



Capire e utilizzare al meglio i modelli GPT



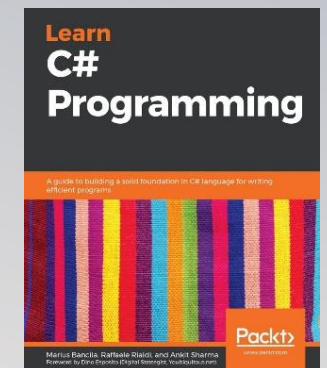
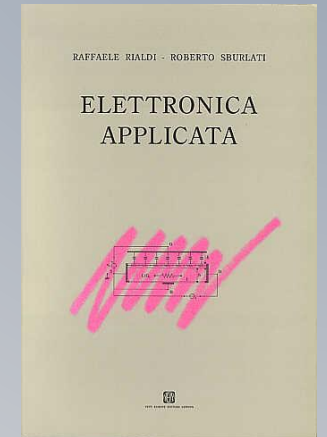
Ing. Raffaele Rialdi

Senior Software Architect - Consultant

@raffaeler - raffaeler@vevy.com

Chi sono

- Laurea Master in Ingegneria Elettronica (Unige)
- Lavoro professionalmente nel software dal 1987
- Insegno saltuariamente a Ingegneria Informatica (Unige)
- Membro della commissione ICT dell'Ordine degli Ingegneri
- Microsoft Most Valuable Professional per 21 anni consecutivi
- Libero professionista, Software Architect in diversi ambiti:
 - Financial, Manufacturing, Healthcare, F1 racing, ...
- Speaker in conferenze nel mondo (più di 200 interventi in 20 anni)
 - Europa, Asia, USA
- Co-Autore del libro "Elettronica Applicata"
- Co-Autore del libro "C# Programming"
- Presidente di DotNetLiguria a Genova



Large Language Models

- Sono algoritmi che fanno uso di reti neurali molto complesse
 - La dimensione è data dalla totalità dei parametri della rete neurale

Modello	Azienda	Open / Closed source	Miliardi di parametri
GPT4	OpenAI	Closed	175
ClaudeAI	Anthropic	Closed	137
Gemini	Google	Closed	1400
Phi-3.5	Microsoft	Open	3.8 (mini), 4.2 (vision), 6.6 (MoE), 7, 14
Llama 3.2	Meta	Open	1, 3, 11, 90

- Durante la fase di training, la rete impara il linguaggio naturale
 - Le **lingue** più diffuse, gli **stili** di linguaggio, i **toni** e i **fatti** su cui è stato addestrato
 - I modelli recenti sono addestrati anche a **ragionare e risolvere passo per passo**
 - Conoscono anche i **linguaggi di programmazione**.

Dal testo ai numeri e vice-versa

- I modelli processano **numeri** ma il linguaggio naturale usa parole
- Il testo in **input** (o **Prompt**) deve essere convertito in numeri (token)

explain quantum computing	[95444, 31228, 25213]
indivisible	[485, 344, 23936]
inducted	[485, 670, 291]

- La rete neurale ha come **output** il prossimo token
- Una funzione matematica "attivazione" chiamata SoftMax calcola i prossimi **possibili token**
- I valori numerici dei token costituiscono la distribuzione di probabilità che la rete produca un determinato **token**.

I modelli LLM stimano i prossimi token

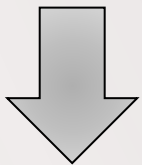
- La capacità effettiva di un LLM è di completare il prompt
 - Prompt: "It's a long way to the top if" ...
 - Il completamento più probabile è: "you wanna rock'n' roll"
- Cosa risponde il modello con "We are the"?
 - Molte frasi su Internet iniziano con le stesse parole!
- The prompt **stabilisce il contesto** del completamento **atteso**
 - **System Prompt:** "Sei un esperto del gruppo musicale Queen. Tu completi solo i titoli forniti dall'utente senza ripetere il titolo per intero"
 - La risposta sarà sempre "Champions"
- Lo scopo dei prompt è di **restringere il contesto**.

