LE COMMUNITY AI TEMPI DEGLI LLM *

Alfonso Piscitelli
Dipartimento di Informatica
Università degli Studi di Salerno

ABSTRACT

Le comunità online hanno svolto un ruolo centrale nello sviluppo del software open-source, contribuendo alla crescita di progetti collaborativi attraverso strumenti come mailing list, IRC e forum. Con l'emergere dei Large Language Models (LLM), come GPT-4, si è osservato un cambiamento significativo nel modo in cui gli utenti interagiscono con piattaforme di condivisione del sapere, come StackOverflow. Questi modelli, in grado di generare risposte tecniche in tempo reale, hanno portato a una drastica riduzione delle domande poste su tali piattaforme e a una diminuzione del traffico complessivo. Questo articolo presenta una panoramica storica delle comunità online, seguita da una discussione sugli LLM e sul loro impatto sul numero di contributi in piattaforme tecniche. Viene poi esplorato il fenomeno dell'"over-trusting", ovvero l'eccessiva fiducia riposta nelle risposte generate dagli LLM, che spesso risultano parziali o inaccurate. L'articolo mette in luce come l'affidamento crescente agli LLM possa minare la creazione di conoscenza condivisa, tipica delle comunità online. Sebbene gli LLM offrano soluzioni rapide e personalizzate, resta cruciale il contributo umano per garantire la qualità e la validità delle informazioni.

***K*eywords** First keyword · Second keyword · More

1 Introduzione

Il mondo del software libero, e di Linux in particolare, deve parte del suo successo alla comunità (*community*). Le *community*, gruppi di persone che si riuniscono intorno ad una tematica di interesse comune, permettono ancora oggi l'avanzamento di tanti software *open-source*; in una community, ogni persona ha un ruolo e ognuno può contribuire con le proprie capacità. Infatti, è possibile contribuire testando il software e segnalando malfunzionamenti, o anche realizzando nuovi strumenti rendendoli a loro volta liberi, risolvendo bug degli strumenti già esistenti e, non meno, finanziando (liddove è necessario) dei progetti.

Sebbene oggigiorno le community utilizzano strumenti moderni, che supportano diversi metodi di comunicazione (come le piattaforme *Slack*, *Discord*, per citarne alcune), le prime community hanno mosso i primi passi attraverso le *mailing-list* e, quasi contemporaneamente, con le chat *IRC* (alcune di queste ancora attive). Molte delle mailing-list sono ancora pubblicamente disponibili nei vari archivi. I programmatori (ma più in generale gli informatici) hanno trovato la loro casa nel portale *StackOverflow*, portale ideato da *Jeff Atwood* e *Joel Spolsky* con l'obiettivo di realizzare una piattaforma dove era possibile scambiarsi domande e risposte (e che tali risposte, alla fine, restassero disponibili per altri che, in futuro, ne avrebbero avuto bisogno). StackOverflow ha raccolto oltre 24 milioni di domande ad Ottobre 2024 e oltre 35 milioni di risposte, tutte disponibili gratuitamente.

L'avvento di modelli di Intelligenza Artificiale come *gpt-3.5* e *gpt-4*, di OpenAI, più comunemente conosciuti con il nome del portale, *ChatGPT*, ha messo in evidenza alcune caratteristiche di questi modelli, come la capacità di generare del testo (anche poesie!), rispondere a domande e finanche generare del codice in diversi linguaggi di programmazione. L'obiettivo di questo articolo è presentare, come l'avvento degli LLM (e l'integrazione di questi negli strumenti di lavoro) stia influenzando le community: analizzeremo in particolare il caso di StackOverflow, che rende disponibili al pubblico i dati sulle domande e le risposte sul suo portale, ma è possibile trasportare queste riflessioni anche su altre community visti i numeri coinvolti.

* *Citation*: Piscitelli Alfonso, *Le community ai tempi degli LLM*. Linux Day 2024. Ottobre 2024. Benevento (Italia).

