



UNIVERSITÀ DEGLI STUDI
DI NAPOLI FEDERICO II

AI System Engineering Project
QuitSmokingSOS

A chatbot to help the smoker in need.

Descrizione progetto

Introduzione

Tale progetto corrisponde a un chatbot LLM in cui è stato utilizzato un modello RAG. Tale chatbot, chiamato QuitSmokingSOS, ha molteplici compiti che vertono nell'assistenza dello smettere di fumare e che saranno successivamente spiegati. Il vantaggio principale di tale sistema è la possibilità di utilizzo sia da parte di specialisti (operatori sanitari o psicologi) sia da parte di utenti senza alcuna conoscenza.

E' possibile fornire in input sia prompt con dati biometrici sia prompt meno schematici, in cui sono espresse anche le proprie difficoltà e i propri miglioramenti da un punto di vista personale.

Prerequisiti

Per permettere l'esecuzione di tale programma, abbiamo utilizzato le seguenti estensioni:

1. Ollama
2. Streamlit
3. Python (da versione 3.8 a 3.11)
4. Tesseract
5. Poppler

Requirement Analysis

- **Stakeholder Identification**

I principali stakeholder per il chatbot comprendono gli utenti finali (persone che cercano di smettere di fumare), gli operatori sanitari (medici, psicologi e specialisti delle dipendenze), i supervisor accademici che valutano il merito del progetto e, potenzialmente, gli amministratori della sanità se il sistema si integra con le piattaforme esistenti.

Ogni categoria di stakeholder porta con sé priorità distinte che devono essere colte attraverso tecniche opportune:

- **Utenti finali**

Gli utenti finali sono i principali beneficiari del sistema, includendo chi sta cercando attivamente di smettere di fumare o sta considerando questa possibilità. Questo gruppo è estremamente variegato per quanto riguarda età, storia del fumo e motivazioni per smettere. Dall'analisi dei requisiti emerge che gli utenti danno la massima importanza a un supporto accessibile e privo di giudizi, disponibile nei momenti critici di voglia di fumare o debolezza. Apprezzano, inoltre, un'assistenza personalizzata che tenga conto dei loro trigger specifici, dei loro schemi di fumo e dei tentativi passati. Molti preferiscono interazioni conversazionali che combinino strategie pratiche di coping con un incoraggiamento motivazionale, adottando un tono empatico e di sostegno piuttosto che autoritario. Gli utenti hanno anche evidenziato l'importanza di garantire la massima riservatezza, considerata la natura delicata delle discussioni su dipendenze e salute.

- **Operatori sanitari**

I professionisti della salute rappresentano un gruppo fondamentale di stakeholder, poiché la loro esperienza clinica assicura che il sistema segua approcci di cessazione basati su evidenze scientifiche. Interviste con specialisti in dipendenze, pneumologi e terapisti comportamentali hanno messo in luce la necessità di una rigorosa accuratezza scientifica e di minimizzare i rischi. Questi esperti hanno sottolineato l'importanza di definire chiaramente i limiti del sistema, inserendo adeguati disclaimer sul ruolo complementare del chatbot rispetto alla cura tradizionale, e di curare attentamente i contenuti per non rinforzare comportamenti negativi. Inoltre, hanno suggerito l'implementazione di meccanismi di screening per identificare tempestivamente utenti che necessitano di un intervento professionale urgente, come quelli con sintomi di astinenza particolarmente intensi o complicazioni relative alla salute mentale. Hanno, infine, evidenziato l'importanza di citare correttamente le informazioni mediche per rafforzare credibilità e fiducia.

- **Psicologi e specialisti del comportamento**

Questa sottocategoria di stakeholder si interessa in maniera specifica al design della conversazione e alle tecniche motivazionali. Le loro esigenze puntano all'integrazione di framework psicologici validati – come il motivational interviewing, le tecniche cognitivo-comportamentali e il modello delle fasi del cambiamento. Essi sottolineano l'importanza di flussi di conversazione adattabili che riconoscano il progresso degli utenti attraverso le diverse fasi della cessazione, dalla fase di contemplazione a quella di mantenimento. Propongono inoltre l'inserimento di strategie di prevenzione delle ricadute e di iniziative volte a promuovere

l'autoefficacia, attraverso la definizione di obiettivi misurabili e il riconoscimento dei successi.

- **Supervisor accademici**

Infine, i supervisor accademici rappresentano l'ultima categoria principale, concentrandosi sulla rigerosità metodologica, sull'innovazione tecnica e sulla completezza della documentazione. Essi richiedono una rendicontazione trasparente dei limiti del sistema, metodologie di valutazione sistematica e un solido ancoraggio alla letteratura scientifica. Per questi stakeholder è fondamentale che il contributo tecnico venga espresso in modo chiaro, evidenziando il valore terapeutico del sistema e, in particolare, le applicazioni innovative della tecnologia RAG nel campo della salute comportamentale.

Requisiti non funzionali

I requisiti non funzionali definiscono gli attributi di qualità fondamentali per garantire l'efficacia, l'affidabilità e l'implementazione etica del chatbot per la cessazione dal fumo. Questi requisiti affrontano aspetti critici che vanno oltre le semplici funzionalità del sistema, stabilendo parametri di riferimento per le prestazioni e per l'esperienza utente.

- **Privacy e sicurezza dei dati**

Vista la natura sensibile delle conversazioni inerenti alla salute, i requisiti di privacy e sicurezza assumono la massima priorità. Il sistema deve implementare una crittografia end-to-end per tutte le interazioni, accompagnata da politiche chiare di conservazione dei dati che limitino l'archiviazione delle informazioni personali relative alla salute. I meccanismi di autenticazione dovranno bilanciare sicurezza e accessibilità, evitando soluzioni troppo restrittive che possano scoraggiare l'uso costante del servizio. L'architettura dovrà conformarsi alle normative vigenti in materia di protezione dei dati sanitari in modo da definire pattern che possano essere estesi a eventuali implementazioni cliniche. In questo contesto è fondamentale garantire trasparenza nell'utilizzo dei dati, gestire adeguatamente il consenso degli utenti e prevedere meccanismi che permettano agli utenti di esportare o cancellare la cronologia delle conversazioni.

- **Accuratezza clinica**

I requisiti di accuratezza clinica assicurano che il sistema offra indicazioni basate su evidenze, rispecchiando le migliori pratiche attuali per la cessazione dal fumo. L'implementazione del RAG deve privilegiare il recupero di informazioni da fonti peer-reviewed e linee guida cliniche riconosciute, associando metadati che indichino la data di aggiornamento e la credibilità delle fonti. La generazione delle risposte dovrà includere meccanismi di attribuzione che colleghino le raccomandazioni ai materiali originali, migliorando la trasparenza e contribuendo all'educazione dell'utente. Inoltre, dovranno essere definiti dei limiti di confidenza per determinare quando il sistema possa fornire una raccomandazione precisa oppure ammettere incertezza, soprattutto nel rispondere a domande mediche che superano le linee guida generali per la cessazione.

