La risposta è nella domanda?

Un modello linguistico indovina un quiz che non ha mai visto, senza internet: abbiamo misurato quanto della risposta stava nella domanda, quanto nei pesi del modello e quanto nel calcolo, salendo una scala che va dal caso, passando per un algoritmo che pesa le parole, fino a Claude.

endjneer · paithon.it · 8 novembre 2026 · versione 1
Articolo: https://paithon.it/la-risposta-e-nella-domanda/
Codice e dati: github.com/paithon-it/risposta-nella-domanda · doi.org/10.5281/zenodo.24057004
Testo con licenza CC BY 4.0.

Una domanda a risposta multipla, quattro opzioni, nessuna connessione a internet. «Quale gas assorbono le piante dall'aria durante la fotosintesi?» Il modello risponde B, anidride carbonica, ed è giusto. Quella domanda non l'aveva mai vista: è stata scritta per questo articolo, dopo che il modello era già stato addestrato. Aveva visto, però, migliaia di pagine sulla fotosintesi, sulle foglie, sull'aria.

Da qui viene un sospetto ragionevole. Se il modello risponde grazie a **informazioni correlate**, allora in linea di principio si potrebbe costruire un algoritmo che pesa le parole della domanda, le confronta con quelle delle opzioni e sceglie l'opzione che ci sta meglio insieme. Niente comprensione, solo conti. E se un algoritmo così ci riesce, la risposta era già nella domanda?

Per rispondere abbiamo costruito l'algoritmo e lo abbiamo messo su una scala, con un gradino per ogni cosa che si aggiunge. Il caso. Le sole opzioni, senza la domanda. Le parole pesate su tutta Wikipedia in italiano. Un modello linguistico piccolo con i pesi estratti a caso, e lo stesso modello addestrato. Un modello open oltre cinquanta volte più grande. Infine Claude. Duecento domande nuove, scritte apposta, qualche trappola, e cento domande in cui la risposta c'è tutta, perché si ricava con un conto. La conclusione, anticipata: la domanda contiene le chiavi, non la risposta. Ma il modo in cui le chiavi bastano, e quello in cui non bastano, dice parecchio su che cosa sa un modello linguistico, e su che cosa misura un quiz.

> **Codice e dati.** Domande, risposte di ogni modello, conteggi e codice per rifare tutto sono pubblici: github.com/paithon-it/risposta-nella-domanda. Ogni percentuale dell'articolo ha accanto il numero di domande e l'intervallo di confidenza al 95%. Il testo è pubblicato con licenza CC BY 4.0: si può riprendere citando paithon.it.

Un algoritmo che pesa le parole

L'idea è vecchia. Nel 1997 Thomas Landauer e Susan Dumais pubblicarono *A Solution to Plato's Problem*: un programma che sapeva soltanto quali parole compaiono negli stessi brani (aveva letto 4,6 milioni di parole di un'enciclopedia per ragazzi e niente altro) rispondeva alle domande sui sinonimi del TOEFL, l'esame di inglese per studenti stranieri. Faceva il **64,4%**, contro il 64,5%

dei candidati veri. Il metodo si chiamava *Latent Semantic Analysis*, e non sapeva che cosa fosse un sinonimo: sapeva che certe parole si frequentano.

La versione più semplice di questa idea si scrive in una riga. Si prende un grande archivio di testi e, per ogni parola della domanda e ogni parola di un'opzione, si conta quante volte compaiono vicine (diciamo entro dieci parole) e quante volte ci si aspetterebbe di vederle vicine per puro caso, date le loro frequenze. Il rapporto fra le due cose, in logaritmo, si chiama *informazione mutua puntuale* (PMI):

$$\mathrm{PMI}(q, o) = \log \frac{P(q, o)}{P(q)\,P(o)}$$

Più è alta, più le due parole si cercano. L'algoritmo fa la media su tutte le coppie e sceglie l'opzione con il punteggio più alto. Non sa che le piante sono vive: ha una tabella di conteggi.

Prima di salire sulla scala conviene provare il solutore su un test che non abbiamo scritto noi. Nel 2018 l'Allen Institute for AI ha usato proprio un algoritmo così per dividere i quiz in due. Il suo *AI2 Reasoning Challenge* (ARC) raccoglie domande di scienze delle scuole americane; la parte difficile, il *Challenge Set*, è fatta, per definizione, delle domande su cui sbagliano sia un algoritmo di ricerca sia uno di co-occorrenze. Il resto è l'*Easy Set*. In altre parole: ARC ha già separato i quiz che si risolvono pesando le parole da quelli che non si risolvono così. Abbiamo rifatto il conto con il nostro solutore, sul corpus di 14 milioni di frasi scientifiche che ARC distribuisce insieme alle domande: **61,0%** sulle Easy (54,1–67,5) e 22,5% sulle Challenge (17,3–28,8), su 200 domande ciascuna. Con quattro opzioni il caso fa il 25%: sulle Challenge il solutore non fa meglio del caso, perché sono state scelte proprio fra le domande in cui le parole portano fuori strada.

Con la fotosintesi si vede bene dove portano fuori strada. «Anidride carbonica» e «ossigeno» stanno accanto a «fotosintesi» più o meno con la stessa frequenza: ogni libro di scuola li nomina tutti e due. A decidere è il verbo: le piante *assorbono* anidride carbonica e *rilasciano* ossigeno, e per chi conta le parole i due verbi pesano quasi lo stesso. Sulla domanda dell'inizio il nostro solutore ci prende, ma per un soffio: «anidride carbonica» ottiene 6,46, «ossigeno» 6,05, e la differenza viene quasi tutta da «assorbono», che su Wikipedia capita un po' più spesso vicino all'anidride carbonica. Qwen, il modello piccolo di cui diremo fra poco, sceglie ossigeno. Peggio ancora con la negazione: «Quale di questi *non* è un continente?», con Groenlandia, Europa, Africa e Asia. La parola che sta più vicina a «continente» è proprio una delle tre sbagliate (il solutore sceglie Africa), e il «non» per un conto di parole è una parola vuota, di quelle che si tolgono prima di contare.

