L'intelligenza si può misurare? Abbiamo rifatto sulle macchine il test che la psicometria non ha mai superato

Il QI non è una quantità, gli indici dell'AI sono panieri, e la misura più usata dell'autonomia delle macchine è una scala, non un metro. Lo dice il test di Luce e Tukey, che abbiamo fatto sui 27.119 tentativi pubblicati da METR. Che cosa si può misurare davvero, e con quale unità.

endjneer · paithon.it · 12 novembre 2026 · versione 1
Articolo: https://paithon.it/lintelligenza-si-puo-misurare-abbiamo-rifatto-sulle-macchine-il-test-che-la-psicometria-non-ha-mai-superato/
Codice e dati: github.com/paithon-it/turing-meter · doi.org/10.5281/zenodo.24153777
Testo con licenza CC BY 4.0.

Prima dell'Unità un mercante di panni che da Firenze andava a vendere a Milano doveva rifare i conti a ogni pezza. Il braccio con cui misurava la stoffa a casa sua era lungo poco più di 58 centimetri; quello di Milano quasi 59 e mezzo, quello di Bologna 64, quello di Venezia più di 68. Si chiamavano tutti allo stesso modo e nessuno era sbagliato: **erano soltanto diversi**. Il Regno d'Italia risolse il problema con la legge del 28 luglio 1861 sui pesi e sulle misure, che estese il sistema metrico decimale a tutto il paese.

Chi oggi legge le classifiche dell'intelligenza artificiale è quel mercante. Ogni laboratorio presenta il suo modello come «il più intelligente», misurato con le sue prove. Gli osservatori indipendenti pubblicano indici: a fine ottobre quello di Artificial Analysis mette in testa Gemini 4 Carbon con 59 punti, poi Claude Opus 5.5 con 58 e, a pari merito con 57, una seconda configurazione di Opus 5.5 e GPT-6 Astra. Epoch AI ne calcola un altro, su una scala senza massimo. Si chiamano tutti intelligenza, e *nessuno* è un metro.

La domanda allora è doppia. L'intelligenza si può misurare? E se sì, con quale unità, in modo che abbia senso dire, come si fa con i byte, che un sistema ne vale mille e un altro un milione? Ci siamo dati due condizioni in più: l'unità deve valere anche fra cent'anni, con tecnologie che oggi non sappiamo immaginare, e deve poter misurare, almeno in linea di principio, anche le persone. Invece di rispondere solo con argomenti, abbiamo fatto una prova. Sui dati pubblici di METR abbiamo rifatto il test che la psicometria non ha mai superato, e la risposta è netta: l'autonomia delle macchine si misura e serve già a decidere, ma non è ancora una quantità. Da lì viene la proposta, sviluppata in un pezzo gemello, «Intelligenza Apparente», di misurare un'altra cosa: non l'intelligenza della macchina, ma quanto una mente mostra di più quando lavora con lei, con un'unità

che abbiamo chiamato **turing**. Per le macchine che lavorano da sole un'unità vera non c'è ancora, ma si può dire che cosa servirebbe perché ci sia.

Prima la grandezza, poi il metro

Chi lavora in un laboratorio di metrologia comincia sempre da una domanda che sembra pedante: qual è il *misurando*? Il Vocabolario internazionale di metrologia lo definisce come la grandezza che si intende misurare. Poi c'è un secondo problema: **che cosa si può fare con i numeri** che lo strumento restituisce.

Nel 1946 lo psicologo Stanley Smith Stevens mise ordine nella questione, su *Science*, con una classificazione che si studia ancora. La scala *nominale* dice soltanto uguale o diverso, come i numeri sulle maglie dei calciatori. La scala *ordinale* dice anche prima o dopo, come l'ordine d'arrivo di una corsa. La scala *a intervalli* dice anche di quanto, ma ha uno zero scelto per convenzione, come i gradi Celsius. La scala *a rapporti* ha uno zero vero, e solo lì ha senso dire «il doppio». La differenza si vede nelle trasformazioni che ciascuna sopporta senza perdere significato:

$$\text{intervalli:}\quad x' = a\,x + b \qquad\qquad \text{rapporti:}\quad x' = a\,x \qquad (a > 0)$$

Venti gradi Celsius non sono il doppio di dieci: in Fahrenheit diventano 68 e 50, e il «doppio» sparisce. Venti kelvin invece sono il doppio di dieci in qualunque unità. Chiedere un «giga di intelligenza» vuol dire chiedere che l'intelligenza stia su una scala a rapporti, con uno zero che significa niente e quantità che si sommano (figura 1).

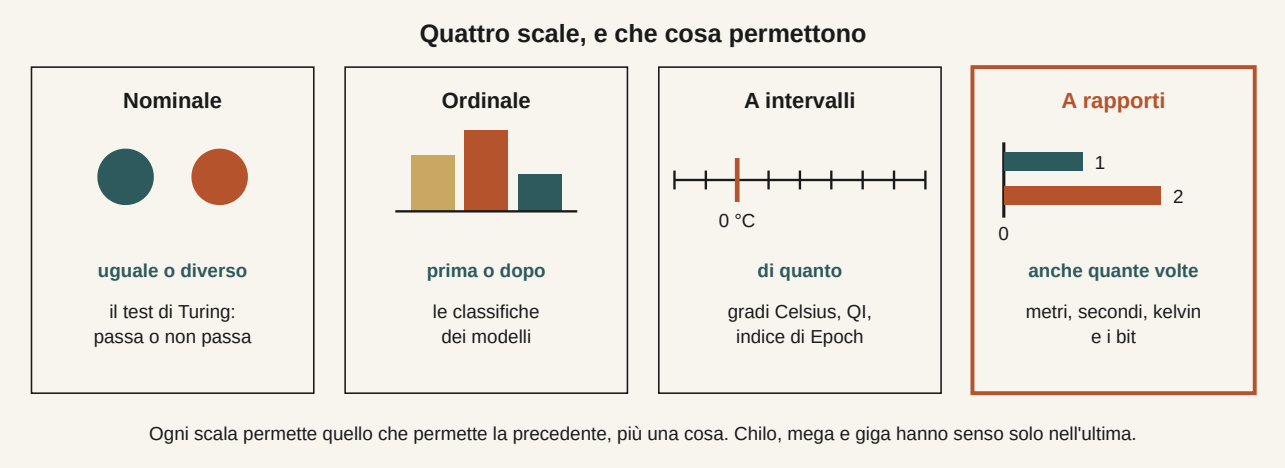


Figura 1. Le quattro scale di Stevens e che cosa permettono: chilo, mega e giga hanno senso solo nell'ultima.