La produzione del prossimo token

It's a long way to the top if

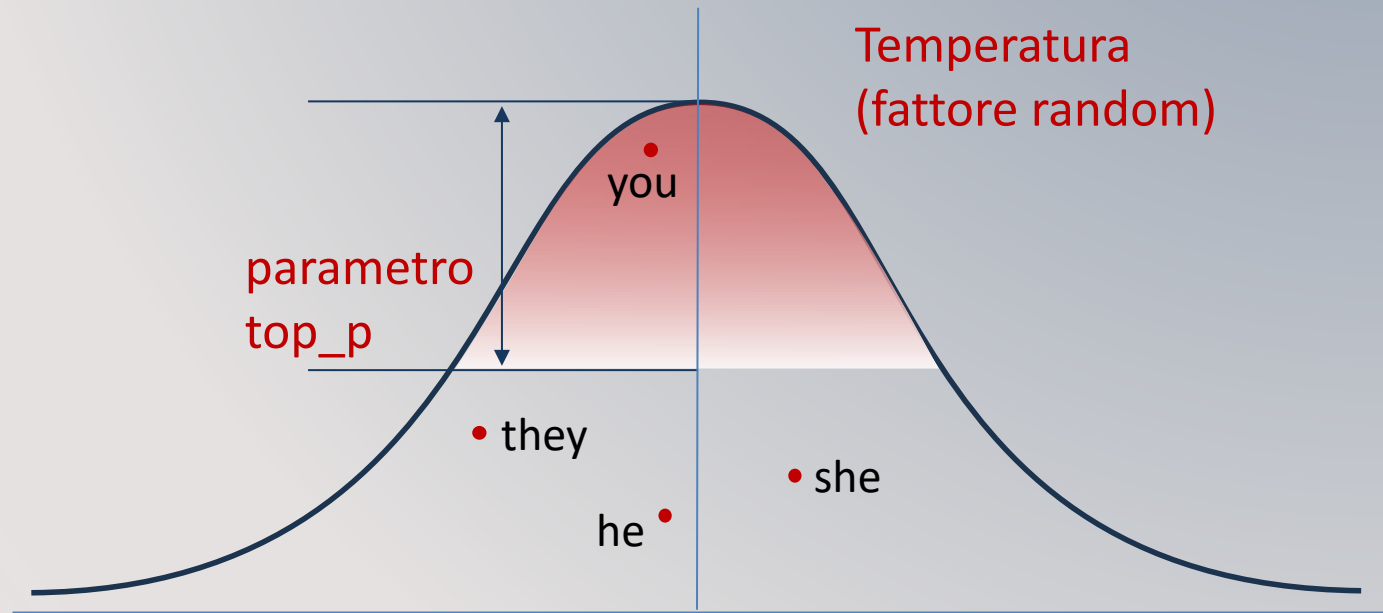
[15834, 261, 1701, 2006, 316, 290, 2344, 538]

you	95%
they	3%
she	1%
he	1%



you

[13320]

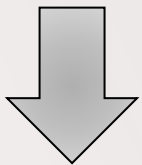


La produzione del prossimo token

It's a long way to the top if you

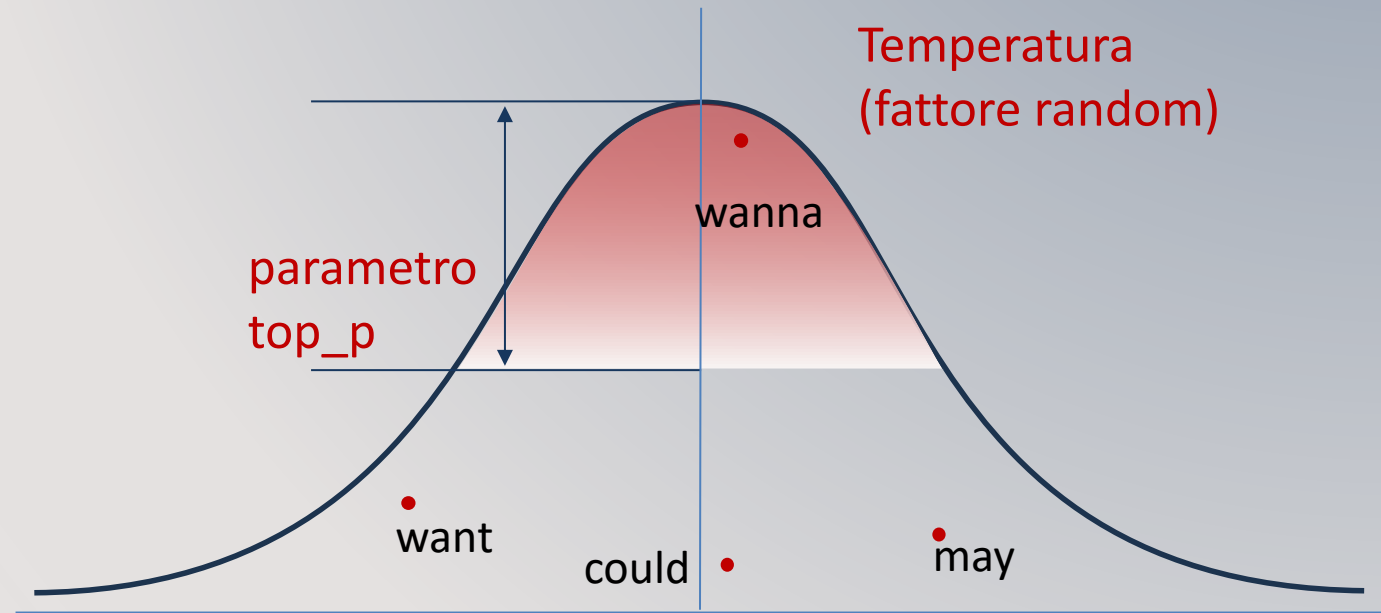
[15834, 261, 1701, 2006, 316, 290, 2344, 538, 13320]

wanna	98%
want	1%
may	0.5%
could	0.5%



wanna

[53399]

