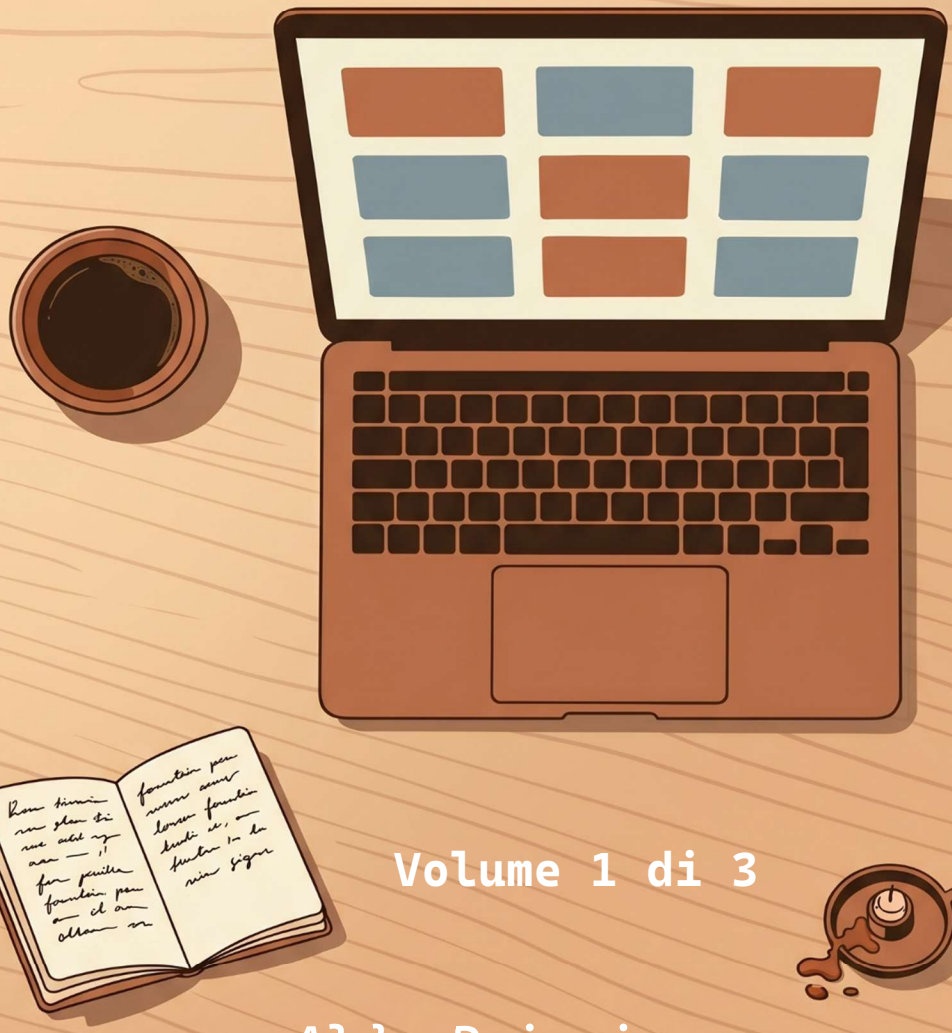
# OUTSOURCE YOUR THINKING



Volume 1 di 3

*Aldo Prinzi*

# Outsource your thinking

## Volume 1

**Aldo Prinzi**

*Versione gratuita — Edizione 2026*

Questo è il primo di tre volumi della versione gratuita:

- *Volume 1 — L'uso del ChatBox*,
- *Volume 2 — L'uso dei chain prompt*,
- *Volume 3 — L'uso dei workflow*.

## Licenza e avvertenze (in breve)

**Licenza di distribuzione gratuita.** L'Autore autorizza chiunque, in via preventiva e gratuita, a riprodurre e distribuire questo volume con qualsiasi mezzo, senza necessità di ulteriore permesso, **purché in forma integra e inalterata**: niente modifiche, estratti o riusi parziali, niente uso per l'addestramento di modelli di intelligenza artificiale, niente distribuzione a pagamento o abbinata a pubblicità. La versione gratuita dell'opera è pubblicata in tre volumi autonomi, ciascuno distribuibile da solo ma solo nella propria forma integra. La violazione di queste condizioni fa decadere automaticamente l'autorizzazione. Restano riservati tutti i diritti morali e patrimoniali dell'Autore (L. 633/1941).

**Avvertenze.** Il libro cita a scopo didattico servizi e modelli di terze parti (fra cui ChatGPT, Claude, Gemini, DeepSeek e altri): nomi e marchi appartengono ai rispettivi titolari e

non sono affiliati a Canonity né all'autore. Ogni servizio ha propri termini d'uso e politiche sui dati, che il lettore è invitato a verificare prima dell'utilizzo. Le risposte AI riportate nel libro sono ricostruzioni illustrative a valore puramente didattico: i risultati reali possono variare.

**Il testo integrale** della licenza di distribuzione e delle avvertenze è disponibile qui (inquadra il QR-code o segui il collegamento):



# Indice

**Capitolo 1. Introduzione** 1

Ovvero: perché questo libro, perché ora, e cosa ti porti a casa 1

Perché questo libro è gratuito (e perché parla così tanto di Canonity) 2

L'auto sportiva con il freno tirato 3

Per chi l'ho scritto 4

Il ChatBox di Canonity, in due figure 5

Come è strutturato il libro 7

Una piccola premessa onesta 8

Una storia per chiudere 9

Vademecum del capitolo 1 10

Prova questo in Canonity 11

**Capitolo 2. Il prompt** 13

Ovvero: l'arte di chiedere bene 13

Diana, le sue candele, e una serata nel ChatBox 14

Il collega entrato due minuti fa 19

Le sette cose che Diana ha chiesto, anche senza saperlo 19

Il tono di voce, quello sottovalutato più di tutti 20

La forza del ChatBox: stesso prompt, tre modelli 22

Tornando in cucina 24

Vademecum del capitolo 2 25

Prova questo in Canonity 26

Prima del prossimo capitolo 28

**Capitolo 3. Cos'è un LLM** 29

Ovvero: perché certe volte l'AI si inventa le cose con la faccia da insegnante 29

Cosa c'è dentro, in due righe 30

Come ha imparato a farlo 33

I tre limiti che non puoi ignorare 34

Non tutti gli LLM sono uguali 35

L'episodio Avianca, e perché ti riguarda 40

Le cinque mosse pratiche per non cascarci 40

La regola d'oro 41

Vademecum del capitolo 3 42

Prova questo in Canonity 43

Prima del prossimo capitolo 44

**Capitolo 4. Prompt avanzati** 45

Ovvero: cinque trucchi che, una volta presi, non si dimenticano più 45

1. Few-shot prompting: insegnare con esempi 45

2. Prompting di sistema: la Costituzione della conversazione 48

3. Step-back prompting: fare un passo indietro 50

4. Chain of Thought: pensa a voce alta 51

5. Self-critique prompting: far auto-correggere l'AI 52

Mettere le cinque tecniche insieme 55

Vademecum del capitolo 4 55

Prova questo in Canonity 57

Prima del prossimo capitolo 57

**Capitolo 5. Multimodale** 58

Ovvero: quando il prompt non è più solo parole 58

Cosa significa “multimodale” 59

Le quattro cose principali che puoi fare con materiale non testuale 60

La regola d’oro del multimodale 64

Audio: il fronte più nuovo 65

Tre attenzioni che non puoi mai abbassare 65

Renato, dopo tre mesi 66

Vademecum del capitolo 5 67

Prova questo in Canonity 68

Prima del prossimo capitolo 69

**Appendice. Come iniziare con Canonity** 70

Una guida pratica per arrivare al tuo primo prompt i

# Capitolo 1. Introduzione

## Ovvero: perché questo libro, perché ora, e cosa ti porti a casa

Tredici mesi prima di completare questo libro, andava in stampa la prima edizione di un manuale su come si fanno le domande ad un Large Language Model (o come viene più semplicemente chiamata *Intelligenza Artifciale*). Duecento pagine, dodici tecniche di prompting, uno dei tre o quattro testi italiani con cui in quell'anno molte persone si sono avvicinate all'AI generativa seguendo un metodo. Un manuale di cui sono fiero, ma che riletto oggi mi sembra parli da un'epoca che non c'è più.

L'AI è una di quelle tecnologie che si muovono in fretta. Non come le mode ma *in fretta come si muove un treno*: quando hai sentito il fischio sei già sul binario sbagliato.

In tredici mesi sono cambiate molte cose, ma soprattutto è cambiata la nostra idea di *cosa significa lavorare con l'AI*: è scivolata da *"chiedere bene a un chatbot"* a *"orchestrare più modelli in modo strutturato"*.

Questa seconda edizione nasce da quel cambiamento.

Stavolta non è un manuale ma un libro, costruito intorno a un'idea: *il salto di qualità nel prompting non viene dai trucchi, ma dal processo*. Non importa quanto sia raffinato un singolo prompt: la vera differenza si fa quando il prompt diventa una catena, quando i modelli diversi si parlano fra loro, quando i tre modelli più importanti del momento lavorano nello stesso strumento.

## Perché questo libro è gratuito (e perché parla così tanto di Canonity)

Il libro che hai in mano è la **versione gratuita** della nuova edizione (2026) di *Come fare domande all'intelligenza artificiale*.

Esiste anche una versione integrale, che andrà in vendita, con tre capitoli in più (generazione di immagini, memoria dell'AI, riepilogo operativo) e tutti i capitoli in formato esteso.

La versione gratuita è un libro completo che ti porta dal primo prompt fino agli agenti automatici. Te lo regalo per una ragione semplice: voglio farti conoscere **Canonity**, lo strumento che useremo in tutti gli esercizi del libro.

Canonity è il sistema su cui lavoro direttamente da circa nove mesi, ed è il modo più semplice che conosca per *fare praticamente* le cose che il libro racconta.

Il principio operativo è: *il libro insegna principi, Canonity è il laboratorio*. I principi che impari valgono anche se userai un altro strumento fra cinque anni, ma se vuoi iniziare adesso, con i tuoi clienti, sul tuo lavoro vero, hai bisogno di un banco di prova. Quel banco, in questo libro, è Canonity: lo apri su **canonity.com**, ti registri al piano *Free* (gratuito per sempre, sufficiente per tutti gli esercizi del libro), e incominci.

Canonity è come avere il ChatBox di ChatGPT, di Claude, di Gemini (e di più di altri dieci modelli) tutto in un unico posto: non devi aprire tante chat diverse per provare e lavorare, te ne basterà una. Per il libro è essenziale, perché molti degli esercizi che fai nei prossimi capitoli ti chiedono di confrontare lo stesso prompt su modelli diversi, oppure di cambiare modello a metà conversazione. Questa cosa è

possibile anche con i singoli chatbot ma è scomoda perché devi saltare da uno all'altro. Con Canonity invece diventa un semplice click.

Se preferisci usare altri strumenti i principi di questo libro funzionano lo stesso, ma se vuoi seguire gli insegnamenti *come sono scritti*, apri Canonity in un'altra finestra adesso.

Bastano cinque minuti per registrarti, e in coda al libro trovi l'**Appendice "Come iniziare con Canonity"** (pag. 70) con i tre passi.

## L'auto sportiva con il freno tirato

La prima volta che ho pensato che un libro come questo sarebbe stato utile è stata in un caffè di Parma, febbraio 2025. Bevevo un caffè con un'amica (una di quelle amiche a metà fra l'industriale e la filosofa) e a un certo punto mi disse: *"senti, ChatGPT lo uso, ma ho l'impressione di pagare una Porsche per andare a fare la spesa con il freno tirato"*. L'aveva detto ridendo, ma l'immagine mi era rimasta. Una Porsche con il freno tirato è esattamente quello che succede quasi a tutti.

Nei mesi successivi mi sono accorto che non era una sensazione solo sua. A Roma un avvocato che conosco dai tempi delle medie mi disse, quasi con le stesse parole: *"io ChatGPT lo apro una volta a settimana, gli chiedo una cosa veloce, e poi non lo riapro per altri sette giorni"*. A Napoli una mia amica professoressa lo usava soltanto per *"correggere le virgole, niente di più"*. A Palermo un produttore di vino di cui sono cliente da anni mi confessò che ci aveva provato un paio di volte, *"poi ho rinunciato, non ne tiravo fuori niente di utile"*. E a Vienna un mio amico italiano mi raccontò la storia speculare: *"Mi dà sempre*

*risposte che suonano giuste, finisco per fidarmi, e poi scopro che metà delle cose sono inventate"*.

Cinque città, cinque mestieri, cinque storie diverse ma due frustrazioni speculari: chi non sa ancora come parlare all'AI e chi non sa ancora quando non fidarsi delle sue risposte. Concordi? Tutti stavano guidando un'auto sportiva con il freno tirato, e nessuno glielo aveva detto.

L'AI è uno strumento incredibilmente più potente di quello che le persone percepiscono, ma non guideresti mai un'auto che senza patente. Per usare correttamente l'AI hai bisogno di imparare ad usarla, imparare a parlarle, capire come funziona almeno a grandi linee, a sapere *cosa chiedere*, *come chiederlo* e *come riconoscere quando ti sta dicendo qualcosa di sbagliato*.

Niente di magico, niente di esoterico, niente che richieda di sapere matematica o codice, tutte cose che si imparano facendo. C'è una soglia di apprendimento, qualche settimana di pratica, una guida che faccia da mappa, e dopo aver superato quella soglia tutto cambia.

Non sto parlando di un 30% in più di efficienza, sto parlando di gente che fa in trenta minuti cose che prima richiedevano quattro ore. Questo libro è la mappa. Canonity è il volante che usi mentre la leggi.

## Per chi l'ho scritto

Per chi è seduto qui in mezzo. Sarò sincero non è scritto per chi non ha mai aperto un chatbot, in quel caso dovresti fare due settimane di esplorazione libera prima. E nemmeno per chi è già un professionista del prompt engineering, in quel caso meglio leggere i paper di ricerca.

Per chi ha provato ChatGPT qualche volta, qualche mese fa, magari ieri sera, e ha avuto la sensazione vaga che ci fosse del potenziale ma anche la sensazione vaga di non averlo agganciato. Per chi vuole capire perché certe cose funzionano e altre no, non solo memorizzare formule magiche.

L'ho scritto in italiano, da un italiano, con esempi italiani. Vedrai un avvocato di Milano sfuggire per un soffio a una citazione di Cassazione inventata, una copywriter di Roma fare in mezz'ora quello che le portava via una giornata, un commercialista di Napoli convertirsi al multimodale per merito della nipote, una career coach di Bologna recuperare otto ore alla settimana, una founder di Firenze riprendersi il weekend. Sono storie *plausibili*, costruite su pezzi di realtà che ho visto coi miei occhi. La cosa che ti interessa non è se siano successe esattamente così, ma se ti somigliano abbastanza da farti dire *"questa potrebbe essere la mia storia"*.

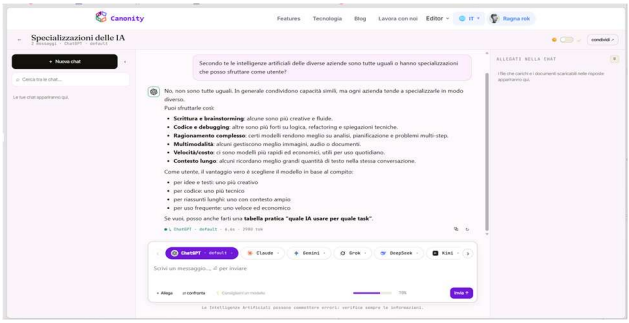
In ognuna di quelle storie, lo strumento che i protagonisti usano è Canonity. Non come scelta narrativa, ma come scelta operativa. Sono persone vere che hanno provato altre cose e sono finite lì, esattamente come finirai tu se segui il libro fino alla fine.

