Appunti - LLM.md

LLM

Cosa sono?

LLM significa *Large Language Model*, cioè un tipo di intelligenza artificiale che ha imparato a "prevedere" quale parola viene dopo in una frase.

Funziona così:

1. Spezza il testo in piccoli pezzi chiamati *token* (che possono essere parole, sillabe o lettere).
2. Ogni token viene trasformato in un numero.
3. Il modello lavora con questi numeri per prevedere quale token (cioè parola o parte di parola) dovrebbe venire dopo, in base a tutto quello che ha visto prima.

```
Large Language Models (LLMs), such as GPT-3 and GPT-4, util
ize a process called tokenization. Tokenization involves br
eaking down text into smaller units, known as tokens, which
the model can process and understand. These tokens can rang
e from individual characters to entire words or even larger
chunks, depending on the model. For GPT-3 and GPT-4, a Byte
Pair Encoding (BPE) tokenizer is used. BPE is a subword tok
enization technique that allows the model to dynamically bu
ild a vocabulary during training, efficiently representing
common words and word fragments. Although the core tokeniza
tion process remains similar across different versions of t
hese models, the specific implementation can vary based on
the model's architecture and training objectives.
```

E i bot che rispondono alle domande come fanno quindi a generare interi testi?

- Semplice, non fanno altro che prevedere la parola successiva, una dopo l'altra. Ne scrivono una, poi guardano tutte quelle già scritte (compresa la tua domanda) e cercano di indovinare la prossima. Ripetono questo processo finché non arrivano alla fine del discorso.

- In pratica, costruisce una frase un pezzo alla volta, cercando ogni volta il token più probabile da aggiungere.

Ma se gli LLM prevedono semplicemente il testo, come fanno a comportarsi da chatbot?

Anche se i modelli linguistici (LLM) sono tecnicamente solo **macchine per prevedere la continuazione del testo**, è possibile farli comportare da chatbot fornendo **istruzioni iniziali (prompt)** che li guidano nel generare risposte in stile conversazionale.

Ecco un esempio di prompt in inglese e in italiano che trasforma un LLM in un assistente virtuale:

Inglese:

```
`You are an intelligent, friendly, and helpful AI assistant. You always
respond clearly, honestly, and concisely. If you do not know the answer to
a question, say so. Engage in natural, conversational dialogue and provide
helpful responses.
User: Hello!
Assistant:`
```

Italiano:

```
`Sei un'assistente virtuale intelligente, amichevole e disponibile.
Rispondi sempre in modo chiaro, onesto e conciso. Se non conosci la
risposta a una domanda, dillo apertamente. Mantieni un tono naturale e
conversazionale e fornisci risposte utili.
Utente: Ciao!
Assistente:`
```

Il modello, vedendo questo contesto, **imita lo stile richiesto** continuando la conversazione come se fosse un assistente.

Quali sono gli LLM a disposizione su internet

- **OpenAI** ha creato **ChatGPT**, probabilmente il più famoso. Questo perché è stato il primo modello rilasciato al pubblico che avesse capacità abbastanza elevate da attirare l'attenzione.
- **Google** ha creato **Gemini**, che prima si chiamava Bard. È pensato per integrarsi bene con gli altri strumenti Google (come Gmail e Drive) e può cercare informazioni direttamente dal web mentre risponde.
- **Meta** (cioè Facebook) ha creato **LLaMA**. La cosa interessante? È *open source*, cioè chiunque può scaricarlo, studiarlo e usarlo gratuitamente. A differenza di quelli di Google o OpenAI, non è una scatola chiusa: è più come un "giocattolo smontabile" per ricercatori e sviluppatori.
- Stanno iniziando anche ad esserci anche casi di aziende più piccole e indipendenti, come:
    i. Mistral – https://mistral.ai
    ii. Nous Research – https://huggingface.co/NousResearch
    iii. Cohere – https://cohere.com
    iv. Anthropic – https://www.anthropic.com
    v. Adept – https://www.adept.ai
    vi. EleutherAI – https://www.eleuther.ai
    vii. xAI – https://x.ai
    viii. DeepSeek – https://www.deepseek.com

Cosa sono i prompt?

