

UNIVERSITÀ DEGLI STUDI DI SALERNO

DIPARTIMENTO DI INFORMATICA



Corso di Laurea Magistrale in Informatica

**Agenti virtuali adattivi: studio e design  
di agenti conversazionali in Extended  
Reality**

Relatore:

**Ch.ma Prof.ssa  
Giuliana VITIELLO**

Candidato:

**Matteo ERCOLINO  
Matr. 0522501462**

Correlatore:

**Dott.  
Andrea Antonio CANTONE**

ANNO ACCADEMICO 2024/2025

---

*Whatever you do in this life,  
it's not legendary,  
unless your friends are there to see it.*

---

## Sommario

La presente tesi indaga l'integrazione tra Extended Reality (XR), Large Language Models (LLM) e riconoscimento automatico delle emozioni vocali, con l'obiettivo di progettare agenti virtuali capaci di comunicare in modo immersivo, naturale ed empatico. La sinergia tra queste tecnologie consente di superare i limiti delle interfacce tradizionali, favorendo interazioni fluide e contestualmente consapevoli che tengono conto anche della dimensione affettiva dell'utente.

Dopo un'analisi critica dello stato dell'arte, viene proposta un'architettura modulare che suddivide l'elaborazione tra un ambiente XR, responsabile del rendering e della raccolta degli input, e un backend intelligente in grado di interpretare sia il contenuto semantico sia i segnali emotivi della voce. Il framework genera risposte vocali coerenti, adattate allo stato emotivo rilevato, garantendo al contempo una bassa latenza.

L'approccio è stato validato attraverso un prototipo impiegato in un contesto educativo simulato in realtà mista, dove un agente virtuale guida l'utente nell'esecuzione di un compito complesso. Uno studio comparativo con utenti reali ha messo a confronto una versione empatica dell'agente e una neutra: i risultati mostrano che l'agente empatico aumenta leggermente la percezione di presenza, riduce il carico cognitivo e peggiora di poco l'usabilità percepita, influendo in negativo su performance oggettive.

La tesi fornisce quindi un modello realistico e replicabile per lo sviluppo di agenti virtuali più reattivi e affettivamente consapevoli, con ricadute promettenti in ambito educativo, formativo e oltre.

# Indice

## Elenco delle figure

## Elenco delle tabelle

|          |  |           |
|----------|--|-----------|
| <b>1</b> | <b>Introduzione</b>  | <b>1</b>  |
| <b>2</b> | <b>Background</b>  | <b>4</b>  |
| 2.1      | Extended Reality . . . . .                                     | 4         |
| 2.1.1    | Virtual Reality, Augmented Reality e Mixed Reality . . . . .   | 5         |
| 2.1.2    | Reale vs Virtuale . . . . .                                    | 8         |
| 2.1.3    | Immersione e presenza . . . . .                                | 8         |
| 2.2      | LLM e AI Generativa . . . . .                                  | 9         |
| 2.2.1    | Modelli linguistici e generazione . . . . .                    | 10        |
| 2.2.2    | Adattamento affettivo . . . . .                                | 12        |
| 2.3      | Emozioni . . . . .   | 13        |
| 2.3.1    | Modelli computazionali delle emozioni . . . . .                | 14        |
| 2.3.2    | Emozioni e riconoscimento vocale . . . . .                     | 15        |
| <b>3</b> | <b>Stato dell'arte</b>   | <b>17</b> |
| 3.1      | Tecnologie immersive . . . . .                                 | 17        |
| 3.1.1    | Ambienti XR e paradigmi interattivi . . . . .                  | 17        |
| 3.1.2    | Agenti virtuali . . . . .                                      | 18        |
| 3.1.3    | Sfide attuali nella progettazione . . . . .                    | 19        |
| 3.2      | LLM in Extended Reality . . . . .                              | 20        |
| 3.2.1    | Esempi applicativi . . . . .                                   | 20        |
| 3.2.2    | Paradigmi di utilizzo . . . . .                                | 22        |
| 3.2.3    | Consapevolezze aumentate dai LLM . . . . .                     | 23        |
| 3.2.4    | Benefici e criticità . . . . .                                 | 24        |
| 3.2.5    | Pratiche progettuali . . . . .                                 | 26        |
| 3.3      | Architetture integrate XR-LLM-Emotion Recognition . . . . .    | 28        |
| 3.3.1    | Architetture multimodali e riconoscimento delle emozioni . . . | 28        |

|          |   |           |
|----------|---|-----------|
| 3.3.2    | Integrazione di LLM per l'adattamento empatico . . . . .  | 28        |
| 3.3.3    | Applicazioni in contesti industriali e sanitari . . . . . | 28        |
| 3.3.4    | Sfide aperte e direzioni future . . . . .                 | 29        |
| <b>4</b> | <b>Sistema sviluppato</b>                                 | <b>30</b> |
| 4.1      | Architettura e progettazione . . . . .                    | 30        |
| 4.1.1    | Linee guida progettuali . . . . .                         | 30        |
| 4.1.2    | Architettura generale del sistema . . . . .               | 35        |
| 4.1.3    | Componenti principali . . . . .                           | 36        |
| 4.1.4    | Flusso di interazione . . . . .                           | 43        |
| 4.1.5    | Considerazioni . . . . .                                  | 45        |
| 4.2      | Implementazione del sistema . . . . .                     | 46        |
| 4.2.1    | Panoramica dell'implementazione . . . . .                 | 46        |
| 4.2.2    | Lato Server . . . . .                                     | 47        |
| 4.2.3    | Lato Client . . . . .                                     | 49        |
| 4.2.4    | Integrazione Client–Server . . . . .                      | 52        |
| <b>5</b> | <b>Studio sperimentale</b>                                | <b>54</b> |
| 5.1      | Obiettivi . . . . .                                       | 54        |
| 5.2      | Domande di ricerca . . . . .                              | 55        |
| 5.3      | Condizioni sperimentali . . . . .                         | 56        |
| 5.4      | Disegno sperimentale . . . . .                            | 57        |
| 5.5      | Partecipanti . . . . .                                    | 58        |
| 5.6      | Procedura . . . . .                                       | 59        |
| 5.7      | Metriche e strumenti . . . . .                            | 61        |
| <b>6</b> | <b>Risultati sperimentazione</b>                          | <b>64</b> |
| 6.1      | Contesto sperimentale . . . . .                           | 64        |
| 6.2      | Analisi descrittiva . . . . .                             | 65        |
| 6.2.1    | Statistiche di base . . . . .                             | 66        |
| 6.2.2    | Visualizzazione delle metriche soggettive . . . . .       | 68        |
| 6.2.3    | Visualizzazione delle metriche oggettive . . . . .        | 69        |

|                       |  |           |
|-----------------------|--|-----------|
| 6.3                   | Analisi inferenziale . . . . .                             | 70        |
| 6.3.1                 | Risultati dei test d'ipotesi . . . . .                     | 71        |
| 6.3.2                 | RQ1 – Presenza percettiva . . . . .                        | 72        |
| 6.3.3                 | RQ2 – Carico cognitivo e <i>cybersickness</i> . . . . .    | 72        |
| 6.3.4                 | RQ3 – Usabilità percepita . . . . .                        | 73        |
| 6.3.5                 | RQ4 – Performance oggettive . . . . .                      | 73        |
| 6.3.6                 | Sintesi inferenziale . . . . .                             | 74        |
| 6.4                   | Trend emergenti e analisi esplorative . . . . .            | 75        |
| 6.4.1                 | Correlazioni tra metriche soggettive e oggettive . . . . . | 76        |
| 6.4.2                 | Sintesi esplorativa . . . . .                              | 76        |
| 6.5                   | Discussione e implicazioni . . . . .                       | 77        |
| 6.6                   | Minacce alla validità . . . . .                            | 77        |
| 6.6.1                 | Validità di conclusione statistica . . . . .               | 77        |
| 6.6.2                 | Validità di costrutto . . . . .                            | 78        |
| 6.6.3                 | Validità interna . . . . .                                 | 78        |
| 6.6.4                 | Validità esterna . . . . .                                 | 78        |
| 6.7                   | Linee di ricerca futura . . . . .                          | 79        |
| <b>7</b>              | <b>Conclusioni</b>   | <b>81</b> |
| <b>Bibliografia</b>   |  | <b>83</b> |
| <b>Ringraziamenti</b> |  | <b>88</b> |

## Elenco delle figure

|    |  |    |
|----|--|----|
| 1  | Continuum of Mixed Reality . . . . .   | 6  |
| 2  | Architettura Transformer . . . . .   | 11 |
| 3  | Chain-of-Empathy (CoE) prompting . . . . .   | 13 |
| 4  | Temi di ricerca identificati dalla review [29] . . . . .   | 20 |
| 5  | Schema ad alto livello dell'architettura del sistema . . . . .   | 35 |
| 6  | Flusso di interazione tra utente e sistema . . . . .   | 44 |
| 7  | Screenshot dell'ambiente MR nelle diverse fasi di preparazione del task di sintesi chimica. . . . .    | 50 |
| 8  | Variazioni visive del globo-agente nei tre stati di interazione. . . . .                               | 51 |
| 9  | Sequenza fotografica di un partecipante durante l'interazione con l'agente. . . . .                    | 61 |
| 10 | Medie con intervallo di confidenza al 95 % per le principali variabili. .                              | 67 |
| 11 | Distribuzione delle metriche soggettive per gruppo. . . . .  | 68 |
| 12 | Distribuzione delle metriche oggettive di performance. . . . .   | 69 |
| 13 | <i>Forest-plot</i> degli <i>effect size</i> . . . . .  | 71 |
| 14 | Scatter-plot con regressione lineare (fascia = IC 95 %) per le principali relazioni esplorate. . . . . | 76 |

## **Elenco delle tabelle**

|   |   |    |
|---|---|----|
| 1 | Confronto tra AR, MR e VR . . . . .   | 7  |
| 2 | Requisiti funzionali del sistema, con riferimento agli scenari d'uso . . . . .              | 32 |
| 3 | Requisiti non funzionali del sistema, con riferimento agli scenari d'uso . . . . .          | 33 |
| 4 | Mappatura tra moduli architetturali e requisiti soddisfatti . . . . .                       | 43 |
| 5 | Ipotesi nulle e alternative per ciascuna RQ. . . . .  | 56 |
| 6 | Metriche soggettive e oggettive utilizzate nello studio . . . . .                           | 63 |
| 7 | Media e deviazione standard ( $\bar{x} \pm s$ ) per gruppo. . . . .                         | 66 |
| 8 | Confronto EMO vs. NEU: statistiche del test, <i>p</i> -value e <i>effect size</i> . . . . . | 71 |
| 9 | Sintesi degli effetti osservati per ciascuna domanda di ricerca. . . . .                    | 75 |

# 1 Introduzione

Negli ultimi due decenni la *Extended Reality* (XR)—un termine ombrello che comprende Virtual Reality, Augmented Reality e Mixed Reality—ha guadagnato un ruolo centrale in domini quali la formazione professionale, la tele-collaborazione industriale e la simulazione scientifica. Parallelamente, l'avvento dei *Large Language Models* (LLM) ha trasformato le capacità dei sistemi conversazionali, rendendoli capaci di comprendere contesti complessi, di generare risposte coerenti e di apprendere nuovi compiti con un ridotto *fine-tuning*. Un terzo filone, quello del riconoscimento affettivo, ha fatto passi avanti grazie alla disponibilità di corpora multimodali e alla maturazione di modelli deep-learning che decodificano segnali non verbali—intonazione, ritmo, intensità—per inferire lo stato emotivo dell’utente.

L’integrazione sinergica di XR, LLM e analisi delle emozioni vocali apre la strada a interfacce immersive *empatiche*, capaci di adattare in tempo reale non solo il contenuto, ma anche il tono e la prosodia della risposta. Tale convergenza, tuttavia, pone sfide tecniche e metodologiche rilevanti: architetture distribuite che minimizzino la latenza pur garantendo moduli riusabili, orchestrazione di flussi asincroni fra percezione e ragionamento, coerenza temporale tra gesti e parlato dell’agente virtuale, oltre a protocolli di valutazione sperimentale che misurino l’effettivo valore aggiunto dell’empatia artificiale.

Questa tesi nasce dall’esigenza di colmare il divario fra le potenzialità offerte da questi tre ambiti e la mancanza di soluzioni integrate, documentate e riproducibili. Il lavoro si propone quindi di progettare, implementare e validare un framework modulare in cui un visore XR agisce da client leggero—responsabile del rendering tridimensionale, della sintesi vocale locale e della gestione degli input utente—mentre un backend a micro-servizi si occupa di *Automatic Speech Recognition*, analisi prosodica, orchestrazione del dialogo e inferenza LLM. La pipeline è pensata per funzionare in modalità strettamente asincrona: la trascrizione del parlato e l’estrazione di *features* emotive avvengono in parallelo; questi due flussi confluiscono in un modulo di *affect fusion* che produce un prompt dinamico per il LLM, il quale restituisce risposte semanticamente rilevanti e vocalmente modulate.

Il prototipo realizzato è stato validato in uno scenario di laboratorio di chimica in mixed-reality: lo studente, guidato da un agente virtuale, esegue in sicurezza una procedura di sintesi potenzialmente pericolosa. Il disegno sperimentale ha seguito un approccio *between-subjects* con due condizioni: una versione dell'agente che adatta il proprio stile comunicativo allo stato emotivo rilevato (EMO) e una controparte neutrale (NEU). Venti partecipanti, bilanciati per genere ed esperienza con la realtà virtuale, hanno completato la sessione; sono state raccolte metriche soggettive e oggettive.

Oltre all'architettura software e al prototipo XR, la tesi fornisce un contributo metodologico delineando un protocollo sperimentale replicabile. Benché il presente lavoro costituisca un primo passo verso agenti immersivi ed empatici, rimangono aperti numerosi filoni di ricerca. Sarà innanzitutto necessario ampliare la stima dello stato affettivo combinando canali multimodali – espressioni facciali, gesture e segnali fisiologici – così da ridurre l'ambiguità insita nell'analisi vocale pura. In parallelo, l'ottimizzazione della latenza attraverso tecniche di *edge computing*, quantizzazione e distillazione dei modelli linguistici potrà rendere l'architettura scalabile a scenari produttivi. Lungo un asse più metodologico, studi longitudinali e *in-the-wild* su campioni eterogenei permetteranno di valutare l'effetto dell'empatia artificiale su engagement, apprendimento e benessere a lungo termine. Infine, l'elaborazione di linee guida etiche e meccanismi di trasparenza rimane cruciale per un'adozione responsabile di interfacce capaci di influenzare lo stato emotivo dell'utente.

La tesi si articola in sette capitoli. Dopo la presente introduzione, il **Capitolo 2** presenta i fondamenti teorici di XR, LLM e modelli computazionali delle emozioni, fornendo il quadro concettuale necessario a comprendere le scelte progettuali successive. Il **Capitolo 3** offre una rassegna critica dello stato dell'arte, evidenziando lacune e opportunità nell'unione delle tre tecnologie. Il **Capitolo 4** descrive l'architettura e l'implementazione del sistema, soffermandosi su scelte tecnologiche, pipeline di *affect fusion* e strategie di ottimizzazione della latenza. Il **Capitolo 5** illustra il disegno sperimentale, le ipotesi di ricerca e le procedure di raccolta dati. Il **Capitolo 6** analizza i risultati, confronta le due condizioni sperimentali e discute le implicazioni pratiche. Il **Capitolo 7** conclude sintetizzando i contributi della

ricerca, riconoscendo le sue limitazioni e proponendo itinerari per studi futuri.

Attraverso la combinazione di un solido impianto teorico, uno sviluppo ingegneristico rigoroso e una valutazione empirica controllata, la presente tesi intende fornire un passo concreto verso interazioni persona-macchina più naturali, adattive ed empatiche all'interno di ambienti immersivi.

## 2 Background

Questo capitolo fornisce il quadro teorico e tecnologico di riferimento per l'analisi e la progettazione di interfacce conversazionali immersive in ambienti di Extended Reality (XR). Tali interfacce si collocano all'intersezione di tre ambiti chiave: le tecnologie immersive che definiscono lo spazio e la modalità dell'interazione; i modelli linguistici di grandi dimensioni (Large Language Models, LLM) che costituiscono il motore dialogico e generativo del sistema; e la dimensione affettiva ed emotiva, che rappresenta un aspetto cruciale per garantire esperienze naturali, coinvolgenti e significative dal punto di vista umano.

L'obiettivo di questa sezione è quello di delineare le basi concettuali e tecniche che caratterizzano ciascuna delle tre componenti, evidenziandone caratteristiche, potenzialità e limiti nei contesti applicativi della Human-Computer Interaction (HCI) e mostrare come queste dimensioni possano essere integrate in modo coerente per dare forma a interazioni avanzate.

Si parte dall'analisi delle tecnologie XR, con particolare attenzione alla differenziazione tra realtà virtuale, aumentata e mista, alla nozione di immersione e presenza, e alle implicazioni progettuali per la costruzione di ambienti interattivi credibili e adattivi. Si prosegue con lo studio dei LLM, approfondendone i principi architettonici, i meccanismi generativi e le strategie di adattamento affettivo, con l'obiettivo di comprenderne il ruolo come agenti conversazionali intelligenti in scenari situati. Infine, si esplora la dimensione emotiva, sia dal punto di vista teorico che computazionale, con un focus sul riconoscimento vocale delle emozioni e sui modelli utilizzati per simularle nei sistemi artificiali.

Queste tre dimensioni – ambienti immersivi, linguaggio generativo ed emozioni – costituiscono il fondamento su cui si basa il progetto di tesi.

### 2.1 Extended Reality

Il termine *Extended Reality* (XR) comprende l'insieme delle tecnologie immersive che spaziano dalla Virtual Reality (VR) alla Augmented Reality (AR), fino alla Mixed Reality (MR), lungo un continuum di integrazione tra reale e virtuale. Que-

ste tecnologie permettono di progettare esperienze interattive in grado di amplificare la percezione sensoriale dell’utente, modificando o arricchendo la rappresentazione dello spazio fisico e sociale.

Nel contesto della Human-Computer Interaction (HCI), XR rappresenta un’opportunità strategica per ripensare l’interazione uomo-macchina in chiave più naturale e contestuale. Le tecnologie XR non si limitano a simulare ambienti digitali, ma abilitano nuove forme di presenza, agency e comunicazione, fondamentali per lo sviluppo di interfacce conversazionali immersive. In questo quadro si collocano la VR, la AR e la MR, ognuna con caratteristiche specifiche e ambiti di applicazione distinti.

### 2.1.1 Virtual Reality, Augmented Reality e Mixed Reality

**Virtual Reality (VR)** La Virtual Reality immerge completamente l’utente in un ambiente generato digitalmente, scolliegandolo dai riferimenti sensoriali del mondo fisico [1]. L’interazione avviene tramite visori immersivi (es. Oculus Rift, Meta Quest 2) e controller tracciati che permettono movimenti nello spazio virtuale. In ambito HCI, la VR consente uno studio controllato del comportamento umano in ambienti simulati, utile per testare interfacce in scenari complessi. Tuttavia, l’isolamento sensoriale può ridurre la naturalezza dell’interazione e causare effetti indesiderati come cybersickness. La progettazione efficace richiede la gestione accurata di latenza, campi visivi e feedback multimodali.

**Augmented Reality (AR)** La Augmented Reality arricchisce la percezione dell’ambiente fisico sovrapponendo contenuti digitali in tempo reale. A differenza della VR, l’AR mantiene il contesto reale come sfondo attivo, supportando un’interazione più naturale e contestuale [1]. I dispositivi utilizzati includono smartphone, tablet, smart glasses (es. HoloLens 2) e visori passthrough (es. Meta Quest 3). L’AR è particolarmente adatta per scenari in cui la continuità con il mondo reale è essenziale: manutenzione industriale, formazione situata, navigazione aumentata, e assistenza vocale contestuale. Dal punto di vista dell’HCI, la progettazione di esperienze AR

richiede attenzione all'allineamento spaziale tra contenuto e ambiente, alla reattività del sistema e alla percezione dell'utente (occlusione, lighting, consistenza visiva).

**Mixed Reality (MR)** La Mixed Reality rappresenta un'evoluzione dell'AR in cui contenuti digitali e reali non solo coesistono ma interagiscono tra loro [1] [2]. Ciò richiede sistemi di tracking avanzati e comprensione semantica dell'ambiente. Dispositivi come Meta Quest 3 abilitano interazioni MR grazie a tecnologie di spatial mapping e passthrough ad alta fedeltà. In contesti HCI, la MR offre potenzialità per esperienze adattive, in cui agenti virtuali rispondono dinamicamente al comportamento e allo stato emotivo dell'utente. Le sfide progettuali riguardano l'ancoraggio spaziale, l'ergonomia cognitiva, e l'integrazione di feedback multisensoriali per mantenere l'illusione di coerenza e presenza.

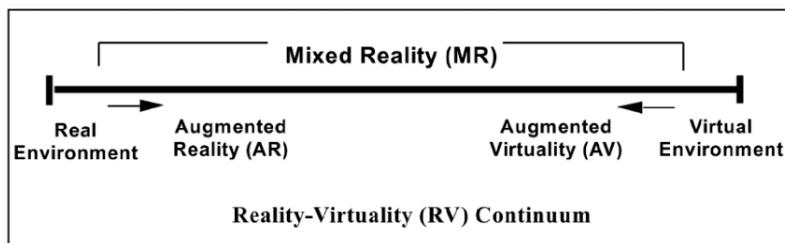


Figura 1: Continuum of Mixed Reality

Tabella 1: Confronto tra AR, MR e VR

| Caratteristica                       | Augmented Reality (AR)                                 | Mixed Reality (MR)  | Virtual Reality (VR)   |
|--------------------------------------|--|---|--|
| <b>Definizione</b>                   | Sovrappone contenuti digitali all'ambiente reale       | Integra contenuti digitali e reali con interazioni bidirezionali      | Sostituisce completamente la realtà con un ambiente simulato |
| <b>Continuità con il mondo reale</b> | Alta   | Media   | Nulla  |
| <b>Interazione con oggetti reali</b> | Presente, ma limitata alla sovrapposizione visiva      | Presente e semanticamente integrata                                   | Assente  |
| <b>Livello di immersione</b>         | Basso-Medio  | Medio-Alto  | Alto   |
| <b>Livello di presenza</b>           | Ancorato al mondo fisico                               | Dipende dalla coerenza tra reale e virtuale                           | Totalmente nel mondo simulato                                |
| <b>Dispositivi tipici</b>            | Smartphone, tablet, smart glasses (HoloLens)           | Visori passthrough avanzati (Meta Quest 3, Magic Leap 2)              | Visori immersivi (Oculus Rift, HTC Vive)                     |
| <b>Esempi d'uso</b>                  | Navigazione, formazione contestuale, assistenza remota | Prototipazione collaborativa, interazione avanzata in ambienti ibridi | Simulazioni, gaming, esposizioni immersive                   |
| <b>Sfide progettuali</b>             | Allineamento spaziale, coerenza visiva                 | Ancoraggio semantico, tracking preciso, fusione percettiva            | Cybersickness, isolamento perettivo, coerenza narrativa      |

### 2.1.2 Reale vs Virtuale

Come discusso da Milgram et al. [1], la distinzione tra reale e virtuale non è sempre netta nei contesti di realtà mista. Sebbene la VR tenda a creare ambienti completamente simulati e l'AR mantenga un ancoraggio diretto al mondo fisico, le tecnologie MR rendono necessaria una riflessione più articolata su cosa definisca un'entità come "reale" o "virtuale".