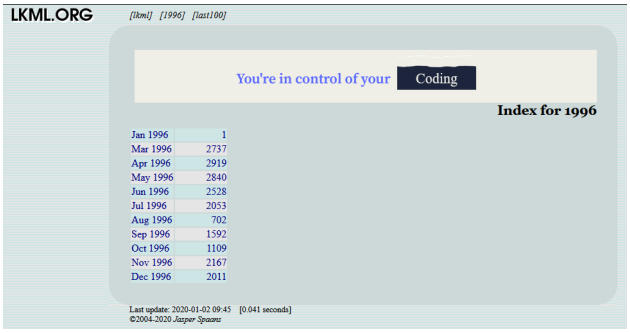

Figure 1: Schermata di LKML con l'archivio dei messaggi del 1996

Il resto di quest'articolo è strutturato nel seguente modo. La Sezione 2 servirà a fare un rapido excurcus storico sugli strumenti che le community hanno utilizzato nel corso degli anni, mentre nella Sezione 3 darò alcune definizioni di base sull'intelligenza artificiale e sugli strumenti oggi disponibili (sia disponibili open source che no). Nella Sezione 4 saranno analizzati i dati su domande e risposte fatte su StackOverflow, mentre nella Sezione 5 evidenzierò alcune limitazioni dei modelli e come queste limitazioni potrebbero ulteriormente influenzare le community. Le conclusioni, nella Sezione 6, riassumeranno i contenuti dell'articolo.

2 Da IRC a StackOverflow, passando per i Forum

Condividere informazioni spinge da sempre l'uomo a cercare metodi, via via sempre più tecnologici, per lasciare traccia ai posteri. E' avvenuto con le forme di comunicazione preistorica (l'arte rupestre), e poi con la scrittura, la stampa ecc. La rete ha aperto la possibilità di comunicare, con tempi impensabili già solo cinquant'anni fa, con tutto il mondo (e, a dire il vero, anche oltre). Una delle prime forme di community in rete sono state le *mailing-list*.

Una mailing-list è un sistema di distribuzione di messaggi via email a un gruppo di persone iscritte, permettendo la condivisione di informazioni su un tema specifico tra i membri. Le mailing-list sono nate negli anni '70, con l'avvento di Internet e delle prime reti di comunicazione digitale. Una delle prime mailing-list famose è stata "SF-LOVERS", creata nel 1975 per gli appassionati di fantascienza. Durante gli anni '90, le mailing-list sono diventate popolari per la discussione di argomenti accademici, tecnologici e sociali. Tra le più note, "**Linux Kernel Mailing List**" (LKML) è stata fondamentale per lo sviluppo collaborativo del sistema operativo Linux. Molte *mailing list* hanno reso pubblici i propri archivi, consentendo anche a chi si iscrive in un tempo successivo (e con l'avvento del Web anche ai non iscritti) di leggere le discussioni. Ad esempio, il sito https://lkml.org contiene l'archivio dei messaggi della mailing list di LKML, dove è possibile leggere tutti i messaggi dal 1996.

Nelle mailing list, come suggerisce il nome, **tutti gli iscritti possono inviare messaggi e tutti gli altri iscritti ricevono il messaggio inviato**. Questo crea un flusso di comunicazione aperto e accessibile a tutti i membri della lista. Tuttavia, la natura **asincrona** della comunicazione tramite mailing list può rallentare la discussione e la formazione di un senso di comunità forte. L'avvento delle chat, in particolare l'IRC (*Internet Relay Chat*) sviluppato da Jarkko Oikarinen nell'agosto del 1988 ha consentito di evolvere la comunicazione online velocizzandola e rendendola sincrona, permettendo di comunicare in tempo reale. Questa immediatezza ha contribuito a migliorare l'idea di comunità, facilitando interazioni più fluide e spontanee. Gli utenti IRC si riunivano in "canali" tematici, replicando in un certo senso la suddivisione per argomenti tipica delle mailing list, ma godendo di una maggiore reattività e dinamicità.

IRC, acronimo di Internet Relay Chat - traducibile come *chat attraverso internet*, è un sistema di messagistica testuale, creato nel 1988 nell'Università finlandese di Oulu. Questo sistema consente sia di scambiare messaggi in un *area pubblica* (l'equivalente un po' di un gruppo WhatsApp per intenderci) sia di scambiarsi messaggi privati. Ma non solo, tramite IRC era possibile (e lo è tutt'oggi) trasferire file. Sebbene oggi l'idea di una chat non sembri così innovativa, va considerato che nel 1988 il Web non era ancora realtà, e che la rete Internet era ad appannaggio esclusivo di enti di ricerca e militari.

