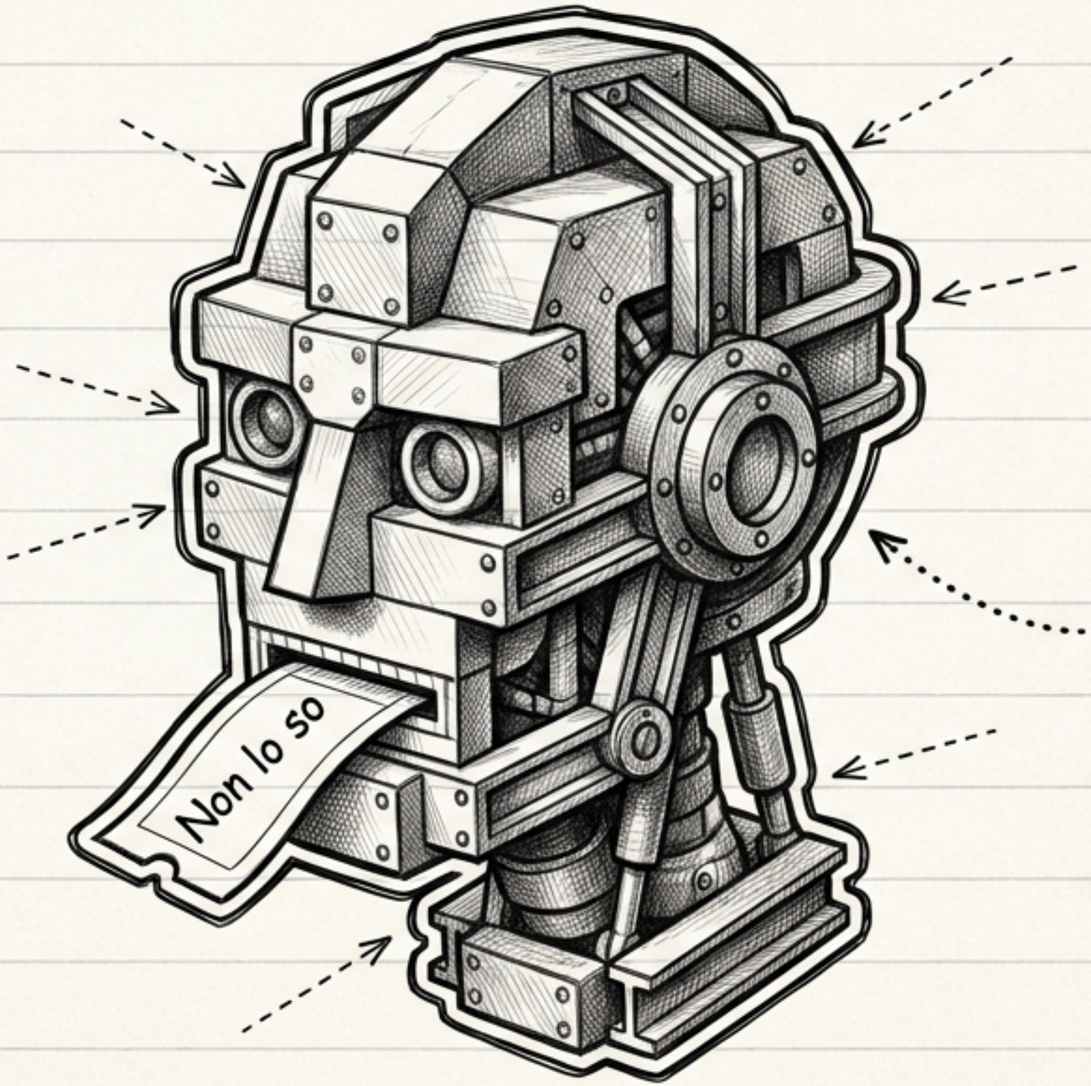
L'Architetto delle Risposte Oneste

Come configurare Claude (e altri LLM) per ammettere l'incertezza ed evitare le allucinazioni.