Milgram propone tre distinzioni operative:

- **Oggetti reali vs virtuali:** gli oggetti reali esistono fisicamente e possono essere osservati direttamente o tramite dispositivi di acquisizione; gli oggetti virtuali esistono solo in quanto simulati computazionalmente e resi visibili tramite un processo di rendering.
- **Immagini reali vs realistiche:** un'immagine può apparire realistica senza rappresentare un oggetto reale. Ad esempio, un video olografico di un oggetto simulato può apparire indistinguibile da uno reale, pur essendo completamente generato.
- **Immagini reali vs virtuali (ottica):** in termini fisici, un'immagine è considerata reale se esiste luce (luminosità) nel punto in cui appare. Un'immagine è virtuale se non emette luce nel punto percepito, come nel caso dei riflessi o delle immagini stereoscopiche.

Queste distinzioni sono fondamentali nella progettazione di esperienze XR: non è sufficiente che un oggetto "sembri reale" per esserlo, né che un'interfaccia "applichi l'AR" per garantire un'interazione significativa. È necessario considerare con precisione il livello di realtà rappresentato, la sua modalità di percezione e le aspettative cognitive ed emozionali che ne derivano.

### 2.1.3 Immersione e presenza

Due concetti fondamentali per la progettazione di esperienze efficaci in ambienti XR sono *immersione* e *presenza*. Essi fanno riferimento a fenomeni distinti ma interconnessi.

**Immersione** L’immersione è una proprietà oggettiva del sistema, che descrive il grado in cui la tecnologia è in grado di sostituire o amplificare gli stimoli sensoriali del mondo reale con stimoli generati artificialmente. Essa dipende da fattori tecnici quali la risoluzione visiva, il campo visivo (*field of view*), la latenza, la qualità del tracking spaziale e la disponibilità di feedback multimodali (audio spaziale, feedback aptici, ecc.). Un sistema altamente immersivo crea le condizioni necessarie affinché l’utente possa sospendere la propria consapevolezza dell’ambiente fisico, facilitando un coinvolgimento cognitivo e percettivo profondo.

**Presenza** La presenza è una risposta soggettiva dell’utente, ovvero la sensazione di “essere lì” (sense of being there) all’interno di un ambiente simulato o aumentato. Essa non dipende solo dal livello di immersione tecnica, ma anche dalla coerenza narrativa, dall’interattività, e dalla rilevanza personale dell’esperienza. Un alto senso di presenza è correlato a un maggiore coinvolgimento emotivo e a una migliore efficacia dell’interazione, sia in termini di usabilità che di impatto cognitivo e comportamentale.

Nella progettazione di interfacce in XR, immersione e presenza non sono meri effetti collaterali, ma obiettivi strategici. Ambienti che inducono un’elevata presenza favoriscono l’apprendimento, la motivazione e la collaborazione, mentre l’immersione controllata consente una manipolazione precisa dei fattori percettivi. In particolare, la presenza è essenziale per la credibilità di agenti conversazionali e avatar reattivi, poiché rafforza l’illusione di agency e intenzionalità dell’agente artificiale. In questo senso, la capacità del sistema di rispondere coerentemente diventa un requisito centrale per la progettazione di esperienze coinvolgenti ed efficaci in ambienti aumentati.

## 2.2 LLM e AI Generativa

Dopo aver analizzato le basi concettuali e tecnologiche della realtà estesa, è ora possibile introdurre una seconda componente fondativa del progetto: i modelli linguistici di grandi dimensioni.

Negli ultimi anni, i modelli linguistici di grandi dimensioni (*Large Language Models*, LLM) si sono affermati come una delle tecnologie centrali nel panorama dell'intelligenza artificiale generativa [3]. Basati su architetture neurali profonde e addestrati su grandi quantità di dati testuali, questi modelli sono in grado di comprendere e generare linguaggio naturale in modo coerente, fluido e contestuale. La loro versatilità li rende adatti a una vasta gamma di applicazioni, dalle chatbot alle interfacce conversazionali intelligenti, fino alla generazione creativa di contenuti.

Nel contesto della Human-Computer Interaction, i LLM rappresentano un cambio di paradigma nell'interazione testuale e vocale, in quanto permettono un dialogo dinamico, personalizzato e potenzialmente empatico tra utente e sistema. In ambienti immersivi come quelli offerti dalla Extended Reality, l'integrazione dei LLM apre nuove possibilità per la progettazione di interfacce conversazionali capaci di adattarsi al contesto e ai bisogni dell'utente.

### 2.2.1 Modelli linguistici e generazione

I *Large Language Models* sono modelli statistici del linguaggio basati principalmente sull'architettura Transformer [4]. Sono addestrati tramite apprendimento auto-supervisionato su corpora di testo di grandi dimensioni, con l'obiettivo di prevedere la parola successiva in una sequenza. Questo approccio, detto *generazione autoregressiva*, consiste nella produzione del testo parola per parola, dove ogni nuova parola è condizionata dalle precedenti.

A livello architetturale, i LLM sono costituiti da una sequenza di blocchi Transformer, ciascuno dei quali include due componenti principali: un meccanismo di auto-attention e una rete feed-forward. Il meccanismo di *self-attention* consente al modello di pesare dinamicamente l'importanza dei vari token nella sequenza di input, in base al contesto. In altre parole, ogni parola viene interpretata tenendo conto della relazione semantica con le altre, anche a lunga distanza.

Durante l'addestramento, i LLM apprendono rappresentazioni distribuite dei token attraverso processi di ottimizzazione che minimizzano la perdita di previsione della parola successiva [5]. Le parole vengono rappresentate tramite *embedding* densi

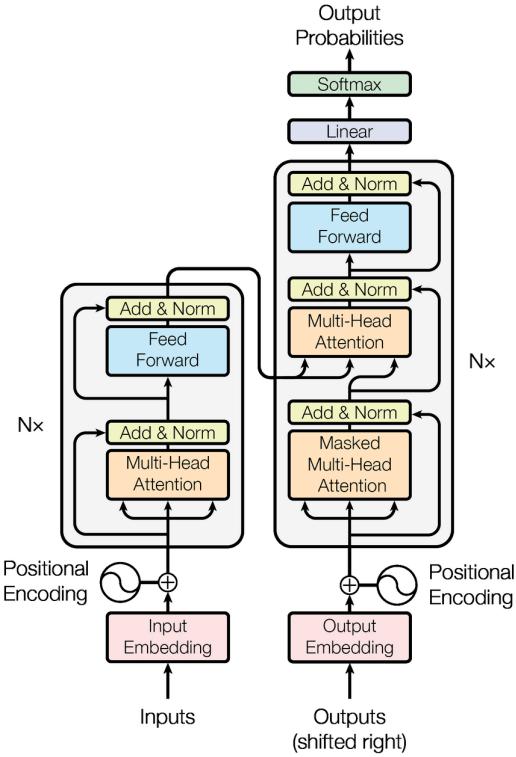


Figura 2: Architettura Transformer

in uno spazio vettoriale ad alta dimensionalità, che codifica similarità sintattiche e semantiche.

Nel caso dei modelli autoregressivi come GPT, la generazione avviene in modo incrementale: dato un contesto testuale iniziale  $x_1, x_2, \dots, x_t$ , il modello calcola la distribuzione di probabilità per il token successivo  $x_{t+1}$  tramite la funzione softmax applicata sull'output del Transformer. Il token più probabile (o campionato stocasticamente) viene poi aggiunto alla sequenza, e il processo si ripete.

Questo processo rende la generazione altamente sensibile al contesto iniziale, ma anche soggetta a deriva semantica o incoerenza se non correttamente vincolata. Per migliorare la coerenza e la pertinenza delle risposte, vengono spesso utilizzati meccanismi di *prompting* avanzato, *fine-tuning* supervisionato e tecniche di *reinforcement learning from human feedback* (RLHF) [6].

Un esempio di LLM è la famiglia di modelli GPT (Generative Pre-trained Transformer), sviluppata da OpenAI, che ha dimostrato capacità avanzate di comprensione semantica, coerenza sintattica, risposta contestuale e adattamento allo stile comunicativo dell'utente. Altri modelli noti includono BERT (Google), T5, LLaMA

(Meta) e Claude (Anthropic), ciascuno con architetture e obiettivi differenti, ma accomunati dalla capacità di operare su sequenze linguistiche complesse.

I LLM possiedono capacità conversazionali che emergono dalla combinazione tra ampia conoscenza linguistica, meccanismi di attenzione contestuale e capacità di mantenere coerenza nel turno di dialogo. Possono rispondere a domande, riassumere testi, tradurre, generare narrazioni, eseguire ragionamenti di base e adattarsi a diversi registri linguistici. Queste proprietà li rendono ideali per interfacce naturali in cui l’interazione verbale è il canale principale.

Tuttavia, l’adozione dei LLM in contesti sensibili richiede attenzione alle loro limitazioni: tra queste, la tendenza alla generazione di contenuti non verificabili (*hallucination*) [7], la mancanza di consapevolezza semantica profonda e la difficoltà nel modellare stati affettivi autentici. Questi aspetti sono particolarmente rilevanti per il presente lavoro, che indaga l’integrazione tra capacità linguistiche generative e intelligenza emotiva in ambienti immersivi.

### 2.2.2 Adattamento affettivo

Nonostante le avanzate capacità linguistiche, i *Large Language Models* (LLM) presentano limitazioni significative nella comprensione e gestione delle emozioni umane. Per affrontare queste sfide, sono state sviluppate diverse tecniche volte a migliorare la sensibilità emotiva dei LLM.

**Tecniche di adattamento affettivo** Una delle strategie emergenti è il *prompting affettivo*, che guida il modello nella generazione di risposte empatiche. Ad esempio, il metodo *Chain of Empathy* (CoE) si ispira a modelli psicoterapeutici come la Terapia Cognitivo-Comportamentale (CBT) per strutturare risposte che riflettano una comprensione profonda degli stati emotivi dell’utente [8].

Un’altra tecnica è l’uso di *fine-tuning* su dataset multimodali contenenti dialoghi emotivi, come nel caso di DialogueLLM, che integra informazioni contestuali ed emotive per migliorare il riconoscimento delle emozioni nelle conversazioni [9].

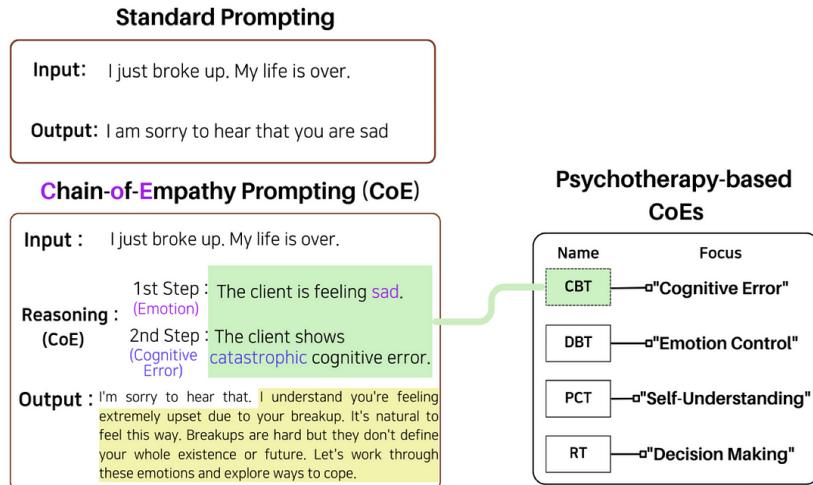


Figura 3: Chain-of-Empathy (CoE) prompting

**Supporto contestuale all’empatia artificiale** Poiché i LLM non possiedono una vera comprensione affettiva, la loro capacità empatica rimane simulata e dipende strettamente dalla qualità del contesto fornito in input. In quest’ottica, un approccio promettente non consiste nel tentare di generare empatia interna al modello, ma nel progettare meccanismi esterni che gli forniscano informazioni contestuali più ricche: lo stato emotivo dell’utente, il tono della voce, l’andamento della conversazione o la situazione ambientale. Integrare questi segnali nella fase di prompting o conditioning può migliorare significativamente la coerenza emotiva percepita delle risposte, rendendo l’interazione più naturale, soprattutto in ambienti immersivi dove l’esperienza è situata nel contesto spaziale, sociale ed emotivo dell’utente.

## 2.3 Emozioni

Le emozioni rappresentano un insieme complesso di risposte psicofisiologiche a stimoli interni o esterni, capaci di influenzare il comportamento, la cognizione e la comunicazione umana. Secondo la letteratura psicologica, le emozioni si distinguono in *primarie* e *secondarie*. Le emozioni primarie, anche dette fondamentali, sono universali, innate e biologicamente determinate: tra queste si annoverano gioia, tristezza, paura, rabbia, sorpresa e disgusto. Paul Ekman ha proposto un modello teorico che identifica tali emozioni fondamentali come riconoscibili attraverso

espressioni facciali universali, indipendenti da cultura e contesto [10].

Le emozioni secondarie, invece, sono apprese e si sviluppano a partire dalle emozioni primarie combinandosi con esperienze personali, contesto sociale e processi cognitivi complessi; esempi ne sono la colpa, la vergogna, l'orgoglio e la gelosia.

Le emozioni si articolano su tre componenti principali:

- **Fisiologica:** include le risposte corporee automatiche, regolate dal sistema nervoso autonomo, come l'aumento della frequenza cardiaca o la sudorazione.
- **Cognitiva:** riguarda l'interpretazione soggettiva dello stimolo emotivo, influenzata da aspettative, esperienze pregresse e contesto.
- **Comportamentale:** comprende l'espressione visibile dell'emozione, come gesti, postura, vocalizzazioni ed espressioni facciali.

Le emozioni svolgono un ruolo cruciale nei processi cognitivi e decisionali. Diverse ricerche hanno dimostrato come gli stati emotivi influenzino l'attenzione, la memoria e il ragionamento, contribuendo a orientare le scelte dell'individuo in situazioni complesse o incerte [11]. Inoltre, nelle dinamiche comunicative, le emozioni facilitano l'empatia e la comprensione reciproca, agendo come canali fondamentali per la regolazione dell'interazione sociale.

Nel contesto dell'interazione uomo-macchina, la modellazione e il riconoscimento delle emozioni assumono un ruolo strategico per migliorare l'efficacia, la naturalezza e l'accettazione delle interfacce.

### 2.3.1 Modelli computazionali delle emozioni

I modelli computazionali delle emozioni sono formalizzazioni matematiche o simboliche che mirano a rappresentare, riconoscere e simulare stati emotivi all'interno di sistemi intelligenti. La loro applicazione è particolarmente rilevante nei contesti di Human-Computer Interaction (HCI), robotica sociale, agenti conversazionali e ambienti immersivi.

Tra i modelli più noti e utilizzati vi sono il modello OCC e il modello PAD.

**Modello OCC** Proposto da Ortony, Clore e Collins (1988), il modello OCC è un approccio cognitivo simbolico che definisce le emozioni come valutazioni (*appraisal*) di eventi, agenti e oggetti rispetto agli obiettivi, standard e preferenze dell'individuo. Il modello identifica 22 categorie emozionali distinte, raggruppate in base al tipo di stimolo (es. evento positivo = gioia; agente responsabile = gratitudine).

**Modello PAD** Il modello PAD (Pleasure - Arousal - Dominance), introdotto da Mehrabian e Russell (1974), descrive gli stati emotivi su uno spazio continuo tridimensionale:

- **Pleasure** (piacere): misura quanto un'emozione è positiva o negativa.
- **Arousal** (attivazione): rappresenta il livello di eccitazione fisiologica.
- **Dominance** (dominanza): esprime il senso di controllo o impotenza percepita.

Questo approccio è ampiamente utilizzato per il riconoscimento e la classificazione automatica delle emozioni da segnali multimodali (voce, espressioni facciali, postura) e consente una rappresentazione continua e dinamica dello stato affettivo, adatta per l'adattamento in tempo reale del comportamento dell'agente.

L'integrazione di modelli come OCC e PAD nei sistemi intelligenti permette non solo di inferire l'emozione dell'utente, ma anche di regolare la risposta dell'IA in modo affettivamente coerente, migliorando la qualità dell'interazione in ambienti complessi e immersivi.

### 2.3.2 Emozioni e riconoscimento vocale

La voce umana è un canale espressivo ricco di informazioni. Il riconoscimento vocale delle emozioni si basa sull'analisi delle caratteristiche prosodiche del parlato, ossia quegli aspetti non linguistici del segnale vocale che riflettono lo stato emotivo dell'individuo. Le principali caratteristiche prosodiche utilizzate per l'inferenza emotiva includono:

- **Intonazione:** variazione della frequenza fondamentale (*pitch*) che può indicare stati come sorpresa (pitch alto) o tristezza (pitch basso).

- **Ritmo:** riguarda la velocità dell'eloquio e la durata delle pause, spesso influenzata da emozioni come ansia o calma.
- **Volume:** l'intensità del segnale vocale, che tende ad aumentare con emozioni attivate come rabbia e a diminuire in stati come paura o tristezza.

Il riconoscimento automatico delle emozioni vocali impiega tecniche di estrazione di caratteristiche acustiche (come MFCC, pitch contour, energia, jitter e shimmer) e algoritmi di classificazione, tra cui reti neurali, modelli HMM (Hidden Markov Models), SVM (Support Vector Machines) e più recentemente reti neurali profonde (DNN, CNN, RNN).

Numerosi studi hanno validato l'efficacia di tali tecnologie in contesti interattivi. Ad esempio, Schuller et al. (2011) [12] mostrano che è possibile raggiungere un'accuratezza significativa nell'identificazione di emozioni da voce spontanea utilizzando corpora annotati e tecniche di apprendimento automatico. Altri lavori, come quelli di Eyben et al. (2016) [13], hanno sviluppato toolkit open-source come *openSMILE*, oggi ampiamente utilizzato per il riconoscimento affettivo multimediale, inclusi gli assistenti vocali.

Nel contesto degli assistenti conversazionali, il riconoscimento vocale delle emozioni è fondamentale per l'adattamento dinamico delle risposte. Ad esempio, Google e Amazon hanno esplorato l'integrazione di moduli di affetto per modulare tono e contenuto in funzione dello stato dell'utente, migliorando l'accettazione e l'efficacia percepita del sistema [14] [15].

L'uso della voce come canale per l'inferenza affettiva si dimostra quindi strategico per l'interazione uomo-macchina, soprattutto in ambienti immersivi dove la comunicazione naturale è centrale.

L'intersezione tra queste tre dimensioni — riconoscimento emotivo, ambienti immersivi e modelli linguistici — costituisce il nucleo sperimentale di questa tesi. Le sfide affrontate non riguardano solo l'efficacia tecnologica, ma soprattutto la qualità relazionale dell'interazione, che in ambienti aumentati assume caratteristiche nuove e complesse.

## 3 Stato dell'arte

In questo capitolo si analizzano criticamente le ricerche più rilevanti relative all'integrazione tra tecnologie immersive, modelli linguistici generativi e sistemi di riconoscimento affettivo, con particolare attenzione ai recenti sviluppi nel campo della Human-Computer Interaction (HCI). L'obiettivo è delineare lo stato dell'arte delle tre dimensioni principali del progetto, per poi esplorare le convergenze che ne derivano.

### 3.1 Tecnologie immersive

#### 3.1.1 Ambienti XR e paradigmi interattivi

Le tecnologie di Extended Reality (XR), che comprendono realtà aumentata (AR), realtà mista (MR) e realtà virtuale (VR), stanno ridefinendo i paradigmi dell'interazione uomo-macchina, introducendo modalità basate su spazialità, multisensorialità e co-presenza [1, 2]. L'integrazione di gesti, movimento spaziale e feedback aptici consente forme di interazione più naturali rispetto ai tradizionali paradigmi WIMP (Windows, Icons, Menus, Pointer) [16]. La presenza di un contesto tridimensionale situato rafforza il coinvolgimento percettivo e cognitivo, facilitando l'apprendimento, la motivazione e la collaborazione.

Gli ambienti XR trovano applicazione in numerosi contesti d'uso:

- **Educazione immersiva:** l'XR offre agli studenti ambienti interattivi e simulati in cui esplorare concetti astratti e complessi attraverso esperienze visive e manipolative. Studi sistematici [17, 18] dimostrano che la VR migliora l'apprendimento esperienziale, la memorizzazione e la motivazione. In particolare, la realtà aumentata ha mostrato di essere efficace nell'apprendimento situato e nel supporto a studenti con difficoltà cognitive [19]. Le applicazioni variano dalla biologia immersiva all'archeologia virtuale, fino alla fisica interattiva.
- **Formazione professionale e simulazione:** XR consente la creazione di ambienti controllati ad alta fedeltà in cui addestrare competenze operative e decisionali senza rischi per la sicurezza. La VR è già impiegata con successo

nella formazione chirurgica, aerospaziale, militare e industriale [20, 21]. Queste simulazioni aumentano la retention, riducono i costi e migliorano l'efficacia rispetto alla sola formazione tradizionale. In particolare, i sistemi immersivi favoriscono l'apprendimento procedurale e il decision-making, rendendo l'XR ideale per la preparazione a scenari rari o critici.

- **Patrimonio culturale e musei:** l'XR permette nuove forme di fruizione e interazione nei musei, abilitando narrazioni dinamiche, ricostruzioni storiche e interazioni multimodali con oggetti digitalizzati [22, 23]. L'aumento del senso di presenza e agency dell'utente rende queste esperienze più coinvolgenti e memorabili.

Questi casi mostrano come la XR non sia solo una tecnologia abilitante, ma un vero e proprio *medium interattivo* che riconfigura i modi di apprendere, lavorare e comunicare. Tuttavia, la progettazione efficace di queste esperienze richiede una profonda comprensione delle dinamiche percettive e cognitive dell'utente, nonché dei vincoli tecnici e contestuali specifici del dominio applicativo.

### 3.1.2 Agenti virtuali

Gli agenti virtuali (VA) rappresentano un elemento chiave nell'evoluzione dell'interazione uomo-macchina all'interno degli ambienti XR. La loro capacità di simulare comportamenti sociali, rispondere in modo multimodale e adattarsi al contesto rende l'esperienza utente più immersiva e naturale.

Studi recenti hanno evidenziato l'importanza delle caratteristiche fisiche e comportamentali dei VA nella percezione dell'utente. Ad esempio, la dimensione e la forma dell'agente influenzano significativamente l'efficienza dell'interazione e il senso di immersione percepito [24]. Inoltre, l'integrazione di modelli generativi per la produzione di comportamenti realistici ha dimostrato di migliorare l'engagement dell'utente [25].

L'interazione multimodale, che combina input vocali, gestuali e visivi, è fondamentale per una comunicazione efficace con i VA. Un framework progettato per ambienti di realtà mista indossabile ha mostrato come l'uso simultaneo di ricono-

scimento vocale, tracciamento dello sguardo e animazioni corporee possa migliorare significativamente l'esperienza dell'utente [26].

Inoltre, la personalizzazione dei VA attraverso l'adattamento ai comportamenti e alle preferenze dell'utente è un'area di ricerca emergente. L'uso di modelli di linguaggio di grandi dimensioni (LLM) per analizzare e prevedere le azioni dell'utente in ambienti XR ha aperto nuove possibilità per la creazione di agenti più reattivi e contestualmente consapevoli [27].

Infine, l'aspetto sociale dei VA è cruciale. La loro capacità di esprimere emozioni, mostrare tratti di personalità distinti e interagire in modo credibile con l'utente contribuisce a creare un senso di presenza sociale, fondamentale per applicazioni educative, terapeutiche e di intrattenimento [28].

### 3.1.3 Sfide attuali nella progettazione

La progettazione di interazioni efficaci in ambienti XR pone numerose sfide che richiedono un approccio multidisciplinare e centrato sull'utente. Tra le principali criticità emergono:

- **Sovraccarico cognitivo e disorientamento:** L'immersione in ambienti virtuali complessi può causare sovraccarico cognitivo e disorientamento, compromettendo l'efficacia dell'interazione.
- **Interazione multimodale e naturale:** L'integrazione di input multimodali (voce, gesti, sguardo) richiede sistemi in grado di interpretare e sincronizzare correttamente le diverse modalità, garantendo un'interazione fluida e naturale. La progettazione di tali sistemi deve considerare la coerenza tra le modalità e la capacità dell'utente di controllarle efficacemente.

