

Documentazione Tecnica Assistente IA Claire (Art. 11 IA ACT)

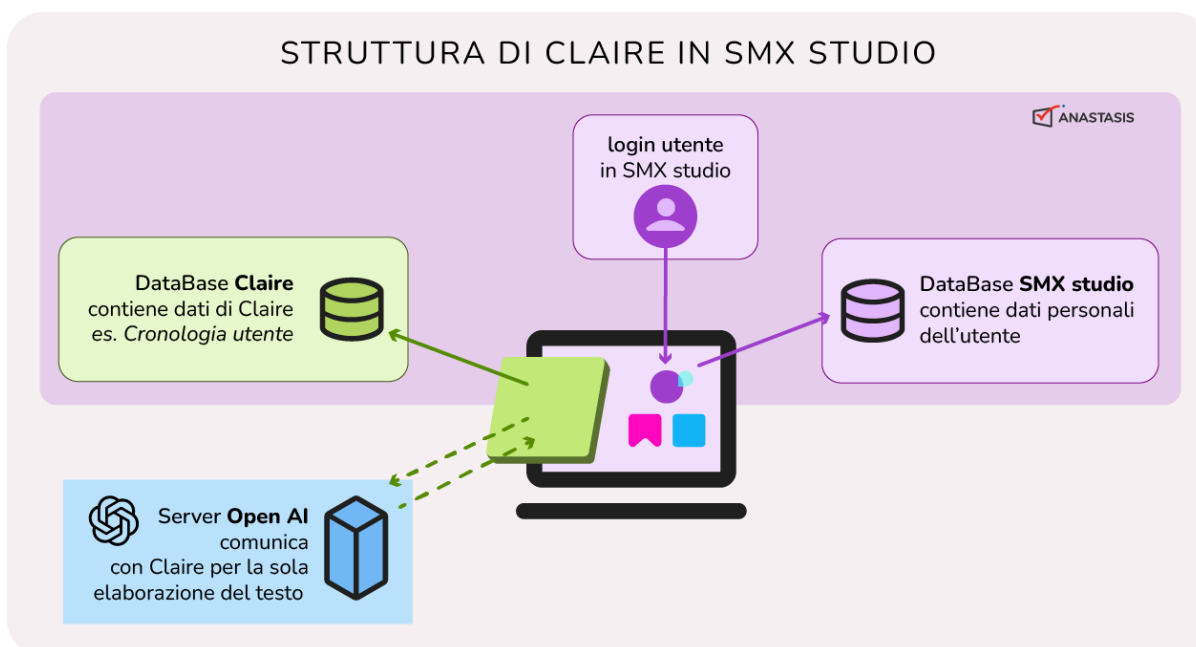
1. Descrizione Generale del Sistema di IA.....	2
2. Architettura del Sistema e Ciclo di Vita Tecnico.....	3
3. Governance dei Dati e Dataset.....	4
4. Monitoraggio, Sorveglianza Umana e Controllo.....	5
5. Gestione dei Rischi e Cybersicurezza.....	5
6. Valutazione dei rischi e accuratezza tecnica.....	6
7. Sistema di logging e tracciabilità dei risultati.....	6

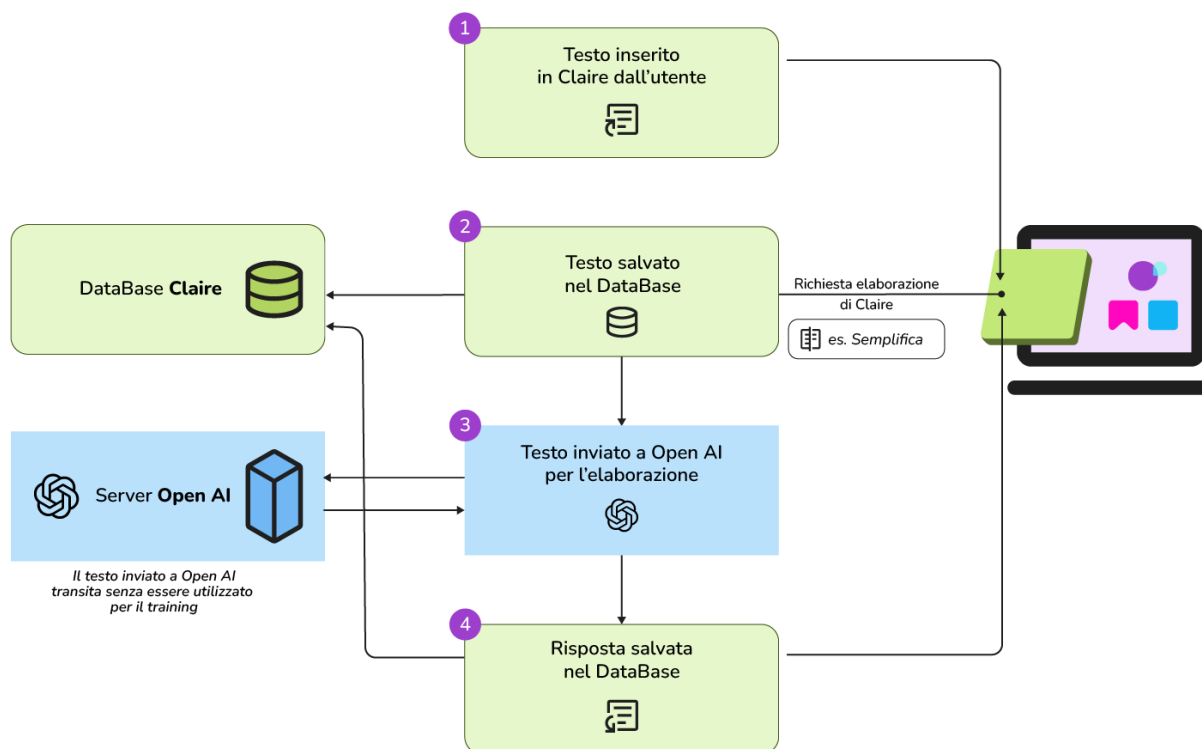
1. Descrizione Generale del Sistema di IA