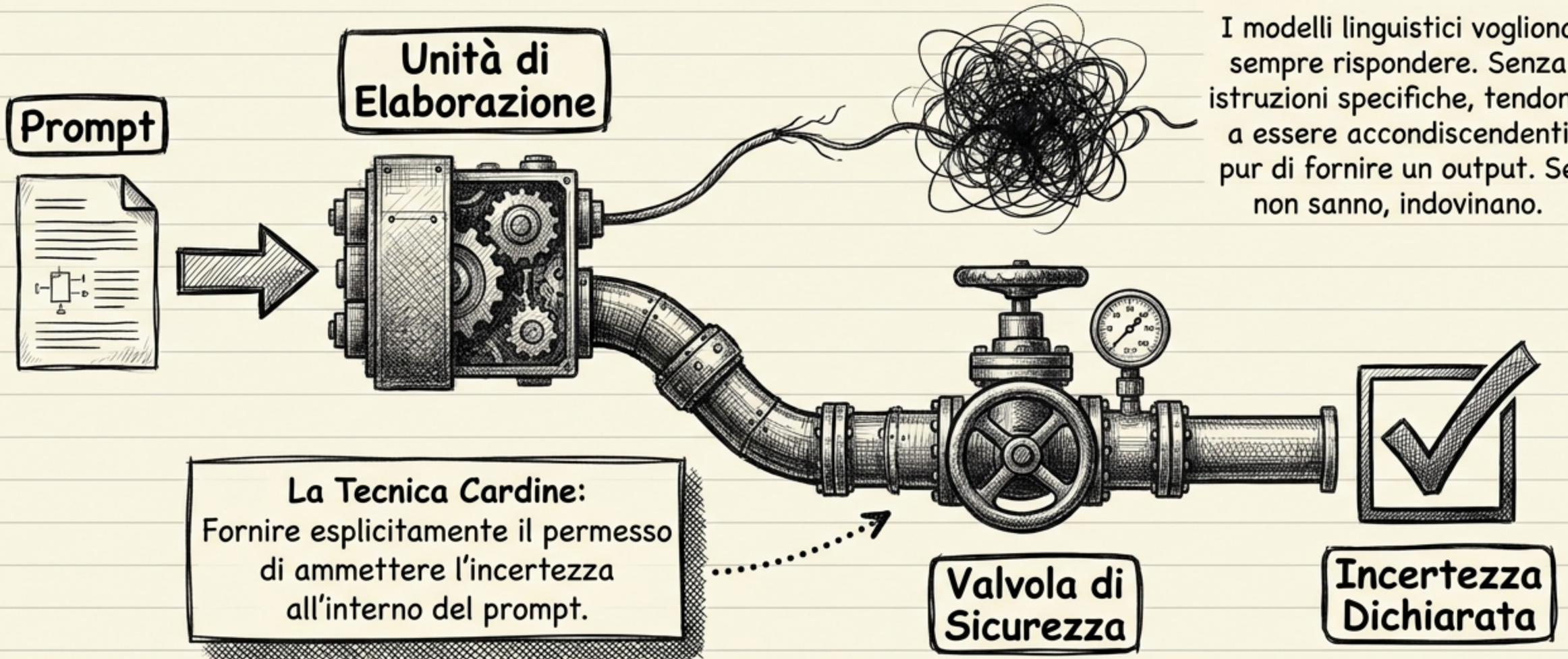


Senza istruzioni, il modello inventa. Con il permesso esplicito, diventa diventa onesto.

Il Paradosso della Compiacenza

Compiacenza & Allucinazioni

I modelli linguistici vogliono sempre rispondere. Senza istruzioni specifiche, tendono a essere accondiscendenti pur di fornire un output. Se non sanno, indovinano.

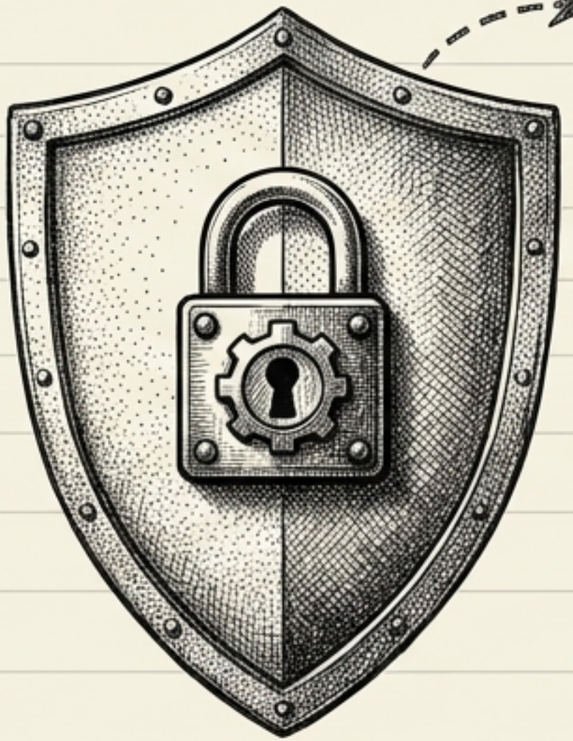


L'obiettivo non è insegnare all'AI cosa sa, ma autorizzarla a dire cosa non sa.