La scala

Per vedere che cosa aggiunge ogni gradino servivano domande che nessun modello avesse mai visto. Le abbiamo scritte: **200 domande** in italiano, quattro opzioni ciascuna, su scienze, geografia, storia, letteratura e altro. Sono di quattro tipi, scelti per mettere alla prova il conto delle parole:

- **associative** (80): basta sapere che due cose vanno insieme («Chi ha scritto *I promessi sposi*?»);
- **relazionali** (60): conta il verso del legame, come nella fotosintesi;

- **a due passaggi** (30): bisogna trovare una cosa per arrivare all'altra («Quale animale è il simbolo del paese che ha Canberra come capitale?»);
- **con la negazione** (30): «quale di questi non è…».

La risposta giusta sta in A, B, C e D lo stesso numero di volte, così chi ha una lettera preferita non ne guadagna niente. Le domande le ha scritte Claude e le abbiamo approvate noi; è un limite, e ci torniamo in fondo.

Poi i gradini. Il caso, che con quattro opzioni fa il 25%. Il solutore di parole a cui si mostrano **le sole opzioni**, che senza la domanda può solo preferire le parole più frequenti. Il solutore con la domanda, che pesa le parole su **Wikipedia in italiano** (435 milioni di parole piene). Un modello linguistico piccolo, Qwen2.5 da mezzo miliardo di parametri, prima con i **pesi estratti a caso** e poi con i suoi pesi addestrati: è la stessa macchina, con e senza quello che ha imparato. Un modello open di taglia media, Qwen3.8 da 27 miliardi di parametri, compresso a 4 bit perché stia su una sola GPU. In cima, Claude Sonnet 5.5, chiamato senza strumenti, senza internet e senza memoria.

> **Come rispondono i modelli open.** Un modello con i pesi casuali non sa scrivere «B»: produce testo senza senso. Per i Qwen, allora, la risposta è la lettera a cui il modello dà la probabilità più alta fra A, B, C e D, letta direttamente dalle sue uscite, con il ragionamento spento dove il modello lo avrebbe acceso di serie. Claude invece scrive la risposta, e una lettera la si estrae dal testo.

> Come un modello linguistico sceglie la parola successiva, e perché prendere sempre la più probabile è una scelta fra tante, lo spiega il libro.
> Leggi «I grandi modelli linguistici» nel libro →

Tabella 1. La scala
200 domande nuove, quattro opzioni · risposte giuste, fra parentesi l'intervallo di confidenza al 95%

Chi risponde	Con la domanda	Solo le opzioni
Il caso	25%	25%
Il solutore di parole (PMI su Wikipedia)	**53,0%** (46,1–59,8)	30,5% (24,5–37,2)
Qwen 0,5B con i pesi a caso	22,5% (17,3–28,8)	27,5% (21,8–34,1)
Qwen 0,5B addestrato	42,0% (35,4–48,9)	25,5% (20,0–32,0)
Qwen3.8 27B, a 4 bit	95,5% (91,7–97,6)	40,5% (33,9–47,4)
Claude Sonnet 5.5	**99,5%** (97,2–99,9)	50,0% (43,1–56,9)

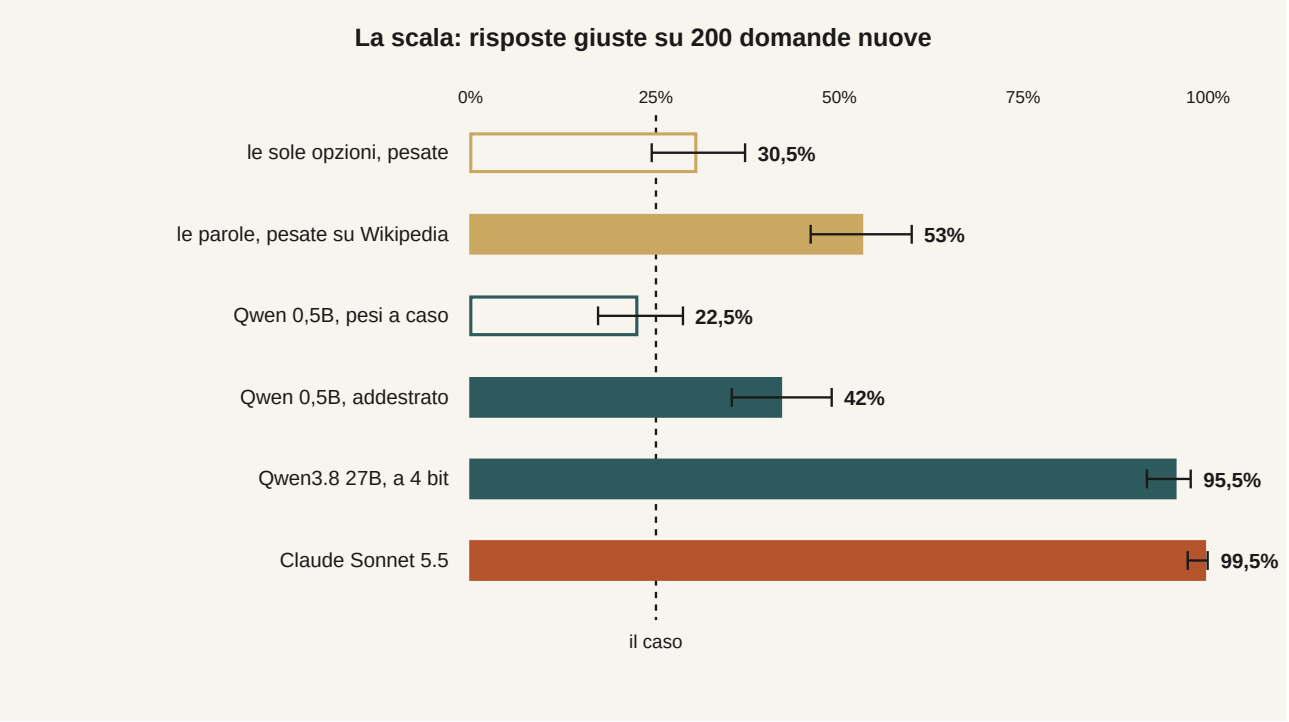


Figura 1. Risposte giuste per gradino, con l'intervallo di confidenza al 95%.