- **Usabilità**

I requisiti di usabilità sono studiati per affrontare le sfide specifiche del supporto agli

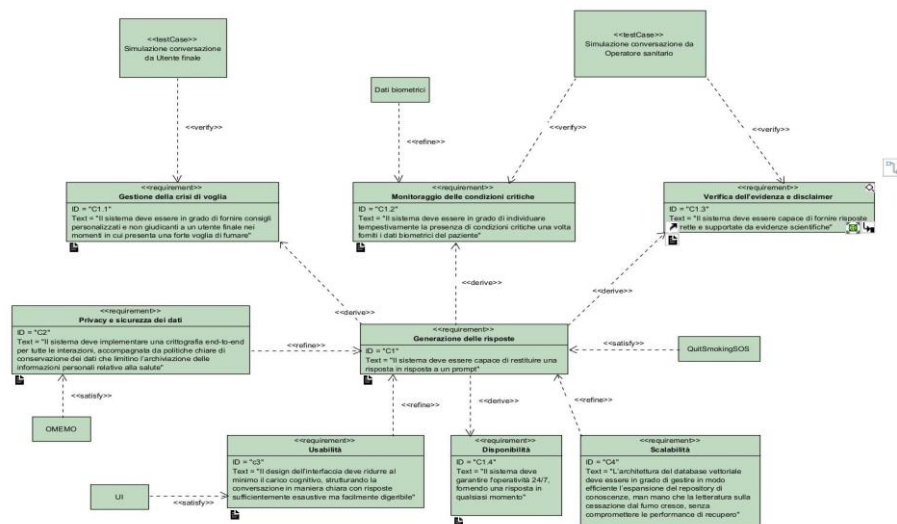
individui nel percorso di cessazione, considerando che gli utenti potrebbero subire effetti cognitivi dovuti al ritiro della nicotina. Il design dell'interfaccia deve ridurre al minimo il carico cognitivo, strutturando la conversazione in maniera chiara e differenziando nettamente tra la semplice fornitura di informazioni e l'offerta di supporto interattivo. La lunghezza delle risposte dovrà essere equilibrata: sufficientemente esaustiva ma facilmente digeribile, per evitare di sovraccaricare l'utente in momenti di vulnerabilità. Il sistema dovrà anche essere capace di adattarsi a differenti livelli di competenza linguistica, pur mantenendo una rigorosa accuratezza scientifica, sfruttando tecniche di generazione del linguaggio naturale che modulino la complessità terminologica in base alle interazioni con l'utente.

- **Disponibilità e affidabilità**

I requisiti di disponibilità e affidabilità riconoscono che il supporto per la cessazione può essere necessario in orari imprevedibili, soprattutto durante episodi di forte craving. Per questo motivo, l'architettura deve garantire un'operatività 24/7 con tempi di inattività minimi, implementando soluzioni di degrado elegante durante le fasi di manutenzione. La latenza delle risposte dovrà essere mantenuta al di sotto dei tre secondi, anche in presenza di richieste complesse, per preservare un flusso conversazionale naturale; si dovranno inoltre adottare strategie di ottimizzazione nei momenti di elevato carico. Considerare funzionalità offline per le operazioni essenziali potrebbe essere utile per ridurre la dipendenza da una connessione continua durante i momenti critici di supporto.

- **Scalabilità**

I requisiti di scalabilità coprono sia l'aspetto tecnico che quello relativo al contenuto del sistema. L'architettura del database vettoriale deve essere in grado di gestire in modo efficiente l'espansione del repository di conoscenze, man mano che la letteratura sulla cessazione dal fumo cresce, senza compromettere le performance di recupero. La gestione della cronologia delle conversazioni deve trovare un equilibrio tra una completa consapevolezza del contesto e l'efficienza computazionale, magari attraverso l'adozione di tecniche di sintesi intelligenti per interazioni prolungate. Infine, la scalabilità in termini di base utenti richiede strategie di bilanciamento del carico e una corretta allocazione delle risorse, anche in ambito accademico, per predisporre il sistema a una potenziale espansione su larga scala.



Requirement Diagram

User stories

Di seguito sono elencati alcuni esempi di user stories che simulano alcuni utilizzi che prevede il chatbot.

- **Utenti finali**

1. Come utente finale che vuole smettere di fumare, voglio ricevere consigli personalizzati e non giudicanti nei momenti di forte voglia, così da affrontare le crisi con maggiore serenità.
2. Come utente finale, desidero interagire con un assistente empatico che rispetti la mia privacy, per sentirmi supportato nel mio percorso di cessazione senza timore di stigmatizzazione.
3. Come utente finale, voglio avere accesso ad articoli e informazioni sui benefici per la salute dello smettere di fumare per mantenermi motivato.
4. Come utente finale, voglio che il chatbot mi raccomandi alternative salutari o distrazioni per i momenti in cui la voglia di fumare si fa sentire.
5. Come utente finale, voglio avere accesso a testimonianze di persone che hanno smesso di fumare con successo per trarre ispirazione.
6. Come utente finale, voglio informazioni sulle opzioni di terapia sostitutiva della nicotina e su altri ausili che possano aiutarmi a gestire efficacemente i sintomi dell'astinenza.

- **Operatori sanitari**

1. Come operatore sanitario, voglio che il sistema fornisca informazioni basate su evidenze scientifiche e includa chiari disclaimer, per integrare in sicurezza il supporto offerto dall'assistente AI alla cura tradizionale.
2. Come operatore sanitario, desidero che il sistema implementi conoscenze di dati biometrici per individuare tempestivamente utenti in condizioni critiche, in modo da poter intervenire prontamente se necessario.

- **Psicologi e specialisti del comportamento**

1. Come specialisti del comportamento, desidero che il sistema proponga obiettivi misurabili e tecniche di coping personalizzate, così da rafforzare l'autoefficacia degli utenti nel percorso di smettere di fumare.
2. Come psicologo, voglio che il chatbot utilizzi framework validati (es. motivational interviewing) e adatti il flusso conversazionale in base al progresso dell'utente, per promuovere strategie efficaci e prevenire ricadute.

Casi d'uso

- **Utenti finali**

1. **Gestione della Crisi di Voglia**

Marco sta attraversando un momento di forte voglia di fumare mentre è al lavoro. Accede al chatbot e, grazie al riconoscimento del tono emotivo, riceve immediatamente un messaggio di supporto non giudicante che include consigli personalizzati (es. esercizi di respirazione e una breve meditazione). Il messaggio lo aiuta a gestire la crisi e a proseguire il percorso di cessazione con maggiore serenità.

2. **Supporto Empatico e Riservatezza**

Giulia, che ha timore di essere giudicata, decide di utilizzare l'assistente AI. Durante una conversazione, il chatbot le fornisce un supporto empatico, rispondendo in modo personalizzato e garantendo il massimo rispetto della privacy. Questo fa sentire Giulia al sicuro e compresa, rafforzando la sua motivazione a smettere di fumare.

3. **Accesso a Contenuti Motivazionali**

Luigi ha bisogno di rinfrescare le sue motivazioni. In pochi secondi, il chatbot gli mostra una selezione di articoli e infografiche che illustrano i benefici per la salute dello smettere di fumare, insieme a testimonianze di successo. Questi contenuti lo ispirano e lo spingono a perseverare nel suo percorso.

4. **Raccomandazione di Alternative Salutari**

Elena si trova in un momento in cui la voglia di fumare diventa preponderante. Il chatbot, riconoscendo il trigger, le propone immediatamente alternative salutari, come fare una breve passeggiata o ascoltare una sessione di mindfulness, aiutandola a distrarsi e a gestire efficacemente l'astinenza.

5. **Ispirazione da Testimonianze**

Riccardo, in un momento di dubbio, chiede al chatbot storie di successo. Il sistema gli presenta brevi testimonianze di ex fumatori che hanno superato la dipendenza, offrendo a Riccardo un modello positivo e aumentando la sua fiducia nel percorso.

6. **Informazioni su Terapie Sostitutive**

Sara, interessata a opzioni alternative, utilizza il chatbot per conoscere le diverse terapie sostitutive della nicotina (come cerotti, gomme, inalatori). Il sistema le fornisce una panoramica chiara dei benefici e dei possibili effetti collaterali, aiutandola a valutare se queste soluzioni possano integrarsi nel suo percorso di cessazione.

