CORSO IA CON CLAUDE · DISPENSA PER L'ALLIEVO

Ragionamento a passi, esempi e caching

Tre tecniche per ottenere risposte migliori spendendo meno

ORA	DURATA	MODULO
06 di 12	60 minuti	3 — Tecniche avanzate

Gianluca Bonomo · www.gianlu.it · info@gianlu.it

Indice

Indice 2
Come usare questa dispensa 3
1. Ragionare prima di rispondere 4
1.1 Perché funziona 4
1.2 Come si presenta, in pratica 5
1.3 I limiti del ragionamento esposto 5
2. Insegnare con gli esempi 6
2.1 Un esempio completo 7
2.2 Quando gli esempi peggiorano le cose 7
2.3 Esempi o istruzioni? Come si sceglie 8
2.4 Dove si prendono gli esempi 8
3. Pagare una volta sola 9
3.1 Il problema 9
3.2 La soluzione 9
3.3 Come si misura il risparmio 10
3.4 Quando il riuso del contesto non serve 10
3.5 Che cosa significa, per chi non programma 11
4. Quale tecnica, per quale compito 11
4.1 Come si combinano 12
5. Esercitazione pratica dell'ora 12
5.1 Un confronto da fare davvero 12
Le domande che arrivano in aula 13
Glossario 15
Per approfondire 16

Come usare questa dispensa

Questa dispensa raccoglie, in forma scritta e ordinata, tutto quello che vediamo insieme durante l'ora di lezione. Non devi studiarla a memoria: serve a farti ritrovare le cose con calma, dopo, quando l'aula è finita e ti resta una domanda in sospeso.

È pensata per chi parte da zero. Ogni termine tecnico viene spiegato la prima volta che compare, e a fine dispensa trovi un glossario con tutte le parole nuove. Se una pagina ti sembra difficile, saltala e torna indietro dopo: nessun concetto di questa dispensa richiede conoscenze informatiche precedenti.

> **COME SONO FATTI I RIQUADRI**
>
> **In parole semplici** (azzurro): la spiegazione senza tecnicismi, quella da portarsi a casa se si ricorda una cosa sola.
>
> **Attenzione** (oro): un errore frequente, un rischio, una cosa da non fare.
>
> **Esempio** (grigio): un caso concreto, preso dalla vita di tutti i giorni o dal lavoro in azienda.

Obiettivo dell'ora

Imparare a far ragionare il modello prima che risponda, a insegnargli un formato con due esempi invece che con dieci righe di istruzioni, e a non pagare due volte lo stesso documento.

Che cosa saprai fare alla fine

- Applicare il ragionamento a passi ai compiti in cui l'errore costa.
- Usare due esempi per fissare tono, struttura e formato.
- Capire che cos'è il prompt caching e come si imposta un prompt perché ne tragga vantaggio.
- Scegliere la tecnica giusta in base al tipo di compito.

1. Ragionare prima di rispondere

La tecnica più semplice del corso, e la più sottovalutata.

Chiedere a un modello la risposta e basta è come pretendere da una persona un calcolo complicato a mente, senza foglio. Talvolta ci azzecca. Spesso no.

Il **Chain-of-Thought**, in italiano *catena di ragionamento*, consiste nel chiedere esplicitamente di ragionare a passi prima di dare la risposta finale. Basta una frase: «prima analizza, poi rispondi». Oppure il tag **<thinking>**, che rende la richiesta esplicita.



Stessa domanda, due percorsi diversi.

1.1 Perché funziona

Torniamo al meccanismo della prima ora: il modello genera **una parola alla volta**, e ogni parola dipende da quelle che ha già scritto. Se gli chiedi solo il risultato, deve arrivarci in un colpo solo. Se gli fai scrivere i passaggi, ogni passaggio diventa materiale su cui appoggiare il successivo.

Non è una furbizia: è dare al modello lo spazio per fare quello che noi faremmo su un foglio di brutta.

> **QUANDO USARLO SEMPRE**

Calcoli, date, scadenze, confronti fra due documenti, verifiche di conformità, analisi di clausole.

In tutti i casi, cioè, in cui la risposta sbagliata non si vede a occhio.

IL COSTO

Ragionare a passi significa scrivere di più, quindi consumare più token e impiegare qualche secondo in più.

Su un contratto, un errore costa infinitamente di più di quei secondi.

