Python Biella Group

LLM e Gen AI

Incontri e serate

01 Introduzione teorica agli LLM

02 Hands On - Come costruire un agente intelligente

Materiale e codice sorgente

https://github.com/PythonBiellaGroup/

Obiettivo della serata

LLM: a kind of magic or...?

- Comprendere **cosa è un LLM**
- Comprendere **come si addestra un LLM**
- Usare un framework per la creazione di agenti

About me

Incontri e serate

Ingegnere informatico

Front End developer @ Frankhood dal 2011

Insegnante di informatica @ ITT Panetti Pitagora

Machine Learning specialist @ Datamasters dal 2020

Code Bubbles
https://giumast.substack.com/

Giuseppe Mastrandrea

giu.mast

LLM

- Algoritmi utilizzati per comprendere e generare linguaggio naturale
- Addestrati su grandi quantità di testo per predire la probabilità di una determinata parola o sequenza di parole
- Comunemente basati su Reti Neurali Artificiali
  - Reti Neurali Ricorrenti
  - Architetture Transformers

LLM

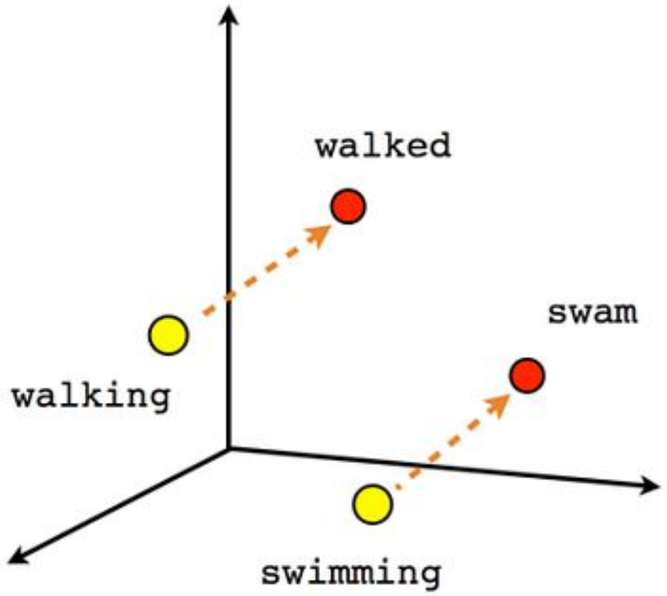
- Generazione di testo
- Completamento automatico
- Traduzione automatica
- Classificazione del testo

Spazi latenti

- Spazio di dimensioni ridotte in cui le informazioni possono essere rappresentate in modo compatto
- Es. un dato di input (come un'immagine) viene compresso in un vettore latente di dimensione ridotta

