



Il tuo partner per la Formazione e la Trasformazione digitale della tua azienda



Note			



SPRING AI

GENERATIVE ARTIFICIAL INTELLIGENCE CON JAVA

Simone Scannapieco

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

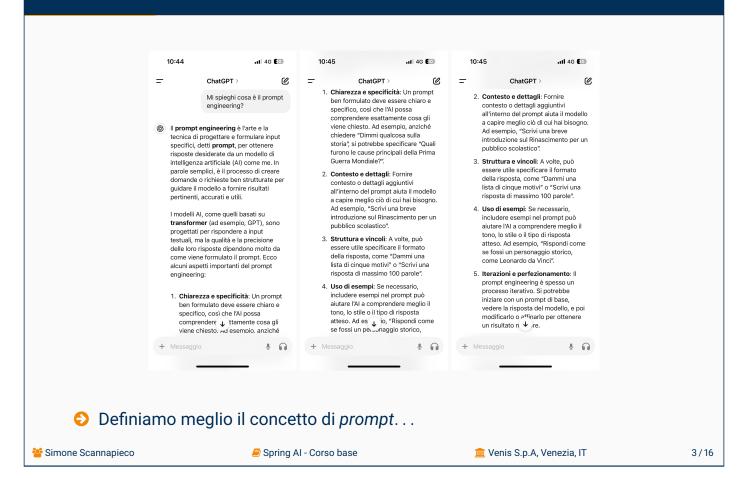
Settembre 2025

Note		

PROMPT ENGINEERING TECNICHE DI BASE	
Note	

PROMPT ENGINEERING DEFINIZIONE





Note

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

©Prime Video

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT €

Note		

PROMPT ENGINEERINGCONTROLLI 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!

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM - II



2 Creatività del modello	: libertà nella scelta del <i>token</i> si	ıccessivo	
	. Ilberta Hella decita del tenen es	30000110	
 Demperatura: Bassa se vogliar Alta per risposte ⚠ Intervallo [0, +∞ 			
•	orimi K <i>token</i> più probabili dalla	distribuzione	
\bigcirc Top-P: seleziona i t \triangle Intervallo $[0,1]$	oken più probabili e la cui proba	bilità cumulativa non supera P	
🛕 Non sempre tutti di	sponibili!		
•	oili, Top-K e Top-P scremano, tei manca, Top-K e Top-P scremano	•	
☆ Simone Scannapieco	■ Spring AI - Corso base	<u>m</u> Venis S.p.A, Venezia, lT	6/16

Note		



Attenzione alle configurazioni borderline

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



Yzma e Kronk ©Disney

m Venis S.p.A, Venezia, IT

Per iniziare (secondo Google)

Simone Scannapieco

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	_	_

Spring AI - Corso base

Note		

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]
- 4 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

🐸 Simone Scannapieco	Spring AI - Corso base	🧰 Venis S.p.A, Venezia, IT

Note		

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 addestra-

re 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

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

PROMPT ENGINEERINGBEST PRACTICE PER COMPONENTE - I



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

Note		

PROMPT ENGINEERING BEST PRACTICE PER COMPONENTE - II



11/16

Contesto

Simone Scannapieco

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

m Venis S.p.A, Venezia, IT

Spring AI - Corso base

Note		

PROMPT ENGINEERING ESISTE UN ORDINE?



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

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

PROMPT ENGINEERING TIPOLOGIE DI PROMPTING BASE



Implicit solo dati di *input*, senza nemmeno definire le direttive

Sfruttare le capacità emergenti del LLM

0-Shot task e dati di input specificati, ma 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

- Esempi, contesto, direttive e indicatore output al system
- Dati di input allo user

Step-back suddividere la richiesta finale al LLM in due fasi

- Porre una domanda generica pertinente alla task finale
- Incapsulare la risposta in un nuovo *prompt* per modellare la *task*
- Attivare dei percorsi neurali e processi di ragionamento

👺 Simone Scannapieco	Spring AI - Corso base	13 / 16

Note		

PROMPT ENGINEERING CONSIDERAZIONI FINALI - I



- Fornire esempi: strumento di insegnamento sempre efficace per allineare il LLM al nostro risultato atteso
- Mirare alla semplicità: utilizzare formulazioni ad alta entropia informativa
 - Se la formulazione è confusa ed ambigua già per l'utente, ...
 - Linguaggio semplice
 - Evitare informazione inutile ("Potresti dirmi quale..." ⇒ "Quale...")
 - ♦ Verbi che descrivono azioni ("Fai una lista...di" ⇒ "Elenca...")
- Struzioni piuttosto che vincoli: dire "cosa fare" piuttosto che "non fare..."
 - Evitare che il LLM debba chiedersi cosa è permesso generare se messo davanti ad una lista di vincoli
 - Oli LLM tendono a trovare scappatoie (ciò che non è esplicitamente vietato è permesso...)
- Mai sottovalutare le tecniche più semplici: possono riservare piacevoli sorprese...
 - Tenersi sempre aggionati sulle capacità emergenti dei modelli

Simone Scannapieco	■ Spring AI - Corso base	<u> </u>	14/1

Note		

PROMPT ENGINEERING CONSIDERAZIONI FINALI - II



- Sperimentare con formati di input e stili di scrittura
 - On caso di utilizzo di LLM in ambiti non prettamente business
 - Wave del JSON prompting e protocolli proprietari (es. OpenAl Harmony)
- Adattamento agli update del modello
 - Modellare lo stile di prompting in maniera retro-attiva
- Affidarsi ad output fortemente strutturati
 - Spring AI si sta muovendo nella giusta direzione...

Simone Scannapieco
 Spring Al - Corso base
 Venis S.p.A, Venezia, IT
 15/16

Note		

PROMPT ENGINEERING CONSIDERAZIONI FINALI - III



- Documentazione dei tentativi di *prompting*
 - Piattaforme specializzate (MLFlow)
- Sperimentazione con altri prompt engineers
 - DLLM istruita su informazioni create da miliardi di persone

Nessun uomo è un'isola, completo in se stesso; ogni uomo è un pezzo del continente, una parte del tutto.

Thomas Merton, 1956

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		