L'impatto di IRC è stato significativo, soprattutto nei primi anni di internet, contribuendo alla creazione di community online e alla diffusione di notizie e informazioni in tempo reale. L'uso esteso del Web a partire dagli anni duemila, l'avvento dei siti web, di piattaforme per la gestione automatica dei contenuti (come *Wordpress, Joomla, ecc.*) ha portato alla nascita dei primi *forum*.

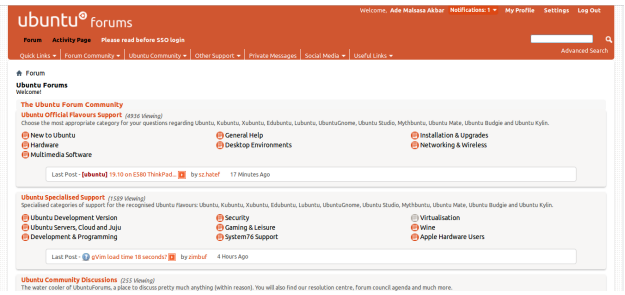

Figure 2: Schermata del forum di Ubuntu, la distribuzione GNU/Linux supportata da Canonical.ltd

Un *forum* è una piattaforma online che consente a persone con interessi comuni di interagire e scambiarsi opinioni su vari argomenti. La struttura dei forum permette agli utenti di avviare discussioni (chiamate *thread*) su un tema specifico, cui altri possono rispondere con commenti, favorendo così un dialogo aperto e collaborativo. Questo modello di comunicazione differisce dai social network in quanto i forum tendono a essere organizzati attorno a tematiche precise, consentendo un approfondimento maggiore.

Gli anni successivi, 2010-2020, hanno rappresentato un ulteriore cambio di paradigma: l'avvento dei social media (Facebook, Twitter tra i tanti), delle applicazioni di messagistica (in particolare di Telegram) hanno permesso di raggiungere altri utenti (specie i meno abituati a canali meno immediati o intuitivi) e di creare nuove community per i settori più disparati. Potremmo dire che oggi sul web esista una community per qualsiasi cosa!

Tutti questi strumenti visti finora hanno in comune una cosa: la persistenza delle informazioni e la condivisione delle stesse ad altri membri della community. Sembra un fatto da poco, ma l'esistenza di una conoscenza condivisa, disponibile per tutti, è un fattore importante per una community, perchè:

- **Promuove un senso di appartenenza**: La condivisione di conoscenze e la partecipazione alle discussioni creano un senso di identità collettiva e rafforzano i legami tra i membri della community.
- **Facilita l'apprendimento collaborativo**: L'accesso a un archivio di informazioni e la possibilità di interagire con altri membri esperti su un determinato argomento favoriscono l'apprendimento e la crescita individuale.
- **Stimola la creazione di nuove idee**: La condivisione di diverse prospettive e la discussione di idee in un ambiente aperto possono portare alla nascita di nuove soluzioni e all'innovazione.
- **Aiuta a risolvere problemi**: I membri della community possono aiutarsi a vicenda nella risoluzione di problemi specifici, attingendo alle conoscenze e alle esperienze condivise.

3 Large Language Models, una overview

In questa sezione presenterò brevemente alcune definizioni collegate all'Intelligenza Artificiale che utilizzerò nel resto di questo articolo, con lo scopo chiarificatore rispetto alla confusione cui generalmente si assiste dalla stampa.

L'Intelligenza Artificiale, di per sè, rappresenta la disciplina che studia strumenti, tecniche (e tecnologie) che consentono di realizzare un sistema artificiale che riproduca l'intelligenza umana. Oggi siamo in grado di realizzare sistemi che funzionano piuttosto bene in task specifici, ma siamo ben lontani dal poter realizzare in tutto e per tutto l'intelligenza umana. Certo, i *Large Language Models* (LLMs) come *gpt-4*, meglio conosciuto al pubblico come *ChatGPT*, sembrano aver *appreso* la capacità di linguaggio degli esseri umani, dimostrando di poter generare testi di senso compiuto, poesie, articoli, con ottimi risultati. Ma è una caratteristica intrinseca di un LLM.

