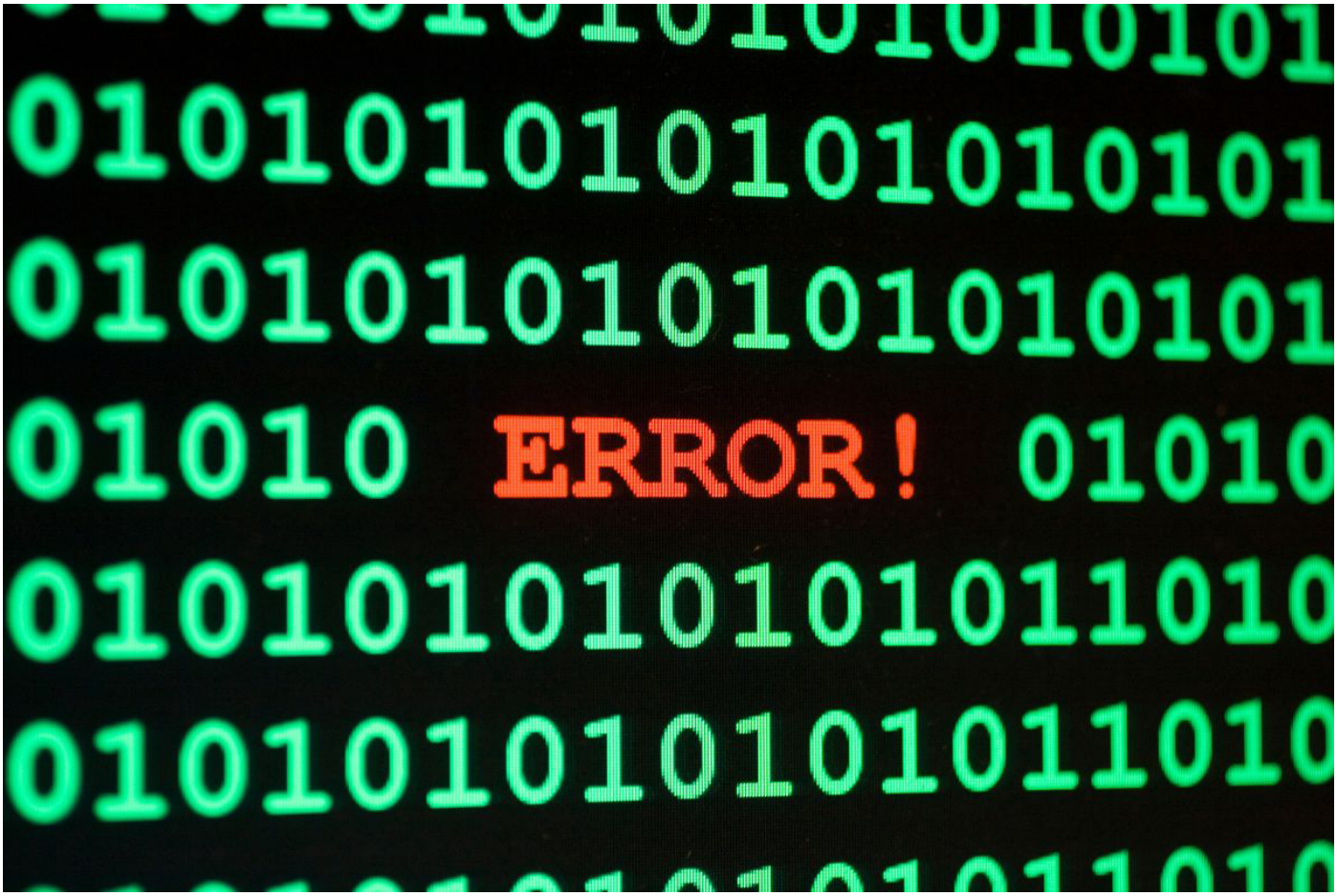
News



E se l'IA fosse uno studente? Come le IA rispondono alle prove INVALSI

L'**intelligenza artificiale (IA)** è diventata un tema centrale nel dibattito educativo e sociale, coinvolgendo tanto la didattica quanto i metodi di valutazione. Se da un lato i sistemi di IA promettono di semplificare alcuni processi di analisi e apprendimento, dall'altro emergono questioni importanti legate all'affidabilità dei modelli e alla loro effettiva "capacità" di ragionare. Le aziende produttrici si pubblicizzano dichiarando l'uscita di modelli sempre più "intelligenti", ma come si misura la capacità di ragionamento di un sistema di IA?

