Politecnico di Torino

Politecnico di Torino

Corso di Laurea Magistrale in Ingegneria Informatica

A.A. 2024/2025

Sessione di laurea Dicembre 2025

Utilizzo di Large Language Models (LLMs) per Generare Storie Interattive in Tempo Reale: Analisi, Design e Sperimentazione

Relatore:

Prof. Andrea Bottino

Candidata:

Flavia Fubini

Abstract

I modelli linguistici di grandi dimensioni (Large Language Models, LLM), come ChatGPT, rappresentano uno strumento promettente per la creazione di narrazioni interattive, in particolare di storie ramificate in cui le scelte dell'utente conducono a percorsi e risultati differenti.
Questa tesi propone di esplorare la progettazione e l'implementazione di un sistema capace di generare tali storie, scena per scena, mantenendo al contempo coerenza stilistica e strutturale lungo l'intera narrazione.

Il nucleo del progetto consiste nella progettazione e realizzazione di un prototipo di applicazione interattiva capace di generare in modo autonomo e in tempo reale una storia coerente a partire da un'introduzione prestabilita. Il sistema produce la narrazione scena per scena, offrendo all'utente tre opzioni tra cui scegliere per determinare l'evoluzione del racconto. Oltre alla scena introduttiva, all'LLM vengono fornite indicazioni relative al tono e all'atmosfera da mantenere lungo tutto lo sviluppo della storia. Ogni scena generata è inoltre accompagnata da un'illustrazione in ASCII art, anch'essa prodotta in tempo reale dal modello.

L'obiettivo principale della tesi è indagare le potenzialità dei Large Language Models (LLM) in un contesto interattivo, esplorandone l'uso come strumento capace di ampliare le possibilità di partecipazione dell'utente, superando i limiti imposti da percorsi narrativi rigidamente predefiniti. Tale indagine riguarda sia la dimensione narrativa — analizzando come il modello possa gestire la direzione e la coerenza della storia — sia quella visiva, valutando la fattibilità di una componente grafica generata in formato testuale, leggera e immediata da produrre.

Per raggiungere questi obiettivi è stato condotto un lavoro di prompt engineering, volto a guidare l'LLM nella generazione di scene che rispettino le specifiche strutturali di una narrazione completa, comprendente introduzione, conflitto e conclusione, garantendo al contempo coerenza stilistica e coinvolgimento interattivo.

Ringraziamenti

> Un primo ringraziamento va al mio relatore, Prof. Andrea Bottino, di cui ho apprezzato moltissimo la tempestività nel rispondere alle mail, l'aver valutato con serietà, e poi l'aver accolto, le mie proposte sulla direzione da prendere per il progetto di tesi e, più in generale, il fatto di non aver contribuito a rendere questa esperienza di tesi un percorso più stressante di quanto non sia già da sé.
Vorrei ringraziare i ragazzi del Lab 1 per l'accoglienza di questi ultimi mesi, in particolare Alessandro per il supporto concreto nella progettazione, implementazione e analisi finale, oltre che per avermi fornito gli strumenti necessari per realizzare questo progetto.

> Un enorme grazie ai miei tre valutatori e scafati revisori di bozze: Jossi, Raffa e Giop, ora massimi esperti di castelli in fiamme e furti di crocchette.

> A Jossi e Aki, per avermi concesso l'enorme lusso di poter studiare anche in Conservatorio e per non avermi imposto tempi che non sarei stata capace di rispettare. Grazie moumix, per avermi fornito cibo in barattolo in questi anni di Poli, grazie al quale mi sono risparmiata, almeno in parte, pasti costituiti esclusivamente da piadine in microonde con olio e sale grosso, per avermi assecondato nell'interpretazione originale del *genitore del fuorisede* facendo tu la spola tra Aosta e Torino oltre che per vederci, per farmi il cambio di stagione.
Grazie per essere stata (complessivamente, dai) il miglior supporto alla mia sopravvivenza.

> A Lori e Sandrino anche loro foraggiatori di pasti completi, marmellate d'eccezione e grandi supporter di tutti i miei percorsi

di studio.

> Alle mie Fra, per questi anni insieme, per avermi offerto una splendida amicizia su cui poter fare affidamento, per la morbidezza e il supporto emotivo enorme e per avermi preparato il cappuccino quando non ero in grado di alzarmi dal letto, senza di voi probabilmente prima o poi mi sarebbe decaduta la carriera. Gli anni della sezione Lagrange sono stati di gran lunga i più divertenti e intensi di questo percorso universitario, con menzione speciale a Gabri, grazie al quale abbiamo potuto apprezzare la superficie originale della cucina, normalmente invasa da oggetti e agglomerati di dubbia natura, oltre ad aver contribuito significativamente all'intrattenimento serale nelle cene casalinghe.

> A Mr. Dibs per tutte le sessioni di studio ed eventi random e in particolare per avermi sbattuto giù dal letto per tutto l'ultimo anno di esami, contributo senza il quale probabilmente prima o poi mi sarebbe decaduta la carriera.
Un plauso a Mr. Dibs anche per avermi fatto conoscere Erika e Sensei, persone di cui amo, tra le varie cose, la capacità di bilanciare studio e svago e senza i quali non avrei mai dato programmazione di sistema e dunque prima o poi mi sarebbe decaduta la carriera.

> A Raffa, la tua spiccata sensibilità è per me un punto di riferimento nel gestire i rapporti umani, a Vale, per esserci sempre (tranne a boogie ma ce lo facciamo andar bene lo stesso) e in generale al gruppo Suore e Suore Estese, Ale G, Jucchee, Ray, Vicky, Edo, Bea Audisio, migliori accollatori di serate e gite collettive (vedi serata powerpoint, grand prix zia Carla, varie gite ad Aosta, ecc...).

> A Lore Frahoo, scafato organizzatore di sciate e gite in montagna, grazie per coinvolgermi in tutte le attività pAzZeh insieme a Edo, Eli, Tony F e Mapi mapi.

> A Vitto per le sessioni di ~~studio~~ gossip, fondamentale nutrimento dell'anima.

> Alla best organizzatrice di pettole day, Sisbet, per tutte le

sessioni di studio e tisanina, nonostante miciotti sia ogni volta un attentato al tuo sistema respiratorio e per il tuo fondamentale aiuto nella ricerca e nella progettazione della tesi, senza di te probabilmente mi sarebbe decaduta la carriera.

> A Diana, compagna pazzesca di vacanze più o meno improvvisate su tavola (indistintamente da sci o da surf), è sempre bello ospitarti (ed essere ospitata :>).

> Alle Giu, Cri e Sabù, ora che (si spera) sarò più libera potrò finalmente venirvi a (ri)trovare.

> A tutto il gruppo boogie, in particolare grazie Oscarito per avermi debuggato il codice dell'ultimo esame (senza di te probabilmente mi sarebbe decaduta la carriera) e per essere, insieme a Ele, i miei best ballerini con cui condividere il percorso agonistico.

> Ai colleghi di progetto, in particolare Marco Cup, Kamy, Matte Biffoni, Marco Russo, Antonio e Nicolae per aver contribuito a rendere questa magistrale un percorso che tutto sommato tornassi indietro rifarei.

> Alla mia piccola mici, che figura nei ringraziamenti di almeno altre tre tesi, dunque non poteva mancare in questa.

> Infine, vorrei ringraziare Gioppelissimo, mi hai accompagnato in questo ultimo anno di disperazione preparandomi cappuccini, ottimo cibo (il king delle cheesecake), offrendo la morbidezza necessaria a non far decadere la carriera :>

Grazie famiglia e amici, vi voglio bene.

Indice

Elenco delle tabelle

Elenco delle figure

Capitolo 1

Introduzione

Contesto generale

Negli ultimi anni la combinazione tra narrazione interattiva e modelli linguistici di grandi dimensioni (Large Language Models, LLM) ha aperto nuove prospettive per lo sviluppo di applicazioni narrative in tempo reale, capaci di rispondere alle interazioni dell'utente in modo dinamico e non rigidamente predefinito. La generazione di contenuti testuali guidata da LLM ha permesso di superare i limiti delle tradizionali narrative ramificate, spesso vincolate a percorsi pre-scritti e a strutture deterministiche. Le ricerche più recenti mostrano infatti come gli LLM siano in grado di mantenere coerenza tematica su narrazioni multi-turno, di adattarsi alle scelte dell'utente e di modellare scenari complessi attraverso un'interazione continua (J. Li et al., 2024; Yuan et al., 2022). In questo contesto, la narrazione non è più un artefatto statico ma un processo generativo emergente, costruito scena dopo scena sulla base dello scambio tra utente e sistema.

Parallelamente, la letteratura ha evidenziato come l'impiego di LLM nei contesti di storytelling interattivo apra la strada a nuove forme di co-creazione uomo–macchina. Sistemi come AI Dungeon (Latitude, Inc., 2019) o InstructGPT-based storytellers (Ouyang et al., 2022) hanno dimostrato la capacità dei modelli di produrre testi creativi e contestualmente pertinenti in risposta a input destrutturati, rafforzando l'interesse per applicazioni narrative adattive e personalizzate. Più di recente, il lavoro di C. Wang, Lin et al., 2023 con Voyager e agenti autonomi per ambienti simulati ha messo in luce il potenziale degli LLM come agenti persistenti capaci di costruire conoscenza e generare comportamenti complessi nel tempo, un risultato che ha ulteriormente consolidato l'idea degli LLM come strumenti per esperienze narrative non lineari.

Questo quadro rende evidente come la convergenza tra scienze del linguaggio computazionale, tecniche di generazione automatica e pratiche di storytelling

digitale stia delineando un campo di ricerca emergente, in cui le narrazioni possono evolvere in modo non previsto e ogni interazione costituisce un'occasione per modellare nuovi scenari, eventi e percorsi.

Motivazioni

L'integrazione di elementi di interattività e gamification consente di ottenere forme di fruizione di contenuti più efficaci e coinvolgenti rispetto ai metodi tradizionali, come lezioni frontali o consultazione autonoma di documentazione (Hamari et al., 2014). Attraverso il coinvolgimento diretto dell'utente, l'apprendimento diventa più attivo, immersivo e personalizzabile. Tuttavia, le applicazioni interattive non generative presentano un limite strutturale: ogni percorso, reazione o contenuto deve essere progettato in anticipo, lasciando all'utente una libertà limitata. Questo approccio garantisce un'esperienza stabile e ben definita, ma comporta anche che, a ogni utilizzo, la dinamica dell'interazione rimanga sostanzialmente invariata, riducendo la varietà e l'adattabilità nel lungo periodo.

Un sistema interattivo generativo real-time è un sistema che combina tecniche di generazione automatica con un'interfaccia che permette all'utente di modificare o influenzare il comportamento del sistema mentre esso è in esecuzione. Lo sviluppo di questo tipo di sistemi risponde a diverse esigenze nel campo della narrazione digitale contemporanea. Da un lato, consente di implementare esperienze narrative personalizzate, in cui la storia si modella in base alle scelte e alle azioni dell'utente, offrendo percorsi narrativi unici a ogni interazione. Dall'altro, si tratta di un laboratorio per sperimentare nuove forme di storytelling, in cui la struttura della narrazione non è più rigidamente definita a priori, ma emerge dinamicamente dal dialogo tra utente e sistema.

In questo contesto, i modelli linguistici di grandi dimensioni (LLM) non sono utilizzati come meri generatori di testo, ma come agenti adattivi in grado di ampliare il contesto narrativo in tempo reale, mantenendo coerenza e creatività su sequenze di brevi scene o episodi (J. Li et al., 2024; Latitude, Inc., 2019).

L'obiettivo non è sostituirsi alla creatività umana, ma offrire nuove possibilità di espressione e interazione in diversi contesti applicativi. In ambito ludico o di intrattenimento, ciò può tradursi nella generazione di storie che evolvono dinamicamente in base alle scelte dell'utente, rendendolo parte attiva e protagonista dell'esperienza narrativa. In contesti educativi o divulgativi, invece, un sistema basato su LLM può sfruttare un database di elementi storici, ambientazioni e personaggi accuratamente documentati per generare narrazioni coerenti e plausibili, capaci di favorire l'immersione e l'apprendimento attraverso la partecipazione diretta.

Questo progetto di tesi è stato pensato come un primo prototipo di un'estensione del progetto CHANGES per il Museo Egizio di Torino (Mensa et al., in press),

dove l'LLM è impiegato come strumento di supporto alla creazione narrativa, integrandosi con l'expertise curatoriale. In questo approccio, dei curatori definiscono la struttura complessiva della narrazione e forniscono descrizioni strutturate delle singole scene; l'LLM interviene successivamente per trasformare tali descrizioni in testi coerenti e stilisticamente rifiniti. Questo processo consente di combinare la conoscenza specialistica dei curatori con le capacità linguistiche avanzate del modello, garantendo al contempo la correttezza storica e la coerenza tematica delle narrazioni. L'utilizzo del modello in un contesto controllato e selettivo permette quindi di valorizzare l'efficienza e la qualità espressiva dei testi, senza compromettere gli standard scientifici richiesti nei contesti di heritage culturale. A partire da questo progetto, la tesi vuole esplorare una forma di maggiore libertà narrativa, tramite l'impiego di un LLM, che non si basi più su un grafo di scene predeterminato, ma generi la storia affidandosi esclusivamente al contesto fornito e al testo prodotto fino a quel momento. In questo scenario, le narrazioni evolvono in maniera unica e imprevedibile a ogni interazione.

Obiettivi e contributi

L'obiettivo generale di questa tesi è l'esplorazione delle potenzialità degli LLM nella generazione di narrazioni interattive, con particolare attenzione alle storie ramificate in cui le scelte dell'utente influenzano in tempo reale l'evoluzione del racconto. Oltre al ruolo di supporto alla generazione testuale, i modelli sono stati utilizzati anche come strumenti di valutazione autonoma (*LLM as judge*, H. Li et al., 2024; Chiang e Lee, 2023), in grado di analizzare le narrazioni prodotte, verificare la coerenza interna delle storie e l'effettivo rispetto delle scelte effettuate dall'utente.

Il lavoro di ricerca si articola in quattro obiettivi principali:

1. **Progettazione e implementazione di un sistema interattivo generativo:**
   Sviluppare un prototipo di applicazione in grado di generare in tempo reale una storia coerente a partire da un'introduzione prestabilita, gestendo la progressione scena per scena. Il sistema deve consentire all'utente di influenzare lo sviluppo narrativo tramite scelte multiple, simulando la struttura delle "storie ramificate" tipiche delle narrative interattive.

2. **Analisi delle capacità narrative degli LLM:**
   Indagare come i modelli linguistici riescano a mantenere coerenza stilistica, continuità tematica e plausibilità narrativa durante la generazione iterativa di brevi scene, verificando in che misura siano in grado di gestire elementi come tono, atmosfera, coesione tra eventi e il rispetto delle scelte dell'utente.

3. **LLM as Judge**:
   Sperimentare l'impiego degli LLM come valutatori autonomi delle narrazioni generate, come supporto alla valutazione umana.

4. **Esplorazione della componente visiva e multimodale:**
   Valutare la fattibilità di integrare una componente visiva leggera e immediata, realizzata mediante illustrazioni in ASCII art, generate dallo stesso modello, per arricchire l'esperienza dell'utente e rafforzare il carattere immersivo dell'interazione.

Per conseguire questi obiettivi, è stato condotto un lavoro di prompt engineering finalizzato a definire strategie efficaci per guidare l'LLM nella generazione controllata dei contenuti, garantendo una struttura narrativa completa (introduzione, conflitto, sviluppo e conclusione) e un livello adeguato di coerenza tra le scene. Inoltre, è stato creato e valutato un dataset di 300 storie, in cui l'interazione con l'utente è stata simulata tramite una funzione automatica che seleziona casualmente una delle opzioni disponibili per la continuazione del racconto. Le storie generate sono state sottoposte sia a valutazione umana che al giudizio degli LLM, permettendo di analizzare la qualità narrativa e la coerenza interna dei testi e di testare l'efficacia dei modelli nel gestire scenari interattivi complessi.

Di seguito vengono elencati i principali contributi di questa ricerca:

- Analisi delle strategie di *prompting* per la generazione narrativa e per la valutazione automatica.
- Progettazione di un'architettura software per applicazioni narrative che seguono l'andamento di un racconto (introduzione, conflitto, sviluppo e conclusione).
- Implementazione di un prototipo minimale che integra generazione testuale e gestione interattiva delle scelte.

