Università di Genova

Corso di laurea in Ingegneria Informatica

Modelli linguistici di grandi dimensioni per il supporto all'apprendimento nel contesto universitario

Candidata:
Azzurra Suffia

Anno accademico 2023/2024

Indice

Capitolo 1

Introduzione

L'obiettivo dell'elaborato è la verifica e l'analisi delle capacità di un modello linguistico di grandi dimensioni di fornire supporto didattico nell'ambito di un corso universitario, in particolare l'insegnamento di 'Reti Logiche'. Quest'ultimo costituisce l'ambiente di test, il quale, tuttavia, è caratterizzato da un alto grado di conoscenze specifiche e specialistiche, nonché da alcune convenzioni proprie. Per tale ragione sarà necessario ricorrere a diversi metodi per rendere simili informazioni accessibili al modello e migliorare i suoi risultati.

1.1 Definizioni preliminari

Un modello linguistico di grandi dimensioni (in inglese *Large Language Model*, spesso abbreviato con LLM) è un modello di intelligenza artificiale in grado di comprendere e generare testo in linguaggio naturale su argomenti di carattere generale. Gli LLMs possono infatti assolvere a compiti diversi quali completare frasi, rispondere a domande, analizzare, classificare e riassumere documenti e tradurre testi in altre lingue. Risultati così notevoli sono dovuti a una vasta quantità di dati di addestramento (libri, articoli etc.), dalla quale apprendono relazioni statistiche e regolano in maniera iterativa una moltitudine di parametri. L'esteso numero di questi ultimi (dell'ordine dei miliardi) ne giustifica il nome. Terminata la fase appena descritta, possono essere facilmente adattati per eseguire più attività ricorrendo a insiemi relativamente ristretti di dati.
Il loro funzionamento si basa sulla previsione continua del simbolo che succede e completa un testo in ingresso. In pratica, il modello analizza il testo fornito e, a partire dallo stesso, predice il simbolo (come una parola o un carattere) che più probabilmente gli farà seguito. Questa previsione avviene iterativamente: il modello aggiorna continuamente la sua predizione basandosi sul testo recepito e sui simboli che di volta in volta ha già generato. Attraverso la ripetizione di questo processo viene costruita una risposta coerente e pertinente alla stringa in input.

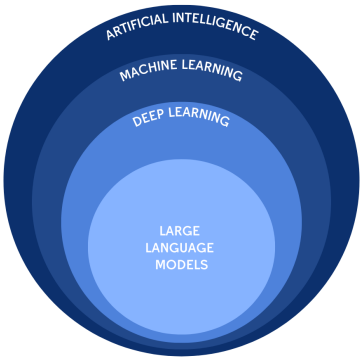

Figure 1.1: Rappresentazione gerarchica delle tecnologie AI.

Come preannunciato, per indicare al LLM di svolgere una determinata funzione tra tutte le possibili, si fornisce in ingresso ad esso una richiesta in linguaggio naturale, detto **prompt** (traducibile con "spunto").

Il modo in cui i prompt vengono formulati influenza notevolmente la qualità e la precisione della replica, e la loro sciente organizzazione consente di ottenere risposte pertinenti e accurate. La massimizzazione delle potenzialità e delle performance del modello è l'interesse del **prompt engineering**, e può essere raggiunta sfruttando alcune tecniche di prompting.
Annoveriamo le seguenti:

- Zero-Shot Prompting, creazione di prompt privi di esempi della mansione per il LLM;
- One-Shot Prompting, creazione di prompt con un solo esempio della mansione per il LLM;
- Few-Shot Prompting, creazione di prompt con un eseguo numero di esempi della mansione per il LLM.

La conoscenza degli LLMs è spesso circoscritta ai dati su cui sono stati preaddestrati: ciò rende non trascurabile l'eventuale generazione di risposte obsolete o inadeguate. In aggiunta, tali limitazioni vincolano l'utilizzo delle capacità di un LLM a determinati settori, contraddistinti da nozioni di ambito generale e di dominio pubblico. Ne segue che una mole di applicazioni che li vede protagonisti sarebbero impraticabili (ad esempio, l'utilizzo di un modello per fare supporto alla ricerca di informazioni nell'archivio di una azienda).
La **Retrieval-Augmented Generation**, o RAG, offre una soluzione al problema. Consiste nel processo di ottimizzazione del responso di un modello linguistico di grandi dimensioni, affinché possa accedere a una base di conoscenza autorevole e affidabile al di fuori delle sue fonti di addestramento prima di produrre una risposta, garantendo output aggiornati e adatti a nuovi contesti.