1.2 Come si presenta, in pratica

Chiedendo il ragionamento a passi, la risposta cambia forma: prima compare l'analisi, poi la conclusione. Su un controllo di scadenze contrattuali, per esempio, si passa da questo:

> *Il contratto scade il 31 dicembre 2026.*

a questo:

> *Analisi: l'articolo 12 fissa una durata di 24 mesi dalla sottoscrizione. La data di firma in calce è il 15 gennaio 2025. Ventiquattro mesi dal 15 gennaio 2025 portano al 15 gennaio 2027. L'articolo 13 prevede però il rinnovo tacito, salvo disdetta con 90 giorni di preavviso. Conclusione: la scadenza è il 15 gennaio 2027; il termine ultimo per la disdetta è il 17 ottobre 2026.*

La seconda risposta è più lunga, ma è **verificabile**: si può controllare ogni passaggio. E nel farla è emersa una clausola di rinnovo tacito che nella prima risposta era semplicemente sparita.

IL VALORE NASCOSTO

Il ragionamento esposto non serve soltanto al modello: serve **a voi**. È il punto in cui potete accorgervi che ha interpretato male un dato, invece di scoprirlo tre settimane dopo.

1.3 I limiti del ragionamento esposto

Chiedere di ragionare migliora molto i risultati, ma non li rende infallibili, e conviene sapere esattamente che cosa non risolve.

- **Un ragionamento ben esposto può essere sbagliato.** La forma ordinata rende più convincente anche una conclusione errata: è un effetto noto, e vale anche fra esseri umani.
- **Non aggiunge informazioni che non ci sono.** Se il dato non è nel documento, nessun ragionamento lo farà comparire: al massimo lo farà comparire in modo più credibile.
- **Non sostituisce il calcolo.** Su operazioni aritmetiche complesse conviene chiedere esplicitamente il passaggio numerico e poi ricontrollarlo con una calcolatrice.

> **L'USO GIUSTO**
>
> Il ragionamento esposto serve a **rendere verificabile** la risposta, non a rendere superflua la verifica.
>
> Se leggendolo non avete tempo di controllare nessun passaggio, l'unico vantaggio che resta è quello statistico: reale, ma non sufficiente per firmare un documento.

2. Insegnare con gli esempi

I modelli linguistici sono straordinari nel riprodurre uno schema. Il modo più rapido per ottenere il formato che vuoi è mostrarglielo.

Il **Few-Shot prompting** consiste nel mettere dentro il prompt due o tre esempi del risultato desiderato: un input tipo e, accanto, l'output che vorresti. Il modello coglie lo schema e lo replica.

> **DESCRIVERE CONTRO MOSTRARE**
>
> **Descrivendo**: «scrivi una scheda prodotto con nome in grassetto, due righe di descrizione in tono informale, poi tre caratteristiche a elenco puntato, senza superlativi, massimo cinquanta parole…»
>
> **Mostrando**: incolli una scheda già fatta come la vuoi. Fine.
>
> Il secondo modo è più breve, più preciso e molto più difficile da fraintendere.

Due o tre esempi bastano. Oltre, si consuma contesto senza guadagno, e si rischia che il modello copi anche i dettagli specifici degli esempi invece dello schema.

Gli esempi vanno dentro il tag **<example>**, separati dal materiale vero: altrimenti il modello può confondere l'esempio con il documento da elaborare.

2.1 Un esempio completo

Immaginiamo di dover classificare le segnalazioni che arrivano al servizio clienti. Descrivere a parole le categorie richiede mezza pagina e lascia comunque dei dubbi. Due esempi le definiscono meglio di qualsiasi spiegazione.

> *<example>*
> *Segnalazione: «Il pacco doveva arrivare martedì, siamo a venerdì e non si è visto nessuno.»*
> *Categoria: LOGISTICA — Urgenza: media — Risposta attesa: entro 24h*
>
> *Segnalazione: «Mi avete addebitato due volte lo stesso ordine.»*
> *Categoria: AMMINISTRAZIONE — Urgenza: alta — Risposta attesa: entro 4h*
> *</example>*

Con questi due esempi davanti, la classificazione delle segnalazioni successive esce nel formato giusto, con gli stessi nomi di categoria e la stessa scala di urgenza. Senza, ogni volta cambia qualcosa: «logistica» diventa «spedizione», «alta» diventa «urgente».

PERCHÉ DUE E NON UNO

Con un esempio solo, il modello tende a copiarne i dettagli: se l'unico esempio riguarda un pacco in ritardo, tende a leggere tutto come un problema di consegna.