Un LLM, come sopra *Large Language Model*, è appunto un modello addestrato su una base di conoscenza (leggi *dataset*) molto ampia (spesso di diversi *TeraByte*), da qui appunto il termine *Large*. Questi modelli, come GPT (*Generative Pre-trained Transformer*), vengono addestrati su enormi quantità di testo per apprendere le regole grammaticali, il significato delle parole e la struttura delle frasi, permettendo loro di rispondere a domande, creare contenuti, tradurre testi e svolgere una vasta gamma di compiti linguistici.

Ne esistono diversi, i più famosi sono sicuramente quelli prodotti da OpenAI come *gpt-3.5, gpt-4o* e l'ultimo in ordine di uscita *gpt-o1*, ma esistono anche gli LLM prodotti da Claude come *claude-3-5-sonnet*, quello di Google, *gemini*, e quelli utilizzabili anche in locale come *llama3* di Facebook e *mixtral*. Per approfondimenti sugli LLM si vedano [1, 2] e quest'articolo di Medium che raccoglie alcuni testi a mio avviso interessanti [3].

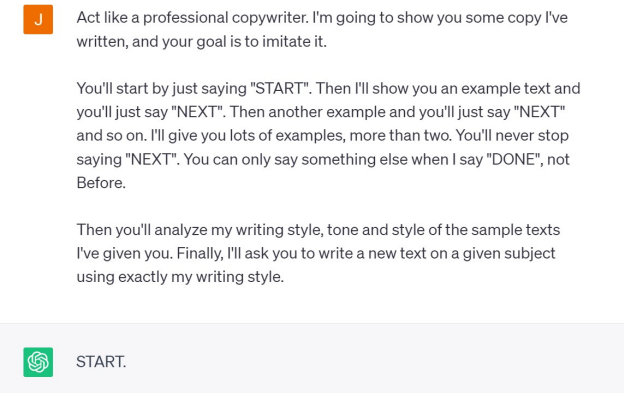

Figure 3: Prompt strutturato di esempio per ChatGPT

L'interazione con questa tipologia di modelli avviene attraverso domande, come in una chat, con il modello che rispode alla domanda posta dall'utente. Queste domande possono essere più strutturate, come i *Prompt*, consentendo ai Large Language Model di risolvere problemi diversi e più complessi della classica generazione di testo [4]. Sui Prompt trovo sia illuminante questo sito [5], che descrive le varie strategie di prompt.

4 StackOverflow e il calo delle domande

4.1 StackOverflow

Stack Overflow [6] è una piattaforma online dedicata alla condivisione di conoscenze tecniche, in particolare nel campo della programmazione e dello sviluppo software. Fondata nel 2008, si basa su un sistema di domande e risposte, dove gli utenti possono porre quesiti su problemi di codice, tecnologie e linguaggi di programmazione, ricevendo risposte da una vasta comunità di esperti e appassionati. Una delle caratteristiche distintive di Stack Overflow è il sistema di reputazione: gli utenti possono guadagnare punti e riconoscimenti in base alla qualità delle loro risposte o domande, incentivando la creazione di contenuti utili e accurati. La piattaforma è molto apprezzata per l'efficacia e la precisione delle risposte, diventando una risorsa essenziale per programmatori di ogni livello, dai principianti ai professionisti.

La forza di Stack Overflow risiede nella struttura comunitaria e nel suo modello di moderazione collaborativa. Gli utenti non solo contribuiscono con risposte, ma hanno anche il potere di votare le risposte migliori e segnalare contenuti non conformi, assicurando così una qualità elevata. Inoltre, grazie all'archiviazione di tutte le domande e risposte, Stack Overflow funge da vasto database di conoscenze tecniche, accessibile a chiunque in cerca di soluzioni. Nell'immagine 4 è possibile riconoscere la Homepage del portale, e il numero di domande ad oggi presenti sul sito: oltre 24 milioni!