Affrontare queste sfide richiede un approccio progettuale che integri principi di Human-Centered Design, considerando le esigenze, le capacità e le limitazioni degli utenti, nonché le specificità tecnologiche degli ambienti XR.

**Design partecipativo e prototipazione iterativa** La progettazione di interfacce conversazionali immersive richiede metodologie HCI user-centered. Il *design*

*partecipativo* coinvolge gli utenti finali nella definizione di funzionalità, flussi e contenuti, garantendo una maggiore aderenza ai bisogni reali. Inoltre, la *prototipazione iterativa* — mediante strumenti low-fidelity (es. storyboard, wizard-of-oz) o ambienti XR simulati — permette di validare precocemente le scelte progettuali riducendo i costi di sviluppo. Tali approcci sono particolarmente efficaci per l’ideazione di comportamenti dialogici dei LLM e per la calibrazione delle risposte emotive in ambienti immersivi.

### 3.2 LLM in Extended Reality

La convergenza tra Large Language Model (LLM) e ambienti di Extended Reality (XR) rappresenta un’area di ricerca emergente nell’ambito HCI. I LLM potenziano l’interazione immersiva rendendola più naturale, adattiva e contestualmente consapevole. Questa sinergia ha aperto nuove prospettive nella progettazione di esperienze interattive avanzate, in cui l’interazione conversazionale diventa adattiva, contestuale e multimodale [29].

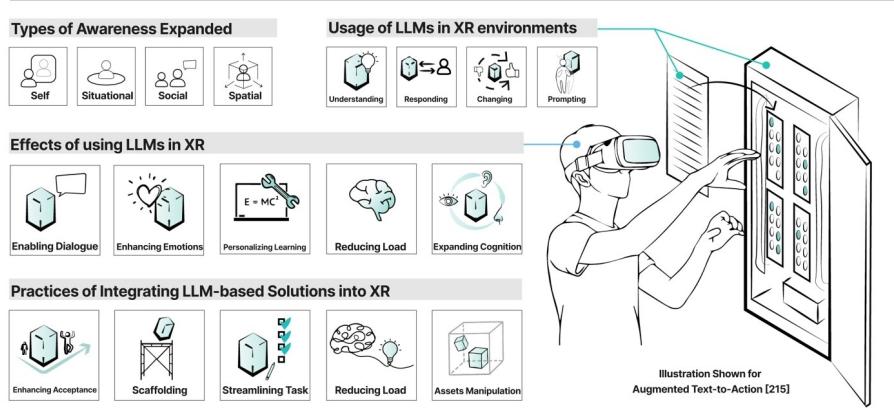


Figura 4: Temi di ricerca identificati dalla review [29]

#### 3.2.1 Esempi applicativi

Diversi progetti recenti hanno esplorato l’integrazione dei Large Language Models (LLM) in ambienti di Extended Reality (XR), focalizzandosi sull’interazione tra esseri umani e agenti virtuali intelligenti. Di seguito, alcuni esempi significativi:

**GesPrompt: Interazione multimodale con co-speech gestures** Il progetto GesPrompt introduce un’interfaccia XR che combina input vocali e gestuali per migliorare la comunicazione con agenti LLM. Gli utenti possono utilizzare gesti sincronizzati con il parlato per fornire riferimenti spaziali e temporali, facilitando l’interazione in ambienti tridimensionali. Uno studio ha evidenziato che questa modalità multimodale riduce il carico cognitivo e migliora la precisione nell’esecuzione dei comandi [30].

**CUIfy: NPC conversazionali in ambienti XR** CUIfy è un pacchetto open-source per Unity che consente l’integrazione di NPC (Non-Player Characters) alimentati da LLM in ambienti XR. Il sistema supporta l’interazione vocale bidirezionale, combinando modelli di riconoscimento vocale, sintesi vocale e LLM per creare dialoghi naturali e contestualmente rilevanti. CUIfy è progettato per essere modulare e facilmente integrabile in diverse applicazioni XR [31].

**Milo: Piattaforma per agenti virtuali basati su LLM** Milo è una piattaforma open-source che permette la creazione di agenti virtuali intelligenti in ambienti XR. Gli agenti possono partecipare a eventi live, sessioni di auto-riflessione e incontri multiutente, adattando il loro comportamento in base al contesto e alle interazioni con gli utenti. La piattaforma è stata utilizzata in vari scenari, dimostrando la versatilità degli agenti LLM in contesti sociali immersivi [32].

**Explainable XR: Analisi del comportamento utente assistita da LLM** Il framework Explainable XR utilizza LLM per analizzare e interpretare i comportamenti degli utenti in ambienti XR. Attraverso la raccolta di dati multimediali e l’uso di descrittori di azioni utente, il sistema fornisce insight dettagliati sulle interazioni, facilitando la comprensione delle dinamiche comportamentali in scenari immersivi. Questo approccio è stato applicato in contesti educativi e collaborativi [27].

**Tell-XR: Sviluppo di automazioni XR tramite dialogo** Tell-XR è un agente intelligente che assiste gli utenti nello sviluppo di automazioni in ambienti XR attraverso interazioni conversazionali. Utilizzando modelli LLM, il sistema guida l’utente

nella definizione di regole Event-Condition-Action (ECA), semplificando la personalizzazione dell'esperienza XR senza necessità di competenze di programmazione. Il sistema è stato valutato in scenari come musei virtuali e smart home [33].

**XaiR: Integrazione di LLM con interazioni nel mondo reale** XaiR è una piattaforma che combina le capacità dei LLM con interazioni nel mondo reale, utilizzando dispositivi XR per fornire assistenza intelligente basata sul contesto fisico dell'utente. Il sistema è progettato per comprendere l'ambiente circostante e fornire risposte o suggerimenti pertinenti, migliorando l'efficacia delle interazioni uomo-macchina in scenari quotidiani [34].

Questi esempi evidenziano le diverse modalità con cui i LLM possono essere integrati in ambienti XR per migliorare l'interazione, la personalizzazione e l'efficacia delle esperienze immersive. La combinazione di input multimodali, comprensione contestuale e generazione di linguaggio naturale apre nuove prospettive per l'evoluzione dell'interazione uomo-macchina in contesti immersivi.

### 3.2.2 Paradigmi di utilizzo

L'integrazione degli LLM in ambienti XR può essere classificata in tre principali paradigmi di utilizzo [29]:

- **Comprensione dell'utente e del contesto:** i LLM vengono impiegati per interpretare input multimodali (voce, gesto, scena visiva) e inferire intenzioni, stato emotivo e bisogni dell'utente.
- **Risposta a richieste contestuali:** i modelli generano risposte linguistiche coerenti con lo stato dell'ambiente XR, adattandosi dinamicamente ai prompt o alle azioni dell'utente.
- **Stimolazione all'azione:** i sistemi guidano l'utente nel compimento di compiti suggerendo azioni pertinenti rispetto al contesto e agli obiettivi dichiarati.

### 3.2.3 Consapevolezze aumentate dai LLM

L'integrazione dei Large Language Models negli ambienti di Extended Reality consente di potenziare diverse forme di consapevolezza cognitiva e situata, che sono centrali per l'efficacia e l'adattività delle interazioni. La letteratura recente identifica quattro dimensioni principali di *awareness* abilitate o amplificate dall'uso dei LLM in XR [29]:

1. **Consapevolezza situazionale** (*situational awareness*): è la capacità dell'utente di percepire, comprendere e anticipare gli eventi in un ambiente dinamico. I LLM, integrati in sistemi XR, possono monitorare il contesto in tempo reale (tramite sensori, input vocali, dati ambientali) e fornire feedback tempestivi o raccomandazioni operative. Questo è particolarmente utile in scenari critici come l'addestramento medico d'emergenza, dove il sistema può riconoscere schemi di azione e supportare decisioni contestuali con suggerimenti proattivi [29].
2. **Consapevolezza spaziale** (*spatial awareness*): riguarda la comprensione delle relazioni geometriche tra l'utente, gli oggetti e lo spazio circostante. I LLM possono contribuire a potenziare questa dimensione interpretando riferimenti spaziali (es. “sulla tua sinistra”) e generando istruzioni linguistiche coerenti con la mappa ambientale o l'orientamento dell'utente. [29].
3. **Consapevolezza sociale** (*social awareness*): è la capacità di cogliere segnali comunicativi, emozioni e intenzioni degli altri interlocutori (umani o virtuali). In ambienti collaborativi XR, i LLM permettono di generare agenti dotati di comportamenti verbali e non verbali realistici, abilitando l'inferenza di stati emotivi e la gestione del turno di parola. Inoltre, possono fornire spiegazioni, parafrasi o sintesi delle interazioni, facilitando l'allineamento semantico tra partecipanti e migliorando la qualità della comunicazione [29].
4. **Consapevolezza di sé** (*self-awareness*): è la capacità dell'utente di riflettere sulle proprie azioni, emozioni e stati mentali. I LLM possono generare feedback personalizzati in tempo reale o a posteriori, aiutando l'utente a mo-

nitorare i propri progressi, a riconoscere errori o bias, e a prendere decisioni più consapevoli. In contesti educativi, questa dimensione si traduce in esperienze di auto-valutazione, introspezione guidata o coaching conversazionale orientato al miglioramento continuo [29].

Queste forme di consapevolezza aumentata rendono l'interazione in XR più adattiva, riflessiva e situata, ponendo le basi per ambienti intelligenti capaci di reagire non solo agli input esplicativi dell'utente, ma anche ai suoi stati impliciti, alle dinamiche sociali e ai cambiamenti ambientali. La capacità dei LLM di mantenere coerenza contestuale e di modellare l'intenzionalità comunicativa rappresenta una componente chiave per la progettazione di agenti conversazionali immersivi e cognitivamente sensibili.

### 3.2.4 Benefici e criticità

L'integrazione dei Large Language Models (LLM) negli ambienti di Extended Reality (XR) apre a scenari altamente innovativi per la progettazione di esperienze conversazionali immersive.

**Benefici** Tra i principali benefici si annoverano:

- **Interazione più naturale:** i LLM, grazie alla loro capacità di comprendere e generare linguaggio naturale, abilitano modalità di dialogo fluide, coerenti e contestualmente rilevanti, migliorando la sensazione di agency e presenza dell'utente [29].
- **Adattamento dinamico al contesto:** attraverso l'elaborazione di input multimodali (voce, sguardo, gesto, contesto spaziale), i modelli possono generare risposte personalizzate e pertinenti rispetto allo stato corrente dell'ambiente XR e del comportamento dell'utente.
- **Riduzione del carico cognitivo:** la capacità dei LLM di semplificare istruzioni, riassumere contenuti o fornire suggerimenti puntuali permette di supportare l'utente durante l'esecuzione di compiti complessi, migliorando la performance e diminuendo l'affaticamento mentale [29].

- **Personalizzazione e accessibilità:** l’adattamento allo stile comunicativo, al livello di competenza e al profilo cognitivo dell’utente consente di costruire esperienze più inclusive, accessibili e coinvolgenti, soprattutto in ambiti educativi, terapeutici o assistivi.

**Criticità** Accanto a queste opportunità, emergono tuttavia alcune criticità tecniche e progettuali che rendono la piena adozione dei LLM in XR ancora sfidante:

- **Allineamento semantico:** mantenere la coerenza tra il linguaggio generato dal modello e lo stato effettivo dell’ambiente virtuale richiede un continuo aggiornamento del contesto. Errori nell’allineamento semantico possono compromettere la credibilità dell’agente e la qualità dell’interazione [29].
- **Gestione dell’ambiguità e prevenzione delle *hallucinations*:** i LLM possono produrre risposte plausibili ma non veritieri. In ambienti XR, questo rischio è amplificato dalla natura immersiva e dalla fiducia che gli utenti tendono a riporre negli agenti conversazionali, specialmente in scenari educativi o medici [29].
- **Sincronizzazione multimodale e latenza:** per garantire un’interazione fluida, è fondamentale che la generazione linguistica sia sincronizzata con i segnali visivi, gestuali e spaziali dell’ambiente. Tuttavia, le attuali limitazioni hardware e software nei visori XR possono introdurre latenze che compromettono la responsività e aumentano il rischio di *cybersickness*.
- **Opacità del modello e mancanza di accountability:** i LLM operano come sistemi a *scatola nera*, rendendo difficile tracciare il rationale delle loro risposte. Questa mancanza di trasparenza pone problemi etici e progettuali legati alla fiducia, alla responsabilità e alla verificabilità dell’interazione [29]. In contesti sensibili, come la salute o l’educazione, ciò può avere implicazioni critiche.

Affrontare tali criticità richiede una progettazione attenta, che integri pratiche di *prompt engineering*, meccanismi di spiegabilità (*explainability*) e strategie di fall-

back, oltre a un'adeguata valutazione empirica del comportamento del sistema in contesti immersivi e dinamici.

### 3.2.5 Pratiche progettuali

La progettazione di esperienze XR guidate da LLM richiede un approccio multidimensionale che coniungi sensibilità interattiva, robustezza tecnica e responsabilità etica. Diversi studi recenti hanno identificato una serie di buone pratiche necessarie per garantire l'efficacia e la sostenibilità di questi sistemi [29].

- **Prompting contestuale:** i LLM possono generare risposte molto variabili a seconda del prompt iniziale. Per ambienti XR, è fondamentale adottare strategie di *prompt engineering* orientate al contesto immersivo, ad esempio attraverso *metaprompt* che includono informazioni spaziali, obiettivi dell'utente, stato del sistema e vincoli comportamentali. In scenari dinamici, l'uso di prompt multi-turno con recupero del contesto (*augmented context retrieval*) ha mostrato di migliorare sensibilmente la coerenza delle risposte [29].
- **Interfacce multimodali:** l'interazione in XR si realizza tramite una combinazione di canali: voce, gesto, sguardo, postura, posizione nello spazio. La progettazione deve prevedere l'integrazione di questi segnali per costruire una comprensione condivisa tra utente e sistema. Tecnologie come eye tracking, sensori aptici, riconoscimento del parlato e motion capture possono essere orchestrate per fornire input ricchi al LLM, migliorando l'adattività e riducendo l'ambiguità interpretativa [29].
- **Metriche di valutazione ibride:** la valutazione dei sistemi LLM-enabled in XR non può limitarsi a metriche convenzionali (e.g., task completion, System Usability Scale). È necessario adottare misure che considerino anche aspetti emergenti come la proattività dell'agente, la qualità dell'integrazione multimodale, la coerenza semantica nel tempo e la consapevolezza del contesto. Alcune metriche proposte includono *contextualized spatial awareness*, *multimodal integration accuracy*, *coherence over time* e *hallucination rate* in ambienti 3D [29].

**Metriche HCI per la valutazione dell'interazione** La valutazione dei sistemi XR-LLM richiede l'integrazione di metriche tipiche dell'HCI per analizzare l'esperienza utente. Oltre a metriche tradizionali come la *System Usability Scale* (SUS) [35], vengono impiegati strumenti specifici per ambienti immersivi:

- **Presence Questionnaire (PQ)** [36]: misura il senso di “essere lì” percepito dall’utente, fondamentale per l’efficacia narrativa e il coinvolgimento.
- **NASA-TLX** [37]: valuta il carico cognitivo percepito (mental, physical, temporal, effort).
- **Speech Interface Usability Questionnaire (SIUQ)**: utile per analizzare la naturalezza del dialogo vocale in interfacce LLM-powered.

L’impiego combinato di queste metriche consente una valutazione quantitativa e qualitativa dell’interazione, supportando decisioni progettuali basate su evidenza empirica.

Accanto alle pratiche tecniche, si impone una riflessione sul ruolo dell’etica nell’adozione dei LLM in ambienti immersivi. L’assenza di trasparenza nei processi decisionali dei modelli, la possibilità di bias sistematici e la raccolta pervasiva di dati biometrici pongono sfide inedite alla responsabilità progettuale. Per questo motivo, viene proposta l’introduzione di una quinta dimensione della consapevolezza: la **consapevolezza etica** (*ethical awareness*) [29]. Questa dovrebbe guidare lo sviluppo di sistemi XR-LLM in modo da:

- garantire la *privacy* dell’utente e dei soggetti terzi, anche in ambienti condivisi o pubblici;
- mitigare i *bias* nei dati e nelle risposte del modello;
- offrire *spiegazioni interpretabili* delle decisioni del sistema;
- preservare l’*autonomia decisionale* dell’utente, evitando la sovra-dipendenza da agenti conversazionali immersivi.

Tali principi non sono solo vincoli progettuali, ma elementi centrali per costruire fiducia e promuovere l'adozione responsabile di agenti intelligenti in contesti XR.

### 3.3 Architetture integrate XR-LLM-Emotion Recognition

La convergenza tra Extended Reality (XR), Large Language Models (LLM) e sistemi di riconoscimento delle emozioni sta delineando nuove frontiere nell'ambito dell'Human-Computer Interaction. Questa integrazione mira a creare ambienti immersivi in grado di percepire, interpretare e rispondere in modo empatico agli stati emotivi degli utenti, migliorando l'efficacia e la naturalezza dell'interazione.

#### 3.3.1 Architetture multimodali e riconoscimento delle emozioni

Recenti studi hanno proposto architetture che combinano input multimodali (voce, espressioni facciali, segnali fisiologici) per il riconoscimento delle emozioni in tempo reale. Ad esempio, l'integrazione di dati EEG con modelli LLM ha dimostrato la capacità di classificare stati emotivi positivi, neutri e negativi, adattando dinamicamente l'ambiente XR alle esigenze dell'utente [38].

#### 3.3.2 Integrazione di LLM per l'adattamento empatico

L'uso di LLM consente una comprensione contestuale avanzata e la generazione di risposte linguistiche empatiche. Modelli come DialogueLLM sono stati ottimizzati per riconoscere e rispondere alle emozioni espresse in conversazioni multimodali, migliorando l'interazione in ambienti XR [9].

#### 3.3.3 Applicazioni in contesti industriali e sanitari

L'integrazione di XR, LLM e riconoscimento delle emozioni trova applicazione in contesti industriali e sanitari. Ad esempio, in ambienti ad alta intensità cognitiva come la manutenzione aerospaziale, sistemi XR-LLM possono monitorare lo stress dell'operatore tramite dispositivi indossabili e fornire supporto adattivo in

tempo reale [39]. In ambito sanitario, assistenti virtuali empatici possono migliorare l’esperienza del paziente, offrendo supporto emotivo personalizzato [40].

### 3.3.4 Sfide aperte e direzioni future

Nonostante i progressi recenti, l’integrazione fluida tra XR, LLM e riconoscimento delle emozioni presenta ancora diverse sfide sul piano dell’interazione naturale:

- **Sincronizzazione multimodale:** la gestione simultanea di input vocali, visivi e fisiologici richiede pipeline di elaborazione ottimizzate per garantire coerenza temporale e semantica. Ritardi o disallineamenti tra i canali possono compromettere la qualità dell’interazione e ridurre l’efficacia del riconoscimento emozionale.
- **Latenza e reattività:** l’elaborazione in tempo reale di dati multimodali e la generazione di risposte empatiche da parte dei LLM impongono requisiti stringenti in termini di latenza. Soluzioni come lo *streaming output* e l’uso di modelli leggeri su dispositivi locali sono esplorate per migliorare la reattività del sistema.
- **Adattamento dinamico al contesto:** la capacità del sistema di adattarsi alle variazioni del contesto e agli stati emotivi dell’utente in modo proattivo e coerente è ancora un obiettivo da raggiungere. L’integrazione di modelli predittivi e meccanismi di apprendimento continuo potrebbe offrire soluzioni promettenti.

Affrontare queste sfide richiede un approccio progettuale centrato sull’utente, che consideri non solo l’efficienza tecnica ma anche la qualità dell’esperienza interattiva. La ricerca futura dovrà concentrarsi sullo sviluppo di architetture più efficienti e sull’implementazione di strategie di adattamento in tempo reale per migliorare l’interazione tra l’utente e il sistema.

## 4 Sistema sviluppato

Questo capitolo descrive l'architettura e l'implementazione del sistema interattivo sviluppato. La prima parte si concentra sugli aspetti progettuali, illustrando la struttura concettuale e i principi architettonici adottati; la seconda descrive la realizzazione concreta del prototipo, incluse le tecnologie utilizzate e l'organizzazione dei moduli software.

### 4.1 Architettura e progettazione

Questa sezione descrive l'architettura del sistema sviluppato nell'ambito della presente tesi, il cui obiettivo è esplorare l'integrazione tra ambienti di *Extended Reality*, LLM e tecniche di riconoscimento affettivo per migliorare il contesto immersivo. Tali componenti concorrono alla costruzione di un'interfaccia conversazionale immersiva, adattiva ed empaticamente sensibile, in linea con gli obiettivi di ricerca.

L'architettura proposta è stata progettata per supportare l'interazione naturale tra utente e agente virtuale in uno spazio immersivo tridimensionale. Essa integra input multimodali (voce, contesto spaziale, segnali emotivi) e genera risposte contestualmente rilevanti, adattando il comportamento dell'agente in base allo stato emotivo percepito dell'utente.

La sezione si apre con una presentazione delle linee guida progettuali che hanno orientato le scelte architettoniche, prosegue con una descrizione della struttura generale del sistema e analizza nel dettaglio ciascuna delle sue componenti principali. Infine, vengono illustrati i principali flussi di interazione, attraverso scenari d'uso esemplificativi, per mostrare il funzionamento concreto del sistema nel contesto applicativo previsto.

#### 4.1.1 Linee guida progettuali

**Principi teorici e metodologici adottati** La progettazione del sistema si basa su un approccio *user-centered*, ispirato ai principi della *Human-Computer Interaction* (HCI), con particolare attenzione all'usabilità, alla naturalezza dell'interazione e alla coerenza multimodale. È stato adottato un paradigma progettuale modulare,

orientato alla separazione delle responsabilità funzionali tra i vari sottosistemi (interazione vocale, motore XR, riconoscimento affettivo, orchestrazione del contesto), al fine di garantire flessibilità, scalabilità e riuso del codice.

Dal punto di vista metodologico, la progettazione ha seguito una logica iterativa e incrementale, con fasi successive di definizione dei requisiti, modellazione architetturale, implementazione prototipale e validazione tramite casi d'uso. L'integrazione delle componenti si è basata sull'uso di interfacce standardizzate (API REST, protocolli locali) e sulla sincronizzazione esplicita degli eventi attraverso un modulo di orchestrazione centrale.

**Scenari d'uso rappresentativi** Per motivare la definizione dei requisiti architettonici, si presentano di seguito tre scenari d'uso che illustrano il comportamento atteso del sistema in contesti applicativi tipici.

**Scenario 1 (S1) – Accoglienza e interazione neutra** Un utente indossa il visore XR ed entra in un ambiente immersivo in cui è presente un agente virtuale. L'agente saluta l'utente, fornisce una breve introduzione all'ambiente e si rende disponibile per domande o assistenza. L'utente chiede informazioni sul funzionamento del sistema. L'agente ascolta, trascrive il parlato, genera una risposta coerente tramite LLM e la comunica all'utente.

*Implicazioni progettuali:* richiede pipeline ASR → LLM → TTS, gestione base del dialogo e visualizzazione dell'agente in ambiente XR.

**Scenario 2 (S2) – Interazione adattiva in funzione dello stato emotivo** Durante una sessione di interazione, l'utente pone domande con tono affaticato o frustrato. Il sistema rileva uno stato emotivo negativo attraverso l'analisi prosodica della voce. L'agente virtuale, pur mantenendo il contenuto informativo, modula il tono della voce e il comportamento visivo per comunicare maggiore empatia e disponibilità.