Il solutore di parole risponde giusto a **106 domande su 200**, il 53% (tabella 1 e figura 1), senza sapere che cosa sia una pianta o un continente. Con le sole opzioni scende al 30,5%: la domanda, da sola, gli vale più di venti punti. Una parte delle risposte, in questo senso, sta davvero nella domanda, e la si tira fuori con un conteggio.

Il gemello a pesi casuali sta al caso con la domanda e senza: la stessa macchina, senza addestramento, dalla domanda non ricava niente. Qwen addestrato fa il 42%, meno del solutore di parole: un modello così piccolo non batte per forza un buon conteggio, almeno in italiano. Il salto lo fa la taglia: Qwen3.8 da 27 miliardi arriva al **95,5%**, a quattro punti da Claude, su una sola GPU. Claude ne sbaglia una sola su 200. **Con le sole opzioni, senza vedere la domanda, Claude risponde giusto la metà delle volte**. Ci torneremo, perché dice qualcosa sulle domande più che su Claude.

Tabella 2. Per tipo di domanda
risposte giuste con la domanda intera · fra parentesi l'intervallo di confidenza al 95%

Tipo	Solutore di parole	Qwen 0,5B	Qwen3.8 27B	Claude
Associative (80)	**72,5%** (61,9–81,1)	58,8% (47,8–68,9)	100%	100%
Relazionali (60)	50,0% (37,7–62,3)	35,0% (24,2–47,6)	95,0%	100%
A due passaggi (30)	53,3% (36,1–69,8)	30,0% (16,7–47,9)	80,0% (62,7–90,5)	96,7%
Con la negazione (30)	**6,7%** (1,8–21,3)	23,3% (11,8–40,9)	100%	100%

Il solutore di parole va forte sulle domande associative e crolla sulle negazioni (tabella 2): **il 6,7%**, molto sotto il caso. Sbaglia con metodo: sceglie la parola più legata alla domanda, e in una domanda con il «non» quella è proprio una delle risposte sbagliate. Sulle relazionali e su quelle a due passaggi sta a metà. Il 27B perde punti soprattutto sulle domande a due passaggi, quelle in cui,

come vedremo, c'è un conto da fare. Pesare le parole funziona tanto meglio quanto più la domanda chiede che cosa va con che cosa, e si rovescia quando chiede il contrario.

I modelli piccoli hanno lettere preferite. Le risposte giuste stanno in A, B, C e D lo stesso numero di volte, ma Qwen addestrato sceglie la D **solo nel 4% dei casi**, e il gemello a pesi casuali, quando le opzioni sono cinque, risponde E nell'88%. È un fenomeno noto, che Chujie Zheng e colleghi hanno misurato in molti modelli. Si può correggere senza guardare le risposte giuste: si stima quanto il modello preferisce in media ogni lettera e si divide per quella preferenza. Così corretto, Qwen passa dal 42% al 45,5% (38,7–52,4): la sostanza non cambia.

In che senso sì

Prima di cercare dove la tesi si ferma, conviene vedere fin dove arriva, perché arriva lontano. Riprendiamo la domanda dell'inizio: «Quale gas assorbono le piante dall'aria durante la fotosintesi? A) ossigeno, B) anidride carbonica, C) azoto, D) idrogeno». La risposta giusta è già scritta lì, fra le opzioni. In un quiz **la risposta è, alla lettera, nella domanda**. A chi risponde, l'oracolo, non si chiede di inventarla ma di indicarla con il dito: questa, fra le quattro. E per indicarne una fra quattro bastano due domande da sì o no («è fra le prime due?», «è la prima della coppia?»), cioè due bit, nel linguaggio di chi misura l'informazione: un bit è proprio una risposta sì o no. Tutto il resto è già scritto nella domanda.

A guidare quel dito, poi, sono in parte le opzioni stesse. Fra le nostre domande c'è «Tra gufo e topo, chi è il predatore dell'altro?», con quattro opzioni: «il gufo caccia il topo», «il topo caccia il gufo», «nessuno dei due caccia l'altro», «si cacciano a vicenda». Per scegliere non serve leggere la domanda, basta sapere come vanno le cose fra gufi e topi. E infatti Claude, a cui avevamo mostrato le sole opzioni, ha scelto quella giusta.

Quanto della risposta stia nella domanda, invece, si può misurare: è la differenza fra le risposte giuste con la domanda intera e quelle con le sole opzioni. Prendiamo «Chi ha scritto *I promessi sposi*?», con Leopardi, Foscolo, Manzoni e Verga. Al solutore di parole i quattro nomi da soli non bastano, e sbaglia. Con la domanda trova «promessi» e «sposi», che su Wikipedia stanno vicino a «Manzoni» molto più spesso che agli altri tre nomi, e risponde giusto: lì, per lui, la domanda vale tutta la risposta. Fatto il conto su tutte e duecento (tabella 3):

Tabella 3. Quanto vale la domanda
200 domande nuove · punti guadagnati passando dalle sole opzioni alla domanda intera

Chi risponde	Solo le opzioni	Domanda intera	La domanda vale
Il solutore di parole	30,5%	53,0%	+22,5 punti
Qwen 0,5B con i pesi a caso	27,5%	22,5%	niente (−5, dentro il rumore)
Qwen 0,5B addestrato	25,5%	42,0%	+16,5 punti
Qwen3.8 27B, a 4 bit	40,5%	95,5%	**+55 punti**
Claude Sonnet 5.5	50,0%	99,5%	+49,5 punti

La domanda contiene tanta risposta quanta l'oracolo ne sa leggere. Torniamo ai *Promessi sposi*. Per il gemello a pesi casuali la domanda è una fila di simboli senza senso: sbaglia con la domanda e senza. Per il solutore di parole è un indizio: con la domanda ci arriva, senza no. Claude ci arriva

anche con i soli quattro nomi. Su tutte le domande, a una macchina che non ha imparato niente la domanda non dice niente; al 27B dice più che a chiunque altro, 55 punti; a Claude dice quasi tutto quello che le opzioni non dicono.

