Cosa abbiamo imparato dal nostro primo (vero) progetto di Generative AI

6 Maggio 2024 - 12 min. read

Generative AI

“Me lo ha detto il CEO” non è una strategia.

Il biennio ‘22 - ’23 ha visto un'esplosione dell'interesse nei confronti dell’Artificial Intelligence (AI) in tutte le sue declinazioni.

Strumenti come ChatGPT, Bard, Claude e Bing AI hanno reso questa tecnologia accessibile ad un vastissimo pubblico e ne hanno messo in luce le innumerevoli potenzialità, in particolar modo riguardo alla Generative AI.

Moltissime aziende si sono già spinte in applicazioni commerciali: dalla personalizzazione di contenuti aziendali e non, alla generazione di materiali promozionali estremamente mirati e personalizzati, dalla produzione di documentazione tecnica, alla traduzione real-time di testi, fino addirittura al miglioramento di sistemi Help Desk e assistenza clienti.

L'entusiasmo suscitato, tuttavia, sembra aver innescato una vera e propria “corsa alla GenAI” portando a un'inversione del processo decisionale: spinti dalla paura di perdere l’opportunità (FOMO), si tende a cercare di adattare la Generative AI ai problemi aziendali più disparati invertendo causa - effetto e rischiando di cadere in investimenti ingenti fini a loro stessi.Nonostante le straordinarie possibilità offerte, dunque, è importante sottolineare che la Generative AI non è sempre la soluzione a tutte le sfide di business: occorre valutare secondo molteplici punti di vista se questa possa realmente offrire un vantaggio competitivo tangibile e giustificare così effort e investimento, ad oggi ancora molto alti.

Ci siamo trovati a dover fare queste e altre importanti considerazioni in occasione della realizzazione del nostro primo (vero) progetto di Generative AI.

Ve ne parliamo in questo articolo partendo dal perché abbiamo ritenuto che la Generative AI fosse lo strumento migliore per massimizzare il risultato, fino a ripercorrere tutto ciò che di prezioso abbiamo imparato durante la realizzazione del progetto.

Next-generation planner

Per un cliente in ambito food, ci siamo occupati della realizzazione di un sistema di pianificazione e raccomandazione. Dato un input contenente criteri personalizzati, la soluzione è in grado di comporre in autonomia un menu su base giornaliera e/o settimanale, pianificando i pasti scegliendo i piatti in base a preferenze, limiti di crediti disponibili sulla piattaforma e necessità alimentari date dell'utente.

Le sfide principali erano:

- Eliminare il rischio per gli utenti di rimanere "a bocca asciutta", come si dice, dimenticando di effettuare l'ordine del pranzo in tempo utile, in particolare, il venerdì per il lunedì;
- Ridurre il tasso di abbandono del carrello in fase di ordinazione a causa del tempo necessario per consultare tutto il menu e selezionare i piatti.

L'obiettivo era quindi duplice: da un lato, **aumentare retention e fidelizzazione** del cliente, riducendo al contempo la friction; dall'altro, **minimizzare il rischio di mancato guadagno** per l'azienda.

Da sempre esistono algoritmi noti di Natural Language Processing (NLP), ovvero algoritmi di AI in grado di elaborare il linguaggio naturale e di interpretarlo per fornire output coerenti con criteri decisionali specifici. Mettendo insieme algoritmi di NLP, Reccomandation e tecniche di best fit, dunque, avremmo potuto raggiungere il medesimo obiettivo ottenuto con l'utilizzo della Generative AI.

La Gen AI, tuttavia, ci avvicinava decisamente di più agli obiettivi di partenza per una serie di motivi:

- **Maggior personalizzazione**
  I motori tradizionali si basano su labelling dei contenuti e algoritmi che utilizzano il filtraggio collaborativo, i quali potrebbero non catturare appieno la complessità delle preferenze individuali degli utenti.

  Quelli basati sull'AI generativa, d'altro canto, possono generare raccomandazioni che tengono conto di una più ampia gamma di fattori, inclusi quelli espressi in linguaggio naturale, offrendo una personalizzazione più profonda

Usando la GenAI, è stato possibile utilizzare i dati a disposizione dal cliente, senza dover effettuare labeling manuale. Tutte le informazioni presenti nella descrizione della pietanza, ed espreesse in linguaggio naturale, vengono automaticamente utilizzate dal modello per categorizzarla e decidere quando e se proporla all'utente in base al prompt fornito.