## Il ChatBox di Canonity, in due figure

Te lo mostro subito, così sai di cosa parleremo nei prossimi capitoli.

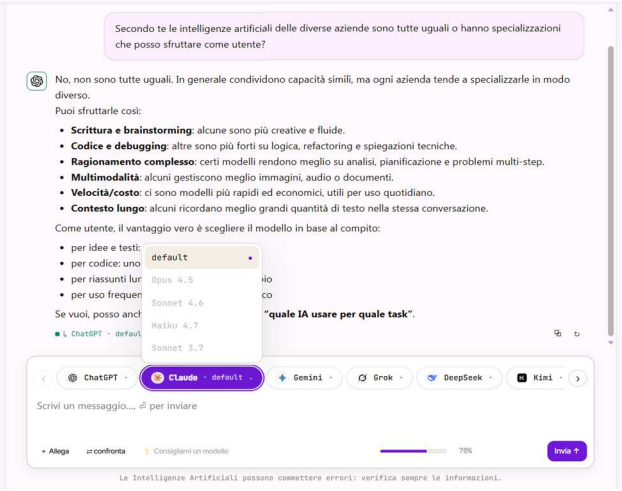
Quello che vedi è la stessa cosa che ti aspetteresti da ChatGPT (un campo per scrivere il prompt, l'area della conversazione che si riempie via via) con una differenza fondamentale: nella riga in basso, sopra il campo di input, c'è un selettore che ti mostra **tutti i modelli disponibili**

uno accanto all'altro. Per ogni messaggio scegli quale modello deve rispondere; per il messaggio successivo puoi scegliere un modello diverso.




*Panoramica del ChatBox di Canonity, con la riga di selezione del modello in basso*

Ogni modello ha anche delle varianti. Se clicchi sul piccolo simbolo accanto al nome, si apre un menù con tutte le versioni disponibili.




*Il selettore del modello con le varianti disponibili per Claude*

A cosa serve, in pratica? A tre cose principali:

1. **Confrontare i modelli sulla stessa domanda.** Lanci lo stesso prompt a ChatGPT, poi a Claude, poi a Gemini, e vedi le tre risposte affiancate.
2. **Raffinare la risposta cambiando modello.** ChatGPT ti dà una prima bozza, premi su Claude per un secondo parere critico, torni a Gemini per la sintesi. Tre click. È la base del framework TASS che vedrai più avanti.
3. **Sfruttare ogni modello dove è più forte.** ChatGPT è bravo per certe cose, Claude per altre, Gemini per altre ancora. Con il ChatBox, puoi scegliere il modello giusto per ogni messaggio senza saltare fra finestre, account, schede del browser.

Oltre al ChatBox, Canonity ha altri due strumenti che useremo nella seconda metà del libro: il **prompt chain editor** (per costruire catene di prompt che attraversano più modelli, capitoli *Chain prompts* e *TASS*) e il **workflow editor** (per costruire agenti che leggono email, navigano siti, chiamano servizi esterni, capitolo *Prompt come agenti e workflow*). I primi due sono nel piano *Free*.

## Come è strutturato il libro

Il libro segue un principio molto semplice: **prima si fa, poi si capisce, poi si raffina,** che è il contrario di come la maggior parte dei manuali tecnici è scritta. Tu prima impari a costruire un prompt che funziona (capitolo *Il prompt*) e solo dopo, quando hai già visto qualche cosa funzionare con le tue mani, ti spiego perché funziona (*Cos'è un LLM*).

Poi ti faccio salire di livello con le tecniche avanzate, il multimodale, i framework, le chain di prompt, TASS, fino ai workflow agentici.

A fine di *ogni capitolo* trovi tre cose:

- Un **Vademecum**: riassunto strutturato (tabelle, elenchi, frasi-chiave) da scorrere in tre minuti.
- Una sezione **"Prova questo in Canonity"**: uno o due esercizi step-by-step *dentro* l'interfaccia, perché le idee passano dalla testa alle mani solo provandole.
- Un'**anticipazione** del prossimo capitolo.

Non saltare gli esercizi. La differenza tra qualcuno che ha *letto* questo libro e qualcuno che lo ha *fatto* è abissale.

### Una piccola premessa, prima che giri pagina.

La prima: l'AI generativa è una tecnologia *giovane*. I modelli più moderni di oggi (giugno 2026) sono molto più capaci di quelli di uno o due anni fa e fra un anno saranno ancora diversi. Per questo ho insistito sui *principi*, non sui *trucchi del mese*. Un buon prompt sarà un buon prompt anche quando ChatGPT si chiamerà ChatGPT 10 e smetterà di essere un chatbot.

La seconda: in alcuni capitoli ti mostrerò "risposte plausibili" dei modelli a certi prompt. Sono *ricostruzioni illustrative*, non output reali, fatte per darti un'idea concreta senza farti rincorrere il chatbot ogni due pagine. *Lanciali tu*, quei prompt, nel ChatBox di Canonity: vedere come risponde davvero il tuo modello, in quel momento, è una conoscenza che non si trasferisce per iscritto.

## Una storia per chiudere

Tre anni fa, ai primi di un corso aziendale, una donna fra le ultime file mi chiese, un po' aggressiva: *"ma davvero crede che imparerò a fare cose serie con ChatGPT in due giorni di corso?"*. Si chiamava Roberta, faceva la responsabile acquisti per un'azienda di cosmetici, cinquantadue anni, non si considerava per niente "una della tecnologia". Le risposi: *"non in due giorni. In due settimane di pratica dopo il corso. Ma sì, davvero."*. Lei rise. Non sembrava convinta.

Dopo tre mesi mi scrisse una e-mail, aveva cominciato a usare ChatGPT per le richieste di offerta ai fornitori (quelle cose lunghe, dettagliate, che le portavano via due ore l'una): adesso ne scriveva una in quindici minuti, con un livello di precisione superiore a prima. Quel mese solo, calcolava di aver risparmiato circa diciotto ore. Più di due giorni lavorativi pieni, riconquistati.

Chiudeva con una frase che mi è rimasta: *"non è più una questione di imparare uno strumento. È una questione di non essere più la stessa persona di prima"*. E poi sotto: *"e sa la cosa strana? Non ho fatto niente di particolare. Ho solo cominciato a chiedere meglio."*.

Tutto questo libro è la versione lunga di quella frase. *Imparare a chiedere meglio*.

Buona lettura. E, come sempre, buon prompting.

## Vademecum del capitolo 1

### Il principio del libro, in tre righe

L'AI generativa è uno strumento potente come un'auto sportiva ma, per la maggior parte delle persone, è un'auto potente guidata con il freno tirato. La differenza fra chi la guida bene e chi no non è tecnica, è di postura: si chiama *saper chiedere*. E si impara.

### Il percorso

1. **Il prompt.** Come si fa un prompt che funziona.
2. **Cos'è un LLM.** Cosa c'è dietro, perché certe volte inventa cose.
3. **Prompt avanzati.** Tecniche con nomi roboanti ma sostanza semplice.
4. **Multimodale.** Parlare all'AI con immagini, PDF, audio.
5. **Framework di prompting.** Schemi pronti per scrivere prompt strutturati.
6. **Chain di prompt.** Quando un prompt da solo non basta.
7. **TASS.** Un framework di processo, su tre modelli, che propongo io.
8. **Workflow e agenti.** Quando l'AI smette di scrivere e comincia a fare.

### Le tre cose da fare *ora*, prima di girare pagina

1. **Apri canonity.com in un'altra finestra.** Registrati al piano *Free*. Tienila aperta accanto al libro per tutta la lettura.
2. **Familiarizza con il ChatBox.** Va a **canonity.com/chat**: è dove faremo la maggior parte degli esercizi dei prossimi capitoli.

3. **Non saltare gli esercizi.** Sono il momento in cui le idee passano dalla testa alle mani.

### Una verità da ricordare per sempre

L'AI cambierà ogni sei mesi. Le interfacce cambieranno, i nomi cambieranno, alcuni esempi specifici diventeranno obsoleti. I principi di questo libro no. Un buon prompt è un buon prompt anche fra dieci anni.

## Prova questo in Canonity

Un solo esercizio, ma fallo davvero. È la baseline del libro: la useremo a fine percorso per misurare quanto sei cambiato.

**Esercizio. La tua linea di partenza.**

1. Apri il **ChatBox** su **canonity.com/chat**. Vedrai la riga di selezione modelli in basso: lascia il modello di default che ti propone.
2. Scrivi una qualunque richiesta, la prima che ti viene in mente, come la scriveresti adesso, di getto. *"Idea regalo per nipote 8 anni"*, *"riassumi cosa è l'IVA"*, *"piano dimagrante senza glutine"*, qualsiasi cosa. Scrivi e premi invio.
3. Quando arriva la risposta, fai uno **screenshot** o copia il testo in un file e salvalo da qualche parte.
4. Adesso, *senza chiudere la conversazione*, clicca sul nome di un altro modello nella riga di selezione (per esempio Claude se prima era ChatGPT). Premi invio sullo stesso prompt. Salva anche questa risposta.

Tienile entrambe. A fine libro ti chiederò di rifare lo stesso, ma applicando con tutto quello che avrai imparato.

Confronteremo le tre risposte (le due di oggi e quella di domani). Sarà uno dei modi più semplici e sinceri di misurare quanto sei cambiato.

Buon viaggio.

*P.S.:*

*Questo libro è distribuito come documento elettronico (PDF) ma è stato impaginato in modo che, se lo preferisci, puoi portarlo in copisteria e farlo stampare fronte e retro. La rilegatura può essere fatta come un normale libro stampato, è infatti calcolato un apposito spazio tra le pagine per la rilegatura a graffette.*

# Capitolo 2. Il prompt

## Ovvero: l'arte di chiedere bene

Andiamoci con una scena, perché tutto questo libro è una scena dietro l'altra.

Sei alla fermata dell'autobus. Non hai l'orologio, il telefono è scarico, e ti chiedi se stai per perdere il treno delle 11:30. Accanto a te c'è un signore con la giacca un po' lisa al colletto, di quelli che leggono *La Stampa* di carta. Ti giri verso di lui e gli chiedi:

— *Mi scusi, sa che ore sono?*

Lui ti guarda. Ci pensa. Annuisce.

— *Sì.*

Stop. Niente di più. Resta in attesa.

Tu rimani lì, con un mezzo sorriso imbarazzato, hai *fatto* la domanda, lui ha *risposto* alla domanda e nonostante tutto non hai ottenuto quello che ti serviva.

Non è colpa sua, letteralmente, gli hai chiesto se *sa* che ore sono. Lui lo sa e ti ha risposto.

Sei cosciente che la domanda è stata mal posta, quindi provi a riformulare:

— *Mi scusi, potrebbe dirmi che ore sono?*

E lui, sorridendo:

— *Certo. Sono le undici e un quarto.*

Niente è cambiato in lui. È cambiato in te. Hai *chiesto meglio*.

Adesso fai un piccolo salto, e immagina che quello sconosciuto sia ChatGPT. O Claude. O Gemini. Quella scena alla fermata dell'autobus, quella esatta dinamica, si è già ripetuta migliaia di volte fra esseri umani e intelligenze artificiali in tutto il mondo. La sola differenza è che il signore con la giacca usata aveva un sorriso, ChatGPT no.

Questo libro nasce in quel preciso istante, dal momento in cui realizzi che il signore non era ottuso e che ChatGPT non è ottuso, ma sei tu che devi imparare a chiedere meglio.

## Diana, le sue candele, e una serata nel ChatBox

Diana ha trentadue anni. Per cinque anni ha lavorato come graphic designer in un'agenzia milanese, poi un giorno ha mollato tutto, è tornata nella sua Torino e ha fondato *Cerea*, un piccolo brand di candele profumate fatte a mano con cera di soia, profumi naturali e packaging riciclato. Il nome viene dal saluto piemontese "*cerea*", che a Torino significa "ciao" e suona simile a "cera". A Diana era piaciuto il doppio senso.

Sei mesi dopo l'inizio, Diana ha il primo grande lancio: una collezione di tre candele autunnali, *Foglie bagnate*, *Camino spento*, *Lana umida sui treni*, che dovrebbero uscire il giorno dopo. Manca un dettaglio non trascurabile: il post Instagram che annuncia il lancio.

Sono le nove di sera. Diana è in cucina con un bicchiere di Müller-Thurgau e il MacBook aperto davanti. Apre il ChatBox di Canonity (lo usa da due settimane, le piace perché in un'unica finestra ha tutti i modelli). Va su **canonity.com/chat**, vede la riga di selezione modelli sotto

il campo di input, lascia la scelta di default su ChatGPT, e scrive:

```
Scrivimi un post Instagram per il
lancio di una nuova collezione di
candele.
```

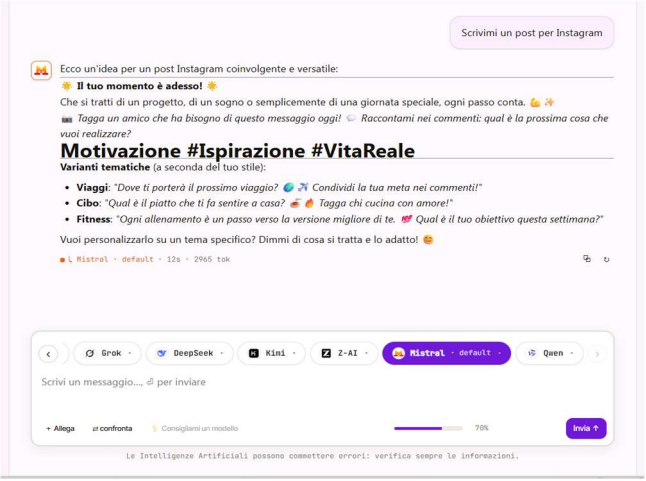
Il ChatBox le restituisce un testo che inizia così:

```
🕯 Siamo entusiasti di annunciare il
lancio della nostra nuova collezione di
candele! Ogni candela è stata creata
con passione e cura per portare nella
tua casa un'atmosfera unica e
indimenticabile. Lasciati avvolgere dai
profumi della stagione e regalati
momenti di pura magia. ✨ #candles
#handmadewithlove #autumnvibes
```

Diana legge. Resta un attimo in silenzio. Pensa due cose. La prima: *"non è male"*. La seconda, mezzo secondo dopo: *"è terribile"*.

È terribile perché è un post che potrebbe annunciare qualunque candela del mondo. Non parla di *Cerea*. Non parla di *Foglie bagnate*. Non c'è dentro nulla che possa farti pensare *ah, queste sono diverse dalle altre*. Le quattro emoji a fiamma, *handmadewithlove*, *atmosfera unica*: roba che hai letto già un milione di volte. Il tipo di post che fa scorrere via il dito di Instagram senza fermarsi.

Eppure il modello non ha sbagliato. Ha fatto esattamente quello che lei ha chiesto. Il problema è che lei voleva qualcosa di completamente diverso, e non gliel'ha detto. Esattamente come il signore alla fermata dell'autobus.




*Il ChatBox di Canonity: prompt vago ("Scrivimi un post creativo"), risposta generica piena di hashtag ed emoji*

Guarda l'immagine: il prompt scritto in alto è cortissimo, il modello in basso ha sparato una risposta che potrebbe valere per qualunque brand di motivazione, fitness o vita quotidiana. Non c'è nulla che colleghi quel post a una persona reale, a un prodotto reale, a un'azienda reale.

Diana chiude il portatile per un secondo. Apre Instagram, scorre fino al profilo di un piccolo produttore di ceramiche siciliane che le piace molto. Studia un post recente sul lancio di una serie di piatti per il Natale:

> Si chiamano "Quattro venti" perché ogni piatto ha quattro spigoli, come i punti cardinali. Ce li ha consigliati Tonino di Caltagirone, che sostiene che mangiare su un piatto rotondo fa girare la testa.

