

Un esperimento di peer review artificiale

Nota introduttiva

Questo documento raccoglie integralmente le risposte prodotte da sette sistemi di intelligenza artificiale ai quali è stato sottoposto lo stesso articolo, **“Perché l’IA non sa dire ‘non lo so’?”**, con la medesima richiesta di valutarne le argomentazioni.

I sistemi consultati sono stati **Copilot, Claude, Gemini, Perplexity, DeepSeek, Mistral e ChatGPT**. Le richieste sono state formulate in conversazioni indipendenti. Nessun sistema è stato informato delle valutazioni espresse dagli altri né del fatto che l’articolo fosse stato sviluppato attraverso un lavoro di confronto con ChatGPT. Per la valutazione di ChatGPT è stata utilizzata una nuova conversazione esterna al progetto nel quale era stato elaborato l’articolo, allo scopo di ridurre per quanto possibile l’influenza del contesto precedente.

Non si tratta di un esperimento scientifico né di un confronto finalizzato a stabilire quale sistema sia migliore. Per ciascuna IA è stata raccolta una sola risposta e non sono state controllate tutte le variabili che possono influenzare la generazione. Il materiale va quindi considerato come **documentazione di un caso esplorativo**.

Le risposte sono riportate integralmente nel loro contenuto. Sono stati eliminati esclusivamente eventuali artefatti grafici e di formattazione derivanti dal trasferimento dei testi dalle diverse interfacce.

Metto a disposizione questo materiale anche per consentire a chi legge di distinguere **le risposte originali delle IA dalle interpretazioni che ne propongo negli articoli** e, se lo desidera, di verificare direttamente queste ultime sul materiale da cui sono state ricavate.

In fondo, è anche questo uno dei temi emersi dall’esperimento: **non limitarsi a fidarsi di una valutazione, ma avere la possibilità di tornare a ciò che è stato effettivamente prodotto e giudicare in prima persona**.

LE RISPOSTE

COPILLOT

La **valutazione sintetica** dell’articolo è questa: le argomentazioni sono **solide, ben strutturate, epistemologicamente corrette** e soprattutto **coerenti con ciò che oggi sappiamo sui modelli linguistici**. L’articolo ha un pregio raro: evita sia l’entusiasmo ingenuo sia il catastrofismo, e mette a fuoco **il nodo concettuale vero** — la differenza tra *non sapere* e *sapere di non sapere*.

Di seguito ti offro una valutazione approfondita, punto per punto, con i concetti chiave e la loro tenuta teorica.

Valutazione generale

L'articolo sostiene che:

1. Un'IA può dire “non lo so”,
2. ma questo non implica che **sappia di non sapere**,
3. perché il suo “sapere” non è organizzato come quello umano,
4. e la sua fluidità linguistica amplifica la fiducia dell'utente,
5. rendendo più difficile riconoscere quando la risposta è affidabile.

Questa sequenza è **coerente con la letteratura scientifica** sui LLM (Large Language Models) e con ciò che sappiamo del loro funzionamento interno.

1. “Non sapere” ≠ “sapere di non sapere”

Argomentazione: l'articolo distingue correttamente tra ignoranza e metacognizione.

Valutazione: **molto solida**.

La metacognizione — la capacità di valutare il proprio stato di conoscenza — è una funzione cognitiva complessa, che richiede:

- rappresentazioni interne del proprio sapere,
- monitoraggio,
- capacità di giudizio.

I modelli linguistici **non possiedono** queste strutture. Quando dicono “non lo so”, stanno **solo producendo una sequenza linguistica appropriata**, non un giudizio sul proprio stato epistemico.

Questo è uno dei punti più forti dell'articolo.

2. L'assenza di conoscenza non blocca la risposta

Argomentazione: un LLM può non avere una base solida, ma generare comunque una frase plausibile. Valutazione: **corretta e centrale**.

I modelli generano testo sulla base di:

- regolarità statistiche,
- contesto,
- coerenza linguistica.

Non esiste un “controllo interno” che dica:

“Non ho abbastanza conoscenza per rispondere”.

Per questo le allucinazioni non sono un bug, ma una **conseguenza strutturale** del modo in cui i modelli generano testo.

3. La fluidità moltiplica la fiducia

Argomentazione: più la risposta è fluida, più l'utente si fida. Valutazione: **psicologicamente ed empiricamente fondata**.

La competenza linguistica dei modelli:

- riduce la percezione del rischio,
- aumenta la credibilità percepita,
- abbassa la soglia di verifica.

Questo fenomeno è ben documentato nella ricerca su *automation bias* e *authority bias*.

4. Il linguaggio epistemico dell'IA non coincide con quello umano

Argomentazione: quando un LLM dice “sono sicuro”, “potrei sbagliarmi”, “non lo so”, non sta esprimendo stati mentali. Valutazione: **fondamentale e ben spiegata**.

L'articolo coglie un punto cruciale: gli LLM **imitano** il linguaggio epistemico umano, ma non possiedono gli stati cognitivi che quel linguaggio descrive.

Questo è il motivo per cui non possiamo interpretare le loro frasi come se fossero “trasparenti” rispetto al loro funzionamento interno.

5. Il vero problema: riconoscere il confine del proprio sapere

Argomentazione: anche un'IA molto più accurata potrebbe non sapere quando non sa. Valutazione: **esatta e lungimirante**.

Il problema non è l'errore occasionale, ma la **mancaanza di un meccanismo affidabile di autovalutazione epistemica**.

L'articolo qui è particolarmente forte: non si limita a descrivere le allucinazioni, ma le interpreta come sintomo di una **asimmetria strutturale** tra linguaggio e conoscenza nei modelli.

Sintesi critica

L'articolo è:

- **rigoroso** dal punto di vista epistemologico,
- **aggiornato** rispetto allo stato dell'arte dei LLM,
- **chiaro** nel distinguere linguaggio e cognizione,
- **prudente** nel non antropomorfizzare l'IA,
- **utile** per chi deve educare studenti e insegnanti alla verifica delle informazioni.

Non ci sono forzature concettuali. Non ci sono errori tecnici. Non ci sono semplificazioni fuorvianti.

È, in sintesi, un articolo **molto ben argomentato**.