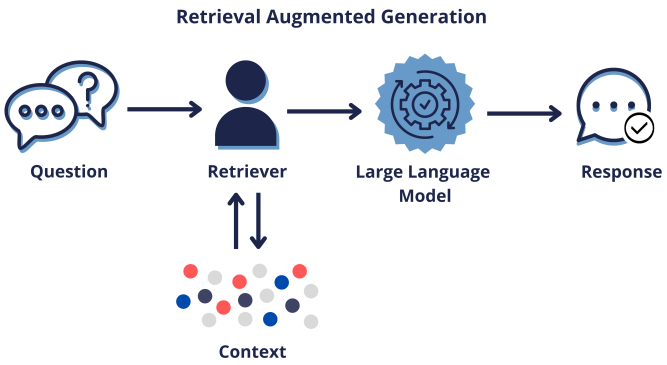

Figure 1.2: Illustrazione della procedura di recupero di informazioni e generazione di testo. Fonte: Moxoff.

Il processo di combinazione prevede l'utilizzo di una *query*. Può essere una domanda diversa dal prompt o coincidere con esso. La query viene fornita in input a un modello di embedding, che cerca le informazioni in grado di rispondere alla stessa. Il contenuto recuperato è infine condiviso al LLM assieme al prompt, il quale potrà, quindi, fare riferimento al contesto aggiuntivo nella produzione dell'output.
Sorge dunque un interrogativo: come possono i modelli di embedding comprendere il significato profondo di una parola? Se si immaginasse di chiedere a un modello di AI, che, in quanto non umano, risulta incapace di comprendere la semantica, una parola simile a 'casa', la risposta che ci si aspetterebbe sarà verosimilmente 'cosa' piuttosto che 'abitazione' o 'dimora' perché la differenza letterale è minore. Come fare dunque?
Ogni parola o frase, in realtà, viene convertita in un vettore, ovvero una serie di numeri che cattura il significato dei termini e le relazioni tra di essi. Perciò i dati esterni sono a priori trasformati e stivati sottoforma di vettori, formando un database vettoriale. Nel momento in cui LLM è richiamato per fornire una risposta, la query viene a sua volta tramutata dal modello di embedding in una rappresentazione vettoriale equivalente. In seguito, quest'ultima è confrontata con le istanze presenti nel database vettoriale e il modello conserva le informazioni la cui forma vettoriale somiglia maggiormante al vettore della query.

Capitolo 2

Google Cloud

Google Cloud (precedentemente noto come Google Cloud Platform, GCP) è una piattaforma pubblica di servizi cloud offerta da Google.
Fornisce una vasta gamma di strumenti che interessano un altrettanto ampio spettro di applicazioni e ambiti, tra cui:

- calcolo, per eseguire applicazioni e container su infrastrutture scalabili (App Engine, Kubernetes Engine, etc.);
- archiviazione, per archiviare e gestire dati, file, oggetti e database (Cloud Storage, Cloud SQL, etc.);
- analisi dati, per l'elaborazione, l'analisi e la gestione di dati, anche di grandi volumi, in tempo reale (Dataflow, BigQuery, etc.);
- machine learning e AI, per sviluppare, addestrare e distribuire modelli (Vertex AI, AutoML, etc.).

Google Cloud poggia sulla stessa infrastruttura che Google adotta per i propri prodotti, garantendo scalabilità e prestazioni elevate. Inoltre, rende accessibili al pubblico tecnologie avanzate e all'avanguardia, la cui disponibilità agevola la trasformazione digitale di aziende e organizzazioni.

2.1 Vertex AI

Vertex AI è una piattaforma completa di machine learning, grazie alla quale è possibile costruire, addestrare e distribuire modelli propri o pre-addestrati all'interno di un ambiente integrato.
Al suo interno, Vertex AI Studio consente di testare, perfezionare e arricchire modelli generativi di Google mediante la sperimentazione di prompt di esempio e la progettazione di nuovi. Rendendo semplice l'interazione con i suddetti modelli, facilita la creazione e diffusione di applicazioni potenziate con AI generativa.

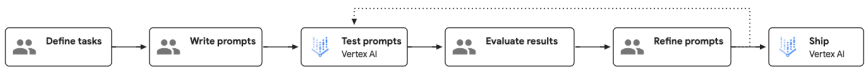

Figure 2.1: Flusso di lavoro per la progettazione dei prompt con Vertex AI. Fonte: Google Cloud.

2.1.1 Scelta del modello

Model Garden è una 'libreria di modelli', un luogo in cui cercare fra un'ampia selezione di modelli di AI generativa non limitati a quelli sviluppati da Google. Propone una sezione per ogni modello, dove sono racchiusi dettagli quali panoramica, casi d'uso e altra documentazione, per guidare la scelta dell'utente sulla base del proprio impiego specifico, livello di esperienza in machine learning e budget. È integrato con Vertex AI Studio, permettendo, infatti, di avviare lo sviluppo di progetti tramite un'interfaccia grafica. Sono presenti tre categorie principali di modelli: di base, specifici per le attività e ottimizzabili oppure open source.