Struttura della tesi

La tesi è organizzata come segue:

- **Capitolo 2 – Stato dell'arte:** analisi dei LLM applicati alla narrazione, tecniche di gestione visuale, architetture real-time, metriche di valutazione e casi d'uso.
- **Capitolo 3 – Metodologia e design del prototipo:** descrizione delle scelte progettuali e delle strategie di prompting.
- **Capitolo 4 – Implementazione:** stack tecnologico, descrizione del prototipo, gestione del prompt e della rappresentazione visiva.

- **Capitolo 5 – Sperimentazione e risultati:** setup dei test, metriche, risultati e analisi critica.
- **Capitolo 6 – Conclusioni e sviluppi futuri:** sintesi dei risultati, contributi della tesi e possibili estensioni.

Capitolo 2

Stato dell'arte

2.1 Introduzione

Negli ultimi anni, l'evoluzione dei modelli linguistici di grandi dimensioni (Large Language Models, LLM) ha reso possibile l'implementazione di nuove forme di narrazione interattiva, in cui la storia viene costruita in tempo reale attraverso l'interazione con l'utente. Queste esperienze, spesso basate su strutture a bivi o sistemi di generazione continua, uniscono elementi di storytelling tradizionale, intelligenza artificiale generativa e interazione uomo-macchina.

In questo capitolo vengono esaminati gli studi e le tecnologie più rilevanti riguardanti la gestione della narrazione generata da modelli linguistici, le soluzioni visuali di supporto e le architetture che permettono un funzionamento *real-time* delle applicazioni interattive.

2.2 Gestione e qualità della narrazione

2.2.1 Prompt engineering

Con *prompt engineering* si intende l'insieme di tecniche e strategie volte a progettare e formulare in modo efficace le istruzioni testuali fornite a un modello linguistico di grandi dimensioni, al fine di ottenere output coerenti con gli obiettivi desiderati. Tale pratica consiste nel definire con precisione il compito del modello, il contesto informativo, il tono, lo stile e il formato dell'output atteso, al fine di orientarne il comportamento generativo. L'efficacia di un prompt dipende dalla chiarezza e dalla specificità delle istruzioni; prompt ambigui o imprecisi possono infatti produrre risultati incoerenti o non pertinenti.
Il prompt engineering assume quindi un ruolo centrale sia nel controllo della qualità e della struttura dei testi prodotti, sia nella regolazione di aspetti stilistici o

narrativi, come il registro linguistico, la lunghezza, la coesione interna o la presenza di dettagli specifici. Il processo richiede generalmente un approccio iterativo, caratterizzato da sperimentazione e affinamento continuo dei prompt, al fine di massimizzare le prestazioni del modello senza intervenire direttamente sulla sua architettura interna. In questo senso, il prompt engineering può essere considerato come un'interfaccia strategica tra l'utente umano e il modello, finalizzata a tradurre esigenze e vincoli concettuali in istruzioni operative efficaci e, in questo contesto, costituisce una delle componenti chiave nella qualità della narrazione generata da LLM, in quanto la capacità di guidare il modello attraverso istruzioni testuali precise influisce direttamente sulla coerenza narrativa, sul tono stilistico e sulla capacità di mantenere un filo logico nel tempo. In questa sezione vengono analizzate le principali strategie di prompting per applicazioni narrative, con riferimento alle ricerche più recenti in letteratura.

Sahoo et al., 2025 offrono una panoramica sistematica delle tecniche di prompt engineering in base alle loro applicazioni pratiche o scenari d'uso. Gli autori individuano dodici ambiti distinti, tra i quali tre risultano particolarmente rilevanti per il presente lavoro: **(1)** riduzione delle allucinazioni, **(2)** miglioramento della coerenza e consistenza e **(3)** gestione delle emozioni e del tono.

1. Reduce Hallucination Con il termine *hallucination* si fa riferimento alla generazione da parte del LLM di informazioni inventate o non supportate dal contesto fornito. Per affrontare tale problema, Sahoo et al., 2025 individuano cinque approcci o architetture principali.

- **Retrieval-Augmented Generation (RAG)** Lewis et al., 2020: combina memoria parametrica e non parametrica pre-addestrata per la generazione testuale. Questo processo consente di arricchire l'output di un LLM con informazioni esterne, garantendo maggiore coerenza e aderenza ai dati di partenza — un aspetto utile anche nel contesto narrativo, dove la coerenza tra gli eventi è cruciale.

- **ReAct Prompting** Yao et al., 2022: integra le fasi di *reasoning* e *action* in un unico processo, permettendo al modello di indurre, tracciare e aggiornare piani d'azione durante l'elaborazione di risposte complesse. Ciò consente un maggiore controllo sullo sviluppo logico della narrazione.

- **Chain-of-Verification (CoVe) Prompting** Dhuliawala et al., 2023: prevede un ciclo di quattro fasi di verifica che mirano ad assicurare la correttezza e la giustificazione delle informazioni generate dal modello.

- **Chain-of-Note (CoN) Prompting** Yu et al., 2023: utilizza note intermedie o annotazioni per guidare il processo di ragionamento del modello. In ambito

narrativo, questa tecnica può supportare l'organizzazione del flusso di idee e trame, favorendo una costruzione coerente della storia.

- **Chain-of-Knowledge (CoK) Prompting** X. Li et al., 2023: si concentra sull'integrazione dinamica di conoscenze eterogenee e di fonti multiple nel processo di ragionamento, garantendo risultati più accurati e informati.

2. Improving Consistency and Coherence Un secondo filone individuato da Sahoo et al., 2025 riguarda la coerenza e la consistenza testuale, aspetti centrali nella generazione narrativa. In questo contesto, la tecnica di **Contrastive Chain-of-Thought (CCoT) Prompting** Chia et al., 2023 prevede di fornire al modello esempi di ragionamento sia corretti sia errati. Tale confronto aiuta l'LLM a riconoscere e generare sequenze di pensiero più coerenti, un meccanismo potenzialmente utile per valutare diverse opzioni narrative e le loro conseguenze sulla trama e sui personaggi.

3. Managing Emotions and Tone La gestione delle emozioni e del tono rappresenta un altro aspetto rilevante nella produzione di testi narrativi. La tecnica di **Emotion Prompting** C. Li et al., 2023 introduce specifici stimoli emotivi all'interno del prompt, ispirandosi a modelli psicologici sul linguaggio e sull'intelligenza emotiva. L'aggiunta di undici frasi di stimolo emotivo ha mostrato un miglioramento significativo nella capacità del modello di adattare tono e stile in diversi contesti comunicativi.

Formati di prompt e impatto sulla narrazione Harmon e Rutman, 2023 esplorano come differenti formati di prompt influenzino la generazione di scelte narrative da parte di LLM open-source. Gli autori valutano le prestazioni di tre modelli basati su LLaMA confrontando tre diverse formulazioni di prompt *zero-shot*, ciascuna contenente la trama di una storia, una decisione del personaggio e la richiesta di proseguire la narrazione coerentemente con le conseguenze di tale scelta. I tre formati analizzati sono:

1. *simple colon*, in cui la richiesta termina semplicemente con due punti;
2. *masterful storyteller*, che esplicita la richiesta di proseguire la storia assumendo il ruolo di un "abile narratore";
3. *answer interrogative*, che presenta la richiesta in forma di domanda seguita dalla dicitura "Answer:".

I risultati indicano che il formato *answer interrogative* produce le risposte più coerenti e adeguate per tutti i modelli analizzati.

In sintesi, la letteratura più recente mostra come il prompt engineering non si limiti a una pratica di formulazione linguistica, ma costituisca una strategia strutturale per modulare il comportamento dei LLM, influenzando la coerenza, la creatività e la qualità emotiva dei testi generati.

2.2.2 Approcci strutturati e riflessivi alla generazione narrativa automatica

Con l'evoluzione tecnologica, la generazione automatica di storie mediante LLM si è progressivamente spostata da approcci *end-to-end*, basati su semplici prompt descrittivi, verso metodologie che introducono **livelli intermedi di rappresentazione narrativa**. Questa transizione riflette un cambiamento di prospettiva: la narrazione non viene più considerata come una sequenza testuale generata in un singolo passaggio, ma come un **processo strutturato**, composto da pianificazione, controllo semantico e revisione riflessiva. I lavori più recenti esplorano in modo crescente la possibilità di modellare e manipolare in modo esplicito gli elementi costitutivi della narrazione — eventi, azioni, morali, valori o schemi di critica — per ottenere storie più coerenti, significative e creative.

Un primo gruppo di ricerche pone l'accento sulla **struttura narrativa interna**. Il lavoro di Hatzel e Biemann, 2024 propone la costruzione di *story embeddings*, ossia rappresentazioni vettoriali delle storie che catturano la loro struttura narrativa, indipendentemente dallo stile linguistico o dai dettagli superficiali. L'obiettivo è permettere ai modelli di distinguere tra storie simili dal punto di vista della *fabula* — la sequenza logico-temporale degli eventi — e storie diverse per struttura o sviluppo. Questa prospettiva quantitativa sulla forma narrativa è affiancata, in *SWAG – Storytelling With Action Guidance* Pei et al., 2024, da un approccio generativo che introduce una **guida d'azione**: la storia viene pianificata come sequenza di eventi chiave che orientano la generazione del testo, garantendo una progressione dinamica e coerente. In entrambi i casi, la narrazione non è più un output spontaneo del modello, ma il risultato di una **pianificazione narrativa controllata**, capace di integrare coerenza e ritmo narrativo.

Un secondo filone di studi amplia il livello di analisi dal piano strutturale a quello **semantico-valoriale**. Hobson et al., 2024 presentano il framework *Story Morals*, in cui i modelli di linguaggio vengono utilizzati per identificare le *morali* e i valori impliciti veicolati dalle storie. Questo tipo di analisi consente di esplorare la dimensione etica e culturale della narrazione, rivelando come i racconti riflettano schemi di valori condivisi o divergenti. Nel contesto della narrativa generativa, la possibilità di estrarre e manipolare la morale di una storia apre prospettive interessanti per la **valutazione qualitativa dei contenuti generati**,

permettendo di verificare se la produzione automatica trasmetta i significati o i messaggi desiderati.

Un terzo gruppo di lavori si concentra invece sulla **riflessività del processo narrativo**, introducendo meccanismi di autovalutazione e revisione iterativa. Nel framework *Collective Critics* Bae e Kim, 2024 la generazione avviene attraverso un sistema multi-agente: una serie di "critici" valuta e commenta la storia prodotta, mentre un "leader" sintetizza il feedback e guida la riscrittura. Il risultato è un ciclo iterativo che migliora creatività, coerenza e qualità stilistica del racconto. Un principio analogo è alla base di *DataNarrative* Islam et al., 2024, che affronta lo *storytelling* basato su dati: due agenti cooperano per generare testi e visualizzazioni coerenti con le evidenze numeriche, alternando fasi di riflessione e verifica. Entrambi i lavori propongono una concezione **meta-narrativa** della generazione, in cui il modello non solo scrive, ma *pensa su ciò che scrive*, introducendo un livello di controllo simile alla revisione umana.

Nel complesso, questi studi convergono verso una visione della narrativa generativa come **processo strutturato, semantico e riflessivo**, in cui la creatività emerge dall'interazione di diversi livelli di controllo. L'azione, la struttura, la morale e la critica diventano dimensioni esplicitamente modellabili, che consentono di guidare la generazione automatica non soltanto sul piano linguistico, ma anche su quello narrativo e valoriale. In questa prospettiva, la ricerca recente mostra come l'efficacia della narrativa generativa non dipenda esclusivamente dalla potenza del modello, ma anche dalla **progettazione del processo**: la definizione di rappresentazioni intermedie, di strategie di revisione e di criteri di coerenza che trasformano la produzione testuale in un atto narrativo consapevole. È su questo terreno — quello della strutturazione e del controllo narrativo tramite *prompting* e rappresentazioni intermedie — che si colloca il presente lavoro di tesi.

2.2.3 Studi sulla qualità della narrazione

La valutazione della qualità narrativa nelle produzioni generate da LLM è un tema emergente. Le metriche utilizzate comprendono sia indicatori quantitativi (lunghezza, varietà lessicale, coerenza semantica) sia valutazioni qualitative (coinvolgimento del lettore, plausibilità, creatività). Particolare rilievo assumono le metriche derivate da studi di *computational creativity* e di *interactive storytelling*, come la coerenza tra scelte narrative e contesto o la consistenza dei personaggi nel tempo.

Gómez-Rodríguez e Williams, 2023 conducono un'analisi comparativa sull'efficacia dei principali LLM nella scrittura creativa. Utilizzando un approccio *zero-shot*,

chiedono ai modelli di generare una storia senza alcuna ottimizzazione tramite *prompt engineering*, apprendimento in-context o altre tecniche accessorie. I risultati mostrano che i modelli commerciali più recenti — in particolare GPT-4 e Claude — ottengono valutazioni paragonabili o superiori a quelle di scrittori umani in diverse categorie narrative, mentre i modelli open-source mostrano prestazioni significativamente inferiori. Sebbene il campione sperimentale non consenta conclusioni definitive, lo studio evidenzia come i LLM di ultima generazione possano raggiungere livelli qualitativi elevati anche senza tecniche avanzate di prompting, ponendo le basi per ulteriori indagini sul rapporto tra istruzioni testuali e creatività automatizzata.

Un ambito strettamente collegato alla valutazione della qualità narrativa riguarda la **creatività nei modelli di linguaggio**. Da un lato, Chakrabarty et al., 2024 analizzano in modo empirico le capacità creative dei LLM attraverso una versione adattata del *Torrance Test of Creative Thinking* (TTCW), mostrando che i modelli, pur essendo in grado di produrre testi fluenti e coerenti, superano solo una minoranza delle prove rispetto agli autori umani. Ciò mette in evidenza un divario significativo tra qualità stilistica e creatività autentica, accentuato dal fatto che gli stessi LLM non risultano valutatori affidabili della creatività, poiché le loro valutazioni non correlano con quelle degli esperti.

Dall'altro lato, Franceschelli e Musolesi, 2025 offrono una cornice teorica più ampia, confrontando i LLM con i principali modelli della creatività — come le categorie di Boden e il framework dei "4P" — per valutare in quale misura tali sistemi possano essere considerati agenti creativi. Gli autori rilevano che i LLM soddisfano parzialmente i criteri di *novità* e *sorpresa*, producendo contenuti che si discostano dai dati di addestramento, ma presentano limiti marcati nella dimensione del *valore*, che richiede consapevolezza del contesto, intenzionalità e capacità trasformativa del dominio creativo.

Nel complesso, questi contributi suggeriscono che la qualità narrativa non può essere misurata soltanto in termini di coerenza o fluidità, ma deve includere anche la **rilevanza creativa** e la **significatività estetica** dei contenuti generati. Di conseguenza, la produzione automatica va interpretata come un processo di co-creazione uomo–macchina, in cui l'apporto del modello rimane potente ma ancora parziale rispetto alle dimensioni profonde della creatività umana.

2.3 Gestione visuale della narrazione

2.3.1 Tecnologie immersive e grafica in tempo reale

L'integrazione tra LLM e tecnologie immersive ha aperto nuove prospettive per la generazione e la gestione di contenuti grafici in tempo reale.

Una delle direzioni della ricerca contemporanea coinvolge l'uso combinato di modelli linguistici e motori grafici tridimensionali, consentendo la creazione dinamica di ambienti interattivi e la modificazione di elementi visivi sulla base di istruzioni espresse in linguaggio naturale, riducendo drasticamente la distanza tra l'intenzione dell'utente e la rappresentazione visiva.

Tang et al., 2025 offrono una rassegna sistematica dell'integrazione tra LLM e tecnologie di *Extended Reality* (XR), evidenziando come tali modelli possano amplificare la capacità immersiva e interattiva delle esperienze virtuali. Gli autori identificano diversi paradigmi di interazione LLM–XR, tra cui la generazione di ambienti guidata da linguaggio, la modellazione dell'utente e la mediazione narrativa in contesti tridimensionali. In questa prospettiva, la grafica in tempo reale non è più un mero supporto visivo, ma un componente semantico che si evolve in funzione delle risposte del modello linguistico, integrando dimensioni spaziali e narrative in un unico spazio cognitivo condiviso.

