

## ICT Training Center







### **SPRING AI**

#### GENERATIVE ARTIFICIAL INTELLIGENCE CON JAVA

Simone Scannapieco

Corso base per Venis S.p.A, Venezia, Italia

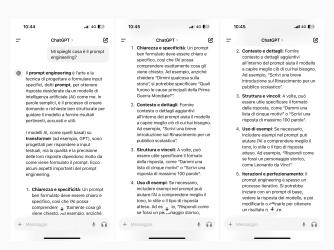
Settembre 2025

### **PROMPT ENGINEERING**

**TECNICHE DI BASE** 

### PROMPT ENGINEERING DEFINIZIONE





Definiamo meglio il concetto di prompt...





## PROMPT ENGINEERING IL PERFETTO PROMPT ENGINEER



- Sceglie uno specifico modello da ottimizzare
- Capisce come è stato addestrato il modello
- Capisce come è stato configurato il modello
- Omprende come riconfigurare il modello
- Si esprime efficacemente con il modello
  - 😜 Stile
  - Tono
  - Struttura
  - Scelta delle parole
- Spinge il LLM verso un comportamento pseudodeterministico



The Mentalist

### PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM - I



- 1 Lunghezza massima: numero di token massimo generato dal modello
  - Evitare alti consumi (energetici, economici, ...)
  - Abbattere i tempi di risposta
  - Cruciale in alcuni tipi di strategie di prompting (ReAct) e per alcune tipologie di task
  - A Non impone modifiche stilistiche al modello!



### PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM - II



- Creatività del modello: libertà nella scelta del token successivo
  - Temperatura:
    - Bassa se vogliamo risposte più deterministiche
    - Alta per risposte "creative"
    - **△** Intervallo  $[0, +\infty]$
  - O Top-K: seleziona i primi K token più probabili dalla distribuzione
    - **△** Intervallo  $[0, +\infty]$
  - Top-P: seleziona i token più probabili e la cui probabilità cumulativa non supera P
    - ▲ Intervallo [0, 1]
  - Non sempre tutti disponibili!
    - Se tutte disponibili, Top-K e Top-P scremano, temperatura decide
    - Se temperatura manca, Top-K e Top-P scremano, processo random decide



#### Attenzione alle configurazioni borderline

Setup	Effetto
$\begin{array}{l} \text{Temp} = 0 \\ \text{Temp} \gg 0 \end{array}$	Top-P, Top-K irrilevanti ( <i>greedy decoding</i> ) Temp irrilevante
$\begin{array}{c} \text{Top-K} = 1 \\ \text{Top-K} \gg 1 \end{array}$	Temp, Top-P irrilevanti (greedy decoding) Decide Temp ma processo decisamente random
Top-P = 0 Top-P = 1	Temp, Top-K irrilevanti (per maggioranza dei modelli) Top-P irrilevante



Yzma e Kronk @Disney

#### Per iniziare (secondo Google)

Obiettivo	Temp	Тор-Р	Тор-К
Risposte formali	.2	.95	30
Risposte creative	.1	.9	20
Risposte molto creative	.9	.99	40
Risposta matematica	0	_	_

# PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM - III



- 3 Sequenze di terminazione: bloccano la terminazione della generazione
  - Generazione di una lista di non più di 10 punti → stop\_sequence=[11]
- Penalità di ripetizione: incentivare il modello a sfruttare la ricchezza del suo dizionario
  - Frequency penalty: penalità calcolata su ciascun token proporzionale al numero di apparizioni nella sequenza generata
  - Presence penalty: penalità globale indipendente da token o numero di occorrenze

### PROMPT ENGINEERING COMPONENTI PRINCIPALI



Direttive istruzioni specifiche relative alla task da eseguire

Esempi per guidare il LLM verso un output più efficace

Ruolo (persona) la prospettiva che il LLM deve adottare per definire tono, stile e contenuto della risposta

"Contesto" informazioni aggiuntive che possono indirizzare e addestrare il modello in modo che generi risposte migliori, pertinenti e coerenti con l'obiettivo

Dati di input il testo contenente l'istanza della task da risolvere

Indicatore di output la tipologia o il formato che si vuole ottenere nella risposta

- Dati di *input* obbligatori
- Altri dati opzionali

### **PROMPT ENGINEERING**BEST PRACTICE PER COMPONENTE - I



- Direttive
  - A Essere chiari e concisi
  - Evitare istruzioni ambigue e vaghe
  - Quando possibile, utilizzare verbi che decrivono le azioni da eseguire ("Traduci", "Fai una lista di ...")
- Esempi
  - Chiari e rilevanti per la task
  - 🛕 Mostrare la struttura o contenuto atteso
  - A Numero di esempi comminato alla complessità della task
- Ruolo (persona)
  - Aggiungere expertise o una prospettiva specifica
  - Ruolo allineato alla task (es., esperto di marketing per generare contenuto promozionale)
  - ▲ Combinare ruolo con il "contesto"



### PROMPT ENGINEERING BEST PRACTICE PER COMPONENTE - II



- Contesto
  - Includere informazione strettamente rilevante
  - ▲ Informazione strettamente collegata alla task
  - A Fornire un background, soprattutto in sotto-casi specifici
- Indicatore di output
  - A Se possibile, sfruttare meccanismi automatici ottimizzati (es. BeanOutputConverter in Spring AI)

### PROMPT ENGINEERING ESISTE UN ORDINE?



- Esempi (se necessari)
- "Contesto" (informazioni aggiuntive)
- Ruolo (persona)
- 4 Direttive
- 6 Dati di input
- 6 Indicatore di output
- Evitare che LLM generi ulteriore contesto prima di seguire le istruzioni

### PROMPT ENGINEERING TIPOLOGIE DI PROMPTING



- Implicit solo dati di input, senza nemmeno definire le direttive
  - Sfruttare le capacità emergenti del LLM
- 0-Shot task specificata, ma la richiesta dell'utente è priva di informazioni sul contesto o priva di esempi
- Few-shots la richiesta dell'utente è seguita da alcuni esempi su cui la LLM tara il template delle risposte da fornire
- Role-based le informazioni del prompt sono suddivise tra una entità superuser (system o admin) e una entità user
  - Ontesto, definizione task e formattazione output al system
  - Input allo user

