Itinerari CmmC - Dal contatto alla conversazione

Gruppo di lavoro sulle tecnologie di AI

L'avvento delle tecnologie di Generative AI ed in particolare dei Large Language Model orientati alle applicazioni di Natural Language Processing è stato accompagnato all'introduzione di nuove tecniche per controllare gli input, gli output e più in generale il comportamento degli LLM, ed in particolare il Prompt Engineering.

Le tecniche di Prompt Engineering consentono di controllare, con relativa facilità e senza necessariamente avere competenze tecniche da Data Scientist, gli input e gli output degli LLM e sono pertanto velocemente diventate una competenza professionale molto richiesta sul mercato.

Il gruppo di lavoro ha affrontato questi due aspetti:

- Le competenze in materia di Prompt Engineering possono diventare una mansione all'interno dell'organizzazione aziendale?
- Padroneggiare queste competenze può portare ad una modificare del rapporto tra aziende committenti e fornitori di tecnologie?

Il Prompt Engineering come mansione

Per rispondere a questo quesito si può partire da due considerazioni.

La prima è che la necessità di allineare le - spesso mutevoli - esigenze di business in materia di Customer Care con la capacità di delivery delle funzioni IT è un'esigenza sempre presente in tutte le aziende, indipendentemente dalla dimensione e dal settore di mercato. E poiché le tecniche di Prompt Engineering non richiedono necessariamente hard skill IT, questo tipo di competenze è di estremo interesse in quanto promette flessibilità, rapidità di implementazione e possibilità di governare con una certa facilità il comportamento delle applicazioni: tutti temi caldi per chi deve gestire un servizio di Customer Care.

Proprio per questo motivo le aziende negli ultimi mesi hanno iniziato a selezionare figure con competenze in materia di Prompt Engineering, ed il termine Prompt Engineer ha rapidamente scalato le classifiche dei siti di selezione del personale.

Questo però ci porta alla seconda considerazione: può chi padroneggia queste tecniche controllare totale in autonomia il comportamento di un LLM e diventarne l'amministratore a 360°?

La risposta è purtroppo no: i fornitori di tecnologia e le società di consulenza e System Integration che hanno iniziato a utilizzare le tecnologie di Generative AI per realizzare applicazioni, progetti pilota e Proof of Concept sanno che nell'esperienza pratica le tecniche di Prompt Engineering sono solo uno degli skill necessari per ottenere risultati convincenti.

Oltre al Prompt Engineering occorre infatti conoscere diverse altre tecniche e strumenti, tra cui:

- Il *Fine Tuning* e *Parameter Efficient Fine Tuning (PEFT)* dei modelli, ovvero l'addestramento dei modelli con documenti e basi di conoscenza specifiche del cliente o di un determinato settore di mercato;
- Le tecniche *LoRA (Low-Rank Adaptation)*, un'altra modalità utilizzata per specializzare e "verticalizzare" il comportamento di un Large Language Model attraverso l'addestramento di modelli più piccoli e compatti che influenzano il comportamento del modello principale;
- Le tecniche RAG (ovvero *Retrieval Augmented Generation*), con cui si sottopongono agli LLM delle "porzioni selezionate" di documenti e contenuti in modo da ottenere risposte il più possibile contestualizzate e minimizzare il fenomeno di allucinazioni;
- I meccanismi di *Reinforcement Learning through Human Feedback (RLHF)*, ovvero la possibilità di correggere le risposte errate di un modello attraverso correzioni applicate da esperti umani coinvolti in attività di verifica - preventiva o ex-post - delle risposte generate automaticamente da un LLM.

Ciascuna di queste tecniche comporta un diverso livello di competenza tecnica: se l'uso dei meccanismi di RLHF è naturalmente demandabile a figure operative esperte di dominio, le tecniche più sofisticate come LoRA o PEFT richiedono un livello di formazione da Data Scientist.

Se, quindi, per realizzare un piccolo prototipo di applicazione può essere sufficiente una buona esperienza con il Prompt Engineering e qualche capacità di scripting, per realizzare applicazioni più complete e affidabili è sostanzialmente necessario uno spettro di competenze molto più ampio, e una capacità di integrazione efficace di tecniche che sono - almeno per il momento e per il prevedibile futuro - di competenza strettamente IT.