In modo complementare, De La Torre et al., 2024 propongono *LLMR* (Large Language Model for Mixed Reality), un framework che utilizza LLM per controllare in tempo reale ambienti tridimensionali interattivi. Il sistema introduce una pipeline modulare in cui diversi agenti, istruiti tramite *meta-prompt*, generano e modificano elementi grafici e comportamentali all'interno di un motore 3D (Unity), traducendo le istruzioni testuali in codice eseguibile e verificato in fase di runtime. I risultati sperimentali mostrano una riduzione significativa del tasso di errore nella generazione automatica di scene, nonché un incremento nella percezione di immersione e controllo da parte degli utenti.

Questi contributi dimostrano come la gestione grafica in tempo reale mediata da LLM costituisca una frontiera cruciale per la progettazione di sistemi generativi immersivi. Nel contesto della presente ricerca, tale approccio è rilevante poiché consente al modello linguistico non solo di produrre contenuti testuali o narrativi, ma anche di orchestrare la rappresentazione visiva in modo coerente con il flusso semantico dell'interazione. Tuttavia, si vuole indagare un approccio in cui la gestione della rappresentazione grafica sia computazionalmente poco onerosa e veloce da generare anche per modelli piccoli e/o open-weight.

2.3.2 Stile visivo minimalista

Nel contesto della tesi, l'attenzione è rivolta a modalità di visualizzazione testuale, come l'ASCII art o le rappresentazioni semigrafiche, che garantiscono leggerezza e immediatezza. Questo stile visivo minimalista offre un notevole vantaggio: riduce l'ingombro visivo e soprattutto computazionale, rendendolo ideale per prototipi leggeri o ambienti con risorse limitate. Contestualmente, concentra l'attenzione dell'utente sull'aspetto narrativo, stimolando l'immaginazione a colmare ciò che è soltanto suggerito visivamente.

```
 /\_/\
( o.o )
 > ^ <
```

Figura 2.1: Esempio di ASCII art: gatto.

ASCII art e narrazione digitale. Con *ASCII art* si intente un mezzo artistico digitale in cui le immagini o i disegni vengono realizzati usando solo i caratteri della tabella ASCII, cioè lettere, numeri, simboli e punteggiatura presenti in un normale set di caratteri del computer.

L'impiego di rappresentazioni testuali come supporto narrativo ha una lunga tradizione. Fin dagli anni Ottanta, titoli come *Rogue* (Toy e Wichman, 1980) e i suoi successori *NetHack* e *Angband* hanno definito il genere *roguelike*, basato su interfacce composte esclusivamente da caratteri ASCII. In questi casi, la limitazione grafica si è trasformata in una forza espressiva: ogni simbolo, pur minimale, assumeva valore semantico (un "@" rappresentava il protagonista, un "D" un drago, ecc.), e l'immaginazione del giocatore suppliva all'assenza di immagini realistiche.

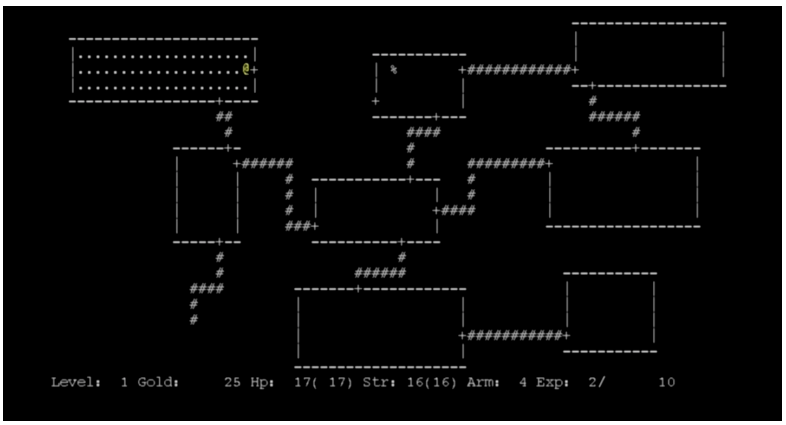

Figura 2.2: Gameplay Rogue.

In anni più recenti, questo approccio è stato ripreso e reinterpretato in chiave estetica o concettuale. *Stone Story RPG* (Martian Rex, 2019) utilizza animazioni composte interamente da caratteri ASCII, dimostrando come un linguaggio visivo minimale possa risultare fluido, dinamico e funzionale a un'esperienza narrativa profonda.

Allo stesso modo, *SanctuaryRPG* (Games, 2014) adotta un'estetica ASCII per costruire un mondo coerente e complesso, dove l'astrazione visiva diventa parte integrante dell'identità ludica.

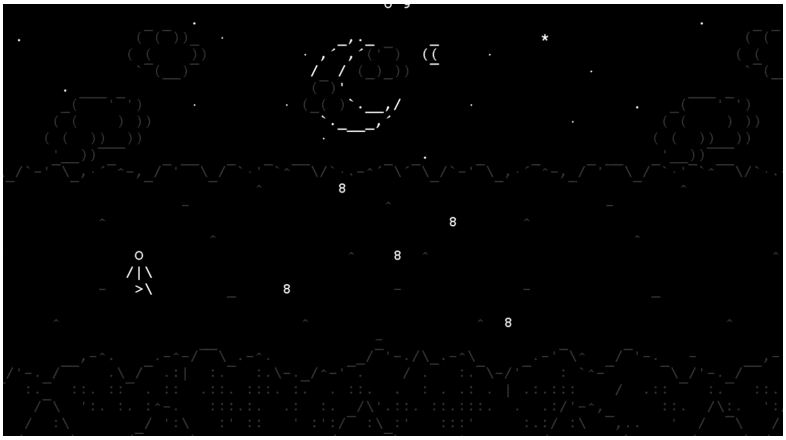

Figura 2.3: Gameplay Stone Story RPG.

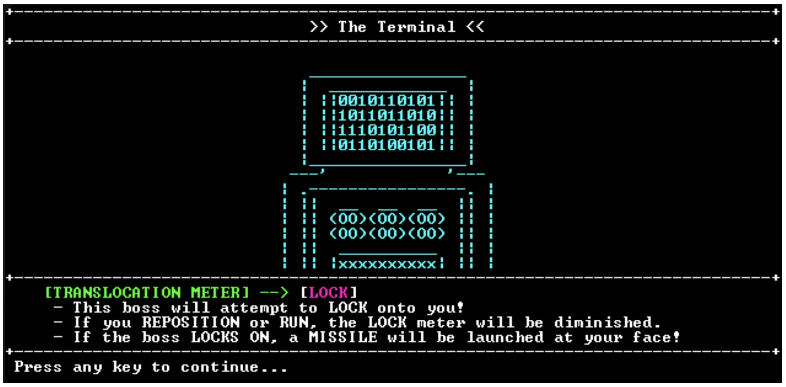

Figura 2.4: Gameplay Sanctuary RPG.

Sperimentazioni contemporanee. L'ASCII art è oggi impiegata anche in progetti indipendenti e sperimentali che esplorano la dimensione narrativa e psicologica dell'interattività. *Being Me* (Lin, 2021) è un esempio in cui l'uso di elementi grafici semplificati rafforza la dimensione introspettiva e la relazione con il testo.

In *ASCIIDENT* (Tsyganov, 2019), la scelta estetica diventa un esercizio di stile: la rappresentazione testuale viene utilizzata per generare atmosfera e tensione narrativa in un universo fantascientifico.

Un ulteriore esempio significativo è *Candy Box!* (Aniwey, 2013), un browser game basato su logiche incrementali, in cui la rappresentazione testuale evolve insieme alla trama. In questo caso, l'ASCII art non funge soltanto da decorazione, ma da interfaccia narrativa che accompagna il giocatore nel percorso di scoperta.

Riflessioni critiche. Dall'analisi di questi casi emerge come l'ASCII art e, più in generale, la grafica testuale, possano essere versatili e adatti anche a risorse

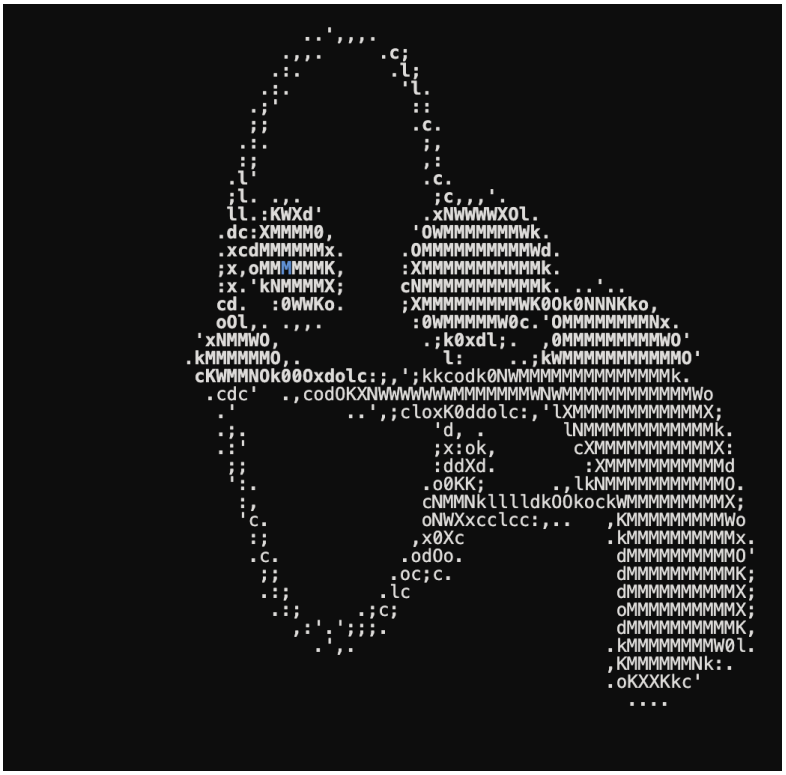

Figura 2.5: Gameplay Being Me.

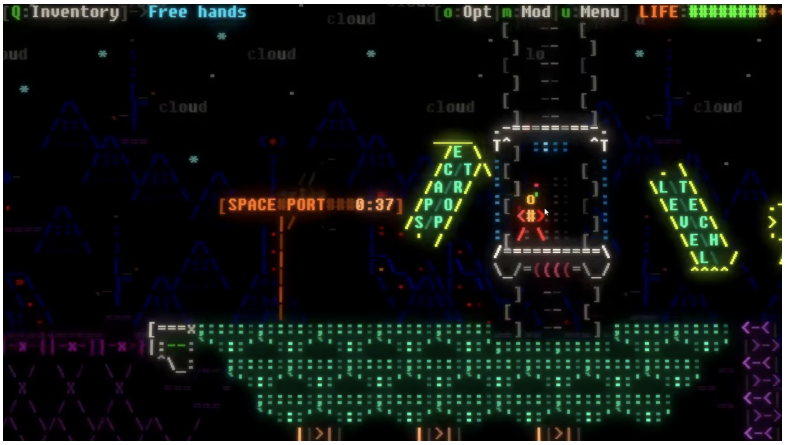

Figura 2.6: Gameplay ASCIINDENT.

limitate e comunque costituire un potente strumento di sintesi narrativa.

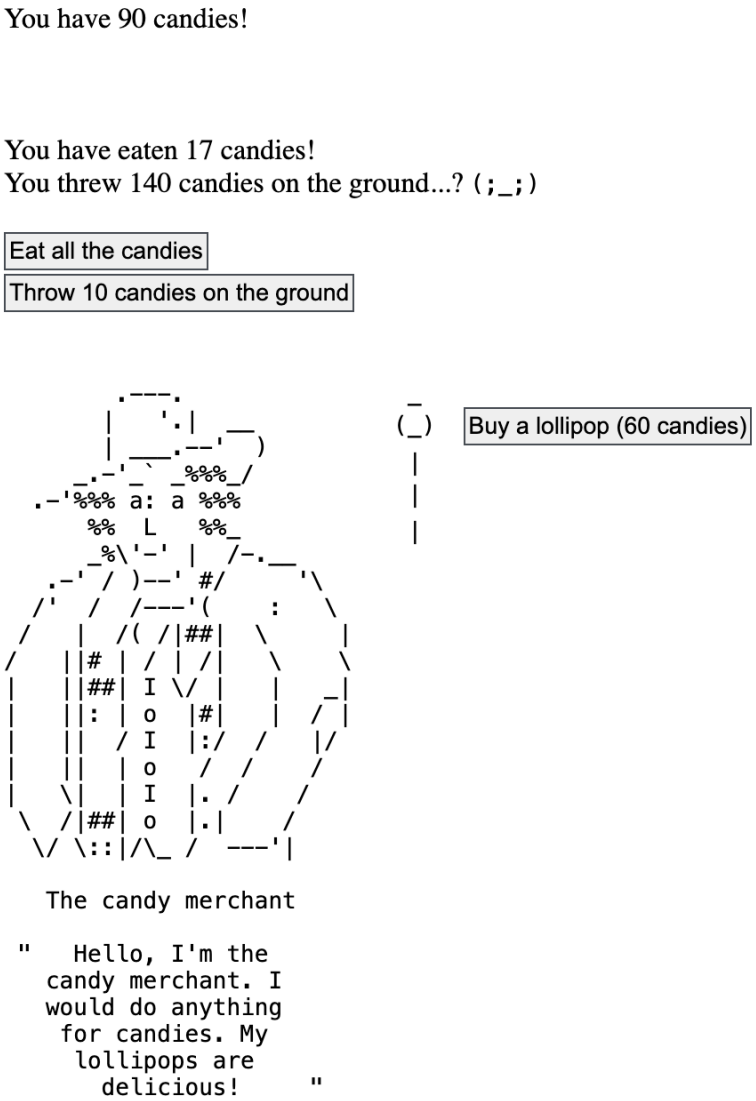

Figura 2.7: Gameplay Candy Box!.

2.3.3 Limiti dei modelli linguistici nella rappresentazione visiva testuale

Diversi studi recenti hanno indagato la capacità dei modelli linguistici e visione–linguaggio di comprendere o generare rappresentazioni che possiedono una componente visiva implicita, come nel caso dell'*ASCII art*. In particolare, lavori quali Bayani, 2024 e Luo et al., 2025 mostrano che, sebbene i LLM siano in grado di produrre e interpretare schemi in ASCII art con una certa coerenza, tali abilità

restano superficiali e non sufficientemenete robuste.

Nel contesto dei test di riconoscimento e trasformazione di diagrammi ASCII, i modelli mostrano buone prestazioni solo in compiti semplici (traslazioni, spostamenti), ma falliscono rapidamente in presenza di variazioni più complesse come rotazioni o ridimensionamenti. Ciò suggerisce che la competenza visiva dei LLM rimane ancorata a pattern puramente testuali piuttosto che a una reale comprensione spaziale della struttura grafica.

L'introduzione di benchmark specifici, come ASCIIBench Luo et al., 2025, ha evidenziato come le rappresentazioni interne (embedding) generate dai modelli siano spesso inadeguate per catturare la somiglianza visiva tra diverse opere ASCII. Il principale collo di bottiglia non risiede nella capacità generativa, bensì nella fase di rappresentazione: i vettori prodotti dai modelli non riflettono efficacemente le relazioni spaziali e morfologiche tra gli elementi della scena. Questo implica che due disegni visivamente simili possano risultare semanticamente distanti nello spazio latente del modello.

Infine, il lavoro di Z. Wang et al., 2025 dimostra che, nei modelli multimodali contemporanei, la componente testuale tende a dominare quella visiva. Quando un input contiene informazioni testuali e visive potenzialmente in conflitto — come accade nelle immagini o nei disegni ASCII che veicolano semantiche contrastanti — il modello privilegia quasi sempre il testo, trascurando la forma visiva. Questo bias testuale costituisce un limite strutturale nella reale integrazione multimodale.

Nel quadro del presente progetto, tali risultati assumono particolare rilievo: la generazione in tempo reale di ASCII art, come estensione della narrazione, impiega una modalità intermedia tra testo e immagine, ma poggia su capacità che i modelli esistenti dimostrano di possedere solo parzialmente. L'obiettivo è pertanto duplice: da un lato, utilizzare la leggerezza e la natura puramente testuale dell'ASCII art come strumento accessibile anche a modelli di piccole dimensioni; dall'altro, verificare sperimentalmente se e in che misura un LLM possa mantenere coerenza semantica e spaziale tra narrazione e rappresentazione visiva, pur in presenza dei limiti evidenziati dalla letteratura.

2.4 Applicazioni interattive real time

2.4.1 Architetture e gestione del tempo reale

Le applicazioni narrative *real-time* richiedono architetture in grado di gestire lo stato della storia e le scelte dell'utente in modo continuo e coerente. Si distinguono generalmente due approcci:

- **Architetture monolitiche**: un singolo modello gestisce input e output in un unico passaggio.

