



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		

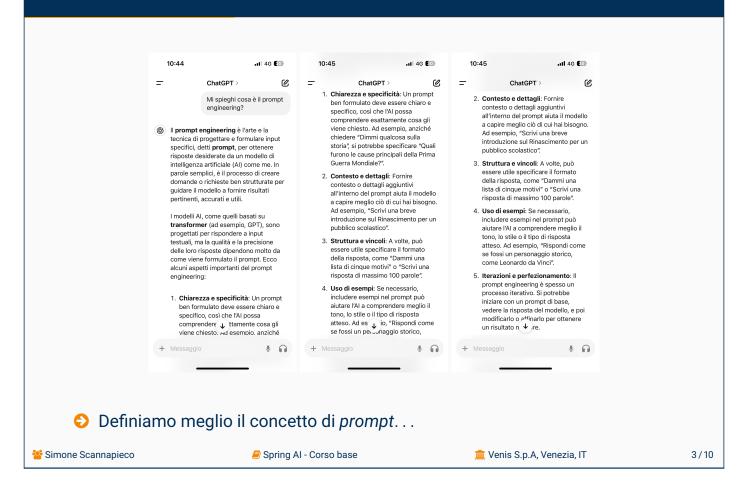
	PROMPT ENGINEERING
Note	

Г

1

## PROMPT ENGINEERING DEFINIZIONE





#### Note

#### **PROMPT ENGINEERING COMPONENTI PRINCIPALI**



Istruzione indicazioni/suggerimenti specifici che si presentano al mo-

dello per raggiungere l'azione desiderata (designano la task

da compiere)

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 ciò a cui si è interessati a trovare una

risposta

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 ENGINEERING IL PERFETTO PROMPT ENGINEER



- Sceglie uno specifico modello da ottimizzare
- Capisce come è stato addestrato il modello
- O Capisce come è stato configurato il modello
- Omprende come riconfigurare il modello
- Si esprime efficacemente con il modello
  - Stile
  - Tono
  - Struttura
  - Scelta delle parole
- Dedica il suo lavoro a tutto questo
- ▲ Spinge il LLM verso un comportamento pseudodeterministico



The mentalist ©Prime Video

👺 Simone Scannapieco

Spring AI - Corso base

🧰 Venis S.p.A, Venezia, IT

Note		

## PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM



- 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!

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

## PROMPT ENGINEERING CONTROLLI DI OUTPUT LLM



Creatività del modello: libertà nella scelta del token successivo
 Temperatura:

 Bassa se vogliamo risposte più deterministiche
 Alta per risposte "creative"
 Intervallo [0, +∞]

 Top-K: seleziona i primi K token più probabili dalla distribuzione

 Intervallo [0, +∞]

 Top-P: seleziona i token più probabili e la cui probabilità cumulativa non supera P

 Intervallo [0, 1]
 Non sempre tutti disponibili!

 Simone Scannapieco

Note	

# PROMPT ENGINEERING COME BILANCIARE TEMPERATURA, TOP-P, TOP-K?



### Attenzione alle configurazioni estreme

Setup	Effetto
Temp = 0	Top-P, Top-K irrilevanti (greedy decoding)
$Temp\gg 0$	Temp irrilevante
$\begin{array}{c} \text{Top-K} = 1 \\ \text{Top-K} \gg 1 \end{array}$	Temp, Top-P irrilevanti Temp, Top-P irrilevanti (greedy deco- ding)
$\begin{array}{c} Top\text{-}P = 0 \\ Top\text{-}P = 1 \end{array}$	Temp, Top-K irrilevanti Top-P irrilevante



Yzma e Kronk ©Disney

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT mild venezia

Note	

#### **PROMPT ENGINEERING** CONTROLLI DI OUTPUT LLM



Creatività del modello quanta libertà lasciamo al LLM nella scelta del token successivo

Temperatura più è bassa, più deterministica la generazione dei token successivi (ovvero, preso sempre il token più probabile)

> Top P definisce una finestra di probabilità entro la quale poter scegliere il token successivo (da abbassare quando si vogliono risposte fattuali, da alzare se si vuole un testo più "fuori dagli schemi")

A Non modificare entrambe!!!

Lunghezza massima numero di token massimo generato dal modello

Stop sequences lista di token che dice al modello di terminare la generazione

Token penalties contro la ripetizione dei medesimi token

Frequence penalty proporzionalmente al n° di volte usato nella sequenza

Presence penalty globale per tutti i token ripetuti

Non modificare entrambe!!!

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		

## PROMPT ENGINEERING TIPOLOGIE DI PROMPTING



Direct solo dati di input, senza nemmeno definire la task obiettivo

Utilizzare il modello secondo la sua natura specifica, senza specializzarlo

*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

Chain-of-Thought il prompt viene costruito affinché il LLM possa "ragionare" per passi sequenziali, sfruttando man mano i dati che ha generato

Utilizzato per le *task* di risoluzione problemi di logica, matematica e simili

Simone Scannapieco

Spring AI - Corso base

m Venis S.p.A, Venezia, IT

Note		