API_NOTES.md 2025-11-19

Note sull'Integrazione API della Chat

La chat è ora collegata a un'intelligenza artificiale esterna per generare le risposte.

1. Stato Attuale (con Groq)

    - **Funzionamento:** La chat utilizza l'API di Groq, un servizio che fornisce modelli di intelligenza artificiale molto veloci.
    - **Vantaggi:** È gratuito per i nostri test e offre risposte quasi istantanee.
    - **Sicurezza:** La chiave di accesso all'API è protetta e non è visibile pubblicamente.
    - **Nota sul Modello:** Per i test iniziali e la verifica della connessione, abbiamo utilizzato modelli base. Ora è configurato con un modello Llama 3.

2. Futura Integrazione con OpenAI (Opzionale)

    - **Flessibilità:** La struttura del codice è già predisposta per passare facilmente all'API di OpenAI, se necessaria.
    - **Vantaggi OpenAI:** Offre modelli di IA molto potenti e ampiamente riconosciuti.
    - **Considerazioni:** L'uso di OpenAI comporta costi, a differenza di Groq che è gratuito per i nostri test. Il passaggio sarebbe rapido e non richiederebbe grandi modifiche alla chat.

3. Gestione delle Risposte

    - **Tipologie:**
        - **Risposta Completa (Non-Streaming):** Il server invia l'intera risposta in un unico blocco. L'utente riceve il testo solo a generazione ultimata.
        - **Risposta in Streaming:** Il server invia la risposta in frammenti man mano che vengono generati, permettendo una visualizzazione progressiva.
    - **Implementazione Attuale:** Attualmente è adottata la **Risposta Completa (Non-Streaming)**.
    - **Complessità Streaming:** L'implementazione dello streaming richiede modifiche a livello di UI, Service Layer e Backend. Offre una User Experience più fluida e reattiva. È un compito fattibile, ma richiede un lavoro coordinato su tutti e tre i livelli architetturali e una buona conoscenza delle API di streaming e dell'ambiente serverless.

4. Gestione degli Allegati

- L'API di intelligenza artificiale attualmente utilizzata è pensata per conversazioni puramente testuali. Esistono API più avanzate, come i modelli multimodali di Google Gemini o l'Assistants API di OpenAI, pienamente capaci di ricevere e interpretare file o altri tipi di allegati.