- **Architetture modulari o a pipeline**: la generazione narrativa è suddivisa in più fasi, come pianificazione, scrittura del testo e validazione.

Un esempio di architettura modulare è proposto nel progetto *GENEVA* Leandro et al., 2024, in cui un LLM è impiegato in una pipeline a due passi: in un primo momento per generare un grafo narrativo, basato su vincoli forniti dal designer, e successivamente per visualizzare e validare tale grafo. Questa separazione tra pianificazione e generazione testuale consente di mantenere un maggiore controllo autoriale e di verificare la coerenza delle ramificazioni prodotte.

Nel contesto della generazione continua, *Unbounded* J. Li et al., 2024 esplora invece un'architettura completamente *real-time* e multimodale, in cui il mondo di gioco, le meccaniche e la narrazione vengono generati dinamicamente da modelli generativi specializzati. Il sistema adotta un approccio a componenti: un LLM distillato (ossia una versione più ridotta, ma efficiente) per la generazione semantica e un modulo adattatore per la parte visiva, consentendo una simulazione aperta in cui ogni interazione modifica in tempo reale lo stato narrativo e ambientale. Questo tipo di architettura evidenzia l'importanza di moduli di controllo e sincronizzazione per garantire coerenza tra le dimensioni testuali e visive della storia.

2.4.2 Memoria e persistenza (episodica, semantica, RAG)

Un aspetto cruciale per la coerenza a lungo termine delle narrazioni generate è la gestione della memoria.

- **Memoria episodica**: conserva informazioni sugli eventi accaduti durante l'interazione.
- **Memoria semantica**: contiene conoscenze generali e regole del mondo narrativo.
- **RAG (Retrieval-Augmented Generation)**: tecnica già esaminata nella sezione di prompt engineering, che consente di arricchire il contesto del modello con informazioni recuperate dinamicamente da un archivio esterno.

Queste soluzioni permettono di superare i limiti di contesto dei modelli e di mantenere la coerenza anche in storie prolungate o ramificate.

Esperienze pratiche come *AI Dungeon* Latitude, Inc., 2019 mostrano come l'assenza di un sistema di memoria strutturato possa portare a fenomeni di *deriva narrativa*, incoerenza tra eventi e perdita di continuità nel lungo periodo. Il gioco, basato inizialmente su GPT-2 e poi evoluto verso modelli più grandi, ha dimostrato la potenza immersiva della generazione aperta, ma anche la necessità di meccanismi di persistenza dello stato e di moderazione dei contenuti. L'introduzione di tecniche di *retrieval* o memorie episodiche permetterebbe di migliorare la coerenza e di

gestire meglio il contesto multi-turn, aspetto centrale nelle narrazioni interattive di lunga durata.

Nel caso di sistemi più complessi come *Unbounded*, la memoria gioca anche un ruolo semantico: mantiene le regole del mondo generato e assicura la consistenza tra le diverse modalità (testo, immagine, stato simulativo). La combinazione di memorie episodiche e approcci RAG risulta dunque fondamentale per supportare esperienze narrative estese e coerenti nel tempo.

2.4.3 Storie ramificate

Le storie a bivi (branching stories) rappresentano un modello classico di narrazione interattiva, oggi reinterpretato in chiave generativa. Si è passati dai sistemi a percorsi fissi (ad esempio *Choose Your Own Adventure*) a modelli dinamici in cui la trama si adatta in tempo reale alle scelte dell'utente, fino alle esperienze basate su LLM che permettono una generazione aperta e non deterministica delle opzioni narrative.

Il lavoro di Summers-Stay e Voss, 2024 analizza proprio questo aspetto, proponendo strategie per gestire narrazioni che ramificano e poi riconvergono. Gli autori discutono i compromessi tra libertà dell'utente e sostenibilità della trama, mostrando come meccanismi di riconvergenza possano ridurre l'esplosione combinatoria delle ramificazioni mantenendo coerenza e controllo autoriale. Tali tecniche sono particolarmente rilevanti nei contesti in cui è necessario bilanciare la percezione di libertà narrativa con la fattibilità computazionale e di design.

Analogamente, *GENEVA* Leandro et al., 2024 fornisce un approccio strutturale alla rappresentazione delle storie ramificate, generando grafi narrativi che consentono al designer di visualizzare e modificare i punti di divergenza e riconvergenza. Questi strumenti offrono un ponte tra generazione automatica e authoring umano, permettendo un controllo più fine della struttura narrativa e delle sue varianti.

Oltre a questi approcci strutturali, il lavoro di Koenitz et al., 2024 introduce una prospettiva complementare, focalizzata non sulla generazione della struttura ramificata in sé, ma sulla capacità degli attuali strumenti di Intelligenza Artificiale generativa di supportare la progettazione di narrazioni interattive. Attraverso un framework di benchmarking sviluppato con designer professionisti, gli autori valutano come i modelli linguistici possano contribuire a compiti tipici dell'authoring di storie ramificate, tra cui la creazione di progressioni narrative, backstory, descrizioni di personaggi e semplici alberi di dialogo.

Il loro studio evidenzia tuttavia limiti significativi nell'uso dei LLM come generatori autonomi di strutture complesse: le ramificazioni prodotte tendono a collassare verso percorsi lineari, a perdere coerenza fra rami paralleli o a richiedere numerose iterazioni di prompt engineering per raggiungere un risultato utilizzabile. In questo senso, il contributo di Koenitz et al., 2024 si colloca in continuità con le

sfide affrontate da Summers-Stay e Voss, 2024, da Mensa et al., in press e dagli strumenti come *GENEVA* Leandro et al., 2024, mostrando che la generazione aperta tramite LLM non è ancora in grado di sostituire i modelli strutturali e gli strumenti di authoring dedicati.

Gli autori sottolineano inoltre come l'efficacia dei modelli dipenda fortemente dal contesto d'uso — ispirazione, formazione o produzione — e dalla qualità delle iterazioni di prompting. Il framework proposto suggerisce quindi un ruolo collaborativo della GenAI nel design delle storie ramificate: utile per accelerare ideazione e prototipazione, ma ancora dipendente dal controllo umano per garantire coerenza narrativa, gestione delle ramificazioni e autenticità autoriale.

2.4.4 Simulazioni e narrazioni continue

Le applicazioni più recenti tendono a fondere il paradigma delle storie a bivi con quello delle simulazioni aperte, in cui non esistono percorsi prestabiliti ma un flusso narrativo continuo, simile a un *virtual pet* o a una *simulazione ambientale*. In questo scenario, la storia non è più una sequenza di scelte discrete, ma un ecosistema narrativo in evoluzione costante.

Unbounded J. Li et al., 2024 rappresenta un caso emblematico di questo paradigma: un mondo generativo illimitato in cui personaggi, ambienti e regole emergono dinamicamente dal comportamento del modello. L'interazione non segue percorsi fissi, ma si sviluppa come una simulazione narrativa continua, nella quale la storia evolve in base alle scelte e al contesto corrente. Questo approccio richiede strategie avanzate di sincronizzazione tra agenti, meccaniche e memoria per evitare derive e mantenere un senso di coerenza globale.

Anche esperienze come *AI Dungeon* Latitude, Inc., 2019 possono essere lette in questa prospettiva: pur essendo testuale, la sua generazione aperta produce un effetto di simulazione continua del mondo narrativo, dove ogni input dell'utente può potenzialmente trasformare radicalmente la direzione della storia. Tuttavia, l'assenza di vincoli semantici espliciti evidenzia la difficoltà di mantenere una coerenza globale in sistemi completamente aperti, spingendo la ricerca verso modelli ibridi che integrino elementi di pianificazione narrativa e controllo semantico.

In sintesi, la tendenza generale è orientata all'unione della flessibilità delle narrazioni generative aperte con la struttura e la coerenza delle architetture modulari, attraverso meccanismi di memoria e controllo che permettano esperienze interattive fluide ma sostenibili nel tempo.

2.5 Metriche di valutazione

La valutazione della qualità narrativa nelle applicazioni interattive in tempo reale basate su LLM rappresenta una sfida complessa, poiché combina aspetti estetici,

cognitivi e funzionali. A differenza di compiti linguistici più strutturati, come la traduzione o il riassunto automatico, la valutazione narrativa richiede di misurare elementi qualitativi quali la coerenza della trama, la caratterizzazione dei personaggi, il coinvolgimento dell'utente e la capacità di adattamento alle interazioni. Negli ultimi anni, la letteratura ha proposto due direzioni complementari per affrontare questa complessità: (i) la definizione di metriche specifiche per la valutazione della storia e (ii) l'impiego degli LLM stessi come strumenti di valutazione, nel paradigma noto come *LLM-as-Judge*.

2.5.1 Metriche per la qualità della storia

Secondo il lavoro di Yang e Jin, 2024, la valutazione della qualità di una storia può essere affrontata attraverso una combinazione di metriche *umane*, *automatiche* e *ibride*, ciascuna mirata a catturare diverse dimensioni della narrazione. Il loro lavoro propone una tassonomia che distingue tra misure di coerenza narrativa, sviluppo dei personaggi, originalità e coinvolgimento emotivo, suggerendo che nessuna singola metrica sia sufficiente a descrivere la complessità del fenomeno narrativo.

Categorie principali di metriche. Tra le dimensioni più frequentemente adottate per la valutazione automatica della narrativa generata da LLM troviamo:

- **Coerenza narrativa:** misura la progressione logica e la connessione tra eventi, evitando contraddizioni e discontinuità;
- **Sviluppo dei personaggi:** valuta l'evoluzione dei personaggi, la consistenza delle motivazioni e la plausibilità delle azioni;
- **Coinvolgimento e interesse:** analizza il grado di immersione del lettore/utente, spesso attraverso metriche derivate da annotazioni umane o modelli di sentiment;
- **Originalità e sorpresa:** quantifica la capacità della storia di introdurre elementi inaspettati ma coerenti.
- **Completezza e chiusura:** verifica la presenza di un arco narrativo concluso o di un equilibrio interno riconoscibile;
- **Adattività e interattività:** misura la capacità della narrazione di rispondere in modo coerente alle scelte o agli input dell'utente, fondamentale nei sistemi in tempo reale;

Yang e Jin propongono inoltre benchmark specifici e dataset annotati per la valutazione automatica della narrativa, evidenziando come molte metriche

tradizionali (BLEU, ROUGE, BERTScore) non siano sufficientemente sensibili agli aspetti semantici e stilistici del testo narrativo. Ne deriva la necessità di approcci più semantici o embedding-based, in grado di catturare relazioni profonde tra segmenti testuali e coerenza globale della storia.

Tabella 2.1: Principali metriche per la valutazione narrativa e loro applicabilità nelle storie interattive in tempo reale.

Metrica	**Descrizione**	**Applicabilità alle storie interattive**
Coerenza narrativa	Progressione logica e consistenza degli eventi	Fondamentale: input utente imprevedibili possono introdurre incoerenze
Sviluppo del personaggio	Evoluzione credibile e motivata dei personaggi	Alta: richiede memoria contestuale
Coinvolgimento	Capacità della storia di mantenere attenzione e curiosità	Essenziale per esperienze interattive
Originalità	Presenza di elementi inaspettati ma plausibili	Media: utile per evitare ripetitività
Completezza	Arco narrativo coerente o bilanciato	Variabile: dipende dalla durata e dal tipo di sessione
Adattività	Reazione coerente alle scelte dell'utente	Cruciale per storie interattive
Consistenza stilistica	Stabilità di tono e registro linguistico	Alta: previene rotture nell'immersione

2.5.2 *LLM-as-Judge*

Un filone crescente di ricerca, infatti, indaga la possibilità di utilizzare gli stessi modelli linguistici come strumenti di valutazione automatica. Questo approccio, noto come *LLM-as-Judge*, mira a sostituire — o almeno integrare — la valutazione umana con il giudizio di un modello, sfruttando la sua capacità di comprendere il linguaggio naturale e di ragionare su criteri astratti.

H. Li et al., 2024 offrono una rassegna sistematica delle metodologie di valutazione basate su LLM, distinguendo fra approcci *prompt-based*, *comparativi* e *ragionati* (chain-of-thought). Thakur et al., 2024 evidenziano i limiti intrinseci di questo paradigma, mostrando come anche i modelli più avanzati tendano a divergere dai

giudizi umani e presentino bias sistematici legati al formato del prompt o alla lunghezza del testo.

Un contributo recente e rilevante è offerto da Chiang e Lee, 2023, che analizza in dettaglio come ChatGPT (GPT-3.5) possa essere utilizzato come valutatore automatico e quali fattori ne influenzino l'affidabilità. Gli autori impiegano due dataset standard di meta-valutazione: SummEval (riassunti) e Topical-Chat (dialoghi guidati da conoscenza), dove sono disponibili giudizi umani su coerenza, rilevanza, naturalezza e coinvolgimento.

Metodologia Vengono testati diversi tipi di prompt:

- **Score-only:** il modello assegna solo un punteggio numerico;
- **Rate-and-explain / Analyze-then-rate:** il modello fornisce sia un punteggio che una spiegazione testuale;
- **Auto chain-of-thought (CoT):** il modello genera autonomamente passaggi di ragionamento prima di produrre il punteggio.

Le correlazioni tra valutazioni umane e automatiche vengono misurate tramite coefficienti di Pearson e Kendall's τ, permettendo di quantificare la vicinanza tra giudizi umani e LLM.

Risultati principali

- I prompt che includono spiegazioni testuali generano correlazioni significativamente più alte con i giudizi umani rispetto ai prompt *score-only*, mostrando l'importanza di rendere espliciti i criteri di valutazione.
- L'uso di auto CoT non migliora sempre la qualità della valutazione e, in alcuni casi, può addirittura ridurla, suggerendo che un ragionamento guidato dal prompt è più stabile.
- Le performance risultano relativamente robuste rispetto a variazioni di temperatura e di formulazione del prompt, purché il compito sia chiaramente strutturato.

Implicazioni pratiche Le conclusioni di Chiang e Lee, 2023 forniscono linee guida utili per la progettazione di sistemi di valutazione narrativa basati su LLM:

- Esplicitare criteri e richiedere spiegazioni testuali aumenta l'allineamento con valutazioni umane;
- L'uso indiscriminato di auto CoT non è consigliato;
- Il design del prompt gioca un ruolo cruciale per stabilità e affidabilità.

2.5.3 Adattamento delle metriche narrative alla valutazione automatica

L'integrazione tra metriche narrative e approcci basati su *LLM-as-Judge* consente di estendere la valutazione tradizionale a una forma più dinamica e iterativa. Seguendo H. Li et al., 2024 e Chiang e Lee, 2023, il successo di un sistema di valutazione automatica dipende dalla chiarezza con cui la metrica viene tradotta in istruzioni per il modello. Ad esempio, per la coerenza narrativa, un prompt efficace può essere formulato come:

> "Valuta quanto la storia mantiene una progressione logica coerente, senza contraddizioni o eventi inspiegabili, assegnando un punteggio da 1 a 5 e fornendo una spiegazione testuale."

La *meta-valutazione* del giudizio automatico, ossia il confronto tra le valutazioni dell'LLM e quelle umane, costituisce un passo necessario per stimarne l'affidabilità. Nel contesto delle applicazioni narrative in tempo reale, l'uso di LLM come giudici può inoltre essere integrato in un ciclo di miglioramento continuo del modello generativo, fornendo segnali di feedback strutturati su coerenza, stile e adattività della storia.

2.5.4 Sintesi

La valutazione narrativa automatica si colloca oggi al crocevia tra linguistica computazionale e estetica narrativa. Mentre le metriche tradizionali forniscono un quadro concettuale utile per descrivere la qualità di una storia, l'uso degli LLM come giudici, supportato da studi empirici recenti, apre la strada a nuove forme di valutazione adattiva, capaci di integrare giudizio umano e analisi automatica. Per le applicazioni narrative interattive in tempo reale, la sfida consiste nel bilanciare accuratezza, trasparenza e tempi di esecuzione, in vista di un futuro in cui la valutazione stessa diventa parte integrante del processo narrativo.

2.6 Casi d'uso museali ed educativi

Diversi progetti recenti hanno esplorato l'uso di LLM per la creazione di esperienze narrative in contesti museali e didattici. Tra questi, il progetto **CHANGES** condotto presso il Museo Egizio di Torino costituisce un caso di particolare interesse, in quanto mira a supportare i curatori nell'implementazione di narrazioni interattive assistite da intelligenza artificiale.