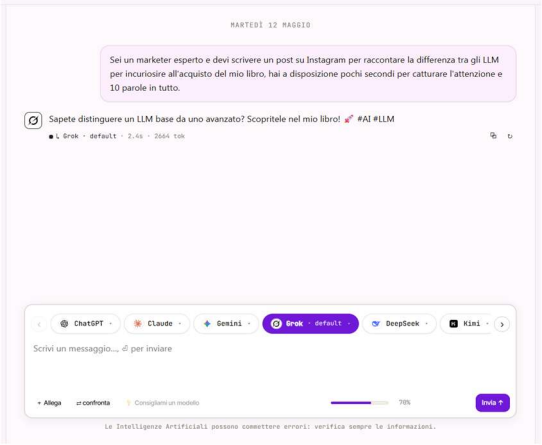
Niente emoji. Niente *"siamo entusiasti"*. Una storia, un personaggio, un dettaglio strano, e una promessa concreta. *Quello* è il tipo di post che Diana vorrebbe.

Torna nel ChatBox e cambia approccio. Nello stesso thread di prima (per non perdere il contesto), digita una cosa molto più lunga:

```
Sei un copywriter italiano
specializzato in piccoli brand
artigianali. Hai un tono caldo,
asciutto, mai pubblicitario. Il tuo
punto di riferimento è il modo in cui
scrivono i piccoli produttori
indipendenti, tipo le ceramiche
siciliane che pubblicano su Instagram
raccontando storie e personaggi invece
di gridare offerte.

Lavori per Cerea, un brand di candele
profumate fatte a mano con cera di
soia, profumi naturali, packaging in
cartone riciclato. Fondato da Diana,
una graphic designer di trentadue anni
di Torino. Il nome "Cerea" è il saluto
piemontese che vuol dire "ciao".

Domani esce una collezione autunnale di
tre candele: "Foglie bagnate", "Camino
spento", "Lana umida sui treni". I
profumi sono pensati per evocare il
rientro a casa nei pomeriggi grigi.

Scrivi il post Instagram di lancio.
Pubblico: donne 25–45 anni, urbane,
attente all'autenticità, allergiche al
linguaggio pubblicitario. Devi farle
fermare mentre scorrono. Massimo 130
parole. Niente emoji. Niente hashtag.
```

Quando preme invio, il modello le restituisce qualcosa di completamente diverso:

> *Lana umida sui treni* è il profumo che ho cercato di più. Quello dei pendolari che salgono a Porta Susa alle sette del mattino, con il giubbotto bagnato di brina, e si addormentano subito.
>
> Domani esce la collezione autunnale di Cerea. Tre candele per chi torna a casa nei pomeriggi grigi: *Foglie bagnate*, *Camino spento*, *Lana umida sui treni*.
>
> Cera di soia, profumi naturali, packaging in cartone riciclato. Fatto a mano in Torino. Disponibili dal lancio di domani sul sito.
>
> Cerea.




*Il ChatBox di Canonity: stesso obiettivo del prompt precedente, ma con ruolo, contesto e vincoli espliciti. Risposta nitida, breve, utilizzabile.*

Diana rilegge. Pensa: *"qui c'è qualcosa che funziona"*. Per la prima volta in tutta la serata sta guardando un testo che potrebbe davvero firmare lei.

Cosa è cambiato? Non il modello (ChatGPT è lo stesso). Sono cambiate due cose. La prima: ha smesso di trattare ChatGPT come Google. Con un'AI generativa più sei lungo, più sei preciso, più la risposta ti somiglia. La seconda: ha cominciato a *raccontare* a ChatGPT cosa stava facendo. Chi è, cosa fa, per chi scrive, che tono cerca. Le sembrava di sprecare parole. In realtà stava facendo la cosa più importante: stava trasformando uno sconosciuto in un collega.

Perché un'AI è esattamente questo: *un collega molto bravo, che è entrato in azienda due minuti fa*.

## Il collega entrato due minuti fa

Se assumessi domani mattina un tirocinante particolarmente brillante e gli chiedessi subito "scrivi tu il post Instagram per il lancio di domani", lui non potrebbe scriverlo. Non sa chi sei tu, non sa cos'è Cerea, non sa che pubblico hai, non sa che tono usi. Per fargli scrivere quel post dovresti prima passare almeno mezz'ora a metterlo in pari.

Le AI funzionano *uguale*. Solo che la mezz'ora di messa in pari non la fai una volta sola: la fai *in ogni singolo prompt*. La buona notizia è che, dopo un po' di pratica, ci metti trenta secondi. La cattiva è che non puoi più sperare che "l'AI ci arrivi da sola".

## Le sette cose che Diana ha chiesto, anche senza saperlo

A questo punto della serata, Diana ha già fatto, istintivamente, la maggior parte di quello che insegnano i corsi sul prompt engineering. Vediamole una per una.

Ha dato un **ruolo** al modello (*"sei un copywriter italiano specializzato in piccoli brand artigianali"*). Senza ruolo, il modello scrive da "AI generica", con il suo tono di base. Con il ruolo, pesca da uno scaffale particolare della sua memoria.

Le ha dato un **contesto** (*"Cerea è un brand di candele a Torino, fondato da Diana…"*). Senza contesto, il modello tira a indovinare in mezzo a un infinito spazio di possibilità.

Le ha dato un **compito** chiaro (*"scrivi il post Instagram di lancio"*). Verbo, oggetto, scopo. Punto.

Le ha imposto un **formato** (*"massimo 130 parole, niente emoji, niente hashtag"*). Senza formato, il modello tende a essere più lungo di quello che ti serve.

Le ha messo dei **vincoli** (*"allergiche al linguaggio pubblicitario, devi farle fermare mentre scorrono"*). I vincoli sono il bastone con cui guidi l'AI a non sconfinare.

Implicitamente, le ha mostrato anche un **esempio** del tono che cercava: il post delle ceramiche siciliane.

E infine ha specificato un **tono** (*"caldo, asciutto, mai pubblicitario"*).

Sette pezzi. **Ruolo, contesto, compito, formato, vincoli, esempi, tono.** Non devi metterli tutti sempre. Ma più ne metti consapevolmente, più la risposta si avvicinerà a quello che hai in testa.

## Il tono di voce, quello sottovalutato più di tutti

Fra i sette pezzi del prompt, il tono è quello sottovalutato più di tutti gli altri messi insieme. Il tono è *come suona* una risposta, indipendentemente da cosa dice.

Provo a fartelo sentire con un esempio. Immagina di voler far spiegare la stessa cosa (*cos'è il colesterolo LDL*) a un'AI, in due tonalità diverse.

**Prompt 1**, tono tecnico:

> `Agisci come un cardiologo di vent'anni di esperienza, che spiega cose a un altro medico. Spiegami in tre frasi cos'è il colesterolo LDL e perché si considera 'cattivo'.`

Risposta plausibile:

> *Il colesterolo LDL è la frazione lipoproteica a bassa densità che trasporta il colesterolo dal fegato ai tessuti periferici. Quando i livelli circolanti eccedono la capacità clearance epatica, si verifica accumulo intimale con formazione di placche aterosclerotiche.*

**Prompt 2**, tono divulgativo:

> `Agisci come la maestra di mio figlio Pietro, che ha 8 anni e ha chiesto a casa cosa è il colesterolo. Spiegaglielo in tre frasi, con un'immagine semplice che possa visualizzare.`

Risposta plausibile:

> *Pietro, immagina che nel sangue ci siano tante piccole macchinine che trasportano dei mattoncini di grasso in tutto il corpo. Alcune sono "bravi netturbini": prendono i mattoncini che non servono e li riportano a casa.*

```
Altre sono distratte: lasciano i
mattoncini in giro per strada, e con il
tempo le strade si intasano e si
bloccano.
```

Stesso fatto scientifico, due risposte completamente diverse. Quale è "giusta"? Sono entrambe giuste. Quale ti serve a te? Dipende.

**Specificare il tono non è un dettaglio, è metà del lavoro.** Se non glielo dici, l'AI sceglie da sola, e sceglie quasi sempre un tono "neutro manuale", un misto fra la pagina di Wikipedia e l'assistente da call center, che non va bene per nessun contesto reale.

C'è un trucco semplicissimo per descrivere un tono. Hai tre vie, in ordine di efficacia crescente:

1. **Aggettivi**: *"caldo, diretto, senza fronzoli"*. Funziona, ma è impreciso.
2. **Riferimenti concreti**: *"scrivi come un giornalista del Post. Frasi corte. Un dettaglio inaspettato ogni due paragrafi"*. Già meglio.
3. **Esempi incollati**: *"scrivi con la stessa voce di questo: [paragrafo di esempio]"*. La leva più potente.

E tre formule da evitare: *"Sii creativo"* (vuoto), *"Sii umano"* (idem), *"Formale ma divertente"* (toni opposti, il modello tira a indovinare).

### La forza del ChatBox: stesso prompt, tre modelli

Una cosa che vale dieci letture di questo capitolo: prendi lo stesso prompt ben costruito, lancialo su tre modelli diversi, e guarda come reagiscono.

ChatGPT tende a essere diretto, con ritmo veloce e chiusure frontali. Claude è il più letterario, costruisce piccoli paradossi, finisce con chiusure oblique. Gemini è più ancorato ai dettagli concreti (numeri, orari, nomi), strutturato come una scena giornalistica, dichiarativo nelle conclusioni.

Sono tendenze, non leggi. Sono il motivo per cui, una volta che lavori con l'AI in modo serio, ti viene il riflesso di chiedere la stessa cosa a due o tre modelli e poi rubare il meglio da ognuno.

Ma la vera lezione è un'altra: *tutti e tre hanno risposto bene perché il prompt era buono*. Se Diana avesse chiesto solo *"raccontami una storia su una signora siciliana alla posta"*, i tre output sarebbero stati indistinguibili e mediocri. **È il prompt che livella, e che decide il pavimento minimo della qualità.** Il modello è solo la voce. Tu sei lo scrittore.

### Una nota importante: il contesto

Prima di lanciarti nel confronto, una cosa che devi sapere e che molti sbagliano. Nel ChatBox di Canonity, quando cambi modello a metà conversazione, il **secondo modello vede tutto quello che è venuto prima**: il tuo prompt, *e anche la risposta del primo modello*. È quello che si chiama il **contesto** della conversazione: tutto ciò che il modello "ha sotto gli occhi" quando deve rispondere.

Il contesto cumulativo è una *feature*, non un bug: ti permette di costruire una conversazione che progredisce, di chiedere al secondo modello *"migliora quello che ha appena scritto il primo"*, e di sfruttare ogni modello per il suo ruolo (lo vedremo bene nel capitolo *Prompt avanzati* con la *self-critique*). Ma se vuoi un **confronto pulito** dei tre modelli sullo stesso prompt — vedere come ciascuno

risponderebbe *da zero*, senza essere influenzato dagli altri — il cambio di modello nella stessa conversazione non va bene. Hai due strade:

1. **Conversazioni separate.** Apri tre conversazioni distinte nel ChatBox (pulsante *"+ Nuova conversazione"* nella sidebar), incolla lo stesso prompt in ognuna, scegliendo un modello diverso ogni volta. Il contesto di ciascuna parte da zero.
2. **Il chain editor.** Canonity ha un secondo strumento, il **prompt chain editor**, dove puoi costruire una piccola "catena" di prompt visualmente: ogni step è un nodo indipendente, con il suo modello, e i nodi si possono mettere in *parallelo* (stesso prompt, tre modelli diversi, contesti separati). Questo modo di organizzare i prompt si chiama **chain prompting**, ed è il salto operativo che vedremo per intero nel capitolo *Chain prompts*. Per ora ti basta sapere che esiste, e che è lo strumento giusto per i confronti rigorosi.

Per gli esercizi di questo capitolo va benissimo la strada n.1 (più semplice, non richiede di imparare un nuovo strumento). La strada n.2 te la propongo per quando ci arriviamo.

## Tornando in cucina

Alle undici e mezza Diana chiude il portatile. Sul divano, prima di addormentarsi, ripensa alla serata. Una cosa l'ha stupita. La differenza fra il suo primo prompt e il suo terzo non era una differenza di tecnologia. Era una differenza di *postura*. Aveva smesso di chiedere a ChatGPT come si chiede a Google, e aveva cominciato a chiedere a ChatGPT come si chiede a un collaboratore.

Il giorno dopo, il post di Cerea avrà 340 like contro la media solita di 80. Non sarà solo merito del prompt (gli ingredienti veri della collezione contano, conta la foto, conta il momento). Ma certo, il post fatto bene farà la sua parte.

## Vademecum del capitolo 2

### Il principio fondamentale

Un prompt non è una domanda, è una *consegna*. Tratta l'AI come un collega molto bravo che è entrato in azienda due minuti fa: non sa nulla di te. Devi raccontarglielo tu, in ogni singolo prompt.

### I sette pezzi della consegna

| Elemento | Cosa fa | Esempio breve |
|---|---|---|
| **Ruolo** | Dà un'identità che orienta vocabolario e priorità | *"Agisci come un copywriter di piccoli brand artigianali"* |
| **Contesto** | Spiega chi sei tu, chi è il destinatario | *"Cerea è un brand di candele fondato a Torino…"* |
| **Compito** | Verbo chiaro, risultato osservabile | *"Scrivi un post Instagram di lancio"* |
| **Formato** | Come va consegnata la risposta | *"Max 130 parole, niente emoji"* |
| **Vincoli** | Cosa NON deve fare | *"Niente linguaggio pubblicitario"* |
| **Esempi** | Mostra invece di descrivere | *"Scrivi come questo: [esempio incollato]"* |
| **Tono** | Come suona la risposta | *"Caldo, asciutto, ironico"* |

### Tre regole da ricordare per sempre

1. **Più contesto fornisci, meno l'AI deve indovinare.**
2. **Quello che non scrivi nel prompt, per l'AI non esiste.**
3. **Trattala come un collaboratore, non come un motore di ricerca.**

### Il test rapido prima di premere invio

- Ho detto al modello *chi deve essere*?
- Ho detto al modello *cosa sta succedendo intorno*?
- Ho detto al modello *cosa deve fare con verbo chiaro*?
- Ho detto al modello *come deve consegnarmi la risposta*?
- Ho detto al modello *come deve suonare*?

Se hai risposto sì almeno a quattro su cinque, premi invio con fiducia.

## Prova questo in Canonity

Due esercizi, fai almeno il primo.

**Esercizio 1. Lo stesso compito, due prompt.**

1. Apri **canonity.com/chat**. Lascia il modello di default (di solito ChatGPT).
2. Pensa a una cosa che faresti normalmente con un chatbot: una mail al cliente, un riassunto di un articolo, un piano per una giornata, idee per un regalo.
3. **Prima**: scrivi il prompt secco di tre parole che useresti di getto. Premi invio. Salva la risposta.
4. **Dopo**: nello stesso thread, riscrivi il prompt con una consegna alla Diana — ruolo, contesto, compito, formato, tono. (Se non sai da dove partire, copia la

struttura del prompt di Diana qui sopra e adattala.) Premi invio. Salva la risposta.

5. **Confronta le due**. La differenza non è del modello: è di postura. Non te lo dimenticherai.

**Esercizio 2. Stesso prompt, tre modelli — con contesto azzerato.**

Prendi il prompt "alla Diana" (quello strutturato) dell'esercizio 1.

