Tavola Rotonda: Quali strategie per gli LLM italiani?

Chair: Roberto Navigli

Il dibattito sui Large Language Model (LLM) si apre con l'intervento del prof. Roberto Navigli, che, in qualità di moderatore, invita sul palco i partecipanti: Il dott. Raniero Romagnoli, CTO di Almawave, il dott. Vincenzo Masucci, Direttore dei progetti finanziati europei e responsabile della sede di Napoli di ExpertAI, il dott. Sanzio Bassini, Direttore del Dipartimento Supercalcolo Applicazioni e Innovazione CINECA, la dott.ssa Roberta Piscitelli, Academic Research Lead EMEA, presso Amazon Web Services, e il dott. Claudio Stamile, Fastweb.

Il prof. Navigli anticipa che porrà tre domande ai partecipanti, focalizzandosi sui loro piani per gli LLM e sull'importanza di disporre di LLM pre addestrati in italiano. Con questa premessa, invita i partecipanti a condividere le loro progettualità e la loro visione in questo ambito, avviando così la discussione della tavola rotonda.

Il dott. Raniero Romagnoli, CTO di Almawave, descrive l'esperienza e l'approccio dell'azienda nel Natural Language Processing. Attiva da quasi 15 anni, Almawave sviluppa tecnologie NLP per grandi aziende e pubbliche amministrazioni. Il dott. Romagnoli sottolinea l'importanza del loro laboratorio di ricerca e sviluppo, che si dedica alla ricerca applicata e industriale. Con l'avvento dei modelli di attenzione e dei transformer, Almawave ha iniziato a implementare propri large language model (LLM). Tuttavia, ha riscontrato problemi nell'uso di LLM, come la gestione di dati confidenziali, il rischio di bias e l'affidabilità dei risultati. Molte aziende usano modelli cloud senza adattarli alle loro esigenze, causando ulteriori problematiche. Per affrontare queste sfide, Almawave sviluppa i propri modelli LLM, offrendoli in modalità open parameter e on premise. Questa strategia permette una maggiore personalizzazione e sicurezza nell'uso dei dati, rispondendo alle esigenze specifiche dei clienti e migliorando l'efficacia delle soluzioni.

Interviene poi il dott. Vincenzo Masucci il quale, da laureato in Fisica indirizzo cibernetico alla Federico II di Napoli, esprime il suo piacere di essere stato invitato ad un convegno sull’intelligenza artificiale a Napoli. Il dott. Masucci ha spiegato che Expert.ai, come azienda italiana concentrata da diverse decine di anni sulla comprensione del linguaggio naturale, ha seguito da vicino l'evoluzione degli LLM sin dai primi sviluppi, integrando queste tecnologie nei propri servizi. L'azienda ha avviato sperimentazioni con ChatGPT, ma ha anche sviluppato un proprio LLM utilizzando Mixtral per superare problemi legati alla privacy. Questo modello è utilizzato in particolare nella generazione di linguaggio, migliorando l'eloquio e l'efficacia dei servizi offerti ai clienti. La possibilità di disporre di LLM pre addestrati specificamente per l'italiano rappresenta un significativo vantaggio per sviluppare servizi che garantiscono una migliore qualità del linguaggio per gli utenti italiani.

La dott.ssa Roberta Piscitelli di AWS, dopo aver ricordato il suo legame con l'Università Federico II di Napoli, che ha organizzato Ital-IA 2024, sottolinea che AWS riconosce l'enorme potenziale dei Large Language Models (LLM) per costruire applicazioni innovative .AWS ha sviluppato i propri LLM (Titan) ma soprattutto si posiziona come partner strategico per lo sviluppo, la commercializzazione e la costruzione di applicazioni basate su LLM.

AWS collabora a livello globale con organizzazioni governative, centri di ricerca, università e aziende, offrendo una suite di servizi di AI generativa per sviluppare e commercializzare nuovi modelli, connettendosi anche all'ecosistema delle startup e alle applicazioni per la pubblica amministrazione. In particolare in Europa, AWS supporta l'addestramento dei modelli interamente sulla propria piattaforma e fornisce la potenza computazionale dei centri di calcolo nazionali in vari paesi, supportando progetti in modo ibrido. Il supporto di AWS si estende anche alle fasi successive, come l'hosting e la specializzazione dei modelli. In Francia, ad esempio, sono state sviluppate applicazioni per la pubblica amministrazione che permettono ai cittadini di chiedere informazioni su vari argomenti, come il rinnovo dei passaporti. Per quanto riguarda l'Italia, AWS riconosce l'importanza di avere modelli addestrati in lingua locale per mantenere la sovranità digitale e gli aspetti culturali di un paese e promuovere l'innovazione a livello nazionale, e sottolinea la collaborazione con vari attori italiani, tra cui FastWeb, che ha annunciato che i modelli pre addestrati per l'italiano saranno disponibili su AWS attraverso Bedrock, il servizio che consente di utilizzare LLM e modelli di base tramite un'API per creare nuove applicazioni.