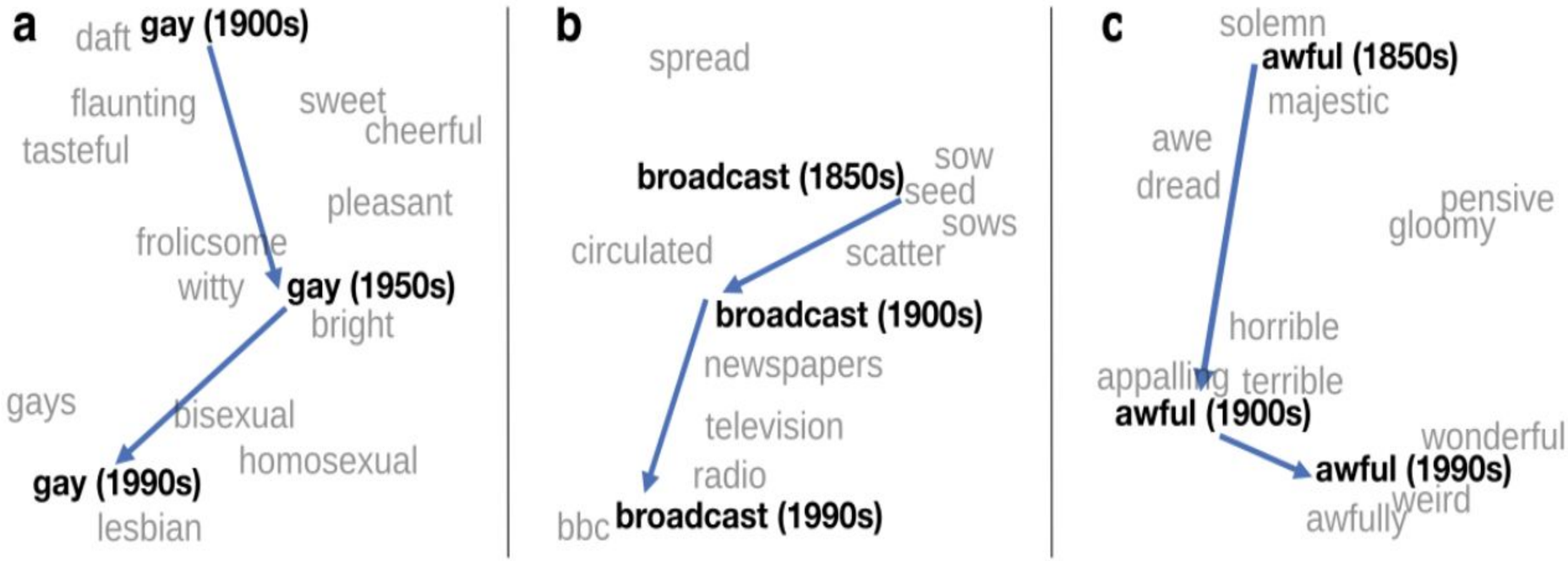
Embedding

- Un tipo particolare di spazio latente sono gli embedding
- Creazione degli embeddings:
  - Rappresentazione del significato delle parole attraverso vettori
  - Addestramento del modello su grandi quantità di dati
  - Catturare le relazioni semantiche tra le parole
- Utilizzo degli embeddings:
  - Calcolo della distanza tra parole
  - Apprendimento della struttura del linguaggio



Esempio pratico: utilizzo degli embeddings in una piattaforma di chatbot per rispondere alle domande degli utenti, fornendo risposte accurate e personalizzate

