Indelli Chatbot AI Assistenza Tecnica

Specifiche Tecniche del Sistema:

"Indelli AI Chatbots - Assistenza Tecnica Avanzata RAG"

Infrastruttura On-Premise, Motore Llama.cpp, Qwen 14B Instruct e RAG Ibrido BM25/Cosine Similarity

1. Introduzione e Obiettivi Architetturali

Il sistema "Indelli AI Chatbots - Assistenza Tecnica" è una piattaforma di intelligenza artificiale generativa progettata specificamente per fornire supporto tecnico specializzato, assistenza sistemistica e gestione dell'ecosistema software (inclusi i moduli Loquimini e le soluzioni Indelli). Interamente auto-ospitato (on-premise) e rilasciato sotto Licenza Open Source Apache 2.0, il sistema garantisce la totale sovranità dei dati, assenza di dipendenze da cloud esterne e conformità ai più rigidi standard di sicurezza della Pubblica Amministrazione.

2. Stack Tecnologico e Componenti di Sistema

L'architettura si basa su componenti software moderni, stabili e ad alte prestazioni:

- - Sistema Operativo: Debian 13 (Trixie) in configurazione minimale e blindata.
- - Front-End Web: Sviluppato in PHP 8.4, ottimizzato per gestire l'interfaccia di interazione utente e il flusso di streaming delle risposte.
- - Motore di Inferenza ed Embedding: Lame.cpp (llama.cpp) configurato ed eseguito come servizio di sistema nativo systemd, operante su una porta interna e non esposta direttamente all'esterno.
- - Modello Linguistico (LLM): Modello della serie Qwen in versione Instruct da 14 miliardi di parametri (14B), ottimizzato per il bilanciamento ideale tra capacità di ragionamento logico, precisione tecnica e consumo di risorse hardware.

3. Sistema di Retrieval-Augmented Generation (RAG) Ibrido

Per superare i limiti dei modelli generativi puri ed eliminare le allucinazioni in ambito tecnico, il chatbot integra un motore di RAG avanzato basato su un approccio ibrido:

- - Indicizzazione BM25: Ricerca basata su keyword e frequenza dei termini, ideale per individuare codici errore esatti, nomi di funzioni, comandi di sistema e parametri di configurazione.
- - Cosine Similarity (Vettoriale): Calcolo della somiglianza semantica tramite vettori di embedding generati dallo stesso motore locale, efficace per comprendere l'intento dell'utente anche a fronte di formulazioni colloquiali o descrittive.
- - Organizzazione dei Chunks: La base di conoscenza (Knowledge Base) è strutturata inserendo file di configurazione, guide operative e istruzioni puntuali per la gestione dell'ecosistema Loquimini/Indelli, affiancata dalla conoscenza tecnica nativa del modello su linguaggi di programmazione, script PHP/Python e infrastrutture Linux.

4. Gestione della Risposta in Streaming

L'interfaccia utente in PHP 8.4 comunica con il servizio di inferenza locale sfruttando flussi di dati asincroni in streaming (Server-Sent Events o chunked transfer). Questo permette all'operatore di visualizzare la risposta del modello token per token in tempo reale, riducendo drasticamente la percezione della latenza durante l'elaborazione di quesiti tecnici complessi.

5. Rilascio sotto Licenza Open Source Apache 2.0

Analogamente alle altre soluzioni della suite Indelli, il chatbot di assistenza tecnica è rilasciato sotto Licenza Apache 2.0. Tale scelta garantisce:

- - Trasparenza e Auditabilità totale del codice front-end e dei moduli di integrazione.
- - Conformità alle linee guida sul riuso del software della Pubblica Amministrazione (CAD).
- - Libertà di integrazione nei sistemi dell'Ente senza vincoli di copyleft o canoni di abbonamento software.
- I modelli da utilizzare sono in formato GGUF e solo ed esclusivamente con licenze permissive MIT o Apache 2
- llama.cpp e' in licenza MIT

6. Matrice di Valutazione Tecnica

Dimensione	**Svantaggio Soluzioni Cloud (es. ChatGPT / API terze)**	**Vantaggio Indelli AI Chatbots On-Premise**
Sovranità e Privacy dei Dati	Invio di codice sorgente e log di sistema a server di terze parti extra-UE (rischio GDPR).	Zero data leakage. Tutti i prompt e i file elaborati restano interamente confinati nell'hardware dell'Ente.
Precisione sul Dominio	Modelli generali che	RAG mirato con istruzioni

Specifico	ignorano le specificità di script custom o architetture proprietarie.	e file dell'ecosistema Loquimini/Indelli per risposte puntuali e prive di allucinazioni.
Costi e Sostenibilità	Canoni a consumo elevati (token-based) o abbonamenti SaaS ricorrenti e imprevedibili.	Costi fissi legati all'infrastruttura hardware esistente, senza costi nascosti per interrogazione.
Disponibilità Operativa	Dipendenza da connettività internet esterna e SLA di fornitori terzi.	Funzionamento continuo e autonomo all'interno della rete locale dell'Ente.