Il dott. Sanzio Bassini, Direttore del Dipartimento Supercalcolo, Applicazioni e Innovazione di CINECA, inizia il suo intervento ricordando il legame con Napoli, dove CINECA sta consolidando la sua presenza. CINECA ha inoltre stabilito una collaborazione strutturata con FAIR, generando risultati significativi, ed ha iniziato un'azione a livello europeo supportando progetti quali Mistral, addestrato per l'80% sui sistemi di CINECA. Il dott. Bassini evidenzia che CINECA ha già avviato collaborazioni con start-up per sviluppare LLM open source, addestrati sui loro sistemi, specificamente testuali e multimodali. L'obiettivo principale è supportare la ricerca scientifica e l'innovazione, inclusa la produzione e i servizi delle pubbliche amministrazioni, rendendole in grado di estrarre valore dai dati per il beneficio dei cittadini. Per la ricerca scientifica, il dott. Bassini sottolinea la crescente applicazione di metodi di machine learning ed AI nei vari domini scientifici, nonostante pochi ricercatori si classifichino come AI. Infatti, quasi il 40% dei progetti di ricerca recenti applicano metodi AI, dimostrando la necessità di un ambiente abilitante. CINECA prevede di supportare la ricerca scientifica di eccellenza e l'innovazione tecnologica, sia per le pubbliche amministrazioni sia per la produzione, promuovendo anche start-up tecnologiche e migliorando settori come la guida autonoma e l'agri-food.

Il dott. Claudio Stamile di Fastweb interviene rispondendo alla domanda iniziale sull'importanza di produrre o predisporre Large Language Models (LLM) in Italia, sottolineando che non è cruciale avere un LLM in sé, quanto piuttosto creare un ecosistema italiano che supporti tali modelli. Fastweb, oltre a promuovere bandi di ricerca, si impegna a supportare questo ecosistema, coinvolgendo università, iniziative private e studenti. Il dott. Stamile sottolinea che l'Italia è ricca di talenti che già lavorano su LLM e altre iniziative tecnologiche. Molti progetti italiani vengono citati in articoli scientifici grazie all'entusiasmo e all'iniziativa di giovani ricercatori. Il suo messaggio è rivolto a questi giovani, incoraggiandoli a restare in Italia per contribuire a queste iniziative piuttosto che cercare opportunità all'estero. Fastweb mira a diventare una "AI factory", seguendo l'esempio di NVIDIA, che ha citato Fastweb come esempio di iniziativa privata per creare un ecosistema di intelligenza artificiale in Italia.

Il prof. Navigli ribadisce l’importanza per l’Italia di produrre i propri LLM per non dipendere esclusivamente da risorse esterne. Evidenzia che, sebbene i modelli attuali siano efficaci, ci

sono rischi legati a cambi di licenza e restrizioni di accesso. Per questo, è cruciale che l'Italia sviluppi le competenze necessarie per creare e gestire i propri modelli di intelligenza artificiale. Pone quindi la seconda domanda ai partecipanti, chiedendo quali applicazioni industriali siano maggiormente influenzate dai LLM, evidenziando l'impatto rivoluzionario che questi modelli hanno avuto sulla percezione dell'intelligenza artificiale da parte del pubblico.

Il dott. Romagnoli sottolinea come l'arrivo di ChatGPT abbia influenzato significativamente il mercato, portando molte aziende a sperimentare con l'idea che un semplice prompt o una serie di prompt potessero risolvere problemi complessi. Tuttavia, nell'ambito industriale delle grandi aziende, è emerso chiaramente che l'applicazione finale richiede un mix di diversi modelli, non limitandosi ad una singola tecnologia. Questo approccio multitecnologico è cruciale perché le grandi aziende devono fornire risposte ai loro clienti in modo sofisticato, evitando risposte troppo semplificate. Specifica quindi che l'uso dell'AI è stato spesso idealizzato come un ragionatore o un oracolo, come nel caso di strumenti quali Copilot o ChatGPT. Almawave ha integrato queste tecnologie in modo che i modelli di ricerca e le risposte possano interagire tra loro, permettendo un dialogo continuo con il sistema. Questa convergenza ha consentito ad Almawave di sviluppare nuovi prodotti verticali che il mercato ha accolto positivamente. Il dott. Romagnoli sostiene che occorre semplificare l'accesso alla conoscenza aziendale tramite tecnologie come quella degli LLM, eliminando la necessità di conoscenze approfondite in SQL o navigazione dati complessa. Questo permette ai dipendenti ed agli utenti di interagire con l'azienda in modo intuitivo e diretto, facilitando le interrogazioni e migliorando l'efficienza complessiva. Infine, riflette sul trend attuale in cui alcune aziende cercano di utilizzare l'AI generativa come una "scatola magica", un approccio che ha portato a diversi proof of concept poco realistici. Almawave cerca di indirizzare le richieste verso soluzioni pragmatiche, sebbene ciò possa rappresentare una sfida data la persistente attrazione per l'innovazione nel campo dell'intelligenza artificiale.