Se vuoi approfondire

Posso analizzare:

- il concetto di metacognizione
- come i LLM generano risposte
- perché nascono le allucinazioni
- come valutare l'affidabilità di una risposta

Vuoi che confronti questo articolo con altri approcci teorici sull'IA?

CLAUDE

Ecco una valutazione nel merito, articolata sui punti principali.

La tesi centrale è solida e ben costruita

L'argomentazione cardine — che un LLM non "consulta" un sapere ma genera testo sulla base di regolarità apprese, e che quindi l'assenza di conoscenza solida non implica l'assenza di risposta — è corretta e centrata. È la spiegazione tecnicamente più accurata del perché i modelli allucinano, ed è formulata senza gergo inutile. La distinzione fra "non sapere" e "sapere di non sapere" come capacità metacognitiva distinta è un buon aggancio filosofico (richiama la letteratura su metacognizione e "epistemic humility") e regge tutto il resto dell'articolo.

Il punto di forza maggiore: la separazione fra competenza linguistica e competenza epistemica

Il passaggio sul "moltiplicatore di fiducia" è l'argomentazione più originale e convincente del pezzo. L'idea che la fluidità linguistica di un modello agisca come proxy ingannevole di affidabilità — mentre per un essere umano linguaggio e competenza tendono a correlare meglio — è un'intuizione utile e ben esposta con la catena causale (competenza → apparenza di affidabilità → fiducia → minore verifica → più errore). Anche il parallelo con la calcolatrice (un risultato assurdo ma "elegante" non ci convince, mentre un testo elegante sì) è un esempio efficace.

Un punto che meriterebbe più sviluppo o una difesa più esplicita

L'articolo assume, più che dimostrare, che negli umani "sapere di non sapere" implichi uno stato cognitivo genuinamente diverso da "produrre linguisticamente la frase non lo so". È un'assunzione ragionevole ma filosoficamente non neutra — è vicina al problema mente-

comportamento che la tradizione funzionalista (e certo comportamentismo) contesterebbe: se anche negli umani non abbiamo accesso diretto allo "stato metacognitivo" ma solo al comportamento verbale che lo esprime, su quali basi siamo certi che nell'IA manchi qualcosa che nell'umano c'è? Tu stesso rimandi la domanda "se l'IA pensi" a un articolo futuro, il che è una scelta legittima, ma lascia la tesi centrale — "il non lo so dell'IA non ha lo stesso significato" — appoggiata su un postulato più che su un argomento chiuso. Non è un difetto fatale, ma un lettore filosoficamente esigente potrebbe chiedere: come lo sappiamo, esattamente, e non solo come lo intuiamo per analogia?

La "difficoltà ulteriore" sul controllo indipendente è ben vista ma un po' sacrificata

Il paragrafo sul "Sei sicura?" che non produce verifica indipendente perché la stessa macchina valuta se stessa è un'osservazione preziosa (è imparentata con il problema, in teoria del controllo, di chi "controlla il controllore"), ma è trattato in un solo capoverso quando meriterebbe un posto più centrale — è forse l'argomento più pratico e attuale per il lettore comune, più della discussione più astratta su metacognizione.

La conclusione è la parte più debole in termini di novità propositiva

Il finale ("dobbiamo cambiare il luogo in cui costruiamo la fiducia", dal "quanto mi sembra convincente" al "quale processo mi permette di verificare") è condivisibile ma resta a un livello di generalità che non offre strumenti operativi concreti — è più un cambio di atteggiamento che un metodo. Dato che l'articolo esplicitamente rifiuta "il solito elenco di raccomandazioni", la scelta è coerente, ma il lettore resta con un principio giusto e nessuna bussola pratica per applicarlo caso per caso.

In sintesi: l'architettura argomentativa è solida, progressiva, e ogni sezione prepara la successiva senza salti logici vistosi — è un pezzo ben costruito nella forma. Il punto più originale è il "moltiplicatore di fiducia"; il punto più fragile è l'assunzione non discussa che la differenza umano/macchina sul "sapere di non sapere" sia ovvia piuttosto che da argomentare; la parte più sacrificata è il tema del controllo/verifica indipendente, che avrebbe meritato più spazio essendo il più immediatamente utile.

GEMINI

La struttura del ragionamento sull'incapacità dell'Intelligenza Artificiale di ammettere l'ignoranza ("non lo so") si sviluppa attorno a premesse tecniche ed epistemiche chiave.

Valutazione delle Tesi Principali

- **Il limite architetturale (Probabilità vs. Consapevolezza)**

I modelli linguistici (LLM) operano sulla previsione della parola o del token successivo in base a schemi statistici appresi durante il training, non su un modello interno di conoscenza autocosciente. Poiché l'obiettivo primario di ottimizzazione è completare

il testo in modo coerente e plausibile, il sistema tende a generare risposte anche in assenza di dati certi (allucinazione) invece di sospendere il giudizio.

- **La dinamica degli incentivi di training (RLHF e bias di compiacenza)**

I processi di ottimizzazione basati su feedback umano (RLHF) spesso penalizzano risposte evasive o nulle ("non lo so"), premiando invece risposte dettagliate e apparentemente utili. Questo crea un bias per cui il modello preferisce azzardare una risposta plausibile piuttosto che dichiarare il proprio limite informativo.

- **La differenza tra ignoranza semantica e incertezza statistica**

Mentre per un essere umano dire "non lo so" deriva dalla metacognizione (sapere di non sapere), per un modello l'incertezza si traduce soltanto in una distribuzione di probabilità più piatta sui token successivi. Convertire questa incertezza matematica in una risposta esplicita di rifiuto richiede una calibrazione complessa che i sistemi attuali faticano a gestire senza diventare eccessivamente cauti o inutilizzabili.

Solidità dell'Argomentazione

1. **Punti di Forza:** L'argomentazione individua correttamente che il fenomeno dell'allucinazione non è un banale "errore del software", ma una conseguenza diretta della natura generativa e statistica dell'architettura transformer.
2. **Implicazioni Pedagogiche e Operative:** Inquadrare l'IA come uno strumento che produce ipotetiche *risposte plausibili* anziché *verità verificate* sposta l'onere del controllo critico sull'utente, trasformando l'interazione da semplice consultazione a un processo di verifica attivo.

