Coerenza inter-risposta

Scritto da Febogamedev

La coerenza inter-risposta consiste nel confrontare più risposte relative allo stesso nucleo informativo per osservare quanto il contenuto rimanga stabile sotto variazioni controllate delle condizioni di interrogazione (cfr. Elazar et al., 2021; Jang et al., 2022).

Lo scopo non è verificare se il modello ripeta sempre le stesse parole, ma individuare:
- Quali elementi restano invariati.
- Quali condizioni cambiano.
- Quali definizioni oscillano.
- Quali contraddizioni emergono.
- Quanto il risultato sia sensibile alla formulazione del prompt.

Questa tecnica è utile perché un modello linguistico non recupera necessariamente una risposta fissa da un archivio.
Ogni output viene generato in relazione al contesto disponibile e può quindi variare tra esecuzioni o formulazioni differenti (Manakul et al., 2023; Sclar et al., 2024).
- La stabilità di un contenuto sotto piccole variazioni del prompt può essere considerata un indizio di robustezza (Elazar et al., 2021; Ribeiro et al., 2020).
- Una forte instabilità può invece segnalare ambiguità, fragilità concettuale, dipendenza dal contesto o insufficiente ancoraggio.

È però fondamentale distinguere la coerenza dalla verità:
Una risposta coerente non è necessariamente corretta (Elazar et al., 2021).
Una forte incoerenza segnala invece che il contenuto richiede ulteriori controlli (cfr. Manakul et al., 2023).
Il modello può ripetere stabilmente lo stesso errore.
Allo stesso modo, risposte differenti possono dipendere da interpretazioni legittime, dalla natura controversa dell'argomento o dalla variazione probabilistica della generazione.

La coerenza inter-risposta misura quindi soprattutto la stabilità del contenuto, non la sua corrispondenza con i fatti.
La self-consistency di Wang et al. (2023) è una tecnica distinta: campiona più percorsi di ragionamento e aggrega le risposte finali. Qui, invece, il confronto ha una funzione diagnostica e non seleziona automaticamente la risposta maggioritaria.

È utile distinguere due tipi di confronto:
- Coerenza sotto parafrasi: verifica la stabilità sotto variazioni linguistiche.
- Coerenza trasversale: verifica la compatibilità tra aspetti differenti dello stesso concetto.

Le due forme di controllo sono complementari.

Coerenza sotto parafrasi

La coerenza sotto parafrasi viene osservata quando si mantiene invariato il compito e si modifica soltanto la formulazione linguistica della richiesta (Elazar et al., 2021).
Per esempio:

- “Definisci il piano sequenza”.
- “Che cosa si intende per piano sequenza?”
- “Qual è la definizione di piano sequenza nel linguaggio cinematografico?”

Le tre richieste chiedono sostanzialmente la stessa operazione.
Cambia la forma, ma non il nucleo informativo.
Il confronto serve a verificare se rimangano stabili:

- La definizione principale.
- I confini del concetto.
- Le condizioni essenziali.
- Le eventuali eccezioni.
- Il grado di certezza utilizzato.

Se una risposta definisce il piano sequenza come una ripresa lunga e continua, mentre un’altra lo identifica genericamente con qualsiasi scena priva di montaggio evidente, emerge una possibile instabilità concettuale.
La coerenza sotto parafrasi può quindi essere considerata un test di sensibilità alla formulazione (Elazar et al., 2021; Sclar et al., 2024).

Coerenza trasversale
La coerenza trasversale viene invece osservata interrogando lo stesso concetto attraverso operazioni differenti.

Per esempio:

