RIFLESSIONI SULL'OTTIMIZZAZIONE DI LLM

Davide Maltoni

Ingegneria e Scienze Informatiche – Cesena

ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

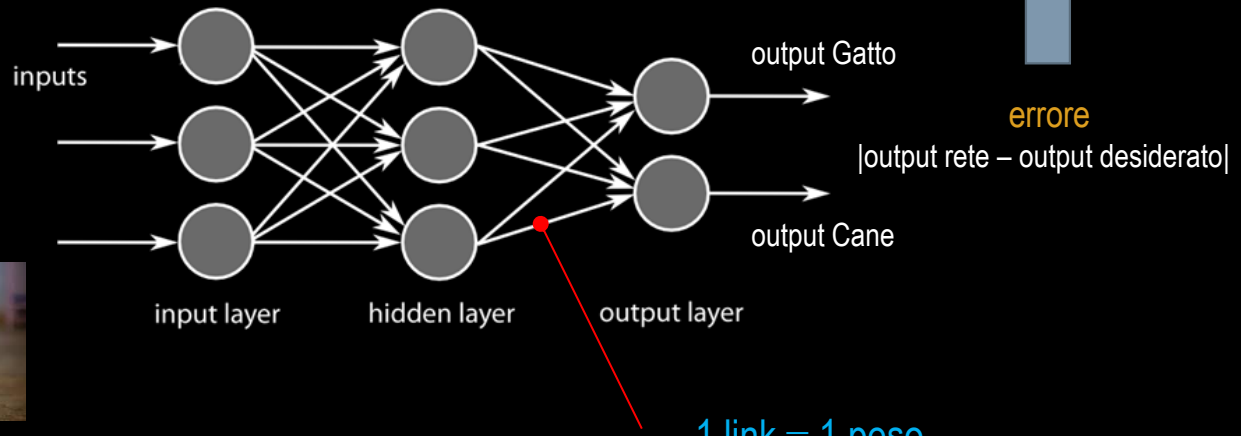
TRAINING (SUPERVISIONATO)

Backpropagation dell'errore
e aggiornamento pesi

[Rumelhart, Hinton & Williams, 1986]