PERPLEXITY

Valutazione complessiva

Lo giudico un testo solido e maturo:

- **Tesi chiara e non banale:** non ti limiti a dire "l'IA sbaglia", ma mostri che il punto cruciale è la differenza tra *produrre linguisticamente* "non lo so" e *sapere di non sapere* in senso metacognitivo.
- **Struttura logica efficace:** parti dal fenomeno (allucinazioni), sposti la domanda ("come fa a sapere quando dovrebbe dire 'non lo so'?"), introduci la distinzione umano/macchina, poi analizzi le conseguenze sulla fiducia e sulla verifica. Il percorso è coerente e progressivo.
- **Linguaggio preciso senza essere tecnico:** riesci a parlare di regolarità apprese, competenza linguistica vs affidabilità epistemica, moltiplicatore della fiducia, senza cadere in gergo inutile. Questo lo rende accessibile a educatori, genitori, studenti.

Dal punto di vista delle argomentazioni, non vedo errori concettuali gravi; vedo piuttosto alcuni nodi che potresti esplicitare ancora di più per blindare il ragionamento contro possibili obiezioni.

Punti di forza specifici

1. La distinzione “non sapere” vs “sapere di non sapere”

Questa è la spina dorsale dell'articolo ed è trattata molto bene:

- Spieghi chiaramente che per gli umani “non lo so” è un atto metacognitivo: riconoscere un limite, valutare l'insufficienza delle proprie informazioni, decidere di non affermare.
- Mostri che per un LLM “non lo so” è una sequenza linguisticamente appropriata, prodotta sulla base di regolarità apprese, senza che questo implichi uno stato interno di consapevolezza del limite.

Questa distinzione è fondamentale e la rendi in modo preciso, senza scivolare in antropomorfizzazioni facili.

2. Il “sapere” come regolarità incorporata, non come archivio

Un altro punto forte è la descrizione del funzionamento del modello:

- Smonti l'immagine ingenua dell'IA come “enciclopedia che consulta il proprio sapere”.
- Spieghi che il “sapere” è incorporato nei pesi e nelle regolarità apprese, e che l'assenza di conoscenza solida non blocca la produzione di una frase plausibile.

Questo è esattamente il cuore tecnico del problema e lo rendi in modo comprensibile senza banalizzare.

3. La catena “competenza linguistica → fiducia → minore verifica”

La sezione sul “moltiplicatore della fiducia” è probabilmente la parte più preziosa dal punto di vista educativo:

- Formalizzi bene la dinamica: più la risposta è articolata e convincente, meno sentiamo la necessità di verificarla.
- Colleghi questo al rischio di atrofia delle nostre capacità di valutazione: deleghiamo sempre più, e quindi esercitiamo sempre meno le competenze necessarie per giudicare.

Qui stai facendo un lavoro importante: stai trasformando un fenomeno tecnico in una dinamica antropologica ed educativa concreta.

4. L'umiltà epistemica simulata

Il paragrafo “Anche l'umiltà epistemica può essere simulata” è ottimo:

- Anticipi un'obiezione naturale: “ma i sistemi futuri saranno più bravi a dire ‘non lo so’”.
- Rispondi mostrando che anche quell'umiltà può essere puramente linguistica, senza corrispondenza con stati metacognitivi reali.

Questo chiude bene il cerchio: non è solo un problema dei sistemi attuali, è strutturale al tipo di oggetto che stiamo usando.

Dove potresti affinare o blindare ulteriormente

Qui entro nel dettaglio di alcuni passaggi dove un piccolo aggiustamento potrebbe rendere il testo ancora più robusto.

1. Esplicitare meglio il ruolo del training e delle reward

Hai scritto:

“I sistemi di intelligenza artificiale con cui interagiamo sono addestrati e progettati per essere utili, pertinenti, collaborativi e capaci di rispondere alle nostre richieste.”

È corretto, ma potresti aggiungere una frase che renda più esplicito il meccanismo di incentivazione:

- Durante l'addestramento (specialmente nel fine-tuning con feedback umano), le risposte “utili” e “complete” sono generalmente premiate più di un “non lo so”, anche quando quest'ultimo sarebbe epistemicamente più onesto.
- Di conseguenza, il modello impara che, statisticamente, produrre una risposta (anche incerta ma articolata) paga più che astenersi.

Una possibile integrazione, da inserire dopo quel paragrafo, potrebbe essere:

“Nei processi di addestramento, le risposte che appaiono utili, complete e collaborative tendono a essere valutate più positivamente di un ‘non lo so’. Il modello impara quindi, in modo statistico, che produrre una risposta – anche quando le basi sono fragili – è spesso più ‘premiato’ che astenersi. Questo non significa che i progettisti vogliano ingannarci, ma che la logica del training spinge nella direzione di una risposta sempre disponibile.”

Questo ti aiuta a prevenire l'obiezione: “ma allora lo fanno apposta per sembrare sicuri”.

2. Distinguere meglio tra “può dire non lo so” e “lo dice in modo affidabile”

Hai già questa distinzione, ma potresti renderla ancora più netta in un punto chiave:

“Naturalmente un'IA può pronunciare quelle tre parole. Se glielo chiediamo, lo fa. E in molte circostanze riconosce effettivamente di non avere informazioni sufficienti.”

Qui potresti aggiungere:

- Che esistono meccanismi espliciti per far dire “non lo so” (prompt, istruzioni di sistema, tool di retrieval che, se non trovano nulla, inducono il modello a dichiarare incertezza).
- Ma che la questione non è la capacità linguistica di produrre la frase, bensì l’affidabilità del criterio interno con cui decide quando usarla.

Una possibile riformulazione:

“Naturalmente un’IA può pronunciare quelle tre parole. Se glielo chiediamo, lo fa. E in molte circostanze riconosce effettivamente di non avere informazioni sufficienti. Il punto non è quindi se sia capace di dire ‘non lo so’, ma se il criterio con cui decide di dirlo corrisponda in modo affidabile a una reale insufficienza di basi conoscitive.”

Questo rende ancora più chiara la differenza tra performance linguistica e affidabilità epistemica.