- **Operatori sanitari**

1. **Verifica dell'Evidenza e Disclaimer**

Il Dr. Rossi, operatore sanitario, utilizza il sistema per assicurarsi che i consigli forniti agli utenti siano supportati da evidenze scientifiche. Analizza la documentazione del

chatbot, che include chiari disclaimer sul ruolo complementare del supporto AI, in modo da integrare in sicurezza il sistema nella cura tradizionale.

2. Monitoraggio di Condizioni Critiche

L'infermiera Maria, incaricata del monitoraggio, osserva che il sistema integra dati biometrici (ad esempio, segnalazioni di stress elevato) per individuare utenti in situazioni critiche. Quando un utente mostra segnali di astinenza particolarmente intensi, il sistema invia un alert, permettendo a Maria di intervenire tempestivamente.

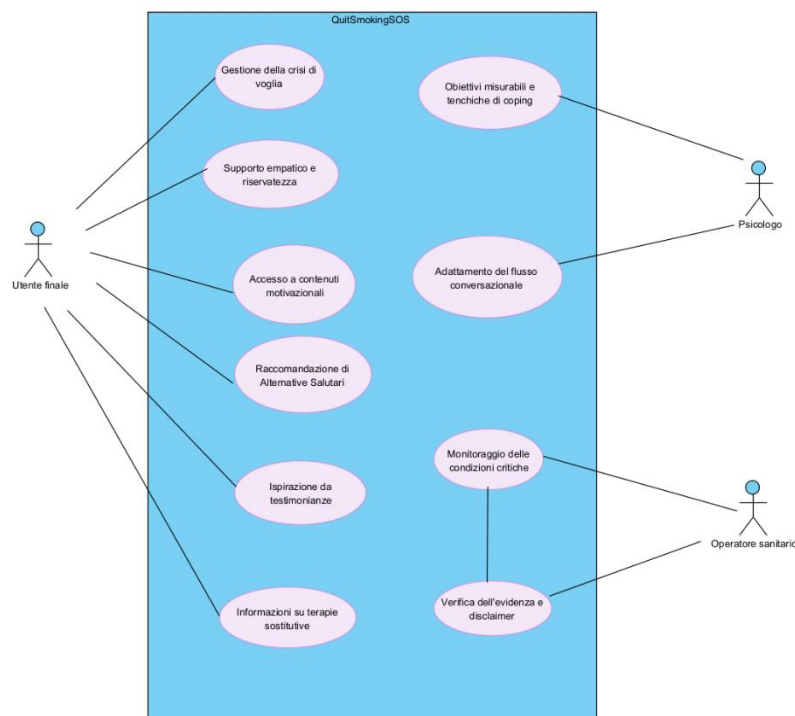
• Psicologi e Specialisti del Comportamento

1. Obiettivi Misurabili e Tecniche di Coping

Il Dott. Bianchi, psicologo specializzato in dipendenze, collabora con il team per impostare obiettivi misurabili nel chatbot. Durante una conversazione, il sistema propone esercizi di coping personalizzati (ad esempio, la registrazione del numero di giorni senza fumare) e traccia il progresso dell'utente, rafforzando così la sua autoefficacia.

2. Adattamento del Flusso Conversazionale

La dottoressa Verdi, esperta in terapia comportamentale, osserva che il chatbot utilizza framework validati come il motivational interviewing. In base alle risposte dell'utente, il flusso della conversazione si adatta dinamicamente: il sistema modifica il tono e propone strategie specifiche per prevenire ricadute, offrendo un supporto sempre in linea con il progresso dell'utente.



Scenari

Caso d'uso	Gestione della crisi di voglia
Attore primario	Utente finale
Attore secondario	-
Descrizione	L'utente finale vuole ricevere consigli personalizzati (es. esercizi di respirazione e una breve meditazione) e non giudicanti nei momenti in cui ha una forte voglia di fumare
Pre-condizioni	Essere connesso alla chatbot
Sequenza di eventi principale	<ol style="list-style-type: none"> 1. Selezionare il pdf da cui ottenere le informazioni <ol style="list-style-type: none"> 1.1 Selezionare il pdf di sample 1.2 Selezionare il pdf da un upload <ol style="list-style-type: none"> 1.2.1 Selezionare il pdf dallo storage locale della macchina 2. Scrivere e inviare un prompt 3. Il sistema ragiona 4. Il sistema fornisce la risposta
Post-condizioni	Aggiunta della domanda e della risposta alla cronologia dei messaggi fino allo scollegamento dall'applicazione
Casi d'uso correlati	Nessuno
Sequenza di eventi alternativi	Se si sceglie di selezionare il pdf da un upload e decidiamo di non inserirlo, possiamo selezionare il pdf di sample, tornando alla UI principale

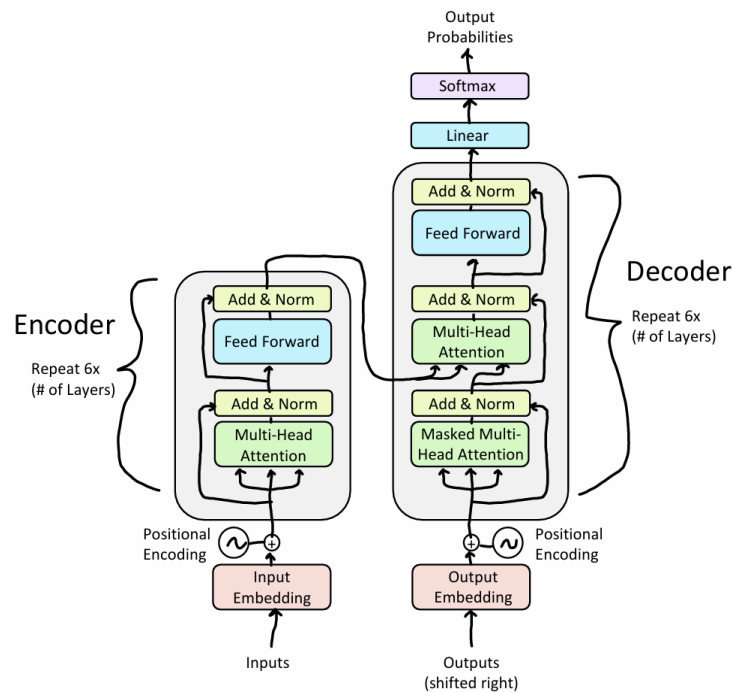
Caso d'uso	Monitoraggio delle condizioni critiche
Attore primario	Operatore sanitario
Attore secondario	-
Descrizione	L'operatore sanitario vuole poter inserire i dati biometrici di un paziente per poter individuare tempestivamente la presenza di condizioni critiche
Pre-condizioni	Essere connesso alla chatbot Avere i dati biometrici dell'utente, forniti dall'utente stesso o ottenuti dopo un'analisi svolta dall'operatore
Sequenza di eventi principale	<ol style="list-style-type: none"> 1. Selezionare il pdf da cui ottenere le informazioni <ol style="list-style-type: none"> 1.1 Selezionare il pdf di sample 1.2 Selezionare il pdf da un upload <ol style="list-style-type: none"> 1.2.1 Selezionare il pdf dallo storage locale della macchina 2. Scrivere e inviare un prompt con i dati biometrici 3. Il sistema ragiona 4. Il sistema fornisce la risposta
Post-condizioni	Aggiunta della domanda e della risposta alla cronologia dei messaggi fino allo scollegamento dall'applicazione
Casi d'uso correlati	Verifica dell'evidenza e disclaimer
Sequenza di eventi alternativi	Se si sceglie di selezionare il pdf da un upload e decidiamo di non inserirlo, possiamo selezionare il pdf di sample, tornando alla UI principale