Gatto

Cane



errore
|output rete – output desiderato|

1 link ≡ 1 peso

Miliardi di pesi in LLM moderni

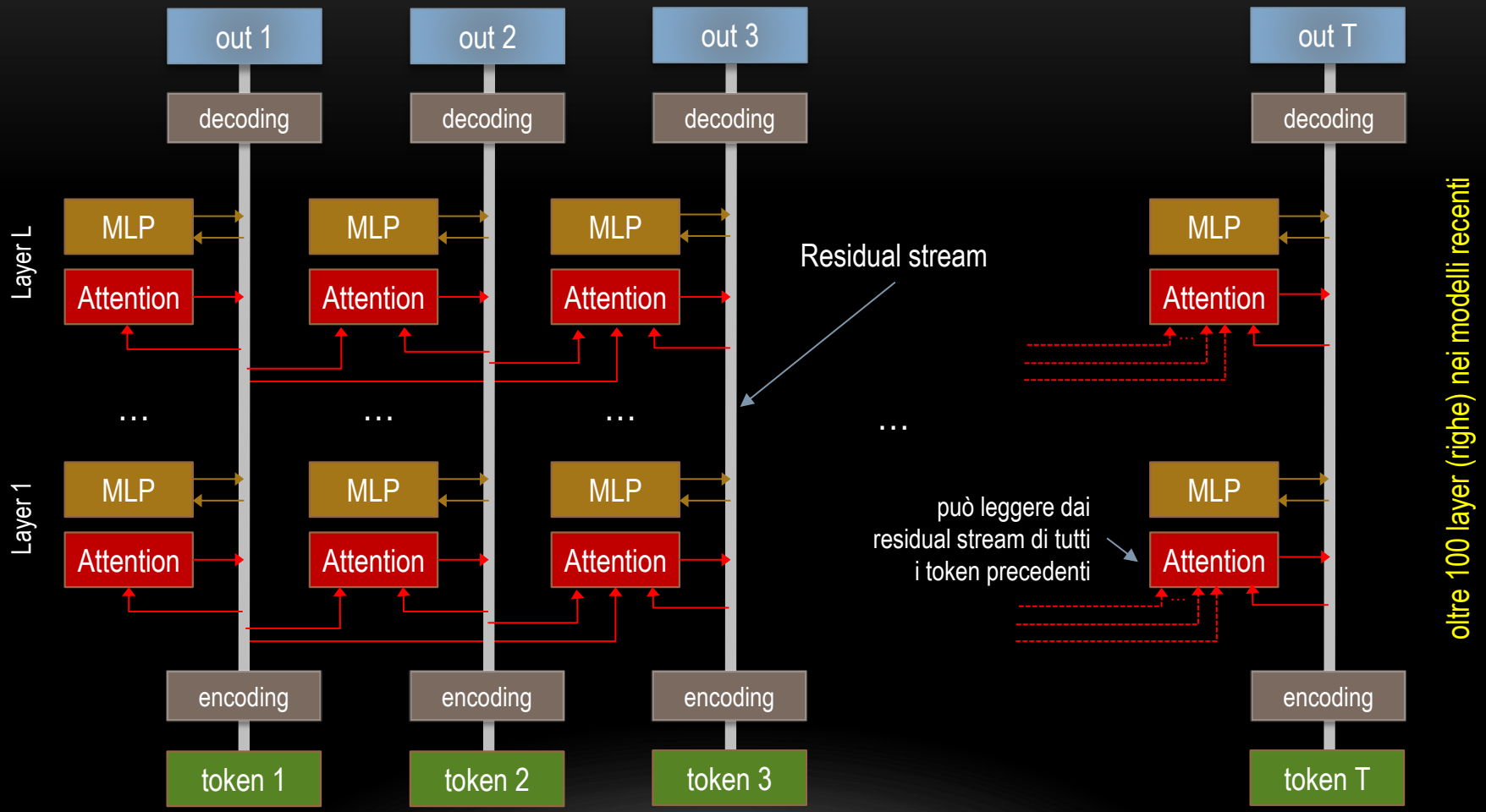
TRANSFORMERS

- Architettura DNN introdotta da Google (Brain) nel 2017
- Sono il modello di riferimento per LLM.
- L'elemento chiave sono i layer di self-attention che mettono in relazione tra loro le parole nella frase per meglio specificarne la semantica.

La
palla
finì
in
rete
al
90
minuto

TRANSFORMERS

Decoder only (es. GPT serie)



oltre 100.000 token (colonne) nei modelli recenti

ADDESTRAMENTO LLM: 3 FASI

L'addestramento di un LLM prevede (solitamente) più fasi di addestramento:

1. Self-supervised pre-training: impara semantica termini, struttura linguaggio, ma non solo.
   - Addestrato facendogli predire il prossimo token nel testo
2. Supervised instruction-tuning: impara a rispondere a domande e a eseguire istruzioni
   - Addestrato su una serie di domande, per le quali è nota la risposta
3. Alignment-tuning: impara a dare risposte allineate con preferenze umane
   - Fornisce risposte multiple che sono valutate da umani che indicano la preferita. (Reinforcement Learning from Human Feedback)

ADDESTRAMENTO (SELF-SUPERVISED)

1	Nel	Nel	Nel	Nel	Nel
2	mezzo	mezzo	mezzo	mezzo	mezzo
3	del	del	del	del	del
4	cammin	cammin	cammin	cammin	cammin
5	di	di	di	di	di
6	nostra	nostra	nostra	nostra	nostra
7	vita	vita	vita	vita	vita
8	mi	mi	mi	mi	mi
9	ritrovai	ritrovai	ritrovai	ritrovai	ritrovai
10	per	per	per	per	per
11	una	una	una	una	una
12	selva	selva	selva	selva	selva
13	oscura	oscura	oscura	oscura	oscura
14	ché	ché	ché	ché	ché
15	la	la	la	la	la

In ARANCIONE le parole fornite in input

In ROSSO la parola da predire

GENERAZIONE (AUTO-REGRESSIVA)