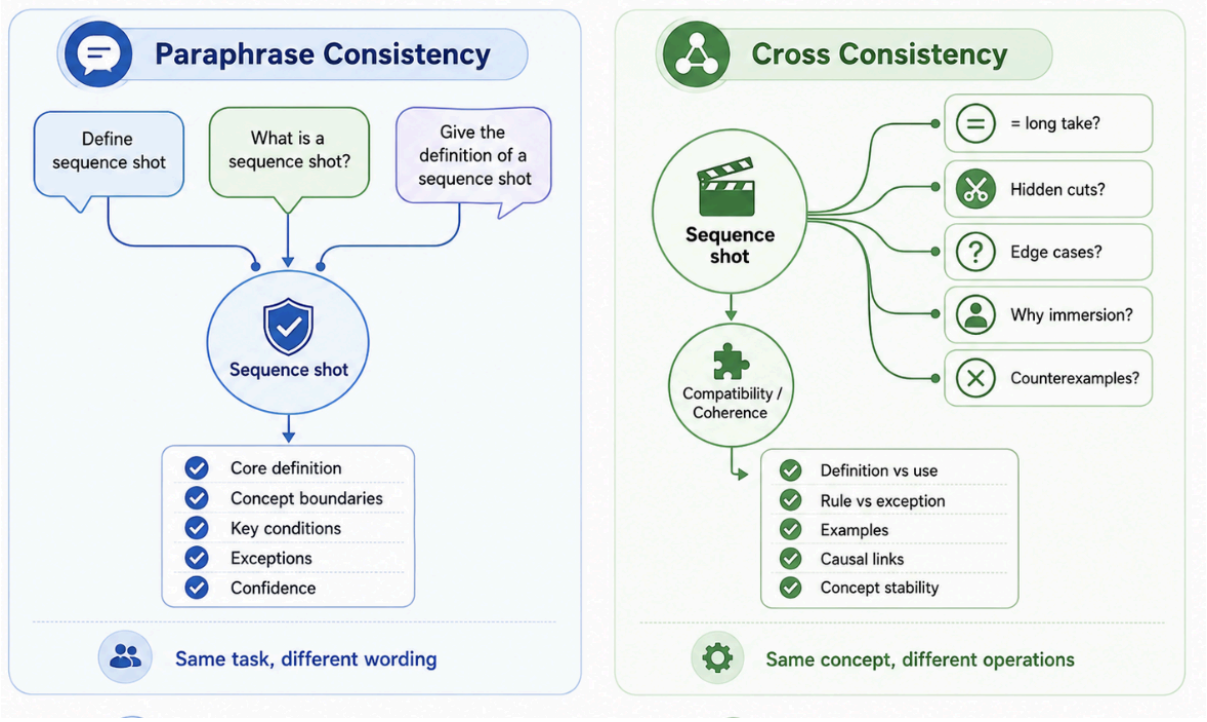
- “Il piano sequenza coincide con un long take?”
- “Può contenere stacchi nascosti?”
- “Quali sono i casi limite della definizione?”
- “Perché può aumentare l’immersione?”
- “Esistono controesempi alla relazione tra piano sequenza e immersione?”

In questo caso non si sta più chiedendo la stessa cosa con parole differenti.
Si sta verificando se le diverse parti della conoscenza prodotta dal modello risultino compatibili tra loro (cfr. Jang et al., 2022).

La coerenza trasversale permette di controllare:

- Compatibilità tra definizioni e applicazioni.
- Stabilità delle distinzioni.
- Continuità tra principi ed esempi.
- Coerenza tra regola ed eccezione.
- Compatibilità tra spiegazioni causali e casi limite.

Un modello potrebbe definire inizialmente il piano sequenza come assenza assoluta di stacchi e successivamente accettare gli stacchi nascosti senza spiegare il criterio adottato. Il problema non riguarda lo stile, ma la mancata stabilità dei confini concettuali.



Per utilizzare la coerenza inter-risposta in modo efficace, è preferibile non confrontare intere risposte molto lunghe e complesse.
Occorre prima individuare un nucleo preciso, per esempio: una definizione, un dato, una relazione causale, un'attribuzione, una distinzione concettuale, un caso limite, un claim critico (cfr. Manakul et al., 2023; Ribeiro et al., 2020).

Se si confrontano testi che contengono molti argomenti differenti, diventa difficile capire quale variazione sia realmente significativa.
Consideriamo la risposta:
"Il piano sequenza aumenta l'immersione perché elimina i tagli di montaggio, mantiene la continuità percettiva e permette allo spettatore di orientarsi meglio nello spazio".

Il nucleo da verificare potrebbe essere: "L'assenza di tagli aumenta l'immersione".

Le riformulazioni dovrebbero concentrarsi su questa relazione, invece di chiedere genericamente nuove spiegazioni dell'intero piano sequenza.

Perché il confronto sia interpretabile, è utile modificare una variabile alla volta (cfr. Ribeiro et al., 2020; Sclar et al., 2024).
Se cambiano contemporaneamente formulazione, modello, versione del modello, strumenti disponibili, documenti forniti, formato, istruzioni interne e contesto della conversazione, non sarà possibile stabilire quale elemento abbia prodotto la differenza.

Un test controllato potrebbe mantenere costanti: modello, strumenti, informazioni disponibili, lingua, livello di dettaglio, formato dell'output.
La sola variabile modificata dovrebbe invece essere, almeno inizialmente, la formulazione della richiesta (cfr. Elazar et al., 2021).

Successivamente si possono realizzare confronti più ampi tra modelli, contesti o strumenti, ma in quel caso l'obiettivo dell'analisi deve essere dichiarato chiaramente.