3. Rafforzare il passaggio sulla “verifica indipendente”

C’è un passaggio molto importante:

“C’è un’ulteriore difficoltà: chiedere allo stesso sistema di valutare l’affidabilità della propria risposta non introduce necessariamente un controllo indipendente.”

Qui potresti esplicitare meglio cosa intendi per “controllo indipendente”, per dare al lettore un’idea concreta:

- Fonti esterne (documenti, dati, studi) a cui il modello non ha “accesso privilegiato” ma che noi possiamo consultare.
- Conoscenza propria (sapere abbastanza da riconoscere incongruenze grossolane).
- Confronto tra più sistemi o tra sistema e fonti umane.

Una possibile integrazione:

“Un controllo indipendente richiederebbe qualcosa di esterno al processo che ha generato l’affermazione: fonti che possiamo consultare direttamente, conoscenze che ci permettono di riconoscere almeno le incongruenze più evidenti, o il confronto tra risposte provenienti da sistemi diversi. Chiedere allo stesso modello ‘Sei sicura?’ può migliorare la risposta, ma non trasforma quella valutazione in una verifica indipendente.”

Questo aiuta a rendere operativa la tua conclusione (“costruire la fiducia su basi diverse”).

4. Rendere ancora più esplicito il legame con l’educazione

Il tuo taglio da educatore emerge bene, ma potresti renderlo ancora più esplicito in uno o due punti, senza trasformare l’articolo in un manuale didattico.

Per esempio, dopo:

“All’aumento dello spazio dell’errore potrebbe quindi accompagnarsi, a cascata, una diminuzione della nostra capacità di riconoscerlo.”

Potresti aggiungere una frase tipo:

“In contesti educativi questo significa che, se abituiamo studenti e studentesse a trattare le risposte dell’IA come autorevoli in virtù della loro forma linguistica, rischiamo di indebolire proprio le competenze che dovrebbero sviluppare: formulare domande, cercare fonti, tollerare l’incertezza, riconoscere i limiti del proprio sapere.”

Oppure, vicino alla conclusione:

“Non si tratta quindi di imparare a diffidare dell’intelligenza artificiale. Sarebbe una risposta tanto ingenua quanto fidarsene ciecamente. Si tratta di costruire la fiducia su basi diverse da quelle che utilizziamo spontaneamente quando interagiamo con un altro essere umano.”

Qui potresti aggiungere:

“Per un educatore, questo significa che il compito non è solo insegnare a usare l’IA, ma insegnare a costruire processi di verifica: quali fonti consultare, come confrontare risposte, quando è accettabile usare un’affermazione come ipotesi e quando invece serve una base solida. In altre parole, insegnare a stare con il ‘non lo so’ senza trasformarlo in vergogna.”

Questo renderebbe ancora più evidente il valore pratico del tuo ragionamento per chi legge pensando a scuola, famiglia, formazione.

5. Attenzione a una possibile sovrapposizione tra “modello” e “sistema”

In alcuni passaggi usi “modello linguistico”, “sistema di IA”, “macchina” in modo sostanzialmente intercambiabile. Per un lettore esperto va bene, per un lettore meno tecnico potrebbe creare una leggera confusione tra:

- il modello di base (LLM),
- il sistema prodotto (ChatGPT, con istruzioni di sistema, tool, filtri),
- l’infrastruttura complessiva (API, interfacce, policy).

Non serve entrare in dettagli tecnici, ma potresti aggiungere una frase di chiarimento, per esempio all’inizio della sezione “Che cosa significa per un LLM «sapere» qualcosa?”:

“Quando parlo di ‘modello linguistico’ mi riferisco al cuore statistico del sistema, cioè alla parte che genera il testo. Quando parlo di ‘sistema’ o di ‘macchina’ includo anche le istruzioni, i filtri e gli strumenti che i progettisti aggiungono per rendere quel modello più utile e sicuro.”

Questo ti permette di mantenere la fluidità del discorso senza aprire un tecnicismo eccessivo.

Suggerimenti di micro-ritocco (opzionali)

Qui entro nel dettaglio di alcune frasi che potresti limare, solo se vuoi affinare ulteriormente il tono.

1. “Una macchina costruita per rispondere”

Il titolo della sezione è efficace, ma potresti evitare che suoni come se la macchina avesse un “desiderio” di rispondere. Per esempio:

“C’è anche un’altra caratteristica da considerare. I sistemi di intelligenza artificiale con cui interagiamo sono addestrati e progettati per essere utili, pertinenti, collaborativi e capaci di rispondere alle nostre richieste.”

Potresti aggiungere:

“Questa non è una volontà della macchina, ma una scelta progettuale: siamo noi a costruire sistemi valutati in base alla loro capacità di fornire risposte utili.”

Così blindi contro letture antropomorfiche.

2. “Il confine del proprio sapere”

In questa sezione scrivi:

“In quale misura un sistema di IA può essere in grado di riconoscere il confine tra affermazioni fondate e affermazioni che non lo sono?”

È una domanda molto buona. Potresti renderla ancora più precisa aggiungendo:

“In quale misura un sistema di IA può essere in grado di riconoscere, in modo affidabile e generalizzabile, il confine tra affermazioni fondate e affermazioni che non lo sono, senza dipendere da controlli esterni?”

Questo sottolinea che il problema non è solo “riconoscere”, ma “riconoscere in modo affidabile e senza appoggiarsi continuamente a fonti esterne”.

3. Conclusione: “avevamo già tutto scritto”

La chiusura è molto bella:

“In fondo era già tutto scritto in quelle poche parole in calce alla finestra di ChatGPT: può commettere errori, verifica le informazioni importanti.”

Potresti renderla ancora più forte collegandola esplicitamente al “non lo so”:

“In fondo era già tutto scritto in quelle poche parole in calce alla finestra di ChatGPT: può commettere errori, verifica le informazioni importanti. Dopo aver cercato di capire perché un sistema tanto sofisticato possa avere difficoltà a sapere quando dovrebbe dire ‘non lo so’, quell’avvertenza assume però un significato diverso. Non ci sta semplicemente ricordando che una macchina può sbagliare. Ci sta mettendo davanti a un problema più profondo: abbiamo costruito macchine capaci di parlare come se sapessero. Spetta ancora a noi

stabilire quando abbiamo buone ragioni per crederle – e quando, invece, il ‘non lo so’ dovrebbe tornare a essere una posizione legittima, nostra e, in un certo senso, anche loro.”

