Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

26 aprile 2024 V1.0[1]

Promemoria sulla sensibilizzazione riguardo ai modelli linguistici di grandi dimensioni basati sull'IA in seno all'Amministrazione federale

Cosa sono i modelli linguistici di grandi dimensioni basati sull'IA?

I modelli linguistici di grandi dimensioni basati sull'IA (*large language models*, LLM) sono tecnologie di IA su cui si fondano i più comuni sistemi di IA[2] utilizzati nell'ambito del linguaggio naturale. Sono infatti la base di molti strumenti di IA generativa.[3]
Gli LLM sono in grado di elaborare, comprendere, interpretare e generare il linguaggio naturale, e sono in grado di svolgere un ampio ventaglio di compiti, come la traduzione, la comprensione del testo e la generazione di testi. Gli LLM più moderni sono addestrati con grandi volumi di dati nei formati più disparati e sanno produrre testi spesso indistinguibili da quelli scritti da un essere umano. È comunque anche possibile utilizzare input acustici o visivi (immagini), poiché in molti casi possono essere convertiti in testo in modo quasi impeccabile. Anche l'emissione vocale dell'IA non è quasi più distinguibile dalla voce umana. Alcuni LLM sono già stati ampliati nei cosiddetti modelli multimodali, che, oltre ai testi, sono ad esempio in grado di elaborare e generare anche immagini e video.[4]
Alcuni esempi di LLM sono la serie di modelli GPT (usati in ChatGPT di OpenAI e in Copilot di Microsoft), Gemini (usato in Gemini di Google, ex Bard), i modelli LLaMA di Meta, la serie di modelli Grok di X e i modelli Claude di Anthropic.[5] Gli LLM acquisiscono la capacità di interpretare e generare contenuti in linguaggio naturale per scopi generali «imparando» le relazioni statistiche tra le parole a partire da documenti di testo nell'ambito di un processo di addestramento iterativo ad alta intensità computazionale. Questi modelli statistici si basano su tecniche e metodi di elaborazione del linguaggio naturale (*natural language processing*, NLP[6]) desumendo così il significato e le correlazioni dal linguaggio umano.

Principi dei modelli linguistici di grandi dimensioni basati sull'IA (LLM) nell'Amministrazione federale[7]

In qualità di utenti di tecnologie di IA generativa (come gli LLM) in seno all'Amministrazione federale, siete invitati a rispettare i seguenti principi per una sperimentazione responsabile. Le tecnologie di IA generativa (come gli LLM) possono fornirvi supporto nelle vostre attività amministrative correnti. Gli LLM sono infatti in grado di eseguire un ampio ventaglio di compiti basati su testo in modo parzialmente o completamente automatico. Provateli e

---

1 Questo promemoria viene sottoposto a revisione periodica per tenere conto dei nuovi sviluppi e permettere una migliore comprensione delle tecnologie di IA generativa (come i modelli linguistici di grandi dimensioni) nell'Amministrazione federale.

2 Un «sistema di intelligenza artificiale» (sistema di IA) è un sistema basato su macchine che, a scopi impliciti o espliciti, deduce a partire dagli input ricevuti in che modo generare degli output nella forma di previsioni, contenuti, raccomandazioni o decisioni in grado di influire sul contesto fisico o virtuale. I sistemi di intelligenza artificiale possono essere più o meno autonomi. Con «tecnologia di intelligenza artificiale» (tecnologia di IA) si intendono singole funzioni implementabili nei computer utilizzate per sviluppare l'intelligenza artificiale (p. es. l'apprendimento automatico). Un sistema di intelligenza artificiale corrisponde quindi a una combinazione strutturata e contestuale di tecnologie di intelligenza artificiale utilizzate per sviluppare quest'ultima. *Fonte:* https://cnai.swiss/it/servizi/terminologia/

3 L'«intelligenza artificiale generativa» (IA generativa) è un concetto molto ampio che si riferisce a sistemi di IA addestrati per gestire grandi quantità di dati provenienti dal mondo fisico e da quello virtuale allo scopo di generare dati autonomamente (p. es. testi, immagini, registrazioni audio, video, simulazioni, codici). Spesso questi prodotti sono multimodali, per esempio con input e/o output in una o più modalità (come testi, immagini, video). *Fonte:* https://cnai.swiss/it/servizi/terminologia/

4 Sono anche noti come grandi modelli multimodali di IA (*large multimodal models*, LMM) con capacità generative (https://huyenchip.com/2023/10/10/multimodal.html, in inglese).