*Implicazioni progettuali:* richiede riconoscimento affettivo, adattamento dinamico del comportamento dell'agente e sincronizzazione multimediale.

### Scenario 3 (S3) – Interazione con feedback e gestione dell’errore

L’utente pronuncia una richiesta parzialmente incomprensibile o fuori contesto. Il sistema rileva un errore di interpretazione e lo comunica all’utente, offrendo opzioni per ripetere o chiarire il messaggio. L’utente utilizza un gesto per richiedere la ripetizione della risposta precedente.

*Implicazioni progettuali:* richiede robustezza dell’interazione, gestione del contesto conversazionale, supporto a comandi vocali e gestuali.

**Requisiti funzionali e non funzionali** I requisiti del sistema sono stati classificati in due categorie principali: *requisiti funzionali* (Functional Requirements, FR) e *requisiti non funzionali* (Non-Functional Requirements, NFR). A ciascun requisito è stato assegnato un identificatore univoco e un livello di priorità (**A** = Alta, **M** = Media, **B** = Bassa).

**Requisiti funzionali** I requisiti funzionali del sistema definiscono le principali capacità necessarie per realizzare l’interfaccia conversazionale immersiva proposta. Ogni requisito è associato a uno scenario d’uso specifico e classificato in base alla priorità. Questi requisiti riguardano principalmente la gestione dell’interazione vocale, la sincronizzazione dell’agente virtuale, e l’adattamento dinamico del comportamento in base agli input emotivi dell’utente.

Tabella 2: Requisiti funzionali del sistema, con riferimento agli scenari d’uso

| ID  | Descrizione  | Priorità | Scenario |
|-----|--|----------|----------|
| FR1 | Il sistema deve visualizzare un agente virtuale tridimensionale in un ambiente XR interattivo. | A        | S1       |
| FR2 | Il sistema deve acquisire il parlato dell’utente in tempo reale tramite microfono.             | A        | S1       |
| FR3 | Il sistema deve convertire l’audio dell’utente in testo tramite riconoscimento vocale (ASR).   | A        | S1       |

| ID   | Descrizione  | Priorità | Scenario |
|------|--|----------|----------|
| FR4  | Il sistema deve analizzare le caratteristiche prosodiche del parlato per stimare lo stato emotivo.   | M        | S2       |
| FR5  | Il sistema deve inviare il testo dell'utente e lo stato emotivo stimato a un modello LLM per generare una risposta coerente e affettivamente adattata. | A        | S2       |
| FR6  | Il sistema deve convertire la risposta testuale generata dal LLM in audio sintetizzato (TTS).  | A        | S1       |
| FR7  | Il sistema deve sincronizzare l'audio generato con l'animazione dell'agente virtuale.  | M        | S1       |
| FR8  | Il sistema deve adattare il comportamento visivo e vocale dell'agente in base allo stato emotivo rilevato.   | M        | S2       |
| FR9  | Il sistema deve mantenere un contesto conversazionale tra i turni di dialogo.  | M        | S2, S3   |
| FR10 | Il sistema deve consentire l'interruzione o la ripetizione del messaggio da parte dell'utente tramite comandi vocali o gestuali.                       | B        | S3       |

**Requisiti non funzionali** I requisiti non funzionali sono stati classificati secondo il modello ISO/IEC 25010, al fine di garantire un'analisi strutturata delle qualità attese dal sistema.

Tabella 3: Requisiti non funzionali del sistema, con riferimento agli scenari d'uso

| ID   | Descrizione  | Categoria  | Priorità | Scenario |
|------|--|------------|----------|----------|
| NFR1 | Il sistema deve garantire una latenza percepita nella risposta al parlato dell'utente inferiore a 3 secondi. | Efficienza | A        | S1       |

| ID   | Descrizione  | Categoria      | Priorità | Scenario |
|------|--|----------------|----------|----------|
| NFR2 | Il sistema deve mantenere almeno 60 FPS durante l'esecuzione su visori XR compatibili.               | Efficienza     | M        | S1       |
| NFR3 | Le componenti software devono essere modulari per consentire aggiornamenti indipendenti.             | Manutenibilità | A        | –        |
| NFR4 | Il sistema deve poter essere eseguito su visori XR consumer e piattaforme compatibili con Unity.     | Portabilità    | M        | S1       |
| NFR5 | I dati sensibili dell'utente non devono essere conservati né trasmessi a terze parti senza consenso. | Sicurezza      | A        | S2       |
| NFR6 | L'interfaccia utente deve essere facilmente comprensibile anche per utenti non esperti.              | Usabilità      | M        | S1       |
| NFR7 | Il sistema deve fornire indicazioni esplicative in caso di errore o mancata comprensione.            | Usabilità      | M        | S3       |
| NFR8 | Il sistema deve includere un modulo di monitoraggio delle prestazioni (latenza, uso risorse).        | Efficienza     | B        | –        |
| NFR9 | Il sistema deve supportare la registrazione dei log di sistema accessibili in fase di debugging.     | Manutenibilità | M        | –        |

#### 4.1.2 Architettura generale del sistema

L’architettura del sistema si basa su un modello modulare a componenti loosely coupled, distribuiti tra client e server. Le funzionalità principali sono suddivise in sei macro-componenti, ciascuna delle quali svolge un ruolo definito all’interno della pipeline interattiva. L’interazione è progettata per sfruttare flussi paralleli asincroni (testo vs audio), al fine di massimizzare la reattività e ridurre la latenza percepita.

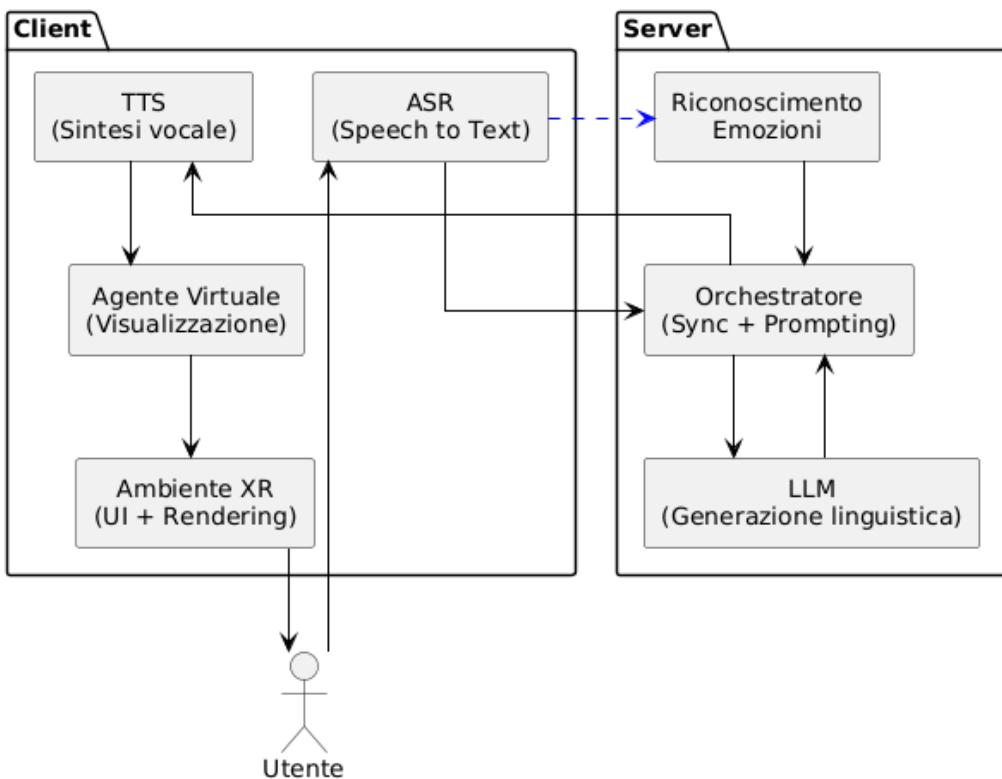


Figura 5: Schema ad alto livello dell’architettura del sistema

**1. Ambiente XR (Client)** Gestisce la scena immersiva tridimensionale e l’interfaccia utente. Visualizza l’agente virtuale, raccoglie input spaziali (gesture, posizione) e rende l’output vocale e visivo del sistema.

**2. Interfaccia vocale (Client)** Acquisisce l’audio dell’utente in tempo reale e lo trascrive in testo tramite un modulo ASR locale. In parallelo, invia il segnale audio grezzo al server per l’analisi emozionale.

**3. Emotion Recognition (Server)** Riceve l'audio asincrono e ne analizza le caratteristiche prosodiche per stimare lo stato emotivo. Il risultato è trasmesso all'orchestratore, se disponibile entro una soglia temporale predefinita.

**4. Dialog Orchestrator (Server)** Coordina il flusso tra i moduli. Riceve il testo dall'ASR e, se disponibile, l'emozione riconosciuta. Costruisce un prompt contestuale per l'LLM, tiene traccia dello stato conversazionale, ed è responsabile della consegna della risposta testuale finale al modulo di Speech Synthesis lato client.

**5. Large Language Model (Server)** Genera una risposta testuale coerente e adattiva sulla base del prompt ricevuto. Può produrre la risposta anche in assenza di dati emozionali, che vengono integrati se presenti in tempo utile.

**6. Speech Synthesis (Client)** Converte la risposta testuale in audio tramite un modulo TTS locale. Se presenti, integra parametri emozionali per modulare prosodia e tono. L'audio viene restituito all'ambiente XR per la riproduzione.

Questa architettura consente l'integrazione fluida di input multimodali, la generazione dinamica del contenuto e l'adattamento al contesto emozionale e dialogico dell'utente. La divisione client-server, unita alla progettazione asincrona, permette di soddisfare i requisiti di reattività, scalabilità e immersività previsti.

#### 4.1.3 Componenti principali

L'architettura del sistema è suddivisa in componenti funzionalmente distinti, ciascuno responsabile di una fase specifica del ciclo interattivo. Questa separazione riflette un approccio modulare che facilita l'analisi, la progettazione e l'evoluzione del sistema. Ogni componente è concepito per operare in modo indipendente, comunicando attraverso interfacce definite e asincrone, nel rispetto dei requisiti di reattività, adattività e coerenza multimodale. Le sezioni seguenti descrivono, da un punto di vista progettuale, le funzionalità attese e le responsabilità principali di ciascun modulo.

**Ambiente XR** Il componente “Ambiente XR” ha il compito di gestire l’interazione spaziale immersiva tra l’utente e il sistema, rappresentando il contesto fisico e semantico in cui si svolge la conversazione. Dal punto di vista progettuale, questa componente incarna il principio della presenza immersiva, garantendo un’esperienza coerente e credibile sotto il profilo visivo, spaziale e comportamentale.

L’ambiente deve offrire:

- una scena virtuale credibile, che rispetti le aspettative percettive dell’utente;
- un agente virtuale ancorato nello spazio e reattivo agli stimoli utente;
- un supporto per input multimodali non verbali (es. postura, sguardo, gesture);
- un canale di output coerente per espressioni vocali, visive e comportamentali.

Dal punto di vista dell’interazione, l’ambiente deve fungere da interfaccia conversazionale naturale, supportando il dialogo tramite:

- comunicazione vocale bidirezionale (input utente, output sistema);
- coerenza tra parlato sintetizzato e feedback visivo dell’agente;
- sincronizzazione tra contenuti linguistici e comportamento visivo (es. cambi di intensità, animazioni contestuali);
- feedback impliciti (prossimità, risposta al comportamento utente).

L’ambientazione XR non è neutra: essa contribuisce alla costruzione di senso e alla percezione di agency del sistema interattivo. Per questo motivo, è progettata per adattarsi in modo dinamico al contenuto del dialogo e allo stato emotivo rilevato, nel rispetto dei requisiti funzionali FR1, FR7, FR8 e non funzionali NFR2, NFR6.

**Rappresentazione dell’agente virtuale** In linea con un approccio orientato alla trasparenza e alla funzionalità dell’interazione, l’agente virtuale non è rappresentato in forma antropomorfa. Si è optato per una rappresentazione simbolica e dinamica, in cui la personalità dell’agente emerge attraverso il linguaggio, il tono e la reattività piuttosto che attraverso tratti umani realistici.

Questa scelta progettuale è motivata da diversi fattori:

- **Evitare l'effetto uncanny valley [41]:** la presenza di un volto umanoide imperfetto può compromettere l'accettazione dell'agente e generare dissonanza cognitiva. Numerosi studi hanno evidenziato che rappresentazioni virtuali quasi umane, ma non perfettamente realistiche, possono suscitare sensazioni di disagio e ridurre la fiducia degli utenti [42, 43].
- **Focalizzazione cognitiva:** la riduzione della complessità visiva favorisce la concentrazione dell'utente sui contenuti e sul comportamento.
- **Espressività astratta:** stati emotivi e intenzioni possono essere trasmessi tramite variazioni di luce, forma, movimento e colore, in modo meno vincolato a stereotipi.
- **Coerenza multimodale semplificata:** la sincronizzazione tra parlato, emozione e output visivo è più gestibile quando non vincolata all'animazione facciale realistica.

L'agente è quindi progettato come una presenza percettiva coerente ma non antropocentrica, che utilizza elementi grafici reattivi per comunicare empatia, intenzionalità e attenzione all'utente. Questo approccio supporta una rappresentazione fluida e adattabile dell'intelligenza artificiale.

**Interfaccia vocale** Il modulo di interfaccia vocale rappresenta l'ingresso principale dell'interazione utente-sistema e si colloca all'intersezione tra acquisizione audio, comprensione semantica e riconoscimento affettivo. La sua progettazione è stata orientata a garantire naturalezza, reattività e robustezza in ambienti immersivi, con particolare attenzione alla latenza percepita.

Dal punto di vista architetturale, l'ASR (Automatic Speech Recognition) è eseguito localmente sul client, in modo da fornire una trascrizione testuale in tempo quasi reale. Questo consente di avviare immediatamente la generazione linguistica da parte del modello LLM, riducendo i tempi di risposta. In parallelo, l'audio grezzo viene trasmesso asincronamente al server, dove viene analizzato dal modulo di Emotion Recognition. Questo flusso parallelo consente di stimare lo stato affettivo

dell’utente e, se disponibile in tempo utile, di adattare la risposta vocale in fase di sintesi (TTS) in termini di prosodia, tono e intensità.

La componente è progettata per supportare:

- **acquisizione audio** da microfono integrato in visore o dispositivo esterno, con normalizzazione e filtraggio;
- **trascrizione del parlato** tramite modello ASR robusto (locale), con gestione del turno conversazionale;
- **invio asincrono del segnale audio** al modulo server-side per l’analisi emozionale;
- **tagging temporale e segmentazione** del parlato per una possibile sincronizzazione multimodale;

La scelta di localizzare l’ASR consente di minimizzare la latenza (NFR1) e garantire fluidità dell’interazione in scenari immersivi (S1, S2). La trasmissione parallela dell’audio permette un arricchimento affettivo non vincolante ma complementare.

**Emotion Recognition** Il modulo di riconoscimento delle emozioni (Emotion Recognition) ha il compito di analizzare il segnale vocale dell’utente al fine di stimare il suo stato emotivo corrente, sulla base di caratteristiche prosodiche non verbali. La presenza di questo componente è cruciale per abilitare un’interazione empatica e adattiva, in cui l’agente virtuale può modulare il proprio comportamento in funzione del tono affettivo percepito.

Dal punto di vista architetturale, l’Emotion Recognition è collocato lato server e riceve in ingresso l’audio grezzo trasmesso asincronamente dal client. Questa scelta consente di impiegare modelli di analisi avanzati e computazionalmente intensivi, preservando al contempo la reattività del sistema grazie alla parallelizzazione con il flusso semantico.

L’elaborazione si basa sull’estrazione di feature acustiche e sull’impiego di un classificatore pre-addestrato, in grado di stimare classi discrete (es. gioia, rabbia,

tristezza). L'output del modulo è uno stato emozionale stimato, che viene inviato all'Orchestratore, dove può essere integrato nel prompt del LLM o influenzare la parametrizzazione del TTS.

Tra le responsabilità principali del modulo vi sono:

- **ricezione asincrona dell'audio** da parte del client;
- **estrazione automatica delle caratteristiche prosodiche** dal segnale vocale;
- **stima dello stato emotivo** tramite classificazione;
- **trasmissione non bloccante** del risultato al modulo di orchestrazione;
- **tolleranza alla latenza**, con integrazione solo se il risultato è disponibile entro soglie predefinite.

Questo approccio consente di arricchire la risposta generata dal sistema con elementi affettivi rilevanti, senza introdurre vincoli rigidi di sincronizzazione. In caso di indisponibilità dell'emozione in tempo utile, il sistema può comunque restituire una risposta neutra e coerente.

L'Emotion Recognition rappresenta quindi un canale affettivo indipendente ma integrabile, che contribuisce a rendere l'interazione più sensibile, situata e umanamente significativa, in linea con gli obiettivi espressi negli scenari S2 e S3.

**Dialog Orchestrator** Il Dialog Orchestrator rappresenta il nodo centrale di coordinamento tra i moduli del sistema. Ha il compito di aggregare e armonizzare le informazioni provenienti dai flussi paralleli (input testuale dal modulo ASR e stato emozionale dal modulo affettivo), gestire il contesto del dialogo, e costruire i prompt per l'interrogazione del modello LLM.

L'orchestratore opera secondo una logica event-driven asincrona. Alla ricezione del testo trascritto, può iniziare la generazione della risposta anche in assenza immediata del dato emozionale. Se l'informazione affettiva è disponibile entro una soglia temporale configurabile, viene integrata nel prompt per arricchire la risposta

del LLM in termini di tono, empatia o contenuto. In caso contrario, la conversazione procede con una modalità neutra ma coerente.

Le principali funzioni del modulo sono:

- **gestione del contesto dialogico**, mantenendo la memoria dei turni precedenti;
- **costruzione dinamica dei prompt** per l'LLM, integrando contenuti semantici e affettivi;
- **coordinamento asincrono** tra i tempi di risposta del LLM e dell'emotion recognition;
- **invio della risposta testuale** al modulo di sintesi vocale sul client per la riproduzione;
- **gestione delle interruzioni o fallback** in caso di incomprensioni o errori (S3).

Questo modulo costituisce il cuore logico del sistema, garantendo coerenza conversazionale, adattività affettiva e sincronizzazione tra le componenti, in conformità ai requisiti FR5, FR9 e NFR1.

**Large Language Model (LLM)** Il modulo LLM costituisce il motore generativo del sistema, responsabile della produzione della risposta linguistica dell'agente virtuale. Integra un modello linguistico di grandi dimensioni (Large Language Model - LLM) in grado di elaborare l'input utente, interpretarlo in modo contestuale e produrre una risposta testuale coerente, fluida e, se possibile, affettivamente modulata.

Il LLM riceve in ingresso un prompt costruito dinamicamente dal modulo orchestratore. Tale prompt contiene:

- la trascrizione del parlato utente (proveniente dall'ASR);
- eventuali informazioni affettive (es. emozione riconosciuta, livello di attivazione);

- il contesto dialogico precedente (stato della conversazione).

Grazie a questa architettura, la generazione linguistica può iniziare immediatamente sulla base del testo, mentre eventuali informazioni affettive vengono integrate se disponibili entro una finestra temporale utile. Il modello è configurato per supportare diverse strategie di prompting, tra cui la modulazione empatica, il rephrasing affettivo e la sintesi coerente in caso di errori o ambiguità (Scenario S3).

Le responsabilità principali del modulo includono:

- **comprendere contestuale** dell'input utente e della situazione dialogica;
- **adattamento stilistico ed emozionale** del contenuto, se richiesto;
- **resilienza agli errori** di trascrizione o ambiguità semantiche.

Il modulo LLM è eseguito lato server per sfruttare risorse computazionali adeguate e mantenere flessibilità nell'aggiornamento del modello. La sua integrazione è centrale nel garantire coerenza narrativa, adattività affettiva e fluidità linguistica, in linea con i requisiti FR5, FR9 e con i principi di interaction design conversazionale.

**Speech Synthesis (TTS)** Il modulo di sintesi vocale (Text-to-Speech, TTS) ha il compito di convertire la risposta testuale generata dal modello LLM in un segnale audio naturale e comprensibile, da restituire all'utente tramite l'agente virtuale. Nel contesto di un'interazione immersiva in XR, la qualità e la coerenza temporale dell'output vocale rivestono un ruolo centrale nella costruzione di presenza e credibilità.

Per ridurre la latenza e migliorare la sincronizzazione con il comportamento visivo dell'agente, la sintesi vocale è eseguita direttamente lato client. Questa scelta consente una risposta immediata, evita ritardi dovuti alla rete e permette di modulare in tempo reale la prosodia dell'output in funzione dello stato emotivo stimato, se disponibile.

Il modulo TTS riceve in ingresso la risposta testuale e, optionalmente, parametri affettivi (es. tono, intensità, velocità) forniti dal modulo orchestratore. Tali parametri vengono utilizzati per adattare l'intonazione della voce sintetica e rafforzare l'effetto empatico percepito.

Le funzionalità principali del modulo includono:

- **conversione del testo in audio**, tramite motore TTS neurale o embedded;
- **parametrizzazione dinamica** della voce sintetica (prosodia, timbro, velocità);
- **sincronizzazione con il comportamento dell'agente** nell'ambiente XR;
- **restituzione del segnale audio** all'ambiente XR per la riproduzione immersiva;
- **fallback neutro** in caso di mancanza di informazioni emotive.

L'esecuzione del TTS lato client è compatibile con architetture moderne per ambienti XR standalone, riduce il carico sul server e migliora l'esperienza percepita. Tale approccio è in linea con i requisiti NFR1 (latenza), FR6 (sintesi) e FR7 (sincronizzazione multimodale), e contribuisce in modo sostanziale alla naturalezza dell'interazione.

Tabella 4: Mappatura tra moduli architetturali e requisiti soddisfatti

| Modulo                   | Requisiti funzionali (FR) | Requisiti non funzionali (NFR) |
|--------------------------|---------------------------|--------------------------------|
| Ambiente XR              | FR1, FR7, FR8             | NFR2, NFR4, NFR6               |
| Interfaccia vocale (ASR) | FR2, FR3                  | NFR1, NFR3, NFR6               |
| Emotion Recognition      | FR4, FR5, FR8             | NFR5, NFR1                     |
| Dialog Orchestrator      | FR5, FR9, FR10            | NFR1, NFR3, NFR7, NFR9         |
| Language Model (LLM)     | FR5, FR9                  | NFR1, NFR3                     |
| Speech Synthesis (TTS)   | FR6, FR7, FR8             | NFR1, NFR2, NFR6               |

#### 4.1.4 Flusso di interazione

Il funzionamento del sistema si articola in un flusso di interazione distribuito tra client e server, progettato per minimizzare la latenza e garantire una risposta

adattiva. Il flusso si basa su un'elaborazione parallela asincrona, in cui il testo e l'audio dell'utente vengono trattati in pipeline distinte e riconciliate in fase di generazione della risposta.

1. L'utente parla attraverso il microfono del visore XR.
2. Il modulo ASR locale (client) trascrive in tempo reale il parlato in testo.
3. In parallelo, l'audio grezzo viene inviato al server per l'analisi emozionale.
4. Il testo viene trasmesso all'Orchestratore, che costruisce un prompt per il LLM.
5. Il LLM genera la risposta testuale.
6. Se disponibile in tempo utile, lo stato emotivo stimato viene integrato nel prompt o usato per modulare la voce sintetica.
7. La risposta testuale viene inviata al modulo TTS locale (client), che la converte in audio.
8. L'audio sintetico viene riprodotto dall'agente virtuale all'interno dell'ambiente XR, sincronizzato con animazioni visive.