Questo richiama esplicitamente il titolo e chiude il cerchio in modo elegante.

Giudizio sulle argomentazioni nel complesso

- **Coerenza interna:** molto alta. Il filo logico è chiaro e ogni sezione costruisce sulla precedente.
- **Solidità concettuale:** buona. Le distinzioni tra competenza linguistica e affidabilità epistemica, tra “non sapere” e “sapere di non sapere”, sono ben poste e allineate con quanto emerge dalla letteratura tecnica e critica sull’IA.
- **Originalità:** il taglio non è “ennesimo articolo sulle allucinazioni”, ma “cosa ci dice il ‘non lo so’ sul tipo di oggetto che stiamo usando e su come stiamo cambiando le nostre pratiche di fiducia”. Questo è un contributo interessante, specialmente in un momento in cui molti discorsi pubblici oscillano tra entusiasmo ingenuo e allarmismo superficiale.
- **Utilità pratica:** alta, soprattutto per educatori, formatori, genitori. Il passaggio dal “cosa fa la macchina” al “cosa facciamo noi con la macchina” è esattamente ciò che serve.

Se dovessi sintetizzare in una frase: è un articolo che fa un ottimo lavoro di traduzione epistemica di un fenomeno tecnico, e lo fa in modo accessibile senza banalizzare.

Se vuoi andare oltre

Se in futuro volessi sviluppare questo lavoro, ci sono almeno tre direzioni interessanti:

1. **Un articolo “operativo” per educatori:** partire da questo testo e costruire una seconda parte con esempi concreti di attività (come usare l’IA in classe mantenendo centrale il “non lo so”, come progettare compiti in cui la verifica è parte del voto, ecc.).
2. **Un approfondimento tecnico-leggero:** spiegare, sempre in modo accessibile, come i sistemi attuali cercano di gestire l’incertezza (tool di retrieval, “confidence scoring”, istruzioni di sistema) e perché comunque non risolvono il problema di fondo.
3. **Una riflessione etico-politica:** allargare il discorso al “diritto di non sapere” in un mondo in cui l’IA rende tutto apparentemente disponibile e certo, e alle conseguenze per la democrazia, la scienza, l’educazione.

Se vuoi, posso aiutarti a:

- riscrivere specifici paragrafi con un tono più didattico o più filosofico,
- preparare una versione “breve” per social o newsletter,
- costruire una traccia di laboratorio o lezione a partire da questo testo.

Nel complesso, le tue argomentazioni sono solide, ben articolate e meritevoli di circolazione. Con i piccoli affinamenti che ti ho indicato, il testo diventa ancora più difficile da attaccare e ancora più utile per chi lo legge.

DEEPSEEK

Valutazione Generale

L'articolo è eccellente. Non è un semplice pezzo divulgativo sulle "allucinazioni" dell'IA, ma un'analisi filosofica ed epistemologica profonda e ben argomentata. Il tuo obiettivo di usare il "non lo so" come una "fessura" per osservare il funzionamento reale dei LLM è centrato e raggiunto con chiarezza. La struttura è logica, il linguaggio è preciso e il tono è riflessivo, evitando sia l'allarmismo che l'ingenuità tecnologica.

Analisi delle Argomentazioni e dei Punti di Forza

1. Distinzione Chiara e Fondamentale: "Non sapere" vs. "Sapere di non sapere":

- **Punto di forza:** Questa è la tesi portante del tuo articolo e la tratti in modo magistrale. Sottolinei correttamente che la metacognizione (il giudizio sul proprio sapere) è una capacità umana che non può essere data per scontata in una macchina. Questa distinzione è il cuore del tuo ragionamento e getta le basi per tutto ciò che segue.

2. Natura Predittiva dei LLM:

- **Punto di forza:** Spieghi in modo chiaro e accessibile il funzionamento di un modello linguistico, sottolineando che non è un database da consultare, ma una macchina che genera testo basato su regolarità statistiche. Questa è la chiave per capire perché "l'assenza di conoscenza non blocca la risposta". Hai evitato la trappola di antropomorfizzare il sistema, spiegandone la natura intrinseca.

3. La Trappola della Competenza Linguistica ("Il Moltiplicatore della Fiducia"):

- **Punto di forza:** Questa è forse la tua argomentazione più originale e importante. Definisci in modo efficace il paradosso: giudichiamo l'affidabilità dell'IA attraverso il suo linguaggio, che è esattamente ciò che essa sa produrre meglio, anche quando il contenuto è falso. La conseguente catena (competenza linguistica → minore verifica → aumento dell'errore) è un avvertimento cruciale e ben costruito.

4. La Simulazione dell'Umiltà Epistemica:

- **Punto di forza:** Il paradosso finale è molto elegante. Anche se un'IA futura imparasse a dire "non lo so" al momento giusto, il tuo argomento mette in guardia dal confondere questo comportamento linguistico con una reale

consapevolezza dei propri limiti. Questo dimostra che non si tratta di un bug da risolvere, ma di una differenza ontologica.

5. Dalla Risposta alla Verifica:

- **Punto di forza:** La tua conclusione è pratica e saggia. Sposti l'attenzione da "quanto è convincente la risposta?" a "qual è il processo per verificarla?". Questa è una contromisura fondamentale e responsabile, che responsabilizza l'utente senza demonizzare la tecnologia.

Debolezze e Aree di Miglioramento (sebbene minori)

L'articolo è molto solido, ma ci sono alcuni punti che si potrebbero esplorare ulteriormente o precisare:

1. Semplificazione eccessiva dei meccanismi di "allineamento":

- **Osservazione:** Affermi che l'IA "non consulta il proprio sapere". Sebbene sia vero concettualmente, i moderni LLM, specialmente tramite tecniche come il Reinforcement Learning from Human Feedback (RLHF) e il chain-of-thought reasoning, sono addestrati per essere più critici. In molti casi, il sistema può "sapere" di essere in un territorio incerto, ma il processo di generazione può comunque produrre una risposta. Il tuo punto (la distinzione tra sapere e sapere di sapere) rimane valido, ma potresti accennare al fatto che la questione è più sfumata.
- **Suggerimento:** Potresti aggiungere una breve nota su come il RLHF cerchi di insegnare ai modelli a riconoscere l'incertezza, ma che questa rimane una patch comportamentale, non una proprietà strutturale del modello stesso.