Cent'anni di QI: un numero che non è una quantità

La psicologia ci ha provato per prima. Nel 1905 Alfred Binet e Théodore Simon, per una commissione del ministero francese dell'istruzione pubblica, costruirono prove per riconoscere i bambini che a scuola avevano bisogno di aiuto; dalla revisione del 1908 il risultato si esprimeva come *età mentale*. Nel 1912 William Stern propose di dividerla per l'età anagrafica, nel 1916 Lewis Terman moltiplicò per cento, e nacque il quoziente d'intelligenza:

$$\mathrm{QI} = 100 \cdot \frac{\text{età mentale}}{\text{età anagrafica}}$$

Era un rapporto fra due durate, ma un rapporto fra durate non fa dell'intelligenza una quantità, e con gli adulti si rompeva: l'età mentale smette di crescere nell'adolescenza, quella anagrafica no. Nel 1939 David Wechsler cambiò la definizione alla radice: il QI divenne una posizione nella distribuzione dei coetanei, riportata su una scala con media 100 e deviazione standard 15.

$$\mathrm{QI} = 100 + 15\,z, \qquad z = \frac{x - \mu_{\text{età}}}{\sigma_{\text{età}}}$$

Da allora il QI non dice quanta intelligenza hai: dice quante persone della tua età stanno sotto di te. Un 200, il numero che a volte si legge nei titoli, corrisponderebbe a una persona ogni 76 miliardi circa: non è una misura, è la coda di una curva prolungata oltre i dati. Negli anni Ottanta James Flynn si accorse poi che i punteggi grezzi salivano di circa **tre punti ogni dieci anni**, e che i test venivano ricalibrati di continuo per riportare la media a 100: lo zero e l'unità del QI sono quelli di una popolazione, e la popolazione cambia.

Sotto il QI c'è un'idea più solida. Nel 1904 Charles Spearman osservò che chi va bene in una prova tende ad andare bene anche nelle altre, e chiamò *g*, fattore generale, ciò che le correlazioni hanno in comune; il modello oggi più accreditato, che porta i nomi di Raymond Cattell, John Horn e John Carroll, gli mette sotto una decina di capacità ampie. Ma *g* è un riassunto statistico, non la lettura di uno strumento, e nel 1923 Edwin Boring lo scrisse con un'onestà che ancora punge: l'intelligenza «è ciò che i test misurano».

La critica più radicale è dell'australiano Joel Michell, uno studioso di teoria della misura. Fra il 1997 e il 2000 ha sostenuto che la psicometria abbia dato per scontata l'ipotesi da cui tutto dipende, cioè che l'intelligenza sia una grandezza *quantitativa*, senza mai metterla alla prova, e l'ha chiamata *scienza patologica*. Lo strumento per metterla alla prova esiste dal 1964, quando Duncan Luce e John Tukey pubblicarono la teoria della *misura congiunta*: se la bravura di chi risponde e la difficoltà di ciò che gli si chiede si combinano in modo additivo, certe relazioni d'ordine fra i risultati devono valere sempre, e si possono controllare sui dati. La psicometria ha una risposta parziale, il modello del danese Georg Rasch, che nella teoria della risposta all'item mette persone e domande sulla stessa scala. Il test di Luce e Tukey, per l'intelligenza, è rimasto in gran parte da fare. Per le macchine lo abbiamo fatto noi.

Le macchine, da Turing agli indici

Nel 1950 Alan Turing propose di sostituire la domanda «le macchine possono pensare?» con un gioco: un interrogatore conversa per iscritto con una persona e con una macchina, e deve capire chi è chi. Previde che entro una cinquantina d'anni un interrogatore medio non avrebbe avuto più del 70 per cento di probabilità di indovinare **dopo cinque minuti di domande**. È una misura su scala nominale: si passa o non si passa. Nel 2025 il test è stato superato: in un esperimento controllato, con conversazioni di cinque minuti, GPT-4.5 istruito a recitare un personaggio è stato giudicato umano nel 73 per cento dei casi. È rimasta la domanda su che cosa misurasse.

Nel 2007 Shane Legg e Marcus Hutter raccolsero una settantina di definizioni di intelligenza e le distillarono in una formula: l'*intelligenza universale* di un agente π è il risultato V che ottiene in tutti gli ambienti μ calcolabili, ciascuno pesato in base alla sua semplicità.

$$\Upsilon(\pi) = \sum_{\mu \in E} 2^{-K(\mu)}\, V_\mu^\pi$$

La complessità di Kolmogorov $K(\mu)$ è la lunghezza in bit del programma più corto che descrive l'ambiente, e il peso $2^{-K(\mu)}$ si dimezza a ogni bit in più: un mondo semplice conta più di uno complicato. Vale per qualunque agente, ma K **non è calcolabile**, e la misura ha un massimo per costruzione. Lo spagnolo José Hernández-Orallo, con David Dowe, propose nel 2010 un test *anytime* che si adatta a chi lo sostiene, pensato per sistemi artificiali e biologici; nel 2015 ha dato alla difficoltà di un compito una misura in bit, lo sforzo di ricerca necessario a trovare e verificare una procedura che lo risolve, e nel 2017 ha raccolto il suo programma in *The Measure of All Minds*. Nel 2019 François Chollet ha ricordato che l'abilità in un compito si può comprare, con conoscenze incorporate o con dati senza limite, e che l'intelligenza è un'altra cosa: l'efficienza con cui si acquisiscono abilità nuove.

Il resto è la storia di ogni benchmark. Su ARC-AGI-3, la prova interattiva nata dalle idee di Chollet, a settembre la ARC Prize Foundation ha misurato GPT-6 Astra al 62,7 per cento con l'impalcatura standard e al 99,9 per cento con un'impalcatura diversa, a un livello di ragionamento più basso e con una spesa di 19.000 dollari: lo stesso modello, due strumenti, due numeri. Una revisione di 445 benchmark per modelli linguistici, fatta da 29 revisori esperti in un lavoro guidato dall'Università di Oxford e uscito nel novembre 2025, ha trovato che solo il 16 per cento usava stime di incertezza o test statistici. E i tentativi di riportare la psicometria umana sulle macchine si sono moltiplicati. David Ilić e Gilles Gignac hanno trovato un fattore generale nei punteggi di 591 modelli; il gruppo di Hernández-Orallo ha proposto diciotto scale di «domanda cognitiva»; Dan Hendrycks e trentadue coautori hanno dato una definizione di intelligenza artificiale generale in cui GPT-4 fa il 27 per cento e GPT-5 il 57, entrambi con **zero nella memoria a lungo termine**.