Il dott. Masucci descrive un scenario in cui attraverso l'uso di tecnologie all'avanguardia si possa perfezionare la realizzazione dei sistemi esperti in auge diversi anni fa. In particolare, usando tecniche di comprensione semantica del linguaggio naturale insieme a tecniche non simboliche come, ad esempio, le reti neurali e gli LLM, si è in grado, con ottimi risultati, di generare dei sistemi esperti su un particolare dominio. Come caso concreto, a partire dalla conoscenza contenuta in manuali d'uso, per apparecchiature mediche in formato pdf, come le macchine per risonanza magnetica, TAC ed altro, si crea un l sistema esperto di macchine di diagnostica che è in grado di fornire risposte in linguaggio naturale sia ai tecnici specializzati sia agli operatori ospedalieri. In questo caso l'uso di LLM consente un eloquio più naturale. Integrando poi, l'uso di tecniche di comprensione del linguaggio naturale basato su regole, si garantisce tracciabilità, spiegabilità e affidabilità. L'approccio usato ha fornito già risultati promettenti in diversi ambiti su cui concentrare gli sviluppi futuri.

La dott.ssa Piscitelli interviene delineando diversi casi d'uso in cui i large language model (LLM) stanno rivoluzionando diverse applicazioni a livello industriale, apportando significativi miglioramenti nella customer experience, nella produttività dei dipendenti e ottimizzazione dei processi aziendali, nella creazione di contenuti. Evidenzia quindi che questi modelli sono cruciali per migliorare l'esperienza del cliente attraverso chatbot e assistenti virtuali su piattaforme come Booking.com e Fox Media, consentendo l'elaborazione di feedback non strutturati per personalizzare raccomandazioni e migliorare il coinvolgimento degli utenti. Menziona poi l'impiego degli LLM per aumentare la produttività aziendale, con assistenti

virtuali per i dipendenti e con la generazione di software a partire da descrizioni, come esemplificato da Amazon Q Developer, citando Accenture che sta utilizzando Amazon Q Developer ed ha ridotto il costo di sviluppo software del 30%, permettendo ai sviluppatori di concentrarsi maggiormente su sicurezza, qualità e prestazioni, e sottolinea quindi l'importanza della Generative Business Intelligence per analizzare dati aziendali e fornire raccomandazioni strategiche, citando Deloitte come utilizzatore di questo strumento. La dott.ssa Piscitelli inoltre illustra come gli LLM siano utilizzati per la generazione automatizzata di contenuti come email di marketing, blog post e contenuti educativi interattivi, come nel caso di Smile & Learn. Infine, sottolinea l'impatto di queste tecnologie sul business di AWS, evidenziando strumenti per l'AI responsabile e la personalizzazione dei modelli attraverso Bedrock e SageMaker. Questi casi d'uso mostrano come gli LLM stiano trasformando vari settori industriali migliorando l'efficienza, l'esperienza utente e l'innovazione nei processi aziendali.

Il dott. Bassini ricorda l'applicazione degli LLM in vari settori industriali e scientifici, evidenziando la complessità e le potenzialità di questi modelli. Egli illustra come un collega nel campo del clima utilizzi gli LLM per integrare dati e formulare equazioni, allo scopo di migliorare le previsioni climatiche. Questo approccio permette di accelerare la risoluzione di problemi complessi non completamente descrivibili attraverso metodi deterministici tradizionali. Prosegue poi esaminando diversi esempi di applicazioni industriali degli LLM. Nel settore manifatturiero, in aziende di packaging, gli LLM aiutano a determinare se i dispositivi necessitano di una nuova progettazione o se possono essere utilizzati componenti standard, migliorando l'efficienza del processo produttivo. Nel settore di energia elettrica: utilizzando dati di consumo e previsione, gli LLM possono ottimizzare il brokeraggio e la produzione di energia elettrica, migliorando la gestione delle risorse energetiche. Nell'industria farmaceutica: combinando i dati relativi ai principi attivi e alle proteine segnale delle patologie, gli LLM facilitano la ricerca e sviluppo di nuovi farmaci, migliorando la precisione e l'efficacia delle cure. Il dott. Bassini specifica che questi esempi rappresentano un tentativo di innovare non solo l'interfaccia utente, ma anche il contenuto dei prodotti stessi, contribuendo a un'evoluzione significativa nei settori della produzione, dell'energia e della sanità. La combinazione di dati sintetici, linguaggio naturale e guida autonoma rappresenta un ulteriore esempio di come gli LLM possano rivoluzionare vari ambiti tecnologici e industriali.