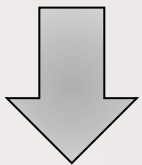


La produzione del prossimo token

It's a long way to the top if you wanna

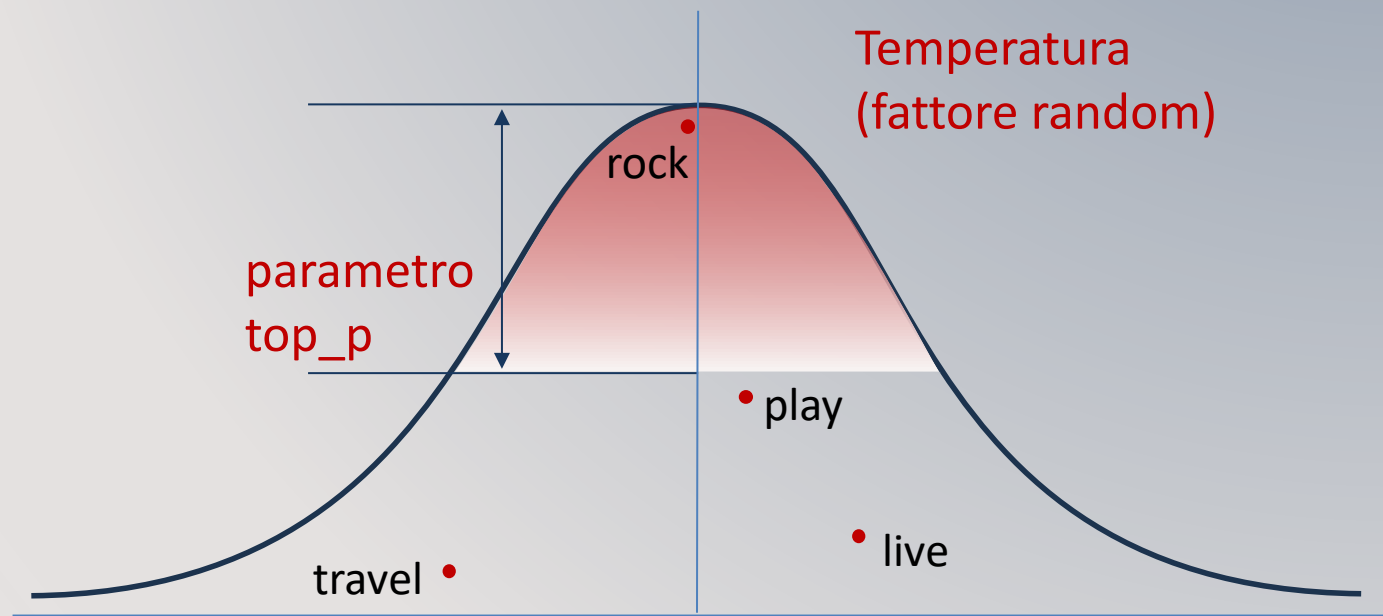
[15834, 261, 1701, 2006, 316, 290, 2344, 538, 13320, 53399]

rock	98%
play	1%
live	0.5%
travel	0.5%



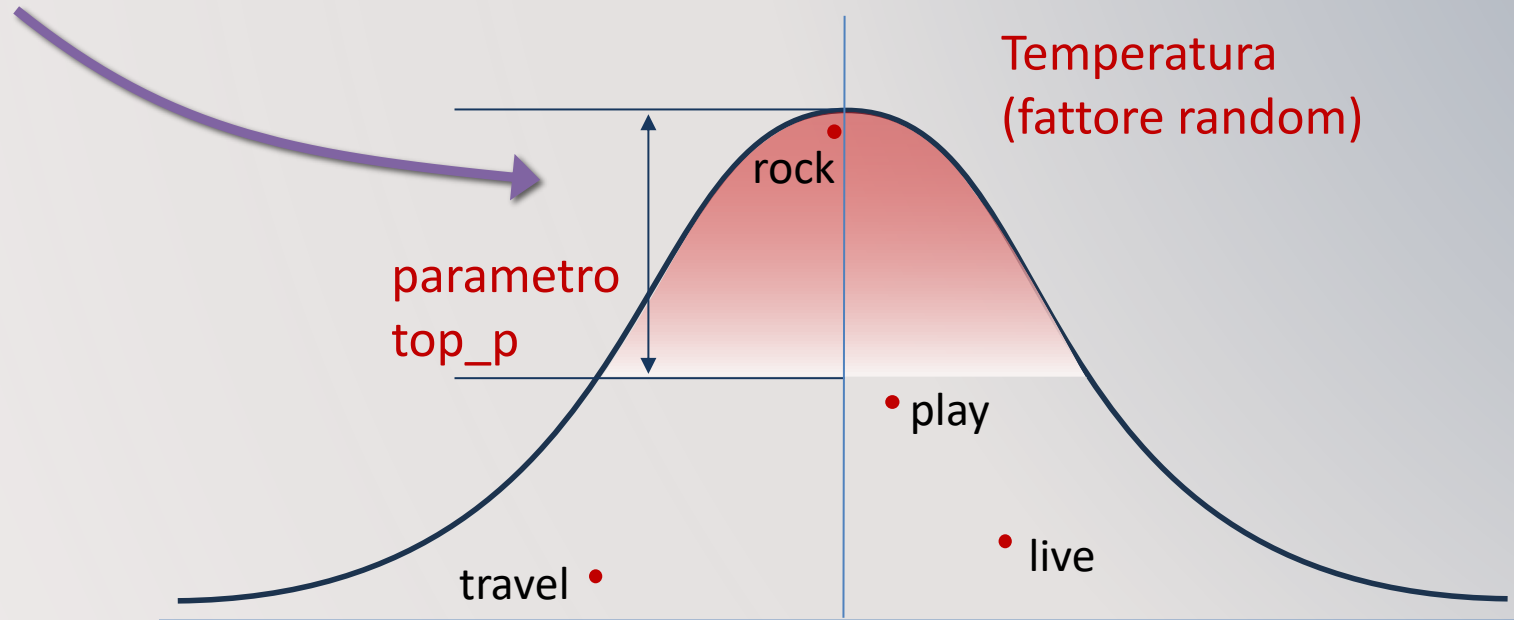
rock'n' roll

[707, 15245,6, 8465]



Perché ChatGPT può dare informazioni errate?

