FONDAZIONE BRUNO KESSLER
MAGAZINE

SHADES, nuovo dataset globale per monitorare come l'AI riproduce (e inventa) stereotipi culturali

14 Maggio 2025

Realizzato con il contributo di Fondazione Bruno Kessler, il progetto coinvolge oltre 50 ricercatori e analizza migliaia di interazioni in 16 lingue per svelare i bias nei modelli generativi

50 ricercatori in tutto il mondo hanno analizzato migliaia di interazioni in 16 lingue diverse per rilevare che l'accentuazione e l'amplificazione di stereotipi già diffusi a livello sociale è uno pericolo concreto anche nei nuovi modelli linguistici generativi.

SHADES è un progetto internazionale e collaborativo di cui **Fondazione Bruno Kessler** è parte e che ha creato un nuovo strumento di monitoraggio e osservazione così esteso e multiculturale da essere il primo nel suo genere.

"La ricerca è nata 4 anni fa all'interno del progetto Big Science, un'iniziativa guidata da Margaret Mitchell, ricercatrice e AI ethics leader ad Hugging Face – spiega **Beatrice Savoldi**, ricercatrice dell' unità **MT** del Centro Digital Industry di FBK**–** *Insieme abbiamo creato un dataset innovativo e unico per estensione linguistica, geografica, culturale e in termini di rappresentazione di categorie di stereotipi. Dalla letteratura sapevamo che i modelli linguistici di grandi dimensioni (LLM) presentano bias, ma i risultati di SHADES si sono rivelati ancora più allarmanti del previsto, soprattutto su categorie meno studiate come etnia e nazionalità, e su lingue finora poco esplorate."*

Il lavoro, frutto della collaborazione di oltre 50 soggetti in tutto il mondo, ha analizzato **migliaia di interazioni in 16 lingue diverse**. Lo studio ha mostrato che, in presenza di prompt stereotipati, i modelli linguistici tendono non solo a riprodurli, ma spesso a giustificarli supportandoli con **fonti pseudo-scientifiche o pseudo-storiche e falsi fatti sociali**. Il linguaggio utilizzato e i riferimenti inseriti sono ingannevoli, ovvero ad un occhio poco esperto potrebbe non risultare subito

chiaro che le informazioni sono inventate. Questo fenomeno rappresenta una forma di inquinamento dell'ecosistema informativo: i contenuti generati promuovono visioni estreme, basate sul pregiudizio piuttosto che su dati reali. In alcuni casi, i modelli arrivano persino a creare nuovi stereotipi, trasformando un singolo pregiudizio in una narrazione sistemica e distorta della realtà.

L'effetto risulta particolarmente accentuato nei modelli meno addestrati in certe lingue, ma nessun sistema analizzato completamente immune, sebbene ci siano variazioni tra modelli e tra categorie di stereotipi.

I dati raccolti pongono interrogativi urgenti sulla responsabilità nello sviluppo e nell'utilizzo come strumenti di informazioni affidabili dei grandi modelli linguistici (LLM), e indicano chiaramente **l'esigenza di strategie più efficaci per la mitigazione degli errori.** Nel contesto attuale, in cui i modelli linguistici sono sempre più presenti in chatbot, assistenti virtuali e strumenti di sintesi automatica, il rischio è che essi contribuiscano a diffondere visioni riduttive, false e pregiudizievoli sul mondo e le persone.

Ma le soluzioni esistono, a patto di affrontare il problema con consapevolezza e responsabilità. Progetti come SHADES offrono strumenti concreti per monitorare e mitigare i rischi, rafforzando una cultura dell'AI trasparente, inclusiva e rispettosa della complessità sociale.

La pubblicazione completa è consultabile qui:
https://aclanthology.org/2025.naacl-long.600/

Il dataset è disponibile (previa richiesta di accesso, data la sensibilità del contenuto) al seguente link: https://huggingface.co/datasets/LanguageShades/BiasShades

LINK

https://magazine.fbk.eu/it/news/shades-nuovo-dataset-globale-per-monitorare-come-lai-riproduce-e-inventa-stereotipi-culturali/

TAG

- #bias
- #big science
- #dataset
- #industriadigitale
- #intelligenzaartificiale
- #large language models
- #lingue
- #llm
- #machine translation
- #modelli linguistici
- #MT
- #shades

AUTORI

- Giovanna Rauzi