Il dott. Stamile discute l'impatto che l'implementazione di chatbot basati su intelligenza artificiale generativa ha avuto sull'azienda, sia in termini di supporto clienti che di ottimizzazione dei processi interni. Fastweb ha lanciato un chatbot AI sulla homepage per due settimane nel mese di novembre 2023. Attraverso questo esperimento, l'azienda ha potuto acquisire una comprensione approfondita delle complessità e delle necessità di guard railing, migliorando la sicurezza e l'efficacia dei propri sistemi AI. Questa esperienza ha influenzato lo sviluppo di sistemi di supporto per gli operatori del customer care, migliorando l'efficienza e la qualità del servizio clienti. Fastweb ha sviluppato servizi analoghi a ChatGPT per uso interno, rispettando le normative sulla privacy e la proprietà intellettuale. Questi strumenti aiutano i dipendenti a svolgere compiti come il riassunto di documenti in modo più rapido ed efficiente, accelerando significativamente i processi di lavoro e riducendo il tempo necessario per svolgere compiti complessi. Un risultato importante è stata la creazione e l'espansione di una cultura dell'AI all'interno dell'azienda. I dipendenti non solo hanno riconosciuto i vantaggi dell'uso degli strumenti AI, ma hanno anche imparato a comprenderne i limiti, migliorando così l'adozione consapevole e responsabile di queste tecnologie. La cultura del dato e dell'AI ha

portato i dipendenti a richiedere nuove funzionalità e a contribuire attivamente al miglioramento degli strumenti interni, evidenziando un coinvolgimento proattivo e un'adozione diffusa della tecnologia. In sintesi, l'introduzione di chatbot e strumenti AI ha trasformato sia il modo in cui Fastweb interagisce con i clienti sia il funzionamento interno dell'azienda, promuovendo una cultura dell'innovazione e dell'efficienza.

Il Prof. Navigli introduce quindi il tema dei dati protetti da copyright, utilizzati sia nell'addestramento iniziale dei modelli che in eventuali fasi di fine tuning successive. Chiede ai partecipanti di esprimere le loro opinioni su questo argomento, invitandoli a discutere ciò che ritengono giusto o sbagliato e le loro strategie o proposte per affrontare questa questione delicata.

Il dott. Romagnoli riprende il tema dei dati introdotto da Fastweb, riconoscendo che molti clienti non comprendono ancora appieno il funzionamento dei modelli di linguaggio di grandi dimensioni. Sottolinea l'importanza della trasparenza nell'uso dei dati e della consapevolezza da parte degli utenti su dove finiscono i loro dati. Almawave ha implementato politiche interne per educare i dipendenti sui confini dell'utilizzo dei dati, soprattutto in relazione alla proprietà intellettuale aziendale e ai dati caricati su piattaforme cloud.
Per quanto riguarda l'addestramento dei modelli, il dott. Romagnoli dichiara che Almawave evita di utilizzare dati protetti da copyright, cercando di pulire i dataset da tali dati e stipulando accordi con i titolari dei diritti, se necessario. Pur essendo un'azienda italiana, Almawave opera in diversi paesi, inclusi Africa e Sud America, dove le regolamentazioni possono variare. Tuttavia, l'azienda adotta la normativa più restrittiva, spesso quella europea, per garantire il rispetto dei diritti d'autore. Conclude affermando che, nonostante la possibilità di errori dovuti alla difficoltà di filtrare grandi quantità di dati, l'intenzione di Almawave è di evitare l'uso di dati protetti da diritto d’autore.