- Se l'area non è riempita di verità, potranno occorrere le **Allucinazioni**



- Come fare se i dati "veri" sono pochi/nulli?
 - Retrieval Augmented Generation (RAG)
 - Uso di Agenti/Tools/Funzioni
 - Reinforced Learning (fine-tuning del modello) che è meglio evitare!

I Prompt

- Le domande al modello si chiamano "User Prompt"
- Le risposte del modello si chiamano "Assistant Prompt"
 - User e Assistant Prompt sono quelli visibili in ChatGPT
- Il System Prompt è un **vincolo** che una applicazione fornisce al modello
 - Evita che il modello possa rispondere a domande fuori dal contesto
 - Si usa per indicare la "Persona" o "Ruolo" dell'assistente virtuale
- Un buon prompt deve:
 - Indicare il compito dell'assistente e restringere il contesto delle risposte attese
 - Specificare le modalità della risposta (testo, JSON, tabelle, grafici, etc.)
 - Precisare la lingua e/o lo stile linguistico in cui rispondere.

Le allucinazioni

Esempio (Modello GPT 3.5 Turbo)

Storia
completamente
sbagliata

Fai un breve riassunto della novella di Pirandello "La Carriola"

La novella "La Carriola" di Luigi Pirandello narra la storia di Gagì, un giovane contadino siciliano che viene ingiustamente accusato di aver commesso un furto. Nonostante le sue proteste di innocenza, viene arrestato e condannato a scontare una pena in prigione.

Durante il tragitto verso la prigione, Gagì viene legato a una carriola, simbolo della sua umiliazione e schiavitù. Durante il viaggio, incontra vari personaggi che lo deridono e lo umiliano ulteriormente, tra cui un amico d'infanzia che lo tradisce.

Esempio (Modello GPT 4 Turbo)

Falso

Fai un breve riassunto della novella di Pirandello "La Carriola"

"La Carriola" è una novella di Luigi Pirandello che racconta la storia di un uomo rispettabile e di successo che si sente intrappolato dalla sua vita monotona e dalle aspettative della società. Per sfuggire a questo senso di oppressione, l'uomo si rifugia in un'attività insolita: ogni giorno, dopo il lavoro, si mette a spingere una carriola nel suo giardino. Questo strano comportamento diventa oggetto di pettegolezzi e critica, ma per l'uomo rappresenta un modo per sfidare le convenzioni sociali e sentirsi finalmente libero. La novella esplora temi come l'identità, la follia e il conflitto tra l'individuo e la società.

Il cane è un elemento fondamentale della storia

È presente un cane nella storia?

No, nella novella "La Carriola" di Luigi Pirandello non è presente un cane. La storia si concentra principalmente sul protagonista e sul suo comportamento insolito di spingere una carriola nel suo giardino come forma di liberazione personale. Non ci sono riferimenti a un cane nella trama.

Esempio 1 (Modello GPT 4o mini)

Fai un breve riassunto della novella di Pirandello "La Carriola"

"La carriola" di Luigi Pirandello è una novella che esplora temi come l'identità e la percezione della realtà. La storia ruota attorno a un personaggio, un uomo che vive un conflitto interiore tra la sua

Il protagonista è un professore di diritto e avvocato

Chi è il protagonista?

Il protagonista di "La carriola" è un uomo di nome **Vincenzo, un bracciante** che vive in un contesto rurale. La sua vita è segnata da una profonda introspezione e da una continua riflessione sul proprio

Cosa succede al cane nella storia?

Falso!

Nel corso della novella "La carriola," il cane di Vincenzo viene colpito da un evento tragico: **muore**. Questa perdita ha un forte impatto sul protagonista, accentuando il suo senso di solitudine e di

Esempio 2 (Modello GPT 4o mini)

Il protagonista è un professore di diritto e avvocato

Non cita il gesto più importante della storia

Dimmi il riassunto della novella di Pirandello "La carriola"

"La carriola" è una novella di Luigi Pirandello che esplora il tema della follia e della percezione della realtà. La storia ruota attorno a un uomo che, per sfuggire alla routine e ai problemi quotidiani, si rifugia in un mondo immaginario.

Il protagonista, un ingegnere, si imbatte in un personaggio bizzarro che trascina una carriola piena di oggetti. Questo incontro lo porta a riflettere sulla sua vita e sul significato della follia, mettendo in

In che modo il cane e la carriola sono correlati?

Il cane e la carriola sono correlati nella novella "La carriola" attraverso il loro simbolismo e il modo in cui entrambi riflettono le emozioni del protagonista.

Esempio di allucinazione (Modello ChatGPT)

Fai un breve riassunto della novella di Pirandello "La Carriola"

"La carriola" è una novella di Luigi Pirandello pubblicata nel 1917. La storia è **narrata in prima persona da un avvocato**, una figura rispettabile e di successo, che racconta un episodio chiave della sua vita interiore.

Corretto!