2.6.1 Il progetto CHANGES e la narrazione curatoriale assistita da LLM

Nella ricerca condotta da Mensa et al., in press è stata sviluppata una piattaforma che consente ai curatori di implementare storie interattive con il supporto di modelli linguistici. L'approccio si fonda su un paradigma di *human-AI co-creation*, in cui l'intelligenza artificiale non sostituisce l'esperto, ma ne amplifica le capacità creative e linguistiche.

- **Obiettivo:** supportare la creazione di narrazioni curatoriali accurate e coinvolgenti.
- **Metodo:** uso selettivo degli LLM per trasformare descrizioni strutturate di scene in prosa narrativa.
- **Risultati:** i modelli hanno mostrato buone capacità linguistiche ma necessitano ancora di revisione umana per garantire coerenza e accuratezza scientifica.

2.6.2 Altri esempi di applicazione di LLM in musei e contesti educativi

Cacce al tesoro interattive nei siti culturali Il lavoro di Gutiérrez-Sánchez et al., 2025 esplora come i LLM possano supportare la generazione automatica di cacce al tesoro in siti culturali, consentendo ai curatori di impostare parametri chiave e lasciando all'AI il compito di creare storie interattive. Lo studio mostra che un approccio basato su una fase di *pre-planning* migliora la coerenza e la qualità narrativa dei contenuti generati, pur richiedendo revisione umana per garantire accuratezza e correttezza dei dati.

Chatbot e avatar virtuali per engagement e apprendimento Zhang et al., 2025 confronta diversi strumenti di supporto ai visitatori in musei virtuali: etichette testuali, chatbot basati su LLM, e avatar guida. I risultati indicano che le interazioni tramite avatar aumentano significativamente l'engagement e l'esperienza utente rispetto a chatbot o etichette, mentre l'effetto diretto sull'apprendimento risulta più contenuto. Questo evidenzia l'importanza dell'interfaccia e della presentazione dei contenuti nella progettazione di esperienze educative digitali.

Tour personalizzati basati sulle preferenze dell'utente Infine, il lavoro di Vasic et al., 2024 propone una guida museale che combina LLM e panorami 3D per offrire tour personalizzati in base agli interessi espressi in linguaggio naturale

dai visitatori. Questo approccio permette di generare percorsi unici e narrativamente coerenti, mostrando come la narrativa generativa possa essere adattata alle preferenze individuali senza compromettere la qualità culturale dei contenuti.

Questi casi evidenziano che gli LLM possono supportare la creazione di esperienze culturali più interattive, personalizzate e coinvolgenti, sia in contesti virtuali sia in siti fisici, pur richiedendo supervisione e revisione dei contenuti da parte di esperti per garantire accuratezza e autenticità.

2.7 Sintesi

Il capitolo ha illustrato le principali direzioni di ricerca e gli strumenti alla base delle applicazioni narrative interattive generate da LLM. Dalla qualità narrativa alla gestione della memoria, dall'integrazione visiva alla valutazione delle prestazioni, emerge una rete di soluzioni in rapida evoluzione, che fornisce il quadro di riferimento teorico per l'indagine sperimentale sviluppata nei capitoli successivi.

Capitolo 3

Metodologia e design del prototipo

3.1 Obiettivi sperimentali

La presente ricerca mira a esplorare le potenzialità della generazione narrativa interattiva in tempo reale, attraverso l'uso di LLM. Gli obiettivi sperimentali sono stati definiti per indagare le dimensioni fondamentali della progettazione di un sistema capace di produrre narrazione coerente e dinamicamente adattiva alle scelte dell'utente.

3.1.1 Analisi della capacità dell'LLM di rimanere coerente

Il primo obiettivo consiste nel verificare la possibilità di generare contenuti narrativi in tempo reale che siano complessivamente coerenti e che complessivamente presentino una struttura tipica di un racconto, ossia gestire uno sviluppo e una conclusione a partire da un'introduzione con un conflitto dati. Questo implica la valutazione della qualità del testo prodotto e dell'esperienza percepita dall'utente durante la fruizione interattiva. L'esperimento si concentra sull'analisi del bilanciamento tra il mantenere una struttura e assecondare l'utente nell'andamento della trama.

3.1.2 Analizzare la gestione dell'interazione con l'utente e delle scelte multiple

Il secondo obiettivo riguarda la capacità del sistema di mantenere la coerenza interna della storia nonostante la presenza di scelte multiple offerte all'utente. Verranno sperimentate strategie di prompting e strutture di memoria per monitorare la continuità di eventi e ambientazioni. Tale obiettivo mira a comprendere come la

progettazione del prompt e la gestione del contesto influenzino la stabilità narrativa e la percezione di immersione da parte dell'utente.

3.1.3 Indagine della fattibilità di rappresentazione grafica in output testuale

Infine, si vuole esplorare la possibilità di integrare elementi visivi descritti in forma testuale, come schemi, mappe o rappresentazioni simboliche, all'interno della narrazione generativa. L'intento è valutare la capacità di un LLM di gestire anche una componente grafica in tempo reale, ma gestita esclusivamente in output testuale, che è veloce e leggero da gestire.

3.1.4 Sintesi

In sintesi, questi obiettivi sperimentali definiscono il quadro operativo entro cui si muove la progettazione del prototipo descritto nei paragrafi successivi, orientando le scelte metodologiche e tecniche verso la costruzione di un'esperienza narrativa interattiva coerente.

3.2 Riferimenti metodologici

La metodologia di sviluppo del prototipo trae ispirazione dal lavoro svolto nell'ambito del progetto CHANGES, presentato nel paper *"There was a scribe, a priest and a thief'..."* Mensa et al., in press, in cui i LLM vengono impiegati per assistere i curatori nella scrittura di narrazioni museali. A differenza di tale approccio, centrato sulla collaborazione uomo–modello, il prototipo sviluppato in questa tesi esplora la generazione narrativa autonoma e continua in tempo reale.

3.3 Scelte progettuali

3.3.1 Struttura narrativa

L'obiettivo principale del prototipo è sfruttare un LLM per offrire all'utente un'interazione caratterizzata da un alto grado di libertà. Nel contesto pratico, ciò significa non definire a priori un grafo delle possibilità narrative, ma generare ogni volta una storia nuova in base alle scelte dell'utente.

Nel progettare la struttura della narrazione, si è cercato di bilanciare due esigenze apparentemente opposte. Da un lato, si vuole mantenere la coerenza tipica di un racconto d'avventura, con una sequenza narrativa composta da introduzione, conflitto, sviluppo e conclusione. Dall'altro, è necessario preservare la libertà

decisionale dell'utente. Lasciare l'LLM completamente libero non garantisce una soluzione interessante, in quanto la narrazione tende a prolungarsi indefinitamente senza giungere a un punto culminante. Viceversa, imporre una struttura narrativa rigida rischia di ridurre l'efficacia delle scelte dell'utente, poiché l'LLM tende a riportare la storia verso gli schemi prefissati.

La soluzione adottata consiste in una breve introduzione iniziale, in cui viene presentato il conflitto principale e vengono offerte tre possibili scelte su come proseguire. Successivamente, la narrazione procede scena per scena, generata dinamicamente dall'LLM, con tre scelte disponibili per l'utente in ciascuna scena. Il numero di scene è stabilito a priori, in modo da garantire il raggiungimento di una conclusione coerente, bilanciando libertà narrativa e struttura del racconto.

3.3.2 Modelli considerati

Nel prototipo sono stati utilizzati due modelli linguistici di grandi dimensioni (LLM) open-weight, scelti per le loro caratteristiche tecniche e per la possibilità di eseguire esperimenti sperimentali:

- **gpt-oss-120b** OpenAI, 2025
- **Gemma 3 27B** DeepMind, 2025

gpt-oss-120b Il modello gpt-oss-120b conta circa 117 miliardi di parametri e utilizza un'architettura di tipo Mixture-of-Experts (MoE), che attiva solo una frazione dei parametri per ciascun token. Supporta finestre di contesto fino a 128 000 token e consente ragionamento avanzato, tool-use e scenari agentici. La sua architettura permette di gestire storie dinamiche con libertà narrativa elevata e ragionamento contestuale continuato.

Gemma 3 27B La famiglia Gemma 3 comprende modelli da 1B fino a 27B parametri, di cui la versione 27B è stata adottata nel prototipo. Tra le caratteristiche principali:

- Finestra di contesto fino a 128 k token, ideale per testi lunghi DeepMind, 2025;
- Supporto multilingue esteso (oltre 140 lingue);
- Capacità multimodale (testo e immagine) tramite encoder visivi integrati;
- Esecuzione efficiente su GPU consumer mediante versioni quantizzate.

Modelli utilizzati nei test preliminari Durante la fase di sviluppo e test preliminare della pipeline narrativa, è stato utilizzato anche il modello **Gemma 3 Latest (4B)**. Questo modello, con circa 4 miliardi di parametri, rappresenta la versione più leggera della famiglia Gemma 3, consentendo di eseguire esperimenti in tempi rapidi e con minori requisiti hardware rispetto a modelli più grandi come Gemma 3 27B. Gemma 3 Latest supporta finestre di contesto estese e capacità di ragionamento avanzato, caratteristiche che lo rendono adatto a sperimentare differenti strategie di prompt e a valutare la capacità dell'LLM di generare scene coerenti e scelte interattive. L'utilizzo di questo modello nei test preliminari ha permesso di ottimizzare la struttura dei prompt, esplorare la coerenza narrativa e ridurre i tempi di iterazione nella progettazione della narrazione, fornendo indicazioni preziose per la scelta dei modelli principali adottati nel prototipo (gpt-oss-120b e Gemma 3 27B).

Motivazione della scelta La selezione combinata dei due modelli principali ha permesso di confrontare approcci differenti in termini di capacità computazionale, libertà narrativa e requisiti hardware:

- un modello a elevata capacità computazionale (gpt-oss-120b), ottimizzato per ragionamento avanzato, uso di strumenti, contesto lungo e deployment efficiente su hardware accessibile;

- un modello più leggero (Gemma 3 27B), ottimizzato per comprensione e generazione multimodale e multilingue, ragionamento e istruzioni complesse, lunghe sequenze di contesto

Questo approccio ha ridotto la dipendenza del prototipo dal singolo modello, garantendo flessibilità nello sviluppo della pipeline e nella sperimentazione della generazione scena-per-scena. Inoltre, ha permesso di valutare l'impatto della capacità computazionale e della complessità del modello sulla qualità narrativa percepita dall'utente.

Considerazioni finali L'adozione di modelli open-weight ha permesso un maggiore controllo sull'intera pipeline di generazione e sulla personalizzazione dei prompt, eliminando la necessità di ricorrere a API proprietarie. Inoltre, considerando una possibile implementazione del prototipo in contesti museali o educativi, i modelli open-weight rappresentano la soluzione più adeguata, poiché evitano la gestione di dipendenze da fornitori esterni, pur richiedendo un impegno hardware relativamente contenuto.

3.3.3 Interfaccia e componente grafica

Il prototipo prevede un'interfaccia essenziale basata su terminale, concepita per consentire una sperimentazione diretta del flusso narrativo generato in tempo reale dal modello di linguaggio. Durante l'interazione, le scene prodotte dal LLM vengono visualizzate sequenzialmente insieme alle opzioni di scelta disponibili per l'utente e alla relativa rappresentazione grafica in formato *ASCII art*, che raffigura un elemento significativo della scena corrente.

Questa soluzione minimale è stata scelta per privilegiare la funzionalità e l'analisi del comportamento narrativo del sistema, riducendo al minimo le dipendenze da componenti grafiche esterne.

In prospettiva futura, potrebbe risultare interessante lo sviluppo di un'interfaccia testuale basata su browser, mantenendo un'estetica coerente con la natura narrativa del progetto ma introducendo elementi di UI più evoluti. Tra le possibili estensioni figurano l'animazione progressiva del testo generato in tempo reale, la possibilità di selezionare visivamente le scelte tramite interfaccia grafica e l'inclusione di transizioni o micro-animazioni per arricchire l'esperienza interattiva pur preservando l'essenzialità dell'impianto testuale.

3.3.4 Architettura software

L'architettura software del prototipo è stata concepita con l'obiettivo di garantire modularità, chiarezza strutturale e facilità di sperimentazione. Il sistema è stato progettato in modo da separare logicamente le responsabilità dei diversi componenti — generazione narrativa, interazione con l'utente e rappresentazione grafica — permettendo di sostituire o estendere ciascun modulo senza modificare l'intera applicazione.

Obiettivi progettuali L'architettura si basa su un principio di semplicità e trasparenza, privilegiando la possibilità di testare differenti LLM e strategie di prompt engineering. Sono stati perseguiti tre obiettivi principali:

- **Modularità:** separazione chiara tra i componenti di gestione narrativa, interfaccia e generazione LLM;
- **Flessibilità:** possibilità di sostituire facilmente il modello di linguaggio o il modulo grafico, senza impatti significativi sul resto del sistema;
- **Riproducibilità:** mantenimento di un'architettura minimale e facilmente eseguibile in ambienti locali, senza dipendenze esterne complesse.

Descrizione generale Il sistema segue un flusso lineare di esecuzione che si ripete per ogni scena generata. In sintesi:

1. L'interfaccia mostra all'utente una scena descritta dal LLM, con la grafica ASCII ad essa associata, e tre possibili scelte su come proseguire la narrazione;
2. L'utente seleziona una delle opzioni proposte;
3. Il sistema costruisce un nuovo prompt basato sullo stato narrativo e sulla scelta effettuata;
4. L'LLM genera la scena successiva e seleziona un elemento significativo da rappresentare in grafica *ASCII*;

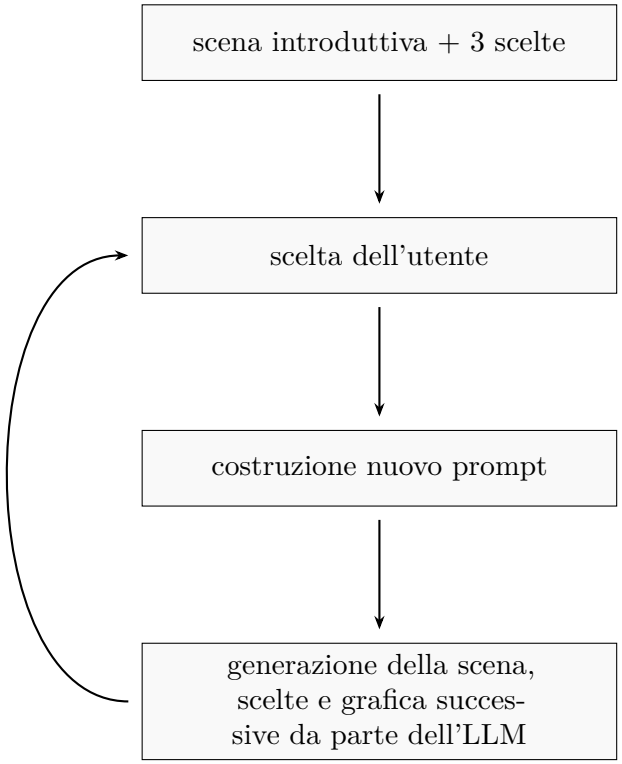

Figura 3.1: Flusso logico di generazione e interazione scena–scelta–risposta.

Componenti principali L'architettura è composta da quattro moduli principali:

- **Core narrativo:** gestisce la progressione della storia, il numero di scene e la coerenza tra le scelte dell'utente e lo sviluppo narrativo;

- **Modulo LLM:** si occupa della costruzione dinamica dei prompt, dell'invio delle richieste ai modelli (gpt-oss-120b e Gemma 3 27B, oltre a Gemma 3 4B per i test preliminari) e dell'elaborazione dell'output testuale;

- **Motore grafico ASCII:** interpreta le indicazioni fornite dal LLM e genera la rappresentazione visiva della scena in formato ASCII art;

- **Interfaccia utente (CLI):** visualizza il testo, la rappresentazione grafica e le scelte narrative, gestendo l'input dell'utente in tempo reale.

Scelte architetturali La decisione di adottare un'architettura a componenti indipendenti ha permesso di mantenere il sistema flessibile e facilmente estendibile. In particolare:

- la separazione tra *core narrativo* e *modulo LLM* consente di modificare la logica narrativa o sostituire il modello di generazione senza alterare l'interfaccia utente;

- la rappresentazione in *ASCII art* è stata gestita come modulo autonomo per favorire futuri sviluppi grafici o multimodali;

- la scelta di un'interfaccia a riga di comando (CLI) riduce la complessità tecnica, consentendo di concentrare l'attenzione sugli aspetti metodologici della generazione narrativa.