Model name	Input data	Output data	Description
Gemini 1.5 Pro	Text, image, video, and audio	Text	Massive context understanding with up to 1M input tokens and robust multimodal input (text, image, video and/or audio)
Gemini 1.0 Pro	Text	Text	The best performing model with features for a wide range of tasks.
Gemini 1.0 Pro Vision	Image and text	Text	The best performing image understanding model to handle a broad range of applications.

Figure 2.2: Tabella riassuntiva di alcuni modelli della famiglia Gemini nel Model Garden. Fonte: Google Cloud.

I modelli di base sono modelli di grandi dimensioni preaddestrati e personalizzabili per attività specifiche. In questa classe rientrano Gemini per l'elaborazione multimodale e la generazione di testo, Imagen per le immagini, Chirp per il parlato e Codey per la generazione di codice.
Gemini 1.5 Pro è un modello di base multimodale, il più avanzato attualmente disponibile in Vertex AI, in grado di offrire buone performance in varie attività come la comprensione visiva, la classificazione, la sintesi e la creazione di contenuti. È in grado processare prompt multiformato contenenti testo, immagini, audio e video. Il corso di 'Reti Logiche' gode di materiali didattici quali libro di testo ed esercitazioni, in cui la componente visuale è importante, e videolezioni: le eccellenti capacità di Gemini 1.5 Pro nell'elaborazione di risorse in diverse modalità hanno determinato la scelta di questo modello per condurre l'analisi.

2.1.2 Interagire con il modello

Esistono tre modi principali con cui interfacciarsi con Gemini, ognuno dei quali raggiunge essenzialmente lo stesso obiettivo con gradi di conoscenza diversi in ambito di programmazione:

- Utilizzando un'interfaccia utente (UI) con la console di Google Cloud. Questa soluzione non richiede la scrittura di codice;
- Utilizzando SDK predefiniti con notebook come Colab e Workbench, integrati all'interno della piattaforma Vertex AI;
- Utilizzando le API di Gemini in combinazione con strumenti da linea di comando come cURL.

Ai fini della ricerca, si è prediletta la seconda modalità: poiché lo scopo dell'elaborato è testare dei prompt e valutarli, conservare tutte le richieste, e relative repliche, in un unico notebook ne esemplifica e velocizza la modifica e il confronto. In aggiunta, Vertex AI Workbench, l'ambiente

di sviluppo basato su Jupyter notebook di Google Cloud, supporta l'intero flusso di lavoro della scienza dei dati, dall'esplorazione e preparazione dei dati fino alla modellazione e al deployment.

Il seguente codice, dopo aver creato un'istanza del modello, mostra come inviare un prompt e memorizza la risposta ricevuta in `response`:

```
model = GenerativeModel("gemini-1.5-pro-001")

responses = model.generate_content(
     prompt,
     generation_config=generation_config,
     safety_settings=safety_settings,
     stream=True,
     )
```

Come precedentemente discusso, lo spunto, *prompt*, è la richiesta in linguaggio naturale inviata a un modello per ricevere una risposta.

La configurazione di generazione, *generation configuration*, contiene i valori di alcuni parametri che regolano la casualità delle risposte intervenendo sulla modalità di selezione dei token di output. Quando un modello deve produrre una risposta, genera una lista di termini che potrebbero succedere al prompt di input e ai simboli già prodotti, ciascuno con una propria probabilità. È allora necessario definire una strategia per estrarre una parola dall'elenco. L'approccio più semplice consiste nello scegliere il simbolo più probabile ogni volta. Tuttavia questa soluzione tende a dare origine a risposte ripetitive; il sorteggio casuale dalla distribuzione restituita dal modello, al contrario, crea repliche improbabili e inaspettate.
La temperatura è una opzione utilizzata per adeguare il grado di casualità: per Gemini 1.5 Pro un valore vicino allo zero restringe le parole possibili a quelle con alte probabilità, mentre un valore vicino a due estende i termini selezionabili per considere anche quelle con probabilità minori. Queste impostazioni sono utili rispettivamente quando ci si aspetta una risposta con meno variabilità e quando si desidera ottenere reazioni più creative e insolite.
Il top K restringe la scelta dell'LLM ai primi K termini in ordine di probabilità (ad esempio, se è pari a cinque, significa che la parola appartiene alle prime cinque possibili). Se, però, la distribuzione di probabilità delle parole è altamente disomogenea con un simbolo molto probabile e gli altri estremamente improbabili, questo approccio restituisce risposte incerte. Simile difficoltà è causata dalla dimensione costante del sottoinsieme di simboli da campionare.
Il top P è un parametro che garantisce che la cardinalità di tale insieme vari dinamicamente in base alla distribuzione di probabilità della parola successiva, selezionando la più piccola lista di termini la cui probabilità cumulativa è uguale a P o superiore.
Se nella configurazione di generazione sono presenti sia top K che top P che temperatura, l'elenco finale di parole candidate è determinato dall'intersezione delle prime due condizioni, mentre l'ultima influisce su come il modello distribuisce le probabilità tra i simboli che hanno superato la selezione. I valori utilizzati per la fase di analisi dei prompt sono:

```
"temperature": 0.5,
"top_p": 0.95,
"top_k": 32,
```