Due esempi **diversi fra loro** mostrano che cosa varia e che cosa resta fisso: ed è esattamente quello che deve imparare.

2.2 Quando gli esempi peggiorano le cose

Ci sono tre casi in cui aggiungere esempi fa danno, e vale la pena riconoscerli.

Il caso	Che cosa succede
Gli esempi sono tutti dello stesso tipo	Il modello deduce una regola che non c'è, e la applica anche dove non c'entra
Gli esempi contengono un errore	L'errore viene replicato con grande coerenza in tutte le risposte successive
Gli esempi sono più lunghi del materiale da analizzare	Si consuma contesto e attenzione per insegnare qualcosa che una riga di istruzione avrebbe detto meglio

IL CONTROLLO PRIMA DI USARLI

Rileggete gli esempi chiedendovi: «se qualcuno imitasse **solo** questi, che cosa concluderebbe?»

Se la risposta contiene qualcosa che non volevate insegnare, l'esempio va cambiato.

2.3 Esempi o istruzioni? Come si sceglie

La stessa cosa si può ottenere descrivendola a parole o mostrandola con un esempio. Non sono equivalenti, e la scelta ha una regola semplice.

Conviene l'istruzione	Conviene l'esempio
La regola è breve e si dice bene a parole: «massimo dieci righe»	La regola è facile da mostrare e difficile da descrivere: un formato, uno stile, un tono
Deve valere in generale, anche su casi che non avete previsto	I casi sono ripetitivi e simili fra loro
Volete poterla cambiare in fretta	Volete che il risultato sia identico tutte le volte

Nella pratica si usano insieme: le istruzioni dicono che cosa fare, gli esempi mostrano come deve venire. Quando le due cose si contraddicono, **vince quasi sempre l'esempio**: è la ragione per cui un esempio sbagliato fa più danni di un'istruzione sbagliata.

2.4 Dove si prendono gli esempi

La difficoltà pratica non è capire che gli esempi servono: è trovarli senza perderci mezza giornata. Tre strade, tutte già percorribili oggi.

1. **Dal lavoro già fatto.** Il documento dell'anno scorso, la risposta a un reclamo che era venuta bene, la scheda compilata correttamente. È la fonte migliore, perché è già nel vostro stile.
2. **Da una prima risposta corretta a mano.** Si chiede senza esempi, si corregge il risultato finché è giusto, e **quella versione corretta diventa l'esempio** per tutte le volte successive.
3. **Da un esempio costruito apposta**, con dati inventati. È la strada obbligata quando il materiale vero contiene dati personali o riservati.

PRIMA DI RIUSARE UN LAVORO VERO COME ESEMPIO

Togliete i dati reali: nomi, importi, riferimenti di pratica. Un esempio viaggia insieme al

> prompt, e un prompt viene condiviso.
>
> È lo stesso avvertimento della decima ora, e in aula è quello che viene dimenticato più spesso.

3. Pagare una volta sola

La parte più tecnica dell'ora, spiegata senza tecnicismi. Serve a chiunque rimandi ogni giorno lo stesso materiale.

3.1 Il problema

Ogni richiesta viene elaborata da sola: il modello non ricorda nulla fra una chiamata e l'altra. Questo significa che se alleghi un manuale di cento pagine e fai dieci domande, quel manuale viene rimandato e rielaborato **dieci volte**. Paghi dieci volte, e aspetti dieci volte.

3.2 La soluzione

Il **prompt caching** mette da parte la porzione iniziale della richiesta che non cambia. Alla richiesta successiva, se quella porzione è identica, viene riletta dalla cache: molto più in fretta e a costo sensibilmente ridotto.



La parte che non cambia si paga a pieno solo la prima volta.

LA REGOLA PRATICA, E L'UNICA COSA DA RICORDARE

Prima ciò che non cambia, dopo ciò che cambia.

Il documento, le istruzioni fisse e gli esempi vanno all'inizio. La domanda del momento va in fondo.

CHE COSA MANDA ALL'ARIA LA CACHE

Il confronto è sul **prefisso esatto**: basta un carattere diverso all'inizio e tutto quello che viene dopo va ricalcolato.

Attenzione quindi a mettere in cima cose che cambiano da sole: una data, un orario, un numero progressivo, il nome dell'utente.

Se sospetti che non stia funzionando, il conto dei token letti dalla cache te lo dice: se resta a zero, qualcosa all'inizio sta cambiando.