Caso d'uso		Verifica dell'evidenza e disclaimer
Attore primario		Operatore sanitario
Attore secondario		-
Descrizione		L'operatore sanitario vuole assicurarsi che i servizi forniti all'utente siano supportati da evidenze scientifiche
Pre-condizioni		Essere connesso alla chatbot Avere i consigli forniti all'utente dal chatbot
Sequenza di eventi principale		<ol style="list-style-type: none"> 1. Selezionare il pdf da cui ottenere le informazioni <ol style="list-style-type: none"> 1.1 Selezionare il pdf di sample 1.2 Selezionare il pdf da un upload <ol style="list-style-type: none"> 1.2.1 Selezionare il pdf dallo storage locale della macchina 2. Scrivere e inviare un prompt fornito dall'utente (sensazioni, dati biometrici, ecc) 3. Il sistema ragiona 4. Il sistema fornisce la risposta 5. Effettuare una ricerca da fonti esterne per avere una certezza assoluta <ol style="list-style-type: none"> 5.1 Previsione corretta 5.2 Previsione errata <ol style="list-style-type: none"> 5.2.1 Ripetere gli step dal passo 2 per effettuare nuove analisi 5.2.2 Ripetere gli step dal passo 1 per effettuare nuove analisi (nel caso in cui si volesse cambiare la fonte)
Post-condizioni		Aggiunta della domanda e della risposta alla cronologia dei messaggi fino allo scollegamento dall'applicazione
Casi d'uso correlati		Monitoraggio di condizioni critiche
Sequenza di eventi alternativi		<p>Se si sceglie di selezionare il pdf da un upload e decidiamo di non inserirlo, possiamo selezionare il pdf di sample, tornando alla UI principale</p> <p>Se non si è soddisfatti di nessuna previsione, non considerare le previsioni della chatbot</p>

Design

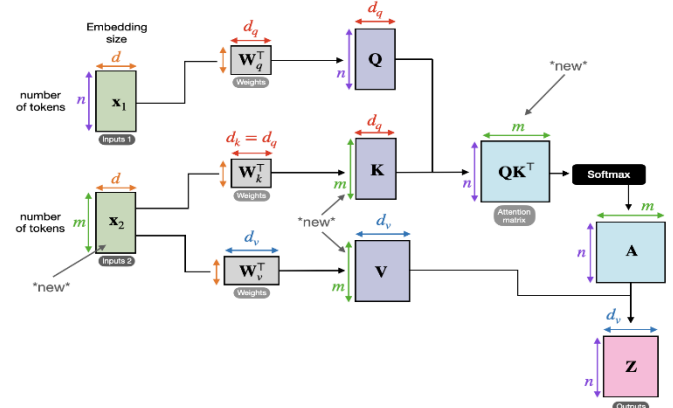
LLM

Il Large-Language-Model (LLM) utilizza principalmente l'architettura dei transformer costituiti da un encoder, seguito da un decoder:



L'encoder prende in input una sequenza di token (ad esempio, le parole di una frase) e la trasforma in una serie di rappresentazioni contestuali. È composto da più strati identici. Ogni strato dell'encoder è formato da due sottocomponenti principali:

- **Multi Head Self-Attention**; meccanismo che assegna a ogni coppia di token interne al testo un valore di attention, che indica quanto una parola è rilevante rispetto a quella corrente. Per ogni token, il modello calcola tre vettori:
 - **Query (Q)**; rappresenta ciò che il token sta cercando nel contesto
 - **Key (K)**; rappresenta le proprietà informative di ciascun token nella sequenza
 - **Value (V)**; contiene le informazioni che verranno aggregate



L'attenzione viene calcolata confrontando il vettore query di un token con i vettori key degli altri token, generando dei punteggi che determinano quanto ogni token debba contribuire all'aggiornamento della rappresentazione. Questi punteggi vengono poi

normalizzati (di solito tramite softmax) e usati per pesare i vettori value, producendo così una nuova rappresentazione per ciascun token.

La self-attention è però di tipologia multi-headed, in modo tale che l'attention sia calcolata molteplici volte in parallelo e i risultati siano combinati

- Feed-Forward Neural Network: Una rete neurale che opera in maniera indipendente su ciascun token per elaborare ulteriormente la rappresentazione, fornendo un'attivazione non lineare
- Connessioni residuali
- Normalizzazione del livello
- Attenzione del prodotto scalare
- Encoding posizionale: Poiché il meccanismo di self-attention non tiene conto dell'ordine dei token (è invariante rispetto alla permutazione), si aggiungono codifiche posizionali agli embedding iniziali. Viene utilizzata una rappresentazione sinusoidale della posizione (concatenando le funzioni sinusoidali di vari periodi)

Il decoder genera l'output (ad esempio, la traduzione di una frase) basandosi sulle rappresentazioni prodotte dall'encoder. Anche il decoder è composto da strati identici, ma con un'ulteriore componente:

- Masked Multi Head Self-Attention: Simile al meccanismo di self-attention dell'encoder, ma con una maschera che impedisce al modello di "guardare" al futuro (ossia, ai token che non sono ancora stati generati)
- Cross-Attention: Consente al decoder di focalizzarsi sulle informazioni rilevanti provenienti dall'encoder, integrando il contesto dell'input nella generazione dell'output
- Connessioni residuali
- Normalizzazione del livello
- Attenzione del prodotto scalare
- Multi Head Self Attention
- Encoding posizionale
- Livello lineare
- Livello softmax

RAG

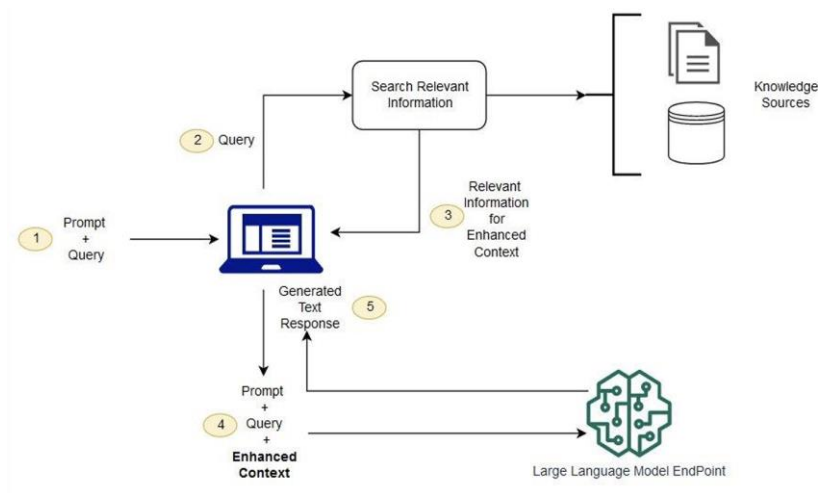
Retrieval-Augmented Generation è un modello ibrido che combina approcci basati su retrieval e quelli generativi. Esso è costituito da:

- Componente di retrieval; recupera documenti o conoscenze rilevanti da un database esterno e utilizza modelli di recupero come BM25. Di contro però manca di creatività o capacità di generare contenuti nuovi
- Componente di generazione; è il modello generativo che sintetizza le risposte o il testo utilizzando il contesto recuperato. Di contro però è limitato dalla possibile mancanza di informazioni aggiornate

