



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

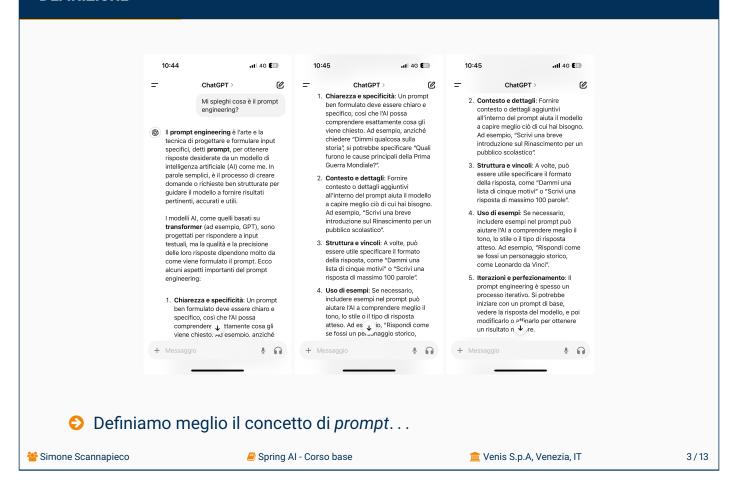
Settembre 2025

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 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!

m Venis S.p.A, Venezia, IT

Note		

PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM - II



☆ Simone Scannapieco	Spring AI - Corso base	<u> </u>	6/1
	sponibili, Top-K e Top-P screma atura manca, Top-K e Top-P scr	no, temperatura decide remano, processo <i>random</i> decide	
▲ Intervallo ▲ Non sempre			
		probabilità cumulativa non supera	Р
Top-K: selezi A Intervallo	ona i primi K <i>token</i> più probabili $[0,+\infty]$	dalla distribuzione	
Bassa se vAlta per ris⚠ Intervallo	v ogliamo risposte più determini sposte "creative" $[0,+\infty]$		
Temperatura		men euccesive	
2 Creatività del m	odello: libertà nella scelta del <i>t</i> o	oken successivo	

Note		



7/13

Attenzione alle configurazioni borderline

Setup	Effetto
$Temp = 0$ $Temp \gg 0$	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	Top-P	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

告 Simone Scannapieco	Spring AI - Corso base	<u></u> Venis S.p.A, Venezia, IT	8/1

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



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

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

PROMPT ENGINEERING ESISTE UN ORDINE?



1	Esempi ((se	necessari)
ш	Locilibi		iicccssaii)

- 2 "Contesto" (informazioni aggiuntive)
- 3 Ruolo (persona)
- 4 Direttive
- 5 Dati di input
- 6 Indicatore di output
- ▲ 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



Implicit solo dati di input, senza nemmeno definire le direttive

Sfruttare le capacità emergenti del LLM

O-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

Simone Scannapieco	Spring AI - Corso base	m Venis S.p.A, Venezia, IT	13 / 13

Note		