Le impostazioni di sicurezza, *safety settings*, definiscono i meccanismi di blocco della generazione di una risposta in base alla probabilità che possa contenere materiali dannosi, categorizzati come:

incitamento all'odio, contenuti pericolosi, contenuti sessualmente espliciti e contenuti molesti. Mediante la selezione di un livello di tolleranza tra tre disponibili (blocco ridotto, blocco limitato, blocco esteso) si imposta la meccanica di arresto della produzione di un output sospetto. Per questo studio è stato utilizzato il blocco limitato per tutte e quattro le classi.

L'abilitazione alla ricezione della risposta in tempo reale, *stream*, ha lo scopo di presentare all'utente i frammenti di testo generati man mano che vengono prodotti, anziché attendere che l'intero output sia pronto per renderlo visibile. Si è scelto di attivarlo per rendere la lettura dei risultati e la loro elaborazione più dinamica.

2.1.3 Personalizzare il modello

Si può ricorrere a diversi metodi per personalizzare e ottimizzare un modello di intelligenza artificiale generativa, quali:

- prompt design;
- adapter tuning;
- reinforcement;
- distilling.

Il prompt desing utilizza il linguaggio naturale senza bisogno di conoscenze pregresse nell'ambito del machine learning, per tale ragione è l'approccio su cui verte l'intero elaborato. Il prompt viene progettato per ottenere un risultato desiderato dal modello e migliorare le sue prestazioni. Quest'ultimo è composto da uno o più elementi:

- l'**input**, obbligatorio, il quale rappresenta l'istruzione sulla mansione che il modello deve svolgere;
- il **contesto**, facoltativo, che può servire a diversi scopi come dettagliare le istruzioni per dirigere il comportamente del modello oppure fornire informazioni aggiuntive che può utilizzare o a cui fare riferimento;
- gli **esempi**, anch'essi facoltativi, i quali si configurano come coppie di input e output che indicano al modello il formato di risposta desiderato e la logica da adoperare per rispondere.

Nell'ambito del prompt design, infatti, contesto ed esempi sono aggiunti se necessario per guidare le repliche. Il prompt design, a differenza delle altre metodologie, non modifica alcun parametro del modello pre-addestrato, tuttavia istruisce lo stesso su come deve reagire. Il suo punto di forza risiede nella rapidità di sperimentazioni e personalizzazioni; d'altra parte, una grave criticità è rappresentata dalla precarietà dei prompt progettati: piccoli cambiamenti nei termini adottati o nel loro ordine possono significativamente influenzare il responso.

2.1.4 Ampliare la conoscenza

Come visto, il contesto gioca un ruolo chiave nella progettazione dei prompt. Poiché gli argomenti del corso di 'Reti Logiche' hanno gradi di diffusione e popolarità disomogenei (passando dall'artimetica binaria, molto rinomata, all'analisi temporale di reti sequenziali, meno conosciuta), si può dedurre che Gemini abbia, a sua volta, conoscenze di base discontinue in merito. Con l'intento di abbattere simili discrepanze e accrescere le performance, nei prompt testati, che

saranno discussi nel dettaglio successivamente, sono presenti informazioni contestuali.
Un vantaggioso sistema per fornire al modello del contesto è la RAG. Due modi con cui concretizzarla sono il **recupero diretto** e il **recupero mediante tool**:

- il recupero diretto genera una stringa di frammenti testuali estrapolati dalla fonte di dati esterna. Questi ultimi vengono selezionati sulla base della retrieval query e di alcuni parametri (distanza vettoriale, top k, etc.). La retrieval query va esplicitamente indicata, quindi non necessariamente coincide col prompt. L'aggiunta della stringa di estratti al prompt consente di abbinare prompt multimediali al recupero dalla sorgente;

- il recupero mediante tool prevede l'aggiunta di un retrieval tool al modello quando questo viene istanziato: è allora sufficiente usare come richiesta per il modello affinato il prompt originario senza aggiunte. Di conseguenza, il passaggio esplicito di ottenimento della conoscenza specifica è assente. Tuttavia, siccome il grounding non ammette input non testuali, non è ammesso l'utilizzo di prompt multiformato con il modello potenziato.

Capitolo 3

Organizzazione dei prompt

3.1 Scrittura dei prompt: contesto ed esempi

I test condotti hanno coinvolto quattro diversi argomenti del corso:

1. compilazione della tabella di verità di una rete combinatoria;
2. costruzione della mappa di Karnaugh e applicazione della sintesi minima a partire da una tabella di verità;
3. completamento di un diagramma temporale di una rete sequenziale;
4. progetto di una macchina a stati finiti, o più di una, con datapath assegnato.