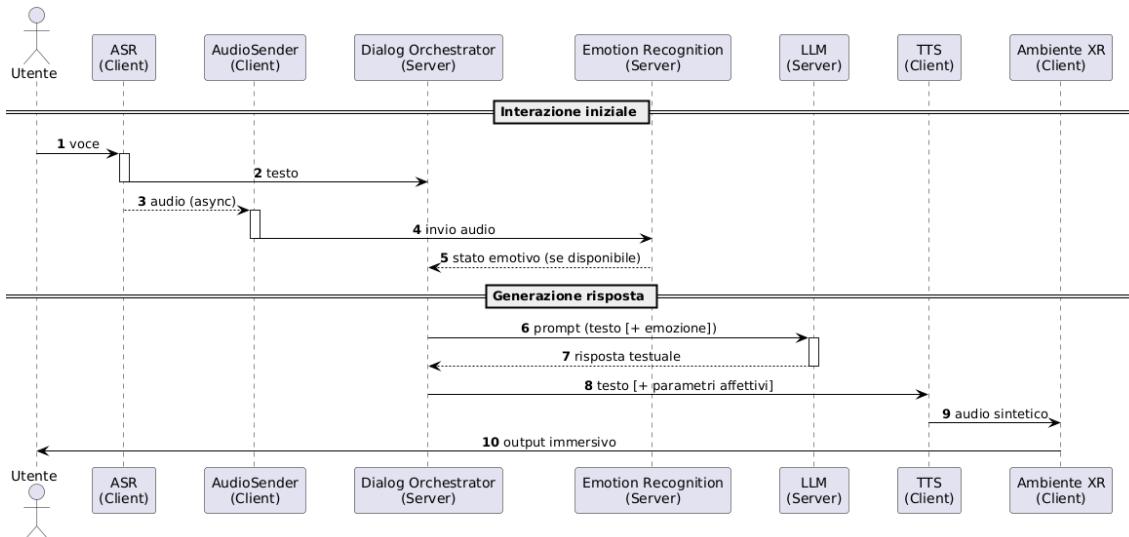


Figura 6: Flusso di interazione tra utente e sistema

Questo schema di interazione riflette una visione architetturale centrata sulla reattività e sull’adattamento in tempo reale, requisiti fondamentali in contesti immersivi. La separazione tra flusso semantico (testo) ed emozionale (audio) consente di iniziare la generazione della risposta non appena il contenuto linguistico è disponibile, riducendo i tempi di attesa percepiti. Allo stesso tempo, l’integrazione opportunistica dell’informazione affettiva consente una modulazione più empatica del comportamento dell’agente senza introdurre blocchi nel flusso.

Le interazioni tra componenti sono progettate secondo un paradigma asincrono basato su eventi, in cui ciascun modulo può proseguire la propria elaborazione indipendentemente dalla disponibilità degli altri. Il modulo orchestratore funge da punto di raccolta e decisione, decidendo se attendere l’analisi emozionale o procedere con una risposta neutra. Questo approccio ”non bloccante” migliora la tolleranza alle latenze e consente un bilanciamento dinamico tra qualità e velocità di risposta.

Dal punto di vista dell’utente, questo flusso si traduce in un’interazione fluida, in cui la risposta dell’agente avviene con minima latenza, ma risulta al tempo stesso personalizzata e contestualmente sensibile. L’integrazione progressiva delle informazioni (prima semantiche, poi affettive) è trasparente all’utente e rafforza la sensazione di dialogo naturale e coinvolgente, elemento chiave per esperienze XR empatiche.

Nel complesso, il flusso operativo riflette i principi di architettura conversazionale immersiva: parallelismo dei canali, tolleranza alla latenza e sincronizzazione flessibile. L’agente virtuale può così rispondere con rapidità e coerenza, arricchendo la propria espressività ogni volta che l’informazione emozionale è disponibile in tempo utile. L’esperienza risultante è percepita come empatica e coerente con gli obiettivi di progettazione centrati sull’utente in ambienti immersivi.

#### 4.1.5 Considerazioni

L’architettura proposta integra tecnologie eterogenee in una struttura modulare, distribuita e asincrona, pensata per massimizzare la reattività e la naturalezza dell’interazione. La separazione funzionale tra client e server consente di sfruttare in

modo ottimale le risorse computazionali disponibili, riducendo la latenza percepita dall’utente e migliorando la qualità dell’esperienza immersiva.

Dal punto di vista del design conversazionale, l’uso di un orchestratore centrale permette di mantenere coerenza nel flusso dialogico, gestire in modo flessibile i turni di parola e adattare dinamicamente le risposte in funzione del contesto emotivo e semantico. L’integrazione asincrona di contenuti affettivi arricchisce il comportamento dell’agente senza compromettere la fluidità della conversazione.

L’impiego di un agente virtuale simbolico e non antropomorfo consente infine di evitare problematiche legate alla *uncanny valley*, mantenendo un forte potenziale espressivo attraverso l’uso coerente di voce, animazioni e stimoli visivi. Nel complesso, l’architettura definita supporta un’interazione in linea con i principi della Human-Computer Interaction e con gli obiettivi di ricerca della tesi.

## 4.2 Implementazione del sistema

### 4.2.1 Panoramica dell’implementazione

**Motivazioni del caso di studio e scelta della MR.** La sintesi del *2,4,6-trinitroresorcinolo* è un esempio didattico ideale perché unisce passaggi tipici della chimica organica (pesata, dissoluzione, aggiunta di acidi forti, controllo termico) ma presenta rischi elevati se condotta realmente. Riprodurla in mixed-reality consente di:

1. eliminare pericoli (sostanze tossiche, reazioni esotermiche);
2. ridurre i costi di materiali di consumo e smaltimento;
3. offrire un training ripetibile ovunque, con il solo requisito di un tavolo reale.

Dal punto di vista HCI, il contesto MR permette di mantenere il riferimento spaziale con l’ambiente reale, favorendo memoria procedurale e trasferimento delle competenze al laboratorio fisico.

**Architettura e tecnologie** Il sistema è strutturato secondo un’architettura client-server. Il **lato server**, implementato in Python tramite il framework Flask, ospita i servizi principali di elaborazione: analisi delle emozioni, orchestrazione del dialogo,

generazione linguistica tramite LLM. Ogni modulo è realizzato come componente indipendente. Il **lato client**, sviluppato in Unity per visori XR standalone (Meta Quest 3), gestisce la scena immersiva, l’agente virtuale e i canali di interazione multimodale (voce, gesture, spazio). Il client è responsabile della rappresentazione visiva e comportamentale dell’agente, della raccolta degli input utente (hand gesture e ASR) e della riproduzione dell’output vocale sintetizzato. Particolare attenzione è stata posta alla qualità dell’esperienza immersiva, alla coerenza percettiva e alla fluidità del dialogo. L’integrazione tra client e server è progettata per essere asincrona e non bloccante, in modo da minimizzare la latenza percepita e permettere una risposta fluida e adattiva da parte del sistema. Il flusso conversazionale è organizzato in pipeline parallele per l’elaborazione semantica (testo) e affettiva (audio), che vengono armonizzate dal modulo di orchestrazione in funzione della disponibilità temporale delle informazioni.

**Riusabilità di backend e agente tramite CONTEXT.** L’implementazione lato server è indipendente rispetto allo scenario XR: tutti i micro-servizi operano su messaggi JSON indipendenti dal dominio e possono quindi servire qualsiasi applicazione immersiva (laboratorio di fisica, officina meccanica, visita museale, *etc.*). Specularmente, il globo-agente lato client non contiene conoscenza del task: al momento dell’inizializzazione il client invia al server una stringa CONTEXT che descrive strumenti, obiettivi e vincoli dell’esperienza corrente. Questa variabile viene prepended al primo prompt utente e resta nel buffer di sistema dell’LLM per l’intera sessione, rendendo l’agente immediatamente ri-utilizzabile in nuovi ambienti semplicemente cambiando il valore di CONTEXT, senza modificare codice né retrain del modello.

#### 4.2.2 Lato Server

**Architettura dei micro-servizi e orchestrazione.** Il backend è organizzato come un insieme di micro-servizi Flask indipendenti, ognuno dedicato a una funzione ben definita (*speech → text*, riconoscimento affettivo, orchestrazione dialogica, inferenza LLM, logging). Questa suddivisione permette modularità: ogni componente

può essere sostituito o aggiornato con basso impatto sugli altri. Il coordinamento logico è affidato alla classe `Orchestrator` (file `components/orchestrator.py`), che arricchisce il prompt con le emozioni rilevate, invoca il modello linguistico e restituisce la risposta all’interfaccia cliente. Il flusso è *event-driven* e non bloccante, così da ridurre la latenza percepita.

**Endpoint di comunicazione e protocollo.** I servizi espongono API REST protette da HTTPS. Gli endpoint principali sono:

- `/upload_audio`: riceve chunk audio *multipart/form-data* e attiva la pipeline emotiva;
- `/chat_message`: accetta JSON con `{user_id, text}` e restituisce la risposta dell’agente;
- `/reset_conversation`: azzerà lo stato di una sessione.

Il payload vocale viaggia in PCM 16 kHz; i messaggi testuali seguono uno schema JSON uniforme (`role`, `content`, `timestamp`, ecc.), facilitando la tracciabilità delle interazioni.

**Pipeline conversazionale (ASR → LLM → TTS).** La conversazione procede in tre fasi asincrone:

1. l’ASR locale sul client trasmette il testo a `/chat_message`;
2. l’Orchestrator richiama l’LLM (`ChatAgent` o `OllamaChatAgent`) e riceve il testo di risposta *token-by-token*;
3. il client richiede al servizio TTS la sintesi audio, che viene riprodotta in streaming per evitare pause percepibili.

In media il round-trip tra utterance utente e inizio della risposta audio è inferiore a 1 s su rete locale, soddisfacendo il requisito di fluidità HCI.

**Modulo di riconoscimento di emozioni.** Il micro-servizio `EmotionRecognizer` utilizza un modello Hugging Face fine-tuned per classificare il segnale vocale in nove categorie discrete (gioia, tristezza, rabbia, ecc.). L'audio viene prima accumulato da `AudioAccumulator` fino a superare una soglia di 25 s: ciò permette inferenze batch che riducono il carico GPU senza interrompere il dialogo. L'output è una mappa `{label: probabilità}` salvata in `EmotionMemory` con TTL configurabile; l'Orchestrator la integra nel prompt solo se “fresca” (<90 s) per non inquinare la risposta con emozioni obsolete.

**Gestione del contesto e memorizzazione.** La persistenza leggera dello storico dialogico è gestita da `ConversationManager`, che mantiene in-memory i turni correnti e li scarica su file JSON (`conversations/user_id.json`) alla fine di ogni scambio. Questi log includono metadati (parole, latenza, emozioni, reset) utili alle analisi post-sessione. Tutta la memorizzazione è *opt-in*: per rispettare i principi di privacy, i dati vengono conservati solo nelle sessioni sperimentali in cui l'utente ha fornito consenso informato.

#### 4.2.3 Lato Client

**Contesto applicativo e obiettivo del task.** L'applicazione è un prototipo didattico di mixed-reality volto a guidare lo studente nella sintesi del *2,4,6-trinitroresorcinolo*. Indossando Meta Quest 3, l'utente deve eseguire – in sicurezza e con l'aiuto di un assistente virtuale – la seguente sequenza operativa: *(i)* posizionare il pallone sul supporto; *(ii)* pesare 500 mg di resorcinolo; *(iii)* trasferire la polvere nel pallone; *(iv)* misurare 50 mL di etanolo e versarlo; *(v)* portare la piastra riscaldante a 60 °C; *(vi)* introdurre acido solforico nel becher B2; *(vii)* aggiungere lentamente acido nitrico; *(viii)* versare la miscela nitrante nel pallone; *(ix)* osservare il viraggio di colore finale. Lo stato di avanzamento non è mostrato in UI: l'agente deduce il contesto unicamente dalle parole dell'utente, per preservare naturalezza conversazionale.

**Scena immersiva e ambientazione XR.** L’esperienza si svolge in passthrough: il mondo reale rimane visibile e viene arricchito da un banco virtuale delimitato da pannelli semi-trasparenti (*workspace walls*). Il *tavolo di lavoro* prefab – una superficie URP opaca blu-grigio – è istanziato in prossimità di un vero tavolo rilevato dall’utente; ciò consente di utilizzare l’app in qualsiasi aula o laboratorio senza setup fisico dedicato. L’altezza del piano virtuale è fissata a  $\approx 1,1$  m per risultare ergonomica sia in piedi sia da seduti.



(a) Vista panoramica del banco di lavoro.



(b) Dettaglio degli strumenti virtuali.



(c) Interazione con il cilindro graduato.

Figura 7: Screenshot dell’ambiente MR nelle diverse fasi di preparazione del task di sintesi chimica.

**Oggetti virtuali e UI.** La vetreria proviene dal pacchetto *3D Laboratory Environment with Apparatus* ed è stata aggiornata con lo shader proprietario “Liquid-Shader” che anima il livello dei fluidi. Gli strumenti disponibili sono: pallone a fondo tondo, becher 250 mL, cilindro graduato 100 mL, piastra riscaldante, supporto con pinza, cucchiaio da laboratorio e vaschette reagenti B1/B2. Ogni contenitore include un *World-Space Canvas* che visualizza in tempo reale la quantità (es. «97/250 mL»); la tipografia e lo stile seguono le linee guida di Meta per garantire leggibilità su sfondo reale. L’interazione fisica è abilitata tramite *Meta Interaction SDK – Grab*

*Interactable.*

**Agente virtuale: rappresentazione e comportamento.** L’assistente è un **globo fluttuante** ( $\approx 18$  cm) generato proceduralmente; un leggero moto di rotazione comunica presenza anche a riposo. Il colore è sempre arancione, ma varia di *luminanza* per indicare lo stato: *arancione chiaro* (idle)  $\rightarrow$  *arancione vivo* (listening)  $\rightarrow$  *arancione intenso* (speaking). Se il riconoscimento affettivo rileva emozioni negative, la pulsazione dello shader accelera e la tonalità si scurisce di alcune unità HSL, offrendo un segnale empatico non verbale. Il globo entra in modalità **listening** quando l’utente effettua una precisa hand gesture; entra in **speaking** quando riceve la risposta dal server e torna **idle** dopo la risposta.



Figura 8: Variazioni visive del globo-agente nei tre stati di interazione.

### Canali di input multimodale.

- **Voce – Meta Voice SDK + Wit.ai.** L’SDK cattura audio a 16 kHz e lo inoltra all’endpoint Wit.ai; dal JSON di risposta vengono estratti transcript, intent e slots, che alimentano l’orchestratore server-side.
- **Gestore – Meta Hand Tracking.** Il *Hand Gesture Recognizer* intercetta open-hand (attiva l’agente), pinch (grab/release oggetti) e point. I raggi di selezione appaiono solo quando necessario, riducendo il disordine visivo.

La combinazione di passthrough, vetreria interattiva e guida vocale/gestuale consente all’utente di completare la sintesi senza mai uscire dal contesto fisico, soddisfacendo i requisiti di immediatezza, bassa latenza cognitiva e alta naturalità richiesti dallo studio HCI.

#### 4.2.4 Integrazione Client–Server

L'integrazione tra client XR e backend Flask è progettata per garantire un'interazione fluida, adattiva e a bassa latenza. Il sistema sfrutta una comunicazione asincrona e modulare, in cui ogni componente può operare e rispondere indipendentemente dagli altri, favorendo la resilienza e l'estensibilità dell'intero stack.

**Protocolli e canali di comunicazione.** I due lati del sistema comunicano principalmente tramite API REST, protetti da protocollo HTTPS. Il canale REST viene utilizzato per le chiamate puntuali (`/chat_message`, `/upload_audio`),

La progettazione degli endpoint favorisce la disaccoppiamento semantico: ogni messaggio JSON è autocontenuto e include metadati sufficienti a essere processato indipendentemente dal contesto ambientale. Questo consente di mantenere la generalità del backend e riutilizzarlo in ambienti diversi, come discusso nel paragrafo sulla variabile CONTEXT.

**Gestione asincrona e tolleranza alla latenza.** L'interazione è orchestrata secondo un paradigma *event-driven*, in cui il client invia richieste senza attendere risposte bloccanti:

- L'ASR locale sul visore invia immediatamente il testo trascritto all'orchestratore;
- In parallelo, l'audio grezzo viene trasmesso al modulo Emotion Recognition;
- Il server restituisce la risposta testuale appena disponibile, senza attendere obbligatoriamente l'analisi affettiva;
- Se il dato emozionale arriva in tempo utile, viene integrato nel prompt o nella modulazione vocale; altrimenti, si procede con una risposta neutra.

Questa strategia riduce i tempi di attesa percepiti e consente all'agente virtuale di rispondere rapidamente, anche in condizioni di latenza di rete o carico elevato.

**Scalabilità e manutenibilità.** L’interfaccia tra client e server è definita in modo esplicito e stabile, con contratti di comunicazione documentati (es. schema JSON, status code, timeouts). Questo facilita:

- l’aggiornamento indipendente dei moduli (es. sostituzione del motore TTS o del classificatore affettivo);
- il porting del client su piattaforme diverse (es. PC VR o AR mobile);
- il deployment del server su ambienti cloud o edge.

L’architettura di integrazione è pensata per garantire compatibilità tra reattività locale (ASR, TTS) e potenza elaborativa remota (LLM, emotion recognition), massimizzando la qualità dell’interazione immersiva. Questo equilibrio tra efficienza, modularità e adattività è fondamentale per supportare esperienze XR naturali e riusabili in contesti applicativi diversi.

## 5 Studio sperimentale

Lo scopo di questo capitolo è presentare in modo sistematico il disegno dello studio condotto per valutare l'impatto dell'adattamento emozionale di un agente vocale in realtà mista sull'esperienza utente. L'esperimento si inserisce nel contesto dello scenario sviluppato, e rappresenta una fase cruciale per validare empiricamente le ipotesi formulate nella fase progettuale.

In particolare, il focus è posto sulla relazione tra il comportamento adattivo dell'agente, basato sull'analisi della prosodia vocale dell'utente, e una serie di variabili psicologiche e comportamentali rilevanti: senso di presenza, carico cognitivo, usabilità del dialogo e performance oggettive nel task.

### 5.1 Obiettivi

L'obiettivo generale dello studio è indagare se e come l'integrazione di un adattamento emozionale nella voce e nei contenuti dell'agente vocale possa influenzare l'interazione in realtà mista, rendendola più coinvolgente, fluida ed efficace.

Per affrontare questo quesito, sono stati definiti quattro obiettivi specifici, che guidano la raccolta e l'analisi dei dati:

- O1** Valutare se un agente con adattamento affettivo è in grado di incrementare il senso di presenza percepito dall'utente, misurato attraverso il *Presence Questionnaire* (PQ) [36].
- O2** Verificare se l'adattamento emozionale riduce il carico cognitivo e il disagio fisico percepiti dagli utenti durante l'esecuzione del compito, tramite il questionario *NASA-TLX* [37] e il *Simulator Sickness Questionnaire* (SSQ) [44].
- O3** Analizzare in che misura la presenza di un comportamento empatico nell'agente migliori la percezione di usabilità del sistema e dell'agente, tramite i questionari *System Usability Scale* (SUS) [35] e *Subjective Assessment of Speech System Interfaces* (SASSI) [45].

**O4** Esaminare l'impatto dell'adattamento su metriche di performance oggettive, quali la durata dell'interazione o il numero di turni dialogici.

Questi obiettivi riflettono un approccio multidimensionale, che integra valutazioni soggettive e misure oggettive per comprendere in profondità l'effetto del comportamento affettivo dell'agente sull'esperienza utente complessiva. Inoltre, viene incluso il SSQ per monitorare eventuali effetti collaterali sul comfort fisico (cyber-sickness), al fine di escludere che le differenze tra condizioni siano attribuibili a malesseri indotti dalla realtà mista.

## 5.2 Domande di ricerca

A partire dagli obiettivi delineati, sono state formulate quattro domande di ricerca che guidano l'intero impianto sperimentale. Ciascuna di esse mira a verificare, in modo specifico, l'effetto della variabile indipendente — ovvero la presenza o meno dell'adattamento affettivo — su aspetti distinti dell'esperienza utente in realtà mista.

**RQ1** *L'agente adattivo (EMO) è in grado di aumentare il senso di presenza percepita rispetto a un agente neutro (NEU)?*

Questa domanda indaga se la responsività affettiva dell'agente favorisce una maggiore immersione e senso di “esserci” nello spazio virtuale condiviso.

**RQ2** *L'agente EMO contribuisce a ridurre il carico cognitivo e il disagio fisico percepito durante l'esecuzione del compito?*

L'ipotesi sottostante è che un comportamento più empatico e contestuale possa rendere l'interazione meno faticosa e più confortevole, sia dal punto di vista mentale che fisico.

**RQ3** *L'interazione con l'agente EMO è percepita come più usabile e affidabile rispetto alla controparte neutra?*

Qui si esplora se l'adattamento emozionale migliora la qualità dell'interfaccia dialogica e aumenta la fiducia dell'utente verso il sistema.

**RQ4** *La presenza di adattamento affettivo migliora le prestazioni oggettive nel task?*

Questa domanda collega il comportamento dell’agente a misure quantitative di efficacia e fluidità dell’interazione.

Tali domande rappresentano il nucleo empirico della ricerca e sono coerenti con un approccio sperimentale controllato, volto a isolare l’effetto specifico del comportamento affettivo dell’agente vocale sul piano sia soggettivo sia prestazionale.

**Ipotesi** Per ciascuna domanda di ricerca (RQ) sono state formulate le seguenti ipotesi statistiche:

Tabella 5: Ipotesi nulle e alternative per ciascuna RQ.

| <b>RQ</b> | <b>H<sub>0</sub> (Ipotesi nulla)</b>                           | <b>H<sub>1</sub> (Ipotesi alternativa)</b>                        |
|-----------|--|---|
| RQ1       | Nessuna differenza tra EMO e NEU nella presenza percepita (PQ) | Esiste una differenza tra EMO e NEU nella presenza percepita (PQ) |
| RQ2       | Nessuna differenza nei punteggi NASA-TLX e SSQ                 | Esiste una differenza nei punteggi NASA-TLX e/o SSQ               |
| RQ3       | Nessuna differenza nei punteggi SUS e SASSI                    | Esiste una differenza nei punteggi SUS e/o SASSI                  |
| RQ4       | Nessuna differenza in tempo e turni (TCT, Turns)               | Esiste una differenza in tempo e/o turni (TCT, Turns)             |

### 5.3 Condizioni sperimentali

Lo studio prevede un confronto tra due condizioni sperimentali, costruite in modo da differire esclusivamente per la presenza o meno dell’adattamento affettivo dell’agente vocale. Questa scelta consente di isolare l’effetto della variabile di interesse senza introdurre confondenti legati all’ambiente o al compito.

**Condizione EMO (emozionale):** In questa modalità, l’agente è progettato per rilevare lo stato affettivo dell’utente analizzando la prosodia della voce. Le

risposte dell’agente sono adattate dinamicamente, sia a livello contenutistico sia prosodico, in funzione delle emozioni stimate. L’obiettivo è simulare un comportamento empatico che si adatti allo stato d’animo percepito dell’interlocutore, rendendo l’interazione più naturale e supportiva.

**Condizione NEU (neutrale):** L’agente mantiene uno stile comunicativo costante, privo di qualsiasi forma di adattamento emotivo. Le risposte vocali sono generate con intonazione neutra e senza modulazioni affettive, indipendentemente dal tono o dalla condizione emotiva dell’utente.