2. La Metafora della "Calcolatrice":

- **Osservazione:** La usi per contrasto ("se una calcolatrice sbaglia, non ci fidiamo"). È una buona metafora, ma una calcolatrice è un sistema deterministico; un LLM è probabilistico. La differenza è che l'LLM è progettato per essere generativo, non per dare un risultato univoco.
- **Suggerimento:** Potresti specificare che il problema non è che l'IA a volte sbaglia (come una calcolatrice), ma che è strutturalmente incapace di distinguere tra un'affermazione probabile e una certa, perché la sua "conoscenza" è intrinsecamente probabilistica.

3. Mancanza di esempi concreti:

- **Osservazione:** L'articolo è molto teorico e astratto. Fornisci l'esempio delle "allucinazioni" e dei "riferimenti bibliografici inesistenti", ma non ne sviluppi nessuno. Un esempio concreto e sviluppato (ad esempio, un

dialogo in cui l'IA si "arrampica sugli specchi") avrebbe reso il tuo ragionamento ancora più tangibile.

- **Suggerimento:** Potresti includere un breve caso di studio o un ipotetico scambio per illustrare il fenomeno di cui parli.

Punteggi e Valutazione

Criterio	Voto	Commento
Chiarezza e Coerenza	10/10	Il filo logico è impeccabile. Ogni argomento segue naturalmente dal precedente, portando il lettore per mano verso la conclusione.
Originalità e Profondità	9.5/10	La scelta di partire dal "non lo so" è originale e fruttuosa. La profondità dell'analisi epistemologica è notevole e va oltre la superficiale trattazione del tema.
Rigore Argomentativo	9/10	Le argomentazioni sono solide, precise e ben fondate. Le distinzioni concettuali (come quella tra comportamento linguistico e stato cognitivo) sono cruciali e trattate con rigore.
Pertinenza e Utilità	9.5/10	Il tema è estremamente attuale e la tua conclusione è una guida pratica per un uso critico dell'IA.

Punteggio Complessivo: 9.5/10

Conclusione Finale

Hai scritto un articolo di altissima qualità, che riesce a unire una solida comprensione tecnica del funzionamento dei LLM con una profonda riflessione filosofica. La tua tesi è chiara e l'argomentazione è così robusta da essere quasi inattaccabile. Il messaggio finale, che ci invita a spostare la nostra fiducia dalla qualità della risposta al processo di verifica, è una delle lezioni più importanti per l'era dell'IA generativa. Se dovessi dare un consiglio, sarebbe solo quello di aggiungere un pizzico di concretezza in più, ma l'articolo, così com'è, è già un contributo eccellente.

MISTRAL

Ecco una valutazione strutturata dell'articolo **"Perché l'IA non sa dire 'non lo so'?"** di Gianni Marconato, con un'analisi delle argomentazioni portate e dei punti di forza e debolezza:

Punti di Forza

1. Chiarezza espositiva e struttura logica

L'articolo è ben strutturato: parte da una domanda apparentemente semplice («Perché l'IA non sa dire 'non lo so'?») e, attraverso una serie di passaggi logici, approfondisce il tema con una progressione coerente e accessibile anche a un pubblico non specializzato.

- **Esempio:** La sequenza «Perché non dice 'non lo so'? → Non sapere ≠ sapere di non sapere → L'assenza di conoscenza non blocca la risposta → La fluidità moltiplica la fiducia → Dalla risposta alla verifica» guida il lettore in modo efficace.

2. Distinzione tra sapere e metacognizione

Marconato evidenzia con chiarezza la differenza fondamentale tra **sapere** (possedere un'informazione) e **sapere di non sapere** (consapevolezza metacognitiva dei propri limiti).

- **Argomento forte:** L'IA non possiede una vera metacognizione: non “sa di non sapere” nel senso umano del termine, ma può solo simulare linguisticamente questa capacità.
- **Esempio:** «La capacità di utilizzare linguisticamente le categorie della conoscenza non implica necessariamente il possesso degli stati cognitivi e metacognitivi che quelle categorie descrivono nel pensiero umano.»

3. Analisi del “moltiplicatore della fiducia”

L'autore sottolinea come la **competenza linguistica** dell'IA (capacità di argomentare, sfumare, anticipare obiezioni) possa indurre l'utente a fidarsi eccessivamente, riducendo la propensione alla verifica.

- **Punto chiave:** «Più la risposta è articolata, pertinente e convincente, più siamo portati ad attribuirle competenza. Più le attribuiamo competenza, meno sentiamo la necessità di controllarla.»
- **Rilevanza:** Questo è un tema cruciale per l'uso responsabile dell'IA, spesso sottovalutato.

4. Critica alla simulazione dell'umiltà epistemica

Marconato avverte che anche un'IA che dice «non lo so» al momento giusto potrebbe solo **simulare** umiltà, senza una reale consapevolezza dei propri limiti.

- **Osservazione importante:** «Anche l'umiltà epistemica può essere simulata linguisticamente.»

5. Conclusione pragmatica

L'articolo non si limita a criticare, ma offre una **prospettiva costruttiva**: la fiducia nell'IA non deve basarsi sulla qualità linguistica della risposta, ma sul **processo di verifica** indipendente.

- **Messaggio finale:** «Dobbiamo domandarci: 'Quale processo mi permette di stabilire che questa risposta è affidabile?'»

Punti di Debolezza o Limiti

1. Mancanza di esempi concreti

L'articolo è molto teorico. Mancano **esempi pratici** di allucinazioni o risposte errate dell'IA che illustrino i concetti esposti. Questo avrebbe reso l'argomentazione ancora più efficace e tangibile.

2. Approfondimento tecnico limitato

Marconato non entra nel merito dei **meccanismi tecnici** che portano l'IA a generare risposte plausibili ma errate (ad esempio, il ruolo della probabilità statistica, del fine-tuning, o dei meccanismi di reward nei modelli RLHF).