In questo contesto, per sfuggire a questa soffocante rigidità, si concede un atto assurdo e liberatorio: un giorno, senza alcun motivo apparente, **prende il suo cane e lo solleva come se fosse una carriola**, spingendolo in giro per casa. Questo gesto stravagante e apparentemente senza senso rappresenta per l'avvocato una sorta di ribellione contro la monotonia e la mancanza di senso della sua esistenza quotidiana.

Esempio (Modello Gemini)

Fai un breve riassunto della novella di Pirandello "La Carriola"

Corretto!

Il protagonista, un uomo apparentemente normale e inserito nella società, cela un'ossessione inconfessabile: il **gioco della carriola con la sua vecchia cagna**. Questo gioco infantile e quasi vergognoso diventa per lui un momento di evasione dalla routine e un modo per ritrovare una semplicità quasi infantile.

In cosa consiste il gioco della carriola?


Poco specifico pur su esplicita richiesta

Il "gioco della carriola" in "La Carriola" di Pirandello non è un gioco come lo intendiamo comunemente. Non ci sono regole precise, né un obiettivo da raggiungere. È piuttosto un gesto semplice, quasi infantile, che il protagonista compie spingendo la sua vecchia cagna in una posizione simile a quella di una carriola.

Esempio (Modello Gemini)

(nuova chat), dopo la prima risposta corretta sul cane

Che mestiere fa il protagonista?

Show drafts 

Sbagliato!

Il mestiere preciso del protagonista de "La Carriola" non viene specificato da Pirandello. L'autore si concentra maggiormente sull'aspetto psicologico del personaggio e sull'atto assurdo che compie, piuttosto che sulla sua professione.

Esempio (Modello LLama 3.2)

Fai un breve riassunto della novella di Pirandello "La Carriola"

Mi dispiace, ma non riesco a trovare alcuna informazione sulla novella "La Carriola" di Luigi Pirandello. Pirandello è noto per le sue opere teatrali, come "Sei personaggi in cerca d'autore", ma potrebbe essere che "La Carriola" sia una novella pubblicata in un contesto meno noto o una sua opera meno conosciuta. Potresti fornire ulteriori informazioni o contesto sulla novella?

Esempio (Modello Claude Sonnet)

R Fai un breve riassunto della novella di Pirandello "La Carriola"
Show drafts

Ecco un breve riassunto della novella "La Carriola" di Luigi Pirandello:

Un giorno, mentre è solo in casa con **la sua cagnolina**, compie un gesto apparentemente assurdo: **prende l'animale per le zampe posteriori e lo fa "camminare" come una carriola.** Questo atto bizzarro diventa per lui un rituale segreto, un momento di liberazione dalla sua identità sociale.

Corretto!

Microsoft Copilot (copilot.microsoft.com)

Fai un breve riassunto della novella di Pirandello "La Carriola"

"La Carriola" di Luigi Pirandello racconta la storia di un **avvocato di successo** che, nonostante l'apparente perfezione della sua vita, si sente intrappolato dalla monotonia e dalle aspettative sociali. Ogni tanto, per sfuggire a questa prigione invisibile, **gioca a fare la carriola con il suo cane**, un gesto infantile che gli permette di liberare le sue frustrazioni e riaffermare la propria individualità. Una riflessione pungente sulla ricerca della libertà interiore in una società rigida e conformista. Notevole, vero?

Corretto!

Microsoft Copilot (copilot.microsoft.com)

Fai un breve riassunto della novella di Pirandello "La Carriola"

"**La Carriola**" di **Luigi Pirandello** racconta la storia di un **avvocato** che, improvvisamente, mentre sta lavorando nel suo studio, si alza e, senza motivo apparente, **inizia a girare in cerchio come una carriola**. Questo gesto, apparentemente folle, rappresenta il desiderio di evadere dalla monotonia della vita quotidiana e dalla rigidità delle convenzioni sociali. Attraverso questo episodio, Pirandello esplora il tema dell'alienazione e della ricerca di libertà personale.

Cosa ti ha incuriosito di più della novella?

Sbagliato!

Cosa succede al cane nella storia?

Il cane nella novella non ha un destino drammatico. Resta un'ombra silenziosa nella vita dell'avvocato, sempre presente ma mai realmente al centro della scena. La sua presenza costante enfatizza la routine quotidiana e l'inevitabile ripetitività della vita dell'avvocato.

Sbagliato!

Il Prompt che risolve totalmente e sempre il problema

Fai il riassunto di questa novella:

La Carriola

di Luigi Pirandello

Quand'ho qualcuno attorno, non la guardo mai; ma sento che lei, mi guarda, mi guarda senza staccarmi un momento gli occhi. Vorrei farle intendere, a quattr'occhi, che non è nulla; che non potevo permettermi con altri questo breve atto, con alcuna importanza e per me è tutto. Lo compio ogni giorno.