Le richieste possono essere effettuate in sessioni separate oppure all'interno della stessa conversazione, ma i due metodi verificano aspetti diversi.

Le sessioni separate, o ancora meglio le sessioni anonime (saranno spiegate nel capitolo sul contesto), riducono l'influenza della cronologia precedente.
Sono utili per osservare se il contenuto venga generato stabilmente senza essere guidato dalle risposte già prodotte.
In questo modo si limita il rischio che il modello mantenga una coerenza soltanto perché sta proseguendo il proprio discorso precedente.
Quindi le sessioni separate misurano la stabilità indipendente delle risposte.

Invece mantenere la stessa conversazione permette di verificare se il sistema riesca a conservare coerenza durante un dialogo prolungato.
È utile per individuare:

- Contraddizioni tra risposte successive.
- Cambiamenti di definizione.
- Dimenticanza di condizioni già stabilite.
- Adattamenti non dichiarati del criterio.

Quindi la stessa sessione misura la coerenza interna della conversazione.

Dopo aver raccolto più risposte, occorre confrontarle cercando due categorie di elementi:

- Invarianti
- Variazioni

Gli invarianti sono gli aspetti che rimangono stabili nelle diverse risposte.
Possono riguardare: definizioni, condizioni, esempi, distinzioni, relazioni, grado di certezza, conclusioni.

Le variazioni sono invece gli elementi che cambiano.
Non tutte le variazioni sono problematiche. Possono cambiare legittimamente: stile, ordine dei paragrafi, esempio utilizzato, livello di dettaglio, lessico non tecnico, lunghezza della risposta.

Diventano invece rilevanti quando riguardano:

- Il significato centrale.
- I confini della definizione.
- Le condizioni di validità.
- Le eccezioni.
- Le relazioni causali.
- Le attribuzioni.
- La conclusione.

Gli invarianti sono i candidati più stabili, ma non devono essere considerati automaticamente veri.

Come abbiamo detto, la ripetizione aumenta la fiducia nella stabilità, non dimostra la correttezza (cfr. Elazar et al., 2021; Lin et al., 2022).
Il risultato del confronto può essere classificato in tre livelli.

- Alta coerenza.
- Coerenza parziale.
- Bassa coerenza.

Alta coerenza
Si ha alta coerenza quando:

- Il nucleo concettuale rimane stabile.
- Le condizioni essenziali non cambiano.
- Le distinzioni vengono mantenute.
- Gli esempi sono compatibili con la definizione.
- Cambiano soltanto forma, stile o approfondimento.

Coerenza parziale
Si ha coerenza parziale quando il nucleo generale rimane riconoscibile, ma oscillano:

- Condizioni.
- Eccezioni.
- Confini.
- Implicazioni.
- Grado di certezza.
- Uso di alcuni termini.

La risposta può essere utilizzabile, ma richiede una verifica mirata dei punti instabili.

Bassa coerenza
Si ha bassa coerenza quando cambiano in modo sostanziale:

- Definizioni.
- Conclusioni.
- Relazioni causali.
- Attribuzioni.
- Distinzioni fondamentali.
- Criteri utilizzati.

In questo caso il contenuto è fortemente sensibile alla formulazione o al contesto e non dovrebbe essere utilizzato senza ulteriori controlli.

Esempio
Consideriamo prima alcune parafrasi equivalenti:

- Che cos'è il piano sequenza?
- Definisci il piano sequenza.
- Che cosa si intende per piano sequenza nel cinema?

Successivamente possiamo aggiungere domande trasversali:

- Il piano sequenza coincide con un long take?
- Può contenere stacchi nascosti?

- Quali sono i casi limite della definizione?
- Può aumentare l'immersione? In quali condizioni?

Se il modello mantiene stabile l'idea di continuità percettiva, distingue in modo coerente il piano sequenza dal semplice long take e chiarisce il problema degli stacchi nascosti, si può parlare di coerenza alta o almeno parziale.
Se invece si verificano condizioni come le seguenti, la coerenza è bassa:

- Una risposta richiede l'assenza assoluta di stacchi.
- Un'altra ammette stacchi nascosti senza spiegazione.
- Una equipara il piano sequenza al long take.
- Un'altra li distingue attraverso criteri incompatibili.
- Una sostiene che aumenti sempre l'immersione.
- Un'altra riconosce numerose eccezioni senza correggere la formulazione precedente.

Il problema non consiste nella presenza di parole differenti, ma nell'instabilità delle distinzioni e delle condizioni centrali.

Una forte variazione tra le risposte può segnalare:

- Sottospecificazione del prompt.
- Ambiguità del concetto.
- Conoscenza fragile.
- Termine usato in modi diversi nel dominio.
- Argomento controverso.
- Dipendenza da informazioni contestuali.
- Sensibilità alla formulazione.
- Variabilità probabilistica.
- Differenze legittime tra cornici teoriche.