Per ogni classe di consegne sono stati selezionati quattro esercizi attinenti, per un totale di sedici richieste. Dopodiché, è stato interrogato Gemini con lo scopo di ottenere dal modello la soluzione per ciascuna. Tuttavia, al posto di inviare un solo prompt per esercizio, ne sono stati inoltrati nove, ognuno con una combinazione differente di fonti di conoscenza aggiuntiva e/o con esempi. L'obiettivo di tale diversificazione è rintracciare e valutare eventuali cambiamenti nella qualità delle soluzioni proposte da Gemini in relazione al contesto fornito. Infine, lo stesso procedimento è stato ripetuto anche in lingua inglese al fine di comprendere se dati di addestramento diversi, e verosimilmente più vasti in inglese, incidessero sulle risposte.

1. prompt senza contesto;
2. prompt con contesto (RAG mediante recupero diretto);
3. prompt con contesto (RAG mediante recupero diretto e video spiegazione dell'esercizio);
4. prompt con contesto (video spiegazione dell'esercizio);
5. prompt con contesto (capitoli di interesse del libro di testo);
6. prompt con contesto (capitoli di interesse del libro di testo e video spiegazione dell'esercizio);
7. prompt con contesto (RAG mediante il tool di recupero);
8. prompt con esempi;

9. prompt con esempi e contesto (capitoli di interesse del libro di testo e video spiegazione dell'esercizio).

In tutti i casi senza la dicitura 'con esempi', la strategia di prompting utilizzata è il zero-shot prompting, altrimenti si è ricorso al few-shot prompting.
Sia in italiano che in inglese la fonte di dati personalizzata e combinata alla tecnica RAG è costituita dal libro di testo del corso e da alcuni appunti esterni in lingua caricati su Google Drive.

3.1.1 Struttura dei prompt

I prompt sono strutturati come segue (in inglese gli elementi sono i medesimi, semplicemente tradotti):

```
prompt = f"""<OBJECTIVE>
{input}
</OBJECTIVE>

<FEW_SHOT_EXAMPLES>
1. Esempio #1
Input: {input1}
Output: {solution1}
2. Esempio #2
Input: {input2}
Output: {solution2}
3. Esempio #3
Input: {input3}
Output: {solution3}
</FEW_SHOT_EXAMPLES>

<CONTEXT>
Per rispondere in maniera corretta utilizza la soluzione in formato video:
{video}
Di seguito è presente del contesto addizionale che dovrebbe essere utilizzato
per produrre la risposta:
{text_retrieved}
Utilizza il file pdf di seguito per fornire la risposta, che contiene
informazioni utili su [argomento]:
{textbook}
</CONTEXT> """
```

Dove:

- `input` è la richiesta in formato testuale, accompagnata da una descrizione delle immagini incluse nella consegna se presenti;
- `inpu1`, `input2` e `input3` accompagnati da `solution1`, `solution2` e `solution3` (inseriti solo in alcune modalità) costituiscono le coppie di esempi ¡compito, risoluzione¿;

- `video` (inserito solo in alcune modalità) è il riferimento al file video in cui è spiegata la soluzione[1];
- `text_retrieved` (inserito solo in alcune modalità) è la stringa ottenuta mediante recupero diretto dalla sorgente di dati personale;
- `textbook` (inserito solo in alcune modalità) è il riferimento al file pdf contenente i capitoli di interesse del libro[1];
- `[argomento]` è una breve descrizione, inserita manualmente, dei temi che coinvolge l'esercizio in analisi.

3.2 Panoramica dei risultati

Per ciascun esercizio è stato predisposto uno schema di valutazione, criterio con cui è stato assegnato un voto a ogni output fornito dal modello adottando la seguente scala:

- Non valutabile (X, viola);
- Completamente o gravemente errata (0, rosso);
- Più errata che corretta (1, arancione);
- Più corretta che errata (2, verde);
- Completamente corretta (3, verde scuro).

I giudizi sono stati archiviati in due tabelle contenute nei file Excel *Evaluations_IT.xlsx* e *Evaluation_EN.xlsx*, dove ogni cella riporta una votazione da 0 a 3, un colore di sfondo pertinente e la giustificazione dell'esito, con riferimento allo schema sopracitato, in un commento.

[1] Il riferimento non è sufficiente a Gemini per elaborare un contenuto multimediale, per cui quest'ultimo viene anche passato come argomento a sé stante al metodo `generate_content()`.