Alcuni cenni sui costi dei modelli

OpenAI e Azure OpenAI

Nome del modello	Costo \$ per milione di token (input)	Costo \$ per milione di token (output)
GPT 4o	2.50	10.00
GPT 4o-mini	0.15	0.60
O1	15.00	60.00
O1-mini	3.00	12.00
ChatGPT-4o-latest	5.00	15.00
GPT-4-Turbo	10.00	30.00
GPT-4	30.00	60.00
GPT-4-32K	60.00	120.00
GPT-4-vision-preview	10.00	30.00
GPT-3.5-turbo-0125	0.50	1.50
GPT-3.5-turbo-instruct	1.50	2.00

Google Gemini

- I seguenti dati si riferiscono ai prezzi *pay-as-you-go*
- C'è inoltre il costo del "contesto" che è decisamente modesto
 - Tra i 2 e 3 centesimi per milioni di token

Nome del modello	Costo \$ per milione di token (input)	Costo \$ per milione di token (output)
Gemini <128K token	0.075	0.300
Gemini > 128K token	0.150	0.600

Ricerca semantica

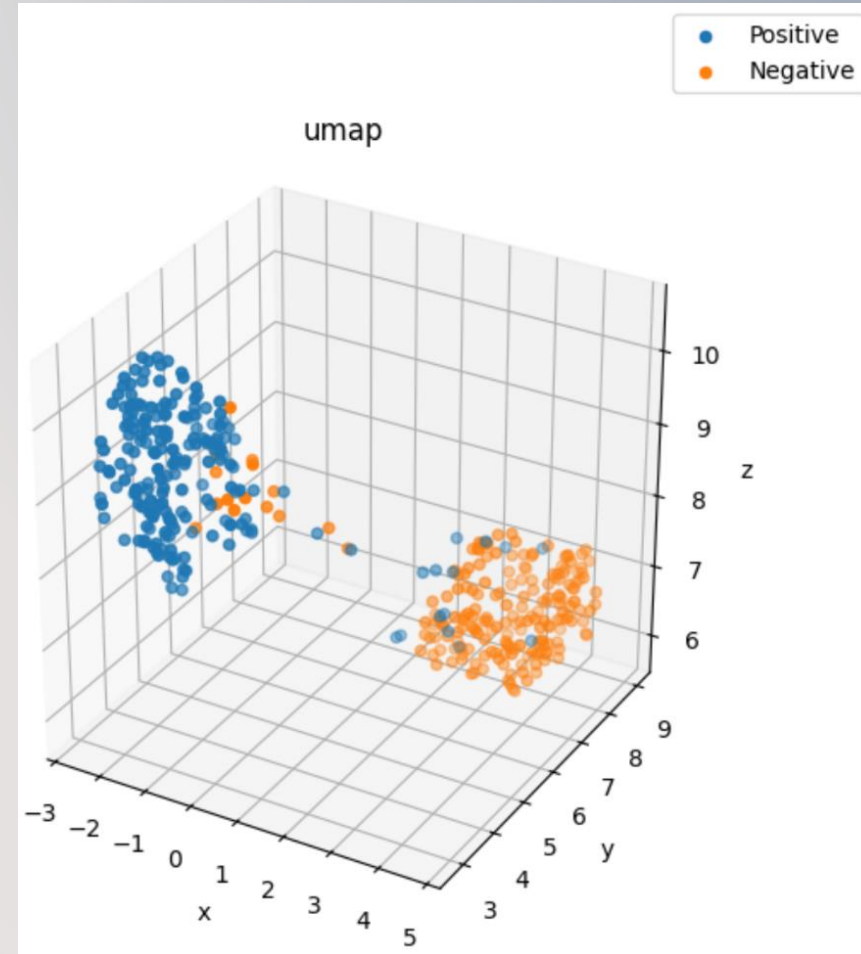
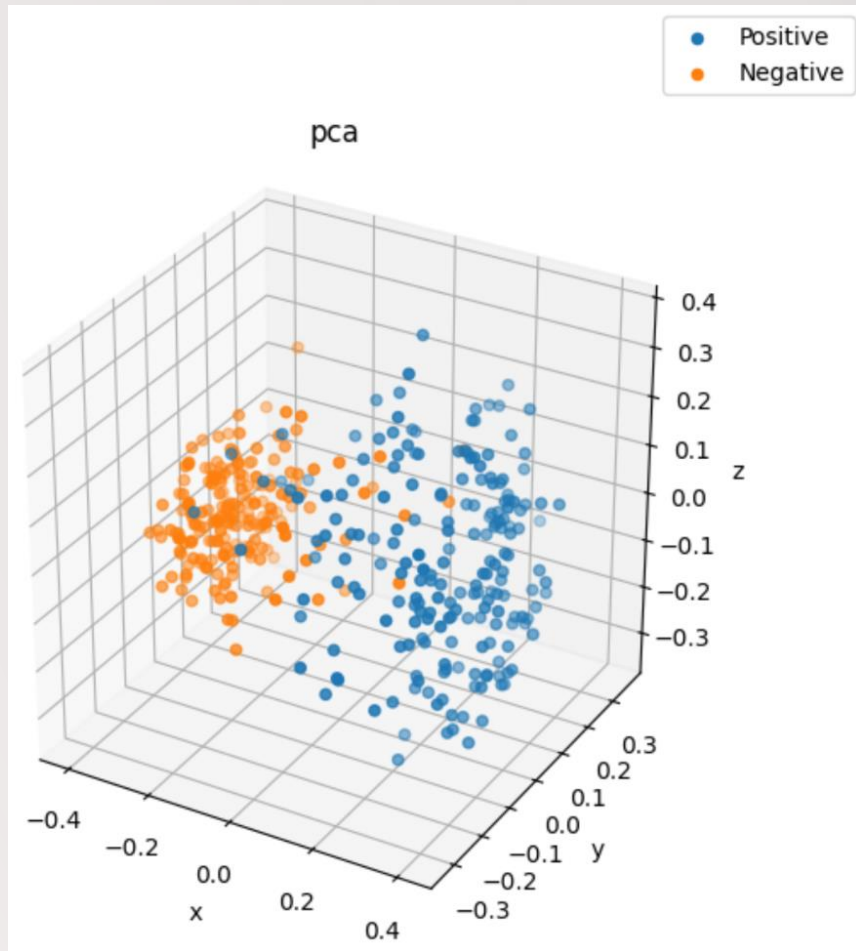
Gli Embedding