Al di fuori del comportamento dell’agente, tutte le altre componenti dell’esperienza in realtà mista — inclusi l’interfaccia, il task assegnato, la durata dell’interazione e la scena virtuale — sono perfettamente equivalenti tra i due gruppi sperimentali. Questo controllo rigoroso garantisce che eventuali differenze osservate nelle misure dipendenti possano essere attribuite con maggiore confidenza alla sola manipolazione del comportamento affettivo dell’agente.

## 5.4 Disegno sperimentale

Il disegno sperimentale adottato è di tipo *between-subjects*, con un singolo fattore indipendente denominato **AFFECT**, che assume due livelli: **EMO** (agente adattivo) e **NEU** (agente neutro). Ciascun partecipante è esposto a una sola condizione, così da evitare effetti di apprendimento o confronti intra-individuali che potrebbero contaminare le misure.

L’assegnazione dei soggetti alle due condizioni avviene in modo randomizzato bloccato, tenendo conto del genere per garantire un bilanciamento equo tra i gruppi e ridurre possibili bias demografici.

Le variabili dipendenti sono suddivise in due categorie principali:

**Misure soggettive** raccolte attraverso questionari standardizzati:

- **Presenza percepita**, valutata tramite il *Presence Questionnaire* (PQ), che misura la sensazione di “essere lì” nello spazio virtuale.

- **Carico cognitivo**, stimato mediante il *NASA-TLX*, che analizza sei dimensioni dello sforzo mentale richiesto.
- **Disagio fisico (cybersickness)**, misurato attraverso il *Simulator Sickness Questionnaire (SSQ)*, somministrato sia prima dell'inizio dell'esperienza sia immediatamente dopo. L'obiettivo è rilevare eventuali sintomi di malessere fisico legati all'esperienza MR, calcolando la variazione complessiva tra pre e post-interazione.
- **Usabilità del sistema**, rilevata attraverso la *System Usability Scale (SUS)*, che fornisce un indice sintetico dell'esperienza d'uso.
- **Usabilità dell'agente**, misurata con il *Subjective Assessment of Speech System Interfaces (SASSI)*, composto da 18 item.

**Misure oggettive** raccolte automaticamente dai log del sistema durante l'interazione:

- **TCT** (*Task Completion Time*): tempo totale impiegato per completare il compito, calcolato come differenza tra il primo e l'ultimo `timestamp`.
- **Turns**: numero complessivo di turni di dialogo (`turn_id`) tra utente e agente.

Questo approccio consente di integrare percezioni soggettive e indicatori comportamentali, offrendo una visione completa dell'efficacia dell'agente adattivo in un contesto operativo realistico.

## 5.5 Partecipanti

Il campione sperimentale è composto da un totale di **20 partecipanti**, suddivisi equamente tra le due condizioni previste: 10 nel gruppo **EMO** e 10 nel gruppo **NEU**.

I partecipanti sono studenti iscritti a corsi di laurea dell'area STEM (*Science, Technology, Engineering and Mathematics*), con un'età compresa tra i 20 e i 30 anni. Per garantire omogeneità nel background, sono stati selezionati soggetti senza esperienza pregressa in contesti di realtà mista applicata alla chimica.

Sono stati inoltre definiti criteri di esclusione per tutelare la salute e la qualità dei dati raccolti. In particolare, non sono stati ammessi partecipanti con disturbi vestibolari noti o condizioni che potessero compromettere la tollerabilità dell'esperienza MR. La verifica dei criteri di esclusione è avvenuta sia tramite due domande specifiche incluse nel questionario preliminare, sia mediante un colloquio verbale iniziale, durante il quale venivano esplicitamente richieste eventuali condizioni mediche rilevanti. Per rafforzare questo controllo, il *Simulator Sickness Questionnaire* (SSQ) è stato utilizzato anche in fase di baseline per individuare eventuali sintomi pre-esistenti e garantire la sicurezza dei partecipanti durante l'interazione.

Questa selezione consente di controllare importanti variabili individuali che potrebbero interferire con la percezione soggettiva o con le prestazioni durante il compito in realtà mista.

## 5.6 Procedura

Ogni partecipante prende parte a una singola sessione sperimentale della durata complessiva di circa 35 minuti, articolata in più fasi pensate per introdurre gradualmente l'utente all'esperienza, raccogliere i dati necessari e ridurre al minimo possibili bias dovuti alla novità della tecnologia.

La procedura si apre con un **briefing iniziale**, durante il quale viene presentato il contesto dello studio e viene raccolto il consenso informato. In questa fase viene anche somministrato un questionario demografico (età, genere, livello di istruzione, esperienza con la realtà mista ed esperienza con agenti vocali o interfacce conversazionali). Il questionario demografico è strutturato con risposte a scelta multipla o con scala Likert, dove i partecipanti possono selezionare l'opzione che meglio descrive la loro situazione. Questo questionario serve per caratterizzare il campione e garantire che il gruppo sperimentale sia omogeneo. Dopo il questionario demografico, viene somministrata la prima compilazione del *Simulator Sickness Questionnaire* (SSQ) per valutare eventuali sintomi di discomfort pre-esistenti (baseline). Questo passaggio è fondamentale per controllare l'insorgenza di cybersickness legata all'esperienza MR.

Segue una fase di **addestramento**, della durata di circa 3 minuti, in cui il partecipante prende confidenza con i comandi vocali e gestuali previsti dal sistema, senza accedere al task reale. Questa fase serve a uniformare il livello di familiarità con l’interfaccia tra i soggetti. L’obiettivo è di far prendere confidenza al partecipante con il sistema e ridurre la possibilità di errori durante l’interazione effettiva con l’agente vocale. I partecipanti devono eseguire le seguenti azioni durante il training:

- Uso dei comandi vocali: Il partecipante deve seguire le istruzioni per interagire con l’agente vocale, ad esempio, chiedendo informazioni o eseguendo semplici comandi.
- Interazione gestuale: Si insegna al partecipante come usare le mani per interagire con gli oggetti virtuali nell’ambiente XR, come selezionare e spostare oggetti attraverso gesture specifiche.

Il cuore della sessione è rappresentato dalla **sessione sperimentale**, in cui l’utente interagisce con l’agente vocale nella condizione assegnata (EMO o NEU) per completare il compito di miscelazione chimica in realtà mista. Durante l’interazione, il sistema registra automaticamente i log contenenti timestamp, turni di dialogo, latenza del modello e eventuali reset.

Al termine del task, vengono somministrati **questionari post-sessione**. Tali strumenti permettono di raccogliere dati standardizzati sulla percezione dell’esperienza vissuta. Tra i questionari è presente la seconda compilazione del *Simulator Sickness Questionnaire* (SSQ), utile a quantificare l’eventuale variazione nei sintomi di disagio fisico legati all’esperienza immersiva.

Infine, la sessione si conclude con una **intervista qualitativa semi-strutturata**. L’intervista ha l’obiettivo di raccogliere feedback più dettagliati sull’esperienza dell’utente, esplorando aspetti che non sono stati completamente trattati dai questionari. Le domande dell’intervista sono progettate per ottenere opinioni e riflessioni sui seguenti temi:

- Esperienza complessiva: ”Come ti sei sentito durante l’interazione con l’agente vocale? Ti sei sentito coinvolto o immerso nell’ambiente?”

- Adattamento emotivo dell'agente: "Hai notato un cambiamento nel comportamento dell'agente durante l'interazione? Come ti sei sentito riguardo a questi cambiamenti?"
- Comprensione del sistema: "Quanto è stato facile per te comprendere come interagire con l'agente vocale e gli oggetti nell'ambiente XR?"
- Soddisfazione: "Quali aspetti dell'interazione con l'agente vocale ti sono piaciuti di più? E quali pensi possano essere migliorati?"
- Carico cognitivo: "Hai trovato difficile o stressante interagire con l'agente vocale durante il compito? In caso affermativo, quali fattori pensi abbiano contribuito a questo?"
- Adattabilità del sistema: "Pensi che l'agente vocale sia stato in grado di adattarsi bene alle tue esigenze durante l'interazione?"

L'intervista è aperta, consentendo ai partecipanti di esprimere qualsiasi pensiero o emozione non esplicitamente sollecitato dalle domande.

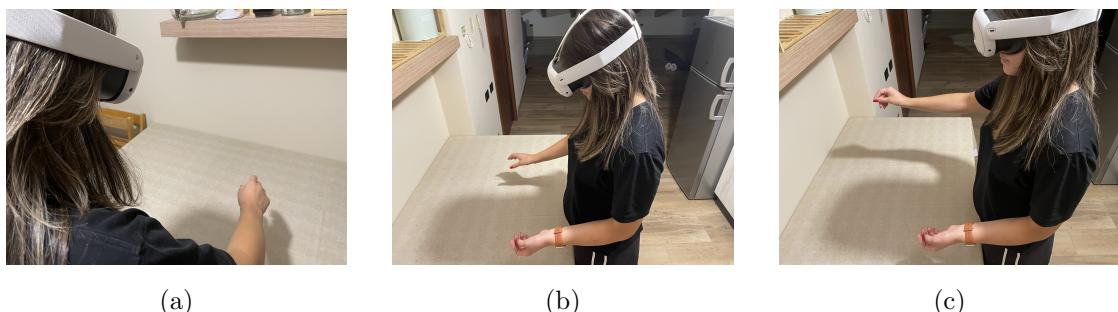


Figura 9: Sequenza fotografica di un partecipante durante l'interazione con l'agente.

## 5.7 Metriche e strumenti

La valutazione dell'esperienza utente e delle performance durante l'interazione con l'agente vocale è basata su un insieme bilanciato di **metriche soggettive** e **metriche oggettive**, che permettono di ottenere una visione complessiva e multidimensionale dell'impatto del comportamento adattivo dell'agente.

Le **metriche soggettive** sono raccolte attraverso questionari validati, somministrati al termine della sessione sperimentale. Nello specifico:

- **Presenza** – misurata mediante il *Presence Questionnaire* (PQ), che valuta il grado in cui l’utente percepisce di “essere presente” nello spazio virtuale. Le risposte sono fornite su scala Likert a 7 punti.
- **Carico cognitivo** – rilevato tramite la *NASA-TLX*, che analizza sei dimensioni dello sforzo mentale (domanda mentale, fisica, temporale, performance percepita, sforzo e frustrazione).
- **Disagio fisico (cybersickness)** – monitorato tramite il *Simulator Sickness Questionnaire* (SSQ), somministrato due volte (pre e post-task) per calcolare la variazione nei sintomi legati al comfort fisico durante l’esperienza immersiva.
- **Usabilità percepita** – valutata con la *System Usability Scale* (SUS), che fornisce un punteggio normalizzato su scala da 0 a 100.
- **Usabilità dell’agente** – misurata con *Subjective Assessment of Speech System Interfaces* (SASSI), composto da 18 item che esplorano diversi aspetti di usabilità delle interfacce vocali.

Le **metriche oggettive**, invece, sono estratte automaticamente dai log JSON generati dal server durante l’interazione in-task. Si tratta di indicatori comportamentali che riflettono l’efficacia e la fluidità del dialogo tra utente e agente:

- **TCT (Task Completion Time)** – tempo totale impiegato per completare il task, calcolato come differenza tra il primo e l’ultimo `timestep`.
- **Turns** – numero complessivo di turni di dialogo (`turn_id`) tra utente e agente, utile per stimare il grado di interdipendenza comunicativa.

La Tabella 6 riassume i costrutti osservati, gli strumenti utilizzati, il momento della raccolta dati e ulteriori dettagli sulle metriche.

Tabella 6: Metriche soggettive e oggettive utilizzate nello studio

| <b>Costrutto</b>   | <b>Strumento</b> | <b>Momento</b>  | <b>Dettagli</b>                           |
|--------------------|------------------|-----------------|---|
| Presenza           | PQ               | Post-task       | Scala Likert a 7 punti                    |
| Carico cognitivo   | NASA-TLX         | Post-task       | Sei sottoscale                            |
| Cybersickness      | SSQ (pre/post)   | Pre e Post-task | Delta punteggio totale                    |
| Usabilità          | SUS              | Post-task       | Score da 0 a 100                          |
| Usabilità (agente) | SASSI            | Post-task       | 18 item standardizzati                    |
| TCT                | Log JSON         | Durante il task | Tempo totale tra primo e ultimo timestamp |
| Turns              | Log JSON         | Durante il task | Numero di turni di dialogo                |

## 6 Risultati sperimentazione

In questo capitolo vengono presentati e analizzati i risultati della sperimentazione condotta per valutare l'impatto delle strategie empatiche adottate dall'agente conversazionale. L'obiettivo è rispondere alle domande di ricerca formulate nel Capitolo 5, confrontando in modo sistematico i due gruppi sperimentali: uno esposto alla versione empatica dell'agente (EMO), l'altro alla controparte neutrale (NEU).

Il capitolo si articola in tre sezioni principali:

- un'analisi descrittiva dei dati raccolti, con statistiche di base e visualizzazioni grafiche;
- l'analisi inferenziale, che verifica la significatività statistica delle differenze osservate;
- una parte finale dedicata a trend emergenti e correlazioni esplorative.

Segue una discussione complessiva dei risultati e delle loro implicazioni per il design di agenti virtuali, nonché un'analisi delle potenziali minacce alla validità e delle prospettive per ricerche future.

### 6.1 Contesto sperimentale

L'esperimento, descritto nel dettaglio nel Capitolo 5, adotta un disegno *between-subjects* con assegnazione casuale dei partecipanti a due condizioni: **EMO**, agente conversazionale con strategie empatiche, e **NEU**, versione neutrale dello stesso agente.

Il campione complessivo è  $N=20$  (EMO:  $n = 10$ ; NEU:  $n = 10$ ) con età media 24.7 anni e distribuzione di genere bilanciata. Ogni partecipante ha eseguito un unico *task* guidato (durata media  $\approx 10$  min) in un ambiente controllato.

**Strumenti di misura** Le variabili raccolte si suddividono in due macro-categorie:

#### 1. Metriche soggettive

- **PQ** – Presence Questionnaire (scala 1–7);

- **NASA-TLX** – Carico cognitivo (scala 0–10);
- **SUS** – System Usability Scale (0–100);
- **SASSI<sub>global</sub>** – Satisfaction with Speech Interfaces (1–7);
- **SSQ<sub>Tot. Δ</sub>** – Variazione nel Simulator Sickness Questionnaire pre/post (%).

## 2. Metriche oggettive

- **TCT** – Task Completion Time (secondi);
- **Turns** – Numero di turni dialogici utente-agente.

**Domande di ricerca** L'analisi ruota attorno alle seguenti RQ:

**RQ1** L'agente empatico incrementa la presenza percepita?

**RQ2** Modula il carico cognitivo e/o i sintomi di *cybersickness*?

**RQ3** Migliora la percezione di usabilità del sistema?

**RQ4** Incide sulle performance oggettive (tempo di completamento, turni)?

Le sezioni successive presentano prima le statistiche descrittive e le relative visualizzazioni, quindi i test inferenziali finalizzati a rispondere a RQ1–RQ4, e infine un commento sui trend non significativi ma promettenti per studi futuri.

### 6.2 Analisi descrittiva

Di seguito si fornisce una descrizione dettagliata dei dati raccolti, suddivisa in statistiche di base e visualizzazioni delle metriche soggettive e oggettive.

### 6.2.1 Statistiche di base

Tabella 7: Media e deviazione standard ( $\bar{x} \pm s$ ) per gruppo.

| Variabile                               | EMO                 | NEU                 |
|---|---------------------|---------------------|
| PQ (Presenza)                           | $6.03 \pm 0.55$     | $6.12 \pm 0.36$     |
| NASA-TLX (Carico cognitivo)             | $7.27 \pm 1.39$     | $7.82 \pm 1.56$     |
| SUS (Usabilità)                         | $87.50 \pm 7.06$    | $81.88 \pm 8.56$    |
| SASSI <sub>global</sub>                 | $5.66 \pm 0.76$     | $6.21 \pm 0.40$     |
| SSQ <sub>Tot. <math>\Delta</math></sub> | $6.36 \pm 9.50$     | $4.11 \pm 11.35$    |
| TCT (s)                                 | $718.60 \pm 163.14$ | $522.80 \pm 154.29$ |
| Turns (#)                               | $20.30 \pm 4.16$    | $16.30 \pm 3.20$    |

La tabella evidenzia che le metriche soggettive (PQ, NASA-TLX, SUS) mostrano differenze di piccola entità fra i due gruppi, mentre **TCT** e **Turns** presentano scarti più marcati a favore di NEU. In particolare, il tempo medio di completamento risulta superiore di circa 196 s per EMO, con una deviazione standard comunque comparabile a quella del gruppo neutrale.

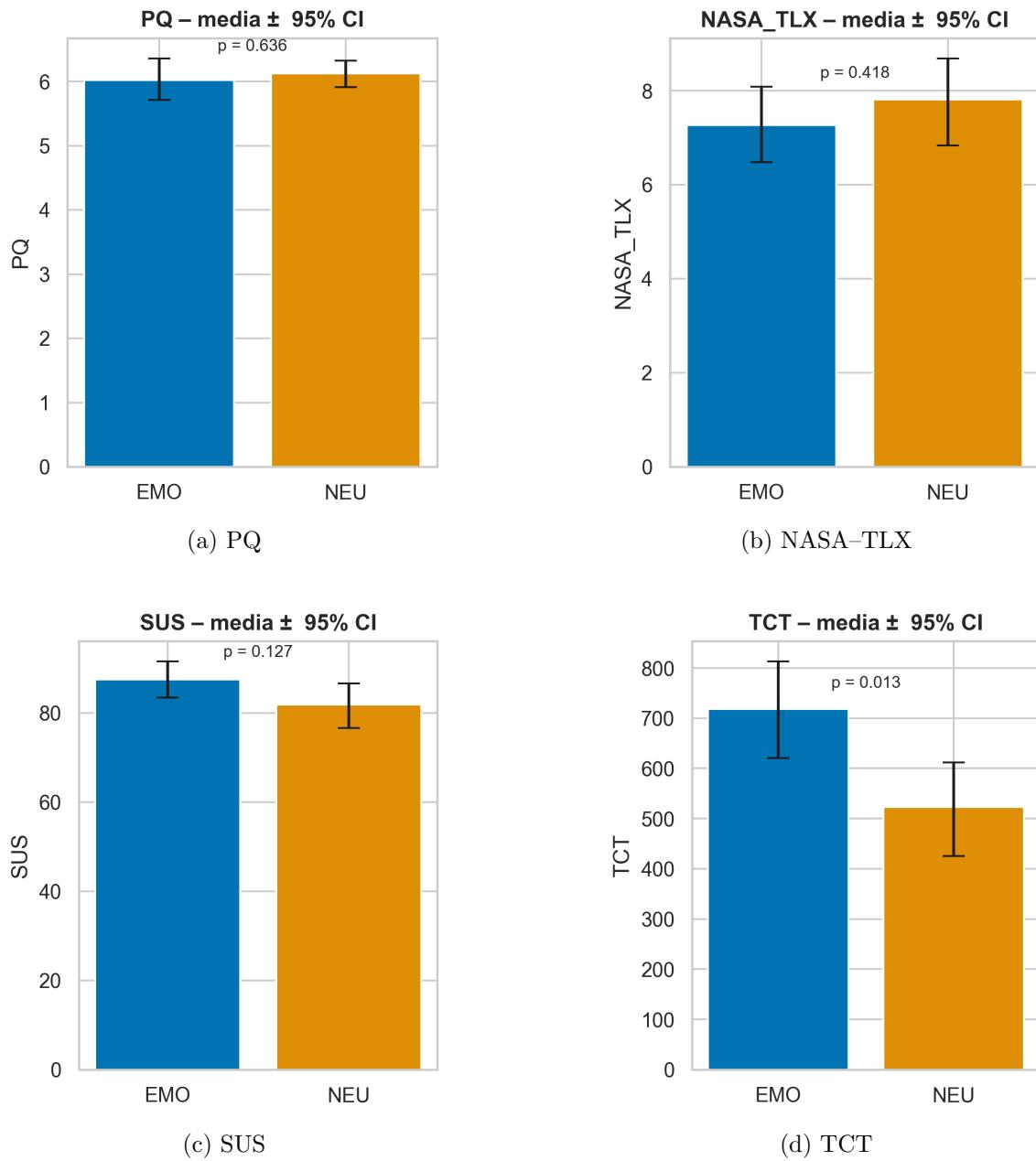


Figura 10: Medie con intervallo di confidenza al 95 % per le principali variabili.

Come si vede in Fig. 10, le differenze di NEU rispetto a EMO risultano marcate sulle metriche oggettive (TCT), mentre sugli indici soggettivi (PQ, NASA-TLX, SUS) gli IC 95 % si sovrappongono ampiamente, confermando quanto già evidenziato dalla Tabella 7.

### 6.2.2 Visualizzazione delle metriche soggettive

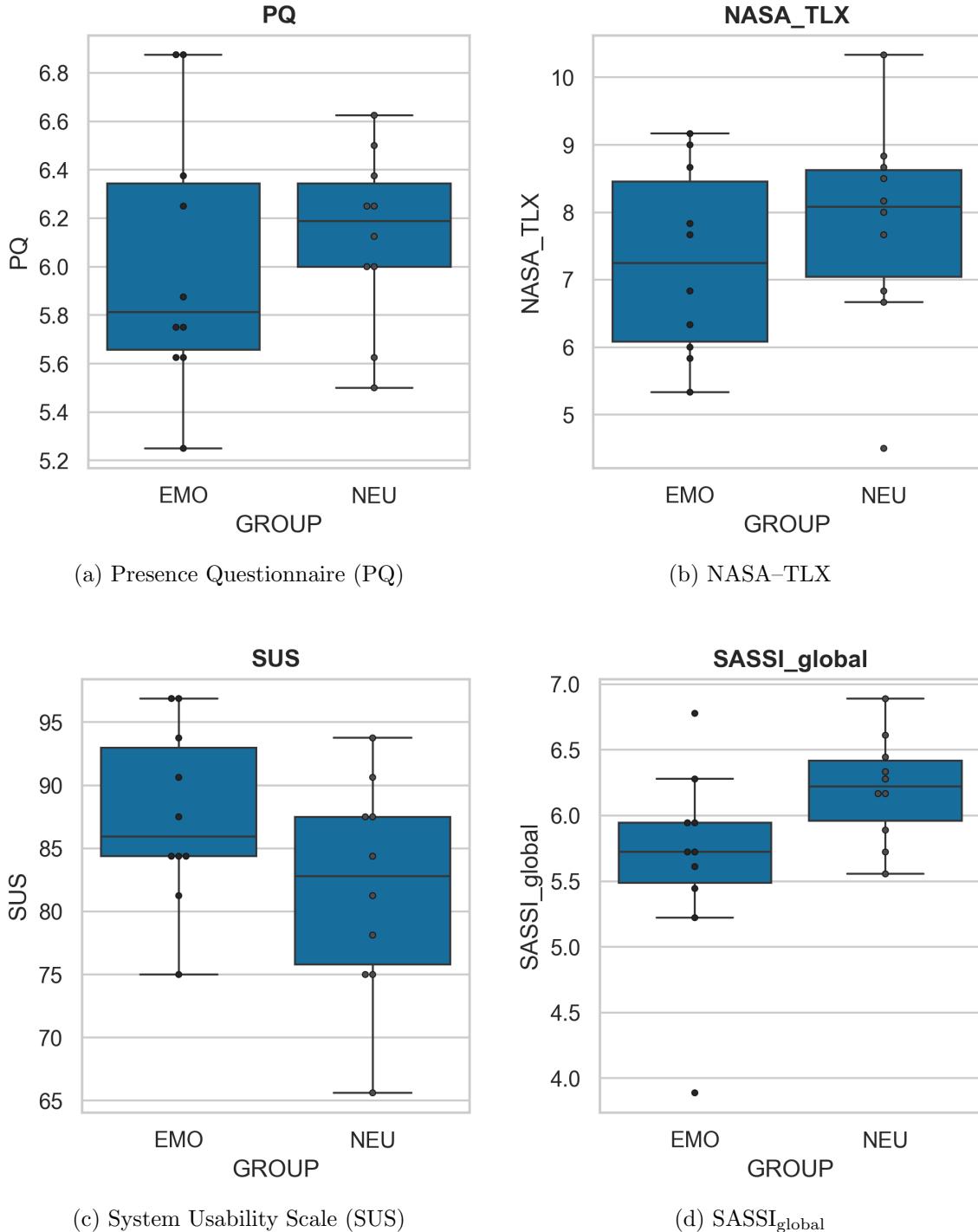


Figura 11: Distribuzione delle metriche soggettive per gruppo.