ID	Categoria	Senza Contesto	Recupero Diretto	Recupero Diretto e Video Soluzione	Video Soluzione	Libro di Testo	Libro di Testo e Video Soluzione	Recupero mediante Tool	Esempi	Esempi con Libro di Testo e Video Soluzione
1	1	3	3	3	3	3	3	3	3	3
2	1	2	2	2	2	1	1	1	2	2
3	1	2	2	2	2	2	2	2	2	2
4	1	2	2	2	2	2	1	2	1	1
5	2	1	1	2	0	0	2	1	1	1
6	2	0	1	1	X	1	0	0	0	0
7	2	1	2	1	2	1	1	1	0	2
8	2	3	2	3	3	0	0	1	0	2
9	3	2	0	2	1	0	2	0	0	0
10	3	0	1	1	1	0	0	1	0	0
11	3	1	1	2	X	1	2	0	0	2
12	3	1	0	1	2	0	1	0	1	2
13	4	1	2	0	1	0	0	2	1	2
14	4	2	1	1	X	1	1	2	1	0
15	4	1	0	1	2	0	0	0	1	0
16	4	2	1	3	X	1	0	0	2	2

Figure 3.1: Tabella riassuntiva delle valutazioni delle risposte in lingua italiana.

ID	Categoria	Senza Contesto	Recupero Diretto	Recupero Diretto e Video Soluzione	Video Soluzione	Libro di Testo	Libro di Testo e Video Soluzione	Recupero mediante Tool	Esempi	Esempi con Libro di Testo e Video Soluzione
1	1	3	3	3	3	3	3	3	3	2
2	1	1	1	2	1	2	2	2	2	2
3	1	2	2	2	0	2	2	1	2	2
4	1	1	2	2	2	2	2	2	1	1
5	2	1	1	1	2	1	1	0	2	0
6	2	0	0	1	1	0	0	1	0	0
7	2	2	1	2	X	1	0	0	0	0
8	2	1	3	2	1	3	0	3	0	1
9	3	1	0	0	2	1	0	0	0	0
10	3	1	0	0	0	1	0	0	0	0
11	3	1	0	2	X	0	1	0	2	2
12	3	0	0	2	1	1	0	0	0	1
13	4	1	1	2	1	1	1	2	2	1
14	4	1	2	2	1	0	1	0	1	1
15	4	0	0	1	2	1	1	1	0	0
16	4	1	0	2	X	1	0	0	0	2

Figure 3.2: Tabella riassuntiva delle valutazioni delle risposte in lingua inglese.

Capitolo 4

Conclusioni

4.1 Analisi delle valutazioni

Nelle tabelle in figura 3.1 e 3.2 si riscontra un dimunizione dell'accuratezza delle soluzioni passando dalla categoria 1 alla categoria 4: la prima classe racchiude mansioni semplici e limitate, la seconda funzioni di media difficoltà, mentre la terza e la quarta incarichi più articolati.
Indipendentemente da quanto appena esposto, le tabelle, in generale, godono di scarsa intelligibilità. Per presentarne il contenuto in modo più intuitivo, le griglie sono state convertite in due file csv e caricate in un nuovo notebook, dove sono state generate rappresentazioni alternative con l'uso della libreria matplotlib.
I grafici a barre raggruppate di seguito, uno per lingua, delineano la distribuzione delle valutazioni per ognuna delle nove tipologie di prompt. La somma delle quattro barre di ogni gruppo è sempre sedici (il numero complessivo di esercizi testati), fatta eccezione per la modalità 'prompt con video soluzione' in quanto unica a restituire alcuni output non valutabili, i quali non sono inclusi nei diagrammi menzionati.