Panieri e termometri

L'Intelligence Index di Artificial Analysis, nella versione 5.0, è una media pesata di nove valutazioni raggruppate in quattro categorie da un quarto ciascuna, riportata su una scala da 0 a 100. Somiglia al paniere con cui l'Istat calcola l'inflazione: utile per confrontare a paniere fermo, muto quando il paniere cambia. A settembre, nella tabella con cui OpenAI presentava GPT-6 Astra, la versione 4.1.1 dell'indice dava al modello 61,2 e a Claude Fable 5.1 65,7; nella versione 5.0 Astra fa 57 e Fable 5.1 fa 49. **L'ordine si è invertito**, e a cambiare è stato il paniere.

L'Epoch Capabilities Index di Epoch AI cuce insieme più di cinquanta benchmark con la stessa famiglia di modelli statistici della psicometria, non ha un massimo ed è dichiarato lineare. Ma è fissato da due modelli di riferimento, Claude 3.5 Sonnet a 130 e GPT-5 a 150, e i valori assoluti, scrivono gli autori, non significano niente da soli. È un termometro Celsius, con un modello vecchio e uno nuovo al posto del ghiaccio e dell'acqua che bolle.

La misura che serve già: l'orizzonte temporale

C'è una misura dell'autonomia delle macchine che non è un paniere e che qualcuno usa davvero per decidere. L'ha introdotta nel marzo 2025 METR, un'organizzazione non profit che valuta le capacità dei modelli di frontiera, e si chiama *orizzonte temporale* dei compiti. Si cronometra quanto ci mettono professionisti esperti a portare a termine compiti di programmazione, di ricerca sul machine learning e di sicurezza informatica, da pochi secondi a molte ore; poi si fanno provare gli agenti, da soli, e per ogni modello si stima con una regressione logistica come scende la probabilità di successo quando i compiti si allungano:

$$p(\text{successo}) = \sigma\big(\alpha_i + \beta_i \log_2 t\big)$$

L'orizzonte al 50 per cento è la durata t dei compiti che il modello porta a termine una volta su due. Ha uno zero vero, è espresso in un'unità che esiste già, e raddoppia con una regolarità notevole: dal 2023 in media ogni **129 giorni**, poco più di quattro mesi (tabella 1).

Tabella 1. L'orizzonte dei modelli di frontiera
METR, suite Time Horizon 1.1; compiti di software, machine learning e sicurezza informatica

Modello	Uscita	Al 50%	All'80%
GPT-4	mar. 2023	4 minuti	1 minuto
Claude 3.7 Sonnet	feb. 2025	1 ora	12 minuti
GPT-5	ago. 2025	3 ore e 23 minuti	38 minuti
Claude Opus 4.6	feb. 2026	12 ore	1 ora e 10 minuti
Gemini 3.1 Pro	feb. 2026	6 ore e 24 minuti	1 ora e 30 minuti
Claude Mythos Preview	apr. 2026	17 ore e 25 minuti	3 ore e 6 minuti

Chi la usa, e per che cosa, si può dire con precisione. **METR**, nelle valutazioni prima del rilascio, ha indicato in via molto provvisoria due soglie: un orizzonte al 50 per cento di oltre 40 ore su compiti di software e machine learning, oppure all'80 per cento di oltre 8 ore su compiti di software che richiedono molto contesto. Superarne una giustificherebbe una valutazione del rischio molto più approfondita, con prove mirate e più sforzo per far emergere le capacità del modello. L'**AI Security Institute** britannico misura con lo stesso metodo l'autonomia nella sicurezza informatica, all'80 per cento e con un tetto fisso di 2,5 milioni di token per compito, perché i risultati restino confrontabili nel tempo; nel rapporto di maggio ricorda la sua stima di febbraio, un raddoppio ogni 4,7 mesi, che i due modelli più recenti hanno già superato. L'istituto lavora con il centro nazionale per la cybersicurezza, che consiglia le imprese. Lo studioso **Tom Harris**, in un lavoro di ottobre, ha combinato le curve di METR con la durata dei compiti dei mestieri americani, e stima che i compiti che valgono il 46 per cento del monte salari degli Stati Uniti siano alla portata dell'ultimo modello, contro il 14 per cento di inizio 2025.

Altrove l'orizzonte non c'è. Le politiche di sicurezza di Anthropic, OpenAI e Google DeepMind definiscono le loro soglie per effetto (accelerare la ricerca, automatizzare un ruolo), non per durata dei compiti. L'AI Act chiede di valutare il «livello di autonomia» di un modello, ma senza un numero. Le assicurazioni e le aziende che scelgono un modello guardano all'affidabilità sul loro compito specifico, al 95 o al 99 per cento, che l'orizzonte non sa misurare. Chi decide con l'orizzonte, quindi, è chi valuta il rischio di frontiera e chi studia il lavoro.

Abbiamo fatto il test di Michell

Resta la domanda di Michell: l'autonomia misurata così si comporta come una quantità? METR pubblica, oltre agli orizzonti, i risultati di ogni modello in ogni compito: per la suite attuale sono 23 agenti, 228 compiti, **27.119 tentativi**, con il tempo umano di ciascun compito. Con questi dati abbiamo fatto tre prove.

La prima riguarda le pendenze. Il modello di METR dà a ogni agente la sua pendenza β_i nella formula. Se l'autonomia fosse una quantità alla Rasch, una pendenza comune dovrebbe bastare, e il confronto fra due agenti non dipenderebbe dalla soglia scelta. Non basta: il modello con pendenze diverse spiega i dati molto meglio, e il solo caso produrrebbe una differenza così grande con una probabilità di circa cinque su mille miliardi. Le pendenze vanno da −0,39 a −0,73 per raddoppio dei minuti nella nostra stima, da −0,41 a −0,69 in quella pubblicata da METR. In pratica, con i parametri pubblicati da METR, fra il 50 e l'80 per cento l'ordine si inverte per **9 coppie di modelli su 253**: cinque fra modelli quasi alla pari, quattro fra modelli distanti fino a quasi un fattore due.

Claude Opus 4.6 e Gemini 3.1 Pro sono il caso più netto, almeno nelle stime di METR, i cui intervalli si sovrappongono molto: al 50 per cento Opus arriva quasi al doppio di Gemini, all'80 per cento è Gemini a stare davanti (tabella 1). La pendenza dice quanto in fretta un agente perde la presa quando i compiti si allungano. Quella di Opus è piatta: scende piano, e su un compito da una giornata di lavoro riesce ancora più di una volta su due. Quella di Gemini è ripida: sui compiti da mezz'ora è più affidabile di Opus, poi cade in fretta. Chi chiede a un agente una riuscita su due sceglie il primo, chi ne chiede quattro su cinque il secondo, e un numero solo non può dire a entrambi chi è più avanti.