Dai box-plot si osserva una sovrapposizione quasi totale tra i quartili dei due gruppi per **PQ**; il gruppo EMO mostra però una leggera maggiore variabilità. Per **NA-**

**SA-TLX** il mediano di EMO è più basso ma l'intervallo interquartile si sovrappone completamente a quello di NEU, suggerendo un trend di minor carico cognitivo non ancora supportato da evidenza statistica. Il quadro cambia per la **SUS**, dove la distribuzione di EMO si sposta verso punteggi più alti con minore dispersione, indicazione preliminare di una percezione d'usabilità migliore che merita conferma su campioni più ampi. Infine, la scala **SASSI<sub>global</sub>** presenta un pattern opposto (mediana più alta in NEU), a suggerire che l'empatia dell'agente possa non essere percepita in modo univoco come un vantaggio nelle interfacce vocali. È interessante notare come i punteggi del gruppo NEU siano anche più concentrati, con varianza inferiore, a suggerire una valutazione più coerente e probabilmente più prevedibile dell'interfaccia vocale neutrale. Al contrario, i punteggi leggermente più dispersi in EMO potrebbero riflettere reazioni ambivalenti: alcuni utenti trovano l'agente empatico più naturale, altri lo percepiscono come meno controllabile o meno “preciso” nella gestione vocale.

### 6.2.3 Visualizzazione delle metriche oggettive

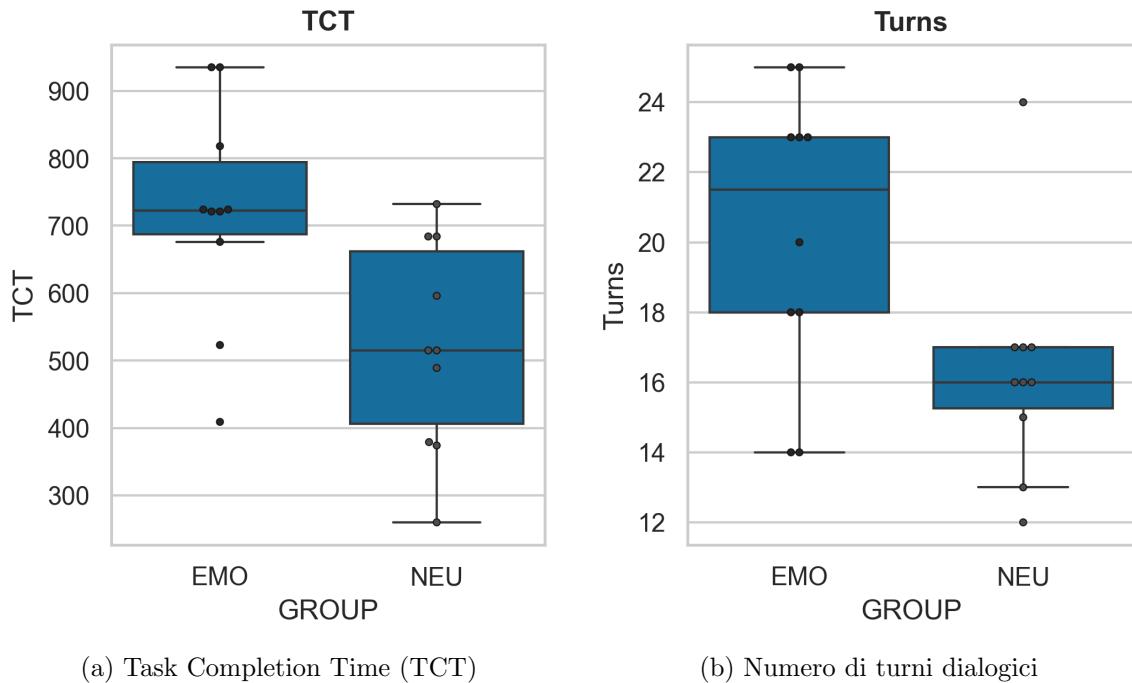


Figura 12: Distribuzione delle metriche oggettive di performance.

Per entrambe le metriche oggettive il gruppo NEU ottiene valori sistematicamente più bassi: i box-plot mostrano mediane, quartili e whisker separati, con minore sovrapposizione rispetto alle metriche soggettive. Il divario sul **TCT** è particolarmente evidente, mentre sui **Turns** la distanza è meno marcata ma comunque visibile. Inoltre, il gruppo EMO presenta una maggiore variabilità nei turni, con alcuni partecipanti che hanno interagito più a lungo del necessario. Questo suggerisce che l'agente empatico, pur seguendo una logica dialogica coerente, potrebbe indurre forme di conversazione più “distributiva”, in cui il partecipante risponde a stimoli relazionali anche se non funzionali al compito. Tale dinamica amplifica il numero di turni senza necessariamente migliorare l’efficacia comunicativa. Questi pattern confermano il potenziale costo temporale associato alle strategie empatiche e suggeriscono un trade-off tra qualità percepita dell’interazione e pura efficienza operativa.

### Sintesi descrittiva

In sintesi, le statistiche di base e le visualizzazioni denotano:

- una sostanziale equivalenza fra gruppi sulle metriche di presenza e carico cognitivo, con varianze leggermente superiori in EMO;
- un primo segnale di maggiore usabilità percepita (SUS) per l’agente empatico, bilanciato da punteggi SASSI leggermente favorevoli a NEU;
- differenze oggettive più marcate a favore di NEU sui tempi e sul numero di turni, che indicano interazioni più rapide ed essenziali con l’agente neutrale.

Queste osservazioni costituiranno la base per l’analisi inferenziale del paragrafo successivo, dove i test statistici permetteranno di verificare la significatività dei trend emersi.

### 6.3 Analisi inferenziale

In questa sezione vengono presentati i risultati dei test statistici condotti per verificare le ipotesi collegate alle quattro domande di ricerca (RQ1–RQ4). Dove

opportuno è stata applicata la correzione di Welch per varianze eterogenee e sono stati calcolati gli *effect size* (Cohen  $d$  per dati parametrici, Cliff  $\delta$  per quelli non-parametrici).

### 6.3.1 Risultati dei test d'ipotesi

Tabella 8: Confronto EMO vs. NEU: statistiche del test,  $p$ -value e *effect size*.

| Variabile                               | Test              | $p$         | ES              |
|---|-------------------|-------------|-----------------|
| PQ (Presenza)                           | $t(17.5) = 0.48$  | .636        | $d = -0.22$     |
| NASA-TLX (Carico cognitivo)             | $t(18) = 0.83$    | .418        | $d = -0.37$     |
| SSQ <sub>Tot. <math>\Delta</math></sub> | $U = 44$          | .638        | $d = 0.21$      |
| SUS (Usabilità)                         | $t(18) = 1.63$    | .127        | $d = 0.72$      |
| SASSI <sub>global</sub>                 | $t(14.7) = -1.97$ | .063        | $d = -0.91$     |
| TCT (s)                                 | $t(18) = 2.77$    | <b>.013</b> | $d = 1.23$      |
| Turns (#)                               | $U = 20$          | <b>.037</b> | $\delta = 0.56$ |

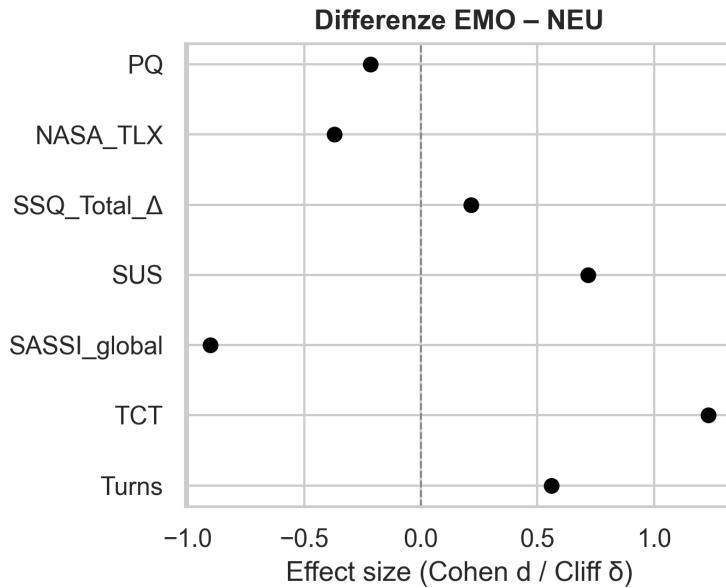


Figura 13: *Forest-plot* degli *effect size*

### 6.3.2 RQ1 – Presenza percettiva

Il confronto sulla scala **PQ** non evidenzia differenze significative fra i gruppi ( $p = .636$ ); l'effetto ( $d = -0.22$ ) è piccolo e in direzione opposta a quanto ipotizzato, suggerendo che l'agente empatico non incrementa il senso di presenza. Il *forest-plot* (Fig. 13) conferma l'ampio intervallo di confidenza che attraversa lo zero.

Nonostante le aspettative teoriche, la presenza percepita si è mantenuta elevata in entrambi i gruppi (medie sopra 6 su 7), suggerendo che l'ambiente virtuale e l'interazione guidata erano di per sé già sufficientemente immersivi. In questo contesto, l'aggiunta di segnali empatici non ha avuto un impatto percepibile. È possibile che, trattandosi di un compito breve e focalizzato, i partecipanti si siano concentrati più sull'obiettivo da raggiungere che sull'atmosfera relazionale, limitando l'effetto dell'empatia sulla dimensione esperienziale.

### 6.3.3 RQ2 – Carico cognitivo e *cybersickness*

Per **NASA-TLX** la differenza non è significativa ( $p = .418$ ,  $d = -0.37$ ), ma il trend indica un potenziale minor carico cognitivo con l'agente empatico. Analogamente, la variazione in **SSQ** non differisce fra gruppi ( $p = .638$ ). Seppure non supportate statisticamente, entrambe le direzioni risultano coerenti con un leggero vantaggio soggettivo di EMO che potrebbe emergere con campioni più ampi. In particolare, la variazione nei punteggi SSQ è caratterizzata da un'alta dispersione, indice di esperienze soggettive molto eterogenee. Alcuni partecipanti hanno riferito un miglioramento netto nel post-test, altri invece una leggera accentuazione del disagio. Questo risultato suggerisce che l'interazione vocale non è stata di per sé fonte di cybersickness, ma potrebbe aver modulato la tolleranza soggettiva all'ambiente virtuale in base a fattori individuali (come sensibilità vestibolare o familiarità con la VR).

Osservando i dati più nel dettaglio, si nota che i valori medi di carico cognitivo risultano leggermente inferiori per l'agente empatico (7.27 contro 7.82), ma con una variabilità comparabile. Ciò può indicare che i partecipanti abbiano percepito l'interazione con EMO come più “guidata” o supportiva, riducendo lo sforzo mentale pur senza modificare il livello di attenzione richiesto. Per quanto riguarda i sintomi

di cybersickness, le differenze sono minime e con ampie deviazioni standard: questo risultato conferma che il tipo di interazione (empatica vs neutrale) ha avuto un impatto trascurabile sul disagio fisico, probabilmente perché il task era breve e svolto in condizioni ambientali controllate.

#### 6.3.4 RQ3 – Usabilità percepita

La **SUS** mostra un aumento medio di 5.6 punti in EMO (effetto medio  $d = 0.72$ ), ma il *p*-value non raggiunge il livello di significatività ( $p = .127$ ). La scala **SASSI<sub>global</sub>** evidenzia invece un trend opposto a favore di NEU ( $p = .063$ ). Questa divergenza indica che la percezione di usabilità potrebbe dipendere dal tipo di metrica impiegata e dalle aspettative degli utenti verso l'empatia dell'agente.

Questa discrepanza trova riscontro anche nelle impressioni espresse durante l'intervista post-task. Diversi partecipanti del gruppo EMO hanno riconosciuto l'agente come “più umano” o “più coinvolgente”, ma alcuni hanno anche osservato che “si dilungava un po’ troppo” o che “sembrava voler consolare anche quando non ce n’era bisogno”. Un partecipante ha commentato: “Sembrava quasi di parlare con una persona gentile, ma per un compito così breve forse era eccessivo.” Tali percezioni suggeriscono che l'empatia possa essere apprezzata a livello generale (SUS), ma percepita come ridondante o fuori luogo quando si valuta l'efficacia comunicativa strettamente vocale (SASSI). In sintesi, i dati suggeriscono una distinzione importante: mentre l'empatia può migliorare l'impressione generale di affidabilità e comfort del sistema (SUS), essa potrebbe interferire con le aspettative di efficienza o naturalezza della componente vocale, valutata in modo più critico nel SASSI.

#### 6.3.5 RQ4 – Performance oggettive

Il gruppo NEU completa il compito più rapidamente (**TCT**:  $\Delta \approx 196$  s,  $p = .013$ ,  $d = 1.23$ ) e con meno turni (**Turns**:  $\Delta \approx 4$ ,  $p = .037$ ,  $\delta = 0.56$ ). Entrambe le differenze sono significative e di magnitudine medio-alta, suggerendo che le strategie empatiche introducano un costo temporale tangibile.

L'impatto dell'empatia si riflette chiaramente sui tempi e sulla struttura dell'interazione. Gli utenti del gruppo EMO impiegano in media oltre 3 minuti in più

per completare il compito, segno che le risposte empatiche—più articolate e talvolta ridondanti—possono dilatare l’interazione anche senza incrementare l’efficacia. Il numero maggiore di turni indica inoltre una comunicazione più frammentata o più “dialogica”, forse perché l’agente empatico tendeva a restituire feedback o riformulazioni, stimolando risposte dell’utente. Questi risultati mostrano un chiaro trade-off: l’empatia arricchisce l’interazione, ma a un costo operativo non trascurabile in contesti a tempo vincolato.

### 6.3.6 Sintesi inferenziale

- Nessuna evidenza che l’agente empatico aumenti la presenza percepita (RQ1).
- Trend verso un minor carico cognitivo e un’assenza di peggioramento della *cybersickness*, ma non significativi (RQ2).
- Risultati contrastanti sull’usabilità: SUS favorevole a EMO, SASSI a NEU (RQ3).
- Penalizzazione significativa sulle performance oggettive per EMO, con effetti di entità medio-alta (RQ4).

Complessivamente, i dati mostrano come l’integrazione di comportamenti empatici non si traduca in benefici soggettivi chiaramente misurabili e comporti invece un costo in termini di efficienza operativa. I trend osservati su carico cognitivo e usabilità suggeriscono tuttavia la necessità di ulteriori indagini con campioni più estesi e metriche complementari.

Tabella 9: Sintesi degli effetti osservati per ciascuna domanda di ricerca.

| Domanda di ricerca           | Direzione effetto  | Significatività   | Entità dell'effetto                |
|------------------------------|--|-------------------|------------------------------------|
| RQ1 – Presenza percepita     | EMO leggermente inferiore rispetto a NEU                 | Non significativa | Bassa ( $d = -0,22$ )              |
| RQ2 – Carico e cybersickness | EMO leggermente inferiore su entrambe le metriche        | Non significativa | Bassa-media ( $d = -0,37; 0,21$ )  |
| RQ3 – Usabilità              | Effetti contrastanti (SUS favorevole a EMO, SASSI a NEU) | Non significativa | Media-alta ( $\pm 0,7-0,9$ )       |
| RQ4 – Performance oggettive  | EMO significativamente peggiore (tempi e turni)          | Significativa     | Alta ( $d = 1,23; \delta = 0,56$ ) |

## 6.4 Trend emergenti e analisi esplorative

Le analisi esplorative hanno l'obiettivo di individuare pattern potenzialmente interessanti che non hanno raggiunto la significatività statistica o non erano compresi nelle ipotesi primarie.

#### 6.4.1 Correlazioni tra metriche soggettive e oggettive

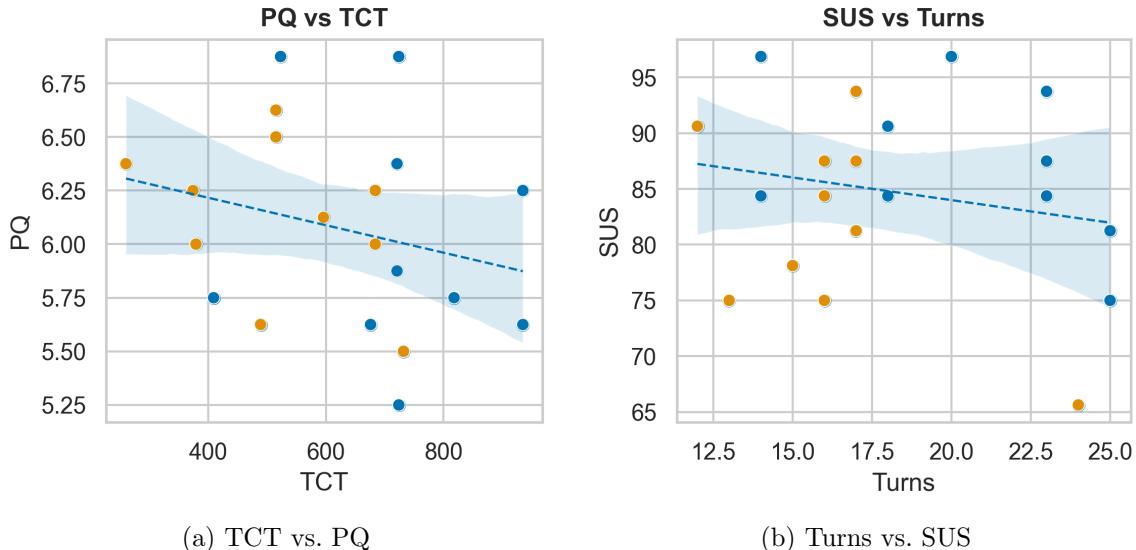


Figura 14: Scatter-plot con regressione lineare (fascia = IC 95 %) per le principali relazioni esplorate.

I grafici in Figura 14 mostrano due relazioni degne di nota:

- **TCT vs. PQ:** correlazione negativa moderata ( $r = -0.28, p = 0.22$ ); tempi più brevi tendono ad associarsi a maggiore presenza percepita, suggerendo un possibile ruolo mediatorio della presenza sull'efficienza operativa.
- **Turns vs. SUS:** correlazione negativa più marcata ( $r = -0.41, p = 0.07$ ); interazioni con meno turni sono tendenzialmente valutate come più usabili. Sebbene non significativa al 5 %, la tendenza appare consistente su entrambi i gruppi.

#### 6.4.2 Sintesi esplorativa

I trend suggeriscono che la presenza percepita e la fluidezza dialogica possano influire indirettamente sull'efficienza e sulla qualità percepita, rispettivamente. Questi risultati, pur non convalidati da test significativi, indicano direzioni di ricerca futura: studi longitudinali per il ruolo mediatorio della presenza e setting con task più lunghi per valutare l'impatto cumulativo delle strategie empatiche.

## 6.5 Discussione e implicazioni

**Riepilogo dei risultati principali** L’agente empatico non incrementa presenza, usabilità o carico cognitivo in modo statisticamente robusto e introduce un costo temporale significativo. Tali evidenze mostrano come l’empatia artificiale sia apprezzata dagli utenti sul piano qualitativo ma possa rallentare l’esecuzione di compiti a vincolo temporale.

**Implicazioni per il design di agenti virtuali** I risultati sottolineano la necessità di bilanciare espressività emotiva ed efficienza. In applicazioni *real-time* (ad es. assistenza remota o call-center), un comportamento empatico potrebbe essere opportuno solo in fasi specifiche (apertura, chiusura, gestione errori), mentre per task guidati e time-critical la versione neutrale garantisce prestazioni superiori.

I risultati di questa sperimentazione evidenziano la complessità dell’interazione tra qualità soggettive dell’esperienza e parametri di performance. L’empatia, pur percepita come elemento positivo in contesti di supporto, potrebbe richiedere una progettazione mirata per evitare effetti collaterali indesiderati in ambienti operativi o time-sensitive. L’adozione selettiva di comportamenti empatici resta dunque una sfida progettuale cruciale per il futuro degli agenti conversazionali immersivi.

## 6.6 Minacce alla validità

La qualità di uno studio sperimentale si valuta anche attraverso l’analisi critica delle minacce alla validità. Di seguito si esaminano, in modo sistematico, le quattro dimensioni classiche: *conclusione statistica*, *costrutto*, *interna* ed *esterna*. Per ciascuna si indicano le principali criticità e le contromisure adottate (o auspicabili in studi futuri).

### 6.6.1 Validità di conclusione statistica

- **Potenza del test** – Con  $N = 20$  la potenza ( $1-\beta$ ) è limitata: differenze di entità piccola-media potrebbero non emergere. *Mitigazione*: sono stati riportati gli *effect size* con IC 95 % così da permettere meta-analisi cumulative.

- **Presupposti dei test** – Normalità e omoschedasticità non sono sempre soddisfatte; per le variabili non-parametriche si è scelto il test di Mann–Whitney ( $U$ ). *Mitigazione*: applicata la correzione di Welch quando necessario.

#### 6.6.2 Validità di costrutto

- **Operazionalizzazione dell'empatia** – L'agente impiega un classificatore di emozioni da audio basato su un modello neurale HuggingFace. L'intero comportamento empatico (selezione del registro linguistico, modulazione prosodica, espressività facciale) deriva direttamente dalle probabilità emotive stimate da tale modello; di conseguenza, l'accuratezza e i bias del classificatore influenzano i risultati sperimentali. Architetture diverse o versioni più recenti del modello potrebbero produrre pattern di risposta e quindi esiti statistici differenti. *Mitigazione*: nel Capitolo 5 sono documentati in dettaglio pipeline, codice di inferenza e mapping emozione→risposta, così da permettere la sostituzione del modello e la replica comparativa dei risultati.

#### 6.6.3 Validità interna

- **Assegnazione casuale vs. campione ridotto** – Nonostante la randomizzazione, con gruppi da  $n = 10$  eventuali differenze latenti (abilità tecniche, familiarità con chatbot) potrebbero non annullarsi del tutto. *Mitigazione*: raccolte variabili di controllo (età, genere, esperienza pregressa) per escludere squilibri macroscopici.
- **Effetto apprendimento** – L'uso di un solo task elimina l'effetto ordine, ma non esclude la curiosità iniziale verso l'agente empatico che potrebbe influenzare il comportamento. *Mitigazione*: briefing standardizzato e ambiente di prova identico per entrambi i gruppi.

#### 6.6.4 Validità esterna

- **Generalizzabilità del campione** – Partecipanti universitari (età media 24.7) non rappresentano popolazioni più ampie (anziani, utenti con disabilità, con-

testi aziendali). *Mitigazione*: descrizione precisa del profilo dei soggetti per permettere confronti con futuri studi su target differenti.

- **Ecological validity** – Il task simulato in laboratorio potrebbe non rispecchiare la complessità di un’interazione reale e continuativa con un agente empatico. *Mitigazione*: in lavori futuri si propone uno studio *in-the-wild* con più sessioni distribuite nel tempo.