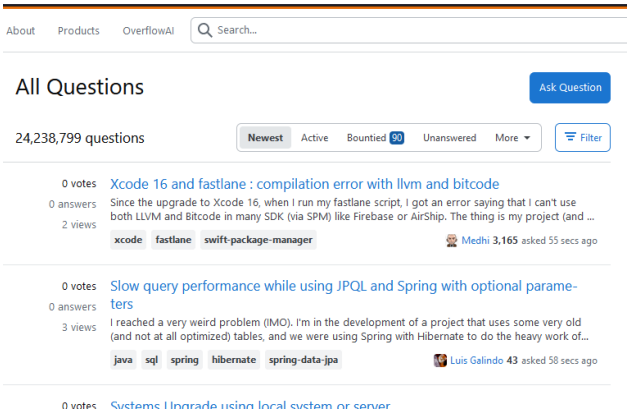

Figure 4: StackOverflow Homepage

4.2 Il calo dei contributi

Le statistiche di StackOverflow consentono di analizzare alcuni parametri come: le pagine viste, il numero di nuove domande, il numero di risposte aggiunte alle domande, i voti ecc. I dati sono pubblicamente disponibili, e in questo articolo farò riferimento ai grafici realizzati da Ayan. [7].

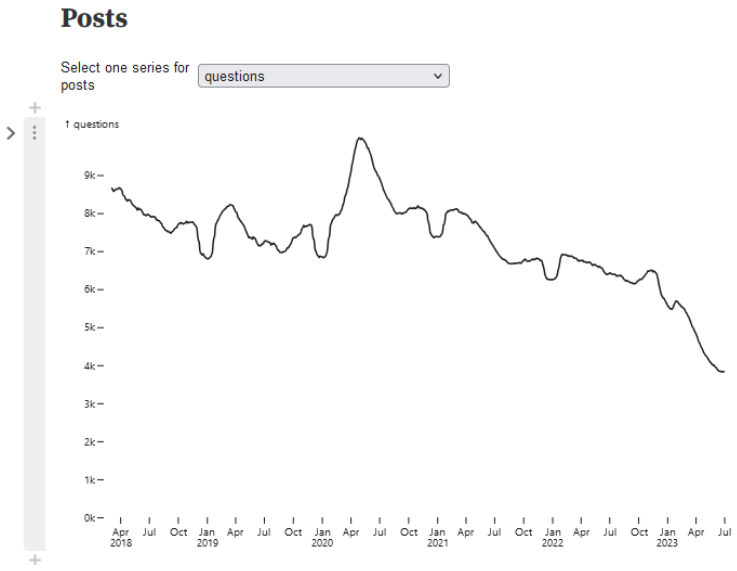

Figure 5: Grafico riguardo il numero di domande presenti su StackOverflow. Fonte: [7]

Il grafico in Figura 5 mostra l'andamento del numero di domande poste su Stack Overflow nel corso del tempo. Si nota che, a parte un picco intorno ad aprile 2020, il numero di domande è rimasto relativamente stabile, oscillando tra le settemila e le ottomila domande al giorno. Tuttavia, da novembre 2022 la curva cambia drasticamente, registrando un crollo significativo: si passa dalle circa settemila domande di novembre a meno di quattromila nel luglio 2023. È prevedibile che i dati relativi al 2024 seguiranno questo trend discendente.

Novembre 2022 segna un momento cruciale per lo sviluppo dei Large Language Models (LLM), con il rilascio di *GPT-3.5* e la diffusione di ChatGPT al grande pubblico. Questo evento ha portato una svolta soprattutto per la sua capacità di rispondere a domande tecniche, riducendo di conseguenza la necessità di porre quesiti su piattaforme come Stack Overflow. Il rilascio di modelli successivi (come *gpt-4o*) sempre più accurati e addestrati su quantità di informazioni maggiori, hanno migliorato la qualità delle risposte tecniche: in molti casi ChatGPT è preferito a StackOverflow, garantendo risposte di qualità in tempi molto più rapidi.

Se il numero di persone che pone domande su StackOverflow sta diminuendo significativamente, anche il numero delle persone che cerca informazioni sul sito si sta riducendo.

