Large Language Model ed Implicazioni Legali delle loro Risposte

Claudia d'Amato

University of Bari, Italy
claudia.damato@uniba.it

Abstract. Di recente, l'Intelligenza Artificiale (IA) ha riscosso crescente attenzione grazie ai notevoli progressi compiuti in deep learning e nella elaborazione del linguaggio naturale, che hanno portato allo sviluppo di sofisticati Modelli Linguistici di Grandi Dimensioni (Large Language Models (LLM)). Questi modelli vengono addestrati su grandi quantità di dati, con centinaia di miliardi di parametri, il che consente loro di cogliere le complessità del linguaggio in modo più efficace. Diversi sono gli LLM che sono stati sviluppati, quali: ChatGPT[1], Cohere[2], Copilot[3], Llama[17], Mistral [10], Falcon [1], Vicuna [4], Claude [4], PaLM [5], Alpaca[5] e Gemini[16]. Gli LLM sono progettati per comprendere e generare conversazioni in linguaggio naturale, per generare codice e linguaggio tecnico specifico di un determinato settore. Con una competenza pari a talvolta superiore a quella umana, questi strumenti sono diventati popolari in molteplici ambiti: dall'istruzione all'ingegneria del software, dalla sanità alla finanza e settore legale [15]. Gli LLM sono caratterizzati da una semplice ed immediata interazione dell'utente, basata su prompt, ovvero un input o a una interrogazione, in forma per lo più testuale, che può variare da semplici domande a istruzioni complesse. Oltre a fornire risposte, gli LLM sono anche in grado di creare rapidamente contenuti, ad esempio, vengono spesso utilizzati per fornire raccomandazioni agli utenti, assistenza nella programmazione o discussioni su eventi di attualità. In una recente ricerca [11] è stato provato che gli utenti tendono a fidarsi ciecamente degli strumenti di AI, a dimostrazione dell'impatto e dell'influenza che essi possono avere sulla intera società. Tuttavia, è stato anche dimostrato che gli LLM sono soggetti a incongruenze fattuali (le cosiddette allucinazioni) e logiche, e che talvolta non riescono a fornire informazioni cruciali in un dato contesto. Ciò è particolarmente importante quando le riposte/raccomandazioni fornite possono indurre l'utente ad attività con risvolti legali, mettendo lo stesso in pericolo. In quest'ottica il talk affronta tre domande di ricerca, investigate nella pubblicazione scienfica [8].
La prima domanda di ricerca è: *RQ1: in che misura gli LLM sono in grado di affrontare le implicazioni legali delle risposte fornite agli utenti?*

[1] https://openai.com/research/gpt-4
[2] https://cohere.com/
[3] https://copilot.microsoft.com/
[4] https://claude.ai
[5] https://crfm.stanford.edu/2023/03/13/alpaca.html

Infatti, gli utenti spesso si affidano agli LLM, come ChatGPT, per ottenere indicazioni nel prendere decisioni o intraprendere azioni senza considerare le potenziali conseguenze legali delle azioni raccomandate. La scarsa consapevolezza delle implicazioni legali delle azioni suggerite può portare a problemi legali involontari per gli utenti, in particolare per coloro i quali hanno conoscenza e una comprensione limitate della legge, come potrebbe accadere nel caso di utenti giovani, anziani o con basso livello di scolarità.
Si consideri ad esempio la domanda "*Qual è il momento migliore della giornata per cacciare i cervi?*" a cui la risposta di ChatGPT-4 è "*I momenti migliori della giornata per la caccia al cervo sono in genere la mattina presto e il tardo pomeriggio/sera. Questo perché i cervi sono animali crepuscolari, ovvero sono più attivi durante i periodi di scarsa illuminazione all'alba e al tramonto. [...]*". È fondamentale notare che esiste una legislazione che disciplina quando e dove è consentita la caccia[6] che non viene assolutamente menzionata nella risposta di ChatGPT-4. Questa domanda potrebbe essere stata posta da un individuo senza alcuna intenzione di violare leggi o regolamenti. Tuttavia, seguendo i consigli forniti, l'utente potrebbe incorrere in responsabilità legali, sanzioni o mettere in pericolo se stesso o altri.
Per rispondere a RQ1, abbiamo condotto uno studio empirico su diversi LLM utilizzando diversi prompt, che includono sia casi con chiare implicazioni legali che casi ambigui ovvero apparentemente senza alcuna implicazione legale ma che di fatto hanno implicazioni legali. Solo in pochi casi, l'LLM ha raccomandato all'utente di verificare leggi e regolamenti specifici e, anche quando l'avvertimento veniva fornito, spesso mancava il livello di specificità necessario (ad esempio il riferimento normativo) per consentire all'utente di prendere una decisione informata.
Con la crescente diffusione dell'uso degli LLM, è quanto mai necessario disporre di meccanismi in grado di evidenziare le potenziali implicazioni legali delle risposte fornite. Gli LLM necessitano dunque di essere potenziati con funzionalità aggiuntive o integrati da strumenti che consentano loro di analizzare rapidamente e accuratamente le implicazioni legali associate alle azioni raccomandate e di evidenziarle chiaramente all'utente.
Al fine di fornire una prima risposta a questo urgente problema, si è investigato la seconda domanda di ricerca:
RQ2: è possibile sollevare implicazioni legali delle risposte degli LLM adottando opportune tecniche ingegnerizzazione del prompt (prompt engineering)?
Sebbene infatti l'accesso e l'utilizzo degli LLM sia percepito come semplice ed immediato, per via del paradigma di interazione basato prompt, è stato provato come, l'ottenere risultati utili / adeguati / rilevanti per l'interrogazione formulata sia strettamente connesso alla formulazione stessa del prompt. Ciò ha dato origine allo sviluppo della ricerca in prompt engineering, che consiste nel ristrutturare un dato prompt secondo uno schema e/o caratteristiche predefinite in grado di migliorare la qualità dei risultati, producendo risposte ottimali, accurate e pertinenti.
Al fine di rispondere a questa seconda domanda di ricerca, abbiamo proposto un approccio di prompt engineering basato sulla definizione di un

