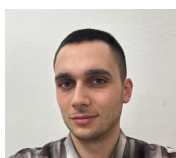


IA di qualità per dati affidabili e di livello

Passata, forse, la prima fase di scoperta da parte del grande pubblico dell'Intelligenza Artificiale (AI), spesso fatta di paure oppure di infatuazioni estreme, oggi la riflessione e lo sviluppo delle tecnologie legate all'intelligenza artificiale generativa si sta spostando sulla sostanza di queste tecnologie, ossia la qualità e la completezza dei sistemi IA. E una grande discussione è sulla base dei dati usati per l'addestramento delle IA.

DOI 10.12910/EAI2025-024



di Sergio Ferraris, Direttore di Nextville.it
e Alessandro Baltaro, Consulente e ricercatore presso ReteAmbiente

Passata, forse, la prima fase di scoperta da parte del grande pubblico dell'Intelligenza Artificiale (AI), spesso fatta di paure oppure di infatuazioni estreme, oggi la riflessione e lo sviluppo delle tecnologie legate all'intelligenza artificiale generativa si sta spostando sulla sostanza di queste tecnologie, ossia la qualità e la completezza dei sistemi IA. E una grande discussione è sulla base dei dati usati per l'addestramento delle IA.

Non è una questione di poco conto ben nota a chi utilizza e addestra AI per i segmenti professionali. Chi possiede basi di dati, o di contenuti, qualificati da tempo rivendica diritti sulle elaborazioni delle IA. È emblematico, sotto a questo profilo il caso che ha contrapposto OpenAI e il New York Times riguardo una controversia legale sul diritto d'autore, specificamente sull'uso non autorizzato di contenuti protetti da copyright per addestrare i modelli di intelligenza artificiale di

OpenAI, come ChatGPT.

Il New York Times ha intentato una causa federale contro OpenAI e Microsoft il 27 dicembre 2023, accusandoli di violazione del diritto d'autore per l'uso non autorizzato di milioni di articoli del NYT, sostenendo che l'uso di questi contenuti senza consenso o pagamento costituisce una violazione del copyright su larga scala e minaccia la capacità del giornale di sostenere il proprio modello di business, inclusi investimenti nel giornalismo di qualità.

OpenAI e Microsoft sostengono che l'uso dei dati rientra nella dottrina del cosiddetto "fair use", che consente l'uso di materiale protetto da copyright per scopi educativi, di ricerca o commento, a condizione che l'elaborazione sia trasformativa e non compete con l'originale.

L'accesso all'archivio del NYT, e comunque di qualsiasi giornale con decenni d'esperienza, rappresentano per le IA un efficace "antidoto" alle

cosiddette allucinazioni che le IA hanno avuto spesso durante la loro prima fase d'addestramento. **Le IA sono a caccia sistematica di contenuti, o dati, di qualità con le quali addestrare i propri sistemi. Rimanendo sempre nel campo dell'editoria, altri editori, come Associated Press e Vox Media, hanno scelto una strada diversa, raggiungendo accordi con OpenAI per condividere i propri contenuti.**

Dati di qualità

Un'evoluzione di questo dibattito è quello delle intelligenze artificiali generative verticali, basate su archivi selezionati e qualitativi che rappresentano un'evoluzione specifica dell'IA generativa, che si concentra su compiti specialistici e settoriali.

Si tratta di sistemi che utilizzano dataset selezionati e qualitativi per addestrare modelli che possono generare contenuti altamente specializzati e pertinenti a un segmento del lavoro della conoscenza specifi-

co, soddisfacendo bisogni precisi e puntuali, come l'ottimizzazione dei processi di produzione, anche sulle filiere della produzione di contenuti oppure sulla consulenza o della catena di approvvigionamento, materiale o immateriale. Per esempio, in ambito legale, l'IA generativa può essere adattata per effettuare il fine-tuning per gestire documentazione specifica, procedure aziendali o di sistema. I vantaggi circa l'utilizzo di archivi selezionati si possono riassumere in tre punti:

- la precisione, visto che i modelli addestrati su archivi selezionati e qualitativi generano contenuti più precisi e pertinenti a un settore specifico;
- l'efficienza, poiché l'uso di dataset mirati riduce il tempo e le risorse necessarie per l'addestramento, migliorando l'efficienza complessiva del sistema;
- l'adattabilità perché diventa possibile un facile adattamento a nuove esigenze specifiche del settore attraverso un fine-tuning.