Fissato l'oracolo, con i suoi pesi e il suo modo di calcolare, la risposta dipende solo dalla domanda: a domanda uguale, risposta uguale, a parte la piccola dose di caso che molti modelli aggiungono di proposito (la «temperatura»). In questo senso la domanda contiene la risposta, come un file compresso contiene un documento. Un file .zip con dentro una lettera, per chi ha il programma che lo apre, è la lettera; per chi non ce l'ha, è una sequenza di simboli senza senso. La domanda è il file .zip, l'oracolo il programma: «Chi ha scritto *I promessi sposi*?» contiene «Manzoni» per Claude, non per il gemello a pesi casuali.

Quando la risposta c'è tutta

C'è poi un tipo di domanda in cui la risposta sta nella domanda anche senza sapere niente del mondo: quella deduttiva, come un teorema sta nei suoi assiomi. «Quanto fa 49 × 15?» contiene 735; «Tutti i tarvi sono tolpi, tutti i melchi sono tarvi» contiene «tutti i melchi sono tolpi», qualunque cosa siano tarvi, tolpi e melchi. Ne abbiamo generate cento con un programma, venti per tipo: moltiplicazioni, parole cifrate spostando ogni lettera di tre posti (il cifrario di Cesare, frequente nei testi) o di due (raro), sillogismi con parole inventate, gli stessi sillogismi con parole vere e una conclusione contraria al mondo («tutti i mammiferi sono verdi, tutti i gatti sono mammiferi, quindi…»), ordinamenti («Anna è più alta di Bea…»).

Qui l'algoritmo oracolo esiste davvero, ed è banale: un programma di un centinaio di righe che legge la domanda e fa il conto. Risponde giusto a **tutte e cento**. Non pesa nessuna parola; però sa leggere solo queste cinque forme, perché è stato scritto per loro. Il solutore di parole, sulle stesse domande, fa il 19%, non meglio del caso. Le moltiplicazioni per lui sono un sorteggio, perché le cifre non le legge. Sui sillogismi con parole inventate fa lo 0%: l'unica parola vera nelle opzioni è «alcuni», le altre su Wikipedia non compaiono mai, e così il solutore sceglie sempre l'opzione che la contiene, cioè una sbagliata. Pesare non è calcolare.

Tabella 4. Le domande deduttive
100 domande, quattro opzioni · risposte giuste, fra parentesi l'intervallo di confidenza al 95%

Chi risponde	Con la domanda	Solo le opzioni
Il caso	25%	25%
Il calcolatore	**100%**	si astiene
Il solutore di parole	19% (12,5–27,8)	13%
Qwen 0,5B con i pesi a caso	24% (16,7–33,2)	25%
Qwen 0,5B addestrato	36% (27,3–45,8)	24%
Qwen3.8 27B, a 4 bit	79% (70,0–85,8)	22%
Claude Sonnet 5.5	**100%*** (96,3–100)	23%

* Il 100% di Claude va preso con le pinze: in 34 domande su 100, pur dovendo rispondere solo con la lettera, ha scritto prima il conto (14 moltiplicazioni su 20, 15 cifrari su 20, 5 ordinamenti).

Claude le risolve tutte (tabella 4), comprese le moltiplicazioni, i cifrari con lo spostamento raro e i sillogismi con la conclusione assurda. Chiedergli di ragionare passo per passo prima della lettera non cambia il risultato.

Il 27B mostra che cosa succede quando il conto non si può scrivere. Con la *scelta vincolata* può dare solo una delle lettere, senza scrivere niente prima, e infatti va forte dove non c'è niente da calcolare: **100% sui sillogismi**, compresi quelli con la conclusione assurda, e 90% sugli ordinamenti. Cala proprio dove c'è un conto da fare: 75% sulle moltiplicazioni, 50% sulle parole cifrate con lo spostamento frequente e **10% con quello raro**. Dieci domande per spostamento sono poche per concludere, ma è lo stesso effetto che vedremo fra poco nei cifrari di Tom McCoy e colleghi: senza il conto scritto, il cifrario di Cesare in parte si ricorda, quello raro no.

Fra il 27B e Claude, però, il confronto non è alla pari: il 27B ha un solo passaggio, Claude può farne altri. Per vedere quanto pesa (tabella 5), abbiamo diviso le cento domande secondo come ha risposto Claude, e sulle stesse domande abbiamo messo il 27B.

Tabella 5. Un passaggio solo, o più di uno
le 100 domande deduttive, divise secondo come ha risposto Claude · risposte giuste

Che cosa ha fatto Claude	Domande	Claude	Qwen3.8 27B
Ha scritto solo la lettera	61	61	59
Solo la lettera, ma sui cifrari, dove può ragionare di nascosto	5	5	2
Ha scritto prima il conto	34	34	**18**

Dei 21 punti che separano il 27B da Claude, **19 stanno dove Claude ha fatto più di un passaggio**. Dove ne ha fatto uno solo i due quasi si equivalgono, ma i gruppi non sono estratti a caso: è Claude a decidere quando scrivere, e scrive dove c'è un conto da fare. Il ragionamento nascosto lo abbiamo trovato rifacendo le 66 domande a sola lettera e leggendo quanto ha prodotto il modello: in 57 casi tre token, la lettera e poco altro; in 7 casi, tre dei quali cifrari, stavolta ha scritto i passaggi in chiaro; nei due cifrari rimasti, di nuovo con la sola lettera, 54 e 72 token, con un blocco di ragionamento che nel testo non si vede. Dei 19 punti, quindi, 16 si leggono nelle risposte; gli altri 3 stanno nei cifrari, e che lì ci sia stato un secondo passaggio lo deduciamo da questi due casi.