1	nostra	nostra	nostra	nostra
2	vita	vita	vita	vita
3	mi	mi	mi	mi
4	ritrovai	ritrovai	ritrovai	ritrovai
5	per	per	per	per
6	una	una	una	una
7	selva	selva	selva	selva
8	oscura	oscura	oscura	oscura
9	ché	ché	ché	ché
10	la	la	la	la
11	diritta	diritta	diritta	diritta
12		via	via	via
13			era	era
14				smarrita

- Dato un LM pre-addestrato si parte con un prompt costituito da 10 parole e si chiede alla rete di predire (generare) l'11-sima parola
- Si appende l'ultima parola generata all'output corrente e si passano al modello le ultime 10 (scartando la prima).
- Si itera in questo modo (auto-regressione) fino ad ottenere output di lunghezza desiderata

FOUNDATION MODEL PREADDESTRATI: LIMITI

- fermi alla conoscenza disponibile a tempo di addestramento
- meno efficaci su lingue meno diffuse
- non conoscono documenti privati
- non conoscono termini e concetti specifici di domini verticali

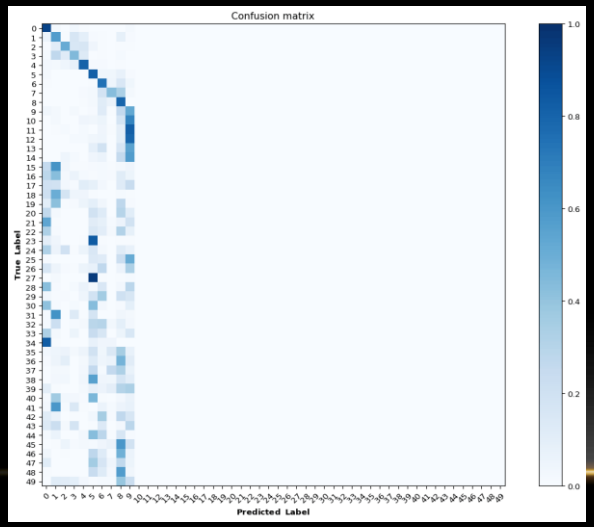
COME SUPERARE I LIMITI: OPZIONI

- Addestrare LLM da zero partendo da pesi random. Complesso e costoso. In quali casi è vantaggioso rispetto a tuning?
- Tuning LLM (tipicamente fase 2 instruction-tuning) attraverso modifica pesi a partire da quelli di un Foundation Model. Possibile e spesso efficace ma valutare effetti collaterali.
- RAG – Retrieval Augmented Generation: pesi fissi, augmentation del prompt con frammenti pertinenti. Strada principale per interrogare base conoscenza.
- ICL – In Context Learning: guidare LLM inserendo esempi nel prompt. Efficacia few-shot dimostrata
- Prompt optimization (es. Meta-prompting). Ottimizzare «automaticamente» formulazione prompt (zero-shot)
- Chain-Of-Thoughs. Forzare LLM a «ragionare» per step intermedi.
- Agentic approaches. Più agenti (LLM) collaborano per risolvere un problema.

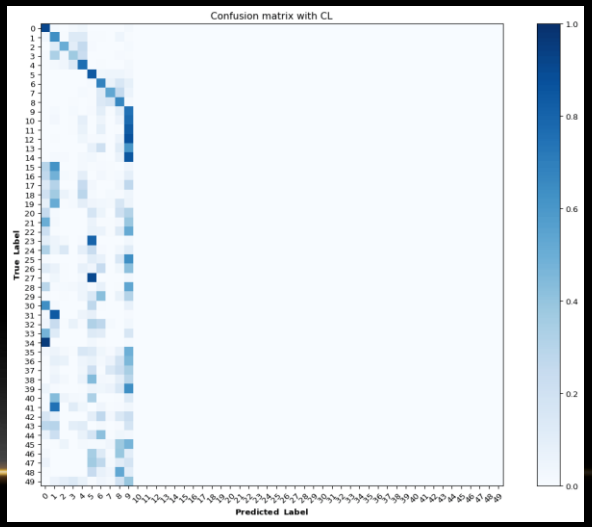
TUNING

- Esistono tecniche efficienti (es. LORA) che non modificano tutti i pesi del foundation model (molto costoso) ma aggiungono componenti addizionali addestrabili (es. matrici low rank o adapter laterali) con numero ridotto di parametri.
- Attenzione: l'addestramento incrementale di una reti neurale porta al catastrofic forgetting della vecchia conoscenza se il sistema è addestrato solo sui dati nuovi. Tecniche di Continual Learning (CL) se occorre preservare conoscenza.