5 V. «Promemoria per l'utilizzo di strumenti di IA generativa nell'Amministrazione federale»; https://cnai.swiss/it/it-servizi-altri-servizi-promemoria-sullia/

6 L'«elaborazione del linguaggio naturale» (*natural language processing*, NLP) è una sottobranca dell'IA che si occupa dell'analisi, della comprensione e della generazione di parole e frasi scritte e parlate (linguaggio naturale). La maggior parte delle tecniche di elaborazione del linguaggio naturale sono basate sull'«apprendimento automatico»: estraggono il significato e il contesto dal linguaggio umano. Le aree di applicazione comprendono ad esempio il riconoscimento del testo (*text recognition*) il riconoscimento vocale (*speech recognition*). i bot, i *chatbot* e gli assistenti digitali. *Fonte:* https://cnai.swiss/it/servizi/terminologia/

7 Ispirato a «*Generative AI Framework for UK Government*» (del 18.01.2024); v. https://www.gov.uk/government/publications/generative-ai-framework-for-hmg (in inglese)

ampliate le vostre conoscenze! Con un po' di creatività, contribuirete così a rendere l'amministrazione innovativa. Procedete però con cautela, rifletteteci bene, fate uso di spirito critico e rispettate le disposizioni vigenti e i principi seguenti.

1) Sapere cos'è l'IA generativa, cosa sono gli LLM e quali sono i loro limiti.
2) Utilizzare le tecnologie di IA generativa (come gli LLM) in modo etico, responsabile e conforme alle normative vigenti. Verificare se esiste una base legale sufficiente per l'utilizzo previsto. I requisiti per la base legale differiscono a seconda che si utilizzino tecnologie di IA generativa (come gli LLM) solo lateralmente per aiutarsi nella traduzione di un testo, oppure cercando un livello di supporto maggiore per redigere testi o ancora se le si utilizzano per prendere una decisione in modo completamente automatizzato.
3) Sapere come mantenere sicuro l'utilizzo delle tecnologie di IA generativa (come gli LLM).[8]
4) Prevedere un controllo da parte di esseri umani quando opportuno e nella giusta fase (dallo sviluppo all'applicazione di tecnologie di IA generativa, come gli LLM).
5) Utilizzare in modo trasparente l'IA in generale e gli LLM in particolare. Il testo creato dagli LLM può dare l'impressione di essere stato scritto da un essere umano. In qualità di dipendenti dell'Amministrazione federale, avete una responsabilità particolare per quanto riguarda la comunicazione interna ed esterna. Si raccomanda di indicare chiaramente quando un testo è stato generato dall'IA (ad es. scrivendo «Il presente testo è stato creato con il supporto dell'IA, poi modificato e controllato»). In caso di decisioni individuali automatizzate ai sensi dell'articolo 21 LPD tale indicazione è addirittura obbligatoria[9].
6) Applicare questi principi insieme alle Linee guida in materia di IA destinate all'Amministrazione federale e insieme ai principi di base della Confederazione per una scienza dei dati (e un'IA) incentrata sull'essere umano e affidabile; assicurare i necessari controlli (di qualità).

Limiti dei modelli linguistici di grandi dimensioni basati sull'IA (LLM)

Gli LLM creano i loro output o risultati scegliendo la successione di parole in base alla probabilità statistica, senza quindi tenere conto del tenore di verità del contenuto prodotto. Questo può portare a mancata fattualità e ad allucinazioni, cioè ad output o risultati che non corrispondono alla realtà. Il problema può essere limitato fornendo un contesto al modello (ad es. un insieme di documenti in formato testuale) in modo che possa utilizzarlo per trovare la risposta. Tuttavia, possono verificarsi allucinazioni anche in questo caso, qualora le riposte includano contenuti al di fuori di detto contesto. Altri limiti degli LLM che ne possono inficiare il corretto utilizzo includono ad esempio:[10]

- output o risultati indesiderati, memoria semantica e distorsioni nel modello (*bias*) derivanti dalla composizione e dal contenuto non equilibrati dei dati di addestramento (ad es. dati personali o protetti da copyright oppure testi dal contenuto discutibile, falso o discriminatorio);
- risultati obsoleti, poiché gli LLM generano testo elaborando i dati di addestramento, che sono per forza di cose limitati ai contenuti già esistenti al momento dell'addestramento del rispettivo modello;
- mancanza di riproducibilità, poiché anche se un LLM riceve ripetutamente lo stesso input, l'output o il risultato generato può essere ogni volta diverso sia in termini linguistici che di contenuto;
- mancanza di sicurezza per il codice generato (se un LLM è stato addestrato a creare codici, ne può anche generare alcuni che potenzialmente potrebbero contenere falle di sicurezza);
- reazione errata a input specifici, poiché anche differenze apparentemente insignificanti negli input possono portare a grandi differenze negli output o nei risultati generati. Ad esempio, nei casi in cui gli input

[8] Per saperne di più sull'utilizzo di queste tecnologie negli strumenti di IA generativa, v. «Promemoria per l'utilizzo di strumenti di IA generativa nell'Amministrazione federale»; https://cnai.swiss/it/it-servizi-altri-servizi-promemoria-sullia/