Possibili estensioni In uno sviluppo futuro, l'architettura potrebbe essere ampliata con:

- un modulo di interfaccia web testuale per la visualizzazione interattiva delle scelte;

- un sistema di persistenza dello stato narrativo per salvare e riprendere le sessioni;

- l'integrazione di un modulo di analisi automatica per valutare coerenza, stile e qualità narrativa delle generazioni.

Queste estensioni manterrebbero la stessa logica modulare, garantendo la continuità metodologica del progetto.

3.4 Strategie di prompting

In questa sezione vengono illustrate le strategie di prompting adottate per la generazione della narrazione e della rappresentazione visiva dell'applicazione interattiva. L'obiettivo è garantire, da un lato, la coerenza e la progressione narrativa e, dall'altro, preservare un alto grado di libertà per l'utente nelle scelte che determinano lo sviluppo della storia. Per questo motivo, sono state progettate due tipologie principali di prompt: una dedicata alla generazione testuale (narrazione scena per scena) e una alla generazione grafica (rappresentazione ASCII degli elementi significativi).

3.4.1 Prompt per la narrazione

Al fine di gestire la generazione narrativa e garantire al contempo un elevato grado di libertà decisionale per l'utente, è stato necessario individuare un compromesso tra due strategie di prompting opposte. Da un lato, un *prompt globale* — ovvero una richiesta generica di proseguire la storia — permette al modello di esplorare liberamente lo spazio narrativo, ma comporta il rischio di ottenere una trama priva di direzione o di un obiettivo narrativo chiaro. Dall'altro lato, un *prompt altamente strutturato* — che descrive in anticipo l'andamento della storia scena per scena — garantisce coerenza e controllo sul percorso narrativo, ma riduce significativamente la libertà creativa e l'impatto delle scelte dell'utente, poiché il modello tende a ricondurre la narrazione verso un percorso predefinito.

Per superare questa dicotomia è stata adottata una strategia ibrida, basata sulla distinzione tra due tipologie di prompt:

- **Prompt per le scene intermedie**, progettati per mantenere aperta la direzione narrativa e valorizzare le scelte dell'utente. Questi prompt vengono generati dinamicamente a ogni nuova scena e contengono:

  1. la storia fino a quel punto;
  2. la scelta effettuata dall'utente;
  3. le informazioni su tono e formato.

  Il modello utilizza tali informazioni per produrre la scena successiva, coerente con il contesto e con l'evoluzione narrativa in corso.

- **Prompt per la scena conclusiva**, con una struttura più vincolante che orienta il modello verso una chiusura coerente e soddisfacente della trama. In questo caso, il prompt specifica esplicitamente che la scena generata deve costituire un epilogo narrativo.

Questa distinzione consente alla storia di evolversi liberamente nelle fasi centrali, pur garantendo un esito narrativo completo e coerente nella fase finale. La costruzione dinamica del prompt a ogni scena rappresenta dunque un meccanismo di *aggiornamento contestuale*, che permette al modello di mantenere la memoria narrativa e di adattarsi continuamente alle scelte dell'utente.

3.4.2 Prompt per la rappresentazione grafica

Parallelamente alla generazione testuale, il sistema produce una rappresentazione visiva in grafica ASCII di un elemento significativo della scena. Anche in questo caso, il prompt destinato alla generazione grafica è costruito dinamicamente: riceve come input la descrizione della scena corrente come contesto. Il modello è quindi istruito a sintetizzare visivamente un singolo elemento iconico che caratterizzi la scena, mantenendo coerenza stilistica e contribuendo alla continuità narrativa complessiva dell'esperienza interattiva.

3.4.3 Test preliminari

Prima di definire la versione finale dei prompt, sono stati condotti test preliminari per valutare l'efficacia delle strategie di prompting e identificare eventuali problemi nella generazione narrativa e delle scelte. I risultati hanno evidenziato alcune criticità importanti:

- L'uso di un **prompt unico** in cui si indicava semplicemente che all'ultima scena la storia dovesse concludersi non era sufficiente a garantire un epilogo coerente. La soluzione adottata è stata quella di distinguere due prompt separati: uno per le scene intermedie e uno per la scena conclusiva;
- La distinzione tra **human prompt** e **system prompt** si è rivelata molto efficace. Il system prompt fornisce istruzioni chiare sul formato, la lunghezza del testo e le regole narrative, mentre l'human prompt comunica al modello il contesto e le scelte dell'utente;
- In generale, la lunghezza del testo e il formato venivano rispettati correttamente, e solo raramente comparivano parole in inglese non desiderate. Più frequentemente si riscontravano errori semantici, come associazioni scorrette tra soggetto e verbo o formulazioni incoerenti all'interno della scena;
- Durante la generazione delle scelte, talvolta venivano introdotti elementi nuovi mai menzionati in precedenza. Per ridurre questo problema, nel system prompt è stato specificato esplicitamente di non aggiungere elementi narrativi nuovi, limitandosi al contesto già definito;

- È stato osservato anche che le opzioni proposte potevano risultare troppo simili tra loro. Per risolvere questa criticità, il system prompt è stato aggiornato con l'indicazione che le tre opzioni devono rappresentare approcci narrativi chiaramente alternativi e distinti;

Per quanto riguarda la componente visiva, la scelta di utilizzare una rappresentazione in *ASCII art* nasce dall'esigenza di ottenere un output grafico leggero e facilmente gestibile, in quanto si tratta di semplice testo formattato. È stata effettuata una prima serie di test utilizzando modelli LLM commerciali online, al fine di valutare le capacità dei diversi modelli nella generazione di elementi visivi coerenti con il contesto narrativo. In alcune specifiche situazioni, i risultati sono stati più che soddisfacenti, fornendo rappresentazioni riconoscibili e coerenti con la scena descritta.

Tuttavia, già in questa fase è emerso come sussistano limiti significativi nella generazione grafica tramite LLM, in linea con quanto discusso nel Capitolo 2. Nonostante ciò, l'esperimento ha indicato che la generazione di ASCII art rappresenta una direzione di ricerca interessante per combinare narrazione e visualizzazione leggera.

Anche per la componente grafica è stata adottata la distinzione tra *system prompt* e *human prompt*. I test preliminari hanno confermato che i modelli seguono correttamente le istruzioni fornite nel system prompt, rispettando nella maggior parte dei casi i formati e le lunghezze richieste e producendo output coerenti con la scena corrente, almeno per quanto riguarda l'elemento da rappresentare.

Questi test preliminari hanno permesso di affinare i prompt e di garantire una maggiore coerenza narrativa, un rispetto più rigoroso del formato e una maggiore diversificazione delle scelte offerte all'utente.

3.5 Gestione delle scelte e branching story

L'interazione tra utente e modello è stata progettata in modo da bilanciare la libertà narrativa con l'efficienza computazionale e la coerenza contestuale. Per ridurre i problemi legati alla gestione della memoria e al mantenimento di un contesto narrativo coerente nel tempo, non è stato adottato un sistema di input testuale libero: invece di permettere all'utente di scrivere liberamente come proseguire la storia, l'applicazione propone tre opzioni predefinite per ogni scena. Queste opzioni, generate dal modello stesso, rappresentano tre possibili direzioni narrative distinte e coerenti con la trama fino a quel momento.

3.5.1 Generazione delle opzioni per l'utente

A ogni scena, il modello riceve nel prompt tutta la storia generata fino a quel punto e l'indicazione di generare tre alternative narrative che rappresentino scelte significative per l'utente. Le opzioni devono essere:

- **Distinte** tra loro, per garantire varietà nelle possibili evoluzioni della storia;
- **Coerenti** con gli eventi precedenti e con il tono narrativo;
- **Sintetiche**, in modo da poter essere presentate in modo chiaro e immediato nell'interfaccia utente.

Questa struttura consente di mantenere un alto livello di interattività pur evitando la complessità derivante dall'elaborazione di input linguistici arbitrari da parte dell'utente, che potrebbero introdurre ambiguità o incoerenze nel contesto narrativo. La generazione controllata delle scelte, inoltre, riduce la possibilità di errori semantici o di deviazioni eccessive rispetto al tema principale della storia.

3.5.2 Limitazioni

Sebbene la struttura a scelte multiple renda l'esperienza più stabile e gestibile, la narrazione non è organizzata come una vera e propria *branching story*. Il sistema non mantiene infatti una rappresentazione grafica o strutturale dei nodi narrativi, né tiene traccia delle scelte precedenti come stati distinti del racconto. Ogni scena viene generata unicamente a partire dal contesto testuale cumulativo e dall'ultima decisione dell'utente, senza la possibilità di "tornare indietro" o di esplorare percorsi alternativi già intrapresi. Un'ulteriore limitazione deriva dal fatto che, a ogni iterazione, il prompt fornisce al modello l'intera storia generata fino a quel momento. Per contenere i costi computazionali e prevenire problemi di gestione del contesto, le singole scene sono state mantenute intenzionalmente brevi, con una lunghezza media di circa 100 parole. Questa scelta ha permesso di rendere le valutazioni sperimentali più gestibili: le storie prodotte sono state infatti analizzate sia da valutatori umani sia da LLM in modalità di autovalutazione. Il mantenimento di scene brevi ha inoltre eliminato la necessità di implementare strutture di memoria articolate o meccanismi di gestione dello stato narrativo persistente.

Capitolo 4

Implementazione

4.1 Stack tecnologico

Per lo sviluppo del sistema di generazione e valutazione automatica delle storie interattive è stato adottato uno stack tecnologico basato su Python e librerie specifiche per il Natural Language Processing (NLP) e il prompt engineering.

4.1.1 Linguaggio di programmazione

Il linguaggio principale utilizzato è **Python**, per la sua versatilità, ampia disponibilità di librerie scientifiche e di NLP, e per la facilità di integrazione con modelli di linguaggio avanzati.

4.1.2 Framework e librerie principali

- **LangChain**: utilizzato per costruire catene di prompt e gestire il flusso di interazione con il modello di linguaggio. LangChain consente di definire prompt strutturati, concatenare più prompt e gestire input/response in maniera modulare.

- **Ollama con Structured Output**: integrato con LangChain per ottenere risposte con una struttura definita, mappata direttamente su schemi Pydantic. Questa funzionalità permette di generare oggetti Python tipizzati (`Scena`, `EvalStoria`) direttamente dal modello, semplificando la gestione dei dati e riducendo la possibilità di errori di parsing.

- **Pydantic**: utilizzata per definire schemi di dati tipizzati per le scene e le valutazioni, garantendo validazione dei dati e coerenza strutturale.

- **Lettura e scrittura JSON**: le storie generate e le valutazioni sono gestite tramite file JSON, permettendo la serializzazione e la memorizzazione progressiva dei dati, utile per operazioni batch e per il salvataggio incrementale.

4.1.3 Workflow tecnologico

Lo stack combinato permette di seguire un flusso di lavoro modulare:

1. Generazione delle scene narrative tramite prompt strutturati con LangChain.
2. Parsing delle risposte in oggetti tipizzati tramite Pydantic e Ollama Structured Output.
3. Salvataggio progressivo delle storie e delle valutazioni in formato JSON, con gestione di errori e ripresa automatica.
4. Possibilità di elaborare batch di storie per valutazioni automatiche, mantenendo consistenza dei dati e integrità strutturale.

Questo stack tecnologico garantisce flessibilità, robustezza e facilità di manutenzione del codice, oltre a semplificare l'integrazione di modelli di linguaggio avanzati in contesti di narrativa interattiva.

4.2 Descrizione e funzionamento del prototipo

Il prototipo sviluppato rappresenta il cuore sperimentale di questo lavoro di tesi. L'architettura complessiva del sistema prevede un flusso iterativo basato su un ciclo di interazione tra modello e utente:

$$\text{Scena} \rightarrow \text{LLM} \rightarrow \text{Scelte} \rightarrow \text{Scena successiva}$$

Questo ciclo si ripete fino al completamento della storia, consentendo di osservare come il modello gestisca la progressione narrativa, la coerenza interna e l'effetto delle scelte dell'utente sull'evoluzione della trama.

4.2.1 Descrizione del prototipo

Il prototipo offre tre scenari distinti, ognuno caratterizzato da un tono narrativo specifico:

1. un contesto **fantasy drammatico**, con atmosfere cupe e tensione narrativa elevata;
2. un'avventura **fantasy leggera**, più orientata all'azione e alla scoperta;

3. un'avventura **comica**, incentrata su un protagonista felino e toni più ironici.

La presenza di questi tre contesti differenti consente di valutare la capacità del modello di mantenere coerenza narrativa e aderenza stilistica in situazioni con registri e obiettivi comunicativi molto diversi. Ogni storia generata è composta da un totale di **sei scene**. Il fatto di avere un numero definito di scene è stato stabilito in fase di progettazione come compromesso tra due esigenze opposte:

- garantire la presenza di un arco narrativo completo, con una struttura definita e una conclusione coerente;
- lasciare al modello un margine di libertà sufficiente per gestire autonomamente l'evoluzione della trama in funzione delle scelte dell'utente.

Questa configurazione consente quindi di bilanciare la struttura narrativa con la spontaneità creativa dell'LLM, evitando di imporre schemi troppo rigidi che ne limiterebbero l'espressività. Per quanto riguarda invece la scelta del numero di scene, è stato deciso di mantenere un numero contenuto per agevolare il processo di valutazione.

4.2.2 Versioni del prototipo

Il sistema è stato sviluppato in due versioni distinte, con obiettivi complementari:

- **Prototipo interattivo:** consente all'utente di interagire direttamente con il modello, scegliendo tra le opzioni proposte al termine di ogni scena. In questo modo il modello genera la narrazione scena per scena, adattandosi dinamicamente alle decisioni dell'utente e producendo un'unica storia completa per sessione di esecuzione.
- **Script di generazione automatica:** utilizza gli stessi prompt e scenari di base, ma genera un numero arbitrario N di storie in maniera completamente automatica. In questo caso, le scelte dell'utente vengono simulate in modo casuale, consentendo la creazione di molteplici narrazioni indipendenti. Questa seconda versione ha uno scopo puramente **valutativo**: permette di analizzare statisticamente le differenze tra le storie prodotte, la varietà delle scelte generate e la coerenza interna mantenuta dal modello. Inoltre, in questa versione non è presente la rappresentazione ASCII in quanto non è stata prevista una valutazione automatica della rappresentazione visiva.

Poiché non esiste una trama predefinita, ma solo uno **scenario iniziale comune**, ogni storia generata risulta unica, offrendo un campione significativo per la valutazione delle capacità narrative e di adattamento del LLM.

4.3 Gestione del prompt e degli scenari

La generazione delle storie è guidata da un insieme di **prompt strutturati** progettati per controllare il comportamento del modello linguistico durante le diverse fasi della narrazione. L'intero processo si basa su due componenti principali: il *prompt di sistema*, che definisce le istruzioni di contesto e il tono narrativo, e il *prompt utente*, che fornisce al modello la situazione corrente della storia e la scelta compiuta dal giocatore, richiedendo la generazione della scena successiva.

4.3.1 Struttura del prompt

Il **prompt di sistema** stabilisce le regole di costruzione della scena e il registro stilistico da adottare. In particolare, le istruzioni specificano:

- il formato testuale, con vincoli su lunghezza, punto di vista (seconda persona singolare) e coerenza con la scena precedente;
- la presenza di tre scelte distinte per l'utente, brevi e logicamente coerenti con la situazione descritta;
- la necessità di mantenere continuità narrativa, evitando l'introduzione di elementi nuovi o estranei;
- il tono narrativo da seguire, che varia in base allo scenario selezionato.

Il **prompt utente**, invece, fornisce al modello la *storia fino a quel momento*, la *scelta effettuata dall'utente* e il *numero della scena corrente*. A partire da queste informazioni, il modello deve generare la conseguenza diretta della scelta, producendo un testo coerente e tre nuove opzioni per proseguire la narrazione.

In fase di chiusura della storia viene utilizzato un *prompt conclusivo* dedicato, che richiede al modello di produrre una scena finale coerente con la scelta ultima dell'utente e in grado di dare un senso di compimento alla vicenda.

4.3.2 Scenari narrativi e tono

Sono stati definiti tre scenari principali, ciascuno associato a un diverso tono narrativo e a un differente contesto diegetico:

1. **Assedio al castello di Varn** – scenario *fantasy drammatico*, caratterizzato da atmosfere cupe, tensione e conflitto. Il tono impostato nel prompt è **"avventuroso, drammatico"**.
2. **Il mistero del bosco di Elarion** – scenario *fantasy leggero*, con enfasi sull'esplorazione e la curiosità. Il tono è **"avventuroso, curioso, leggero"**.