Embedding

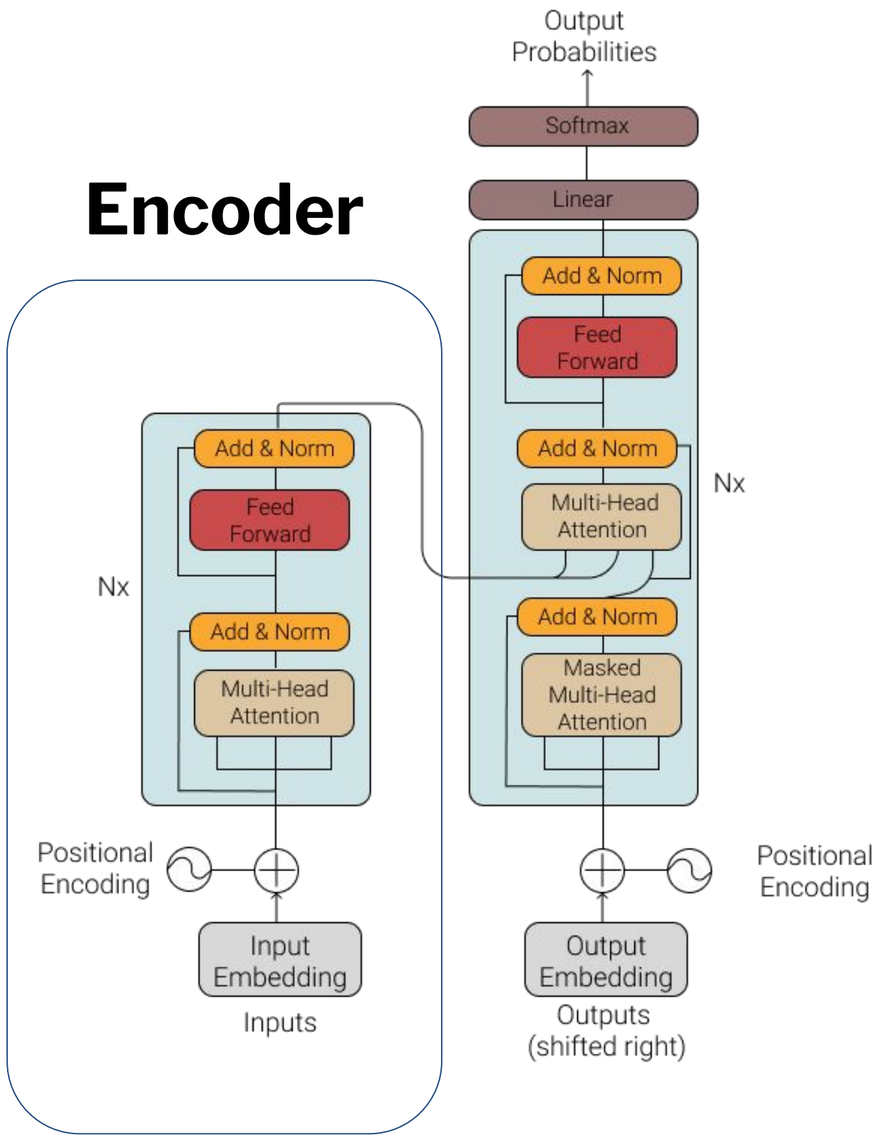

a
daft
gay (1900s)
flaunting
sweet
cheerful
tasteful
pleasant
frolicsome
witty
gay (1950s)
bright
gays
bisexual
homosexual
gay (1990s)
lesbian
b
spread
broadcast (1850s)
sow
seed
sows
circulated
scatter
broadcast (1900s)
newspapers
television
radio
bbc
broadcast (1990s)
c
solemn
awful (1850s)
majestic
awe
dread
pensive
gloomy
horrible
appalling
terrible
awful (1900s)
wonderful
awful (1990s)
weird
awfully

Transformer

- Paper di riferimento: **Attention is all you need**
- Architettura composta da:
  - Encoder: codifica del testo nello **spazio latente** del modello
  - Lo spazio latente è il modo in cui il modello rappresenta le informazioni in input
  - Decoder: *trasforma* i vettori di embedding definiti nello spazio latente del modello nell'output desiderato
- Originariamente proposta per traduzione