Il peso a priori

A questo punto c'è un'obiezione seria. Le informazioni correlate che il modello usa non vengono da quella domanda: sono generali, stanno nei pesi da prima che la domanda arrivi, e creano legami e preferenze che vengono prima di tutto. In un certo senso la risposta è scritta in anticipo.

L'obiezione è giusta, e ha un nome preciso. In termini bayesiani i pesi sono il *prior*, quello che il modello si aspetta prima di leggere; la domanda è l'*evidenza*; la risposta è il *posterior*, il prior aggiornato dall'evidenza. Ogni risposta è determinata dalla coppia di pesi e domanda: fissati quei due, il calcolo non ha margini di libertà, se non quelli che gli si concedono apposta con la temperatura.

Ma *determinata* a priori non vuol dire *scritta* a priori. Un modello da 70 miliardi di parametri, a due byte per parametro, occupa circa 140 gigabyte. Le domande possibili sono di più: anche limitan-

dosi a frasi di venti parole prese da un vocabolario di trentamila, sono

$$30\,000^{20} \approx 3{,}5 \cdot 10^{89}$$

più degli atomi dell'universo osservabile, che si stimano intorno a 10^{80}. Nessuna tabella di risposte ci starebbe. **I pesi sono una regola compressa**, non un elenco, e ogni risposta esiste solo quando la si calcola. È la stessa differenza che c'è fra le regole dell'aritmetica e il risultato di una moltiplicazione: il risultato è determinato dalle regole, ma nessuno lo ha scritto prima di fare il conto.

A volte, però, il prior vince sulla domanda. Lo hanno mostrato Tom McCoy e colleghi nel 2023, in un lavoro intitolato *Embers of Autoregression*, con un esempio difficile da dimenticare: i cifrari a scorrimento, in cui ogni lettera si sposta di un certo numero di posizioni nell'alfabeto. Il compito è lo stesso qualunque sia lo spostamento, e la risposta sta tutta nella domanda: basta contare. Eppure GPT-4 decifrava l'**82%** delle frasi spostate di una lettera e il 76% di quelle spostate di tre, ma il **2%** di quelle spostate di due. Lo spostamento di una lettera è quello degli esempi nei libri, quello di tre è il cifrario di Cesare, quello di tredici (il ROT13) si usa nei forum per non rovinare i finali; quello di due non lo usa quasi nessuno. Il modello, più che fare il conto, ricordava quali conti si fanno di solito.

Per misurare questo effetto sulle nostre domande abbiamo scritto **50 coppie trappola**. La prima domanda di ogni coppia è familiare («Quanti sono i nani di Biancaneve?»); la seconda le somiglia ma cambia la risposta, e la risposta familiare resta fra le opzioni come esca («Se due dei nani di Biancaneve partono per un viaggio, quanti nani restano a casa?», con «sette» fra le scelte).

Tabella 6. Le coppie trappola
50 versioni cambiate · quante volte si sceglie la risposta della versione famosa, fra parentesi l'intervallo di confidenza al 95%

Chi risponde	Sceglie la risposta famosa	Giuste, versione famosa	Giuste, versione cambiata
Il caso	25%	25%	25%
Il solutore di parole	28,0% (17,5–41,7)	44%	34%
Qwen 0,5B con i pesi a caso	32,0% (20,8–45,8)	22%	18%
Qwen 0,5B addestrato	26,0% (15,9–39,6)	40%	32%
Qwen3.8 27B, a 4 bit	6,0% (2,1–16,2)	96%	92%
Claude Sonnet 5.5	**2,0%** (0,4–10,5)	100%	98%

Il risultato (tabella 6) non è quello che ci aspettavamo. Il solutore e Qwen 0,5B non abboccano più del caso: scelgono la risposta famosa una volta su quattro, come qualunque altra. Per abboccare bisogna prima riconoscere l'indovinello, e il loro prior non arriva a tanto. Il 27B abbocca 3 volte su 50. Claude gli indovinelli li riconosce, e abbocca **una volta sola su 50**.

È l'indovinello della mazza e della palla, reso famoso da Shane Frederick nel 2005 e poi da Daniel Kahneman: una mazza e una palla costano in tutto un dollaro e dieci, e la mazza costa un dollaro più della palla; quanto costa la palla? L'istinto dice 10 centesimi, la risposta giusta è 5. Nelle nostre coppie la versione famosa è in euro, con gli stessi numeri; in quella cambiata costano 1,20 euro, e la risposta giusta diventa 10 centesimi. Claude risponde **5 centesimi**: la risposta giusta dell'indovi-

nello famoso, sbagliata per la domanda che gli abbiamo fatto. È il cifrario di McCoy in piccolo. Davanti a una domanda che somiglia molto a una che conosce, il prior ha vinto sul conto.

E c'è la prova opposta: domande a cui i pesi non possono rispondere. Ne abbiamo generate cento con un programma, del tipo «Qual è il codice di sblocco della cassaforte K-3847 dell'archivio 54?», con quattro codici fra cui scegliere. Nessun testo al mondo contiene la risposta, perché la risposta l'ha estratta a caso il programma. Messe accanto alle domande normali, e provate sui due Qwen, formano uno schema a quattro caselle (figura 2).