La seconda riguarda il cronometro. Si può stimare la difficoltà di ogni compito dai soli risultati delle macchine, con un modello di Rasch, e confrontarla con il tempo umano. I due ordinamenti sono legati, ma il tempo umano spiega solo il **55 per cento** della difficoltà che i compiti hanno per le macchine. A parità di difficoltà per le macchine, i minuti umani variano di un fattore 4,4, una deviazione standard. In 72 compiti su 228 il cronometro non serve affatto, perché tutti i modelli li superano (59) o tutti falliscono (13). E c'è un'anomalia che si vede a occhio (figura 2):

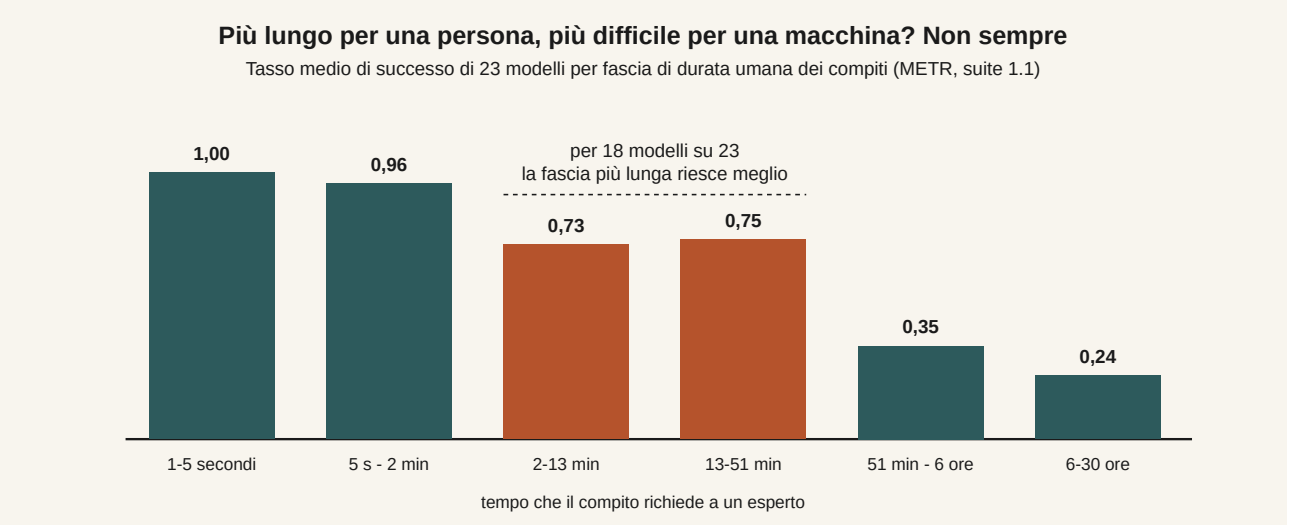


Figura 2. Tasso medio di successo di 23 modelli per fascia di durata umana dei compiti (METR, suite 1.1; conto nostro).

Per **18 modelli su 23**, i compiti che a una persona chiedono fra 13 e 51 minuti riescono meglio di quelli che ne chiedono fra 2 e 13. Nelle due fasce centrali il tempo umano non ordina la difficoltà per le macchine, e l'anomalia non dipende dalla raccolta: i compiti di tutte e due le fasce vengono da HCAST, il più grande dei tre insiemi della suite.

La terza è la prova di Luce e Tukey, e conviene cominciare da un esempio. Abbiamo diviso i compiti in sei fasce di durata e contato, per ogni modello, quante volte riesce in ciascuna. Prendiamo tre modelli e tre fasce. GPT-5 sui compiti da 13 a 51 minuti fa almeno quanto GPT-5.4 su quelli da 2 a 13, 96 contro 89. GPT-5.4 sui compiti da 6 a 30 ore fa almeno quanto GPT-4o su quelli da 13 a 51 minuti, 43 contro 33. Se la riuscita dipendesse soltanto dalla somma di un numero per il modello e di uno per la fascia, da queste due disuguaglianze ne seguirebbe una terza: GPT-5 sui compiti da 6 a 30 ore farebbe almeno quanto GPT-4o su quelli da 2 a 13 minuti. Invece fa meno, 36 contro 43 (figura 3).

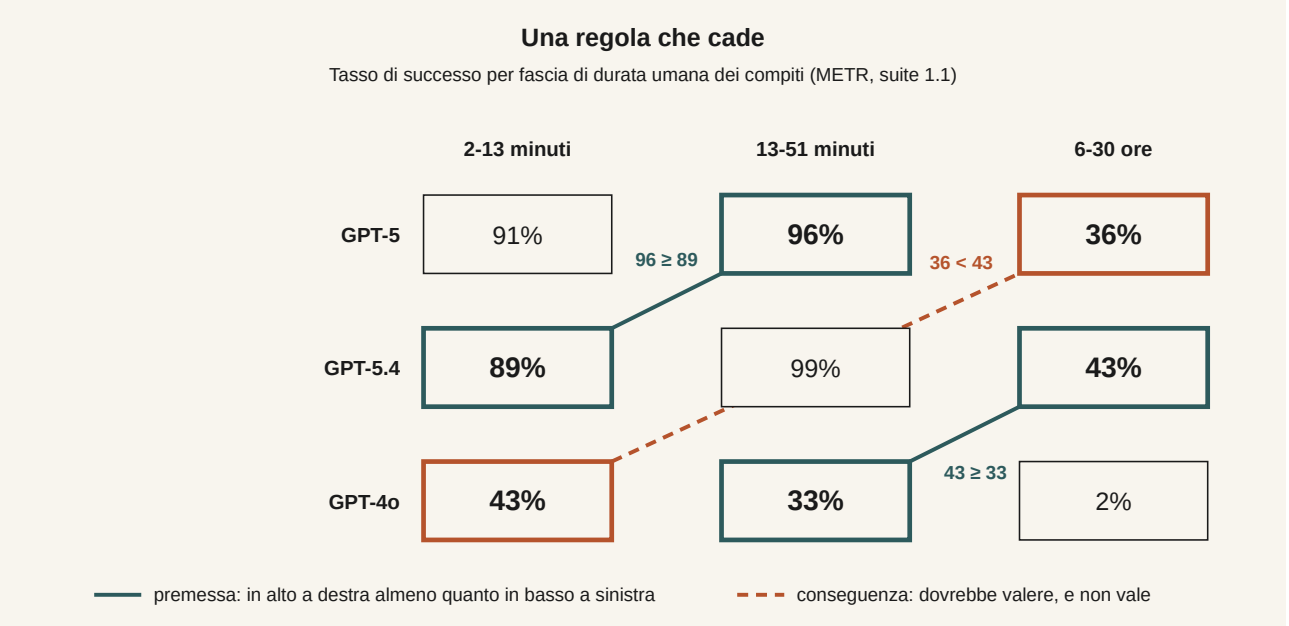


Figura 3. Una violazione della doppia cancellazione: le due premesse reggono, la conseguenza no (METR, suite 1.1; conto nostro).