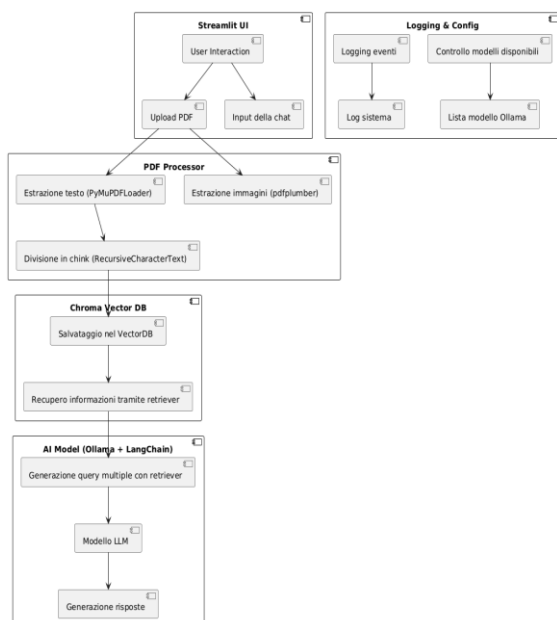
- Processo end-to-end; RAG opera in una pipeline Retrieve→Encode→Generate

Gli step che svolge sono i seguenti:

1. Formulazione della query; la query di input viene utilizzata per cercare documenti rilevanti in un corpus di recupero
2. Recupero dei documenti; il modello di retrieval recupera i primi k documenti che sono semanticamente rilevanti per l'input
3. Codifica del contesto; i documenti recuperati vengono codificati insieme alla query originale per formare un contesto ricco
4. Generazione; il modello generativo produce una risposta o un output utilizzando sia la query che il contesto recuperato