- **Identificazione del sistema:** Claire è un'assistente basata sull'intelligenza artificiale, progettata per accompagnare studentesse e studenti nella comprensione dei testi. È presente esclusivamente all'interno della piattaforma SMX Studio.
- **Uso previsto:** Claire è pensata per studenti e studentesse, docenti, educatori, tutor e per contesti educativi eterogenei. È particolarmente utile quando il testo diventa un ostacolo: in presenza di difficoltà di comprensione, DSA, BES o semplicemente quando serve più tempo per capire. La sua interfaccia consente all'utente di inserire un testo, di lunghezza tra i 300 e i 10000 caratteri, per applicarvi 11 possibili funzionalità che restituiscono un nuovo testo, rielaborazione di quello inserito. Non è possibile per l'utente modificare i prompt relativi alle 11 funzionalità, fruibili esclusivamente tramite 11 bottoni.
- **Limitazioni d'uso:** Il sistema fornisce sempre una risposta, specificando quando il testo fornito non sia adatto alla funzionalità applicata e mostrando un avviso apposito nel caso in cui compaia nell'input e/o nell'output un lessico potenzialmente problematico.
- **Interazione con l'utente:** L'output non può essere modificato all'interno di Claire, può soltanto essere copiato e incollato altrove per la rielaborazione.
- **Specifiche dell'hardware:** Claire è ospitata sulla Google Cloud Platform e le funzionalità basate su AI Generativa si basano sulle API di OpenAI e di conseguenza non necessitano di infrastrutture dedicate.
- **Conformità** con requisiti obbligatori previsti dal Regolamento (UE) 2024/1689 (AI Act)
L'assistente IA Claire è sviluppata e aggiornata nel pieno rispetto del Regolamento UE 2024/1689 (AI Act) e della normativa nazionale italiana di recepimento (Legge 23 settembre 2025, n. 132), con particolare riferimento ai requisiti di trasparenza per i sistemi di IA destinati a interagire con persone fisiche.
Inoltre, **Il sistema è stato sottoposto** preliminarmente ad un **accurato processo di classificazione del rischio**, il cui **esito finale** è stata la collocazione nella fascia di **rischio basso**, in quanto svolge esclusivamente un **compito limitato/preparatorio** in **coerenza** con quanto previsto dall'**art. 6.3 del Reg. AI ACT**.

2. Architettura del Sistema e Ciclo di Vita Tecnico

- Logica computazionale:** Qui di seguito il link (costantemente aggiornato) relativo al modello attualmente in uso:
<https://sites.google.com/anastasis.it/claire-guidaoperativa#h.vm8izkpf89ze>.
- Flusso dei dati:** Una volta effettuato il login a SMX studio, l'utente può interagire con Claire digitando o inserendo del testo all'interno di un campo di input. Tramite degli appositi bottoni, può chiedere a Claire alcune rielaborazioni predefinite del testo inserito. Il testo inserito viene inviato a OpenAI per la rielaborazione, al termine della quale Claire visualizza la risposta. Durante la chiamata alle API di OpenAI non viene inviato alcun dato personale dell'utente. Il testo inserito e le risposte ricevute vengono salvate da Claire sul database, per consentirne l'uso da parte dell'utente, che ha comunque la possibilità di eliminarli in qualunque momento. Il solo testo inserito, e nessuna altra informazione dell'utente, transita sui server di OpenAI per consentire l'elaborazione del risultato, senza che esso venga utilizzato per il training.
- Scelte di progettazione:** La scelta di OpenAI e del modello in uso è stata subordinata a dei test sul rapporto qualità/prezzo sia automatizzati tramite il framework Promptfoo, sia qualitativi ed empirici grazie al supporto di figure esperte in ambito educativo.