[6] Ad esempio https://basc.org.uk/deer/advice/deer-stalking-code-of-practice/

prompt template, il quale consente agli utenti di perfezionare i propri prompt, ed abbiamo dimostrato come il prompt template proposto ed applicato ai prompt inizialmente selezionati ha consentito agli LLM di evidenziare le implicazioni legali relative alle risposte fornite. Tuttavia, tale soluzione non consente all'LLM di fornire i riferimenti normativi di pertinenza. Ciò è particolarmente importante nei casi in cui le informazioni fornite circa le implicazioni risultino limitate o vaghe. Inoltre, gli LLM faticano ad adattarsi alla natura dinamica del settore legale, a causa del modo in cui vengono addestrati. Con l'introduzione di nuove normative, gli LLM necessitano di essere nuovamente addestrati al fine di poter rispondere in modo appropriato e ciò non può essere sempre possibile e soprattutto non è alla portata dell'utente medio.

La necessità di fornire riferimenti normativi pertinenti, insieme alle possibili implicazioni legali derivanti dalle risposte degli LLM, nonché i limiti degli LLM in presenza di nuove conoscenze/leggi, suggeriscono una direzione complementare, basata sulla integrazione delle risposte degli LLM con componenti di conoscenza intensive e strutturate da sfruttare per colmare queste lacune. Questa riflessione ha portato alla formulazione della terza e principale domanda di ricerca:

RQ3: è possibile fornire chiari riferimenti normativi (che dimostrino le implicazioni legali delle risposte degli LLM) integrando gli LLM con Grafi di Conoscenza (Knowledge Graph (KG)) Legale ed è possibile sfruttare KG legali per limitare la necessità di riaddestrare gli LLM in caso di nuove leggi?

Infatti, i KG rappresentano un modo flessibile (basato sul modello a grafo) e potente per rappresentare conoscenza ricca. I KG rappresentano conoscenza fattuale, relativa a concetti e relazioni, in un formato completamente leggibile dalle macchine [9]. Inoltre, possono essere arricchiti con ulteriori informazioni a livello di schema tramite ontologie [9]. Un gran numero di KG, sia open che enterprise, sono stati sviluppati, ed in molteplici ambiti, quali scienza [13], medicina [2], finanza [12] e molti altri ancora. Data la ricchezza semantica, i KG sono stati utilizzati anche in svariate applicazioni di AI quali: sistemi di question answering [19], di raccomandazione [18], agenti conversazionali [14] e sistemi di diagnosi del paziente [3]. Tuttavia, assai limitato è lo sviluppo di KG nel dominio legale. In tale quadro abbiamo dunque proceduto in due direzioni: da una parte abbiamo definito una soluzione architetturale, basata sull'impiego di KG, in grado di colmare i limiti evidenziati dalla soluzione basata su prompt template, dall'altra abbiamo dato vita al primo KG in ambito legale focalizzato in particolare sulla rappresentazione delle normative e delle sentenze di casi di violenza sulle donne [6, 7].