Il grafico in Figura 6 illustra l'andamento del numero di pagine viste su Stack Overflow nel tempo. Anche in questo caso, si osserva una riduzione significativa a partire da novembre 2022: il numero di pagine viste è sceso da circa 16 milioni nel novembre 2022 a circa 9 milioni nel luglio 2023. Questo rappresenta un calo di oltre il 40%, suggerendo una tendenza sempre più marcata degli utenti a rivolgersi a strumenti come ChatGPT, piuttosto che cercare se la loro domanda sia già stata risolta su Stack Overflow. Questo cambiamento di comportamento suggerisce una preferenza per soluzioni più rapide e personalizzate offerte dagli LLM, che riduce la necessità di esplorare forum di domande e risposte

5 Implicazioni - Il fenomeno dell'over-trusting

I dati presentati nella Sezione 4 mostrano la tendenza degli utenti ad evitare l'uso delle community come primo posto dove ricercare informazioni e chiedere aiuto, tendendo ad una privatizzazione dei propri problemi e della ricerca della soluzione. Questa tendenza all'isolamento richiede ulteriori approfondimenti e ricerche, che non sono oggetto di questo articolo.

Di contro assisteremo in futuro a modelli di intelligenza artificiale (non solo LLMs) in grado di supportare il lavoro e di risolvere una maggiore gamma di problemi che oggi in ogni caso richiedono il confronto tra persone (che avviene

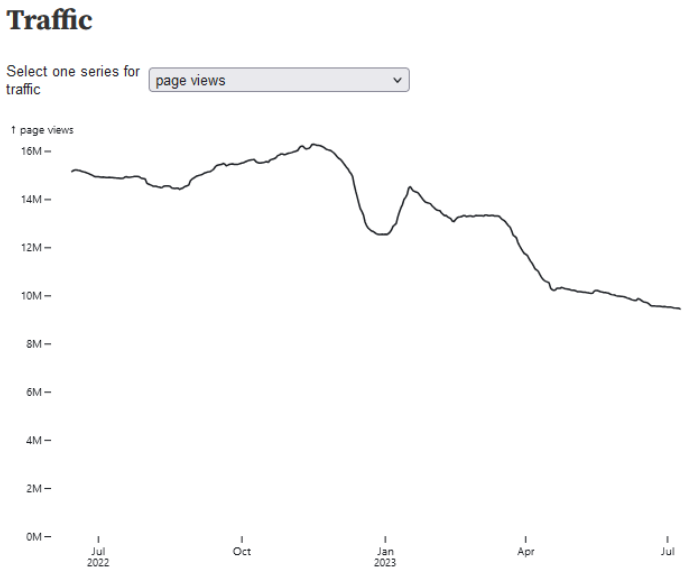

Figure 6: Grafico riguardo il numero di visitatori su StackOverflow. Fonte: [7]

prevalentemente nelle community). La crescita delle capacità di questi strumenti sta già portando al fenomeno dell'*over-trusting*, dell'eccessiva fiducia dell'essere umano data alla risposta generata. Nell'articolo di Prather e altri autori [8], pubblicato nella rivista *ACM Transactions on Computer-Human Interaction*, è stato condotto uno studio con studenti del corso di programmazione, che ha evidenziato come si tenda a fidarsi molto delle risposte degli LLM al punto da chiedere direttamente le soluzioni di un determinato problema. Questo fenomeno è stato visto maggiormente in dettaglio in un altro studio, condotto dal mio gruppo di ricerca, che sarà pubblicato a breve nei *Proceedings of the 16th International Conference on Education Technology and Computers* [9].

L'eccessiva fiducia comporta anche, dall'altro lato, la difficoltà nella comprensione delle risposte generate e di poter eventualmente aggiustarle (o nel caso del codice, trovare errori/bug e correggerli). C'è poi un discorso legato all'affidabilità delle risposte ricevute. Su questo fronte, lo studio condotto da Kabir Samia presentato alla conferenza *ACM Conference on Human Factors in Computing Systems* [10], ha evidenziato che molto spesso le risposte ottenute da ChatGPT contengono delle informazioni errate (circa il 50% delle volte) e che nel 70% dei casi queste risposte sono particolarmente prolisse. Questo stile, comunque, aiuta a generare fiducia da parte degli utenti che hanno preferito in molti casi le risposte (incorrette) di ChatGPT.