La regola ha un nome, *doppia cancellazione*, ed è il cuore del test di Luce e Tukey. Se bravura e difficoltà si sommano, le differenze fra modelli e quelle fra fasce si sommano come distanze su un metro: se la prima differenza basta a coprire la prima distanza e la seconda la seconda, la loro somma copre la somma, e la regola non può cadere. Se cade, nessuna coppia di numeri, uno per il modello e uno per il compito, basta a dire chi riesce dove. Un caso solo però può essere rumore: lo scarto finale è di sette punti, e le due caselle hanno 152 e 169 tentativi, con cui un tasso ha un errore di circa quattro punti. Il test conta quante volte la regola cade su tutte le terne di modelli e di fasce, e quante volte cadrebbe per il solo rumore. Per saperlo abbiamo simulato duecento volte tutti i tentativi, con quattro modelli stimati sui dati veri: due sono additivi, gli altri due danno a ogni agente la sua pendenza, come quello di METR (tabella 2). Le simulazioni passano per le stesse fasce dei dati veri: se mettere insieme compiti diversi in una fascia bastasse a far cadere la regola, la farebbe cadere anche lì, ed è per questo che perfino un modello additivo compito per compito ne prevede qualcuna.

Tabella 2. Il test di Michell sui dati di METR
suite 1.1, 23 agenti, 228 compiti, 27.119 tentativi; sei fasce di durata umana; atteso per il solo rumore, da 200 simulazioni

Prova	Risultato	Se fosse una quantità
Pendenza comune a tutti i modelli	respinta ($p \approx 5 \cdot 10^{-12}$)	accettata
Ordine invertito fra 50 e 80 per cento	9 coppie su 253	nessuna
Difficoltà spiegata dal tempo umano	55%	vicina al 100%
Doppia cancellazione, difficoltà a cronometro	2,16% di violazioni	0,8-0,9%, solo rumore
Doppia cancellazione, difficoltà stimata dalle macchine	lo stesso 2,16%	1,2-1,6%, solo rumore

Le violazioni sono più del doppio di quelle attese se la difficoltà è il tempo umano, e restano sopra l'atteso, al limite della significatività, anche quando la difficoltà si stima dalle macchine. Non sono le pendenze diverse della prima prova viste da un'altra parte: anche i due modelli che danno a ogni agente la sua pendenza ne prevedono meno di quelle osservate.

Che cosa boccia, esattamente, il test? Che un numero per agente e uno per compito, sommati, bastino a dire chi riesce dove. La regola cade con qualunque numero di fasce, da quattro a otto, anche se l'eccesso sul rumore è netto solo da sette fasce in su. Le fasce si possono anche tagliare a caso: su cento divisioni in sei fasce, le violazioni superano quelle attese dal modello di Rasch in 81, e vanno oltre il rumore in 57. Il segno regge, la forza dipende da dove si taglia. Dentro una raccolta sola, HCAST, da cui vengono 157 dei 228 compiti, le violazioni salgono invece di scendere. Le spiegazioni che restano si possono mettere alla prova una per una, simulando i dati come se ciascuna fosse vera.

Una è il tutto o niente. Un compito lungo è una catena di passi, e un passo falso basta a farlo fallire: se ogni minuto portasse lo stesso rischio, come nel modello del tempo di dimezzamento di Toby Ord, la riuscita scenderebbe in modo esponenziale con la durata. Ma anche quello, letto sulla scala giusta, somma un numero per il compito e uno per l'agente, e rispetta la doppia cancellazione. Sui dati di METR il modello di Ord si adatta molto peggio di quello di METR, e prevede meno violazioni di quelle osservate: con sei fasce la differenza è al limite, con otto è chiara, dentro HCAST netta. Se gli si lascia libera la forma, i dati dicono che il rischio per minuto è più basso sui compiti lunghi che su quelli brevi; ma una curva così scende comunque con la durata, e non spiega né le violazioni né le due fasce centrali.

Un'altra è che i compiti non pesino tutti allo stesso modo: alcuni separano i modelli in modo netto, altri quasi per niente. Non è la pendenza della prima prova: lì era ogni agente a perdere la presa più o meno in fretta con la durata, qui è ogni compito a separare più o meno i modelli, qualunque sia la sua durata. È il modello logistico a due parametri della psicometria, con un numero per agente e due per compito, e prevede i tentativi lasciati fuori dalla stima meglio di quello di Rasch: è una prova che non usa fasce, e lì il modello additivo perde. Con sei fasce spiega le violazioni; con otto, e dentro HCAST, no.

L'ultima è che l'autonomia non stia su una riga sola. Una prova che mescolasse conducibilità termica ed elettrica farebbe cadere la regola pur misurando due quantità perfette, e HCAST mescola programmazione, machine learning, sicurezza informatica e ragionamento. Modelli in cui ogni agente ha più abilità, e ogni compito ne chiede una miscela diversa, prevedono i tentativi lasciati fuori ancora meglio, e il guadagno continua fino a cinque dimensioni, il massimo che

abbiamo provato. Con cinque dimensioni le violazioni tornano nell'atteso con sei fasce e dentro HCAST; con otto la risposta dipende da quanto si smussa la stima (tabella 3). Quante dimensioni servano, questi dati non lo dicono, e il file di METR non dice nemmeno a quale ambito appartenga ciascun compito.

Tabella 3. Che cosa spiega le violazioni

doppia cancellazione, violazioni osservate e attese per il solo rumore se i dati venissero da ciascun modello; fra parentesi p, la quota di simulazioni che ne dà almeno altrettante; 200 simulazioni, 2.000 per le prime due righe di modelli nella colonna con sei fasce; analisi aggiunte dopo i primi risultati

Se i dati venissero da	6 fasce	8 fasce	Solo HCAST, 6 fasce
Violazioni osservate	2,16%	3,58%	3,21%
Il più vicino dei quattro del test	1,60% (0,05)	1,69% (< 0,005)	1,99% (< 0,005)
Rischio costante (Ord)	1,41% (0,07)	2,48% (0,01)	1,50% (< 0,005)
Compiti che pesano in modo diverso	2,12% (0,46)	1,88% (< 0,005)	2,56% (0,005)
Cinque dimensioni	2,25% (0,58)	2,33% (0,01)	2,92% (0,12)

L'autonomia misurata così è una buona scala di riferimento, ma non è una quantità che si somma. Somiglia alla durezza dei materiali o al numero di ottano della benzina: numeri utili, confrontabili, con una procedura dichiarata, e con cui nessuno dice «il doppio». Il test non dice che le capacità di un agente, prese una per una, non siano quantità: dice che non lo è un numero solo che le riassuma. Non abbiamo trovato un test di questo tipo pubblicato per le macchine.