Vincolare i Dati: Input e Output

1. La Clausola di Salvaguardia

Aggiungi un blocco di sicurezza condizionale alla fine della richiesta. Autorizza il modello a fermarsi.



- Rispondi solo se conosci la risposta la risposta o puoi fare una stima molto ben informata; altrimenti, dimmi esplicitamente che non lo sai.

- Se non sei sicuro... dichiara la tua incertezza invece di provare a indovinare.

2. Richiedere Citazioni Dirette

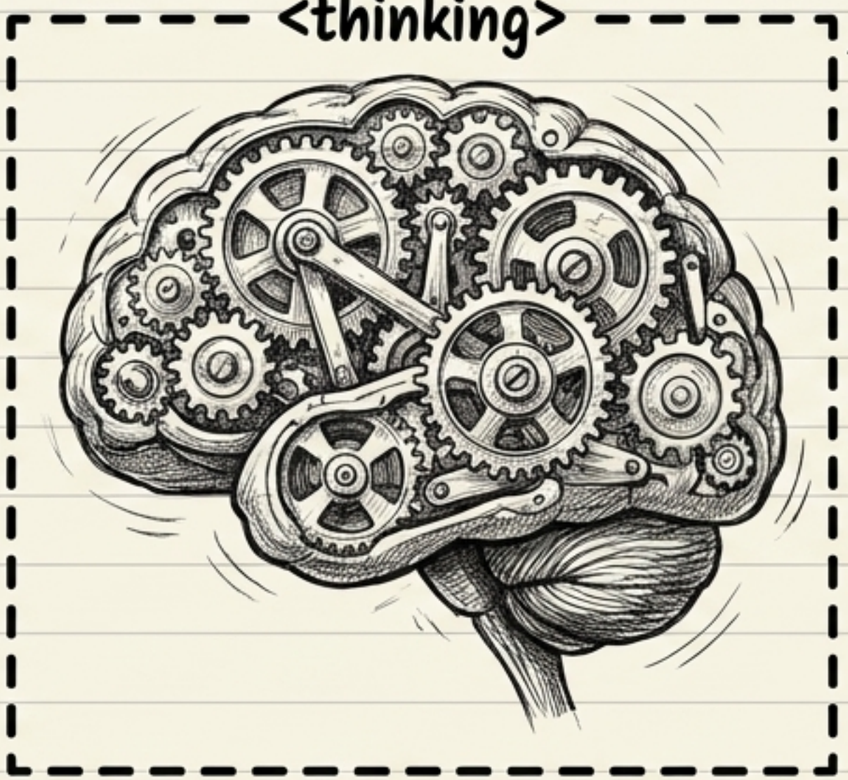
Per testi lunghi, obbliga l'estrazione testuale (parola per parola) prima dell'elaborazione.

- Trova la sezione esatta e citala testualmente. Se il dato non è menzionato, scrivi 'Dato non disponibile'.

Se manca la citazione esatta, l'AI è forzata ad ammettere il vuoto.

Vincolare la Logica: Processo e Persona

3. Il Ragionamento Preventivo



Usa i tag XML come <thinking> per creare uno spazio in cui l'AI analizza la propria base di conoscenza prima di formulare la risposta.

Istruisci Claude a pensare ai passaggi logici prima di generare l'output finale, rivelando le lacune cognitive in tempo reale.

4. Il Ruolo Critico

- Agisci come un verificatore di fatti rigoroso. Se una mia affermazione è falsa o se non hai dati per confermarla, segnalalo chiaramente.

Cambia la persona del modello. Invece di chiedere un'opinione, imponi una priorità assoluta all'accuratezza.