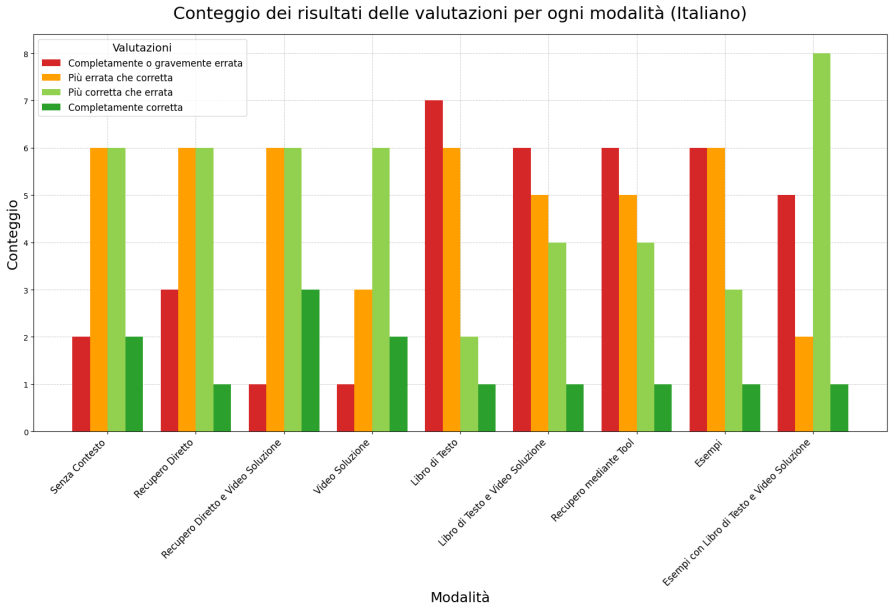

Figure 4.1: Distribuzione della qualità delle risposte per modalità in italiano

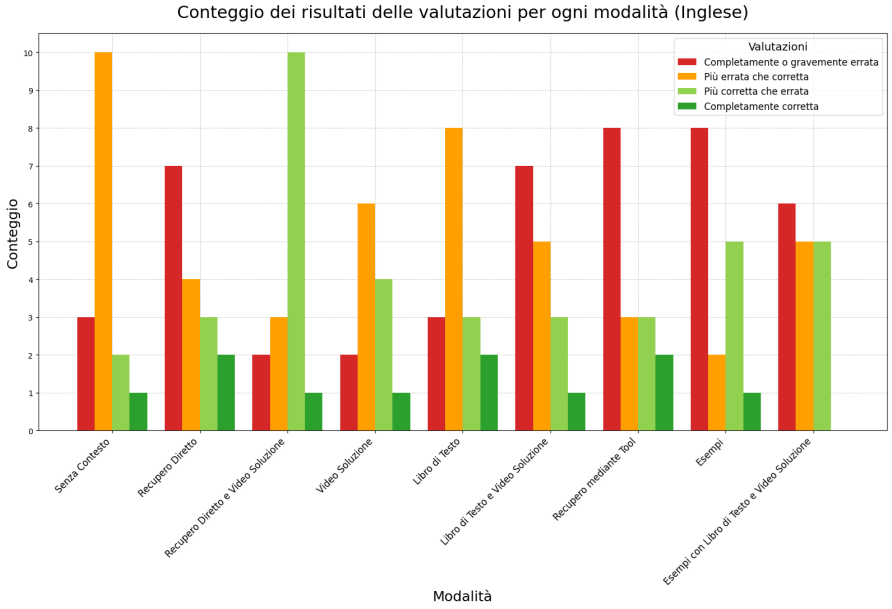

Figure 4.2: Distribuzione della qualità delle risposte per modalità in inglese

Mettendo a confronto le due visualizzazioni, si può osservare che:

- si registrano distribuzioni diverse rispetto al caso senza contesto, sia con miglioramenti che con peggioramenti;

- l'andamento delle soluzioni in inglese è in linea con il corrispettivo italiano, se non peggiore;
- in entrambi i casi il tasso di risposte completamente corrette è basso;
- in ambedue le lingue l'uso del libro di testo e video soluzione o del recupero mediante tool o di soli esempi produce responsi con una elevata percentuale di affermazioni completamente o gravemente errate.

Il livello di repliche perfettamente accurate non supera le tre unità, le quali sono, peraltro, concentrate sugli esercizi più semplici. Di conseguenza, può risultare più utile fissare come modalità di riferimento quella che presenta il numero maggiore di risposte corrette e perlopiù corrette. In inglese quest'ultima si configura come il 'prompt con recupero diretto e video soluzione' con 11 valutazioni positive, mentre in italiano si rivelano essere, a pari merito, il 'prompt con esempi, libro di testo e video soluzione' e il 'prompt con recupero diretto e video soluzione', entrambi con 9 output soddisfacenti.
Bisogna tenere presente che lavorando su dati così limitati (il cui ordine non supera 10) non si ha un significativo grado di fiducia nel generalizzare le osservazioni appena descritte.
Applichiamo lo stesso ragionemento di individuazione del prompt migliore per gruppo di consegne:

- **Categoria 1: Reti Combinatorie**
  - *Best case (Italiano)*: prompt senza contesto, prompt con recupero diretto, prompt con recupero diretto e video soluzione, prompt con video soluzione.
  - *Best case (Inglese)*: prompt con recupero diretto e video soluzione, prompt con libro di testo, prompt con libro di testo e video soluzione.
- **Categoria 2: Mappe di Karnaugh**
  - *Best case (Italiano)*: prompt con recupero diretto e video solzuione, prompt con video soluzione.
  - *Best case (Inglese)*: prompt con recupero diretto e video soluzione.
- **Categoria 3: Reti Sequenziali**
  - *Best case (Italiano)*: prompt con recupero diretto e video solzuione, prompt con libro di testo e video soluzione, prompt con esempi, libro di testo e video soluzione.
  - *Best case (Inglese)*: prompt con recupero diretto e video soluzione.
- **Categoria 4: Macchine a stati finiti**
  - *Best case (Italiano)*: prompt senza contesto, prompt con recupero mediante tool, prompt con esempi, libro di testo e video soluzione.
  - *Best case (Inglese)*: prompt con recupero diretto e video soluzione.