3. **Le avventure del gatto** – scenario *comico e surreale*, incentrato su un protagonista felino. Il tono specificato è **"avventuroso, comico, divertente."**

La variazione del tono nel prompt consente di osservare come il modello linguistico adatti la propria produzione testuale in funzione del contesto narrativo. Ogni tono influisce non solo sullo stile della narrazione (scelte lessicali, ritmo, umorismo), ma anche sulla struttura delle scelte proposte al giocatore, che risultano più drammatiche, esplorative o ironiche, a seconda dello scenario.

4.3.3 Considerazioni progettuali

La gestione modulare dei prompt permette di riutilizzare lo stesso schema di generazione per contesti diversi, modificando unicamente la sezione dedicata al tono e alla scena iniziale. Questa impostazione consente di mantenere costante la struttura di interazione, pur variando la direzione narrativa e il tipo di esperienza offerta all'utente. L'approccio adottato garantisce quindi un elevato grado di **generalizzabilità** e facilita l'estensione del prototipo a nuovi scenari futuri.

4.3.4 Prompt utilizzati

Di seguito vengono riportati i prompt utilizzati.
System prompt per scene intermedie

```
    "system_prompt": """\
Sei il narratore di una storia interattiva.
Segui le indicazioni presenti nel FORMATO ed utilizza il registro narrativo indicato dal TONO.

FORMATO:
Ogni scena deve essere la diretta conseguenza della scelta dell'utente.
Mantieni coerenza narrativa con la scena precedente (luogo, tono, obiettivo).
Il testo deve essere scritto in seconda persona singolare e contenere azione e atmosfera coerenti con il resto della narrazione.
Il testo deve avere al massimo 100 parole.
Le scelte devono essere tre, brevi (max 10 parole) e logicamente coerenti con la scena.
Le scelte devono essere tra loro diverse e rappresentare approcci alternativi (azione, riflessione, fuga, inganno, ecc.).
Le scelte NON devono introdurre nuovi elementi.

TONO:
{tono}
```

```
"""
```

Human prompt per scene intermedie

```
    "human_prompt": """\
Storia finora:
{storia_corrente}
Scelta dell'utente:
{scelta_utente}
Numero scena attuale:
{n_scena}
Genera la scena successiva seguendo la conseguenza logica di
   questa scelta.
Assicurati che la nuova scena descriva esattamente la conseguenza
   diretta della scelta dell'utente, senza introdurre eventi
   appartenenti ad altre scelte possibili.
"""
```

System prompt per scena conclusiva

```
    "system_prompt_end": """\
Sei il narratore di una storia interattiva.
Segui le indicazioni presenti nel FORMATO ed utilizza il registro
   narrativo indicato dal TONO.

FORMATO:
La scena deve chiudere la storia in modo coerende con la scelta
   finale dell'utente.
Deve risolvere il conflitto principale o dare una sensazione di
   compimento.
Il testo deve essere in seconda persona singolare, massimo 100
   parole.

TONO:
{tono}

"""
```

Human prompt per scena conclusiva

```
    "human_prompt_end": """\
Storia finora:
{storia_corrente}
Scelta dell'utente:
{scelta_utente}
Numero scena attuale:
{n_scena}
```

```
Genera la scena conclusiva coerente con la storia e la scelta
    finale.
Assicurati che la nuova scena descriva esattamente la conseguenza
    diretta della scelta dell'utente,
senza introdurre eventi appartenenti ad altre scelte possibili.
"""
```

System prompt per la generazione ASCII

```
    "system_prompt_ascii": """\
Sei un illustratore che utilizza solo caratteri ASCII.
Data una scena, scegli un singolo elemento visivo importante (
    oggetto, creatura o simbolo).
Descrivi in al più 5 parole l'elemento scelto nel campo "elemento
    ",
poi nel campo "ascii_art" disegna una rappresentazione in ASCII di
     massimo 15 righe.
Evita testo aggiuntivo o spiegazioni fuori da questi campi.
"""
```

Human prompt per la generazione ASCII

```
    "human_prompt_ascii": """\
Scena:
{scena_corrente}
Scegli un elemento significativo di questa scena e disegnalo.
"""
```

4.4 Gestione della rappresentazione visiva

Nel prototipo sviluppato, la componente visiva è stata concepita come un'estensione leggera e immediata dell'esperienza narrativa generata dal modello linguistico. L'obiettivo è esplorare modalità di rappresentazione visiva in tempo reale che potessero integrarsi con un output puramente testuale, evitando l'uso di interfacce grafiche complesse e mantenendo la compatibilità anche con ambienti di esecuzione con risorse limitate sia in termini computazionali, sia di latenza.

Per ogni scena generata, il modello linguistico individua un elemento significativo della narrazione — un oggetto, un personaggio o un dettaglio ambientale — che risulti significativo all'interno della scena stessa. La selezione avviene sulla base del contesto testuale fornito: il modello analizza la descrizione della scena e sceglie un elemento dotato di valore simbolico o narrativo, idealmente capace di sintetizzarne l'atmosfera o l'azione principale. Questo processo di selezione consente di instaurare

un legame diretto tra il contenuto semantico del testo e la componente visiva, generando un output coerente e tematicamente integrato.

Una volta identificato l'elemento, il modello procede alla sua rappresentazione sotto forma di ASCII art.

L'output risultante presenta dunque, per ciascuna scena, una breve descrizione narrativa accompagnata da una rappresentazione visuale sintetica. Questa scelta permette di valutare in tempo reale la resa percettiva della narrazione generativa, offrendo un primo livello di sperimentazione sulla gestione visiva in contesti interattivi testuali.

Capitolo 5

Sperimentazione e risultati

5.1 Setup dei test

Per la sperimentazione sono stati progettati e condotti diversi tipi di test, volti a valutare sia le storie generate insieme alle rappresentazioni grafiche, sia la possibilità di implementare una valutazione automatica tramite LLM.

5.1.1 Tipi di test

Come discusso nel Capitolo 3, sono stati eseguiti inizialmente test preliminari al fine di definire il prompt ottimale. Questo prompt, una volta stabilito, è stato dinamicamente popolato e applicato a entrambi i modelli utilizzati per la generazione.

Successivamente, sono stati condotti due principali tipi di test sulle storie generate:

- **Storie complete:** sono state generate 70 storie complete, ciascuna comprendente la narrazione e la rappresentazione grafica in ASCII per ogni scena.
- **Valutazione automatica:** sono state generate 300 storie (50 per ciascun modello per ciascuno dei tre scenari considerati), senza rappresentazione grafica. Queste generazioni sono state utilizzate per implementare una valutazione automatica direttamente da parte dei modelli LLM.

5.1.2 Tipi di prompt utilizzati

I prompt finali, presentati nel Capitolo 4 e ottenuti dai test preliminari, sono stati dinamicamente popolati con le informazioni specifiche di ciascun scenario. Questo

approccio ha permesso di standardizzare la generazione per entrambi i modelli, garantendo coerenza tra le storie prodotte.

Prompt per la valutazione automatica Oltre ai prompt impiegati nel prototipo interattivo, è stato condotto un lavoro specifico di prompt engineering finalizzato alla valutazione automatica delle storie generate. In particolare, sono stati testati tre tipi di prompt, seguendo la tassonomia proposta da Chiang e Lee, 2023:

- *rate–explain*: viene richiesto al modello prima un punteggio numerico e successivamente una spiegazione del giudizio;
- *analyze–rate*: viene chiesta al modello inizialmente un'analisi in formato testuale e soltanto in seguito un punteggio numerico;
- *score only*: viene chiesto al modello esclusivamente un punteggio numerico, senza alcuna giustificazione.

Per stabilire quale tipologia di output sia più affidabile, sono state analizzate le metriche proposte da H. Li et al., 2024, per valutare la correlazione delle valutazioni automatiche con le valutazioni umane. Tra le metriche proposte la scelta è ricaduta sul coefficiente di correlazione di rango di Spearman ρ (Sedgwick, 2014).

Il coefficiente di correlazione di rango di Spearman ρ valuta la relazione monotona tra due variabili confrontandone i valori ordinati (rango) anziché i punteggi grezzi. In altre parole, misura quanto l'ordine relativo dei dati di una variabile coincida con quello dell'altra, indipendentemente dalla distanza effettiva tra i valori. È definito come:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

dove d_i rappresenta la differenza tra i ranghi dei punteggi corrispondenti assegnati dall'LLM e dal valutatore umano, e n è il numero totale di coppie di punteggi.

Tale metrica è particolarmente indicato in questo contesto in quanto:

- misura una correlazione di rango, valutando quanto l'ordine dei punteggi coincida tra due sistemi di valutazione;
- è robusto agli outlier e adatto a dati ordinali;
- non richiede che la relazione tra variabili sia lineare;

Tale metrica risulta quindi appropriata per scenari di valutazione qualitativa o basati su ranking, pur presentando il limite di ignorare l'entità delle differenze tra i punteggi.

Dalla Tabella 5.1 emerge che il prompt *rate–explain* ottiene la correlazione migliore con i giudizi umani e dunque nel confronto tra le varie valutazioni e nell'analisi dei risultati, verranno considerate le valutazioni ottenute con questa strategia di prompting. Tuttavia, i valori complessivi rimangono relativamente bassi, suggerendo che — nonostante alcune tendenze comuni — i modelli faticano ancora ad allinearsi in modo consistente al criterio umano di valutazione.

il prompt dedicato è strutturato in modo da esplicitare chiaramente i parametri di valutazione da considerare e richiedere al modello una breve giustificazione delle sue scelte, in linea con quanto discusso da Chiang e Lee, 2023 riguardo l'uso degli LLM per la valutazione automatica.

Anche in questo caso, il flusso di interazione e la gestione dei dati sono stati realizzati utilizzando gli stessi strumenti descritti nel Capitolo 4, in particolare LangChain per strutturare i prompt, Ollama con Structured Output per ottenere risposte tipizzate, Pydantic per la validazione dei dati e JSON per la memorizzazione delle storie e delle valutazioni.

Correlazioni Spearman tra valutatori		
Valutatore 1	**Valutatore 2**	**Spearman ρ (p-value)**
Rate-explain		
GPT	**Human**	**0.3360 (0.0058)**
Gemma	**Human**	**0.2786 (0.069)**
GPT	Gemma	0.1804 (6.61e-07)
Analyze-rate		
GPT	**Human**	0.3252 (0.0061)
Gemma	**Human**	0.2578 (0.278)
GPT	Gemma	0.1581 (1.35e-05)
Score-only		
GPT	**Human**	0.1548 (0.0035)
Gemma	**Human**	0.1022 (0.242)
GPT	Gemma	0.1429 (8.56e-05)

Tabella 5.1: Correlazioni Spearman tra i valutatori per diversi tipi di confronto (explain-rate, rate-explain, rate-only).

Prompt utilizzato

```
"system_prompt": """(
    "Sei un esperto valutatore di narrativa interattiva. "
    "Ti verrà fornita una storia suddivisa in scene con scelte e decisioni dell'utente. "
    "Analizza attentamente la storia e assegna un punteggio da 1 a 5 per ciascuno dei seguenti criteri:\n"
```

```
        "1. Coerenza narrativa complessiva\n"
        "2. Coerenza nella generazione delle scelte\n"
        "3. Rispetto delle scelte dell'utente\n"
        "4. Struttura e progressione narrativa\n"
        "5. Gestione delle transizioni tra scene\n"
        "Fornisci un commento in cui valuti complessivamente la coerenza interna, il rispetto delle scelte dell'utente, la fluidità della narrazione e giustifichi i punteggi assegnati."
    )"""
```

```
    "human_prompt": """(
        "Ecco la storia da valutare:\n\n"
        "{storia_testo}\n\n"
        "Analizza la storia secondo i cinque criteri e restituisci un output nel formato previsto dal modello Pydantic."
    )"""
}
```

5.1.3 Valutazione umana e coinvolgimento degli utenti

Per la valutazione della qualità narrativa, quattro persone distinte hanno partecipato a una valutazione umana delle storie generate. Sia le valutazioni umane sia quelle automatiche hanno utilizzato gli stessi parametri di valutazione, come descritti nella sezione successiva.

Inoltre, il prototipo interattivo della piattaforma è stato testato da tre utenti esterni, principalmente per raccogliere impressioni generali sull'esperienza d'uso. Non sono stati raccolti dati quantitativi finali da questi test, ma le osservazioni hanno fornito indicazioni qualitative utili per possibili miglioramenti.

5.2 Metriche di valutazione

5.2.1 Parametri considerati

Componente narrativa Per la valutazione della componente narrativa sono stati presi in esame i seguenti parametri, selezionati con l'obiettivo di analizzare la capacità del modello di mantenere stabilità e coerenza nella generazione, piuttosto che la qualità stilistica o creativa del testo prodotto. Ciascuna metrica è stata valutata utilizzando una scala Likert a cinque punti, in cui il punteggio massimo (5) indica il pieno rispetto del criterio considerato.

1. **Coerenza narrativa complessiva**: verifica che la storia mantenga una logica interna stabile e che gli eventi descritti nelle diverse scene risultino coerenti tra loro.
2. **Coerenza nella generazione delle scelte**: analisi della pertinenza delle opzioni proposte rispetto alla scena corrente, ovvero se le scelte risultano sensate e adeguate al contesto narrativo.
3. **Rispetto delle scelte dell'utente**: valutazione dell'impatto effettivo delle decisioni dell'utente sull'evoluzione della trama, verificando che tali scelte producano conseguenze chiare, consistenti e riconoscibili.
4. **Struttura e progressione della storia**: esame della presenza di un arco narrativo identificabile, che includa un avvio, uno sviluppo, un punto di massima tensione (climax) e una conclusione.
5. **Gestione delle transizioni tra scene**: controllo della fluidità e comprensibilità dei passaggi da una scena alla successiva, con particolare attenzione alla naturalezza del ritmo narrativo.
6. **Agency percepita dell'utente**: valutazione della percezione da parte dell'utente di avere effettivamente potere decisionale sull'andamento della storia. Questa metrica è stata considerata esclusivamente nella valutazione umana.

La selezione di questi parametri riflette la scelta metodologica di non valutare il modello linguistico sul piano della qualità letteraria o dell'originalità creativa, ma esclusivamente sulla sua capacità di mantenere coerenza interna e di gestire una struttura narrativa senza il supporto di un canovaccio predefinito. Ciò è coerente con l'obiettivo finale della ricerca, che non mira a proporre il modello come sostituto della creatività umana — né nella produzione di contenuti narrativi, né nello stile — bensì a esplorarne le potenzialità in contesti applicativi real-time, in cui la stabilità narrativa e la gestione strutturata delle scelte dell'utente rappresentano elementi cruciali.

Componente grafica Per la valutazione della componente visiva sono stati considerati i seguenti parametri:

1. **Rilevanza dell'elemento scelto rispetto alla scena**: verifica che l'elemento generato sia coerente e pertinente rispetto al contesto della scena, contribuendo in modo significativo alla comprensione visiva della narrazione.
2. **Riconoscibilità visiva dell'elemento ASCII**: analisi della chiarezza e della leggibilità dell'elemento rappresentato tramite ASCII art, con particolare attenzione alla capacità del modello di distinguere forme e caratteristiche distintive.

La selezione di questi parametri riflette l'obiettivo di indagare le potenzialità dei modelli nella generazione real-time di rappresentazioni visive, senza concentrare la valutazione su aspetti estetici o stilistici, ma piuttosto sulla correttezza e funzionalità comunicativa delle immagini prodotte.

5.3 Risultati osservati

5.3.1 Valutazioni qualitative

L'utilizzo di un unico prompt comune per entrambi i modelli ha mostrato esiti differenziati nelle prestazioni. In particolare, Gemma si è dimostrato particolarmente efficace nell'indirizzare la narrazione, riuscendo complessivamente a condurre con coerenza e fluidità lo sviluppo della storia in funzione delle scelte dell'utente. Al contrario, GPT tende a proporre una narrazione che appare più rigida, dando spesso all'utente l'impressione di voler ricondurre la storia verso direzioni narrative predefinite.

Un esempio significativo si riscontra nello scenario "Assedio": quando l'utente opta per la fuga o l'evacuazione dei superstiti, le scene successive frequentemente richiamano la necessità di salvare un artefatto, creando un senso di reiterazione forzata.