- Alcuni modelli LLM sono fatti per creare solo Embedding
- Un Embedding è un vettore di numeri decimali
- È la **coordinata del concetto** in uno spazio multidimensionale
 - Gli Embedding vicini identificano concetti simili
- Come possiamo usare gli Embedding?
 - Calcolo un Embedding per ciascun documento
 - Misuro la **distanza tra gli Embedding** ricavati dai documenti
 - Importante: gli errori di sintassi non alterano il concetto espresso
- In altre parole è possibile usarli per la ricerca semantica
 - Si contrappone alla ricerca classica basata su parole chiave

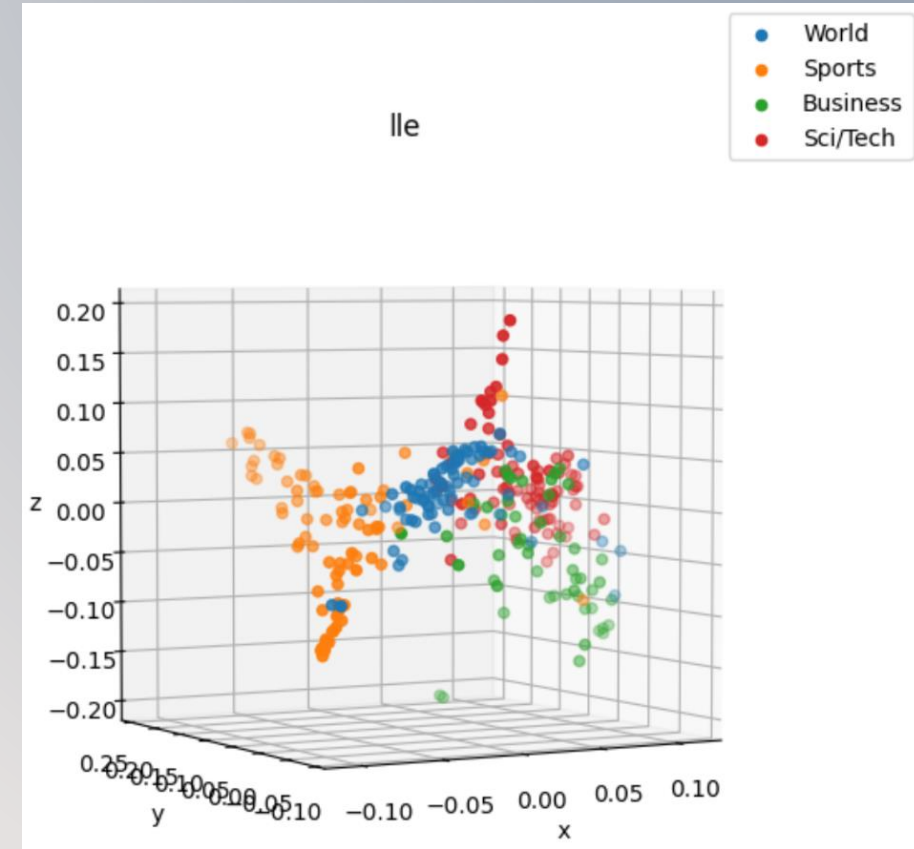
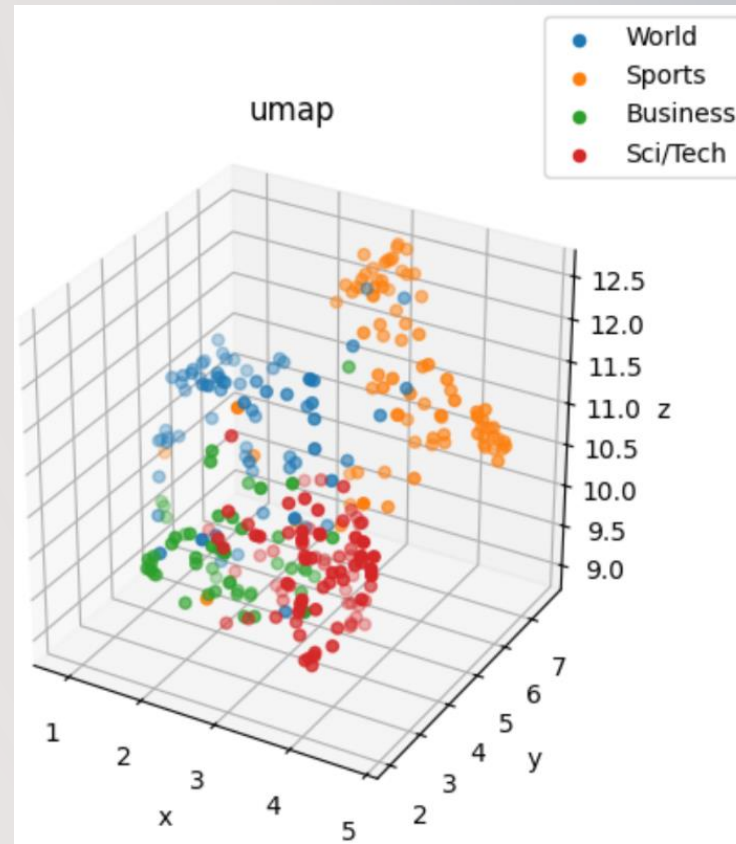
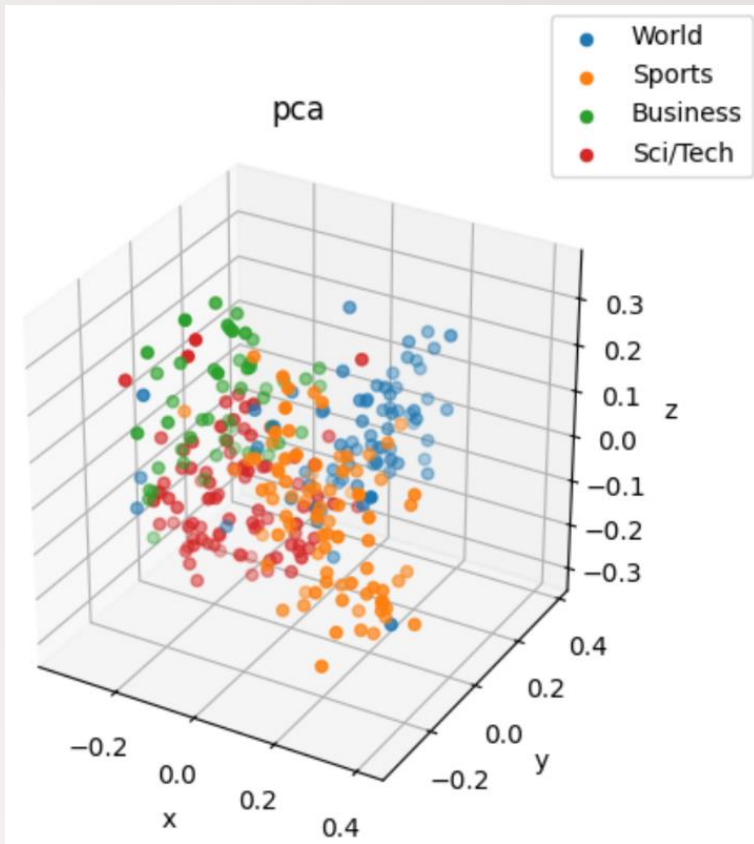
La ricerca semantica