3.3 Come si misura il risparmio

Chi lavora tramite interfaccia di programmazione vede i numeri: ogni risposta riporta quanti token sono stati riletti dalla cache e quanti sono stati elaborati da zero. Se il primo numero resta a zero, la cache non sta funzionando, e quasi sempre il motivo è che qualcosa cambia all'inizio della richiesta.

Chi lavora dall'interfaccia normale non vede quei numeri, ma può misurare due cose altrettanto utili:

- **Il tempo di risposta.** Su un documento lungo, la differenza fra la prima richiesta e le successive si nota a occhio.
- **Quanto in fretta si esaurisce il limite d'uso.** È l'indicatore più concreto che avete a disposizione: se lavorando tutto il giorno sullo stesso materiale arrivate a sera senza interruzioni, state impostando bene le richieste.

3.4 Quando il riuso del contesto non serve

Non tutti i lavori ne beneficiano, e insistere dove non serve complica il prompt senza vantaggio.

- **Se il materiale cambia a ogni richiesta**, non c'è nulla da riusare: ogni volta si riparte comunque da zero.

- **Se il materiale è breve**, il guadagno è impercettibile: su mezza pagina non c'è niente da risparmiare.
- **Se fate una domanda sola e chiudete**, non c'è una seconda richiesta che possa approfittarne.

Il caso in cui invece rende moltissimo è sempre lo stesso: **un materiale grande e fisso, molte domande diverse**. L'analisi di un fascicolo, le verifiche su un capitolato, le domande su un manuale.

3.5 Che cosa significa, per chi non programma

Tutto questo capitolo parla di un meccanismo che si governa in modo esplicito solo tramite interfaccia di programmazione. Ma il principio che ne sta sotto cambia il modo di lavorare anche dalla pagina web, e vale la pena tradurlo in tre abitudini.

1. **Caricate il materiale una volta sola, all'inizio.** Poi fate tutte le domande che volete nella stessa conversazione, invece di ricominciare da capo per ognuna.
2. **Non modificate il materiale a metà strada.** Sostituire il documento iniziale con una versione corretta annulla il vantaggio e, soprattutto, confonde: da quel momento non è più chiaro a quale versione si riferiscano le risposte precedenti.
3. **Tenete separate le conversazioni per documento.** Una conversazione per fascicolo funziona molto meglio di una conversazione unica in cui i fascicoli si accumulano.

> **IL MODO GIUSTO DI PENSARCI**
>
> Non è «risparmiare sul costo»: è **montare il lavoro nell'ordine giusto**.
>
> Prima quello che non cambia, poi quello che cambia. È la stessa regola della quinta ora, vista dal lato della macchina invece che dal vostro.

4. Quale tecnica, per quale compito

Se il compito è…	La tecnica giusta	Perché
Analizzare, confrontare, verificare	Ragionamento a passi	L'errore nasce dal saltare un passaggio.
Produrre tante cose uguali	Due esempi (few-shot)	Il problema non è capire, è la forma.
Interrogare sempre lo stesso materiale	Caching	Il costo sta nel rimandare ogni volta la stessa cosa.

Se il compito è...	La tecnica giusta	Perché
Tutte e tre insieme	Si sommano	Documento e esempi all'inizio, ragionamento e domanda in fondo.

4.1 Come si combinano

Le tre tecniche di quest'ora non si escludono: il loro effetto migliore si vede quando lavorano insieme, ciascuna al suo posto nel prompt.

1. **Gli esempi vanno prima.** Stanno con il materiale stabile, in testa al prompt: sono la parte che non cambia e che conviene far riusare.
2. **La richiesta di ragionare va nelle istruzioni.** Una riga: «prima analizza, poi concludi».
3. **Il formato va per ultimo.** È la cosa che il modello deve avere più fresca quando comincia a scrivere.

Il risultato è un prompt che ha la stessa struttura della quinta ora, con due aggiunte: un blocco di esempi in alto e una riga di ragionamento nelle istruzioni. Non è una tecnica nuova: è la stessa impalcatura, usata meglio.

5. Esercitazione pratica dell'ora

Un prompt solo, con due tecniche insieme, su un caso concreto.

Obiettivo: estrarre le clausole rilevanti da un testo contrattuale.

1. Scrivi il prompt con i tag della quinta ora.
2. Aggiungi il **ragionamento a passi**: «individua prima le sezioni del contratto, poi analizza ciascuna sezione una per una».
3. Aggiungi **un esempio**: una clausola già estratta nel formato che vuoi, dentro <example>.
4. Metti il contratto e l'esempio **prima**, la domanda **dopo**.