SGD classico



con CL



INTERROGARE UN LLM SU BASE DOCUMENTALE

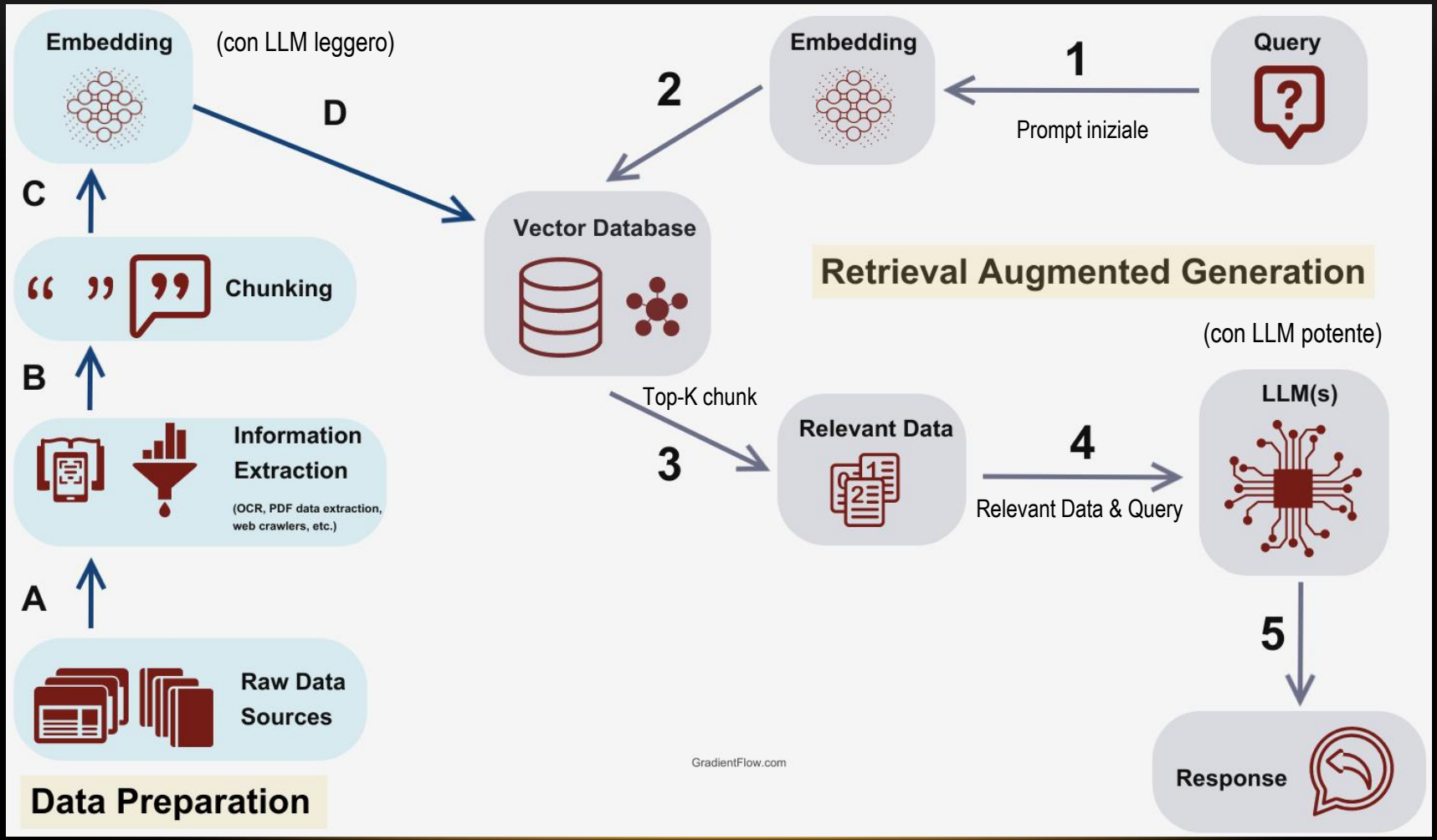
Se ho un solo documento, posso inserirlo (tutto o un suo estratto o una sua sintesi) in testa al prompt, prima della domanda.