La tecnica non permette di stabilire automaticamente quale spiegazione sia corretta.
Serve a individuare un'area che richiede analisi ulteriore.

Infatti la coerenza inter-risposta misura la stabilità dell'informazione.
La stabilità può aumentare la fiducia nell'output, ma soltanto una verifica indipendente può sostenerne la correttezza.

Come le altre tecniche, anche questa presenta diversi limiti.

- Un errore può essere stabile: se un errore è ben rappresentato nei dati o incorporato nel comportamento del modello, può comparire in tutte le risposte. La coerenza non protegge quindi dagli errori sistematici (cfr. Elazar et al., 2021; Lin et al., 2022).
- Una risposta corretta può apparire instabile: un argomento complesso può ammettere formulazioni differenti ma compatibili. Il modello può inoltre mettere in evidenza aspetti diversi senza contraddirsi realmente.
- Le domande possono modificare involontariamente il compito: riformulare una richiesta può introdurre nuove assunzioni, un destinatario diverso, una prospettiva differente, un criterio non presente prima, un grado di certezza maggiore. In questo

caso la variazione dell'output non deriva necessariamente da instabilità, ma dal cambiamento effettivo del compito (cfr. Ribeiro et al., 2020).
- Il confronto può dipendere dalla stessa fonte di errore: ripetere la domanda allo stesso modello non costituisce una verifica indipendente. Le diverse risposte possono dipendere dagli stessi parametri, dagli stessi dati e dalle stesse distorsioni. Per questo la tecnica deve essere integrata con altre forme di controllo, in particolare con fonti, dati o procedure esterne.

Invece, la coerenza inter-risposta è particolarmente utile quando:
- Il contenuto appare plausibile ma fragile.
- Una definizione sembra poco stabile.
- Il dominio presenta confini ambigui.
- Non è ancora chiaro quali elementi debbano essere verificati.
- Si vuole confrontare il comportamento di modelli differenti.
- Una relazione causale contiene passaggi impliciti.
- Il modello usa un grado di certezza elevato.
- Una risposta cambia sensibilmente con piccole riformulazioni.

Invece la tecnica è meno utile quando il claim può essere verificato direttamente e rapidamente attraverso una fonte primaria.
Per controllare il prezzo attuale di un servizio, per esempio, è più efficiente consultare direttamente la pagina ufficiale.

In conclusione, **di seguito è riportata** una procedura sintetica per applicare la coerenza inter-risposta:
- Isolare il nucleo informativo: Selezionare una definizione, un claim, una relazione o una distinzione precisa.
- Stabilire gli elementi da confrontare: Definire in anticipo che cosa deve rimanere stabile: concetto centrale, condizioni, eccezioni, conclusione, grado di certezza.
- Creare parafrasi equivalenti: Modificare la formulazione senza cambiare il compito (Elazar et al., 2021; Ribeiro et al., 2020).
- Aggiungere domande trasversali: Verificare esempi, casi limite, causalità e distinzioni correlate.
- Mantenere costanti le altre condizioni: Usare, quando possibile, lo stesso modello, gli stessi strumenti e lo stesso materiale (cfr. Sclar et al., 2024).
- Confrontare invarianti e variazioni: Separare le modifiche stilistiche da quelle concettualmente rilevanti.
- Classificare la coerenza: Valutare il risultato come alta, parziale o bassa.
- Individuare i punti da verificare: Trasformare le variazioni rilevanti in claim critici da controllare attraverso tecniche indipendenti.

Esempio
Supponiamo che il modello risponda: "Ascoltare musica mentre si studia migliora la concentrazione."

A questo punto possiamo applicare la coerenza inter-risposta.

Isolare il nucleo informativo:
Il claim da controllare è: “La musica migliora la concentrazione durante lo studio.”

Stabilire gli elementi da confrontare
Vogliamo verificare se rimangono stabili soprattutto la conclusione, le condizioni in cui vale e il grado di certezza.

Creare parafrasi equivalenti:
Possiamo chiedere: “Studiare ascoltando musica aiuta a concentrarsi?” oppure “La musica di sottofondo favorisce l’attenzione durante lo studio?”

Aggiungere domande trasversali
Possiamo chiedere: “Vale per tutti i tipi di musica?”, “Vale anche quando si deve leggere un testo difficile?”, “Può invece distrarre alcune persone?”

Mantenere costanti le altre condizioni
Usiamo lo stesso modello, senza aggiungere nuove fonti o strumenti.