In conclusione, la risposta al quesito iniziale è piuttosto evidente: il Prompt Engineering è uno skill sicuramente importante e la sua vicinanza - in termini, ad esempio, di "soft skill" come la capacità di comunicazione e la comprensione delle priorità del business - lo rende di particolare interesse per qualunque azienda che voglia utilizzare concretamente tecnologie di Generative AI e i Large Language Model.

Ma difficilmente questo tipo di competenza potrà ritagliarsi un ruolo organizzativo autonomo all'interno di un'organizzazione: lo scenario più probabile è che i Prompt Engineer diventeranno parte integrante di gruppi più ampi composti da un mix di professionalità che vedrà coinvolti da una parte Subject Matter Expert con conoscenza pratica delle esigenze del business e dall'altra Data Scientist e sviluppatori con capacità e competenze IT molto più specializzate.

Il rapporto con il fornitore di tecnologie

Come richiamato sopra, l'esigenza di avvicinare business e IT non è una novità, e le aziende più grandi hanno da tempo iniziato ad adottare modelli organizzativi basati su competence center che integrano tutte le competenze necessarie per rendere più flessibili e rapide le attività di progetto.

L'avvento della Generative AI può inserirsi in questo solco grazie a tecnologie che sono, almeno in parte, più semplici da utilizzare e personalizzare, in particolare per le attività di prototipazione e sperimentazione - non va dimenticato, tra l'altro, che ad oggi la stragrande maggioranza delle iniziative legate alla Generative AI va inquadrata come Proof of Concept o al più come progetto pilota.

Tuttavia, lo spettro di competenze effettivamente necessarie per realizzare applicazioni complete è tale da richiedere comunque il contributo di specialisti IT le cui competenze, tranne poche eccezioni legate alle grandissime realtà aziendali, sono patrimonio dei partner tecnologici, dei system integrator e delle società di consulenza.

Uno sguardo al futuro

La quantità di investimenti e le dinamiche di rapida evoluzione della Generative AI sono sotto gli occhi sia delle aziende committenti, sia dei fornitori di soluzioni che sono interessati ad inserire nei propri prodotti funzionalità di Intelligenza Artificiale più avanzate, affidabili e versatili.

Gli stessi skill professionali sono in costante evoluzione: se la conoscenza delle tecniche di Prompt Engineering è emersa come un "best buy" in termini di esperienze ricercate dalle aziende da inizio 2023 ad oggi, alcuni osservatori ritengono che queste tecniche potrebbero avere vita breve: i nuovi Large Language Model in via di apparizione tendono a diventare più capaci nell'interpretare le richieste dell'utente, nel mantenere il contesto in conversazioni più articolate, nell'analizzare documenti più complessi e nell'interfacciarsi con le basi dati aziendali.

E' un percorso ancora complesso (nuove versioni di LLM si succedono a distanza di poche settimane le une dalle altre, e l'intero settore è in uno stato estremamente turbolento in termini di trend tecnologici, accordi industriali e scenari competitivi) ma tutto sommato delineato.

Così come il mondo della Data Science si sta più o meno lentamente aprendo a nuove figure come i Citizen Data Scientist - per ora presenti più sulla carta e nelle presentazioni degli analisti di mercato che nella realtà

- e si iniziano ad applicare concretamente soluzioni di AI ispirate dai principi del *low coding* o *no coding*, è facile immaginare che nel giro di poco tempo anche gli strumenti di Generative AI diventeranno progressivamente più semplici da utilizzare e governare.

Se nel breve periodo l'esperienza concreta dei technology vendor rimane un ingrediente chiave per il successo di progetti che vadano oltre il semplice Proof of Concept, con l'emergere di figure organizzative - queste sì dedicate - come il Chief Artificial Intelligent Officer sarà possibile acquisire e coordinare talenti specializzati e selezionare soluzioni via via più complete, sofisticati e semplici da integrare all'interno dei processi di Customer Care delle aziende.