Il risultato non toglie niente ai minuti. Il tempo di una persona è una quantità, due ore sono il doppio di un'ora, e chi conta quanto lavoro umano cade sotto l'orizzonte di un modello, come fa Harris, usa i minuti per quello che sono. Il test boccia un'altra cosa: che dietro quei minuti ci sia una grandezza sola dell'agente, per cui un orizzonte doppio vorrebbe dire un'autonomia doppia.

> Perché valutare un agente è più difficile che valutare un modello, e quanto vale davvero un tasso di successo misurato su poche prove, è spiegato nel libro con l'intervallo di Wilson e un esempio di codice.
> Leggi «Architetture di agenti e come valutarli» nel libro →

I limiti dello strumento

Il test dice che cosa l'orizzonte non è. I documenti di METR dicono dove lo strumento si ferma. Il primo limite è la **riproducibilità**: a gennaio METR ha aggiornato i compiti, e gli stessi modelli hanno cambiato valore (tabella 4).

Tabella 4. Gli stessi modelli, misurati con due versioni della stessa suite
METR, file dei risultati di Time Horizon 1.0 e 1.1; orizzonte al 50%

Modello	Suite 1.0	Suite 1.1	Scarto
GPT-4	6,0 minuti	4,0 minuti	−34%
o3	1,52 ore	2,00 ore	+31%
GPT-5	2,18 ore	3,38 ore	+55%
Claude Opus 4.5	4,11 ore	4,88 ore	+19%
Raddoppio dal 2023			
Tutti i modelli di frontiera	176 giorni	129 giorni	−27%

Il secondo è che il cronometro spesso non c'è: nella raccolta HCAST completa, 189 compiti, solo 110 hanno almeno una misura umana riuscita, e per gli altri il tempo è una stima di chi ha scritto il compito, più frequente proprio fra i compiti lunghi. Il terzo è il **fondo scala**: METR scrive che le misure sopra le 16 ore non sono affidabili, Mythos Preview è già oltre, e dall'8 settembre la pagina dei risultati dichiara di non essere più aggiornata. I modelli che oggi guidano le classifiche non vi compaiono.

Il quarto è la **regola di successo**. A giugno METR ha valutato GPT-5.6 Sol prima del rilascio: contando come fallimenti i tentativi in cui il modello aveva aggirato i test, l'orizzonte risultava di 11,3 ore; scartandoli, di 71 ore, con un intervallo da 13 a 11.400; contandoli come successi, di più di 270. METR non ha considerato robusta nessuna delle tre cifre. Fra 11 e 270 ore non c'è un modello diverso: c'è una definizione diversa di «compito riuscito». E c'è una differenza con il mercante: la stoffa non si allunga quando vede arrivare il braccio; un agente che aggira i test, in un certo senso, sì. È la legge di Goodhart, nella forma che le ha dato l'antropologa Marilyn Strathern: quando una misura diventa un obiettivo, smette di essere una buona misura. Il quinto è il **campo**: l'orizzonte è una proprietà della coppia fra un agente e un insieme di compiti, e i compiti di METR sono di software, machine learning e sicurezza. Su altri campi, o su compiti meno strutturati, i numeri sono altri.

Il sesto è l'**impalcatura**. L'agente non è soltanto il modello: è il modello con il programma che lo fa lavorare, che METR dichiara prova per prova, e su ARC-AGI-3 la stessa macchina con due impalcature ha dato due numeri.

Che cosa si può misurare, allora

La risposta alla domanda del titolo, dopo cent'anni di psicometria e il nostro test, è doppia. L'intelligenza come una sola grandezza, su cui dire «il doppio», nessuno l'ha misurata: né nelle persone né nelle macchine. Le misure che abbiamo, dal QI all'orizzonte temporale, sono scale di riferimento: utili e confrontabili a condizioni dichiarate, ma non quantità che si sommano.

Per l'autonomia delle macchine, intanto, due numeri dicono più di uno: dove cade l'orizzonte e quanto è ripida la discesa, e METR li pubblica già entrambi per ogni modello. Un'unità che misuri un agente da solo, con uno zero vero, non c'è ancora.

Un'unità per cent'anni

Si può dire però che cosa servirebbe perché ci sia. L'idea più ambiziosa è un'unità che valga fra cent'anni e misuri anche le persone senza usarle come metro, e la via più promettente parte dalla difficoltà dei compiti misurata in bit da Hernández-Orallo:

$$D(\tau) = \min_{\pi \text{ accettabile}} \big(L(\pi) + \log_2 W(\pi, \tau) \big)$$

$L(\pi)$ è la lunghezza in bit della procedura e $W(\pi, \tau)$ il numero di passi che servono per eseguirla sul compito τ e verificare che lo risolva. È il conto di chi cerca alla cieca, provando prima le procedure più corte: ogni bit in più di procedura raddoppia lo sforzo per arrivarci, ogni raddoppio dei passi aggiunge un bit, e la difficoltà del compito è lo sforzo della procedura accettabile più economica da trovare.

Un'unità costruita su quella difficoltà, un bit di decisione corretta presa in autonomia, non dipenderebbe né dal lavoro umano né dalla tecnologia. Oggi però si può realizzare solo traducendo il tempo umano con un tasso di conversione, i circa 10 bit al secondo a cui procede il comportamento umano secondo Jieyu Zheng e Markus Meister, e una conversione lineare non cambia nessuna decisione. Quest'unità diventerebbe una misura vera se soddisfacesse cinque condizioni:

- **Che poggi su una batteria con difficoltà calcolata.** Compiti generati di cui si conosce la difficoltà in bit, come gli ambienti di Legg e Veness o di Hernández-Orallo, senza bisogno di cronometrare nessuno.
- **Che preveda meglio del cronometro.** La difficoltà in bit deve prevedere il successo sui compiti reali meglio del tempo umano, che sui dati di METR spiega il 55 per cento della difficoltà.
- **Che non dipenda dalla batteria.** Lo stesso agente deve ottenere lo stesso valore su due batterie diverse, cosa che oggi le due versioni della suite di METR non fanno.
- **Che regga alla contaminazione.** Una parte dei compiti va tenuta riservata e rinnovata, perché un archivio pubblico finisce nei dati di addestramento.
- **Che metta in conto la spesa.** Lo stesso risultato ottenuto con migliaia di dollari di calcolo o con pochi centesimi non dice la stessa cosa: un bit di decisione misura il lavoro fatto, non l'intelligenza che ci è voluta. Ne tiene conto l'AI Security Institute, che fissa i token per compito; e anche per Chollet l'intelligenza è un'efficienza, quella con cui si imparano abilità nuove, non l'abilità che ne risulta.