- Un **prompt** è semplicemente il testo che scriviamo per far partire una risposta da un LLM, come una domanda, un'istruzione o una richiesta.
- È come dare un "input" al modello: più è chiaro e preciso, più la risposta sarà utile.
    - Ad esempio, scrivere *"Scrivimi una ricetta veloce con quello che ho in frigo"* è un prompt.
    - Anche quando parli con un assistente virtuale o chiedi "spiegami come funziona qualcosa", stai dando un prompt.
- Alcuni utenti diventano bravissimi a scrivere prompt dettagliati, quasi come se parlassero una nuova lingua: si chiama proprio *prompt engineering*, cioè l'arte di chiedere bene per ottenere risposte migliori.

In cosa sono bravi gli LLM

- Un LLM è bravissimo a **lavorare con le parole**:
    - *capire* e *scrivere* testi [[#Esempio 1]]
    - *spiegare* concetti [[#Esempio 2]]
    - *riassumere* [[#Esempio 3]]
    - *tradurre* (è persino meglio di google traduttore oramai, perché riesce a comprendere il contesto della frase) [[#Esempio 4]]
    - *generare idee* [[#Esempio 5]]
    - *aiutarti a capire* qualcosa in modo *semplice*. [[#Esempio 6]]
- È quindi utile per studiare, scrivere email, creare contenuti o farsi venire spunti creativi.

In cosa NON sono bravi gli LLM

- Non è infallibile:
    - può inventare dati
    - sbagliare su argomenti tecnici
    - dare risposte sicure ma in realtà sbagliate
- Non ha "esperienza del mondo", capisce solo dai testi che ha visto. Quindi è perfetto come assistente o “seconda opinione”, ma non va usato per prendere decisioni importanti, legali, mediche o finanziarie senza verificarle.
- Ha un numero massimo di token che riesce a gestire, superato quelli inizierà a scordarsi le cose

Un aspetto importante da tenere in considerazione

- Anche se un LLM non ha esperienza diretta del mondo (cioè non *vive* o *sperimenta* come un essere umano), ha letto una quantità enorme di testi scritti da persone. Questo include libri di matematica e fisica, documentazione di strumenti online, manuali tecnici e tantissimo codice da forum, guide e tutorial. Quindi non "capisce" davvero come funziona il mondo, ma ha visto così tanti esempi ben scritti che riesce a **riconoscere i modelli** e a **riprodurre risposte molto accurate**, soprattutto su argomenti ben documentati come scienza, tecnologia e programmazione. In pratica, non sa *per esperienza*, ma sa *per lettura massiccia*.

- Un LLM non “sa” davvero le cose: non ha una memoria con fatti salvati come un’enciclopedia. Fa previsioni, una parola dopo l’altra, cercando di costruire una risposta *plausibile*, cioè che **sembri vera** in base a quello che ha imparato leggendo tanti testi. Questo funziona bene in molti casi, ma può portare a **"allucinazioni"**, cioè risposte inventate ma scritte con tono sicuro. Succede soprattutto quando la domanda è troppo vaga, riguarda informazioni molto specifiche o contiene qualcosa di sbagliato. Il modello, pur di completare la frase in modo sensato, a volte preferisce “riempire i buchi” con la sua fantasia piuttosto che dire “non lo so”.

Esempi di Prompt

Provate pure a copiare e incollare questi prompt in un chatbot e vedere che succede!