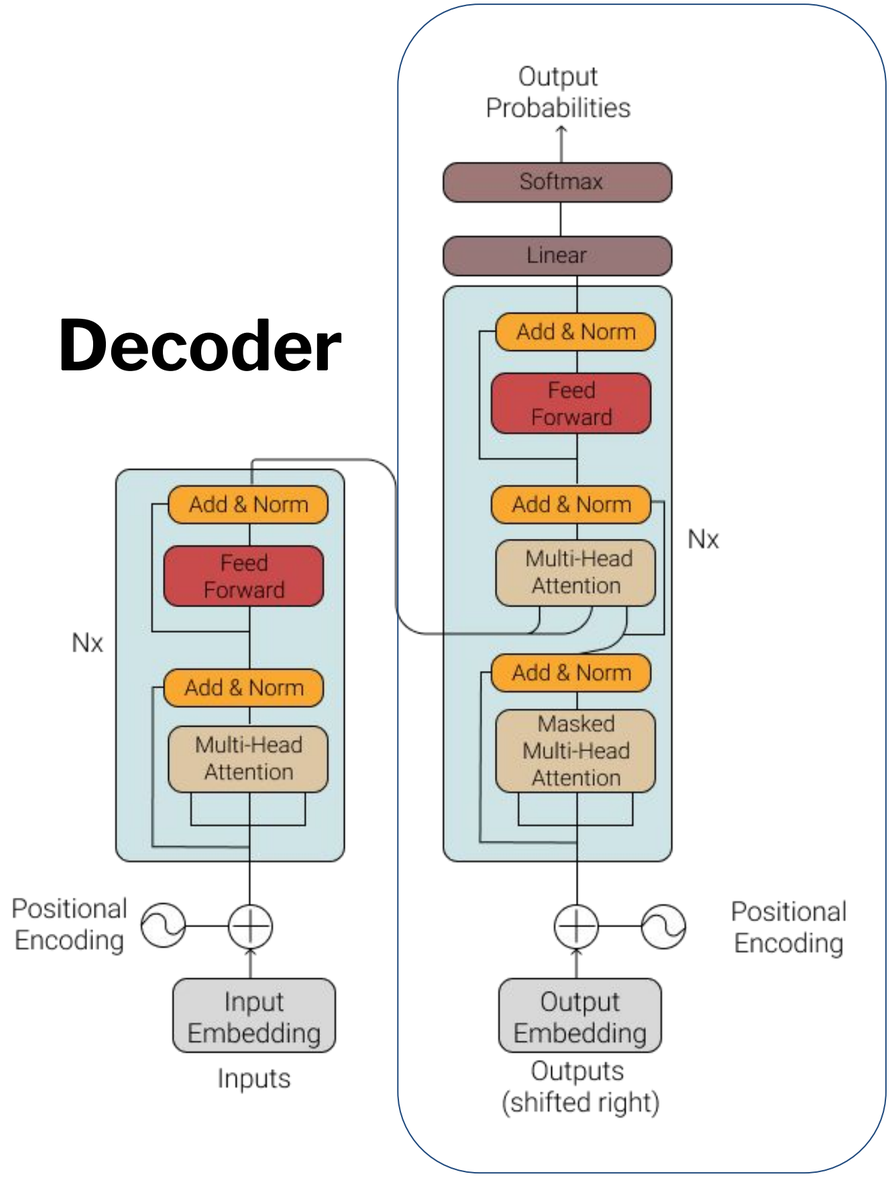
Transformer

- Un encoder lavora nello spazio latente del modello usando:
  - Embedding su cui si è **addestrato**
  - Codifica **posizionale**
  - **Multi-head attention**
  - Connessioni **residuali**
  - **Normalizzazione**