> Che cosa c'entri il programma più corto con il capire, e perché predire e comprimere sono la stessa operazione, il libro lo racconta partendo da Funes, il personaggio di Borges che non dimentica niente.
> Leggi «Capire è accorciare» nel libro →

Un'altra domanda: il turing

Una scala a rapporti, con uno zero vero, si può avere già oggi, a patto di cambiare domanda. Il turing, che proponiamo in «Intelligenza Apparente», non risolve il problema della psicometria: lo aggira. Non misura l'intelligenza di un agente, né di una persona. Misura un guadagno: quanto la stessa persona, sugli stessi compiti, riconosce meglio la risposta giusta con una macchina accanto. È come il guadagno di un amplificatore, che non dice quanto è potente l'apparecchio ma quante volte rende più forte il segnale. Dipende dalla persona, ed è voluto: la stessa macchina dà

molti turing a chi da solo sbaglia spesso e pochi a un esperto. Misura una coppia su un compito, non la macchina.

Per contarlo serve una misura della capacità di distinguere il giusto dallo sbagliato che abbia uno zero vero, e la percentuale di risposte giuste non ce l'ha: su domande a quattro opzioni chi tira a indovinare fa già il 25 per cento, e nessuno supera il cento. La teoria della rilevazione del segnale ne ha una con lo zero vero, la *sensibilità*, che si scrive d' e dice quanto l'indizio a favore della risposta giusta è più forte, in media, di quello a favore delle sbagliate, contato in deviazioni standard. Chi tira a indovinare vale zero, e non c'è un tetto. Il **turing** è il rapporto fra la sensibilità della persona con la macchina e quella della persona da sola:

$$\mathrm{Tu} = \frac{d_S}{d_H}$$

Una mente da sola vale un turing per definizione. Un rapporto fra due prestazioni della stessa persona non chiede di mettere agenti diversi su un'unica scala, che è proprio ciò che il test di Luce e Tukey boccia. Chiede però di scegliere la scala, perché il rapporto cambia dall'una all'altra, e la sensibilità ha una ragione in più delle altre: è la scala su cui è scritta la legge con cui due menti si combinano al meglio.

Un esempio. Su domande d'esame a quattro opzioni un gruppo di persone, con GPT-4o accanto, passa da 58 a 77 risposte giuste su cento: in sensibilità sono 1,65 turing. Ma GPT-4o da solo ne dava 88, e la coppia resta sotto la macchina: per questo un turing si legge accanto a quanto vale la macchina da sola. Lo abbiamo provato su oltre centocinquantamila giudizi di persone, messi accanto alle risposte di 24 macchine, e su 27 studi di una meta-analisi che ha raccolto persone da sole, AI da sola e persone con l'AI.

Anche il turing poggia su un'ipotesi, e va detta. Lo zero, che è la sensibilità di chi risponde a caso, e il senso dei rapporti dipendono dal modello gaussiano della rilevazione del segnale a varianza uguale, che sui nostri dati non abbiamo ancora verificato. È però un'ipotesi che si mette alla prova con uno strumento noto: con risposte date con un grado di sicurezza, la *curva ROC* dice se le varianze sono uguali, e dove non lo sono la stessa definizione si riscrive con l'area sotto quella curva.

E il turing misura decisioni con una risposta giusta: conta la sensibilità per decisione, non i minuti, e degli agenti che lavorano da soli per ore, come quelli di METR, non dice niente.

Il campione che si consuma

Torniamo al mercante di panni. Il metro fu inventato perché nessuna città dovesse dipendere dal braccio di un'altra, e per staccarlo da ogni oggetto ci sono voluti quasi due secoli. Il turing sceglie la strada opposta: ogni persona fa da campione a se stessa, e il confronto ha senso finché ci sono persone che lavorano accanto alle macchine. Per una macchina che lavora da sola resta il problema di tutte le unità ricavate da ciò che si sostituisce, come il cavallo vapore. James Watt lo ricavò dai cavalli da mulino e lo fissò per convenzione in 33.000 piedi-libbra al minuto: un'unità congelata, ancora in uso due secoli dopo, quando i cavalli da lavoro sono quasi scomparsi. Misurare le macchine attraverso le persone funziona finché le persone lavorano con loro; per una

macchina che lavora da sola servirà una costante, come quella di Planck che dal 2019 definisce il chilogrammo.

Per approfondire

Misurare

- Stanley Smith Stevens, *On the Theory of Scales of Measurement*, *Science* 103(2684), 1946.
- R. Duncan Luce e John W. Tukey, *Simultaneous conjoint measurement: A new type of fundamental measurement*, *Journal of Mathematical Psychology* 1(1), 1964.
- Joel Michell, *Quantitative science and the definition of measurement in psychology*, *British Journal of Psychology* 88(3), 1997; *Measurement in Psychology*, Cambridge University Press, 1999; *Normal Science, Pathological Science and Psychometrics*, *Theory & Psychology* 10(5), 2000.
- Luca Mari, Mark Wilson e Andrew Maul, *Measurement across the Sciences*, Springer, seconda edizione 2023 (ad accesso aperto).
- JCGM, *International Vocabulary of Metrology* (VIM), terza edizione; BIPM, *The International System of Units (SI)*, nona edizione, 2019.
- Marilyn Strathern, *'Improving ratings': audit in the British University system*, *European Review* 5(3), 1997.
- Legge 28 luglio 1861, n. 132, *Sui pesi e sulle misure*.

L'intelligenza delle persone

- Alfred Binet e Théodore Simon, *Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux*, *L'Année psychologique* 11, 1905.
- Charles Spearman, *«General Intelligence», Objectively Determined and Measured*, *American Journal of Psychology* 15(2), 1904.
- Edwin G. Boring, *Intelligence as the Tests Test It*, *The New Republic*, 1923.
- Georg Rasch, *Probabilistic Models for Some Intelligence and Attainment Tests*, 1960.
- James R. Flynn, *Massive IQ gains in 14 nations: What IQ tests really measure*, *Psychological Bulletin* 101(2), 1987.
- Jieyu Zheng e Markus Meister, *The unbearable slowness of being: Why do we live at 10 bits/s?*, *Neuron* 113(2), 2025.