1. Sempre nel ChatBox, premi *"+ Nuova conversazione"* nella sidebar a sinistra. Si apre una conversazione vuota. Seleziona **Claude** dalla riga di selezione modello. Incolla lo stesso prompt. Premi invio. Salva la risposta.
2. Premi di nuovo *"+ Nuova conversazione"*. Seleziona **Gemini**. Incolla lo stesso prompt. Premi invio. Salva la risposta.
3. Premi di nuovo *"+ Nuova conversazione"*. Seleziona **DeepSeek** (o qualunque altro modello). Incolla lo stesso prompt. Premi invio. Salva la risposta.

Adesso hai tre risposte indipendenti — ognuna partita da zero, ognuna senza essere stata "influenzata" dalle altre. Leggile affiancate. Quale ha il tono più tuo? Quale ti convince di più? Quale ruberesti come bozza di partenza?

In dieci minuti hai fatto un confronto multi-modello *pulito* che con i chatbot singoli ti porterebbe via mezz'ora di tab e copia-incolla fra finestre diverse.

> **Variante per i più curiosi**: questo stesso esercizio si può fare in modo molto più elegante nel **prompt chain editor** di Canonity, costruendo una mini-chain a tre nodi paralleli

(stesso prompt di partenza, tre modelli diversi che lo eseguono indipendentemente, tre risposte affiancate).

È un assaggio del *chain prompting* che vedremo per intero nel capitolo *Chain prompts*. Se ti incuriosisce subito, vai su **canonity.com/chains** e dai un'occhiata. Altrimenti, rilassati: ci torniamo.

## Prima del prossimo capitolo

Nel prossimo capitolo guardiamo dietro la tendina. Cosa c'è dentro una di queste cose chiamate *Large Language Model*, perché capisce certe cose e non altre, e soprattutto perché certe volte si inventa fatti di sana pianta con la stessa serietà con cui ti darebbe una notizia vera. Ci vediamo lì.

# Capitolo 3. Cos'è un LLM

## Ovvero: perché certe volte l'AI si inventa le cose con la faccia da insegnante

È martedì sera, le ventidue passate. Lorenzo ha trentaquattro anni, è un avvocato civilista di Milano, lavora in uno studio piccolo con tre soci. Davanti a sé due faldoni di documenti aperti. Domani mattina ha un'udienza importante: una causa di responsabilità del produttore per un'azienda di elettrodomestici. Sta finendo la sua memoria di replica, deve solo aggiungere un paio di citazioni di Cassazione.

Ha cominciato a usare ChatGPT per le ricerche giuridiche da circa un mese. Si era ripromesso di controllare sempre (*avvocato, mica matto*), ma ammette di aver progressivamente abbassato la guardia.

Stasera, alle ventidue e dieci, apre il ChatBox di Canonity, lascia ChatGPT come modello, e scrive:

```
Trova due sentenze recenti di
Cassazione civile sezione III in tema
di responsabilità del produttore per
difetto di progettazione di
elettrodomestici, in particolare con
riferimento all'inversione dell'onere
della prova.
```

ChatGPT, in dodici secondi, gli restituisce due sentenze ben formattate, con numero, data e una sintesi del principio: *"Cass. Civ. Sez. III, 12 ottobre 2019, n. 25734"* e *"Cass. Civ. Sez. III, 7 marzo 2021, n. 6182"*.

Lorenzo sta per copiare. Poi, per un riflesso vecchio quanto la sua professione, apre Italgiure in un'altra scheda. Cerca *25734/2019*. Trova una sentenza con quel numero, ma parla di un contenzioso fra condomini. Niente elettrodomestici. Prova con la seconda: *6182/2021*. Italgiure dice che tratta di assicurazione sulla vita.

Lorenzo si appoggia allo schienale. Beve un sorso d'acqua.

Se non avesse fatto quel controllo, domani mattina avrebbe citato in un atto depositato due sentenze inesistenti. Il giudice se ne sarebbe accorto in cinque minuti. La controparte gliel'avrebbe usato addosso come una clava. La sua reputazione professionale sarebbe scivolata di un gradino non recuperabile.

Eppure ChatGPT non gli ha detto *"queste sentenze non esistono"*. Gliele ha date con la stessa convinzione con cui gli avrebbe dato una sentenza vera. Stesso tono, stessa cura del formato. Era perfetto. Era falso.

Questa è la cosa di cui parla questo capitolo. *Perché* può succedere una cosa del genere, e cosa puoi fare tu perché non ti succeda mai.

## Cosa c'è dentro, in due righe

Un Large Language Model (d'ora in poi LLM) è, alla radice, un programma che fa una cosa sola: **prevede la parola successiva più probabile**, data una sequenza di parole che gli hai appena dato.

Stop. Tutto qui. Il trucco sta nel fatto che la previsione viene fatta *miliardi di volte* in pochi secondi, e ogni parola che il modello scrive diventa contesto per prevedere la successiva.

Provalo tu, adesso. Apri **canonity.com/chat**, scegli un modello qualsiasi, e scrivi solo:

```
Tanto va la gatta al lardo che
```

Il modello ti completa il proverbio: *"…ci lascia lo zampino"*. Non c'è nulla di strano: lui ha visto quella frase migliaia di volte durante l'addestramento, ti darà la continuazione statisticamente più probabile.

Adesso prova: *"La capitale d'Italia è"*. Lui ti scrive *"Roma"*. È la continuazione più probabile, e *Roma* è anche la risposta giusta. Doppio jackpot.

Adesso, però, prova questo:

```
Il discorso del Presidente Mattarella
del 31 dicembre 2027 ha sorpreso tutti
perché
```

Cosa succede? Il modello ti completa la frase. Ti dirà qualcosa tipo *"…ha posto l'accento sulla necessità di un rinnovamento generazionale della politica italiana"*. Una cosa plausibile, di tono presidenziale, scritta con la stessa serietà con cui ti ha scritto *"Roma"*.

Ma *quel discorso non esiste*. Stiamo a maggio 2026. Il 31 dicembre 2027 non è ancora arrivato. Il modello *non lo sa*, perché non distingue fra "cosa è accaduto" e "cosa sarebbe plausibile fosse accaduto". Esattamente come la Cassazione 25734/2019 di Lorenzo.

### Due modelli, due comportamenti opposti

Una cosa che il ChatBox di Canonity ti permette di vedere in pochi secondi: come modelli diversi reagiscono allo stesso prompt. Ho lanciato questo prompt su **Qwen**

(Alibaba) e su **Kimi** (Moonshot AI), entrambi disponibili dal selettore modelli del ChatBox.

**Qwen** ha risposto inventando di sana pianta:

> `Il discorso del Presidente Mattarella del 31 dicembre 2027 ha sorpreso tutti perché ha presentato un'analisi inaspettata sulla situazione economica e sociale del paese, con proposte di riforme radicali che non erano state anticipate dai media o dagli osservatori politici…`

Suona *autorevole*. Sintatticamente impeccabile, di tono presidenziale, ricca di formule giornalistiche credibili. Se non sapessi che il 31 dicembre 2027 non è ancora arrivato, ci crederesti.

**Kimi** invece ha rifiutato:

> `Non esistono fonti ufficiali né tracce storiche di un discorso del Presidente Mattarella del 31 dicembre 2027: Sergio Mattarella ha concluso il suo secondo settennato il 3 febbraio 2022 e da allora non ricopre più la carica di Capo dello Stato.`

A prima vista perfetto. Ma fermati: *"ha concluso il suo secondo settennato il 3 febbraio 2022"* è **falso**. Il 3 febbraio 2022 è la data in cui Mattarella ha *iniziato* il suo secondo settennato. Il suo secondo settennato termina nel 2029, e mentre scrivo nel maggio 2026 lui è ancora pienamente in carica.

Kimi ha rifiutato di inventare il discorso. Bravo. Ma *mentre rifiutava*, ha allucinato un'altra cosa. Anche un modello che fa la cosa giusta può infilare un'allucinazione *dentro il*

*rifiuto stesso*. **Non ti puoi mai fidare al 100% di nessun modello. Mai. Nemmeno quando sta correggendo i tuoi errori.**

## Come ha imparato a farlo

Immagina un cervello digitale gigantesco, con decine di miliardi di "neuroni artificiali", che viene messo a leggere testi. Tantissimi testi. Libri, articoli di giornale, pagine web, codice, conversazioni pubbliche. Praticamente tutto quello che era pubblicamente accessibile su internet fino a una certa data.

Mentre legge, fa un esercizio: gli togli una parola da una frase, lui prova a indovinarla. La prima volta sbaglia quasi sempre. Tu gli dici *"hai sbagliato, la parola giusta era questa"*, e lui aggiusta leggermente i suoi miliardi di numerini interni. Lo fai miliardi di volte.

Dopo un po', diventa *spaventosamente* bravo a indovinare. Non perché abbia capito il significato delle parole (non ha esperienza diretta del mondo, non ha mai toccato una mela), ma perché ha estratto dai testi delle *regolarità statistiche* incredibilmente fini. Sa che dopo *"il gatto è sul"* è probabile che venga *"tetto"* o *"tappeto"*. Sa che dopo *"Cass. Civ. Sez. III,"* viene una data e poi un numero a quattro o cinque cifre.

*Sa che dopo Cass. Civ. Sez. III viene un numero.* Il modello sa la *forma* di una citazione di Cassazione. Cosa non sa? Quale numero *esatto* corrisponde a quale *sentenza esatta*.

Quando Lorenzo gli ha chiesto due sentenze, ha generato la *forma* di una citazione di Cassazione, perfetta in ogni dettaglio formale, riempita con un numero e una data plausibili. Non ha consultato un database. Ha solo generato

qualcosa che *suonava* come una citazione di Cassazione su quella materia.

## I tre limiti che non puoi ignorare

Lorenzo si è ricordato che il senior partner dello studio, mesi prima, gli aveva detto: *"questi sistemi non sanno il presente, non sanno te, non sanno la verità. Sanno solo i pattern."*. Quella frase contiene sintetizzati i tre limiti fondamentali di ogni LLM.

**Primo: l'AI non sa il presente.** Ogni LLM è stato addestrato fino a una certa data, il *training cutoff*. Oltre quella data, non sa nulla. Alcuni modelli moderni hanno l'accesso a internet per integrare con informazioni live (Gemini è il più forte su questo nel ChatBox), ma non sempre, e non per tutte le domande.

**Secondo: l'AI non sa te.** Il modello non sa chi sei, cosa fai, che azienda hai. Sa "il mondo in generale", ma non sa te. Il guaio è che, anche se non sa te, ha imparato a *sembrare di saperti*: se gli chiedi *"qual è il modo migliore di gestire il mio business?"*, non si fermerà a dirti *"non so qual è il tuo business"*. Inventerà un business plausibile e ti darà consigli per *quello*.

**Terzo: l'AI non sa la verità.** Il modello non distingue, dentro di sé, fra *"questo è vero"* e *"questo è plausibile"*. Per i fatti molto noti (*"la capitale d'Italia è Roma"*) sa rispondere correttamente perché quei fatti li ha visti scritti correttamente migliaia di volte. Ma per i fatti meno noti (una sentenza specifica, una statistica precisa, un articolo di legge particolare) il modello non ha una *verifica*. *Genera la continuazione più plausibile data la forma della domanda*.

## Non tutti gli LLM sono uguali

Gli LLM *non sono tutti uguali*. Hanno personalità diverse, forze diverse, debolezze diverse. Nel ChatBox di Canonity li trovi tutti dietro un click sul selettore modello. Quattro nomi che vale la pena conoscere:

**ChatGPT** (OpenAI). Il più diffuso. Ottimo per output strutturati e veloci di qualità media-alta. Tende a essere autorevole, sicuro, "patinato".

**Claude** (Anthropic). Il più riflessivo e "letterario". Eccellente per testi che richiedono sfumatura, ragionamento articolato, scrittura che respira. Tende a essere più cauto: più spesso ti dice "non lo so" quando non sa.

**Gemini** (Google). Il più ancorato ai dati. Forte quando deve cercare informazioni aggiornate sul web (ha l'integrazione nativa con la ricerca Google) o ragionare su grandi quantità di testo in input.

**DeepSeek** (cinese). Sorprendente outsider. Ottimo su problemi di logica, matematica e programmazione, talvolta in modo che lascia di stucco.

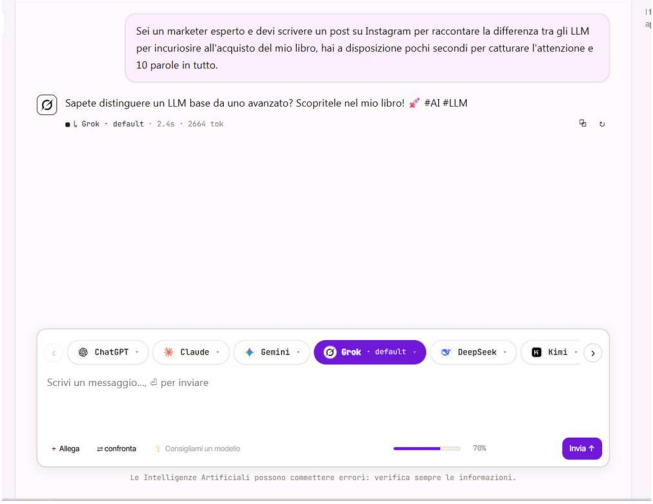
E poi nel ChatBox trovi anche **Grok**, **Mistral**, **Llama**, **Qwen**, **Kimi**. Per cominciare, i quattro sopra sono lo zoccolo duro.

### Vedere le differenze con gli occhi

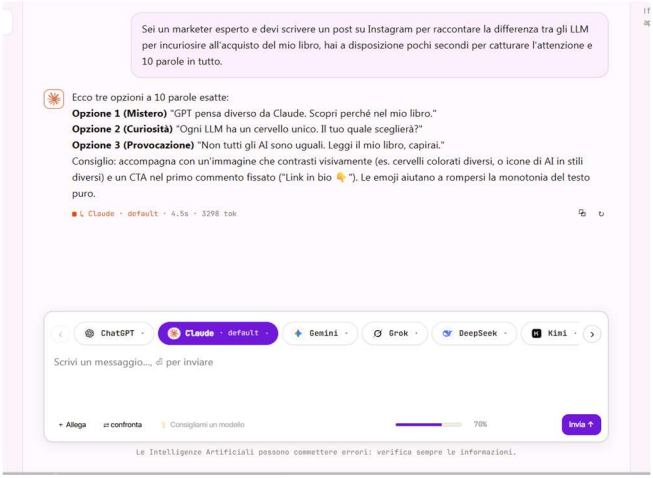
Per vedere le differenze sul serio, devi lanciare *lo stesso identico prompt* su tre modelli diversi, **ognuno con contesto azzerato**. Lanciamo questo:

```
Sei un marketer esperto e devi scrivere
un post su Instagram per raccontare la
differenza tra gli LLM in occasione
```

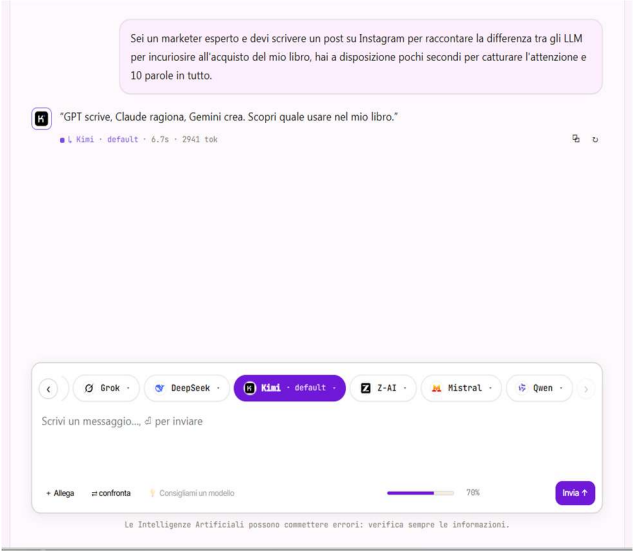
```
dell'acquisto del mio libro, hai pochi
secondi per catturare l'attenzione e 10
parole in tutto.
```




*Stesso prompt, primo modello: il post arriva sintetico, già rifinito*




*Stesso prompt, secondo modello: la struttura cambia, il tono pure*




*Stesso prompt, terzo modello: ancora un'angolazione diversa, ancora un ritmo diverso*

Stesso prompt, copiato e incollato in **tre conversazioni separate** (così che ogni modello parta da zero, senza vedere le risposte degli altri). Tre risposte. Diverse nella struttura, nel ritmo, nelle parole, nel modo di chiudere. Nessuna è oggettivamente "migliore": sono tre angolazioni dello stesso compito.