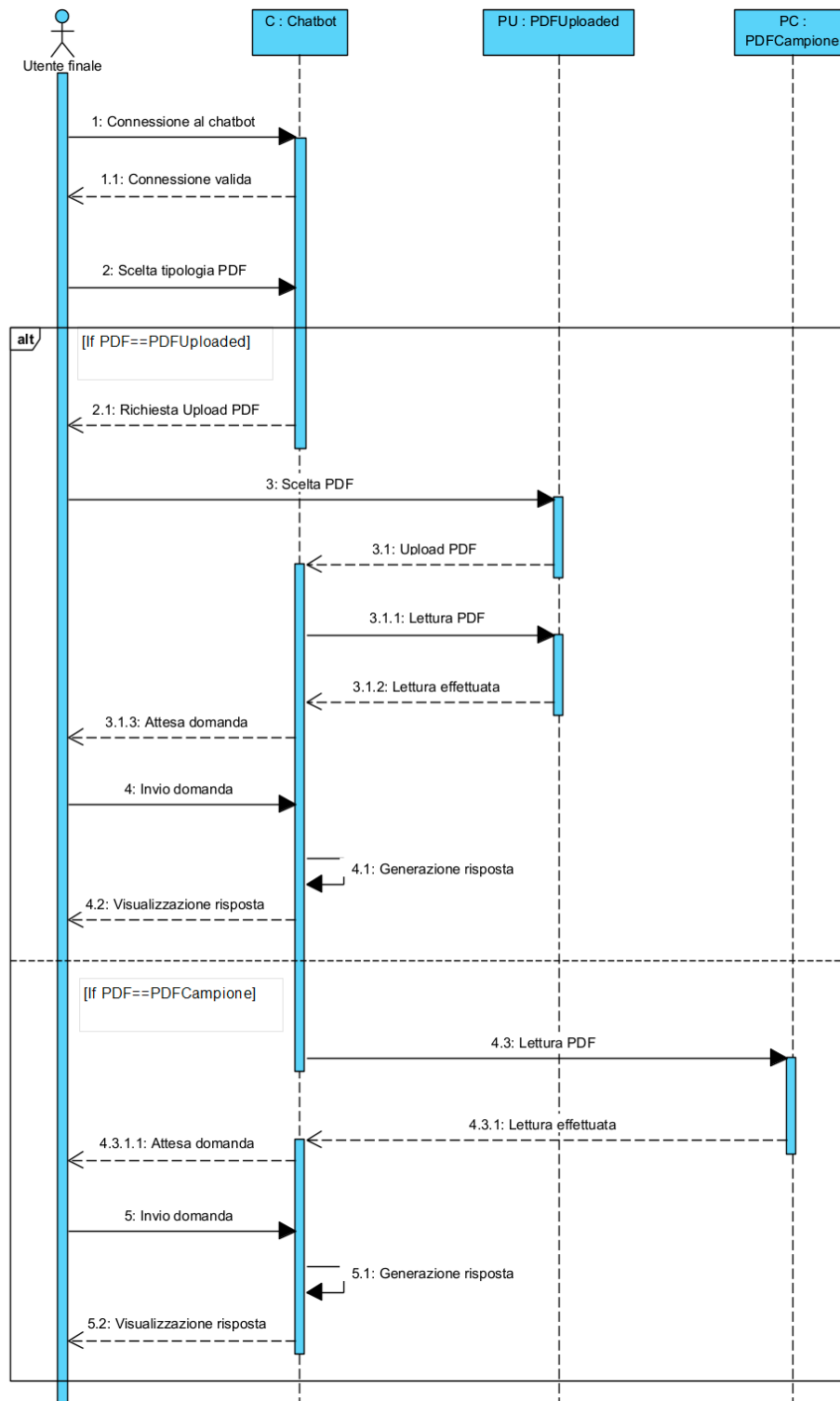


Internal Block Diagram

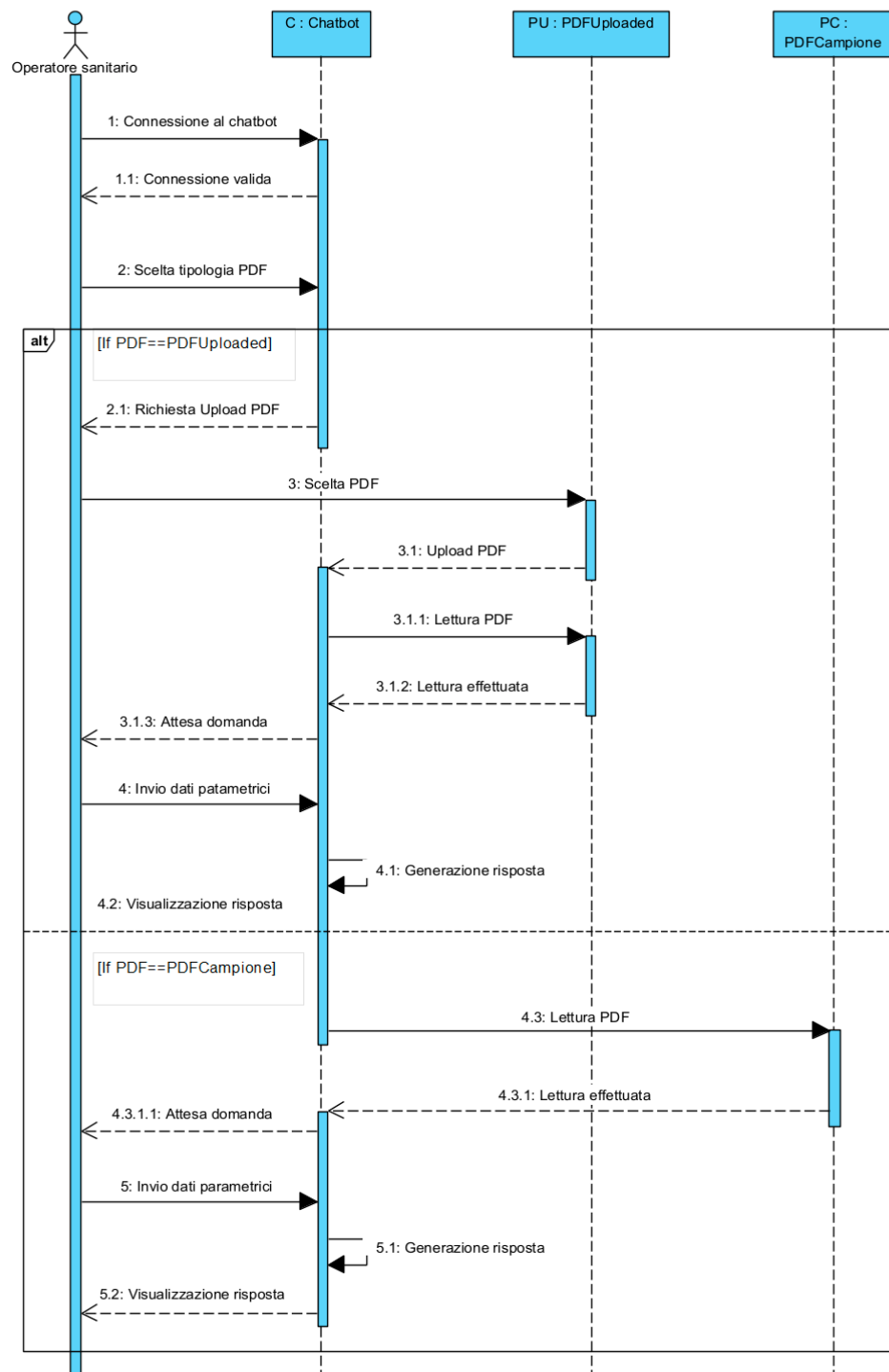


Sequence Diagrams

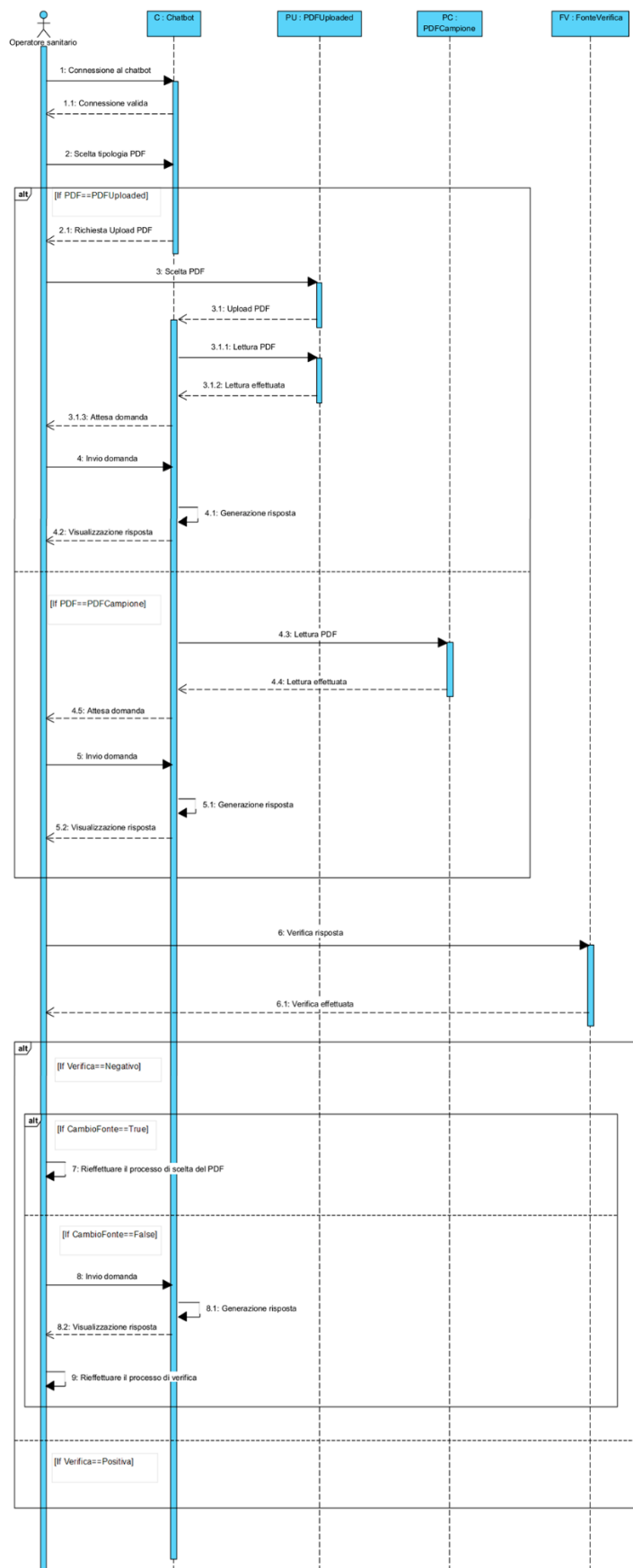
1. Gestione della crisi di voglia



2. Monitoraggio delle condizioni critiche



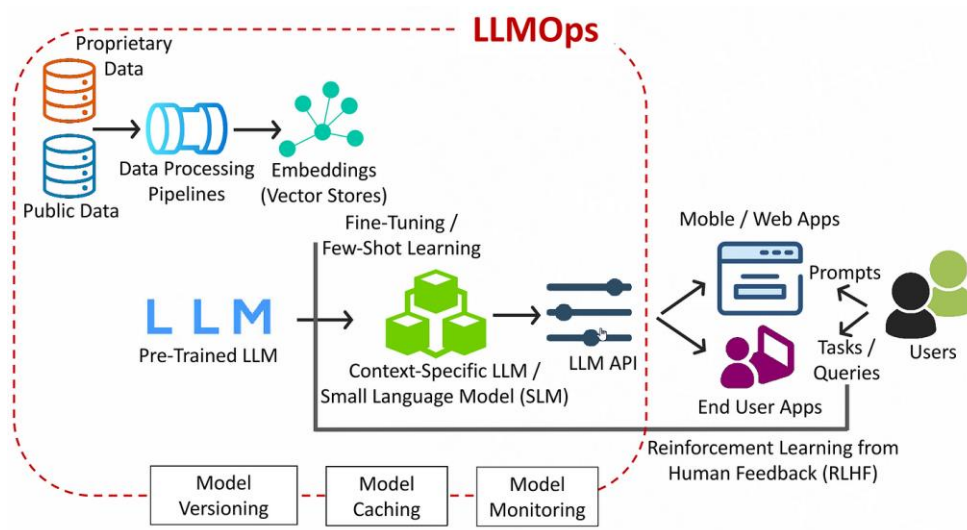
3. Verifica dell'evidenza e disclaimer



Development

LLMOps

Il ciclo di sviluppo che abbiamo eseguito è il seguente:



- Caricamento dei dati; si ha la possibilità di usare 2 tipologie di dati
 - File PDF di sample fornito stesso da noi
 - File PDF caricati dagli utenti stessi

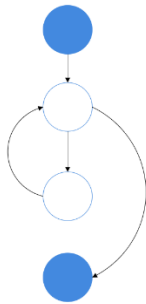
Abbiamo scelto di fornire un file in cui sono raggruppati molteplici libri riguardanti il fumo ed esempi di persone che sono riuscite a smettere di fumare e raccontano la loro esperienza (soggettiva, quindi limitata, ma non meno importante), in modo tale da dare una base a utenti inesperti. Inoltre, abbiamo deciso anche di dare la possibilità di caricare PDF in modo tale da accedere ad altre fonti, specialmente in casi in cui gli utenti sono gli psicologi e i medici (in generale persone con maggiori conoscenze)