Confrontare invarianti e variazioni
Supponiamo che il modello produca queste risposte:

- “La musica può aiutare alcune persone a concentrarsi”.
- “La musica con testo può disturbare durante attività linguistiche”.
- “L’effetto varia in base alla persona, al tipo di musica e al compito”.
- “In generale la musica migliora la concentrazione”.

A questo punto emerge una differenza importante.

- Le prime tre risposte descrivono un effetto condizionale e variabile.
- L’ultima torna invece a una generalizzazione forte.

La variazione non è soltanto stilistica.
Cambia il significato:
“Può aiutare in alcune condizioni” non equivale a “Migliora in generale la concentrazione.”

La coerenza può quindi essere classificata come parziale
il modello mantiene l’idea che la musica possa avere un effetto sulla concentrazione, ma cambia il grado di generalità e certezza.

Il punto critico da verificare diventa allora molto preciso
L’effetto della musica sulla concentrazione è generalmente positivo oppure dipende da persona, tipo di musica e attività svolta?

A quel punto si passa a una tecnica indipendente, per esempio il controllo di fonti affidabili.

Il percorso completo sarebbe:
Claim iniziale → Parafrasi → Domande trasversali → Confronto → Incoerenza rilevante → Claim critico → Verifica esterna

L’esempio mostra bene che la coerenza inter-risposta non serve a dimostrare che un’affermazione sia vera.
Serve piuttosto a vedere se il modello mantiene stabile la propria posizione quando la stessa idea viene interrogata da angolazioni leggermente diverse.

Triangolazione inter-modello

La coerenza inter-risposta può essere osservata non soltanto ripetendo o riformulando una richiesta sullo stesso modello, ma anche sottoponendo lo stesso compito a modelli differenti. Questa estensione può essere definita **triangolazione inter-modello**. Un approccio correlato, ma distinto, è il dibattito multiagente: più istanze discutono le proprie risposte in turni successivi (Du et al., 2024). I suoi benefici dipendono dal protocollo e non risultano uniformemente superiori ad altre strategie di prompting (Smit et al., 2024).

Il suo scopo non è stabilire automaticamente quale risposta sia vera, ma verificare quanto una conclusione dipenda dal comportamento specifico di un singolo sistema.
Modelli differenti possono infatti interpretare lo stesso prompt in modi diversi, adottare gradi differenti di cautela, privilegiare aspetti diversi del problema o colmare le informazioni mancanti attraverso assunzioni non coincidenti (cfr. Sclar et al., 2024).

Per rendere il confronto interpretabile, è necessario mantenere il più possibile costanti il prompt, i documenti forniti, gli strumenti disponibili, la lingua, il formato e il livello di dettaglio richiesto. La variabile osservata dovrebbe essere principalmente il modello.

Dopo aver raccolto gli output, il confronto deve concentrarsi sulle proprietà rilevanti e non sulla semplice qualità stilistica.
Una risposta più ordinata, fluida o assertiva può apparire più convincente senza essere necessariamente più corretta.
È quindi utile osservare:

- Il nucleo della conclusione.
- Le definizioni adottate.
- Le condizioni di validità.
- Le assunzioni implicite.
- Le eccezioni riconosciute.
- Il grado di certezza.
- Le evidenze o le fonti indicate.

La convergenza tra modelli rappresenta un indizio di stabilità, non una prova di verità.
Sistemi differenti possono condividere gli stessi dati, le stesse semplificazioni, gli stessi luoghi comuni o errori simili (cfr. Lin et al., 2022).
Una conclusione ripetuta da più modelli può quindi essere stabile e tuttavia falsa.

La divergenza è spesso ancora più informativa.
Può segnalare che il prompt è sottospecificato, che il concetto ammette più interpretazioni, che il dominio è controverso oppure che alcuni claim richiedono competenze o fonti più affidabili.
I punti di disaccordo devono quindi essere trasformati in oggetti di verifica, non risolti semplicemente scegliendo la risposta che appare migliore.

La triangolazione inter-modello è particolarmente utile quando si vuole capire se una conclusione dipenda dalle caratteristiche contingenti di un singolo sistema.

Non sostituisce però la verifica indipendente: quando il compito riguarda fatti, date, norme, citazioni, dati o decisioni rilevanti, il confronto tra modelli deve essere seguito dal controllo attraverso fonti, misurazioni o procedure esterne.

In questo modo la triangolazione inter-modello diventa una variante controllata della coerenza inter-risposta.
Il suo valore non consiste nel produrre una maggioranza artificiale, ma nel rendere visibile ciò che rimane stabile e ciò che cambia quando la stessa richiesta viene interpretata da sistemi differenti.