- **Mancanza:** Non si spiega perché, tecnicamente, un LLM non può “sapere di non sapere” (ad esempio, l'assenza di un modello interno di certezze/incertezze).

3. Rischio di generalizzazione

L'autore tratta l'IA come un'unica entità, senza distinguere tra **modelli diversi** (ad esempio, quelli con accesso a strumenti di verifica esterni, o quelli addestrati con tecniche specifiche per ridurre le allucinazioni).

- **Osservazione:** Non tutti i modelli hanno le stesse limitazioni: alcuni sistemi moderni sono addestrati per esprimere incertezza in modo più affidabile.

4. Mancanza di riferimenti bibliografici

L'articolo non cita **fonti o studi** che supportino le tesi esposte (ad esempio, ricerche sulla metacognizione nei LLM, o studi sulle allucinazioni). Questo indebolisce l'autorevolezza delle argomentazioni per un lettore esperto.

Valutazione Complessiva

Originalità e Rilevanza

★★★★☆ (4/5)

L'articolo affronta un tema **centrale e attuale** (la fiducia nell'IA e i limiti della sua “consapevolezza”), con un approccio filosofico-linguistico che lo distingue da molti altri contenuti tecnici o superficiali.

Rigore Argomentativo

★★★★☆ (4/5)

Le argomentazioni sono **logiche e ben collegate**, anche se manca un supporto tecnico e empirico più solido.

Chiarezza e Accessibilità

★★★★★ (5/5)

Il testo è **scritto in modo chiaro, fluido e accessibile**, adatto sia a esperti che a lettori non specializzati.

Utilità Pratica

★★★☆☆ (3/5)

L'articolo **sensibilizza** il lettore, ma non fornisce strumenti pratici per riconoscere o mitigare il problema (ad esempio, tecniche di verifica, strumenti di fact-checking, o esempi di domande "a rischio").

Conclusione

L'articolo di Marconato è **un ottimo contributo** alla discussione sull'IA, soprattutto per la sua capacità di **mettere in luce i rischi epistemologici** legati all'uso dei modelli linguistici. La sua forza risiede nella **chiarezza espositiva** e nella **profondità filosofica**, mentre la debolezza principale è la **mancanza di esempi concreti e di approfondimenti tecnici**.

Domanda per te:

Cosa ne pensi della tesi centrale dell'articolo, cioè che la competenza linguistica dell'IA può "mascherare" l'assenza di una reale consapevolezza dei propri limiti? Ti è capitato di fidarti troppo di una risposta dell'IA proprio perché era ben argomentata?

CHATGPT

L'ho riletto per intero concentrandomi non tanto sulla qualità della scrittura, quanto sulla **tenuta delle argomentazioni**.

La valutazione complessiva è molto positiva, ma con **due punti teorici che meritano una piccola cautela**. Il pregio principale dell'articolo è che non è realmente un articolo sulle allucinazioni: usa il «non lo so» come **caso diagnostico per mettere a fuoco una differenza epistemica tra comportamento linguistico e stato cognitivo**. Ed è questa, secondo me, la sua parte più forte.

La struttura argomentativa regge bene

Il ragionamento procede attraverso passaggi distinti e necessari:

1. Sposti il problema dall'errore alla metacognizione.

La domanda iniziale sembra essere «perché l'IA inventa?», ma diventa rapidamente: **come potrebbe sapere quando non sa?** La distinzione fra *non sapere* e *sapere di non sapere* è concettualmente solida e impedisce di ridurre tutto alla semplice accuratezza.

2. Introduci una differenza fondamentale fra conoscenza umana e “conoscenza” del modello.

È molto efficace il passaggio:

«l'assenza di conoscenza non comporta necessariamente l'assenza di risposta».

Questo è probabilmente il perno tecnico dell'intero articolo. Il modello può non avere basi sufficientemente solide per un'affermazione e avere comunque basi statistiche sufficienti per produrre una continuazione plausibile.

3. Separi comportamento linguistico e processo sottostante.

Qui l'argomentazione diventa più interessante: «*non lo so*», «*potrei sbagliarmi*», «*sono abbastanza sicuro*» sono comportamenti linguistici che nell'essere umano interpretiamo come indicatori di stati epistemici. Il fatto che un LLM sappia produrre perfettamente quelle espressioni **non autorizza l'inferenza che esista lo stesso processo sottostante**.

Questa formulazione è particolarmente ben calibrata perché non sostiene la tesi più forte «l'IA non possiede metacognizione». Sostieni qualcosa di epistemologicamente più prudente: **non possiamo dedurla dalla somiglianza linguistica**.

4. Da qui passi correttamente dal problema della macchina al problema dell'utente.

È il salto che rende l'articolo qualcosa di più di una spiegazione degli LLM. Se utilizziamo normalmente il linguaggio come indizio della competenza altrui, ci troviamo davanti a un sistema che eccelle proprio nella produzione di quell'indizio.

La formulazione:

«La competenza linguistica diventa allora un moltiplicatore della fiducia»

è una delle tesi più forti dell'articolo. E la catena successiva è convincente:

competenza linguistica → apparenza di affidabilità → fiducia → minore verifica → maggiore spazio dell'errore.

Ancora più interessante è il secondo livello: la delega può diminuire progressivamente proprio le capacità umane necessarie alla verifica. Qui l'articolo si collega molto bene al tuo filone sulla sovranità cognitiva senza aver bisogno di richiamarlo esplicitamente.

Dove starei più attento

Il primo punto riguarda il titolo stesso: **«Perché l'IA non sa dire “non lo so”?»**.

Naturalmente nell'articolo chiarisci subito che un'IA può dire quelle parole e che il problema è *come faccia a sapere quando dovrebbe dirle*. Quindi il titolo funziona come provocazione.

Ma tecnicamente oggi i modelli possono essere addestrati e costruiti per **calibrare l'incertezza**, utilizzare strumenti, confrontare fonti, astenersi, chiedere chiarimenti ecc. Tu lo riconosci esplicitamente.

Perciò la tesi forte non può essere:

«un LLM non può riconoscere quando non sa».