- Elaborazione dei dati; i dati sono elaborati attraverso una pipeline in modo tale da ottenere degli embedding. Tale embedding è svolto da Nomic Embed Text
 - Nomic Embed Text è un modello di embedding del testo che converte il testo in vettori numerici. Utilizza un'architettura Mixture-of-Experts, in cui invece di utilizzare tutti i parametri su un input, sono attivati solo i parametri per elaborare l'input. Inoltre presenta un supporto multilingue
- LLM preaddestrato; è possibile utilizzare come modello preaddestrato Deepseek
- Fine-tuning; il modello viene addestrato utilizzando i parametri creati durante il preaddestramento
- Modello contestuale; il modello perfezionato è accessibile tramite un'API LLM
- Integrazione nelle applicazioni; il modello viene integrato in un'applicazione per gli utenti
 - Gli utenti possono interagire con il modello tramite l'utilizzo di un prompt (task/query)

- Monitoraggio continuo; si utilizza l'apprendimento tramite feedback umano (RLHF) per migliorare il modello
 - RLHF è una modalità di apprendimento in cui viene fornita una ricompensa ai vari output a seguito di una query/task. Tale ricompensa è basata su feedback umani; ovviamente, più è alto la ricompensa, migliore è la risposta. Da aggiungere alla ricompensa è un valore di malus, nel caso in cui il modello si distacchi troppo dai parametri ottimizzati

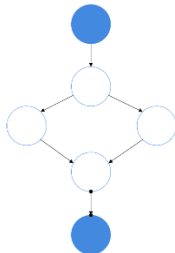
Control Flow Graph

- typewriter()



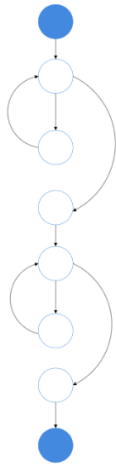
$$\text{numero cicломatico} = \# \text{archi} - \# \text{nodi} + 2 = 4 - 4 + 2 = 2$$

- extract_model_names()



$$\text{numero cicломatico} = \# \text{archi} - \# \text{nodi} + 2 = 6 - 6 + 2 = 2$$

- create_vector_db_multiple()



numero cicломatico = $\#archi - \#nodi + 2 = 9 - 8 + 2 = 3$

- `process_question()`



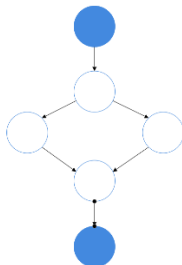
numero cicломatico = $\#archi - \#nodi + 2 = 2 - 3 + 2 = 1$

- `extract_all_pages()`



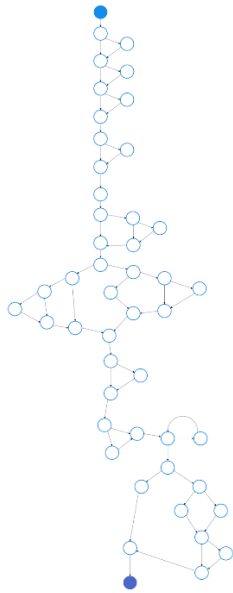
numero cicломatico = $\#archi - \#nodi + 2 = 2 - 3 + 1 = 1$

- `delete_vector_db()`



numero cicломatico = $\#archi - \#nodi + 2 = 6 - 6 + 2 = 2$

- `main()`



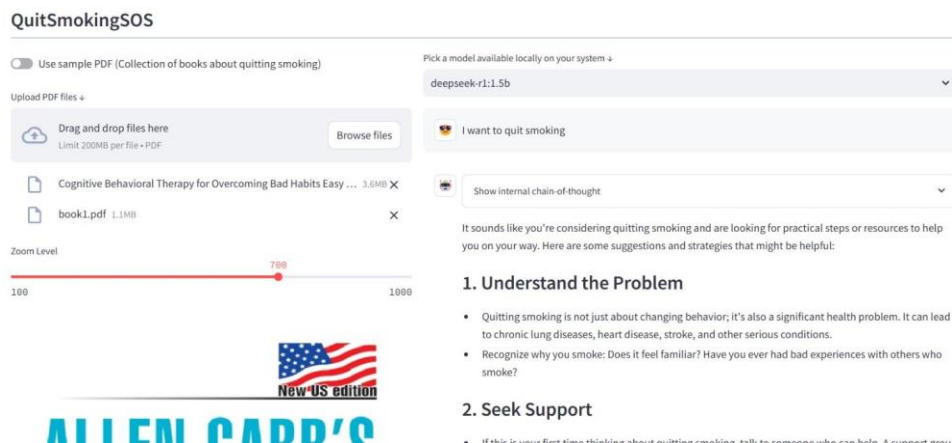
numero cicломatico = $\#archi - \#nodi + 2 = 55 - 56 + 2 = 1$

Testing

Input testing

Abbiamo testato l'applicazione su diversi possibili input:

- Input riguardante lo scopo di utilizzo del nostro chatbot, quindi in relazione allo smettere di fumare



[User-Input]: "I want to quit smoking"

[Oracolo-Output]: Consigli e risorse sullo smettere di fumare

[System-Output]: "It sounds like you're considering quitting smoking and are looking for practical steps or resources to help you on your way. Here are some suggestions and strategies that might be helpful:

1. Understand the Problem

Quitting smoking is not just about changing behavior; it's also a significant health problem. It can lead to chronic lung diseases, heart disease, stroke, and other serious conditions.

Recognize why you smoke: Does it feel familiar? Have you ever had bad experiences with others who smoke?

2. Seek Support

If this is your first time thinking about quitting smoking, talk to someone who can help. A support group or counselor can provide guidance and emotional support as you work through this.

The most effective way to quit smoking is often through professional programs that are tailored to your specific situation.

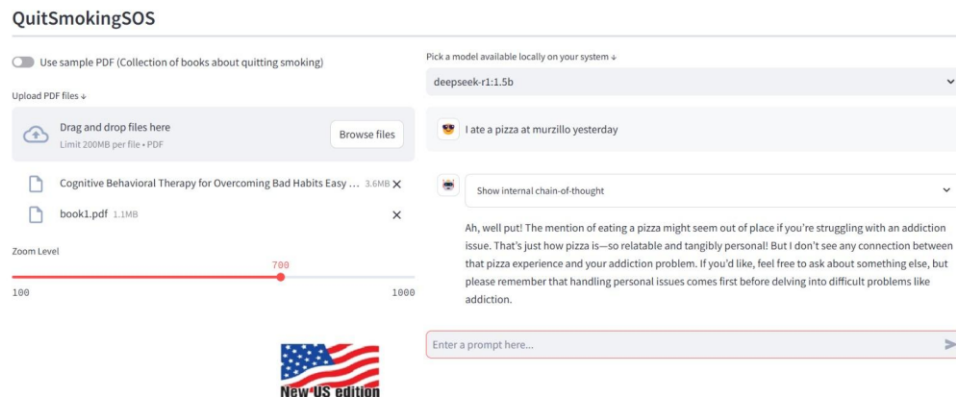
[...]

10. Consider External Support

If you can't take on the responsibility alone, consider reaching out to organizations that specialize in addiction treatment, such as addiction recovery centers or support groups.