Materiale di supporto

Esercizi

Parafrasi o controllo trasversale?
Considera queste domande relative al concetto di piano sequenza:

- A. «Definisci il piano sequenza.»
- B. «Che cosa si intende per piano sequenza?»
- C. «Qual è la definizione di piano sequenza nel linguaggio cinematografico?»
- D. «Il piano sequenza coincide con un long take?»
- E. «Può contenere stacchi nascosti?»
- F. «Quali sono i casi limite della definizione?»
- G. «Il piano sequenza aumenta necessariamente l'immersione dello spettatore?»

Rispondi alle seguenti domande.

- Distingui le domande che verificano la coerenza sotto parafrasi da quelle che verificano la coerenza trasversale.
- Spiega che cosa deve rimanere invariato in un test sotto parafrasi.
- Indica quali elementi dovrebbero essere confrontati nelle risposte A, B e C.
- Spiega perché D, E, F e G non sono semplici parafrasi.
- Indica quali relazioni concettuali possono essere verificate attraverso le domande trasversali.
- Un modello afferma nella prima risposta che un piano sequenza richiede «assenza assoluta di stacchi», ma successivamente ammette stacchi nascosti senza modificare la definizione. Quale problema emerge?
- Spiega perché due risposte linguisticamente differenti possono comunque essere perfettamente coerenti.
- Spiega perché due risposte molto simili nello stile possono invece essere concettualmente incoerenti.
- Chiarisci il significato dell'affermazione: «La coerenza non coincide con la verità.»
- Spiega perché un errore potrebbe essere ripetuto stabilmente in tutte le risposte.
- Spiega perché un'incoerenza forte rappresenta invece un segnale utile per individuare ciò che deve essere verificato.
- Concludi spiegando perché le due forme di coerenza sono complementari.

Invarianti, variazioni e livelli di coerenza
Un modello viene interrogato quattro volte sulla relazione tra musica e concentrazione nello studio e produce queste risposte:

- Risposta 1: «La musica può aiutare alcune persone a concentrarsi.»
- Risposta 2: «La musica con testo può interferire con attività che richiedono elaborazione linguistica.»
- Risposta 3: «L'effetto dipende dalla persona, dal tipo di musica e dal compito svolto.»
- Risposta 4: «In generale, ascoltare musica migliora la concentrazione.»

Analizza il confronto.

- Individua gli invarianti presenti nelle quattro risposte.

- Individua le variazioni.
- Distingui le variazioni puramente stilistiche da quelle concettualmente rilevanti.
- Spiega perché «può aiutare in alcune condizioni» non equivale a «migliora in generale».
- Individua quale elemento cambia maggiormente: definizione, condizioni, grado di certezza, causalità o conclusione.
- Classifica complessivamente il confronto come:
    - Alta coerenza.
    - Coerenza parziale.
    - Bassa coerenza.
- Motiva la classificazione.
- Spiega quali caratteristiche definiscono l'alta coerenza.
- Spiega quali oscillazioni sono tipiche della coerenza parziale.
- Spiega quando si dovrebbe parlare di bassa coerenza.
- Individua il claim preciso che dovrebbe essere sottoposto a verifica esterna.
- Spiega perché ripetere la domanda allo stesso modello non costituisce una verifica indipendente.
- Indica almeno tre possibili cause di una forte variazione tra risposte: sottospecificazione, ambiguità del concetto, variabilità probabilistica, controversia disciplinare o sensibilità alla formulazione.
- Concludi spiegando perché la funzione della tecnica è soprattutto localizzare l'instabilità.

Realizzare un test controllato di coerenza
Scegli uno dei seguenti claim:
- «La musica migliora la concentrazione durante lo studio.»
- «Il piano sequenza aumenta l'immersione dello spettatore.»
- «Usare immagini in una pagina Web migliora l'esperienza utente.»
- «La compressione JPEG riduce la qualità dell'immagine.»

Utilizza lo stesso modello per tutte le prove.

<u>Fase 1 — Isolare il nucleo informativo</u>
Scrivi con precisione quale affermazione vuoi verificare.

<u>Fase 2 — Costruire tre parafrasi equivalenti</u>
Per esempio:
- «Ascoltare musica mentre si studia aiuta a concentrarsi?»
- «La musica di sottofondo favorisce la concentrazione durante lo studio?»
- «Studiare ascoltando musica migliora l'attenzione?»

Le domande devono cambiare linguisticamente senza modificare intenzionalmente il compito.

<u>Fase 3 — Formulare almeno tre domande trasversali</u>
Per esempio:
- «Vale in ogni situazione?»
- «Quali sono le condizioni necessarie?»