- **Comprensione del Linguaggio Naturale**
  I motori tradizionali hanno numerose limitazioni nella comprensione delle richieste fatte in linguaggio naturale.

  La Gen AI, al contrario, permette di elaborare e interpretare il linguaggio colloquiale, e di rispondere accuratamente a query complesse "con la stessa moneta".

- **Generazione di Contenuti**
  I motori tradizionali si limitano a suggerire contenuti o prodotti esistenti sulla base di metriche e modelli conosciuti. I motori che sull'IA generativa, invece, hanno il potenziale per creare nuovi contenuti, come descrizioni di prodotti personalizzate, migliorando ulteriormente l'esperienza dell'utente.

- **Indipendenza da Dati Storici**
  I motori di raccomandazione tradizionali, infine, possono dipendere strettamente dai dati storici e dalle interazioni passate. In mancanza di questi, la loro capacità di scoprire nuove tendenze o raccomandare nuovi prodotti può essere limitata. I sistemi basati su Gen AI possono, in teoria, superare tali limitazioni generando raccomandazioni innovative, anche in assenza di grandi volumi di dati storici.

Vediamo ora quali sono stati i principali takeaway che abbiamo fatto nostri durante l'implementazione della soluzione e che saranno fondamentali per il successo di tutti i progetti futuri che prevederanno l'utilizzo della Generative AI.

Takeaway e lesson learned per il successo di un progetto di Generative AI.

Qualità del risultato e tempi di go-to-market sono rapportati allo sforzo

Utilizzare i LLM facilita enormemente la vita in tutti quei contesti dove un certo grado di imprecisione non è una criticità e dove fornire un nuovo servizio all'utente è già di per sè un vantaggio competitivo.

Questo perchè non bisogna investire tempo e risorse nell'individuazione di un modello specifico per risolvere un determinato problema, ma soprattutto non bisogna investire per trainarlo. L'addestramento dei modelli resta infatti una delle pratiche più costose in assoluto in ambito AI.

Grazie a un training già sufficientemente ampio da non richiedere aggiustamenti avanzati, infatti, i LLM ci permettono di arrivare ad un risultato soddisfacente in tempi relativamente brevi rispetto all'utilizzo di algoritmi classici.

Il rovescio della medaglia sta nel fatto che, per quanto buono, il risultato che andremo ad ottenere non sarà il miglior risultato ottenibile in assoluto. In un'ottica di miglioramento incrementale, tuttavia, si possono ottimizzare i tempi di go-to-market servendosi di LLM per poi agire a posteriori con con algoritmi noti con cui effettuare best match.

Non ci sono limiti alla perfezione!

I benefici tangibili per l'utente finale e per l'azienda

Quando si parla di user experience, spesso i benefici per l'utente finale sono intangibili. In questo caso il planner ha un impatto concreto su due dimensioni: tempo e denaro.

Il tempo che l'utente passa a pianificare i pasti per la settimana successiva varia tra i 5 e 15 minuti (a seconda delle esigenze, della scelta dei piatti e del confronto di specifici valori nutizionali).

Già con il primo test, il modello impiega 38 secondi a presentare all'utente una pianificazione. Come ci siamo detti, ci sono diversi aspetti perfettibili sia per ridurre i tempi, sia per migliorare l'output. Tuttavia, in meno di 2 minuti l'utente è in grado di aggiustare la pianificazione proposta e ritrovarsi con un menù completo.

Questo si traduce in un risparmio di tempo pari al 60%-86%.

Si supera inoltre in automatico il problema dell'utente che si dimentica di ordinare da una settimana con l'altra, trovandosi così costretto a dover uscire o ordinare tramite servizi di consegna a domicilio. Se il disagio di arrivare un giorno in ufficio e non trovare il proprio pasto non è quantificabile, si può misurare che mediamente pranzare fuori ha un costo superiore del 30%-45% (a parità di cibo consumato). Indirettamente quindi non c'è solo un fattore di user experience migliorata, ma anche di risparmio dei costi.

Dal punto di vista dell'azienda il planner crea un rapporto diretto tra soddisfazione del cliente e guadagno. Ogni volta che un utente si dimentica di ordinare si crea uno scenario lose-lose: l'utente vive un disagio e l'azienda ha un mancato guadagno.

Con il planner l'azienda non solo si assicura che ogni giorno ci sarà un guadagno, ma anche che questo venga massimizzato, dal momento che il modello cercherà di utilizzare il massimo numero di crediti possibili nel pianificare i pasti.

Bello sì. Ma sostenibile?