Notiamo che nell'analisi in inglese è riscontrabile una scelta di contesto ed esempi con un rendimento nettamente superiore alle altre, come si evince dallo schema sopra e dalla figura 4.2, mentre in lingua italiana simile picco è meno accentuato. Ciò è imputabile a una media di valutazioni positive per modalità più alta (la somma degli output più corretti che errati è 47 in italiano e 38 in inglese).

Recupero diretto e recupero mediante tool

Sia il recupero diretto che il recupero mediante tool riescono a elaborare della fonte di dati soltanto la componente testuale, per cui le immagini, fondamentali nel libro, aggiungono una informazione nulla, o minima, ai prompt.
Infatti, nel primo caso è stato riscontrato che i frammenti recuperati sono spesso sconnessi, principalmente per effetto della presenza di figure, le quali, se ignorate o non correttamente interpretate, rendono gli estratti discontinui.
D'altra parte, il recupero mediante tool si è rivelato limitante per l'impossibilità di abbinarlo a richieste multimodali.
Questo osservazioni hanno suscitato l'introduzione dei prompt con libro di testo, in quanto, i pdf, se inseriti direttamente all'interno degli input, sono interpretati dal modello come immagini. Di conseguenza le figure nel libro sono maggiormente considerate.

4.2 Possibili migliorie

Sono elencate di seguito alcune tecniche e strategie che potrebbero apportare benefici allo studio condotto:

1. **Prompt chaining:** Invece di chiedere a Gemini la risoluzione completa di un esercizio complesso, è possibile scomporre il compito in prompt minori e consequenziali. In questo modo, il modello deve portare a termine mansioni più ristrette e con minore margine di errore. Le richieste sono inviate in successione e, a ogni passo, si aggiunge all'input successivo l'output precedente formando una catena. Un esempio di prompt chaining è domandare, dato un circuito, di stabilire i valori delle uscite un fronte di clock alla volta, basandosi sulla porzione di diagramma ricorsivamente creata fino a quel punto.

2. **Estensione degli esercizi:** Ampliare il numero di consegne per categoria e le tipologie di compiti disponibili garantirebbe risultati più affidabili e renderebbe l'analisi più robusta. In più, sarebbe interessante osservare se i risultati estesi confermino o meno i correnti.

3. **Elaborazione delle immagini:** Risulterebbe utile definire una strategia per preprocessare le immagini presenti nel libro e negli appunti affinché Gemini possa apprendere anche da esse e tratte maggiormente vantaggio dall'uso della RAG.

Sitografia e videografia

[1] Ahmed Sahin, *What is Retrieval-Augmented Generation(RAG) in LLM and How it works?*, https://medium.com/@sahin.samia/what-is-retrieval-augmented-generation-rag-in-llm-and-how-it-works-a8c79e35a172, (Ultima modifica: 22 aprile 2024).

[2] Amazon Web Services, *Cosa sono i modelli linguistici di grandi dimensioni (LLM)?*, https://aws.amazon.com/it/what-is/large-language-model/#:~:text=I%20modelli%20linguistici%20di%20grandi%20dimensioni%2C%20conosciuti%20anche%20come%20LLM,con%20capacit%C3%A0%20di%20auto%2Dattenzione., (Ultimo accesso: 31 agosto 2024).

[3] Amazon Web Services, *Cos'è la RAG (Retrieval-Augmented Generation)?*, https://aws.amazon.com/it/what-is/retrieval-augmented-generation/, (Ultimo accesso: 31 agosto 2024).

[4] Boraso, *Strategie di Prompting per LLM: Zero Shot, Few Shot e Chain of Thought*, https://www.boraso.com/blog/strategie-di-prompting-per-llm-zero-shot-few-shot-e-chain-of-thought/, (Ultima modifica: 23 maggio 2024).

[5] Google Cloud, *Introduction to Vertex AI Studio*, https://www.youtube.com/watch?v=KWarqNq195M, YouTube, (Data di pubblicazione: 8 aprile 2024).

[6] Google Cloud, *Panoramica delle strategie di prompt*, https://cloud.google.com/vertex-ai/generative-ai/docs/learn/prompts/prompt-design-strategies?hl=it, (Ultima modifica: 22 agosto 2024).

[7] Google Cloud Tech, *What is Google Cloud?*, https://www.youtube.com/watch?v=kzKFuHk8ovk, YouTube, (Data di pubblicazione: 10 aprile 2022).

[8] Google Cloud Tech, *What is Vertex AI Model Garden?*, https://www.youtube.com/shorts/kPKb3yPhlwU, YouTube, (Data di pubblicazione: 27 settembre 2023).

[9] Wikipedia, *Google Cloud Platform*, https://it.wikipedia.org/wiki/Google_Cloud_Platform, (Ultima modifica: 26 luglio 2024).

[10] Wikipedia, *Modello linguistico di grandi dimensioni*, https://it.wikipedia.org/wiki/Modello_linguistico_di_grandi_dimensioni, (Ultima modifica: 23 marzo 2024).