Transformer

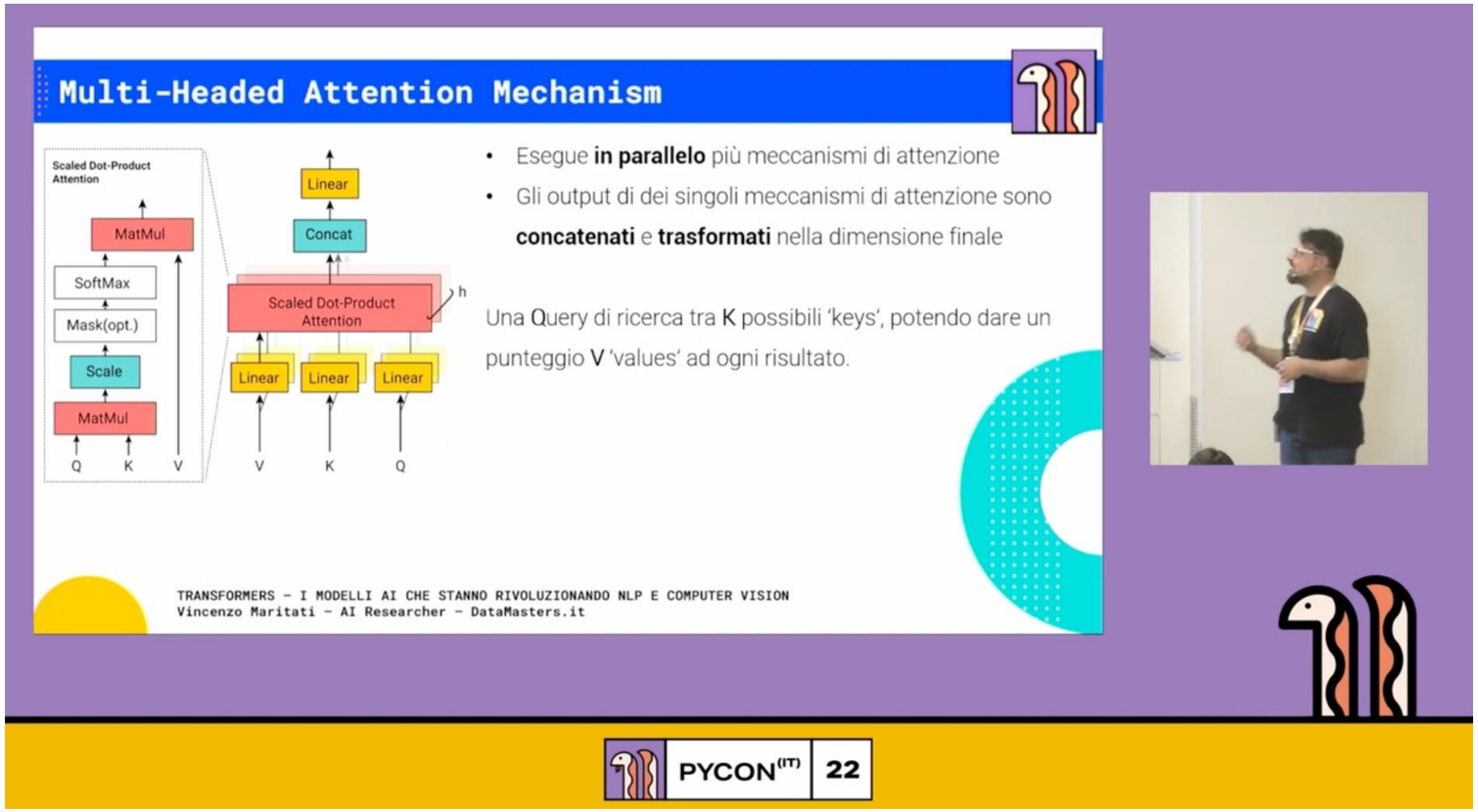
- Un decoder parte da un token standard detto **start-of-sequence**
- Anche qui usiamo **positional encoding**
- Sfrutta gli embedding dell'encoder per generare la prima parola



Transformer

Dove approfondire?

[Transformer e Meccanismi di Attenzione: modelli che rivoluzionano l'AI - Vincenzo Maritati]