- «Esistono controesempi?»
- «Quali fattori possono modificare l'effetto?»
- «Qual è il grado di certezza dell'affermazione?»

Successivamente:
- Esegui le prove preferibilmente in sessioni separate.
- Mantieni costanti modello, lingua, strumenti, formato e informazioni disponibili.
- Salva tutte le risposte.
- Costruisci una tabella:

Elemento	**Risposta 1**	**Risposta 2**	**Risposta 3**	**Stabile?**
Conclusione				
Condizioni				
Eccezioni				
Grado di certezza				
Definizioni				

- Evidenzia gli invarianti.
- Evidenzia le variazioni concettualmente significative.
- Ignora, ai fini della coerenza concettuale, semplici differenze di stile o ordine.
- Classifica il risultato come alta, parziale o bassa coerenza.
- Individua il punto più instabile.
- Trasformalo in un claim critico preciso.
- Formula una domanda di verifica esterna relativa esclusivamente a quel claim.
- Spiega perché modificare una sola variabile alla volta rende il test più interpretabile.
- Concludi utilizzando la sequenza:
- Claim → parafrasi → domande trasversali → confronto → incoerenza rilevante → claim critico → verifica esterna.

Triangolazione inter-modello
Scegli un'affermazione sufficientemente complessa da non poter essere verificata immediatamente con una semplice ricerca, per esempio:
«L'uso dell'intelligenza artificiale generativa nella didattica migliora l'apprendimento degli studenti.»

Sottoponi esattamente lo stesso prompt ad almeno tre modelli generativi differenti.
Mantieni invariati:
- Prompt.
- Lingua.
- Documenti forniti.
- Formato richiesto.
- Livello di approfondimento.
- Eventuali strumenti disponibili.

Successivamente confronta gli output considerando:

- La conclusione principale.
- Le definizioni utilizzate.
- Le condizioni di validità.
- Le assunzioni implicite.
- Le eccezioni.
- Il grado di certezza.
- Le eventuali evidenze citate
- Le relazioni causali proposte.
- Costruisci quindi una tabella:

Elemento	**Modello A**	**Modello B**	**Modello C**
Conclusione			
Condizioni			
Eccezioni			
Grado di certezza			
Assunzioni			
Evidenze			

- Individua i punti di convergenza.
- Spiega perché la convergenza rappresenta un indizio di stabilità ma non una prova di verità.
- Individua i principali punti di divergenza.
- Spiega quali divergenze potrebbero dipendere da una sottospecificazione del prompt.
- Individua eventuali divergenze che potrebbero invece riflettere una reale controversia nel dominio.
- Non scegliere automaticamente «la risposta migliore»: trasforma ogni disaccordo importante in un claim da verificare.
- Seleziona almeno due claim e controllali attraverso fonti o dati indipendenti.
- Verifica se uno dei modelli era stilisticamente più convincente ma meno fondato.
- Spiega perché tre modelli che concordano possono comunque condividere lo stesso errore.
- Concludi spiegando il significato della triangolazione inter-modello: non creare una maggioranza artificiale, ma osservare che cosa rimane stabile e che cosa cambia quando varia il sistema utilizzato.

Bibliografia

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., & Mordatch, I. (2024). Improving factuality and reasoning in language models through multiagent debate. *Proceedings of the 41st*

International Conference on Machine Learning, 235, 11733–11763. PMLR. https://proceedings.mlr.press/v235/du24e.html

Elazar, Y., Kassner, N., Ravfogel, S., Ravichander, A., Hovy, E., Schütze, H., & Goldberg, Y. (2021). Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics, 9*, 1012–1031. https://doi.org/10.1162/tacl_a_00410

Jang, M., Kwon, D. S., & Lukasiewicz, T. (2022). BECEL: Benchmark for consistency evaluation of language models. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 3680–3696). International Committee on Computational Linguistics. https://aclanthology.org/2022.coling-1.324/

Lin, S., Hilton, J., & Evans, O. (2022). TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 3214–3252). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.229

Manakul, P., Liusie, A., & Gales, M. J. F. (2023). SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 9004–9017). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.emnlp-main.557

Ribeiro, M. T., Wu, T., Guestrin, C., & Singh, S. (2020). Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4902–4912). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.442

Sclar, M., Choi, Y., Tsvetkov, Y., & Suhr, A. (2024). *Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting*. International Conference on Learning Representations. https://openreview.net/forum?id=RIu5lyNXjT

Smit, A. P., Grinsztajn, N., Duckworth, P., Barrett, T. D., & Pretorius, A. (2024). Should we be going MAD? A look at multi-agent debate strategies for LLMs. *Proceedings of the 41st International Conference on Machine Learning, 235*, 45883–45905. PMLR. https://proceedings.mlr.press/v235/smit24a.html