Per rispondere a questa domanda sono stati costruiti numerosi benchmark, indici misurabili, con l'idea di confrontare fra di loro i vari Large Language Model (quali ad esempio ChatGPT, Claude o LLaMa). Tuttavia, gli attuali indicatori presentano spesso alcuni di questi limiti:

- i benchmark vengono ideati e applicati dalle stesse aziende produttrici;
- test e addestramento attingono a una sovrapposizione sostanziale di fonti, minando la validità dei risultati, in quanto

il modello “riconosce” le risposte invece di “ragionare”;
- i benchmark si riferiscono unicamente a domande e risposte fornite in lingua inglese.

Inoltre, molti di questi indicatori misurano principalmente la capacità di completare frasi o di riconoscere pattern statistici, senza valutare in modo adeguato la profondità del ragionamento, l'adattabilità a contesti specifici e la coerenza.

I SISTEMI DI AI GIUDICATI ATTRAVERSO LE PROVE INVALSI

Alcuni gruppi di ricerca italiani, fra cui uno dell'Università di Milano Bicocca, hanno proposto e stanno sperimentando l'utilizzo dei test INVALSI pubblici come benchmark per confrontare la performance dei vari sistemi di IA disponibili. A prima vista potrebbe sembrare insolito sottoporre un'IA a un test pensato per gli studenti, eppure proprio studi recenti (come quello citato) mostrano che questi strumenti standardizzati possono costituire un benchmark efficace per capire quali modelli di IA siano più adatti a specifici compiti linguistici, logico-comprensivi e di ragionamento.

Le prove INVALSI, concepite per misurare competenze quali la **comprensione del testo,** le **capacità di riflessione linguistica** e le **abilità logiche e matematiche**, stanno aiutando i ricercatori a evidenziare molti limiti degli attuali sistemi di IA.

GLI ERRORI COMMESSI DAI SISTEMI IA CRESCONO ALL'AUMENTARE DEL GRADO SCOLASTICO

Le analisi condotte mostrano una tendenza piuttosto chiara: le intelligenze artificiali ottengono **risultati generalmente migliori sulle prove INVALSI di grado inferiore**, ad esempio la scuola primaria rispetto a quelle di scuola secondaria, dove si richiedono competenze più complesse e una maggiore capacità di ragionamento.

Ma dov'è che le IA incontrano maggiore difficoltà?

I modelli di IA sembrano gestire bene i quesiti a risposta multipla semplice su testi o nozioni di base, mentre incontrano difficoltà evidenti in attività che richiedono elaborazioni linguistiche approfondite, interpretazioni articolate o conoscenze più strutturate, tipiche dei livelli di istruzione superiori.

RAGIONAMENTI SBAGLIATI CHE POSSONO PORTARE ALLA RISPOSTA GIUSTA

Un'evidenza degna di nota riguarda poi i quesiti di riflessione linguistica o di logica avanzata, dove il modello è chiamato ad applicare più passaggi di ragionamento per arrivare alla soluzione. In alcuni casi, le IA analizzate sono giunte alla risposta esatta, ma attraverso un processo tutt'altro che lineare o coerente: ad esempio, per stabilire se in una frase fosse corretta o meno l'aggiunta di una lettera “h”, un modello ha iniziato inspiegabilmente a tradurre il testo in francese e a confrontare la presenza della “h” in varie parole francesi, ignorando del tutto la regola ortografica italiana. In altri frangenti, gli algoritmi hanno mescolato nozioni matematiche fuori contesto o generato paragrafi senza alcun nesso logico, salvo poi “approdare” al risultato giusto quasi per combinazione. Questo meccanismo di “ragionamento contorto” solleva **interrogativi sulla validità delle spiegazioni che l'IA fornisce** e su quanto la correttezza della risposta sia frutto di un processo statistico casuale.