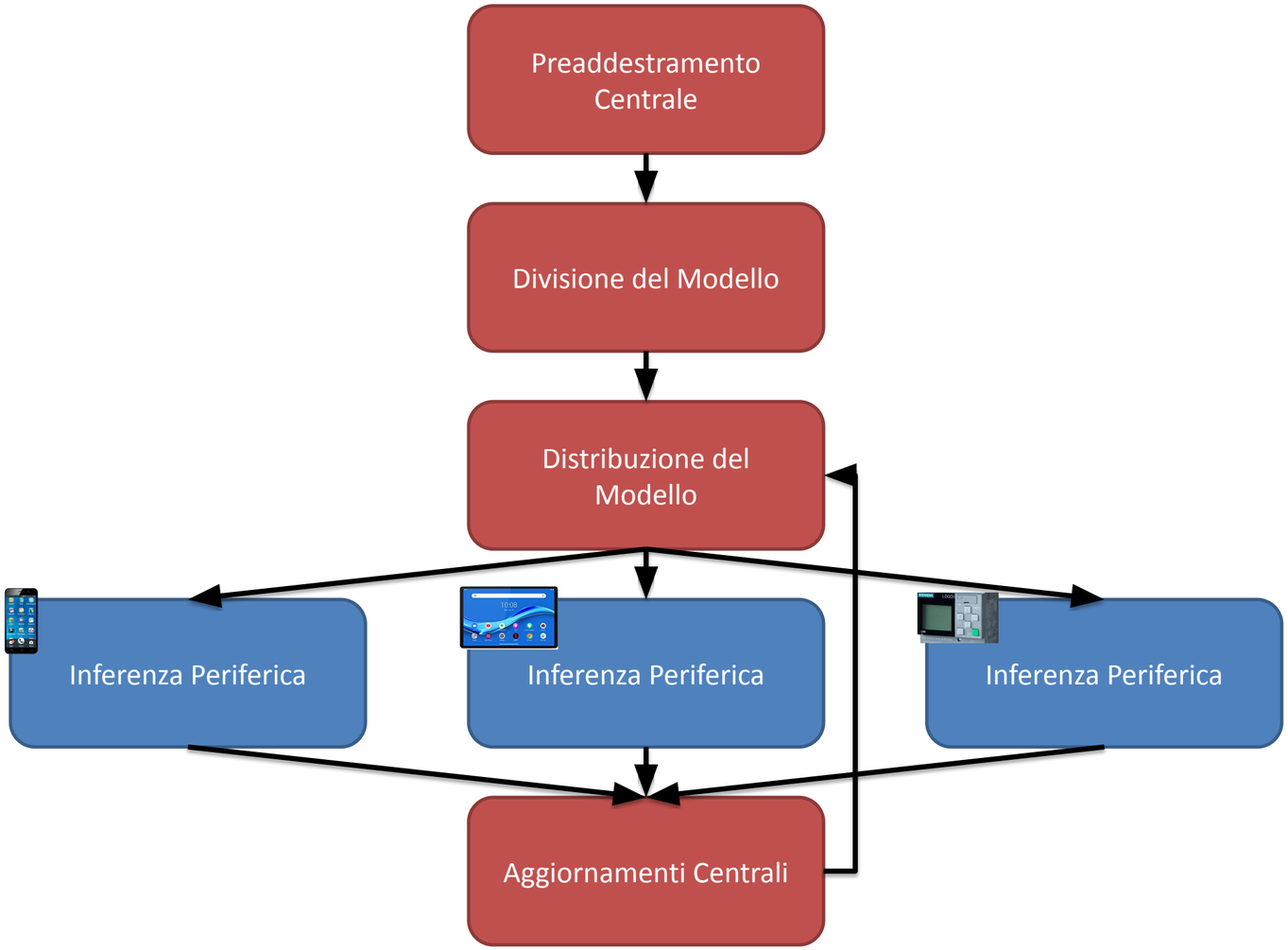
Apprendimento di un LLM

- Addestramento pregresso
- Rappresentazione delle parole (embeddings)
- Architettura del modello
- Fine-tuning
- Contesto e relazioni semantiche
- Generazione di risposte
- Fine-tuning specifico
- Contestualizzazione
- Apprendimento Continuo