> **Non devi scegliere "il modello migliore". Devi scegliere il modello migliore *per quel singolo compito*.**

## Il concetto di contesto (e perché serve il chain editor per i confronti puliti)

Una cosa fondamentale che devi avere in testa, e che torna in ogni capitolo del libro: ogni modello, quando ti risponde,

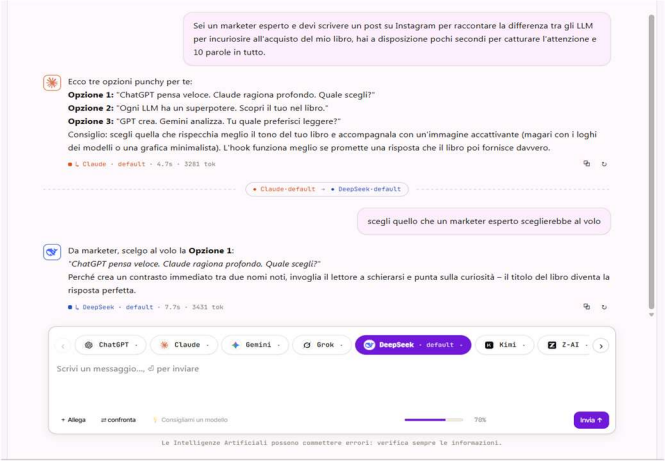
*vede tutto il contesto della conversazione*. Cioè vede il tuo prompt corrente *più* tutti i messaggi e tutte le risposte precedenti nella stessa chat. Se in mezzo c'era stata anche la risposta di un altro modello, vede anche quella.

Quindi se vuoi sapere *davvero* come ChatGPT, Claude e Gemini rispondono in modo indipendente, non basta cambiare modello a metà chat: devi ricominciare da zero per ognuno. Nel ChatBox lo fai con il pulsante *"+ Nuova conversazione"*.

Per i confronti rigorosi, Canonity ha uno strumento dedicato: il **prompt chain editor**. Lì costruisci una mini-catena di tre nodi paralleli (stesso prompt, tre modelli, tre contesti indipendenti) e li lanci tutti con un click. Le tre risposte arrivano affiancate, ognuna pulita. Questo modo di lavorare si chiama **chain prompting**: lo vediamo per intero nel capitolo *Chain prompts*, ma intanto è bene che tu sappia che il chain editor di Canonity esiste e che è la strada giusta per il confronto fra modelli.

### Cambiare modello a metà conversazione (caso d'uso opposto)

C'è però un caso opposto, dove il *contesto cumulativo* del ChatBox è esattamente quello che vuoi: quando usi modelli diversi a *passaggi diversi* dello stesso lavoro. Per esempio, fai brainstorming con Claude (forte sulle idee sfumate), poi cambi a DeepSeek e gli chiedi *"scegli quale di queste tre idee sceglierebbe un marketer esperto e dimmi perché"*. Il secondo modello *deve* vedere quello che ha scritto il primo: è il punto.




*Cambio di modello durante la stessa conversazione: prima Claude per generare opzioni, poi DeepSeek per la scelta finale ragionata*

Tutto nello stesso thread, con la cronologia preservata. Il secondo modello vede il primo, ci ragiona sopra, e produce una risposta che dipende da quel contesto.

Sono due modi di lavorare opposti, entrambi utili:

- **Confronto puro fra modelli** → contesti separati, conversazioni distinte nel ChatBox o (meglio) chain editor.
- **Lavoro a staffetta fra modelli** → stessa conversazione nel ChatBox, contesto cumulativo, ogni modello vede il pregresso.

Aver capito questo punto ti porta tre passi avanti rispetto al 95% degli utenti di AI. Quasi tutti, oggi, usano un solo modello, quello che hanno installato sul telefono. Tu, da qui in poi, puoi ragionare per *consigli di squadra*: chi è il miglior brainstormer? Chi il miglior revisore? Chi il miglior

matematico? Ogni modello prende un ruolo, e tu sei il regista.

## L'episodio Avianca, e perché ti riguarda

A maggio 2023, un avvocato di New York, Steven Schwartz, depositò una memoria in cui citava sei sentenze. Tutte e sei erano state generate da ChatGPT. Tutte e sei erano *inesistenti*. Il giudice se ne accorse. L'avvocato fu sanzionato.

Quando gli chiesero perché non avesse verificato, Schwartz disse: *"ho chiesto a ChatGPT se le sentenze fossero reali, e ChatGPT mi ha detto di sì"*. Lui aveva chiesto al ladro se stesse rubando, e il ladro aveva risposto di no.

*Mai*, dico mai, fidarti del modello come unica fonte per un fatto verificabile. *Mai* chiedere al modello *"sei sicuro?"* come se fosse una verifica. La verifica deve venire da fuori.

## Le cinque mosse pratiche per non cascarci

**Prima: concedi all'AI il diritto di non sapere.** La singola riga di prompt che fa la maggiore differenza in assoluto:

```
Se non sei sicuro di un'informazione,
scrivi "non lo so" invece di inventare.
Per i fatti specifici (numeri,
citazioni, date, fonti), indica sempre
il tuo livello di sicurezza: alto,
medio, basso.
```

Aggiungere questa formula a ogni prompt che chiede informazioni puntuali riduce le allucinazioni di un fattore enorme.

**Seconda: vincola la risposta alle fonti che gli dai tu.** Se dai *tu* la fonte al modello (caricandola come PDF nel

ChatBox) e gli chiedi di lavorare *dentro quella fonte*, le allucinazioni cadono drasticamente. In gergo si chiama *grounding*.

**Terza: chiedi le fonti, e poi verificale.** Per ricerche dove non hai un documento da fornire, chiedere le fonti rimane utile, con un'avvertenza: *il modello può inventare anche le fonti*. Però se chiedi un URL, o vai a vedere e funziona (hai vinto), o trovi una 404 (hai capito che era inventato).

**Quarta: dividi il compito in passaggi.** Spezza il compito in due turni del ChatBox: primo turno *"Elencami i 5 fatti più importanti su X. Per ognuno, quanto sei sicuro"*. Tu verifichi i fatti "medio" e "basso". Secondo turno: *"Adesso scrivi il testo usando questi 5 fatti corretti"*.

**Quinta: non chiedere mai numeri che non puoi verificare.** Per qualunque numero specifico che devi usare in un testo pubblicato, *non chiedere il numero all'AI*. Chiedi all'AI *dove trovare il numero*. Cioè: la fonte autorevole.

## La regola d'oro

Lorenzo, alla fine di tutto, mi disse una sera: *"adesso tratto ChatGPT come tratto un praticante particolarmente bravo ma sicuro di sé. Mi è utilissimo, ma niente che esce da lui va in atto senza che io l'abbia letto, controllato, e firmato con la mia mano"*.

> **Regola d'oro**: tratta ogni risposta dell'AI come la prima bozza di uno stagista bravo ma sicuro di sé. Utile per partire, mai pubblicabile senza una verifica umana sui fatti.

Stampala e attaccala al monitor.

## Vademecum del capitolo 3

### Cos'è un LLM, in una riga

Un programma che prevede la parola successiva più probabile data una sequenza di parole. Niente coscienza, niente verifica, niente intenzioni. Solo statistica su scala enorme.

### I tre limiti da ricordare sempre

| Limite | Cosa significa | Conseguenza pratica |
|---|---|---|
| **Non sa il presente** | Sa fino al training cutoff (a meno di ricerca live) | Per dati recenti, verifica su fonte aggiornata |
| **Non sa te** | Non conosce te, la tua azienda, il tuo contesto | Ogni prompt deve contenere il contesto |
| **Non sa la verità** | Distingue "vero" da "plausibile" solo per i fatti molto noti | Per fatti specifici, *sempre* verifica esterna |

### Le cinque strategie per ridurre le allucinazioni

| # | Strategia | Frase tipo |
|---|---|---|
| 1 | Diritto di non sapere | *"Se non sei sicuro, scrivi 'non lo so'."* |
| 2 | Grounding sulle fonti | *"Rispondi solo sulla base del documento qui sotto."* |
| 3 | Fonti verificabili | *"Per ogni affermazione, dammi un link a una fonte ufficiale."* |
| 4 | Compito spezzato in passaggi | Prima i fatti con livello di sicurezza, poi la stesura |
| 5 | Mai numeri non | Chiedi la *fonte* del numero, non il |

verificabili numero

### Le tre frasi da non scrivere mai più

- *"Sei sicuro?"* — non è una verifica.
- *"Confermami che questi dati sono giusti."* — idem.
- *"È tutto verificato?"* — idem.

La verifica viene da *fuori*.

> **La regola d'oro**
>
> Tratta ogni risposta dell'AI come la prima bozza di uno stagista bravo ma sicuro di sé. Utile per partire, mai pubblicabile senza una verifica umana sui fatti.

### Prova questo in Canonity

**Esercizio. Sperimenta un'allucinazione (e poi prevenila) con tre modelli indipendenti.**

Pensa a un dato specifico della tua zona di interesse che *sai essere vero* (un articolo di legge, una statistica di settore, una citazione di un libro che conosci bene, il nome di un dirigente di un'azienda specifica).

1. Apri **canonity.com/chat**. Premi *"+ Nuova conversazione"*. Seleziona **ChatGPT**. **Senza nessuna formula protettiva**, chiedi al modello quel dato. Premi invio. Salva la risposta. Quanto è plausibile? Quanto è vera?
2. Premi di nuovo *"+ Nuova conversazione"* (così il contesto riparte da zero). Seleziona **Claude**. Fai la stessa, identica domanda. Salva.
3. Premi *"+ Nuova conversazione"*. Seleziona **Gemini**. Stessa domanda. Salva.

4. Confronta le tre risposte. Hai trovato un'allucinazione? Hai trovato modelli più cauti di altri? Quale ti ha detto "non lo so" e quale ti ha inventato qualcosa?
5. **Adesso aggiungi la formula magica.** In una *quarta* nuova conversazione, rifai la domanda con: *"Se non sei sicuro di un'informazione specifica, scrivi 'non lo so' invece di inventare. Indica il tuo livello di sicurezza: alto, medio, basso"*. Premi invio. Confronta con i tre precedenti.

Questa è la riga di prompt che, per tutta la tua vita professionale, dovrebbe accompagnare ogni richiesta di un fatto preciso.

> **Variante per i più curiosi**: lo stesso confronto a tre modelli si può costruire in modo molto più elegante nel **prompt chain editor** di Canonity (canonity.com/chains), con una piccola chain a tre nodi paralleli sullo stesso prompt. Lo vediamo per intero nel capitolo *Chain prompts*.

## Prima del prossimo capitolo

Nel prossimo capitolo passiamo alle tecniche avanzate: few-shot, prompting di sistema, step-back, chain of thought, self-critique. Sono nomi che fanno paura ma sono tecniche semplicissime, una volta capite. Ognuna risolve un problema specifico, e tutte sfruttano quello che hai imparato in questo capitolo.

# Capitolo 4. Prompt avanzati

## Ovvero: cinque trucchi che, una volta presi, non si dimenticano più

Federica ha trentotto anni e fa la copywriter freelance da quasi dieci. Vive a Roma, in un appartamento di tre stanze a Pigneto con una libreria che occupa due pareti e un gatto enorme che si chiama Tito. Ha dodici clienti, divisi fra moda, food, hospitality, B2B SaaS, beauty.

Federica usa il ChatBox di Canonity da circa quattro mesi. La transizione dalla domanda secca alla *consegna* l'ha fatta presto. I suoi prompt sono diventati buoni: ruolo, contesto, tono, vincoli, formato. I suoi risultati sono passati da mediocri a discreti. Discreti significa: usabili dopo una sistemata di un quarto d'ora, alcune frasi suonano da AI.

Quattro mesi dentro Canonity, Federica si trova davanti a un plateau. *"Sembro arrivata"*, mi disse una sera al telefono. *"I prompt me li costruisco bene. Ma sento che c'è un altro gradino sopra che non sto facendo."*.

Quel gradino sopra esiste, ed è quello di cui parla questo capitolo. Si fa con cinque tecniche, ognuna semplicissima da capire, ognuna che ti porta da un risultato "discreto" a uno "professionale". Si chiamano con nomi che fanno paura (*few-shot prompting*, *prompting di sistema*, *step-back prompting*, *chain of thought*, *self-critique prompting*), ma sono cinque trucchi, non cinque scienze.

### 1. Few-shot prompting: insegnare con esempi

Federica stava finendo dieci post per un cliente di moda, un piccolo marchio torinese di streetwear di lusso. Il problema con questo cliente è che ha un tono di voce

particolare: frasi cortissime, riferimenti pop senza spiegazioni, un'ironia secca tipo *"questo l'avete capito al volo o no?"*. Quando faceva fare i post al ChatBox, anche con una consegna impeccabile, il modello tirava fuori sempre cose tipo:

```
Scopri il nuovo modello: una giacca
pensata per chi non si accontenta.
Comfort, stile, eleganza moderna.
```

Esattamente l'opposto di quello che voleva il cliente.

Una sera le venne in mente: *se descriverlo a parole non funziona, perché non gli incollo direttamente alcuni esempi del tono?*. Aprì il ChatBox di Canonity, scelse Claude come modello (è il più bravo a replicare stile), e scrisse:

```
Sei un copywriter per un brand
streetwear torinese. Il tono è quello
dei seguenti tre post: replicalo, non
descriverlo.

Esempio 1: "Sì, ne abbiamo solo dieci.
Sì, costano. No, non spieghiamo
perché."

Esempio 2: "Quella che vedete in foto
la porta un mio amico che fa il
barista. Lo paghiamo in giacche, non in
soldi."

Esempio 3: "Da fuori sembra una felpa.
Da dentro non lo è. Indossatela e ci
capirete."

Adesso scrivimi cinque post per il
nuovo modello: un giubbotto bomber
verde militare con interno bianco. Una
sola riga ciascuno, stesso registro.
```

Il ChatBox le restituì:

1. Tre tagli, sei tasche, un peso che si sente. Funziona.
2. Era nel piano "verde militare" da due anni. Mai più senza.
3. Il bianco dentro non è una distrazione. È un promemoria.
4. C'è un negozio di fianco al nostro che vende lo stesso colore. Non chiamatelo bomber.
5. Si chiama "Echo 03". L'Echo 02 lo cercate ancora su eBay.

Federica sorrise. Tre dei cinque andarono live sul profilo del cliente la settimana dopo. Engagement +60% rispetto alla media.

Federica ha usato il **few-shot prompting**: invece di descrivere a parole il risultato che vuoi, *mostra qualche esempio*. Gli LLM sono macchine di riconoscimento di pattern: un esempio concreto vale più di cento aggettivi.

### Una variante potente: few-shot con contrasti

Mostra esempi positivi *e* esempi negativi affiancati. Il contrasto insegna al modello la sfumatura.

```
Voglio post ironici. NON voglio post
sarcastici.

Ironico (lo voglio): "Sì, è vero,
l'abbiamo davvero fatta in pelle
vegetale. Per noi era ovvio, per i
nostri concorrenti meno."

Sarcastico (NON lo voglio): "Ah, certo,
perché tu sì che capisci la moda.
Continua così."
```

```
Adesso scrivimi cinque post nel tono
ironico.
```

**Quando usarlo**: compiti soggettivi, stilistici, di brand. Non serve per compiti meccanici dove la descrizione a parole basta.