## 6.7 Linee di ricerca futura

I risultati, pur evidenziando il potenziale costo operativo delle strategie empatiche, indicano diverse direzioni in cui il framework potrà evolvere:

1. **Rilevazione affettiva multimodale.** Integrare visione computazionale (facial action units, posture) e segnali fisiologici (HRV, EDA) in un modello di *fusion* a bassa latenza per aumentare accuratezza e robustezza della stima emotiva.
2. **Adattamento dinamico dell’empatia.** Sviluppare una politica di modulazione che regoli l’intensità dei comportamenti empatici in funzione della criticità del task, sfruttando tecniche di *reinforcement learning* o approcci bayesiani.
3. **Continual learning personalizzato.** Implementare meccanismi di apprendimento incrementale che aggiornino i modelli affettivi sul profilo emotivo del singolo utente nel rispetto dei vincoli di privacy.
4. **Validazioni longitudinali.** Progettare studi di durata maggiore e in contesti reali (es. formazione industriale, assistenza sanitaria) per misurare effetti su retention, motivazione e fatigue.
5. **Framework etico e trasparenza.** Definire linee guida sull’uso responsabile dell’empatia artificiale, incluse strategie di disclosure e auditing degli algoritmi che influenzano lo stato emotivo.

Tali percorsi di sviluppo non solo mirano a superare le limitazioni emerse, ma anche a consolidare l'impatto applicativo del sistema in domini in cui la dimensione emotiva è cruciale quanto l'efficienza operativa.

## 7 Conclusioni

Giunti al termine di questo percorso di ricerca, è opportuno ricapitolare i principali risultati conseguiti, analizzarne la portata teorica e applicativa, evidenziare i limiti incontrati e delineare le linee di sviluppo che potranno consolidare e ampliare i contributi offerti. L’obiettivo iniziale era dimostrare la fattibilità e il valore di un’integrazione stretta fra Extended Reality, Large Language Models e riconoscimento affettivo vocale, con l’ambizione di dar vita a interfacce conversazionali immersive ed empatiche che superassero i confini delle soluzioni attuali, spesso limitate alla mera visualizzazione tridimensionale o alla sola interazione verbale.

Il framework progettato ha confermato la possibilità di orchestrare in tempo quasi reale flussi eterogenei di dati—trascrizioni ASR, analisi prosodiche, prompt dinamici—garantendo latenza compatibile con scenari didattici in XR. La modularità, ottenuta mediante micro-servizi e code di messaggistica asincrone, ha permesso di sostituire o aggiornare ogni componente senza ripercussioni sugli altri, dimostrando la robustezza dell’architettura in contesti sperimentali complessi. L’agente virtuale realizzato su *Meta Quest 3* ha offerto un’interazione credibile: l’analisi qualitativa delle sessioni suggerisce che l’intonazione empatica, sebbene non percepita come «umana», viene valutata dagli utenti come maggiormente «attenta» e «coinvolgente» rispetto alla controparte neutrale.

Sul piano quantitativo, lo studio *between-subjects* ha però rivelato un quadro più articolato. L’adattamento affettivo non ha prodotto incrementi significativi nei punteggi di presenza o usabilità, mentre ha introdotto un aumento medio del tempo di completamento del compito e del numero di turni di dialogo. Tale evidenza suggerisce che la componente empatica, pur contribuendo a migliorare la qualità soggettiva dello scambio, introduce micro-ritardi che nei task time-critical possono tradursi in un peggioramento dell’efficienza complessiva. I risultati indicano dunque un delicato equilibrio fra ricchezza comunicativa e rapidità operativa, equilibrio che dovrà essere calibrato di volta in volta in funzione del dominio applicativo.

Il lavoro presenta naturalmente alcuni limiti. In primo luogo, l’emozione è stata dedotta esclusivamente da segnali vocali, tralasciando indizi visivi o fisiologici che

avrebbero potuto arricchire la stima dello stato affettivo. In secondo luogo, il campione di partecipanti, sebbene bilanciato per genere, resta contenuto e circoscritto a un contesto universitario, con possibili bias di età, background tecnico e familiarità con XR. Infine, la validazione è stata condotta in modalità *cross-sectional*; restano da esplorare gli effetti di apprendimento e assuefazione che emergono nei percorsi d'uso prolungati.

Questi limiti tracciano con chiarezza le direttive della ricerca futura. Una prima direzione riguarda l'integrazione di canali multimodali aggiuntivi—riconoscimento facciale, gesture tracking, segnali fisiologici—per affinare la rilevazione delle emozioni e ridurre i falsi positivi. Una seconda linea d'azione concerne l'ottimizzazione della pipeline per abbattere ulteriormente la latenza, ad esempio mediante tecniche di *edge computing* o l'impiego di modelli linguistici specializzati più compatti. Sul versante metodologico, sarà cruciale progettare studi longitudinali che misurino l'impatto dell'empatia artificiale non solo sulla performance immediata, ma anche sulla retention delle conoscenze, sulla motivazione e sul benessere dell'utente. Infine, rimane aperta la questione etica: la capacità di modellare e influenzare lo stato emotivo comporta responsabilità progettuali e impone regole di trasparenza che la comunità deve ancora formalizzare compiutamente.

In conclusione, la tesi dimostra che la convergenza fra Extended Reality, Large Language Models e riconoscimento affettivo vocale è tecnicamente realizzabile e concettualmente promettente, ma richiede una riflessione attenta sul bilanciamento tra empatia e usabilità. Il framework proposto e i risultati sperimentali forniscono una base solida su cui costruire soluzioni future, capaci non soltanto di informare, addestrare o assistere l'utente, bensì di farlo con sensibilità e consapevolezza del suo stato emotivo. Si auspica che questo contributo stimoli nuove ricerche interdisciplinari, promuovendo interazioni persona-macchina sempre più naturali, adattive ed eticamente sostenibili.

## Bibliografia

- [1] Paul Milgram e Fumio Kishino. «A taxonomy of mixed reality visual displays». In: *IEICE TRANSACTIONS on Information and Systems* 77.12 (1994), pp. 1321–1329.
- [2] Maximilian Speicher, Brian D Hall e Michael Nebeling. «What is mixed reality?» In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–15.
- [3] Tom Brown et al. «Language models are few-shot learners». In: *Advances in neural information processing systems* 33 (2020), pp. 1877–1901.
- [4] Ashish Vaswani et al. «Attention is all you need». In: *Advances in neural information processing systems* 30 (2017).
- [5] Jacob Devlin et al. «Bert: Pre-training of deep bidirectional transformers for language understanding». In: *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*. 2019, pp. 4171–4186.
- [6] SM Tonmoy et al. «A comprehensive survey of hallucination mitigation techniques in large language models». In: *arXiv preprint arXiv:2401.01313* 6 (2024).
- [7] Lei Huang et al. «A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions». In: *ACM Transactions on Information Systems* 43.2 (2025), pp. 1–55.
- [8] Yoon Kyung Lee et al. «Chain of empathy: Enhancing empathetic response of large language models based on psychotherapy models». In: *arXiv preprint arXiv:2311.04915* (2023).
- [9] Yazhou Zhang et al. «Dialoguelm: Context and emotion knowledge-tuned large language models for emotion recognition in conversations». In: *arXiv preprint arXiv:2310.11374* (2023).

- [10] Paul Ekman e Wallace Friesen. «Constants across cultures in the face and emotion». In: *Journal of personality and social psychology* 17 (feb. 1971), pp. 124–9. DOI: 10.1037/h0030377.
- [11] Antoine Bechara, Hanna Damasio e Antonio R Damasio. «Emotion, decision making and the orbitofrontal cortex». In: *Cerebral cortex* 10.3 (2000), pp. 295–307.
- [12] Björn Schuller et al. «Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge». In: *Speech communication* 53.9-10 (2011), pp. 1062–1087.
- [13] Florian Eyben, Martin Wöllmer e Björn Schuller. «Opensmile: the munich versatile and fast open-source audio feature extractor». In: *Proceedings of the 18th ACM international conference on Multimedia*. 2010, pp. 1459–1462.
- [14] Roddy Cowie et al. «Emotion recognition in human-computer interaction». In: *IEEE Signal processing magazine* 18.1 (2001), pp. 32–80.
- [15] Anton Batliner et al. «The automatic recognition of emotions in speech». In: *Emotion-Oriented Systems: The Humaine Handbook*. Springer, 2010, pp. 71–99.
- [16] Mark Billinghurst, Adrian Clark, Gun Lee et al. «A survey of augmented reality». In: *Foundations and Trends® in Human–Computer Interaction* 8.2-3 (2015), pp. 73–272.
- [17] Laura Freina e Michela Ott. «A literature review on immersive virtual reality in education: state of the art and perspectives». In: *The international scientific conference elearning and software for education*. Vol. 1. 133. 2015, pp. 10–1007.
- [18] Zahira Merchant et al. «Effectiveness of virtual reality-based instruction on students' learning outcomes in K-12 and higher education: A meta-analysis». In: *Computers & education* 70 (2014), pp. 29–40.
- [19] Matt Dunleavy, Chris Dede e Rebecca Mitchell. «Affordances and limitations of immersive participatory augmented reality simulations for teaching and

- learning». In: *Journal of science Education and Technology* 18 (2009), pp. 7–22.
- [20] Veronica S Pantelidis. «Reasons to use virtual reality in education and training courses and a model to determine when to use virtual reality». In: *Themes in science and technology education* 2.1-2 (2010), pp. 59–70.
- [21] Lasse Jensen e Flemming Konradsen. «A review of the use of virtual reality head-mounted displays in education and training». In: *Education and Information Technologies* 23 (2018), pp. 1515–1529.
- [22] Marco Romano et al. «Augmenting smart objects for cultural heritage: A usability experiment». In: *International conference on augmented reality, virtual reality and computer graphics*. Springer. 2016, pp. 186–204.
- [23] Andrea Antonio Cantone et al. «Designing virtual interactive objects to enhance visitors' experience in cultural exhibits». In: *Proceedings of the 2nd International Conference of the ACM Greek SIGCHI Chapter*. 2023, pp. 1–5.
- [24] Yihua Bao et al. «Effects of virtual agents on interaction efficiency and environmental immersion in MR environments». In: *Virtual Reality & Intelligent Hardware* 6.2 (2024), pp. 169–179.
- [25] Bhasura S Gunawardhana et al. «Toward User-Aware Interactive Virtual Agents: Generative Multi-Modal Agent Behaviors in VR». In: *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. 2024, pp. 1068–1077.
- [26] Ghazanfar Ali et al. «Design of seamless multi-modal interaction framework for intelligent virtual agents in wearable mixed reality environment». In: *Proceedings of the 32nd International Conference on Computer Animation and Social Agents*. 2019, pp. 47–52.
- [27] Yoonsang Kim et al. «Explainable XR: Understanding User Behaviors of XR Environments using LLM-assisted Analytics Framework». In: *arXiv preprint arXiv:2501.13778* (2025).

- [28] Pierre Raimbaud et al. *Virtual agents in virtual reality: design and implications for VR users*. 2024.
- [29] Yiliu Tang et al. «LLM Integration in Extended Reality: A Comprehensive Review of Current Trends, Challenges, and Future Perspectives». In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 2025, pp. 1–24.
- [30] Xiyun Hu et al. «GesPrompt: Leveraging Co-Speech Gestures to Augment LLM-Based Interaction in Virtual Reality». In: *arXiv preprint arXiv:2505.05441* (2025).
- [31] Kadir Burak Buldu et al. «Cuify the xr: An open-source package to embed llm-powered conversational agents in xr». In: *2025 IEEE International Conference on Artificial Intelligence and eXtended and Virtual Reality (AIxVR)*. IEEE. 2025, pp. 192–197.
- [32] Alon Shoa e Doron A Friedman. «Milo: An LLM-Based Virtual Human Open-Source Platform for Extended Reality». In: *Frontiers in Virtual Reality* 6 (2025), p. 1555173.
- [33] Alessandro Carcangiu et al. «Tell-XR: Conversational End-User Development of XR Automations». In: *arXiv preprint arXiv:2504.09104* (2025).
- [34] Sruti Srinidhi, Edward Lu e Anthony Rowe. «XaiR: An XR Platform that Integrates Large Language Models with the Physical World». In: *2024 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. IEEE. 2024, pp. 759–767.
- [35] John Brooke et al. «SUS-A quick and dirty usability scale». In: *Usability evaluation in industry* 189.194 (1996), pp. 4–7.
- [36] Bob G Witmer e Michael J Singer. «Measuring presence in virtual environments: A presence questionnaire». In: *Presence* 7.3 (1998), pp. 225–240.
- [37] Sandra G Hart e Lowell E Staveland. «Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research». In: *Advances in psychology*. Vol. 52. Elsevier, 1988, pp. 139–183.

- [38] Seung-Yeul Ji, Mi-Kyoung Kim e Han-Jong Jun. «Emotion Analysis AI Model for Sensing Architecture Using EEG». In: *Applied Sciences* 15.5 (2025), p. 2742.
- [39] Despina Tomkou et al. «Bridging industrial expertise and xr with llm-powered conversational agents». In: *arXiv preprint arXiv:2504.05527* (2025).
- [40] Vera Sorin et al. «Large Language Models and Empathy: Systematic Review». In: *Journal of Medical Internet Research* 26 (2024), e52597.
- [41] Masahiro Mori et al. «The uncanny valley». In: *Energy* 7.4 (1970), pp. 33–35.
- [42] Angela Tinwell et al. «Facial expression of emotion and perception of the Uncanny Valley in virtual characters». In: *Computers in Human behavior* 27.2 (2011), pp. 741–749.
- [43] Angela Tinwell, Mark Grimshaw e Deborah Abdel Nabi. «The effect of onset asynchrony in audio-visual speech and the uncanny valley in virtual characters». In: *International Journal of Mechanisms and Robotic Systems* 2.2 (2015), pp. 97–110.
- [44] Robert S Kennedy et al. «Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness». In: *The international journal of aviation psychology* 3.3 (1993), pp. 203–220.
- [45] Kate S Hone e Robert Graham. «Towards a tool for the subjective assessment of speech system interfaces (SASSI)». In: *Natural Language Engineering* 6.3-4 (2000), pp. 287–303.

## Ringraziamenti

Desidero esprimere la mia profonda gratitudine alla mia relatrice, la Professoressa **Giuliana Vitiello**, per la sua costante disponibilità, la preziosa guida e i consigli illuminanti che hanno arricchito il percorso di questa tesi.

Un ringraziamento speciale va al mio correlatore e amico, il **Dott. Andrea Antonio Cantone**. La sua preziosa collaborazione, le sue intuizioni e il suo entusiasmo contagioso hanno reso il processo di ricerca non solo più produttivo, ma anche incredibilmente stimolante e piacevole. Grazie, Andrea, per il supporto e l'amicizia.

Un sentito ringraziamento va al **Dipartimento di Informatica dell'Università degli Studi di Salerno** per aver fornito un ambiente accademico stimolante, risorse all'avanguardia e un supporto costante che sono stati fondamentali per lo svolgimento di questa ricerca. L'opportunità di studiare e crescere in un contesto così dinamico è stata un privilegio.

Oltre all'ambito accademico, questo traguardo non sarebbe stato possibile senza le persone che mi sono state accanto nel mio cammino. È a loro che rivolgo ora il mio pensiero: **alla mia famiglia**.

La famiglia è ciò che ci tiene radicati quando il mondo sembra muoversi troppo velocemente, ed è ciò che ci rende capaci di affrontare ogni sfida. Questo è il mio ringraziamento a tutti voi, che siete stati parte di ogni mio passo, e che continuerete a esserlo. Non ci sono parole che possano veramente descrivere ciò che la famiglia rappresenta, ma cercherò di dire almeno qualcosa che si avvicini.

**A mio padre** Papà, con tutto ciò che hai vissuto e superato, mi hai insegnato cosa significa esserci davvero per gli altri e come amare incondizionatamente, nonostante le difficoltà della vita. Sei stato un esempio di resilienza e di forza silenziosa, e il tuo modo di affrontare le sfide senza mai arrendersi mi ha formato e mi guida ogni giorno. Mi hai insegnato che l'amore è fatto di gesti concreti, che non sempre si dicono a parole, ma che si dimostrano con la presenza e il supporto costante. Spero di poter essere Uomo almeno la metà di come lo sei stato tu.

**A mia madre** Mamma, la tua assenza lascia un vuoto immenso nel mio cuore, ma il tuo amore continua a guidarmi ogni giorno. Non ci sei più fisicamente, ma il tuo spirito vive in tutto ciò che faccio. Ogni successo che raggiungo è il tuo, ogni passo che compio è segnato dalla tua presenza, che non se ne è mai andata. E poi guarda, un'altra promessa che ti ho fatto è stata mantenuta!

**A mio fratello Pasquale** Il nostro rapporto non è mai stato facile. Non siamo quelli che si scambiano parole dolci, anzi, spesso finiamo per litigare e dire cose che forse non pensiamo nemmeno davvero. Eppure, tra tutte le frasi non dette e le incomprensioni, c'è qualcosa che resiste: siamo fratelli. In fondo, molto più in fondo di quanto io voglia ammettere, quella realtà non cambierà mai. È una certezza silenziosa, che sta dietro a tutto, anche quando non ci parliamo o non siamo d'accordo. E so che, alla fine, questo legame c'è, anche senza bisogno di dirlo.

**Ai miei nonni Elena, Immacolata, Carmine e Pasquale** A voi, nonni, va tutta la mia gratitudine. Mi avete insegnato con il vostro esempio l'importanza della famiglia, dei valori semplici e dell'amore incondizionato. Con ognuno di voi ho imparato qualcosa di unico, e ogni ricordo che mi avete lasciato è un tassello che arricchisce la mia vita. Nonostante il passare degli anni, il vostro affetto continua a guidarmi e a farmi sentire radicato, più forte di qualsiasi distanza o tempo che ci separi. Grazie per avermi dato tanto, senza mai chiedere nulla in cambio.

**Ai miei zii** A voi, zii, voglio semplicemente dire grazie. Ci siete sempre stati, e so che ci sarete sempre, in ogni momento, in ogni passo. Non importa dove ci porterà la vita, sono certo che non ci allontaneremo mai, perché il legame che ci unisce è più forte di qualsiasi distanza o cambiamento. Siete una parte fondamentale della mia vita, e lo sarete sempre.

**Ai miei cugini** A voi, cugini, che ho visto crescere passo dopo passo, voglio dire solo una cosa: potete sempre contare su di me. Anche se il tempo passa e ci cambia, il nostro legame rimarrà sempre lo stesso. Vi sono vicino e, qualunque cosa accada, sapete che ci sarò sempre per voi.

Spesso, quando si raggiunge un traguardo come questo, si celebra la tenacia, la perseveranza, i sacrifici e le notti insonni. In effetti, c'è stato un po' di tutto questo. Ma, a essere sinceri, non è stata la mia forza a condurmi fin qui. Se dovessi davvero indicare ciò che ha fatto la differenza, direi senza esitazione che il vero motore di questo percorso è stata la **fortuna**.

Ma non la fortuna intesa come un colpo di caso, un momento fortunato o una coincidenza favorevole. Parlo di una fortuna più rara, più vera: la fortuna di aver incontrato persone straordinarie. Amici sinceri, presenti, capaci di rendere più leggere le salite e più ricchi i giorni normali.

Questa fortuna — *la mia più grande fortuna* — ha volti, nomi e storie. Sono proprio loro che voglio ringraziare, con profonda gratitudine, in queste righe.

**A Simone** Sei stato, sin dai banchi di scuola, il mio compagno più grande. Sedermi accanto a te è stata una di quelle fortune silenziose che capisci davvero solo col tempo. In te ho trovato un amico sincero, leale, presente. Abbiamo condiviso momenti che hanno segnato la nostra crescita, e spero che continueremo a crescere insieme, perché certe cose non dovrebbero finire mai. E poi, lo sai: continuo a sperare che un giorno tu faccia i milioni... e ci mantenga entrambi. Tienimi aggiornato!

**A Carmine** Di persone buone, davvero buone, ce ne sono poche. Tu sei una di queste. In mezzo a risate e snappate, mi hai aiutato a rimanere in piedi anche quando tutto sembrava troppo. C'è chi ringrazia i compagni per le lunghe ore di studio condivise, ma sappiamo bene che, nel nostro caso, sarebbe una bugia. Quello che abbiamo condiviso vale molto di più di migliaia di libri e dispense.

**A Emanuele** A un certo punto ho capito che la vera forza di un'amicizia non sta in quanto bene vuoi a qualcuno, ma in quanto riesci a volerti bene quando sei con lui. E tu, Emanuele, hai sempre avuto questo effetto raro: farmi stare bene con me stesso. Senza dirmelo, senza farmi troppe domande, senza cercare di fare nulla. Solo stando lì. E se oggi mi sento più solido, è anche perché ci sei stato tu.

**A Matteo** Con te basta uno sguardo, una parola, un riferimento assurdo, e ci ritroviamo a ridere come se fossimo ancora lì, in quella notte di anni fa. C'è qualcosa di magico in quell'intesa che non ha bisogno di spiegazioni, né di racconti: esiste e basta. Sei uno di quegli amici rari con cui ogni momento diventa una storia da ricordare, e ogni distanza si annulla al primo "oh, ti ricordi?". *"E un giorno saremo entrambi più grandi. La vita diventerà più rumorosa, più piena. Le chiamate forse si faranno più rare, ma nulla potrà mai sostituire quella sensazione. Sapere che, da qualche parte nel mondo, c'è ancora qualcuno che si ricorda esattamente chi sei."*

**Agli avventurieri** Bruno, Emanuele, Salvatore, Pasquale. Con voi ho condiviso un mondo fantastico — fatto di mappe, dadi, *Eldritch Blast*, il vecchio Jim, gnomi odiosi e palle di fuoco di 10° livello. Ma in realtà, quello che costruivamo davvero, sessione dopo sessione, era qualcosa di molto più concreto: un rifugio. Un luogo dove essere noi stessi, dove ridere senza filtri, che ogni venerdì sera diventava una piccola certezza.

E a **Bruno** — nostro DM e molto più che quello — grazie. Perché D&D è solo l'ultima delle passioni che abbiamo condiviso. Ma la tua creatività, la tua amicizia — e anche la tua lentezza — c'erano ben prima dei dadi. E so che ci saranno anche dopo.

**A tutti voi, amici miei** Con voi ho condiviso tutto: le risate che fanno male alla pancia, i ricordi che non smettono di tornare, i dolori che sembravano troppo grandi da affrontare da soli. Siamo cresciuti, cambiati, a volte forse anche un po' persi — ma mai lontani.

Dicono che nella vita si è fortunati se si trova un vero amico. Io non sono solo fortunato. Io sono *ricco*. Ricco di persone che mi fanno sentire visto, voluto bene, capito. Ricco di mani che so che ci sono, anche quando non le cerco. Ricco di legami che non chiedono niente, ma danno tutto.

So che senza di voi io non sarei nemmeno la metà di quello che sono. Non è retorica. Voi siete la mia **fortuna**.

**Ad Alessia** E infine, a te. L'amore della mia vita.

Non sei entrata nella mia vita per caso, né in un momento preciso. Sei entrata ovunque, in ogni pensiero, in ogni cosa che ho iniziato a desiderare davvero.

Con te ho capito che l'amore vero non ha bisogno di essere spiegato o messo in scena. Si costruisce giorno per giorno, nel modo in cui ci si guarda, ci si ascolta, ci si resta accanto.

Ogni cosa che faccio, ogni scelta che prendo, ogni idea che inseguo — ha dentro un po' di te. E se oggi mi sento pronto per tutto ciò che verrà, è perché ci sei tu.

Questo traguardo è mio, ma porta le tue iniziali *ovunque*.

E sì, lo so: per voi resto sempre il **Presidente**. Ma la verità è che siete voi a guidarmi ogni giorno.

*Ognuno di voi è un pezzo preciso di ciò che sono.*