[9] https://www.fedlex.admin.ch/eli/cc/2022/491/it#art_21

[10] Per la stesura di questo elenco è servito come base il capitolo «2.3 Risiken von LLMs» della versione 1.1 dell'11.03.2024 del rapporto dell'Ufficio federale tedesco per la sicurezza informatica (Bundesamt für Sicherheit in der Informationstechnik) sulle opportunità e i rischi degli LLM; v. https://www.bsi.bund.de/SharedDocs/Downloads/DE/BSI/KI/Generative_KI-Modelle.pdf

forniti a un LLM si discostano dai testi utilizzati per l'addestramento, spesso il modello non è più in grado di elaborarli correttamente e genera output o risultati errati;

- Generalmente gli LLM generano testi impeccabili dal punto di vista linguistico, convincenti in termini di contenuto e pertinenti in un'ampia varietà di ambiti. Questo può dare l'impressione che le capacità della macchina siano simili a quelle degli esseri umani e quindi incoraggiare l'utente a riporre troppa fiducia nei risultati e nelle capacità del modello (*automation bias*);
- frequente riutilizzo dei dati di input come dati di addestramento (il che comporta quindi l'utilizzo per l'addestramento anche di dati oggettivamente errati immessi dagli utenti);
- predisposizione a interpretare il testo come un'istruzione, poiché gli LLM interpretano tutti gli input allo stesso modo e non distinguono tra istruzioni e altri testi;
- mancanza di riservatezza dei dati inseriti, in quanto gli LLM sono spesso offerti come servizio su Internet attraverso apposite interfacce (ad es. tramite un browser web). L'inserimento di dati personali in tale servizio via Internet può costituire una trasmissione di dati a terzi. Questi casi sono disciplinati dalla legge sulla protezione dei dati;
- lo sviluppo e il funzionamento di LLM da parte di fornitori sulla loro infrastruttura può comportare un elevato grado di dipendenza (portando ad es. alla mancanza di sovranità sui dati o alla mancanza di controllabilità sul modello);
- generazione di fonti talvolta imprecise o addirittura inesistenti da parte degli LLM, alle quali gli LMM poi si attengono in modo persistente e convincente.

Le tecnologie di IA generativa (come gli LLM) offrono un ampio ventaglio di opportunità e possibili applicazioni, e stanno attualmente vivendo un periodo di intenso sviluppo. Da ciò consegue la necessità di imporre nuovi requisiti in termini di sicurezza, sia per quanto riguarda lo sviluppo, che per il funzionamento e l'utilizzo di questi modelli. Per utilizzare queste tecnologie di AI in modo sicuro, è essenziale analizzare sistematicamente i rischi.

A tal fine, ogni utente che lavora presso l'Amministrazione federale è tenuto a effettuare un'analisi dei rischi per ogni tecnologia di IA generativa (come gli LLM) che desidera integrare nei propri processi di lavoro quotidiani. Per farlo, può utilizzare i principi appena elencati come quadro di riferimento per effettuare una sperimentazione responsabile.

Per porre domande:

- sull'IA nell'Amministrazione federale: gruppo di lavoro IA e poli di competenza all'interno della Rete di competenze per l'IA (CNAI)
- sulla sicurezza dell'informazione e la protezione dei dati: ISIO e CPDO della vostra unità amministrativa
- su servizi specifici nell'ambito della scienza dei dati e dell'IA: DSCC

Altre indicazioni

Linee guida in materia di IA destinate all'Amministrazione federale

Le sette linee guida in materia di IA destinate all'Amministrazione federale sono sempre in vigore: centralità dell'essere umano, garantire condizioni quadro favorevoli allo sviluppo e all'utilizzo dell'IA, trasparenza, tracciabilità e comprensibilità, definire chiaramente le responsabilità, garantire la sicurezza, promuovere la partecipazione attiva alla governance del settore IA coinvolgendo tutti gli attori importanti a livello nazionale e internazionale.

Codice di comportamento della Confederazione per una scienza dei dati (e un'IA) incentrata sull'essere umano e affidabile

Attraverso il Codice di comportamento, una sorta di guida orientativa, le unità amministrative della Confederazione vengono sensibilizzate nei confronti dei principi fondamentali della scienza dei dati (e di un'IA) incentrata sull'essere umano e affidabile, da una parte fornendo loro spiegazioni pratiche, e dall'altra rendendole capaci di attuare tali principi nel loro lavoro quotidiano. I principi fondamentali sono: protezione dei dati e delle informazioni, sicurezza dei dati e delle informazioni, governance dei dati, non discriminazione, spiegabilità, tracciabilità, trasparenza, riproducibilità, neutralità, oggettività e trattamento etico di dati e risultati.

Il promemoria è stato messo a punto dal gruppo di lavoro «IA nell'Amministrazione federale» del CNAI con la partecipazione di rappresentanti di tutti i Dipartimenti e della CaF.