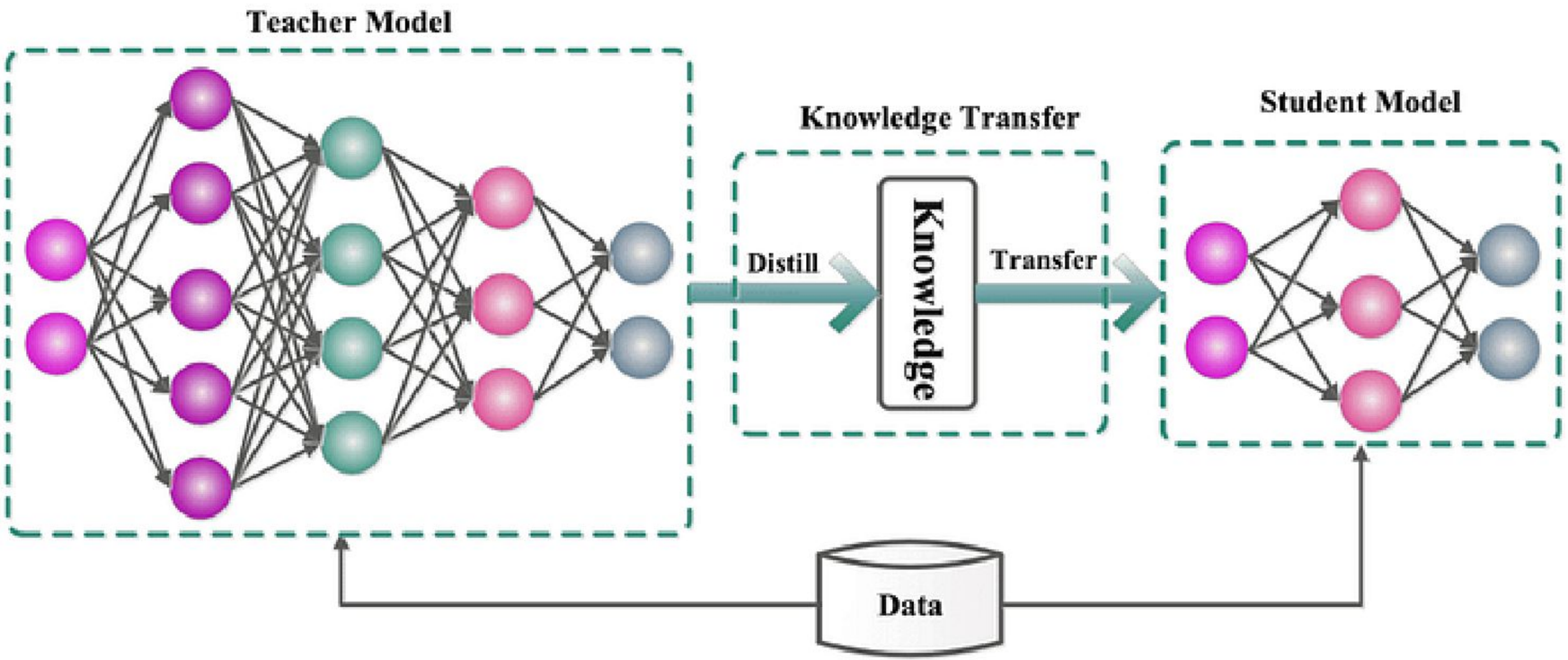
GPT

- Trasformazione Attiva
- Pre-training della parola chiave
- GPT-3 di OpenAI
  - Split-Learning
  - Distillation
- GPT-4 di OpenAI
  - Multimodale

Split Learning



Distillation



Hallucinations

- L'allucinazione nei LLM è un fenomeno in cui il modello produce output inventato senza una base di conoscenza reale
- L'allucinazione può manifestarsi con informazioni dettagliate su argomenti o eventi inesistenti
- L'output può sembrare plausibile, ma è il risultato di associazioni linguistiche casuali all'interno del modello
- Approcci per mitigare l'allucinazione includono:
  - il controllo della temperatura
  - l'uso di filtri post-generazione
  - processi di valutazione e revisione umana
- È necessaria un'attenzione continua per migliorare la qualità e l'affidabilità dei risultati generati dagli LLM

Sfide

- Hallucination
- Utilizzo Etico
- Auto-Addestramento
- Controllo dei Fatti
- Limite: gli LLM non sono a conoscenza di fatti avvenuti **dopo** il loro addestramento

LangChain

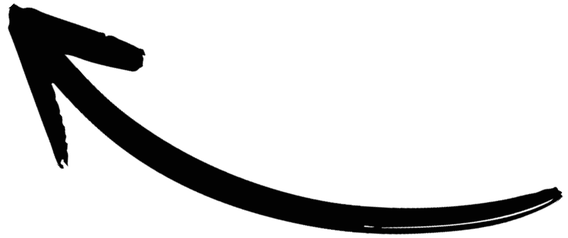
- Un framework per **sviluppare applicazioni** powered by **LLM**
- 74K stars su GitHub
- Offre fra le altre cose un **abstraction layer**
- Offre **molto altro**:
  - Context
  - Parser
  - Prompt Engineering
  - Agents

LangChain

https://bit.ly/langchain-prompt-intro

Join Us on Discord

launchpass.com/datamasters



You’re lucky lucky, you’re so lucky!

50% off su **Generative AI**
codice **PYBGEN**

Sitografia e link utili

- https://openai.com/enterprise-privacy
- Inside the secret list of websites that make AI like ChatGPT sound smart
- Il Manifesto blocca ChatGPT e le altre IA
- Making AI less thirsty