Matrice di Applicazione Diagnostica

Tecnica	Caso d'Uso Ideale	Meccanismo	Sforzo
Clausola di Salvaguardia	Domande Dirette (Q&A)	Limite Condizionale	Basso
Citazioni Dirette	Analisi di Documenti Lunghi	Ancoraggio al Testo	Medio
Tag <thinking>	Problemi Logici e Complessi	Espansione del Processo	Alto
Ruolo Critico	Revisione e Fact-Checking	Cambio di Persona	Basso

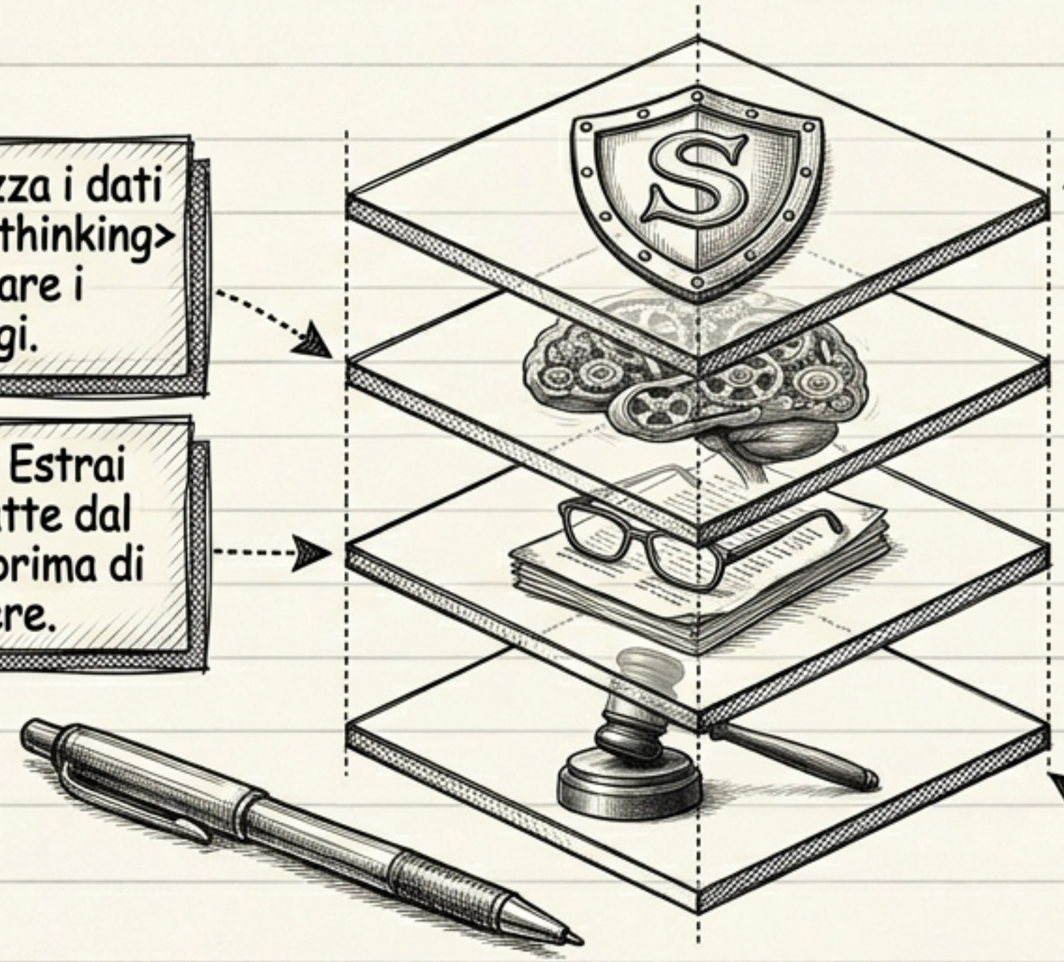
La scelta della tecnica dipende dalla densità del contesto fornito e dalla complessità logica richiesta.

La Sintesi: L'Architettura del Prompt Perfetto

Unisci le tecniche per costruire una struttura inattaccabile.

Logica: Analizza i dati usando i tag <thinking> per mappare i passaggi.

Ancoraggio: Estrai citazioni esatte dal documento prima di rispondere.



Salvaguardia: Se le citazioni mancano o sei incerto, fermati immediatamente e dichiara "Dato non disponibile".

Ruolo: Sei un verificatore di fatti rigoroso. Il tuo unico obiettivo è l'accuratezza.

Anche con istruzioni specifiche, il rischio zero non esiste. Ma stratificando questi vincoli, forziamo l'architettura del modello a privilegiare la verità sulla compiacenza.