E se i documenti sono tanti e voluminosi?

Anche se la lunghezza massima del prompt in LLM allo stato dell'arte consente di inserire molte informazioni (oltre 1 milione di token): prompt troppo lunghi sono costosi, inefficienti (risposta lenta) e anche meno efficaci (peso diverso a parti diverse)

Soluzione: RAG (Retrieval Augmented Generation), che prevede di inserire in testa al prompt non tutti i documenti completi ma solo estratti pertinenti alla domanda posta nel prompt iniziale.

RAG: SCHEMA



Fonte: https://gradientflow.substack.com/p/best-practices-in-retrieval-augmented

RAG: COME FUNZIONA ?

Fase iniziale di preparazione dati

- A, B. Si estraggono informazioni (testo) dai documenti, utilizzando opportuni estrattori.
- C. Il testo viene suddiviso in frammenti (chunk).
- D. Per ogni chunk si calcola, utilizzando un LLM «leggero», un embedding (vettore numerico multidimensionale) che ne riassume in modo compatto e semanticamente rilevante il contenuto.
- Chunk e relativi embedding si memorizzano in un Vector Database (locale).

Successivamente si possono eseguire interrogazioni

- 1. Si calcola l'embedding della query (o prompt) iniziale (con l'LLM leggero)
- 2, 3. Si recuperano dal vector database i top-K chunk i cui embedding sono più simili all'embedding della query.
- 4. Si inseriscono in testa al prompt iniziale i top-K chunk come «premesse»; il prompt così aumentato viene processato da un LLM «potente» che può reperire dalle premesse le informazioni che gli servono per rispondere (5).

RAG: ESEMPIO

36324383836	Nella riunione di martedì è stato discusso il bilancio aziendale, evidenziando la cresciata di fatturato.
87437653465	Il negozio ha un bel arredamento e il personale è gentile e accogliente.
00344232439	Il corso di formazione sulle nuove teniche di marketing sarà tenuto il giorno 15 aprile.
43675674333	I prezzi del nuovo negozio sono adeguati alla qualità della merce.
...	...

Chunk
+
Vettori

Embedding (LLM di embedding)

Nella riunione di martedì è stato discusso il bilancio aziendale, evidenziando la cresciata di fatturato.

Il negozio ha un bel arredamento e il personale è gentile e accogliente.

Il corso di formazione sulle nuove teniche di marketing sarà tenuto il giorno 15 aprile.

I prezzi del nuovo negozio sono adeguati alla qualità della merce.

...

Chunk (frammenti)

Chunking

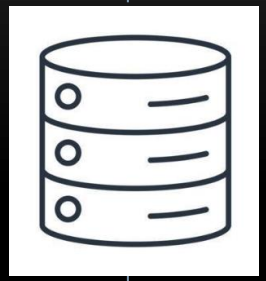
Nella riunione di martedì è stato discusso il bilancio aziendale, evidenziando la cresciata di fatturato. Il negozio ha un bel arredamento e il personale è gentile accogliente. Il corso di formazione sulle nuove teniche di marketing sarà tenuto il giorno. I prezzi del nuovo negozio sono adeguati alla qualità della merce...

Estrazione testo

Mio archivio documenti

Preparazione Dati

Vector Database



RAG

Prompt iniziale

Vorrei una sintesi dei commenti ricevuti sul nuovo punto vendita

Embedding (LLM di embedding)

053836437328

Frammenti semanticamente pertinenti inseriti in testa al prompt iniziale

Il nuovo punto vendita è molto facile da raggiungere e dotato di comodo parcheggio.