3. Governance dei Dati e Dataset

All'interno di Claire vengono salvati solo i dati relativi all'utente e al suo utilizzo del software. In particolare è presente una cronologia delle rielaborazioni ottenute, i cui elementi possono essere eliminati in qualunque momento dall'utente, qualora lo desiderasse. L'utente visualizza i 10 elementi più recenti della propria cronologia di utilizzo, ed eliminandoli visualizza progressivamente quelli meno recenti. Gli elementi di cronologia che non siano tra gli ultimi 10 dell'utente vengono eliminati dopo 90 giorni dal loro ultimo utilizzo.

Claire non invia mai ad OpenAI dati personali dell'utente, ma soltanto il testo di input da rielaborare. Tale testo viene utilizzato da OpenAI esclusivamente per elaborare la risposta e non viene impiegato per il training.

Per quanto riguarda gli LLM di OpenAI, è possibile consultare i dettagli sulla raccolta dei dati per il training e sul processo di campionamento e data cleaning a [questo link](#). In breve

- **Provenienza dei dati:** si tratta di un dataset di oltre 10 miliardi di token di testi multilingua disponibili pubblicamente, dati ottenuti tramite partnership, dati sintetici, testi generati da umani, incluso testo dal web, materiali di riferimento di carattere tecnico, codice e altri testi preparati appositamente.
- **Metodologia di campionamento:** una parte dei dati di training è sintetica, ovvero creata tramite modelli di AI Generativa appositamente, per bilanciare i dati di

training in caso di scarsità o sbilanciamento. Questo consente la mitigazione di alcuni bias potenzialmente presenti nei dati disponibili.

- **Flusso di data-cleaning:** i testi vengono preparati tramite operazioni di filtraggio automatico e tramite modelli appositi, per rimuovere contenuti indesiderati, illegali o o per cui sia segnalato che non devono essere impiegate per l'addestramento di modelli.

4. Monitoraggio, Sorveglianza Umana e Controllo

- ➔ **Interfaccia di sorveglianza:** L'interfaccia di Claire consente di dare una valutazione positiva o negativa alla rielaborazione testuale ricevuta dal sistema, oppure di inviare una segnalazione tramite i canali di assistenza tecnica.
- ➔ **Metriche di prestazione:** Accuratezza e robustezza di Claire vengono misurati tramite test automatizzati con il framework Promptfoo e da tester umani. I test automatizzati si basano su testi scelti in modo mirato, in modo da poter monitorare la qualità dei risultati ottenuti in un contesto controllato. In parallelo, avviene un lavoro di controllo qualitativo da parte di esperti umani, che verificano l'attendibilità delle rielaborazioni ottenute e la loro qualità rispetto al tipo di supporto specifico che Claire deve fornire sulla comprensione del testo. La valutazione qualitativa da parte di esperti umani si avvale di un questionario che ciascun tester compila a seguito di ciascuna iterazione che possa alterare il funzionamento di Claire (i.e. cambiamento di modello, modifiche ai prompt, ecc.) e che consente di monitorare in modo quantificabile eventuali variazioni. Oltre al monitoraggio delle prestazioni nelle casistiche più frequenti, vengono effettuate delle verifiche mirate in merito a:
 - ◆ linguaggio problematico: Claire deve segnalare con un disclaimer apposito la presenza di linguaggio discriminatorio, violento, o sessualmente esplicito all'interno della rielaborazione fornita
 - ◆ testo fornito non adatto alla funzionalità applicata: Claire segnala con una risposta apposita che la funzionalità che l'utente intende applicare non è adatta al testo fornito
 - ◆ lingua della risposta: Claire ha l'indicazione di rispondere nella lingua del testo fornito in input

5. Gestione dei Rischi e Cybersicurezza

Tutte le informazioni sulla gestione dei rischi e della cybersicurezza da parte di OpenAI sono disponibili alle pagine:

- [OpenAI Trust Portal](#)
- [OpenAI Sicurezza a ogni passo](#)
- [OpenAI Product compliance status](#)

6. Valutazione dei rischi e accuratezza tecnica

Essendo Claire parte della piattaforma SMX studio, si rimanda alla [documentazione del medesimo](#) in merito alla valutazione dei rischi e all'accuratezza tecnica.

7. Sistema di logging e tracciabilità dei risultati

- **Tracciabilità delle operazioni (Log di sistema):** Claire implementa un sistema di logging automatico che registra gli eventi tecnici fondamentali. Questi log sono accessibili esclusivamente agli amministratori di sistema per attività di manutenzione, monitoraggio delle performance e troubleshooting, e non contengono informazioni direttamente riconducibili all'identità anagrafica dell'utente.
- **Registrazione dei feedback e delle anomalie:** Le valutazioni (positive/negative) e le segnalazioni inviate dagli utenti tramite l'interfaccia di sorveglianza vengono registrate e storicizzate separatamente. Questo flusso di dati permette al team di sviluppo e agli esperti umani di tracciare nel tempo i casi di malfunzionamento, i potenziali bias o le risposte non conformi, garantendo un registro utile al miglioramento continuo e alla verifica della robustezza del sistema.