Un ulteriore aspetto da considerare è la privatizzazione delle problematiche e delle relative soluzioni. Quando si chiede aiuto su un forum come Stack Overflow, la domanda e la risposta rimangono disponibili nel tempo, permettendo ad altri utenti di usufruirne anche in seguito. Questo approccio favorisce la creazione di una conoscenza collettiva: le discussioni tra esperti (e non) permettono di esaminare diverse soluzioni e di considerare aspetti inizialmente non presi in esame, arricchendo il valore delle risposte.

Questo flusso collaborativo viene a mancare nell'interazione con ChatGPT e modelli simili. Le risposte fornite rimangono salvate nell'account personale dell'utente, ma sono inaccessibili a terzi, a meno che non vengano esplicitamente condivise. Inoltre, queste risposte non vengono indicizzate dai motori di ricerca, rendendole invisibili a chi potrebbe trovarsi, successivamente, di fronte allo stesso problema. In questo senso, l'approccio basato sui LLM riduce il contributo alla conoscenza pubblica, creando soluzioni individualizzate, ma non accessibili alla comunità nel suo insieme.

6 Conclusioni

In questo articolo è stato presentato un excursus storico sulle community, in particolare del mondo Linux, per poi fare riferimento ad una delle community più famose nel campo della *Computer Science*, StackOverflow. L'articolo ha mostrato i dati relativi agli accessi al portale e al numero di domande poste al giorno, mostrando una significativa riduzione iniziata con la diffusione di ChatGPT. Le implicazioni di questi dati sono state discusse nella Sezione 5 di questo lavoro, analizzando come le persone tendano a fidarsi troppo delle risposte ottenute dai vari LLM, nonostante questi spesso diano informazioni imparziali o scorrette [10]. Le community, quindi, rivestono ancora un ruolo

fondamentale per la condivisione delle informazioni e la crescita individuale di ogni persona. Le capacità dei modelli di linguaggio (gli LLM) potrebbero essere fornite a supporto della comunità, magari integrati in dei forum, dove altri utenti possano poi ritrovare quell'informazione generata, valutarla (in pieno stile StackOverflow) e magari integrarla, generando nuove idee e discussioni.

Ringraziamenti

Un ringraziamento sincero al *Laboratorio per l'Informatica Libera Sannita* per aver accettato questa proposta di intervento su un tema ritengo centrale per il futuro delle community e del Software Libero. Ringrazio poi ciascuno di voi che è arrivato fin qui: per ogni commento, delucidazione o suggerimenti vari, è possibile contattarmi via mail *alfonso.piscitelli91@gmail.com*.

References

[1] https://www.elastic.co/what-is/large-language-models.

[2] Gokul Yenduri, Ramalingam M, Chemmalar Selvi G, Supriya Y, Gautam Srivastava, Praveen Kumar Reddy Maddikunta, Deepti Raj G, Rutvij H Jhaveri, Prabadevi B, Weizheng Wang, Athanasios V. Vasilakos, and Thippa Reddy Gadekallu. Generative pre-trained transformer: A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions, 2023.

[3] Abonia Sojasingarayar. Books on llm and nlp in 2024 - https://medium.com/@abonia/books-on-llm-and-nlp-in-2024-aea05617d557, April 2024.

[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[5] https://www.promptingguide.ai/.

[6] https://https://stackoverflow.com/.

[7] Ayhan Ç. The fall of stack overflow - https://observablehq.com/@ayhanfuat/the-fall-of-stack-overflow?ref=blog.pragmaticengineer.com, July 2023.

[8] James Prather, Brent N. Reeves, Paul Denny, Brett A. Becker, Juho Leinonen, Andrew Luxton-Reilly, Garrett Powell, James Finnie-Ansley, and Eddie Antonio Santos. "it's weird that it knows what i want": Usability and interactions with copilot for novice programmers. *ACM Trans. Comput.-Hum. Interact.*, 31(1), November 2023.

[9] Alfonso Piscitelli, Gennaro Costagliola, Mattia De Rosa, and Vittorio Fuccella. Influence of large language models on programming assignments – a user study. *Proceedings of the 16th International Conference on Education Technology and Computers*, 2024.

[10] Samia Kabir, David N. Udo-Imeh, Bonan Kou, and Tianyi Zhang. Is stack overflow obsolete? an empirical study of the characteristics of chatgpt answers to stack overflow questions. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.