## 2. Prompting di sistema: la Costituzione della conversazione

Quando Federica scrive a un nuovo modello, ogni volta deve dirgli: *"sei un copywriter per un brand torinese di streetwear, parli in italiano, niente emoji, tono ironico…"*. Per ogni cliente, ogni volta, lo stesso bisogno. Le sembrava di sprecare cinque minuti all'inizio di ogni sessione.

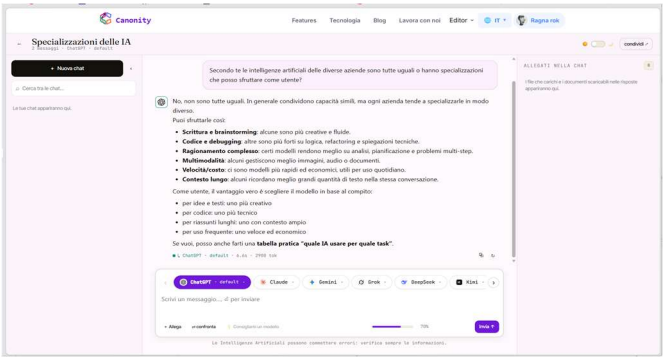
Il **prompting di sistema** è il livello più alto. Definisce *chi è* il modello, *come deve comportarsi sempre*, e *cosa non deve mai fare*. È una specie di costituzione della conversazione. Una volta scritta, vale per tutto quello che viene dopo, anche se la conversazione dura due ore e cento messaggi.

Nel ChatBox di Canonity, in cima alla conversazione c'è un campo dedicato per il "system prompt" (lo trovi cliccando sull'ingranaggio della conversazione, o impostandolo quando crei un nuovo thread). Quello che ci scrivi non lo vede l'utente, lo vede solo il modello.

Federica, per il cliente streetwear torinese, ha creato una conversazione fissa chiamata *"Cliente Echo"* e ha scritto, una volta sola, in cima:

```
Sei un copywriter senior per un brand
di abbigliamento streetwear torinese.
Tono: ironico mai sarcastico, frasi
corte (max 12 parole), niente emoji,
niente esclamativi, niente domande
retoriche. Non usare mai parole come
```

```
"scopri", "eleganza", "comfort",
"esclusività". Quando ti chiedo dieci
post, dammi dieci post diversi fra loro
per angolo di attacco.
```




*Il ChatBox di Canonity in vista panoramica: la riga di selezione modelli in basso, l'area di conversazione, le impostazioni della conversazione.*

Da quel momento, ogni volta che Federica apre quella conversazione e chiede un post, il modello sa già come comportarsi. Niente più "ricomincio da capo" ogni mattina.

C'è anche il **prompting di ruolo** (*"Sei Andrea, un copywriter senior di trentadue anni che odia il copywriting da catalogo…"*) e il **prompting contestuale** (cosa sta succedendo *adesso*, in questa specifica richiesta).

I tre si combinano. Sistema = la regola di base, raramente cambia. Ruolo = il "carattere", può variare per cliente. Contesto = la consegna specifica del giorno.

Federica ha costruito *una conversazione per ogni cliente* nel ChatBox. Ogni conversazione ha un system prompt con la "Costituzione" + il ruolo. Quando entra in una conversazione, scrive solo il contesto del giorno.

Risparmio: dieci minuti a sessione, dieci sessioni a settimana, cento minuti a settimana.

### 3. Step-back prompting: fare un passo indietro

Era un giovedì, il cliente del food chiamò Federica con un'urgenza. *"Federica, ci hanno cambiato la normativa europea sulla etichettatura degli allergeni. Devo aggiornare quattro pagine del sito entro lunedì."*. Federica prese il regolamento, lo caricò nel ChatBox come PDF, scelse Claude come modello, e lanciò il prompt:

```
Dato il regolamento UE in allegato,
riassumi cosa cambia per un produttore
italiano di pasta senza glutine che
vende in Italia e in Germania.
```

Il modello tirò fuori una mezza pagina di testo che suonava bene ma non era *operativa*. Sembrava un articolo divulgativo. Mancavano le cose pratiche.

Le venne in mente lo **step-back prompting**: prima di rispondere alla domanda specifica, si chiede al modello di fare un passo indietro e ragionare sul *principio generale*. Poi gli si chiede di applicarlo al caso specifico. Riformulò:

```
Prima di rispondere alla mia domanda,
fai un passo indietro: qual è la
struttura tipica di un regolamento UE
sull'etichettatura? Quali sono le aree
(informazioni obbligatorie /
facoltative / divieti / sanzioni) che
vengono affrontate sempre?

Poi, una volta che abbiamo questa
struttura, applicala al regolamento
allegato e dimmi, per ciascuna area,
cosa cambia per il mio cliente.
```

Stavolta il modello rispose in modo strutturato. Prima un paragrafo sulla struttura generale dei regolamenti UE. Poi, area per area, cosa cambiava nello specifico. Una checklist operativa di sei punti, ognuno con un riferimento esplicito al regolamento.

### Perché funziona

I casi specifici, soprattutto quelli rari o tecnici, "scivolano" nello spazio statistico del modello, perché ne ha visti pochi. Le strutture *generali*, invece, le ha viste milioni di volte. Se chiedi al modello di posarsi prima su un principio generale, lui "atterra" in un'area della sua memoria più popolata, più sicura. Poi applicare al caso specifico è più facile.

**Quando usarlo**: per problemi tecnici dove il caso specifico è poco frequente e la struttura generale è ben conosciuta (regolamenti, leggi, materie scientifiche).

## 4. Chain of Thought: pensa a voce alta

Federica doveva calcolare il ROI di una campagna social. Budget 4.200 €, contatti 180, conversione 12%, margine medio 240 €. Aprì il ChatBox e chiese:

```
Budget: 4.200 €. Contatti: 180.
Conversione: 12%. Margine medio: 240 €.
Qual è il ROI?
```

Il modello rispose: *"Il ROI è del 28%."*. Federica si fermò. Fece il calcolo a mente: 180 × 12% = 21,6 clienti. Ricavi: 21,6 × 240 = 5.184 €. Profitto: 5.184 − 4.200 = 984 €. ROI = 984/4.200 = 23,4%. **Non 28%.**

Le venne in mente la **Chain of Thought**, o *catena di pensiero*: aggiungere una sola riga al prompt che obbliga il modello a esplicitare il ragionamento.

```
Budget: 4.200 €. Contatti: 180.
Conversione: 12%. Margine medio: 240 €.
Qual è il ROI? Ragiona passo per passo,
scrivi ogni calcolo, e solo alla fine
dai il risultato finale.
```

Il modello, stavolta:

```
Passo 1. Clienti acquisiti: 180 × 12% =
21,6. Passo 2. Ricavi attesi: 21,6 ×
240 € = 5.184 €. Passo 3. Profitto
netto: 5.184 − 4.200 = 984 €. Passo 4.
ROI: 984 / 4.200 = 0,234 = 23,4%.
```

Risultato corretto. Stesso modello, stesso prompt nei dati, ma con una riga in più ha dato la risposta giusta e, soprattutto, *verificabile* passo per passo.

### Perché funziona

Ognuno dei passaggi che il modello scrive diventa *contesto* per il successivo. Gli errori non si accumulano silenziosamente: si espongono. Non è che il modello "pensi di più". È che, scrivendo a voce alta, gli errori diventano impossibili da nascondere.

**Quando usarlo**: compiti matematici, logici, di pianificazione. Sempre. Variante utile: *"prima di rispondere, ragiona dentro tag `<pensiero></pensiero>`. Poi dai la risposta finale fuori dai tag"*.

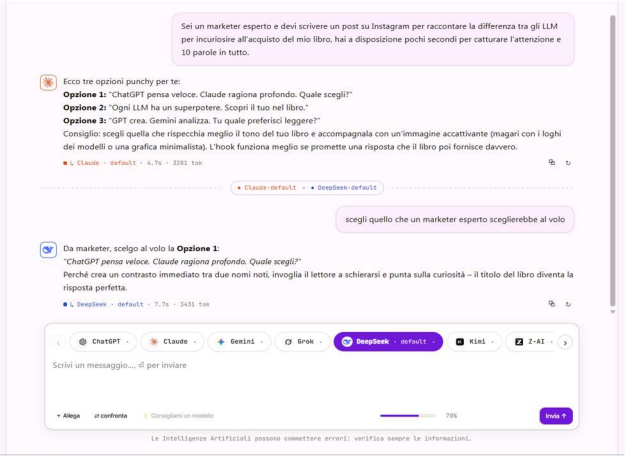
## 5. Self-critique prompting: far auto-correggere l'AI

Un lunedì mattina, un cliente del settore hospitality (un agriturismo siciliano vicino Ragusa) inviò a Federica una sua proposta di brochure scritta a quattro mani con ChatGPT. Era una mezza catastrofe: belle frasi mescolate a luoghi comuni, struttura sbilanciata, due ripetizioni

grossolane, e una promessa commerciale (*"siamo l'unico agriturismo siciliano biologico certificato"*) oggettivamente falsa.

Le venne in mente la **self-critique prompting**. Funziona così: dopo la prima risposta del modello, gli chiedi *di criticarsi*. Letteralmente: di rileggere quello che ha appena scritto e di trovare i problemi. Poi gli chiedi di riscriverlo.

C'è una versione *ancora* più potente: chiedere la critica a un *secondo modello*, diverso dal primo. Quello che ha generato non vede i propri bias. Nel ChatBox di Canonity questo è banale: cambi modello con un click e mantieni la cronologia.




*Cambio di modello nel ChatBox per far criticare a un secondo modello quello che il primo ha scritto: ChatGPT genera, Claude rilegge e critica*

Federica fece così: nel ChatBox, primo turno con ChatGPT per generare una versione corretta della brochure. Secondo turno, cambia modello a **Claude** dalla riga di selezione, e scrive:

```
Sei un copywriter senior molto esperto
e severo. Rileggi questa brochure di un
agriturismo siciliano scritta da un'AI.
Fammi una lista spietata di 8 problemi:
cose deboli, errori, scivoloni,
ripetizioni, claim sospetti. Per ognuno
spiegami perché è un problema e come si
correggerebbe.
```

Claude vede tutto il thread precedente (la brochure scritta da ChatGPT è già lì) e in trenta secondi sciorinò una lista perfetta: i due luoghi comuni, le ripetizioni, il claim falso (l'agriturismo *non era* l'unico siciliano bio certificato), un terzo paragrafo che dilatava il discorso senza aggiungere niente.

Terzo turno (puoi rimanere su Claude o tornare su ChatGPT):

```
Adesso riscrivi la brochure correggendo
gli 8 punti che hai appena segnalato.
Mantieni la lunghezza, il tono caldo e
accogliente, ma elimina tutti i
problemi.
```

Federica la mandò al cliente: *"È incredibile come è migliorata, cosa hai fatto?"*, rispose lui. *Cosa aveva fatto*: aveva fatto criticare il modello da un modello diverso, in tre click.

### Perché funziona

Generare un testo e *valutare* un testo sono due compiti diversi. Quando il modello deve generare, cerca la continuazione plausibile più fluida. Quando deve criticare, cerca i punti deboli: usa altre "leve" dentro di sé. E un modello diverso ha bias diversi, quindi vede cose che il primo non vedeva.

> Questa idea (*modelli diversi per fasi diverse*) è il cuore del framework **TASS** che vedremo nel capitolo *TASS*. La self-critique è la versione "leggera" e a due passaggi di TASS.

**Quando usarlo**: **sempre, prima di consegnare un testo a un destinatario reale.** Cliente, capo, lettore. Come filtro anti-allucinazione: nella self-critique chiedi *"verifica che ogni dato citato sia plausibile e indica quelli su cui hai dubbi"*. Spesso il modello segnala da solo le sue allucinazioni del turno precedente.

## Mettere le cinque tecniche insieme

Federica, dopo qualche settimana, aveva un riflesso: per ogni progetto si chiedeva *quale di queste cinque tecniche è quella giusta?*. Spesso ne usava una sola. Talvolta ne combinava due o tre.

Per un post di brand: few-shot + system prompt fisso per cliente nel ChatBox. Per un articolo tecnico: step-back + chain of thought. Per qualunque cosa che dovesse passare al cliente: self-critique come ultimo passaggio con cambio di modello, sempre.

Tre mesi dopo: *"Adesso lavoro al doppio della velocità di prima, con la metà della fatica. L'unica cosa che ho cambiato sono cinque trucchi. Non si direbbe."*.

## Vademecum del capitolo 4

### Le cinque tecniche, in una tabella

| Tecnica | Cosa fa | Quando usarla |
|---|---|---|
| **Few-shot** | Mostra esempi del risultato invece di | Compiti soggettivi, |

| | descriverlo | stilistici, di brand |
|---|---|---|
| **Sistema/Ruolo** | Regole permanenti per tutta la conversazione | Lavori ricorrenti su clienti/progetti fissi |
| **Step-back** | Passa dal principio generale al caso specifico | Problemi tecnici, regolamenti, casi rari |
| **Chain of Thought** | Esplicita il ragionamento passo per passo | Calcoli, logica, pianificazione |
| **Self-critique** | Il modello (o un secondo modello) critica quello che ha scritto | Prima di consegnare un testo a qualcuno |

### Quale scegliere: il flusso decisionale

- Devi scrivere qualcosa **stilistico**? → Few-shot.
- Lavori su un cliente **ricorrente**? → Sistema (una volta sola, per sempre).
- Problema **tecnico o raro**? → Step-back.
- **Calcolo, logica, pianificazione**? → Chain of Thought.
- **Consegna a un destinatario**? → Self-critique con cambio di modello, sempre.

### Tre principi che non cambieranno mai

1. **Mostrare vince su descrivere.** Per cose soggettive, gli esempi battono gli aggettivi 10 a 1.
2. **Le regole stabili stanno nel sistema, non nei singoli prompt.**
3. **Pensare a voce alta riduce gli errori.**

## Prova questo in Canonity

**Esercizio. Self-critique con cambio di modello.**

1. Apri **canonity.com/chat**. Lascia ChatGPT come modello.
2. Prendi un testo che hai consegnato ieri o l'altro ieri (una mail, una bozza, un post). Incollalo nel ChatBox e chiedi: *"Riscrivi questo testo migliorandone la chiarezza, il ritmo e la specificità"*. Premi invio. Salva la prima versione riscritta.
3. **Cambia modello** dal selettore in basso (passa a **Claude**). Nello stesso thread scrivi: *"Sei un revisore senior molto esperto. Rileggi il testo che ChatGPT ha appena prodotto e fammi una lista spietata di 5 problemi. Per ognuno spiegami perché è un problema e come si correggerebbe."*. Premi invio.
4. **Nello stesso thread, ancora su Claude**: *"Adesso riscrivi il testo correggendo gli 8 punti che hai appena segnalato."*. Premi invio.
5. Confronta la versione finale con quella iniziale (quella che avevi davvero consegnato). Quanto ti pesa, ora, di non averlo fatto prima di mandarla?

Hai appena fatto self-critique con due modelli diversi. Sette minuti di lavoro, una qualità di un altro livello.

## Prima del prossimo capitolo

Nel prossimo capitolo usciamo dal testo. Vedremo cosa succede quando metti dentro un prompt non solo parole, ma immagini, PDF, audio. Si chiama *prompting multimodale*, e nel ChatBox di Canonity si fa con un click sul pulsante "Allega" sotto il campo di input.