Approcci verticali

In sintesi, le intelligenze artificiali generative verticali, basate su archivi selezionati e qualitativi, offrono un approccio personalizzato e efficiente per generare contenuti altamente specializzati, migliorando la precisione e l'efficienza in vari settori.

Se pensiamo ai vantaggi portati da queste tecnologie all'interno del settore, sicuramente il mondo normativo spunta subito tra i maggiori candidati per applicazioni future di modelli AI verticali, che possano essere una

guida affidabile in uno scenario come quello della normativa italiana, le cui luci e ombre possono essere paragonate, senza dubbi, a quelle di un inferno dantesco.

Considerando la delicatezza degli argomenti trattati, risulta difficile ai più affidarsi a sistemi artificiali per consigli di carattere giuridico e normativo, un campo minato in cui ad ogni passo si rischiano sanzioni o cause legali; c'è però una realtà, storica nel settore, che ha trovato in questo campo una nuova frontiera da esplorare, e che nei test condotti con alcuni gruppi ristretti di esperti in occasione di Key Energy a Rimini, ha già mostrato risultati più che sorprendenti.

Esperienza diretta

Si tratta di Nextville, dell'Osservatorio Normativa Ambientale e della Rivista Rifiuti, facenti tutte parte del gruppo ReteAmbiente, società che è stata tra le prime a creare database ordinati nei minimi dettagli dedicati alla normativa energetica nel primo caso, e a quella ambientale negli altri, con testi coordinati curati e aggiornati quotidianamente dalle diverse redazioni, che possono contare su figure autorevoli e di spessore.

Partendo da questa base, grazie alla collaborazione con un team di sviluppo dedicato, l'implementazione di modelli IA diventa, nella sua complessità, molto più semplice; tramite una ricerca vettoriale fatta su misura, questi modelli navigano all'interno dei database proprietari che coprono la normativa comunitaria, nazionale e

regionale, con sezioni dedicate a bandi, sanzioni, giurisprudenza e così via, fornendo non solo tutti i riferimenti di legge necessari, ma garantendo anche un'armonizzazione tra tutti i testi, nelle relazioni reciproche e nei "combinati disposti", affidabile, precisa e dettagliata, che in questa fase sperimentale è riuscita a impressionare i più esperti e i più scettici.

Lavoro integrati

Quello che a prima vista potrebbe sembrare un altro sistema IA comparabile ai più famosi citati già in precedenza, risulta invece soltanto un lontano parente: le prime sono dei tuttologi, come quei soggetti che su Facebook pensano di essere campioni in ogni materia, che fanno un po' di tutto ma nulla in particolare, e che basano le loro risposte su contenuti presenti nei propri database ma spesso di dubbia provenienza. In questo caso invece, ci troviamo di fronte a due esperti di normativa che elaborano risposte, consigli e analisi altamente affidabili basate sui materiali prodotti ogni giorno, è bene ricordarlo, dalle redazioni umane, come commenti, guide, focus e dossier dedicati.

Una sorta di nuovo modello ibrido nel quale il lavoro umano si fonde con quello dell'AI, coniugando la produttività dell'AI con il controllo e le competenze degli esperti. Una logica oltretutto contrapposta alla vulgata che vuole il lavoro delle persone in competizione con quello delle IA. Una strada che con ogni probabilità è quella da seguire per aumentare l'accettabilità sociale delle IA.