ZSL o FSL il modello pre-addestrato rimane immutato
- Problematiche
 - Il modello potrebbe essere non aggiornato
 - I dati su cui è stato addestrato potrebbero non essere specifici per un determinato problema o contesto di analisi
 - Ad es. potrebbe aver visto pochi esempi di testi nella lingua target
 - Il prompt che uso per istruire il modello ha una lunghezza limitata
 - Posso fornire un numero limitato di esempi
 - Il testo generato potrebbe contenere allucinazioni
 - Testo offensivo
 - Stile non appropriato
 - Contenuti troppo generici
 - Contenuti inventati o ripetuti

Fine-tuning

- Specializzazione del modello attraverso una nuova fase di addestramento
 - Parziale o totale
- Richiede
 - GPU
 - Molti dati annotati (esempi domande-risposte)
 - Memoria e tempo macchina
- Costi generalmente molto elevati rispetto a in-context learning

Introduzione all'analisi dei dati

Il task Text2SQL

Demo Text2SQL su ChatGPT (<https://chatgpt.com/>)

INPUT

Dato la base di dati relazionale avente come il seguente schema logico: "Si consideri una base di dati relazionale composta da una tabella fornitori denominata F avente come chiave primaria il campo «codice fornitore» denominato CodF e come campi NomeFornitore e Sede (campo opzionale)". Data la seguente query in linguaggio naturale da eseguire sulla basi di dati precedentemente descritta: "Visualizzare i codici dei fornitori in ordine alfabetico di nome". Formulare query in linguaggio SQL.

OUTPUT



Per ottenere i codici dei fornitori ordinati alfabeticamente per nome dalla tabella fornitori, possiamo utilizzare la seguente query SQL:

sql

Copia codice

```
SELECT CodF
FROM F
ORDER BY NomeFornitore ASC;
```

Spiegazione della query:

- **SELECT CodF:** seleziona il campo `CodF`, che rappresenta il codice del fornitore.
- **FROM F:** indica la tabella `F` da cui prelevare i dati.
- **ORDER BY NomeFornitore ASC:** ordina i risultati in ordine alfabetico crescente in base al campo `NomeFornitore`.



Text2SQL

- Esempio di task da risolvere mediante LLM
- Dato lo schema logico relazionale di una base di dati e un'interrogazione formulata in linguaggio naturale generare la corrispondente interrogazione in linguaggio SQL
- Esempio di prompt
 - *«Dato la base di dati relazionale avente schema logico [...] e data la query in linguaggio naturale [...] da eseguire sulla basi di dati precedentemente descritta, formulare query in linguaggio SQL»*
- Esempi di descrizioni dello schema logico relazionale da inserire nel prompt
 - Es. 1: «Si consideri una base di dati relazionale composta da una tabella fornitori denominata F avente come chiave primaria il campo «codice fornitore» denominato CodF e come campi NomeFornitore e Sede (campo opzionale)»
 - Es. 2: «Si consideri il seguente schema logico relazionale: tabella F(CodF, NomeF, Sede), CodF chiave primaria, Sede campo opzionale»
- Esempio di query in linguaggio naturale da inserire nel prompt
 - Es. «Visualizzare i codici dei fornitori in ordine alfabetico di nome»
- Esempio di output atteso
 - Query SQL: «SELECT CodF from F ORDER BY NomeF;»

Text2SQL

- Esempi di FSL
 - Esempi di domande-risposte (da aggiungere in coda al prompt testuale precedente):
 - «Si considerino i seguenti tre esempi di coppie query testuale - query SQL:
 1. Testo: [Visualizzare i codici dei fornitori] SQL: [SELECT CodF FROM F;]
 2. Testo: [Visualizzare i dati dei fornitori ordinati per nome] SQL: [SELECT * FROM F ORDER BY NomeF;]
 3. Testo: [Visualizzare i codici e nomi dei fornitori] SQL: [SELECT CodF, NomeF FROM F;]

Esempio di query più complessa

- Prompt

Data il seguente schema logico relazionale di una base di dati:

- Fornitore(CodiceFornitore, NomeFornitore, Sede) Chiave Primaria: CodiceFornitore

- Prodotto(CodiceProdotto, NomeProdotto, Prezzo) Chiave Primaria: CodiceProdotto

Forniture(CodiceFornitore, CodiceProdotto, QuantitàFornita)
Chiave Primaria: CodiceFornitore, CodiceProdotto

Rispondere alla seguente query in linguaggio SQL:

Trovare i fornitori che hanno fornito i prodotti con prezzo medio massimo.

Esempio di query più complessa

- La soluzione prodotta dall'LLM (in questo caso ChatGPT) è giusta?
 - Laboratorio su LLM #1

```
Query SQL:

sql Copia

WITH PrezzoMedio AS (
  SELECT p.CodiceProdotto, AVG(p.Prezzo) AS PrezzoMedio
  FROM Prodotto p
  GROUP BY p.CodiceProdotto
),
PrezzoMax AS (
  SELECT MAX(PrezzoMedio) AS PrezzoMedioMax
  FROM PrezzoMedio
)
SELECT f.CodiceFornitore, f.NomeFornitore
FROM Fornitore f
JOIN Forniture fo ON f.CodiceFornitore = fo.CodiceFornitore
JOIN Prodotto p ON fo.CodiceProdotto = p.CodiceProdotto
JOIN PrezzoMedio pm ON p.CodiceProdotto = pm.CodiceProdotto
JOIN PrezzoMax pmx ON pm.PrezzoMedio = pmx.PrezzoMedioMax;
```

Esempio di query più complessa

- La soluzione prodotta dall'LLM (in questo caso ChatGPT) è giusta?

Query SQL:

```
sql Copia  
  
WITH PrezzoMedio AS (  
    SELECT p.CodiceProdotto, AVG(p.Prezzo) AS PrezzoMedio  
    FROM Prodotto p  
    GROUP BY p.CodiceProdotto  
),  
PrezzoMax AS (  
    SELECT MAX(PrezzoMedio) AS PrezzoMedioMax  
    FROM PrezzoMedio  
)  
SELECT f.CodiceFornitore, f.NomeFornitore  
FROM Fornitore f  
JOIN Forniture fo ON f.CodiceFornitore = fo.CodiceFornitore  
JOIN Prodotto p ON fo.CodiceProdotto = p.CodiceProdotto  
JOIN PrezzoMedio pm ON p.CodiceProdotto = pm.CodiceProdotto  
JOIN PrezzoMax pmx ON pm.PrezzoMedio = pmx.PrezzoMedioMax;
```



Esempio di query più complessa

- Come ha «ragionato» il modello?
 - Laboratorio su LLM #2

Per risolvere la query, dobbiamo innanzitutto trovare il prezzo medio massimo tra i prodotti e successivamente identificare i fornitori che hanno fornito questi prodotti.

Passaggi:

1. Calcolare il prezzo medio per ciascun prodotto.
2. Trovare il prezzo medio massimo.
3. Identificare i fornitori che hanno fornito prodotti con il prezzo medio massimo.

Esempio di query più complessa

- Come ha «ragionato» il modello?
 - Laboratorio su LLM #2



Per risolvere la query, dobbiamo innanzitutto trovare il prezzo medio massimo tra i prodotti e successivamente identificare i fornitori che hanno fornito questi prodotti.

Passaggi:

1. Calcolare il prezzo medio per ciascun prodotto.
2. Trovare il prezzo medio massimo.
3. Identificare i fornitori che hanno fornito prodotti con il prezzo medio massimo.

Copyright

Attribution + Noncommercial + NoDerivative