In generale, si nota che — sebbene l'IA arrivi a indicare alcune risposte in maniera efficace — **non sempre è in grado di approfondire** il perché di quelle risposte. Questo è un aspetto cruciale per i docenti, che nell'interpretare il risultato di una prova valutano non solo la correttezza, ma anche i processi di apprendimento e di ragionamento sottostanti.

QUESITI A CUI LE IA SI RIFIUTANO DI RISPONDERE

Alcuni modelli, specie quelli a carico di grandi aziende, si rifiutano di rispondere nei casi in cui la domanda includa parole percepite come violente o discriminatorie. Possono essere un esempio i testi che parlano di guerra o di contestazione. Questo è dovuto a sistemi di protezione che vengono inseriti al fine di evitare usi impropri delle IA o la generazione di contenuti discriminatori o pericolosi. Tuttavia, i sistemi proprietari, non avendo una reale comprensione del contesto, tendono a “censurare” contenuti che percepiscono come inadeguati, anche quando tali passaggi sono assolutamente leciti e funzionali all'esercizio di comprensione.

L'INTERESSE DELL'ISTITUTO INVALSI PER I RISULTATI DI QUESTE RICERCHE

L'Istituto nazionale per la valutazione del sistema educativo di istruzione e di formazione (INVALSI) segue con grande interesse l'utilizzo delle sue prove come benchmark per valutare la capacità di ragionamento delle IA, poiché sono necessarie riflessioni non solo su come questo strumento stia influenzando e influenzerà la didattica ma anche su come in risposta dovrà evolvere la valutazione.

Uno dei temi sarà legato anche al modo in cui le prove INVALSI, ma anche i docenti, dovranno cercare di misurare

competenze sempre più “umane” e meno aggirabili da procedure puramente statistiche.

Se la direzione fosse quella di puntare a limitare l’utilizzo dell’IA per lo svolgimento delle prove, i dati indicano una strategia che punta a domande aperte, con quesiti che mettano alla prova la coerenza del percorso risolutivo, anziché limitarsi alla verifica del risultato finale (come nel caso de quesiti a risposta multipla semplice o di tipo vero/falso), oppure sottoponendo nuovi tipi di prove che richiedano riformulazioni creative e ragionamenti articolati. Su questo aspetto vale la pena ricordare che molti LLM, in questo momento, non integrano sistemi in grado di “tracciare linee” o interpretare adeguatamente grafici e schemi, ciò vuol dire che problemi di questo tipo necessitano di una rielaborazione per affinché l’IA possa proporre una soluzione.

PORTARE L’IA IN CLASSE

Gli esempi riportati mostrano un quadro dello stato attuale del “ragionamento” dei sistemi di IA, ma bisogna tenere presente che questi tool evolvono in fretta e le correzioni ai modelli sono continue, e permetteranno a questi strumenti di rispondere sempre meglio a ogni tipo di test. Questo non significa che i chatbot saranno capaci di “ragionare” come un essere umano ma solo che **sarà sempre più difficile smascherare il loro comportamento statistico**. Oggi, nel momento in cui questa tecnologia è ancora da perfezionare, è possibile mostrare agli studenti casi di ragionamento errato mostrando che affidarsi totalmente a questo tipo di tecnologia è rischioso**. Lavorare in classe su un utilizzo consapevole** e sullo sviluppo di competenze che permettano agli alunni di verificare la veridicità e la qualità di quello che viene proposto da ChatGPT, Copilot, Claude e tutti gli altri sistemi di IA potrebbe rivelarsi la migliore delle strategie.

VIDEO PER APPROFONDIRE

Laboratorio pratico di Intelligenza artificiale | Irene Fabbri

L’intelligenza artificiale strumenti per l’insegnante | Giuliana Barberis

L’AI alla prova: valutazione dei modelli Linguistici di Grandi Dimensioni sui test INVALSI

ARTICOLI ACCADEMICI (PREPRINT) DI RIFERIMENTO

“Disce aut Deficere: Evaluating LLMs Proficiency on the INVALSI Italian Benchmark” – Fabio Mercorio, Mario Mezzanzanica, Daniele Potertì, Antonio Serino e Andrea Seveso

“INVALSI – Mathematical and Language Understanding in Italian: A CALAMITA Challenge” – Giovanni Puccetti, Maria Cassese e Andrea Esuli