I consumi della Gen AI relativamente ad aspetti economici, energetici, di potenza di calcolo e CPU sono notevoli.

Va tenuto a mente che stiamo utilizzando uno strumento estremamente potente per risolvere problemi molto specifici e circoscritti, spesso potenzialmente risolvibili con l'impiego di molte meno risorse... se pur con uno sforzo implementativo maggiore.

Per utilizzare una metafora, "sparando alla formica col bazooka", si ottiene sicuramente il risultato di eliminarla, ma con tutti gli effetti collaterali dell'aver utilizzato un'arma così potente per un obiettivo così circoscritto.

Questo solleva un tema importante, la sostenibilità, su cui la Generative AI ha un impatto significativo da diversi punti di vista:

- **Consumo energetico e utilizzo delle risorse**
  Addestramento dei modelli e inferenza richiedono una quantità enorme di potenza di calcolo, contribuendo in modo consistente al consumo di risorse non rinnovabili e all'emissione di gas serra.

  Inoltre, gli sviluppi rapidi nella Generative AI portano a una rapida obsolescenza dei dispositivi e delle infrastrutture necessarie per eseguire questi potenti modelli. Questo provoca un aumento di rifiuti elettronici notoriamente difficili da riciclare, con impatto sul problema del carbon footprint.

  Si tratta di una questione che riguarda principalmente i grandi provider su cui noi possiamo avere poco impatto.

  Dobbiamo però ricordarci che tanto noi come system integrator, quanto gli utenti come fruitori del servizio, siamo parte attiva di questa value chain.
- **Sostenibilità economica**
  Sul fronte della sostenibilità economica, sperimentare con la Generative AI comporta costi significativi. Un'azienda che si avvicina a questa tecnologia deve affrontare costi elevati, talvolta proibitivi, fin dalle prima fasi. Ad esempio, nel caso del nostro planner, già in fase di test ciascuna chiamata alla Gen AI aveva dei costi tutt'altro che irrisori e per un'azienda nei suoi primi anni di attività il costo del servizio è un elemento che va considerato.

  Per arrivare alla sostenibilità economica di un progetto di Gen AI-based, bisogna prima superare i costi delle fasi di test e arrivare in produzione con una soluzione in grado di

generare un valore “vendibile” per l’utente finale, tale da giustificare gli investimenti.

Per quanto importante, ad oggi resta difficile fare in modo che l’utente finale abbia contezza del costo aziendale che sta dietro alla generazione del grande valore di cui può usufruire. La ragione risiede probabilmente nella diffusione di strumenti "AI Powered" che offrono benefici simili gratuitamente.

In conclusione, allocare un budget sostenibile e pianificare a lungo termine è essenziale per garantire la sostenibilità economica dell'utilizzo della Generative AI, insieme alla sensibilizzazione degli utenti riguardo al pagamento di piani premium.

Gioie e dolori del calcolo dei Costi

Utilizzando la Generative AI, il contesto del workload può giocare a nostro favore o a nostro sfavore (o, come vedremo, addirittura entrambe le cose), rendendo la previsione dei costi un’incognita.

Da un punto di vista puramente informatico e di programmazione, l'utilizzo dei LLM è relativamente semplice, poiché l'elaborazione del dato che genera l'output avviene in gran parte in modo automatico, come in una black box.

Non essendo richieste elaborazioni avanzate (almeno in una prima fase), il grande sforzo, nonché l’ago della bilancia sul controllo dei costi, si sposta sulla system integration e sulla comprensione logica del contesto in cui il LLM opera.

Nell’utilizzo dei LLM, il costo dipende dal numero di token (unità di misura assimilabili alle sillabe testuali) in ingresso nel prompt e in uscita dal risultato generato. Per evitare spese eccessive e fuori controllo, è utile limitare l'incertezza nell'uso del prompt, quando possibile.

Torniamo al nostro planner: come menzionato in precedenza, ciascun input testuale, ha un costo per nulla trascurabile per l’azienda. Il costo elevato è dovuto al fatto che, affinché la soluzione funzioni, ogni prompt deve includere non solo le richieste dell'utente, ma anche altre informazioni "nascoste", come l'intero menu (meta-prompt). Poiché la lunghezza e la variabilità del prompt sono difficili da ottimizzare, il costo non può essere previsto con certezza.