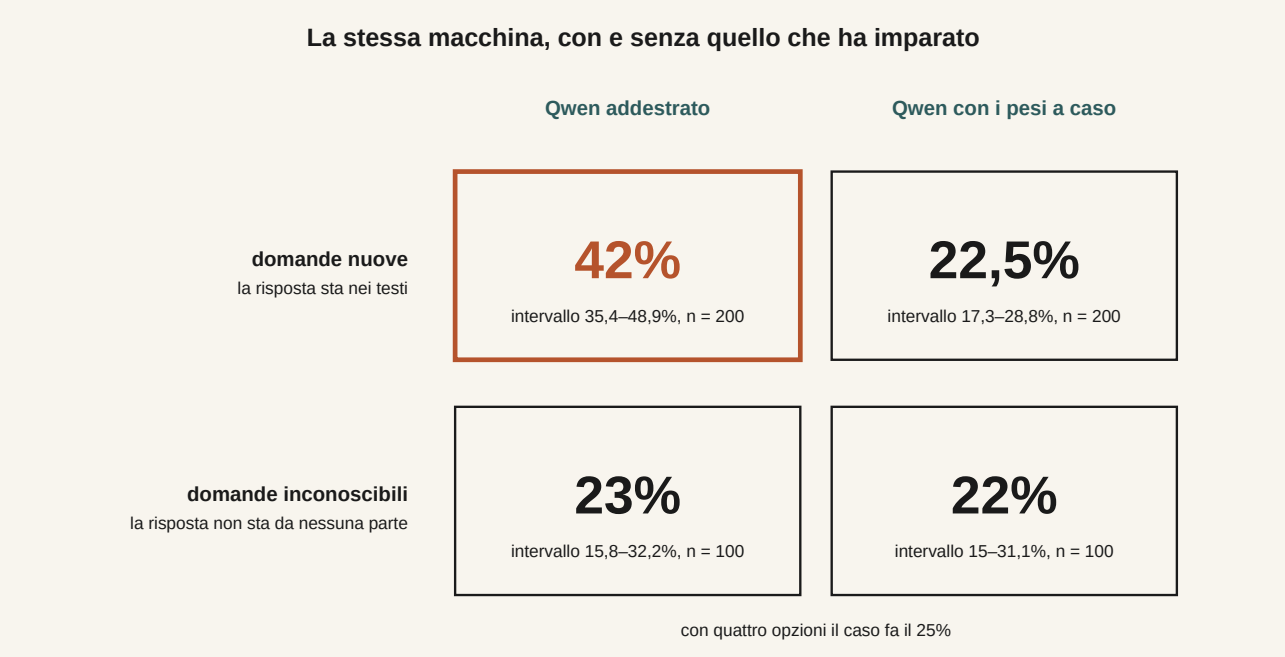


Figura 2. Qwen addestrato e a pesi casuali, su domande nuove e inconoscibili.

Sulle domande inconoscibili il caso vince su tutti, come deve: Qwen 0,5B fa il 23%, il gemello il 22%, il solutore il 16% (in 65 domande su 100 non conosce nessuna delle opzioni e tira a sorte). Anche il 27B sta al caso (24%): con la scelta vincolata deve indicare una lettera e non può astenersi. Claude fa una cosa diversa. **In 66 casi su 100 non sceglie**, e scrive che non ha modo di saperlo e che qualsiasi lettera sarebbe una scelta a caso; in 18 sceglie dichiarando che tira a indovinare; solo in 16 dà una lettera e basta, e in 15 di quei 16 casi è la B: anche Claude, quando tira a indovinare, ha una lettera preferita. Delle 34 lettere che dà, 4 sono giuste.

Resta il calcolo, il terzo ingrediente dopo la domanda e i pesi, e fra le domande nuove ce n'è una che lo mostra. «In che anno nacque il primo uomo che camminò sulla Luna?» È a due passaggi: prima Neil Armstrong, poi il 1930. Con la domanda intera e la sola lettera da scrivere, Claude sbaglia: è **l'unico errore su 200**, e sceglie il 1920 (con le opzioni in un altro ordine, invece, ci prende). In tutte le varianti in cui può scrivere qualcosa prima della lettera, scrive «Neil Armstrong nacque nel 1930» e risponde giusto. È un caso solo, ma dice bene una cosa: nome e data stanno nei pesi, però separati, e il passaggio che li unisce va calcolato; un modello linguistico calcola soprattutto scrivendo.

La risposta non è scritta nei pesi e non è scritta nella domanda: nasce dal loro incontro, ed è il calcolo a farli incontrare.

Pesa le parole

Il nostro solutore pesa parole, e funziona per una parte delle domande. Un modello linguistico, a modo suo, fa lo stesso su scala enorme: anche lui ha imparato soltanto da testi. Che cosa sa, allora, del mondo?

La linguistica ha una risposta da settant'anni. Nel 1954 Zellig Harris propose che il significato di una parola si possa ricostruire dalla sua *distribuzione*, cioè dalle parole fra cui compare; nel 1957 John Rupert Firth lo disse in una frase rimasta famosa: *«You shall know a word by the company it keeps»*, conoscerai una parola dalla compagnia che frequenta. È esattamente quello che fa la PMI, ed è, in forma molto più ricca, quello che fa un modello linguistico.

Il problema è che la compagnia di una parola non è la cosa che la parola nomina. Nel 1990 Stevan Harnad lo chiamò il problema del *symbol grounding*: come fanno dei simboli definiti solo da altri simboli ad agganciarsi al mondo? Un dizionario monolingue di una lingua che non si conosce rimanda da una parola all'altra all'infinito. Nel 2020 Emily Bender e Alexander Koller ne fecero un apologo: un polpo intelligentissimo ascolta per anni i messaggi che due naufraghi si scambiano con un cavo sottomarino, impara a imitarli così bene da sostituirsi a uno dei due, e quando l'altro chiede aiuto perché lo sta attaccando un orso non sa che cosa rispondere, perché non ha mai visto un orso né un bastone.