Esempio 1

```
Leggi il seguente testo e dimmi di cosa parla in modo chiaro e preciso. Poi
scrivine una versione più semplice adatta a uno studente delle medie.

Negli ultimi decenni, l’intelligenza artificiale ha compiuto enormi
progressi grazie alla crescita esponenziale della potenza di calcolo e alla
disponibilità di grandi quantità di dati. Le reti neurali profonde, in
particolare, hanno rivoluzionato il modo in cui le macchine apprendono,
rendendo possibile il riconoscimento vocale, la traduzione automatica, la
guida autonoma e molto altro. Tuttavia, l’uso crescente di queste
tecnologie solleva anche importanti questioni etiche e sociali, come la
trasparenza degli algoritmi, la privacy dei dati e l’impatto sul lavoro
umano.
```

Esempio 2

```
Spiegami in modo semplice cos’è il machine learning, come se avessi 12
anni. Usa esempi concreti.
```

Esempio 3

```
Riassumi il seguente articolo in massimo 5 punti chiave.

La città e la diocesi di Pisa in festa per il patrono san Ranieri, un laico
vissuto nel Medioevo, che ha accompagnato per molti secoli la vita
spirituale dei suoi concittadini.
```

Nato da una famiglia di mercanti tra il 1115 ed il 1120, Ranieri visse un’adolescenza indifferente alla storia umana e divina di Gesù. Fin quando incontrò sulla sua strada Alberto Leccapecore, un nobile corso che, vivendo in povertà, andava in giro per il mondo ad annunciare il Vangelo: lui lo seguì fino al monastero di San Vito, chiese di parlargli, si confessò. Fu quello l’inizio di una conversione.
Arrivato in Terra Santa a bordo di una nave mercantile, si spogliò delle ricche vesti indossando una ruvida pilurica: un gesto che Ranieri compì nel giorno del Venerdì Santo del 1138 nella basilica del Santo Sepolcro.
Da quel momento si dedicò alla preghiera e alla meditazione sulla vita di Gesù, visitando i principali luoghi santi. Tornerà a Pisa nel 1154, trovando ospitalità dai monaci vittorini di Sant’Andrea in Kinzica. Poi si trasferì a San Vito e vi rimase da laico, ospite della foresteria, dedicandosi alla predicazione e alla guida spirituale di numerosi fedeli che lo seguivano. Morì il 17 giugno 1161.
Nel ricordo della sua traslazione dalla chiesa di San Vito al Duomo (dove le sue spoglie sono conservate in un urna in un apposito transetto) i lungarni pisani, alla vigilia della festa patronale, dunque stasera, si illuminano grazie a oltre 100mila lampanini – ovvero lumini ad olio – distesi sulla biancheria – cornicioni bianchi appesi alle facciate di chiese e palazzi, dando vita alla Luminara, che si chiuderà con uno spettacolo pirotecnico capace di tenere con il naso all’insù decine di migliaia di pisani.
Domani, martedì 17 giugno, il giorno della festa: celebrazioni eucaristiche alle ore 8, 9.30 e 17 in cattedrale. Alle ore 11 la solenne messa in pontificale animata dalla cappella musicale del Duomo e presieduta da mons. Saverio Cannistrà, arcivescovo di Pisa, tornato nelle scorse ore – insieme agli altri vescovi della Toscana – dal pellegrinaggio in Terra Santa. Letture e omelia saranno interpretate, per la prima volta, nel linguaggio dei segni.
Sulla torre di Pisa saliranno i campanari barghigiani e della Lucchesia per tirare a mano le storiche campane.
Nel pomeriggio, sulle acque dell’Arno, il palio remiero dedicato al santo: le quattro imbarcazioni in gara rappresentano gli antichi quartieri storici della città, dedicati a san Martino e sant’Antonio (a sud dell’Arno), san Francesco e santa Maria (a nord).

Esempio 4

Traduci in inglese questo testo tenendo conto del tono formale e del contesto scolastico.

"Gentili professori, vi invio in allegato il documento aggiornato."

Esempio 5

```
Devo preparare una presentazione per la scuola sull’intelligenza
artificiale. Suggeriscimi 5 idee originali per renderla interessante e
coinvolgente.
```

Esempio 6

```
Non riesco a capire la differenza tra "correlazione" e "causalità". Me la
puoi spiegare con parole semplici e con un esempio?
```