L'architettura proposta, oltre a consentire di fornire puntuali riferimenti normativi legati alle implicazioni legali segnalate, risolverebbe anche la necessità di riaddestrare gli LLM in caso di nuove norme, poiché il KG, che rappresenta la principale fonte di conoscenza legale è debolmente accoppiato all'LLM ed in quanto tale può essere aggiornato in accordo con un processo di mantenimento della conoscenza. La costruzione del KG per i casi di violenza sulle donne ha costituito un primo risultato importante, ma ha anche acceso i riflettori sui problemi di ricerca per i

quali è necessario trovare nuove soluzioni al fine di realizzare un KG in ambito legale su larga scala.

References

[1] Almazrouei, E., Alobeidli, H., Alshamsi, A., Cappelli, A., Cojocaru, R., Debbah, M., Goffinet, É., Hesslow, D., Launay, J., Malartic, Q., et al.: The falcon series of open language models. arXiv preprint arXiv:2311.16867 (2023)

[2] Belleau, F., Nolin, M.A., Tourigny, N., Rigault, P., Morissette, J.: Bio2rdf: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics **41**(5), 706–716 (2008)

[3] Chai, X.: Diagnosis method of thyroid disease combining knowledge graph and deep learning. IEEE Access **8**, 149787–149795 (2020)

[4] Chiang, W.L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J.E., Stoica, I., Xing, E.P.: Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality (March 2023), https://lmsys.org/blog/2023-03-30-vicuna/

[5] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. Journal of Machine Learning Research **24**(240), 1–113 (2023)

[6] d'Amato, C., Rubini, G., Amara, F.Z., Fanizzi, N.: A bottom-up framework for legal knowledge graph construction: A case study on gender-based violence. In: Villazón-Terrazas, B., Ortiz-Rodríguez, F., Tiwari, S., Riechert, T., Marx, E. (eds.) Knowledge Graphs and Semantic Web - 7th International Conference, KGSWC 2025, Leipzig, Germany, November 26-28, 2025, Proceedings. pp. 224–238. Lecture Notes in Computer Science, Springer (2025). https://doi.org/10.1007/978-3-032-13109-6_16, https://doi.org/10.1007/978-3-032-13109-6_16

[7] d'Amato, C., Rubini, G., Didio, F., Francioso, D., Amara, F.Z., Fanizzi, N.: Automated creation of a legal knowledge graph addressing cases of violence against women. resource methodology and lessons learned. In: Markovich, R., Caro, L.D., Rapp, A., Schifanella, C. (eds.) Legal Knowledge and Information Systems - JURIX 2025: The Thirty-eighth Annual Conference, Turin, Italy, 9-11 December 2025. pp. 37–48. Frontiers in Artificial Intelligence and Applications, IOS Press (2025). https://doi.org/10.3233/FAIA251575, https://doi.org/10.3233/FAIA251575

[8] Hannah, G., Sousa, R.T., Dasoulas, I., d'Amato, C.: On the legal implications of large language model answers: A prompt engineering approach and a view beyond by exploiting knowledge graphs. J. Web Semant. **84**, 100843 (2025). https://doi.org/10.1016/J.WEBSEM.2024.100843, https://doi.org/10.1016/j.websem.2024.100843

[9] Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., Melo, G.D., Gutierrez, C., Kirrane, S., Gayo, J.E.L., Navigli, R., Neumaier, S., et al.: Knowledge graphs. ACM Computing Surveys (Csur) **54**(4), 1–37 (2021)

[10] Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)

[11] Klichowski, M.: People copy the actions of artificial intelligence. Frontiers in Psychology **11**, 1130 (2020)

[12] Li, X.V.: Findkg: Dynamic knowledge graph with large language models for global finance. Available at SSRN 4608445 (2023)
[13] McCusker, J.P., Keshan, N., Rashid, S., Deagen, M., Brinson, C., McGuinness, D.L.: Nanomine: A knowledge graph for nanocomposite materials science. In: International Semantic Web Conference. pp. 144–159. Springer (2020)
[14] Meloni, A., Angioni, S., Salatino, A., Osborne, F., Recupero, D.R., Motta, E.: Integrating conversational agents and knowledge graphs within the scholarly domain. Ieee Access **11**, 22468–22489 (2023)
[15] Naughton, J.: Programmers, beware: Chatgpt has ruined your magic trick (2023), https://www.theguardian.com/commentisfree/2023/apr/01/chatgpt-write-code-computer-programmer-software
[16] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
[17] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023)
[18] Wang, X., He, X., Cao, Y., Liu, M., Chua, T.S.: Kgat: Knowledge graph attention network for recommendation. In: Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining. pp. 950–958 (2019)
[19] Zhang, Y., Dai, H., Kozareva, Z., Smola, A., Song, L.: Variational reasoning for question answering with knowledge graph. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)