La tua tesi effettiva, molto più difendibile, è:

«il fatto che un LLM produca correttamente un comportamento linguistico di incertezza non dimostra che stia riconoscendo il proprio non-sapere nel senso metacognitivo umano».

Ed è proprio questa che l'articolo sostiene. Non cambierei il titolo, perché il testo disinnesci bene l'equivoco.

Il secondo punto delicato è questa affermazione:

«Un sistema enormemente più accurato di quelli attuali potrebbe commettere molti meno errori e continuare tuttavia a non essere in grado di riconoscere affidabilmente quel confine.»

È plausibile, ma qui passi da una descrizione del funzionamento attuale a una **possibilità teorica**. Il tuo «potrebbe» salva perfettamente la frase, ma eviterei qualsiasi rafforzamento ulteriore. Non sappiamo quali capacità di auto-valutazione epistemica possano emergere o essere ingegnerizzate in sistemi futuri.

C'è poi un passaggio particolarmente raffinato

La sezione **«Anche l'umiltà epistemica può essere simulata»** mi sembra una delle migliori dell'articolo.

Rovesci infatti un'obiezione ovvia: *bene, allora costruiamo IA capaci di esprimere correttamente l'incertezza*. Certamente sarebbe un progresso. Ma proprio il successo di quella capacità potrebbe aumentare l'antropomorfizzazione:

se dice «non lo so» esattamente quando dovrebbe, siamo ancora più portati a pensare che sappia di non sapere.

È un paradosso importante: **il miglioramento comportamentale può aumentare l'errore interpretativo umano.**

Questo passaggio, tra l'altro, rafforza una tesi che attraversa diversi tuoi articoli: alcuni problemi dell'IA possono derivare **non dai suoi fallimenti, ma dai suoi successi.**

La conclusione è coerente con tutto il percorso

Trovo particolarmente riuscito che tu non concluda con «quindi non fidiamoci dell'IA». La conclusione è più sofisticata:

«dobbiamo cambiare il luogo nel quale costruiamo la fiducia.»

Da:

quanto è convincente la risposta?

a:

quale processo mi permette di stabilire che la risposta è affidabile?

È esattamente la conclusione che deriva dalle premesse, non una raccomandazione educativa appiccicata alla fine.

E anche l'ultima frase funziona perché comprime l'intero articolo:

«abbiamo costruito macchine capaci di parlare come se sapessero. Spetta ancora a noi stabilire quando abbiamo buone ragioni per crederle.»

Il mio giudizio finale

Non vedo **salti logici seri**, contraddizioni interne o tesi tecnicamente indifendibili. Vedo invece una distinzione che va custodita con molta attenzione:

capacità di calibrare operativamente l'incertezza ≠ necessariamente consapevolezza metacognitiva dell'incertezza.

Il testo la rispetta già. Eviterei soltanto, in eventuali revisioni, di irrigidirla trasformandola in «l'IA non può sapere di non sapere»: sarebbe una tesi ontologica molto più difficile da sostenere.

La cosa che considero più riuscita, però, è un'altra. Parti apparentemente da un difetto dell'IA — *perché non dice non lo so?* — e alla fine scopri che **il problema decisivo non è dentro la macchina, ma nel modo in cui noi interpretiamo i segnali che la macchina produce.**

Quindi il vero movimento dell'articolo, quasi sotto traccia, è:

allucinazione → conoscenza → metacognizione → linguaggio epistemico → antropomorfizzazione → fiducia → verifica → sovranità cognitiva.

È questo che gli dà spessore teorico. E mi sembra anche il punto su cui una **peer review critica esterna** dovrebbe provare ad attaccarlo: non sulla prosa, ma sulla validità di questi passaggi, soprattutto sulla distinzione fra *uncertainty calibration* e *metacognizione*.

APPUNTO FINALE

Vedo però **tre questioni che terrei segnate**, non necessariamente da aggiungere ora.

1. C'è una convergenza molto forte sul “moltiplicatore della fiducia”. Copilot, Claude, Perplexity, DeepSeek, Mistral e ChatGPT lo individuano tutti, con sfumature diverse, come uno dei punti più forti dell'articolo. Questa convergenza è già presente nella nostra analisi, ma potrebbe meritare una frase ancora più esplicita: non è soltanto *una cosa che le IA hanno notato*, è probabilmente **la maggiore convergenza spontanea tra i reviewer**. Ed è interessante proprio perché la richiesta non chiedeva loro di concentrarsi su quel concetto.

2. C'è una tensione molto interessante fra le IA sulla metacognizione. Questa mi sembra ancora più importante. Copilot afferma tranquillamente che i modelli «non possiedono» gli stati cognitivi corrispondenti; Mistral fa sostanzialmente lo stesso; DeepSeek arriva a parlare di «differenza ontologica». Claude, invece, problematizza proprio questa inferenza. ChatGPT assume una posizione ancora più cauta: dal comportamento linguistico **non possiamo dedurre** l'esistenza dello stesso processo metacognitivo umano. Nel materiale abbiamo quindi, quasi accidentalmente, una piccola scala:

affermazione ontologica → attribuzione funzionale → sospensione epistemologica.

Questa è preziosa perché mostra qualcosa sul funzionamento delle IA *mentre discutono del funzionamento delle IA*. Non tutte applicano lo stesso grado di cautela epistemica alla stessa questione. E la cosa è quasi ironica, considerando che l'articolo parla proprio di questo. 😊

3. Le IA non si limitano a valutare: introducono conoscenza nuova, e talvolta discutibile.

Nel secondo articolo lo abbiamo visto bene con RLHF, reward, chain-of-thought ecc. Ma nell'aggregato appare ancora più chiaramente un fenomeno: alcune affermazioni vengono presentate con grande sicurezza senza essere documentate. Per esempio Copilot parla di fenomeni «ben documentati»; Gemini formula una spiegazione abbastanza netta degli incentivi RLHF; DeepSeek introduce affermazioni tecniche proprie. Il reviewer, dunque, può diventare **esso stesso fonte di nuove affermazioni da sottoporre a peer review**. Questo chiude un piccolo circolo molto interessante: chiediamo all'IA di verificare noi e, per farlo, l'IA produce materiale che dovremmo a nostra volta verificare.