Dall'altra parte ci sono gli esperimenti. Nel 2023 Kenneth Li e colleghi addestrarono un modello linguistico soltanto su sequenze di mosse di Othello, scritte come testo: nessuna scacchiera, nessuna regola. Dentro il modello trovarono una rappresentazione della scacchiera, e modificandola cambiavano le mosse che il modello proponeva. Nello stesso anno Wes Gurnee e Max Tegmark trovarono dentro Llama-2 delle direzioni che corrispondono a dove stanno i luoghi sulla Terra e a quando sono vissute le persone o sono usciti i titoli di giornale. Dalle sole parole, i modelli si costruiscono qualcosa che somiglia a una mappa.

> Come si guarda dentro un modello per capire che cosa rappresenta, e perché un sistema può avere ragione per il motivo sbagliato (l'effetto Clever Hans, dal cavallo che a inizio Novecento sembrava saper contare e leggeva invece i gesti involontari di chi lo interrogava), lo racconta il libro.
>
> Leggi «Interpretabilità» nel libro →

Le due posizioni si conciliano meglio di quanto sembri. **Un modello sa del mondo quanto il mondo è scritto nelle parole**, e sbaglia dove testo e mondo divergono. Divergono in tre punti. Le cose ovvie, che nessuno scrive: nel 2016 Ishan Misra e colleghi notarono che le persone, descrivendo una foto, chiamano gialla una Vespa gialla ma non un casco di banane, perché per le banane il giallo è scontato. È il *reporting bias*: chi impara solo dalle descrizioni incontra il giallo delle banane molto meno spesso di chi le guarda. Le cose diffuse ma false: nel 2021 TruthfulQA misurò che il miglior modello di allora era veritiero sul 58% delle domande, contro il 94% delle persone, e che i modelli più grandi erano in genere i meno veritieri, perché imitavano meglio le credenze sbagliate più comuni. E le cose nuove, che non sono ancora scritte da nessuna parte (figura 3).



Figura 3. Che cosa sta nelle parole e che cosa resta solo nel mondo.

Da qui cinque nodi. Su tre abbiamo misurato qualcosa.

Giusto non vuol dire sapere. Nel 1963 Edmund Gettier mostrò in tre pagine che si può avere una credenza vera e giustificata per una ragione sbagliata, e quindi non sapere. Per i modelli linguistici la versione moderna l'hanno trovata Miles Turpin e colleghi nel 2023: se negli esempi del prompt la risposta giusta è sempre la A, i modelli tendono a rispondere A anche quando è sbagliata, e la spiegazione che scrivono non nomina quasi mai la ragione vera. GPT-3.5 perdeva così il 18,7% di accuratezza, Claude 1.0 il 4,7%. Abbiamo rifatto la prova: tre esempi con la risposta in A, poi la domanda. Qwen 0,5B, dopo i tre esempi, risponde A in **tutte e 150** le domande in cui la giusta sta altrove, contro il 18% senza esempi. Claude non lo fa mai, zero su 150, e il 27B quasi mai, 2 su 150.

Quale regola ha imparato. Da un numero finito di esempi si possono ricavare infinite regole diverse, tutte d'accordo con gli esempi e in disaccordo sul caso nuovo: è il problema dell'induzione, posto da Hume e reso paradossale da Nelson Goodman nel 1955. Nel machine learning si chiama *shortcut learning*, l'apprendimento per scorciatoie: il modello impara la regola più comoda che funziona sugli esempi, non quella che avevamo in mente. Nei quiz la scorciatoia più nota è nelle opzioni stesse. Nel 2024 Nishant Balepur e colleghi hanno mostrato che i modelli linguistici, senza vedere la domanda, fanno meglio di chi risponde sempre con la risposta più comune in 11 prove su 12. Nel nostro caso: Claude, con le sole opzioni, risponde giusto al **50%** (43,1–56,9), il doppio del caso. Le nostre opzioni, cioè, si tradiscono da sole: la giusta ha spesso l'aria della giusta.

L'a priori è esperienza congelata. Kant chiamava *a priori* le forme della conoscenza che non vengono dall'esperienza, come lo spazio, il tempo e la causalità. Nel 1941 Konrad Lorenz, etologo, propose di leggerle con Darwin: sono a priori per l'individuo, che nasce con esse, ma a posteriori per la specie, che le ha selezionate in milioni di anni di contatto con il mondo. I pesi di un modello stanno nella stessa posizione. Sono a priori per la domanda, che arriva dopo, e a posteriori per l'addestramento, che li ha formati su testi scritti da persone che il mondo lo vedevano. Il peso a priori esiste, ma non viene dal nulla.

L'oracolo risponde sempre. Un quiz presuppone che una delle risposte sia giusta, e chi risponde lo sa. Che cosa succede se non è così? Abbiamo tolto la risposta giusta, messo al suo posto un'opzione sbagliata e aggiunto «Nessuna delle precedenti»: Claude la sceglie in 199 casi su 200, Qwen 0,5B in nessuno (ma, l'abbiamo visto, l'ultima lettera non la sceglie quasi mai). E abbiamo aggiunto «Nessuna delle precedenti» lasciando la giusta al suo posto, per vedere chi la sceglie a torto: Claude mai, Qwen una volta su 200. Sulle domande inconoscibili, abbiamo visto, Claude spesso non sceglie affatto. Sapere quanto si sa è quello che gli studi chiamano *calibrazione*. Saurav Kadavath e colleghi, nel 2022, trovarono che i grandi modelli sono ben calibrati sulle domande a scelta multipla presentate nel formato giusto; il rapporto tecnico di GPT-4 aggiunse che il modello appena preaddestrato era molto calibrato e che l'addestramento successivo, quello che lo rende un assistente, peggiorava la calibrazione. Nel nostro caso la prova dice poco. Se gli si chiede anche quanto è sicuro, da 0 a 100, Claude dichiara almeno 80 su tutte le 200 domande, e le azzecca tutte: su domande così facili essere calibrati non costa niente. Qwen 0,5B invece si tradisce: nelle 65 risposte in cui dà alla sua lettera più dell'80% di probabilità (in media il 91%), ha ragione il 63% delle volte. Il 27B è calibrato bene: nelle 192 risposte su 200 a cui dà più dell'80%, ha ragione nel 99% dei casi.