By taking these steps, you can work toward quitting smoking and making a positive impact on your life. Remember, change is gradual, but it's worth working towards”

- Input riguardante argomenti differenti dallo scopo di utilizzo del nostro chatbot, di seguito

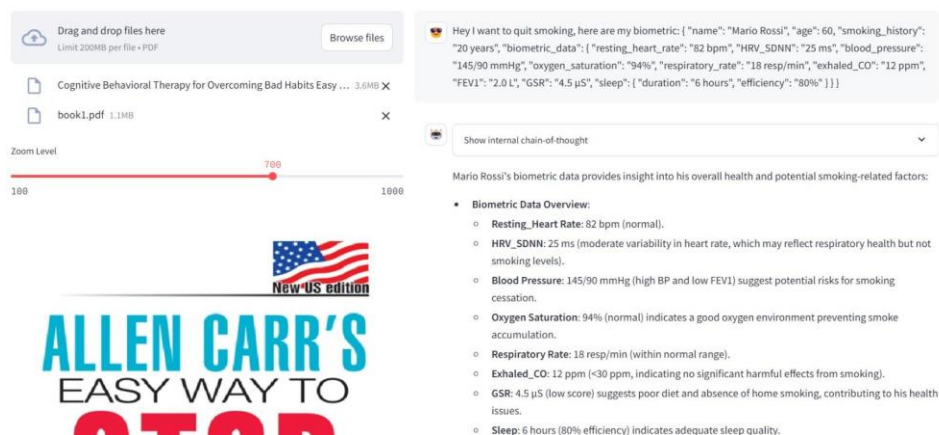


[User-Input]: "I ate a pizza yesterday at murzillo"

[Oracolo-Output]: Chatbot confuso

[System-Output]: "Ah, well put! The mention of eating a pizza might seem out of place if you're struggling with an addiction issue. That's just how pizza is—so relatable and tangibly personal! But I don't see any connection between that pizza experience and your addiction problem. If you'd like, feel free to ask about something else, but please remember that handling personal issues comes first before delving into difficult problems like addiction."

- Input con parametri biometrici



[User-Input]: "Hey I want to quit smoking, here are my biometric: { "name": "Mario Rossi", "age": 60, "smoking_history": "20 years", "biometric_data": { "resting_heart_rate": "82 bpm", "HRV_SDNN": "25 ms", "blood_pressure": "145/90 mmHg", "oxygen_saturation": "94%", "respiratory_rate": "18 resp/min", "exhaled_CO": "12 ppm", "FEV1": "2.0 L", "GSR": "4.5 uS", "sleep": { "duration": "6 hours", "efficiency": "80%" } } }"

[Oracolo-Output]: Analisi dei dati biometrici e conclusioni su stato salutare del paziente

[System-Output]: "Mario Rossi's biometric data provides insight into his overall health and potential smoking-related factors:

Biometric Data Overview:

Resting_Heart Rate: 82 bpm (normal).

HRV_SDNN: 25 ms (moderate variability in heart rate, which may reflect respiratory health but not smoking levels).

Blood Pressure: 145/90 mmHg (high BP and low FEV₁) suggest potential risks for smoking cessation.

Oxygen Saturation: 94% (normal) indicates a good oxygen environment preventing smoke accumulation.

Respiratory Rate: 18 resp/min (within normal range).

Exhaled_CO: 12 ppm (<30 ppm, indicating no significant harmful effects from smoking).

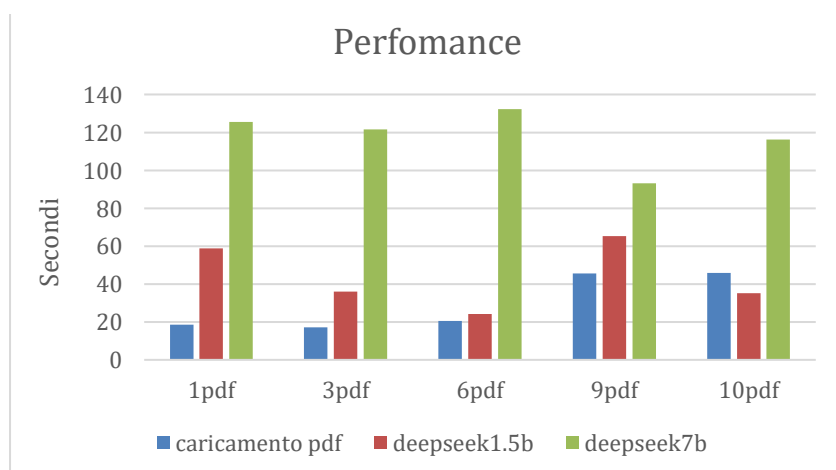
GSR: 4.5 μ S (low score) suggests poor diet and absence of home smoking, contributing to his health issues.

Sleep: 6 hours (80% efficiency) indicates adequate sleep quality.

Conclusion: Mario's high blood pressure, low FEV₁, and moderate GSR score are key indicators of smoking-related risks. His age with a significant smoking history and poor diet likely contribute to these findings. Continued efforts to quit smoking may improve his overall health profile."

Performance testing

Abbiamo fatto una misurazione sulla velocità del chatbot a seguito di varie operazioni al variare del numero di pdf caricati e del modello preso in considerazione (in tutti e 2 i casi è stato usato deepseek, ma il numero dei parametri presi in considerazione è differente)



Si constata la regolarità dell'andamento, come facilmente prevedibile:

- All'aumentare dei pdf, aumenta il tempo di caricamento (anche se leggermente in alcuni casi)
- All'aumentare dei parametri utilizzati dal modello, aumenta il tempo per la generazione della risposta

L'unica eccezione a questo andamento regolare è la diminuzione del tempo di generazione della risposta nell'utilizzo di deepseek1.5b in corrispondenza di 10 pdf

Ethics Guidelines for Trustworthy AI

Sono rispettate le seguenti linee guida etiche per AI definite dall'UE:

- Agenzia e supervisione umana; le risposte possono essere convalidate da professionisti
- Sicurezza e robustezza tecnica; il nostro sistema è stato sottoposto a rigorosi test di stress e sicurezza per garantire che sia resistente agli attacchi e alle situazioni di fallimento
- Privacy e controllo dei dati; rispettato tramite l'utilizzo del modello di crittografia OMEMO
- Trasparenza; dai risultati del testing, possiamo notare come le risposte generate risultano essere comprensibili, inoltre la possibilità di leggere chain-of-thought mostra il ragionamento dietro la generazione della risposta
- Diversità; uno degli obiettivi principali del nostro sistema è quello di non giudicare le persone durante il loro percorso
- Benessere ambientale e sociale; il nostro progetto è utilizzato per dare supporto morale alle persone in modo tale da migliorare la propria salute (psicologica e fisica)
- Responsabilità; gli utenti possono segnalare eventuali problemi e ottenere supporto

Deployment

Docker

In tal caso abbiamo deciso di utilizzare un container Docker:

- dockerfile:

```
FROM python:3.11

WORKDIR /app

COPY . /app

RUN pip install --no-cache-dir -r requirements.txt

EXPOSE 8501

CMD ["python", "run.py"]
```

- docker-compose:

```
services:
  ollama:
    image: ollama/ollama
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: all
              capabilities:
                - gpu
    container_name: ollama
    ports:
      - "11434:11434"
    volumes:
      - ollama_data:/root/.ollama

  streamlit_app:
    build:
      context: .
      dockerfile: Dockerfile
    container_name: streamlit_app
    ports:
      - "8501:8501"
    depends_on:
      - ollama
    environment:
      - OLLAMA_HOST=http://ollama:11434

volumes:
  ollama_data:
```

- requirements.txt:

```
ollama==0.4.4
streamlit==1.40.0
pdfplumber==0.11.4
langchain==0.3.14
langchain-core==0.3.29
langchain-ollama==0.2.2
langchain_community==0.3.14
langchain_text_splitters==0.3.5
unstructured>=0.16.12
unstructured[all-docs]>=0.16.12
onnx>=1.17.0
protobuf==5.29.2
chromadb>=0.4.22
Pillow==10.4.0
numpy==1.26.4
pytest==7.4.4
pytest-cov==4.1.0
coverage==7.4.0
pydantic==2.10.4
pymupdf
```