Il negozio ha un bel arredamento e il personale è gentile accogliente.

I prezzi del nuovo negozio sono adeguati alla qualità della merce.

...

Vorrei una sintesi dei commenti ricevuti sul nuovo punto vendita

Prompting (LLM finale)

Risposta

Il nuovo punto vendita è in ottima posizione, ben arredato, con personale gentile e prezzi competitivi.

RAG: COME ESTRARRE I CHUNK

- La soluzione di default (frammenti di ugual lunghezza parzialmente sovrapposti) non adeguata in molti casi
  - Non posso spezzare il testo in punti qualunque (articoli, commi, …)
  - Documenti strutturati: figure, tabelle, riferimenti esterni
  - Formati diversi (PDF, MsWord, HTML, JSON, ecc.)
- Associare metadati ai chunk:
  - Da quale documento, capitolo, sezione è stato estratto
  - Data del documento, a quale macchinario si riferisce, ecc.
- I metedati possono essere sfruttati in combinazione con gli embedding per:
  - Pre-filtrare chunk non rilevanti
  - Ordinare i chunk per rilevanza
  - Inserire nel prompt informazioni utili a recuperare l'origine
- Evoluzione: uso Knowledge Graphs, Ontologie, ecc.

RAG: ERRORI COMUNI

- Carico tutti i documenti e interrogo… ci pensa l'AI …
- Non dimentichiamo decenni di tecniche di ingegneria del software e sistemi informativi.
  - Un RAG efficace è spesso un mix di LLM e informatica tradizionale.
  - Se tutti i dati sono strutturati perché usare un LLM
  - Non disincentiviamo le buone prassi di strutturazione dei dati
- Come misurare le prestazioni?
  - Troppo spesso con qualche query estemporanea valutata soggettivamente da personale coinvolto nella sperimentazione. Non ripetibile, non confrontabile.
  - Questa fase può essere la più dispendiosa in termini di investimento (creazione di un test set, etichettatura).
  - Per alleviare carico si possono usare dataset sintetici generati a partire dal corpus documentale (es. Ares: https://arxiv.org/abs/2311.09476)