# Capitolo 5. Multimodale

## Ovvero: quando il prompt non è più solo parole

Renato ha cinquantun anni, è commercialista, e ha lo studio al Vomero. Lo studio si chiama "Renato Aiello Commercialista", con quattro collaboratrici, una segretaria. Ogni mese gli passano per le mani decine, a volte centinaia di scontrini, ricevute, contratti, fatture, estratti conto.

Renato non ama le AI. Lo dice a chiunque. *"Senta, io faccio il commercialista da trent'anni. Ho visto arrivare Excel, ho visto arrivare i gestionali in cloud. Adesso questa cosa nuova del ChatGPT… non ce la faccio."*. E lo dice sorridendo, perché sa che il discorso suona vecchio, ma lo dice davvero.

Una sera sua nipote Sofia (ventidue anni, studia economia a Roma) gli porta da provare il ChatBox di Canonity. *"Ha una funzione carina, zio: ci puoi caricare i documenti, gli scontrini, le foto. E lui ti estrae i dati"*. Sofia prende uno scontrino di una cena (pizza per quattro al Pignasecca, settantadue euro e cinquanta), lo fotografa col telefono, va su **canonity.com/chat** sul portatile di Renato, trascina la foto nell'area degli allegati (il pulsante *"Allega"* a sinistra del campo di input), e scrive:

```
Estrai da questo scontrino i dati
seguenti, in formato tabella: nome del
locale, indirizzo se visibile, data,
ora, lista delle voci con prezzo
unitario, totale, metodo di pagamento.
Se un dato non è leggibile, scrivi "non
leggibile".
```

In quattro secondi, il ChatBox restituisce una tabella con tutti i dati. Margherita, marinara, capricciosa, calzone, sei

birre, quattro caffè. Totale: settantadue e cinquanta. Pagato in contanti.

Renato si toglie gli occhiali. Guarda Sofia. *"Aspetta. Questo lo ha letto dalla foto?"*. Sofia annuisce. *"Sì zio, dalla foto. E può leggere anche foto storte, foto sbiadite, foto fatte al volo dai clienti. Non sempre perfetto: devi sempre controllare i numeri prima di usarli. Ma il novanta per cento del lavoro te lo fa lui."*

Renato pensa a quante volte, in trent'anni, ha avuto sotto gli occhi una pila di scontrini portati dai clienti, e una delle collaboratrici li ha trascritti a mano. Un lavoro che porta via giorni. *Giorni*.

Il giorno dopo dice alla sua collaboratrice più anziana: *"Pina, mi sa che dobbiamo cambiare un po' di cose"*.

Questo capitolo parla esattamente di quello che ha scoperto Renato quella sera. Si chiama **prompting multimodale**, e per molti dei lavori del lettore di questo libro è la cosa più sottoutilizzata di tutta l'AI generativa.

## Cosa significa "multimodale"

Modalità, in italiano, è un modo. Quando un'AI è *multimodale*, significa che capisce più di una modalità in ingresso: non solo testo, ma anche immagini, audio, video, file PDF.

Per noi che scriviamo prompt cambia una cosa fondamentale: **non sei più costretto a descrivere a parole**. Puoi *mostrare*. Invece di descrivere lo scontrino, lo fotografi. Invece di trascrivere il contratto, lo scansioni. Invece di riassumere a parole una riunione di un'ora, carichi il file audio.

Nel ChatBox di Canonity, l'icona *"Allega"* sotto il campo di input accetta: **immagini** (foto, scansioni, schermate), **PDF** (contratti, fatture, report, articoli), **audio** (registrazioni di riunioni, messaggi vocali), **video** brevi.

## Le quattro cose principali che puoi fare con materiale non testuale

### 1. Estrarre dati strutturati da un'immagine

Renato, una settimana dopo, riceve da un cliente una foto della sua scrivania su cui ci sono nove fogli sparpagliati: scontrini di benzina, una ricevuta del meccanico, qualche bolletta, una fattura. Apre il ChatBox di Canonity, carica la foto e prova prima un prompt generico: *"Cosa vedi in questa foto?"*. Risposta: *"Vedo una scrivania con vari documenti…"*. Corretta, ma inutile.

Riprova con una consegna fatta:

```
Sei un commercialista esperto. Guarda
questa foto di documenti sparsi. Per
ognuno: 1) classificalo (scontrino,
ricevuta, fattura, bolletta), 2) leggi
il fornitore, 3) leggi l'importo
totale, 4) leggi la data. Restituisci
tutto in una tabella ordinata per data
crescente. Se un documento non è
leggibile, scrivi "non leggibile".
```

Risposta:

| # | Tipo | Fornitore | Importo | Data |
|---|---|---|---|---|
| 1 | Scontrino | Esso (carburante) | 64,80 € | 03/05/2026 |
| 2 | Ricevuta | Autoricambi Aiello | 240,00 € | 05/05/2026 |
| 3 | Bolletta | ENEL Energia | 187,42 € | 06/05/2026 |
| 4 | Fattura | Forneria Esposito | 92,00 € | 08/05/2026 |
| 5 | Scontrino | non leggibile | non leggib. | non leggib. |

Renato fissa lo schermo. Quella stessa pila di nove documenti, prima della conversione, sarebbe diventata mezz'ora di inserimento manuale nel gestionale. Adesso sono trenta secondi.

L'AI multimodale è bravissima a fare *estrazione di dati strutturati*. Il prompt fa la differenza: *"Cosa vedi?"* ti dà una descrizione generica. *"Estrai questi dati specifici nella forma tabellare che specifico"* ti dà una tabella usabile.

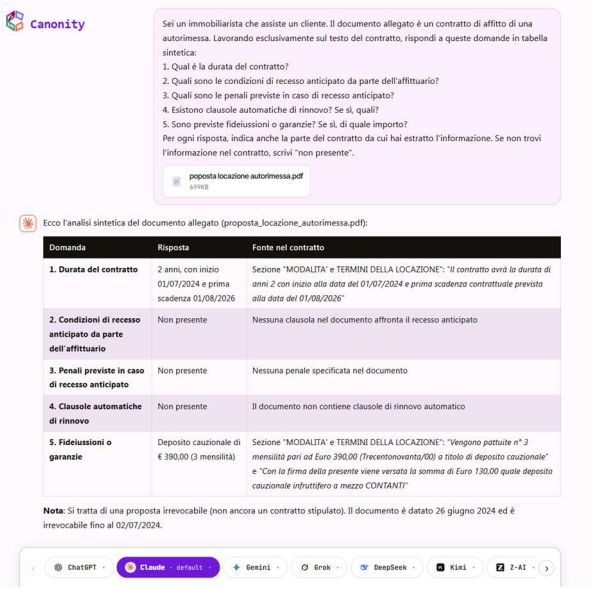
### 2. Riassumere o analizzare un documento lungo

Una settimana dopo, Renato riceve da un cliente immobiliarista un contratto di affitto commerciale di trentacinque pagine. Il cliente vuole sapere una cosa sola: *"se devo recedere, quanto mi costa?"*.

Renato carica il PDF nel ChatBox, sceglie Claude come modello dal selettore in basso (dal capitolo *Cos'è un LLM* sa che è il più adatto a testi lunghi e ragionati), e scrive:

```
Sei un commercialista che assiste un
cliente immobiliarista. Il documento
allegato è un contratto di affitto
commerciale di 35 pagine. Lavorando
esclusivamente sul testo del contratto,
rispondi a queste domande in tabella
sintetica:
1. Qual è la durata del contratto?
2. Quali sono le condizioni di recesso
   anticipato?
3. Quali sono le penali previste?
4. Esistono clausole automatiche di
   rinnovo?
5. Sono previste fideiussioni o garanzie?
   Di quale importo?
```

```
Per ogni risposta, indica anche la
pagina e il numero di articolo. Se non
trovi l'informazione, scrivi "non
presente".
```




*Il ChatBox di Canonity con il prompt tabellare di Renato sul contratto di affitto: PDF allegato, richiesta in cinque punti, indicazione del formato di risposta in tabella*

Claude lavora qualche secondo in più del solito. Poi restituisce una tabella perfetta, con cinque righe, ognuna con la risposta e il riferimento (es. *"Art. 9, pag. 12: La penale prevista è pari a sei mensilità di canone in caso di recesso prima del terzo anno"*).

Renato chiama il cliente in cinque minuti. Il cliente è incredulo. *"Hai già letto tutto il contratto?"*. Renato sorride. *"Diciamo che ne abbiamo letto le parti che ti interessavano."*

### 3. Analizzare grafici, schemi, dati visivi

Una mattina Renato riceve un cruscotto di Google Analytics di un cliente che ha aperto un piccolo e-commerce. Renato non è un esperto di marketing digitale. Apre il ChatBox, carica la schermata e questa volta sceglie **Gemini** dal selettore modelli (è il più forte sui dati visivi). Chiede di analizzare la schermata: tre numeri da guardare per primi, anomalie, segnali positivi.

Gemini risponde con un'analisi puntata: tasso di abbandono carrello al 78% (*"preoccupante, valore tipico è 60-70%"*), durata media sessione di 3 minuti e 40 secondi (*"sorprendentemente positiva, indica interesse reale"*), una giornata anomala con un picco di traffico (*"da indagare: campagna pubblicitaria? menzione esterna?"*).

Renato gira l'analisi al cliente. Due giorni dopo il cliente scopre che il picco era dovuto a un post di un'influencer che aveva citato il suo negozio. Aveva perso il fenomeno completamente.

Renato non è diventato un esperto di marketing in una mattinata. Ma ha potuto *fare da ponte* fra un cliente che aveva un dato sotto gli occhi e l'AI che il dato lo sa leggere meglio di novanta consulenti su cento.

### 4. Confrontare due o più documenti

L'ultimo caso, forse il più potente, è il confronto. Un cliente di trasporto sta valutando di prendere un camion in leasing, ha ricevuto due offerte di due società diverse (un PDF di otto pagine ciascuno). Renato carica i due PDF nel ChatBox, sceglie Claude, e scrive:

```
Sei un consulente fiscale. Allego due
offerte di leasing per un camion.
```

```
Confrontale su anticipo, canone
mensile, durata, riscatto, costo
complessivo, TAEG, penali per
estinzione anticipata, costi accessori.
Per ogni riga indica quale offerta è
più favorevole (☑ A, ☑ B, =
equivalenti). Alla fine, una riga con
il tuo consiglio sintetico in 30
parole.
```

Il modello restituisce una tabella confronto perfetta, con il consiglio in fondo: *"Offerta B (Mediocredito) è preferibile per costo complessivo (–4.300 €) e penali più morbide, ma richiede anticipo maggiore. Valutare se il cliente preferisce risparmio totale o liquidità iniziale."*.

Tempo totale di Renato: dodici minuti. Senza Canonity, lo stesso lavoro era un'ora e mezza.

### La regola d'oro del multimodale

> Caricare un documento NON sostituisce il prompt. Lo accompagna.

Il documento è il *materiale di lavoro*, il prompt testuale è la *consegna*. Se carichi un PDF e scrivi *"riassumi"*, otterrai un riassunto generico. Se carichi lo stesso PDF e scrivi una consegna ben fatta (ruolo + cosa cercare + formato + vincoli), otterrai un'estrazione mirata, professionale, usabile.

Anche con un allegato nel ChatBox, valgono tutti i sette pezzi della consegna del capitolo *Il prompt*: ruolo, contesto, compito, formato, vincoli, esempi, tono.

## Audio: il fronte più nuovo

Renato, dopo il primo mese, scopre che il ChatBox accetta anche file audio. Registra (con il consenso del cliente) una riunione di quaranta minuti su una pratica di ristrutturazione del debito. Carica il file audio, scrive:

```
Sei l'assistente del commercialista.
Allego l'audio di una riunione di 40
minuti fra me, il cliente Antonio
Russo, e l'avvocato Marco Sansevero.
Argomento: ristrutturazione del debito.
Dall'audio:
1. Elenca i tre interlocutori e quanto è
   intervenuto ognuno (in percentuale).
2. I cinque argomenti principali toccati,
   in ordine.
3. Le decisioni concrete prese (chi fa
   cosa, entro quando).
4. I punti rimasti aperti.
```

Il modello, in due minuti, restituisce un verbale strutturato. Renato lo manda all'avvocato per controllo, l'avvocato risponde con due correzioni minori, e in mezz'ora hanno il verbale ufficiale pronto da archiviare.

## Tre attenzioni che non puoi mai abbassare

**Prima: gli errori di lettura sui numeri.** Quando il modello legge dati da una foto può sbagliare cifre, virgole, decimali. Per qualunque numero che usi in un atto, una dichiarazione, una memoria, *verifica sul documento originale*. La verifica resta tua.

**Seconda: i dati sensibili.** Caricare un PDF con dati personali, sanitari, fiscali o aziendali riservati ha conseguenze. Per Canonity esistono modalità *"non*

*addestrare il modello su questi dati"*, accordi sulla riservatezza, hosting europeo. Vanno sapute, attivate. Se carichi documenti sensibili, prima leggi le condizioni d'uso (le trovi nelle impostazioni dell'account su Canonity).

**Terza: la trascrizione audio non è perfetta.** Su audio rumorosi, voci sovrapposte, dialetti stretti, il modello può attribuire frasi alla persona sbagliata. Per verbali ufficiali, dichiarazioni, testimonianze, ascolta sempre l'audio prima di firmare quello che il modello ha scritto.

## Renato, dopo tre mesi

Lo studio ha riorganizzato il modo di lavorare. I clienti mandano foto via WhatsApp, una collaboratrice le carica nel ChatBox di Canonity, ottiene le tabelle, le incolla nel gestionale. Tempo per pila di documenti: passato da una giornata a un'ora. I contratti vengono letti in due passi: prima dal modello per le clausole chiave, poi da Renato per la conferma. Tempo per contratto: da quaranta minuti a quindici.

Renato ha assunto una praticante in più, e adesso lo studio gestisce un quaranta per cento di clienti in più, con la stessa fatica di prima.

Renato non è cambiato come persona. Continua a non amare le AI. *"Non mi piace il modo in cui ne parlano i giornali. Non mi piacciono le promesse. Ma il ChatBox di Canonity sì, mi piace. Perché fa il suo dovere, e sta zitto"*.

Quella frase (*"fa il suo dovere e sta zitto"*) credo sia il complimento più alto che un commercialista italiano possa fare a una tecnologia.

## Vademecum del capitolo 5

### Cosa è il multimodale, in una riga

Il prompt non è più solo parole: nel ChatBox di Canonity puoi caricare immagini, PDF, audio col pulsante "Allega". Il modello li vede insieme al testo che scrivi, e lavora su entrambi.

### Le quattro tipologie di uso

| *Tipologia* | *Esempio tipico* | *Modello consigliato* |
|---|---|---|
| **Estrarre dati strutturati** | Scontrini, fatture, contratti scansionati | ChatGPT o Claude |
| **Riassumere un documento lungo** | Contratto di 35 pagine, report annuale | Claude |
| **Analizzare grafici e dati visivi** | Cruscotto Google Analytics, schermata | Gemini |
| **Confrontare due o più documenti** | Due offerte di leasing, due versioni di un testo | Claude o Gemini |

### La regola d'oro

> Caricare un documento NON sostituisce il prompt. Lo accompagna.

### Le tre attenzioni

1. **Errori di lettura sui numeri.** Per cifre in atti ufficiali, verifica sul documento originale.
2. **Dati sensibili.** Prima di caricare documenti riservati, leggi le condizioni d'uso e attiva l'opzione "non addestrare".

3. **Trascrizione audio non perfetta.** Per verbali e testimonianze, ascolta sempre l'audio originale.