Siamo comunque riusciti a evitare l’esplosione incontrollata dei costi con un paio di accortezze. Una dettata da noi, ovvero un limite forzato di 200 caratteri (quindi un numero di token in ingresso più o meno stabile) per esprimere i propri criteri di scelta, e una data dalla piattaforma stessa (quindi dal contesto): attraverso l’utilizzo del meta-prompt, ovvero della parte di informazioni non visibili all’utente, siamo stati in grado di passare il numero

totale di crediti disponibili per pasto (80 nel nostro caso, corrispondenti a una spesa di 8€) da cui viene man mano sottratto il costo di ciascun piatto scelto dal LLM. Nonostante notoriamente i LLM non eccellano nei calcoli matematici, il risultato ottenuto è stato buono, in quanto il sistema stesso si occupa di non mostrare eventuali piatti che eccederebbero il residuo dei crediti.

Sebbene in modo approssimativo, siamo riusciti a ottenere un risultato per l'utente all'interno dei limiti dei punti disponibili e a limitare la spesa grazie a una gestione per l'azienda prevedibile del numero di token in output.

Il Concetto di linguaggio

Uno degli aspetti che abbiamo sottovalutato nell'utilizzo dei LLM è senz'altro il concetto di "linguaggio". È importante sottolineare che i LLM non si limitano ad accettare e generare testo in linguaggio naturale *umano*. In realtà, possono accettare input in qualsiasi linguaggio formale dotato di una grammatica e produrre un output nella stessa forma.

I linguaggi di programmazione sono ovviamente compresi in questa definizione più ampia: il JSON, ad esempio, essendo un formato di interscambio, è a tutti gli effetti un linguaggio strutturato e può quindi essere utilizzato sia come input in un prompt, sia come output richiesto. È infatti possibile chiedere al LLM di produrre il risultato in un linguaggio diverso da quello naturale, ad esempio, di rappresentare il menù proposto usando un JSON. Ci si può spingere - ed è bene - ad indicare l'esatta forma del JSON che ci si aspetta in output. In questo modo, il modello fornirà un output direttamente utilizzabile dal sistema del cliente per portare il valore all'interno delle applicazioni e restituire agli utilizzatori finali una risposta integrata nell'esperienza utente attesa. Nel caso specifico, il risultato viene utilizzato dal sistema per comporre il menù nella medesima interfaccia che userebbe l'utente.

In generale la possibilità di usare linguaggi vicini alla macchina rende possibile sfruttare gli LLM per avvicinare le richieste utente a sistemi tradizionali, unificando e migliorando la UX.

Per concludere

Le potenzialità della Generative AI sono oggi sotto gli occhi di tutti. A beneficiarne maggiormente sono sicuramente gli utenti finali grazie alle innumerevoli opportunità di toccare con mano i risultati di strumenti ormai di uso comune.

Tuttavia, per coloro che, come noi, operano nel settore, l'hype che circonda questa tecnologia solleva domande e incertezze. Se da un lato la possibilità di evitare le

complessità tipiche di un più tradizionale utilizzo dell'AI è un grosso vantaggio, dall'altro ci si chiede se il gioco possa veramente valere la candela.

A fronte di interessi e investimenti sempre crescenti nel settore, le aziende, anche le più grandi, stanno ancora cercando di capire come monetizzare appieno le tecnologie basate su AI.

In conclusione, nonostante la popolarità tra gli utenti, sembra che il percorso verso una completa integrazione e un utilizzo sostenibile sia ancora piuttosto lungo e incerto. È essenziale mantenere un approccio critico e bilanciato, valutando attentamente i benefici e i rischi associati a queste tecnologie emergenti.Voi cosa ne pensate?

Avete già realizzato soluzione Gen AI-based?

Fatecelo sapere nei commenti!

About Proud2beCloud

Proud2beCloud è il blog di beSharp, APN Premier Consulting Partner italiano esperto nella progettazione, implementazione e gestione di infrastrutture Cloud complesse e servizi AWS avanzati. Prima di essere scrittori, siamo Solutions Architect che, dal 2007, lavorano quotidianamente con i servizi AWS. Siamo innovatori alla costante ricerca della soluzione più all'avanguardia per noi e per i nostri clienti. Su Proud2beCloud condividiamo regolarmente i nostri migliori spunti con chi come noi, per lavoro o per passione, lavora con il Cloud di AWS. Partecipa alla discussione!

Alessio Gandini

Cloud-native Development Line Manager @ beSharp, DevOps Engineer e AWS expert.Computer geek da quando avevo 6 anni, appassionato di informatica ed elettronica a tutto tondo. Ultimamente sto esplorando l'esperienza utente vocale e il mondo dell'IoT.Appassionato di cinema e grande consumatore di serie TV, videogiocatore della domenica.