- Caso d'uso:
Cercare i documenti contestuali al Prompt espresso in linguaggio naturale
- **Fase preparatoria:**
 - Trasformo ogni documento in uno o più Embedding
 - Salvo questi vettori numerici dentro un Database Vettoriale
- **Fase operativa:**
 - L'utente scrive una domanda
 - Il software converte il testo della domanda in un Embedding
 - Il software cerca nel Database Vettoriale i vettori più vicini alla domanda.

Visualizzazione degli Embedding (PCA e UMAP)



Visualizzazione degli Embedding (PCA, UMAP e LLE)



I costi degli Embedding

- Sono decisamente bassi
- OpenAI (ad oggi):

Modello	Costo per 1 milione di token
text-embedding-3-small	\$0.020
text-embedding-3-large	\$0.130
ada v2	\$0.100

RAG (Retrieval Augmented Generation)

- Utilizza la Ricerca Semantica per arricchire il Prompt delle informazioni che servono a rispondere.
 1. Il Prompt dell'utente viene convertito in Embedding
 2. Si cercano gli Embedding più vicini nel DB Vettoriale
 3. I risultati sono costituiti da un punteggio (score) e dal nome del file del documento. Vengono tenuti solo i punteggi più alti (più vicini)
 4. I documenti più simili vengono aggiunti al Prompt
 5. Il modello riceve la domanda e i dati necessari per rispondere
- La RAG è molto efficiente e fornisce risposte su **dati non-pubblici**.

Per concludere

- I Prompt sono estremamente importanti
- I modelli LLM non vanno usati come motore di ricerca
- I modelli più vecchi ed economici possono comunque essere sufficienti a certi scopi
- Manca ancora un tassello importante: gli Agenti (ne parliamo più tardi).

Discussione



Domande?

Grazie!

@raffaeler

raffaeler@vevy.com



Generato offline (GPU)
con Stable Diffusion

Prompt:

create a picture of a crowd of
happy smiling people in a
room cheering at the speaker

Domande?

Grazie!

@raffaeler

raffaeler@vevy.com



Generato offline (GPU)
con Stable Diffusion

Prompt:

create a picture of a crowd of
happy smiling people in a
room cheering at the speaker

Domande?

Grazie!

@raffaeler

raffaeler@vevy.com



Generato offline (GPU)
con Stable Diffusion

Prompt:

create a picture of a crowd of
happy smiling people in a
room cheering at the speaker

Domande?

Grazie!

@raffaeler

raffaeler@vevy.com



Generato offline (GPU)
con Stable Diffusion

Prompt:

create a picture of a crowd of
happy smiling people in a
room cheering at the
speaker.