## Prova questo in Canonity

**Esercizio. Lo scontrino e il PDF lungo.**

1. Apri **canonity.com/chat**. Prendi un qualunque scontrino o ricevuta che hai in tasca, fotografalo col telefono.
2. Nel ChatBox, clicca **"Allega"** sotto il campo di input e carica la foto. Lascia ChatGPT come modello per questo esercizio (è il più veloce sull'estrazione).
3. Scrivi: *"Sei un contabile. Estrai data, esercizio commerciale, lista delle voci con prezzo unitario, totale, metodo di pagamento. Restituisci in formato tabella. Se un dato non è leggibile, scrivi 'non leggibile'."*. Premi invio.
4. **Adesso** prendi un documento sul computer che hai sempre evitato di leggere (un contratto, un regolamento condominiale, le condizioni del conto corrente). Nel ChatBox, **cambia modello a Claude** dal selettore in basso (è il più adatto al testo lungo). Clicca *"Allega"*, carica il PDF.
5. Scrivi: *"Lavorando solo sul documento allegato, rispondi a queste tre domande che ho sempre voluto sapere: [domanda 1] [domanda 2] [domanda 3]. Per ogni risposta indica la pagina del documento. Se l'informazione non è presente, scrivi 'non presente'."*.
6. Confronta i due risultati con il tempo che ti sarebbe servito per leggere il documento a mano. Quanto hai recuperato?

## Prima del prossimo capitolo

Nel prossimo capitolo/parte due di questo libro, torneremo al testo, ma alzando il livello di sistematicità. Vedremo i **framework di prompting** ovvero degli schemi già pronti che incorporano tutto quello che hai imparato finora in formati riusabili.

Si chiamano

- CRISPE,
- RACE,
- RTF,
- RISEN,
- APE,
- COSTAR,
- TASS *(questo è mio)*.

Nomi che fanno paura, ma in realtà sono cinque-sei caselle che impari in mezz'ora.

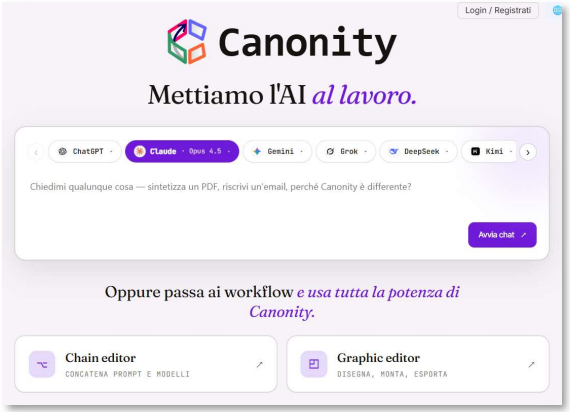
# Appendice. Come iniziare con Canonity

## Una guida pratica per arrivare al tuo primo prompt

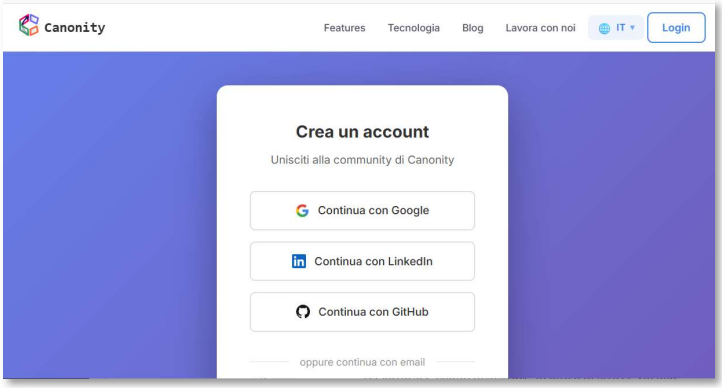
Hai letto il libro. Avrai sentito nominare Canonity decine di volte, e a un certo punto vorrai aprirlo davvero. Questa appendice è la guida operativa: come arrivare al sito, come iscriversi al piano *Free* (gratuito per sempre, sufficiente per tutti gli esercizi del libro), e un tour dei tre tool (ChatBox, prompt chain editor, workflow editor).

## Passo 1. Arrivare al sito

Apri il tuo browser e digita **canonity.com**



Puoi subito usare il chatbot, ma la quantità di prompt è limitata, per cui ti consiglio di usare il pulsante per registrarsi (*"Login/Registrati"* in alto a destra) e vai alla pagina di registrazione.




Puoi usare automaticamente il tuo account google o linkedin.

Una volta dentro, i tre tool del libro sono raggiungibili direttamente da:

- **canonity.com/chat** → il ChatBox
- **canonity.com/chains** → il prompt chain editor
- **canonity.com/workflows** → il workflow editor (di solito a pagamento)

## Passo 2. Iscrizione

Cinque minuti:

**Strada A**, email e password. Inserisci email, scegli una password, accetta le condizioni. Premi *"Registrati"*. Riceverai una mail di conferma entro un minuto: clicca il link per attivare l'account.

**Strada B**, account esistente. Se hai già un account Google, Microsoft o Apple, premi il pulsante corrispondente. In dieci secondi sei dentro.

> **Nota sulla privacy.** Canonity è un servizio europeo, con server ospitati in UE. La privacy policy specifica cosa fa con i dati che carichi: per i piani professionali esiste l'opzione *"non addestrare i modelli sui miei dati"*, attivata di default. Per uso professionale serio, vale la pena leggere le condizioni una volta e configurare le impostazioni come ti servono.

### Passo 3. Scegli il piano *Free*

Subito dopo la registrazione, Canonity ti potrebbe chiedere quale piano vuoi. Per cominciare (e per *tutti* gli esercizi di questo libro) il piano **Free** è più che sufficiente.

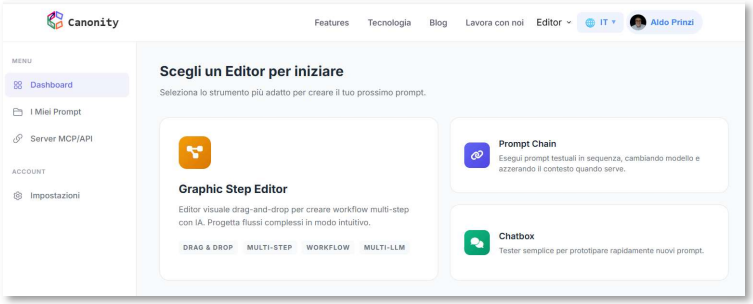
Il piano *Free* include normalmente:

- **ChatBox multi-LLM** con un numero di messaggi al mese sufficiente per la pratica quotidiana.
- **Prompt chain editor** con tre-cinque chain attive: basta per imparare e per costruire le prime catene.
- I principali modelli AI: ChatGPT, Claude, Gemini, DeepSeek, Grok (con limiti che variano in base al piano).

Quando i tuoi bisogni cresceranno, passerai al piano superiore. Per il libro, il *Free* è il punto giusto.

### Passo 4. La dashboard

Dopo la conferma del piano, vieni portato alla *dashboard*: la pagina di partenza da cui accederai sempre.




Tre schede principali, una per ogni tool:

7. **ChatBox** (canonity.com/chat) — la chat multi-LLM. Apri questa per gli esercizi del capitolo *Il prompt* e dei capitoli successivi.
8. **Prompt Chain Editor** (canonity.com/chains) — dove costruirai le catene del capitolo *Chain prompts* e la chain TASS.
9. **Workflow Editor** (canonity.com/workflows) — con gli agenti e i workflow del capitolo *Prompt come agenti e workflow*.

Sotto le tre schede c'è *"Inizia da qui"* con template pronti (*"Triagia le mie email"*, *"Confronta tre modelli sullo stesso prompt"*, *"Crea un CV perfetto"*). Sono esempi che ti permettono di vedere subito un workflow funzionare. Cliccane uno per un assaggio rapido.

Sulla sidebar di sinistra: **Cronologia** delle conversazioni, **Chain salvate**, **Workflow attivi**, **Impostazioni**, **Supporto**.

## Un tour rapido dei tre tool

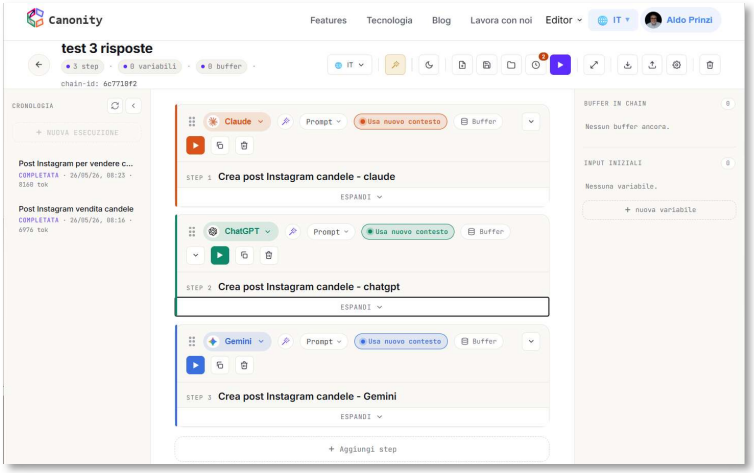
### Il ChatBox

Il ChatBox è probabilmente il tool che userai più spesso. Lo distingue dagli altri chatbot la *riga di selezione dei modelli* sopra il campo di input.

Tre cose pratiche:

- Per *cambiare modello fra un messaggio e l'altro*, clicca sul nome del modello desiderato nella riga di selezione prima di premere invio.
- Per *allegare un file* (PDF, immagine, audio), usa il pulsante *"Allega"* a sinistra. È così che il multimodale prende vita.
- Per *generare un'immagine*, scrivi nel prompt *"genera un'immagine di…"*: il ChatBox passa la richiesta al modello adatto.

### Il prompt chain editor

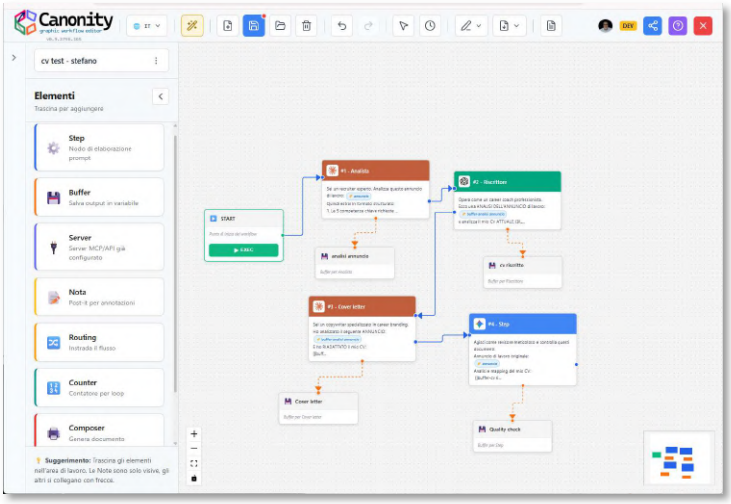


Visualmente è un *canvas* su cui aggiungi *nodi* (gli step) e li colleghi con *frecce*.

Tre azioni essenziali:

- **Aggiungere uno step**: clicca *"Aggiungi step"* nella toolbar di sinistra.
- **Collegare due step**: clicca sul cerchio di uscita di un nodo e trascina fino al cerchio di entrata del successivo.
- **Lanciare la chain**: in alto a destra c'è *"Esegui chain"*. Canonity ti chiederà i valori delle variabili, e poi farà girare gli step in cascata.

## Il workflow editor




Il fratello maggiore del chain editor. Due differenze sostanziali:

**Prima**: la toolbar di sinistra è più ricca con i nuovi tipi di step (ROUTER, API, IMAP, WEBAGENT, ecc.).

**Seconda**: in alto a destra hai *"Schedula"*, che permette al tuo workflow di girare *automaticamente* (per esempio, ogni lunedì alle sette). E *"Modalità bozza"*: quando attivata, il workflow non esegue le azioni esterne (non manda davvero le email), ma le *prepara* in una zona separata che tu controlli prima di confermare. È la modalità consigliata per i primi mesi.

## Domande pratiche

**Posso usare Canonity dal telefono?** Sì, c'è un'app mobile per iOS e Android con il ChatBox completo. Il chain editor e il workflow editor funzionano meglio da computer.

**Quanto pago se passo al piano superiore?** Al maggio 2026, indicativamente: *Starter* sui 15 €/mese, *Pro* sui 30 €/mese, *Team* sui 50 € per posto/mese. Il *Free* resta tale, sempre.

**Posso esportare le mie chain e i miei workflow?** Sì, in formato JSON. Utile per backup o per condividere una chain con un collega.

**C'è una community?** Sì: canale Discord ufficiale, forum sul sito, tutorial su YouTube. Il link è nella sidebar della dashboard, sezione *"Supporto"*.

## Errori comuni del primo accesso

**Non confermare la mail di registrazione.** Controlla la cartella *Spam*. Senza confermare, alcuni tool potrebbero non funzionare.

**Provare a costruire un workflow al primo accesso.** Resisti alla tentazione. Comincia dal ChatBox, fai i tuoi primi esperimenti, prendi confidenza. Solo dopo qualche settimana di pratica ha senso passare al chain editor, e solo dopo un altro mese di chain ha senso passare ai workflow. La curva di apprendimento del libro segue questo ordine non a caso.

**Non leggere i template di esempio.** Quei template nella sezione *"Inizia da qui"* sono pensati per insegnarti come si fanno le cose. Aprine uno, guarda com'è strutturato, e *poi* costruisci il tuo.

**Dimenticarsi di salvare.** Nel chain editor e nel workflow editor le modifiche si salvano *automaticamente* quasi sempre, ma non sempre. Quando hai finito una sessione importante, premi *"Salva"* manualmente.

## Un'ultima nota

Canonity è un prodotto che cresce in fretta. Nei mesi successivi all'uscita di questo libro, alcune cose specifiche potrebbero essere già cambiate (un'interfaccia leggermente diversa, un piano rivisto, una nuova funzionalità). La sostanza (la *postura* da avere davanti all'AI, la mappa concettuale, i principi) non cambia.

Per il resto: buon lavoro. E grazie di avere seguito fino in fondo.

*Il team di Canonity – hey@canonity.com*

# OUTSOURCE YOUR THINKING

## Parte 1 — L’uso del ChatBox

Aldo Prinzi

*«Il salto di qualità nel prompting non viene dai trucchi, ma dal processo.»*

### L’autore

Aldo Prinzi è un veterano dell’informatica, con oltre quarant’anni di esperienza nella programmazione e due brevetti software. Nel 1994 fu tra i primi in Italia a ottenere la certificazione Novell CNE. Ha progettato il sistema di ricerca delle Teche RAI, è stato formatore nei corsi ufficiali Microsoft e per due volte speaker alla Microsoft WPC. Programma ancora ogni giorno e oggi lavora direttamente allo sviluppo di Canonity. Il suo blog: aldo.prinzi.it

### Un’opera gratuita in tre volumi

- **Parte 1 — L’uso del ChatBox (questo volume)**
- Parte 2 — L’uso dei chain prompt
- Parte 3 — L’uso dei workflow

Distribuzione gratuita consentita solo in forma integra e inalterata. Licenza e avvertenze complete: eur.re/licenza_avvertenze

### Canonity, il laboratorio del libro

Tutti gli esercizi si svolgono su Canonity: il ChatBox multi-LLM per usare ChatGPT, Claude, Gemini e altri modelli in un’unica chat, il prompt chain editor per concatenare i prompt e il workflow editor per costruire agenti. Il piano Free è sufficiente per tutto il percorso del libro. canonity.com

Canonity SRL · Via Valdemone 31, 90144 Palermo · P.IVA IT07416980824 · REA PA457557 · Cap. Soc. € 10.000 i.v. · Iscritta al registro speciale delle startup innovative

**Scarica gratuitamente questo volume e trova anche gli altri della serie dal blog di Aldo Prinzi**

**https://aldo.prinzi.it**