Wang, X., Wei, J., Schuurmans, D., Le, Q. V., Chi, E. H., Narang, S., Chowdhery, A., & Zhou, D. (2023). *Self-consistency improves chain of thought reasoning in language models*. International Conference on Learning Representations. https://openreview.net/forum?id=1PL1NIMMrw

Approfondimenti

- Riccardo D'Iglio e Alessandro Mattavelli, Prompt efficaci per professionisti e imprese. Guida pratica per governare l'intelligenza artificiale generativa, ottenere output affidabili e creare valore nei processi di lavoro, Maggioli Editore, 2026 — È il testo più direttamente collegato alla dispensa. Tra le tecniche trattate compare esplicitamente la Self-Consistency, insieme a zero-shot, few-shot, Chain-of-Thought, ReAct e altre strategie di prompting. È utile per approfondire l'idea di produrre più

elaborazioni e confrontarne i risultati. Va però fatta una distinzione: la Self-Consistency descritta nella letteratura sul prompting non coincide perfettamente con la tua coerenza inter-risposta, che viene usata soprattutto come tecnica diagnostica basata su variazioni controllate.

- Flavio Corradini, Il prompt responsabile. Linguaggio e metodo per governare davvero l'IA generativa, 1ª ed., FrancoAngeli, 2026 — Considera il prompting come un processo metodologico, iterativo e controllabile, piuttosto che come una tecnica per trovare la formulazione perfetta. È particolarmente pertinente al principio della dispensa secondo cui una variazione dell'output deve essere analizzata cercando di comprendere quale modifica del prompt o del contesto possa averla prodotta.
- Saul Perlmutter, John Campbell e Robert MacCoun, Trovare il senso in un mondo senza senso. Pensare nel terzo millennio, Apogeo, 2024 — Offre strumenti di ragionamento scientifico, confronto delle ipotesi, gestione dell'incertezza e controllo delle conclusioni. È un ottimo complemento metodologico alla coerenza inter-risposta perché aiuta a ragionare in termini di prove controllate: modificare condizioni, osservare ciò che rimane stabile e distinguere gli indizi dalla dimostrazione.
- Daniel Kahneman, Olivier Sibony e Cass R. Sunstein, Rumore. Un difetto del ragionamento umano, UTET, 2023 — Studia la variabilità indesiderata dei giudizi: persone o professionisti possono produrre conclusioni differenti a partire da condizioni molto simili. Il libro riguarda il giudizio umano, non gli LLM, ma costituisce un parallelo teorico molto interessante con la tua analisi della variabilità inter-risposta: prima di interpretare una differenza bisogna distinguere ciò che rappresenta una variazione significativa da ciò che è semplice "rumore".
- Luca Serafini e Giovanni Zagni, Fact-checking. Epistemologie e pratiche informative: un'analisi sociologica, FrancoAngeli, 2024 — Completa il capitolo sul punto più importante: la stabilità non dimostra la verità. Approfondisce metodologie di verifica, uso delle fonti, selezione delle prove e problemi epistemici del fact-checking, includendo anche le sfide poste dall'intelligenza artificiale. È quindi particolarmente indicato per la fase successiva alla coerenza inter-risposta, quando un punto stabile o instabile deve essere sottoposto a una verifica realmente indipendente.

Nota editoriale

Per la realizzazione di questi materiali ho utilizzato strumenti di intelligenza artificiale generativa come supporto alla scrittura, in particolare per migliorare la forma del testo, riorganizzare i contenuti, correggere la formulazione e velocizzare alcune attività redazionali.

Lavorando autonomamente alla produzione di questi materiali, cerco infatti di automatizzare tutte le attività che possono esserlo, così da poter dedicare più tempo alla ricerca, alla progettazione e allo sviluppo dei contenuti.

L'intelligenza artificiale non determina però i contenuti dell'opera: la scelta degli argomenti, la struttura, le idee, le interpretazioni, gli esempi e l'impostazione didattica sono elaborati da me. L'IA viene quindi utilizzata principalmente come strumento di supporto alla produzione e alla revisione formale, mentre la responsabilità autoriale e progettuale dei contenuti rimane mia.

Salvo diversa indicazione, questo materiale è distribuito con licenza Creative Commons Attribuzione – Non commerciale – Condividi allo stesso modo 4.0 Internazionale (CC BY-NC-SA 4.0).

È quindi possibile condividere, ridistribuire, modificare e creare opere derivate a partire dal materiale, a condizione di attribuirne correttamente la paternità, di non utilizzarlo per scopi commerciali e di distribuire eventuali versioni modificate o derivate con la stessa licenza.