Il dott. Masucci affronta la questione copyright creando un distinguo tra dati utilizzati per l'addestramento dei modelli linguistici e i dati dei clienti usati per fare inferenza, spiegando come caso concreto che, per utilizzare i dati del cliente, come il manuale per la risonanza magnetica, Expert.ai ha dovuto sviluppare un modello linguistico interno, poiché il detentore del copyright del manuale ne ha vietato la pubblicazione online. Questo ha comportato la necessità di lavorare localmente con i dati. Per quanto riguarda i dati di training dei large language model, il dott. Masucci sottolinea l'importanza di rispettare le leggi sulla proprietà intellettuale. Afferma che i modelli linguistici attuali non generano nuova conoscenza, quindi, è essenziale proteggere i diritti di chi crea e condivide conoscenza. Tuttavia, esprime la speranza che un giorno i modelli linguistici possano effettivamente produrre nuova conoscenza, il che giustificherebbe la vendita di tale conoscenza a chi ne usufruisce. L’uso delle leggi sulla proprietà intellettuale proteggerebbe in questo caso anche gli sviluppatori dei modelli stessi.

La dott.ssa Piscitelli spiega che AWS gestisce il fine tuning dei modelli esistenti creando una copia del modello in una rete privata virtuale per il cliente. I dati utilizzati per il fine tuning non vengono usati da AWS per riaddestrare il modello principale. Per quanto riguarda l'addestramento su dati protetti da copyright, AWS richiede che gli utenti abbiano le licenze necessarie e seguano le policy per l'AI responsabile. Sottolinea che l'addestramento di un modello su dati protetti da copyright non implica automaticamente una violazione dei diritti di copyright nell'output del modello. Questo dipende dal tipo di dati usati, da come vengono

elaborati e dall'output generato. I modelli sviluppati da AWS ed i servizi di AI gestiti da AWS (come Amazon Q, Q developer professional e Titan su Bedrock), sono costruiti in modo da minimizzare il rischio di violazione di copyright. Nella remota ipotesi che ci sia una violazione dei diritti di copyright da terze parti, AWS offre un indennizzo e copertura legale per i propri clienti. Infine, la dott.ssa Piscitelli chiarisce che AWS adotta un modello di responsabilità condivisa: AWS è responsabile della qualità del servizio, mentre la responsabilità di rispettare le policy di AI responsabile, avere le licenze appropriate e gestire l'output ricade sui clienti.

Il dott. Bassini descrive l'infrastruttura dati del CINECA, che comprende due data lake da 50 petabyte ciascuno, uno a Bologna e uno al centro di San Giovanni a Teduccio, oltre ad un sistema di storage di svariate centinaia di petabyte. Questi data lake fungeranno da repository per dati pubblici, accessibili alla comunità e conformi alle certificazioni necessarie (ISO 27001, ISO 28000, GDPR). Per quanto riguarda i dati, verrà pubblicato un catalogo di dati pubblici, che includerà metadati e dati di valore, provenienti da fonti come Istat, Protezione Civile, Banca d'Italia e camere di commercio. È in corso un progetto legato al PNRR che mira a combinare questi dati per generare modelli di linguaggio di grandi dimensioni (LLM) e produrre semilavorati utili, come valutazioni dell'impatto del rischio di eventi estremi. Un esempio pratico presentato dal dott. Bassini è l'uso dei dati per analizzare gli effetti di un terremoto, includendo dati sull'evento, la ricostruzione e i relativi costi. Questi dati possono essere utilizzati per sviluppare strumenti a valore industriale, come polizze assicurative contro il rischio. Il progetto prevede anche la digitalizzazione del patrimonio culturale del Ministero dei Beni Culturali, che comprende cataloghi di pubblicazioni dall'Unità d'Italia ad oggi. Questa iniziativa è destinata a supportare il sistema pubblico e la ricerca, con il Cineca come consorzio interuniversitario a coordinare queste attività. L'obiettivo è favorire l'integrazione dei dati pubblici e creare semilavorati che possano stimolare l'innovazione.

Il dott. Stamile, a conclusione dell’evento, presenta la strategia di gestione dei dati adottata da Fastweb. L’idea di fondo è di stabilire partnership con editori. Sebbene esistano molti dati pubblici, ci sono due problemi cruciali: il copyright e la qualità dei dati, sintetizzata nel principio "garbage in, garbage out". I dati pubblici possono non essere sufficienti per addestrare un LLM di qualità. Per affrontare questo problema, Fastweb sta creando partnership con editori, come quella già annunciata con Bignami. Queste collaborazioni garantiscono che i dati utilizzati siano di alta qualità, peer-reviewed e creati da esseri umani, riducendo il rischio di dati generati da strumenti di intelligenza artificiale che potrebbero compromettere l'integrità del modello. Questa strategia assicura anche che gli editori ricevano il giusto compenso per i loro dati, indipendentemente dalle attuali leggi sul copyright. La priorità di Fastweb è garantire che i dati utilizzati per l'addestramento siano certificati e affidabili.