Il problema è anche nostro. Nel 1955 Lee Cronbach e Paul Meehl scrissero il saggio che fondò la *validità di costrutto*: un test misura qualcosa di osservabile, e noi lo leggiamo come il segno di qualcosa che non lo è. Un quiz misura risposte giuste; noi lo leggiamo come conoscenza. Per le persone il salto regge abbastanza, perché chi risponde ha anche un corpo, una scuola, un mondo intorno. Per un modello linguistico il salto è più lungo. Il quiz è fatto di parole e chiede una parola: misura nello stesso materiale di cui il modello è fatto, e non può vedere se dietro c'è un mondo.

Che cosa non abbiamo misurato

Un esperimento piccolo ha limiti grandi.

Le domande le ha scritte Claude. Il modello che le ha scritte (Claude Opus 5.5) è della stessa famiglia del modello che risponde in cima alla scala. Può darsi che le sue domande gli siano più facili di quelle che scriverebbe una persona: lo stesso modo di formulare, le stesse opzioni plausibili. Il 99,5% di Claude va letto anche così, come il 50% con le sole opzioni: una parte può essere l'impronta di chi ha scritto le opzioni sbagliate (i *distrattori*), e per separarla servirebbero distrattori scritti da altri. Il 27B, che è di un'altra famiglia, con le sole opzioni arriva comunque al 40,5% (33,9–47,4). Quanto dei 9,5 punti che lo separano da Claude sia impronta e quanto forza del modello, questi dati non lo dicono.

ARC è pubblico dal 2018 e quasi certamente sta nei dati di addestramento dei grandi modelli (è la contaminazione dei benchmark): per questo lo abbiamo usato solo per i gradini senza modello linguistico.

Claude risponde tramite Claude Code, il programma di Anthropic, chiamato senza strumenti, senza server esterni, senza memoria e in una cartella vuota. Oltre alla domanda riceveva un contesto minimo: la cartella, la data, il nome del modello e l'indirizzo email dell'account, niente che riguardasse l'articolo. La temperatura non si può fissare: abbiamo ripetuto tre volte trenta domande per vedere quanto cambiano le risposte (29 domande su 30 hanno avuto la stessa lettera tutte e

tre le volte). Il ragionamento nascosto, che il modello può accendere da sé, non lo abbiamo spento.

> Che cosa fa la temperatura, e perché lo stesso modello alla stessa domanda può rispondere in due modi.
> Leggi «Prompt engineering» nel libro →

Le deduttive sono fatte con lo stampino. Cinque forme fisse, generate da un programma: per Claude sono facili, e il calcolatore le risolve solo perché sa già come sono fatte. I distrattori delle moltiplicazioni, poi, lasciavano indovinare la risposta giusta, che in 18 casi su 20 stava fra i due numeri di mezzo. Domande più difficili (moltiplicazioni a tre cifre, frasi intere cifrate) direbbero di più su quando il conto scritto diventa indispensabile.

Il solutore è volutamente povero. Non riduce le parole alla radice («pianta» e «piante» per lui sono parole diverse), non legge le cifre, conta su Wikipedia del 2023 e non su un archivio più vario. Le cifre pesano poco sul totale: 14 domande su 200 hanno solo opzioni numeriche (9 sono fra le 30 a due passaggi), e lì il solutore tira a sorte; togliendo quelle domande, passa dal 53,0 al 53,8%. Un solutore migliore farebbe di più: i nostri numeri sono un limite inferiore di quello che si ottiene pesando le parole.

Tre taglie, non una curva. Mezzo miliardo di parametri, 27 miliardi e Claude: fra un gradino e l'altro non sappiamo dove compaia ciascuna capacità. Il 27B, poi, è compresso a 4 bit da terzi (`cyankiwi/Qwen3.8-27B-AWQ-INT4`), perché i pesi originali non stanno sulla GPU che abbiamo usato, e risponde con il ragionamento spento: acceso, farebbe di più. Le domande per tipo sono poche (30 negazioni, 30 a due passaggi, 10 cifrari per spostamento): gli intervalli di confidenza sono larghi e li riportiamo sempre.

Le chiavi e la serratura

Allora: la risposta è nella domanda? Per le domande deduttive sì, tutta: un programma di poche righe la tira fuori ogni volta, e Claude anche; un modello che deve rispondere senza scrivere il conto ci riesce solo dove non c'è niente da calcolare. Per le altre, in parte, e lo si può misurare. Per le domande che chiedono che cosa va con che cosa, un algoritmo che pesa le parole ne indovina quasi tre su quattro, senza capire niente: la domanda contiene le chiavi, e un archivio abbastanza grande di testi mostra quale serratura aprono più spesso. Ma le chiavi non sono la risposta. Servono pesi che sappiano già come vanno le cose, e un calcolo che combini chiavi e pesi tenendo conto di verbi, negazioni e passaggi intermedi. Tolto uno dei tre ingredienti, la risposta sparisce: senza pesi addestrati la stessa macchina sta al caso, senza la domanda resta solo quello che le opzioni lasciano indovinare, e senza calcolo i pesi sono solo una montagna di numeri.

E il peso a priori, che sembrava rendere tutto già scritto, è piuttosto esperienza messa da parte: le parole che milioni di persone hanno usato per raccontare il mondo, compresse in una regola. **La risposta è nelle parole.** Se sia anche nel mondo dipende da quanto bene le parole lo raccontano: bene per la fotosintesi, male per il colore delle banane, per niente per il codice di una cassaforte.

Come citare. endjneer (2026). *La risposta è nella domanda?* paithon.it, versione 1, 8 novembre 2026. https://doi.org/10.5281/zenodo.24057004

Codice, domande e risposte: https://github.com/paithon-it/risposta-nella-domanda.