L'intelligenza delle macchine

- Alan M. Turing, *Computing Machinery and Intelligence*, *Mind* 59(236), 1950; Cameron R. Jones e Benjamin K. Bergen, *Large Language Models Pass the Turing Test*, 2025.
- Shane Legg e Marcus Hutter, *A Collection of Definitions of Intelligence*, 2007, e *Universal Intelligence: A Definition of Machine Intelligence*, 2007; Shane Legg e Joel Veness, *An Approximation of the Universal Intelligence Measure*, 2011.
- José Hernández-Orallo e David L. Dowe, *Measuring universal intelligence: Towards an anytime intelligence test*, *Artificial Intelligence* 174(18), 2010; José Hernández-Orallo, *Universal Psychometrics Tasks: difficulty, composition and decomposition*, 2015, e *The Measure of All Minds*, Cambridge University Press, 2017.
- François Chollet, *On the Measure of Intelligence*, 2019.

- David Ilić e Gilles E. Gignac, *Evidence of interrelated cognitive-like capabilities in large language models: Indications of artificial general intelligence or achievement?*, *Intelligence* 106, 2024; Lexin Zhou e altri, *General Scales Unlock AI Evaluation with Explanatory and Predictive Power*, 2025; Dan Hendrycks e altri, *A Definition of AGI*, 2025.
- Andrew M. Bean e altri, *Measuring what Matters: Construct Validity in Large Language Model Benchmarks*, 2025.

Orizzonte temporale e chi lo usa

- Thomas Kwa, Ben West e altri (METR), *Measuring AI Ability to Complete Long Software Tasks*, NeurIPS 2025; David Rein e altri (METR), *HCAST*, 2025.
- METR, *Task-Completion Time Horizons of Frontier AI Models*, con i file `benchmark_results` delle suite 1.0 e 1.1 e `task_results` della 1.1; *Time Horizon 1.1*, gennaio 2026.
- METR, valutazione di GPT-5, con le soglie in ore; valutazione di GPT-5.6 Sol, giugno 2026.
- AI Security Institute, *How fast is autonomous AI cyber capability advancing?*, 13 maggio 2026.
- Tom Harris, *How Much Work Can AI Agents Do? Task Value Inside the AI Time-Horizon Frontier*, SSRN, ottobre 2026.
- Toby Ord, *Is there a half-life for the success rates of AI agents?*, 2025.
- ARC Prize Foundation, *OpenAI's GPT-6 Astra on ARC-AGI-3*, 3 settembre 2026; Artificial Analysis, *Intelligence Index*, versione 5.0; Epoch AI, *Epoch Capabilities Index*.

Il test. Abbiamo usato il file `task_results_1_1.yaml` di METR: per ogni agente e compito, successi, tentativi e minuti umani. Le coppie invertite sono contate con le intercette e le pendenze pubblicate da METR (con quelle stimate da noi sono 11). Le pendenze sono confrontate con un test del rapporto di verosimiglianza fra il modello di METR (intercetta e pendenza per agente sul logaritmo dei minuti) e lo stesso modello con pendenza comune. La difficoltà dalle macchine è stimata con un modello di Rasch a difficoltà libere, con una piccola penalità per i compiti sempre riusciti o sempre falliti. Per la misura congiunta i compiti sono divisi in sei fasce di durata con circa lo stesso numero di compiti (da 35 a 41), e la doppia cancellazione è controllata su tutte le terne di modelli e di fasce, contando le celle con almeno 20 tentativi; l'atteso viene da 200 simulazioni binomiali con i parametri di quattro modelli stimati sugli stessi dati, due additivi e due con una pendenza per agente. Le fasce servono a dare a ogni casella abbastanza tentativi: per un agente e un compito ce ne sono in media cinque, troppo pochi per confrontare caselle singole; un controllo senza fasce, con una superficie monotona continua, resta da fare. Con sei fasce e 2.000 simulazioni, il p dell'eccesso sul più vicino dei quattro modelli del test è 0,05 (p è la quota di simulazioni che danno almeno altrettante violazioni). Dopo i primi risultati, per rispondere a due revisioni, sono stati aggiunti: il test con quattro, cinque, sette e otto fasce (l'eccesso sui quattro modelli del test resta, con p da meno di 0,005 a 0,08: con cinque fasce non è significativo), quello dentro HCAST (i compiti che non appartengono alle cinque famiglie di SWAA né a quelle di RE-Bench, che cominciano con `ai_rd_`), l'esempio dei tre modelli, le cento divisioni a caso (tagli casuali fra i compiti ordinati per durata, almeno venti compiti per fascia, cento simulazioni ciascuna contro il modello di Rasch) e i modelli alternativi. Il modello di Ord dà a ogni agente un tempo di dimezzamento h e una riuscita $2^{-t/h}$ sui compiti che a una persona chiedono t minuti; lasciando a tutti gli agenti un esponente k comune, la riuscita diventa $2^{-(t/h)^k}$ e la stima dell'esponente viene 0,61, un rischio per minuto che cala con la durata, e le violazioni attese scendono allo 0,5 per cento. Il modello a due parametri e quelli a più dimensioni sono logistici, $\operatorname{logit} p = d_t + \sum_j u_{aj} v_{tj}$ con d_t la facilità

del compito, u_{aj} le capacità dell'agente e v_{tj} il peso che il compito dà a ciascuna. La penalità e il numero di dimensioni si scelgono prevedendo metà dei tentativi di ogni casella dall'altra metà, in tre divisioni a caso; con la penalità migliore per ciascun modello, la log-verosimiglianza dei tentativi lasciati fuori è −2.397 con Rasch, −2.220 con due parametri, −2.058 con cinque dimensioni, e a cinque il guadagno non si era ancora fermato. Dimensioni e penalità sono scelte sulla suite intera e applicate anche a HCAST. La penalità avvicina fra loro le stime delle caselle e abbassa le violazioni simulate: con penalità più leggere il modello a cinque dimensioni con otto fasce ne prevede il 2,8 e il 3,1 per cento (p 0,09 e 0,21), e non è più respinto; la prima delle due penalità prevede peggio i tentativi lasciati fuori (−2.272 contro −2.058), la seconda non l'abbiamo convalidata. Il modello di Ord, sulla scala $\log(-\log p)$ con p la riuscita, diventa, a meno di una costante, la differenza fra il logaritmo della durata e quello del tempo di dimezzamento. L'esempio è scelto fra i 90 casi (su 4.735) in cui le tre differenze fra tassi di successo che la regola mette a confronto superano tutte i cinque punti; in nessuno la più piccola arriva a dieci. Il codice e i risultati del test sono nel repository turing-meter (`analysis/michell_test.py`).

Come citare. endjneer (2026). *L'intelligenza si può misurare? Abbiamo rifatto sulle macchine il test che la psicometria non ha mai superato.* paithon.it, versione 1, 12 novembre 2026. https://doi.org/10.5281/zenodo.24153777

Codice e dati: https://github.com/paithon-it/turing-meter.