RAG SU MICROSOFT GRAPH

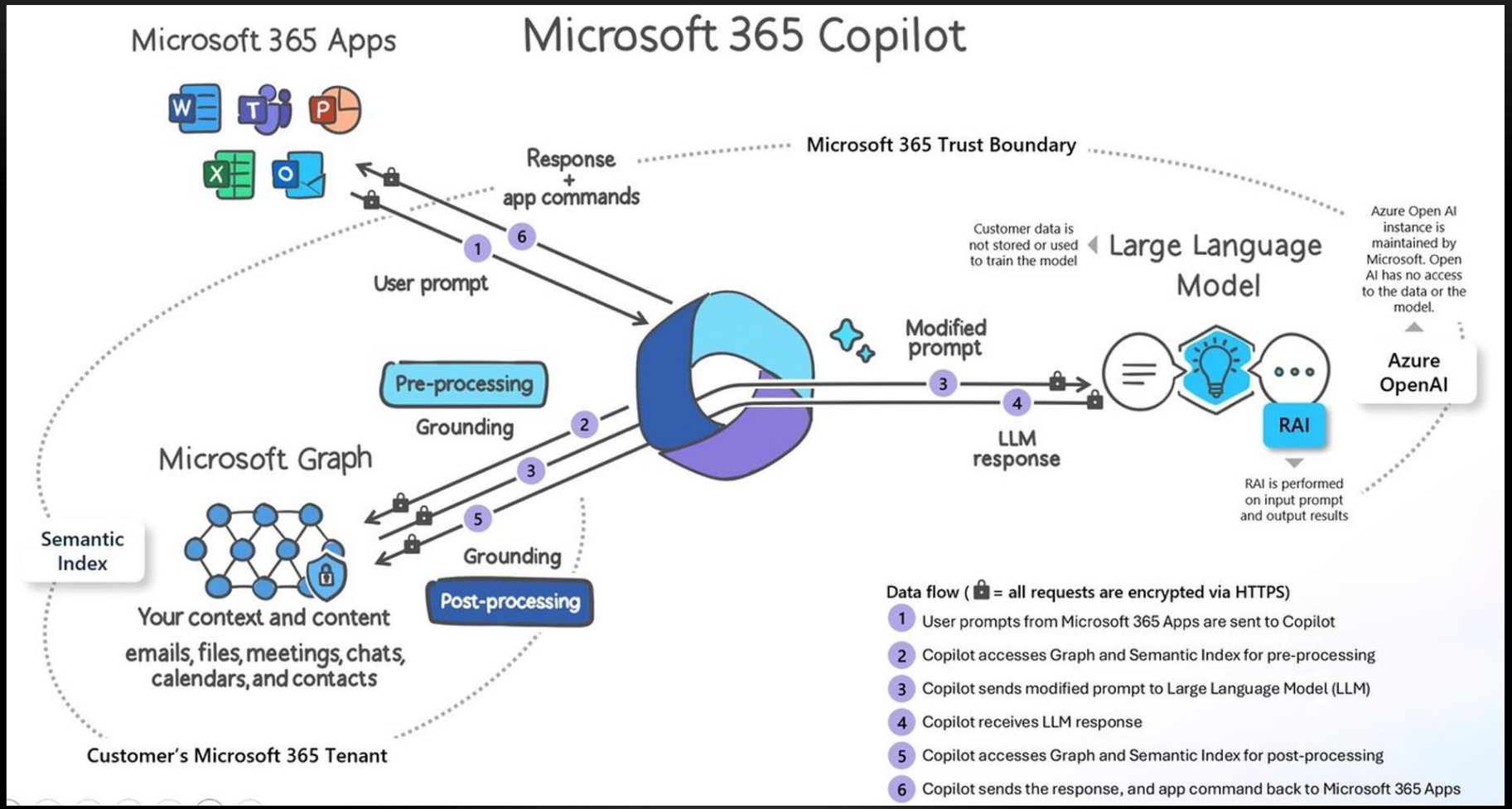
Microsoft Copilot per Office 365 fa RAG sul vostro Microsoft Graph

Cos'è il Microsoft Graph di un utente? l'insieme delle informazioni da lui prodotte utilizzando applicativi Microsoft Office 365 (archiviate nel cloud):

- Documenti Word, Excel, ecc. su OneDrive o SharePoint
- Mail ed eventi nel Calendario
- Contatti e Gruppi di lavori
- Riunioni Teams

La fase di preparazione dati (semantic indexing) avviene dietro le quinte quando il microsoft graph è modificato (es. aggiunta nuovi file).

RAG IN COPILOT



ICL – IN CONTEXT LEARNING

Si inseriscono alcuni (few-shot) esempi «strutturati» in testa al prompt.

Esempio prompt (3-shots):

Review: Questo film è noioso. Sentiment: negativo.

Review: Amo questo film. Sentiment: positivo.

Review: Questo non è il tipo di film che io amo. Sentiment: negativo.

Review: Questo è il più bel film che abbia mai visto. Sentiment:

Domanda: posso inserire in testa al promp (contesto) un dizionario di termini (spiegati) per il mio dominio verticale, invece di apprenderli con sessione di tuning?

ESEMPIO SUMMARIZATION - ICL CON COPILOT

Prompt (1-shot): Il documento /… contiene un esempio di descrizione estesa di un provvedimento, preceduta dal marcatore “Esempio 1:" e una sua sintesi preceduta dal marcatore “Sintesi Esempio 1:”. Il documento prosegue con il testo esteso di un secondo provvedimento, preceduto dal marcatore “Esempio 2:”. Adottando lo stesso stile di sintesi usato per l’esempio 1, puoi fare una sintesi per l’esempio 2? Nella sintesi concentrati sulle novità introdotte dal provvedimento e mantieni il testo compatto senza utilizzare un elenco puntato di paragrafi.

Più efficace se fatto da Bing in modalità «lavoro» perché usa un LLM più potente di quello che serve Word.

Possibile ottimizzazione (e automazione): scegliere gli esempi da inserire few-shot in base a simularità semantica con il provvedimento corrente da sommarizzare.