Poi confronta il risultato con quello ottenuto dal prompt secco della scorsa ora. La differenza è il senso di questa lezione.

5.1 Un confronto da fare davvero

L'esercizio più utile di quest'ora è anche il più semplice: la stessa richiesta, fatta due volte, e i due risultati messi uno accanto all'altro.

1. **Primo giro.** Prendete un compito vostro che richiede qualche passaggio — un controllo di scadenze, un confronto fra due offerte, una classificazione — e chiedetelo così come vi viene.
2. **Secondo giro.** La stessa identica richiesta, con due aggiunte: «prima analizza passo per passo, poi concludi» e un esempio del risultato che volete.
3. **Confronto.** Incollate i due risultati in due colonne e guardate tre cose: che cosa è comparso nel secondo e mancava nel primo; che cosa è cambiato nel formato; quanto tempo vi servirebbe per verificare l'uno e l'altro.

È l'esercizio che convince più di qualsiasi spiegazione, perché il confronto lo fate su un lavoro di cui conoscete già la risposta giusta.

Le domande che arrivano in aula

Sono le stesse in ogni corso. Vale la pena avere pronta una risposta breve e onesta, anche per quando le farà un collega.

«Se gli chiedo di ragionare, mi costa di più?»

Sì: qualche parola in più e qualche secondo in più. Ma un errore su un contratto o su una scadenza costa infinitamente di più.

E sui compiti in cui il materiale si ripete, il riuso del contesto recupera gran parte della differenza.

«Devo per forza mostrare il ragionamento al cliente?»

No. Potete chiedere di ragionare e poi di riportare **solo** la conclusione: «analizza passo per passo, poi scrivi soltanto la risposta finale in tre righe».

Il beneficio sulla qualità resta; cambia solo quello che vedete.

«Quanti esempi conviene dare?»

Due o tre. Con uno solo il modello rischia di copiarne i dettagli specifici invece dello schema; oltre i tre si consuma contesto senza guadagno.

Se con tre esempi il formato ancora non torna, il problema non sono gli esempi: è che il formato non è stato deciso bene nemmeno da voi.

«Il riuso del contesto vale anche nella chat normale?»

Il meccanismo è pensato soprattutto per l'uso tramite interfaccia di programmazione, dove lo si governa in modo esplicito.

Ma la regola che ne discende — mettere prima ciò che non cambia e dopo ciò che cambia — è comunque il modo giusto di scrivere un prompt lungo, e nei progetti diventa naturale.

Glossario

Le parole che incontri in questa dispensa, spiegate in una riga. Tienilo a portata di mano: torneranno in tutte le ore successive del corso.

Termine	Che cosa significa
Cache	Memoria temporanea in cui si tiene qualcosa già elaborato per non doverlo rifare.
Chain-of-Thought	Catena di ragionamento: chiedere al modello di esporre i passaggi prima della risposta finale.
Few-shot	Fornire nel prompt due o tre esempi del risultato voluto.
Prefisso	La parte iniziale della richiesta. È quella che viene confrontata per decidere se la cache vale ancora.
Prompt caching	Meccanismo che riusa la parte iniziale invariata di una richiesta, riducendo tempi e costi.
Zero-shot	Chiedere senza dare esempi. È il modo predefinito.

Per approfondire

Le fonti usate per costruire questa dispensa. Sono tutte pubbliche e consultabili: se un dato ti interessa, vai alla fonte e verificalo. È esattamente quello che chiediamo di fare anche all'intelligenza artificiale.

La documentazione sul caching

La guida tecnica di Anthropic descrive il funzionamento del prompt caching, l'ordine consigliato delle parti e come verificare che stia funzionando.

Il ragionamento a passi nel libro

Niente paura, è solo IA, capitolo 6.3, tratta chain-of-thought e few-shot fra le tecniche intermedie di prompting.

Il libro di riferimento del corso

Gianluca Bonomo, *Niente paura, è solo IA. La guida completa per capirla, usarla e restare umani*.

UNA NOTA SULLE DATE

Le norme sull'intelligenza artificiale cambiano in fretta. I dati normativi di questa dispensa sono verificati a **settembre 2026**. Prima di usarli in una decisione aziendale, controlla sempre la fonte ufficiale: il calendario dell'AI Act è già stato modificato una volta.