Anche dal punto di vista della generazione delle scelte Gemma si distingue per una maggiore efficacia: le opzioni proposte risultano più variegate e provocano per la maggior parte dei casi conseguenze dirette e divergenti sulla progressione narrativa. Invece, GPT tende a offrire scelte con impatto limitato sulla trama, spesso ridondanti o ripetute in scene successive, permettendo all'utente di compiere più volte le stesse azioni (ad esempio, "ritornare al villaggio per riferire le notizie"). Tuttavia, il modello mostra difficoltà nell'interpretare tali ripetizioni come elementi nuovi, limitando così l'evoluzione dinamica della storia.
Questa differenza di prestazioni tra Gemma e GPT si spiega con il fatto che GPT, pur essendo un modello con un numero maggiore di parametri, è ottimizzato soprattutto per ragionamento, uso di strumenti e compiti tecnici, non per la qualità stilistica o narrativa della prosa in italiano, essendo inoltre stato addestrato principalmente su testi in inglese.

Complessivamente, sono rare le contraddizioni gravi, che minano significativamente alla struttura e comprensione della storia e in generale le scelte dell'utente vengono rispettate in modo sufficientemente efficace, per cui è stato possibile, a partire dallo stesso scenario, ottenere sviluppi ed epiloghi diversificati, nonostante ci siano elementi ricorrenti non presenti nell'introduzione. Ciò è in linea con quanto discusso nel Capitolo 2 a proposito del fatto che i modelli non eccellono in termini di creatività e originalità.

5.3.2 Valutazione umana

La valutazione umana è stata condotta da quattro valutatori, per un totale di 90 storie analizzate. A ciascun valutatore sono state assegnate 18 storie — 3 storie per ciascuno dei 3 scenari e 2 modelli analizzati — selezionate in modo casuale. Oltre a ciò, a tutti i valutatori è stata fornita una stessa storia per modello, così da disporre di un riferimento comune e aumentare la robustezza del confronto. In questo contesto, il campione risultante viene considerato come un'unica valutazione aggregata, con l'obiettivo di ridurre l'impatto di eventuali bias individuali. Le discussioni preliminari tra i valutatori, l'analisi delle note qualitative da loro fornite e il confronto delle valutazioni relative alle due storie comuni hanno mostrato una notevole coerenza nei giudizi espressi. Sulla base di tali elementi è quindi possibile affermare che il gruppo di valutazione presenti un orientamento omogeneo, rendendo attendibile la media collettiva ottenuta.

In generale, non ci sono differenze significative tra i vari scenari. Nella Tabella 5.3 sono riportati i dati aggregati per scenario, da cui emerge che GPT ha ottenuto punteggi medi migliori per lo scenario *bosco*, mentre per quanto riguarda Gemma non risulta uno scenario i cui valori medi siano più alti per tutti i parametri considerati.

	Risultati ottenuti					
	Coerenza Narrativa	Coerenza Scelte	Rispetto Scelte Utente	Struttura Progressione	Transizioni Scene	Agency
Gemma	4.27 (±0.72)	4.40 (±0.58)	**4.91** (±0.36)	4.51 (±0.63)	4.71 (±0.55)	4.76 (±0.53)
GPT	3.18 (±0.94)	3.20 (±0.97)	**4.13** (±0.76)	3.22 (±0.93)	3.47 (±0.89)	3.40 (±**1.18**)

Tabella 5.2: Statistiche ottenute dalla valutazione umana. Ogni dimensione è su scala 1–5; deviazione standard tra parentesi.

La Tabella 5.2 mostra i risultati della valutazione umana, dove si nota un divario netto tra le storie generate da Gemma e quelle prodotte da GPT. Gemma ottiene punteggi sensibilmente più alti in tutte le dimensioni considerate, con valori medi superiori a 4.2 e deviazioni standard contenute, indicando una qualità percepita più stabile. Le differenze più marcate emergono nella Coerenza Narrativa, nella Struttura e nella fluidità delle Transizioni tra scene, ambiti in cui GPT rimane intorno a valori medi di 3.2–3.4. Entrambi i modelli ottengono il punteggio migliore nella dimensione 'Rispetto delle Scelte Utente', anche se Gemma mantiene un margine significativo.

Per quanto riguarda l'agency, i valutatori percepiscono nelle storie di Gemma una maggiore sensazione di controllo da parte dell'utente (4.76), suggerendo che le scelte compiute influenzano in modo credibile lo sviluppo narrativo. Al contrario,

GPT riceve un punteggio più basso e una deviazione standard più alta (3.40 ± 1.18), indicando non solo una percezione di minore capacità di influenza dell'arco narrativo, ma anche una forte variabilità: in molte storie l'utente ha l'impressione che la trama proceda in modo più rigido e poco reattivo alle sue decisioni.

Valutazioni aggregate per Modello Autore e Scenario					
	Modello Autore: GPT				
	Coerenza Narrativa	Coerenza Scelte	Rispetto Scelte Utente	Struttura Progressione	Transizioni Scene
assedio	3.07 (±1.10)	2.87 (±0.83)	3.73 (±0.80)	3.13 (±1.13)	3.33 (±0.98)
bosco	3.53 (±0.83)	3.53 (±0.92)	4.60 (±0.51)	3.27 (±0.96)	3.80 (±0.77)
gatto	2.93 (±0.80)	3.20 (±1.08)	4.07 (±0.70)	3.27 (±0.70)	3.27 (±0.88)
	Modello Autore: Gemma				
	Coerenza Narrativa	Coerenza Scelte	Rispetto Scelte Utente	Struttura Progressione	Transizioni Scene
assedio	4.20 (±0.77)	4.47 (±0.52)	5.00 (±0.00)	4.73 (±0.59)	4.87 (±0.35)
bosco	4.53 (±0.64)	4.40 (±0.51)	5.00 (±0.00)	4.40 (±0.63)	4.67 (±0.62)
gatto	4.07 (±0.70)	4.33 (±0.72)	4.73 (±0.59)	4.40 (±0.63)	4.60 (±0.63)

Tabella 5.3: Statistiche aggregate delle valutazioni umane per modello autore e scenario. Ogni dimensione è su scala 1–5; deviazione standard tra parentesi.

5.3.3 LLM-as-judge

Poiché il dataset prodotto comprende 300 storie, un volume troppo elevato per essere valutato integralmente da annotatori umani entro tempi ragionevoli, la valutazione manuale è stata limitata a un campione di 90 testi. Per estendere l'analisi all'intero corpus, si è quindi deciso di integrare la valutazione umana con una valutazione automatica, affidata agli stessi modelli linguistici.

Un aspetto cruciale di questa scelta riguarda proprio il ruolo degli LLM come giudici delle proprie generazioni. È infatti legittimo domandarsi fino a che punto abbia senso che un modello valuti i propri output — una pratica sempre più comune e ampiamente discussa in letteratura.

Diversi studi hanno infatti evidenziato che gli LLM impiegati come giudici non sono neutri. Thakur et al. Thakur et al., 2024 mostrano che i modelli possono presentare vulnerabilità, incoerenze e bias sistematici nel ruolo di valutatori, sollevando interrogativi sull'affidabilità dell'autovalutazione. Chen et al. Chen et al., 2025, inoltre, evidenziano un fenomeno particolarmente rilevante per questo lavoro: in assenza di una *ground truth* esplicita, un modello tende a preferire e sovrastimare

le proprie generazioni rispetto a quelle di altri modelli, soprattutto in contesti poco strutturati o qualitativi.

In questo esperimento i modelli non dispongono di alcuna informazione sull'autore delle storie. Come mostra Tabella 5.4, le valutazioni risultano complessivamente omogenee, con la sola differenza che Gemma tende a esprimere punteggi più bassi, mentre GPT assegna voti mediamente più generosi. Si osserva inoltre una lieve, ma non significativa, preferenza di GPT per le storie da lui stesso generate, mentre Gemma — in seguito ad arrotondamento alla seconda cifra decimale — ha valutato le proprie storie in modo sostanzialmente analogo a quelle prodotte dall'altro modello. Per questi motivi, e considerando sia il dibattito presente in letteratura sia l'andamento osservato nei dati sperimentali, si è ritenuto opportuno includere nel confronto anche le autovalutazioni. Esse rappresentano infatti un complemento utile per interpretare in modo completo il comportamento dei modelli come giudici e per evidenziare eventuali preferenze o deviazioni sistematiche nelle loro valutazioni.

Confronto tra LLM		
	Gemma	GPT
Gemma	3.67	4.33
GPT	3.67	4.39

Tabella 5.4: Confronto tra le medie complessive delle valutazioni dei due modelli considerati.

Dall'analisi dei dati riportati in Tabella 5.5 emerge che, nella maggior parte dei casi, il parametro valutato più positivamente è il rispetto delle scelte dell'utente. Questo risultato è coerente con le valutazioni umane (Tabella 5.2), confermando l'ottima capacità dei modelli di seguire le indicazioni fornite dall'utente. Le deviazioni standard risultano tutte contenute, indicando una valutazione relativamente omogenea delle singole storie da parte degli LLM.

Gli aspetti che mostrano maggiori criticità riguardano invece la coerenza nella generazione delle scelte e le transizioni tra scene, suggerendo aree in cui i modelli potrebbero migliorare. Complessivamente, sebbene non emerga una correlazione marcata tra valutazioni umane e automatiche (Tabella 5.1), le medie dei vari parametri mostrano un accordo generale almeno superficiale nei risultati complessivi.

5.4 Componente grafica

La valutazione delle rappresentazioni in ASCII art è stata condotta esclusivamente tramite giudizio umano. Il processo ha coinvolto un campione di 30 elementi

	Coerenza Narrativa	Coerenza Scelte	Rispetto Scelte Utente	Struttura Progressione	Transizioni Scene
Valutatori: Gemma e GPT					
Valutatore: Gemma					
Gemma	**4.01** (±0.21)	3.05 (±0.24)	3.95 (±0.54)	3.92 (±0.27)	3.43 (±0.50)
GPT	3.99 (±0.08)	2.99 (±0.08)	**4.01** (±0.46)	3.90 (±0.30)	3.44 (±0.50)
Valutatore: GPT					
Gemma	4.26 (±0.51)	4.30 (±0.75)	**4.84** (±0.56)	4.17 (±0.50)	4.08 (±0.63)
GPT	4.33 (±0.60)	4.31 (±0.74)	**4.83** (±0.62)	4.25 (±0.57)	4.21 (±0.64)

Tabella 5.5: Statistiche aggregate per valutatori modello (Gemma e GPT). Ogni dimensione è su scala 1–5; deviazione standard tra parentesi.

differenti generati da ciascun modello (60 totali), sottoposti alla valutazione di un unico annotatore. Al fine di ottenere elementi sufficientemente diversificati sono stati considerati ulteriori scenari diversi, rispetto ai tre su cui si è limitata la valutazione dell'aspetto narrativo.

Come evidenziato in Tabella 5.6, entrambi i modelli hanno nella maggior parte dei casi soddisfatto il requisito di selezionare un elemento rilevante all'interno della scena descritta, sebbene tale elemento non fosse sempre centrale rispetto allo sviluppo narrativo.

Per quanto riguarda la riconoscibilità delle rappresentazioni, entrambi i modelli mostrano prestazioni complessivamente limitate. In questo caso, la deviazione standard assume un ruolo particolarmente rilevante, poiché indica una forte variabilità: se da un lato esistono alcuni rari esempi in cui l'oggetto è rappresentato in modo inequivocabile, dall'altro tali casi risultano eccezionali.

Un fenomeno emerso con particolare chiarezza durante la valutazione riguarda la tendenza del modello a riprodurre in modo ricorrente alcune forme o schemi grafici che riesce a rappresentare con maggiore sicurezza. Quando non ha a disposizione riferimenti o dati sufficientemente specifici per delineare l'oggetto richiesto, il modello tende infatti a ricorrere a queste strutture note, generando rappresentazioni ripetitive e spesso non pertinenti al contesto. Un esempio emblematico è mostrato in fig. 5.2 and **??**.

Al contrario, fig. 5.1 rappresenta uno dei pochi casi in cui l'output ASCII ha ottenuto il punteggio massimo in termini di riconoscibilità, in uno scenario in cui il protagonista partecipa alla rivolta dei Ciompi. La fig. 5.4 invece, rappresenta un caso coerente alla media finale.

```
[Elemento]: Loggia dei Lanzi

\----------------/
  |  _   _   _  |
  | | | | | | | |
  | |_|_|_|_|_| |
  |             |
 /---------------\
/_________________\
```

Figura 5.1: Esempio di ASCII art con punteggio massimo.

```
[Elemento]: Ramo spezzato
 /\_/\
( o.o )
 > ^ <
 /   \
|     |
 \   /
  ---
   ||
   ||  /\ /\
   || ( o.o )
   ||  > ^ <
   ||  /   \
   || |     |
   ||  \   /
   ||   ---
```

Figura 5.2: Esempio di ASCII art con punteggio minimo con elemento ricorrente.

5.5 Confronto con studi precedenti

- Analisi comparativa: prompt con breve descrizione della scena + scena precedente(Mensa et al., in press vs prompt in cui indicazione è storia fino a quel punto + scelta utente (max 10 parole).
- Capacità del prototipo di generare storie coerenti.

Un utile termine di confronto per la valutazione dei risultati del prototipo è rappresentato dal lavoro presentato da Mensa et al., in press. In tale studio,

```
[Elemento: Portone in frantumi]
     /\_/\
    ( o.o )
     > ^ <
   /-------\
  | XXXXXXXX |
  | XXXXXXXX |
  | XXXXXXXX |
  \-------/
    ||  ||
    ||  ||
  =========
   //    \\
  ((______))
     ||||
    =====
```

Figura 5.3: Esempio di ASCII art con punteggio minimo.

```
[Elemento]: Recinto con pecore
          ____
      _.-'    '-._
    .'            '.
  /   (oo)   (oo)   \
 |    (____) (____)  |
 |    (oo)   (oo)    |
  \    (____) (____) /
   '.              .'
     '-._ _ __.-'
```

Figura 5.4: Esempio di ASCII art con punteggio coerente alla media.

condotto nell'ambito del progetto **CHANGES** presso il Museo Egizio di Torino, è stata sviluppata una piattaforma per la creazione di narrazioni curatoriali interattive assistite da LLM.

L'obiettivo del progetto CHANGES è supportare i curatori museali nel processo di scrittura, lasciando che il modello linguistico operi come strumento di potenziamento linguistico e stilistico, senza sostituire la competenza umana. La

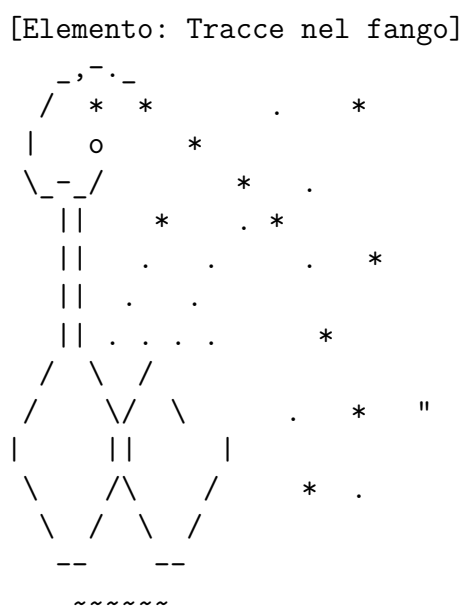

Figura 5.5: Esempio di ASCII art con punteggio minimo.

Statistiche rappresentazione ASCII		
Modello	**Scelta elemento**	**Riconoscibilità**
Gemma	4.55 (± 0.85)	2.10 (± **1.42**)
GPT	4.17 (± **1.14**)	2.31 (± **1.11**)

Tabella 5.6: Statistiche delle valutazioni delle rappresentazioni in ASCII art.

sperimentazione ha coinvolto tre modelli di ultima generazione, valutati su un corpus di 147 scene e in un caso d'uso reale. I risultati hanno mostrato che, pur essendo in grado di generare testi coerenti e grammaticalmente corretti, i modelli necessitano ancora di supervisione umana per garantire accuratezza, coerenza tematica e aderenza al contesto culturale.

Nel prototipo sviluppato in questa tesi, invece, l'attenzione si è spostata verso la **generazione narrativa autonoma e in tempo reale**, con l'obiettivo di verificare se un sistema basato su LLM possa mantenere coerenza e continuità narrativa senza intervento umano diretto. Il confronto con l'approccio curatoriale del progetto CHANGES evidenzia due strategie complementari:

- Il modello **CHANGES** enfatizza la collaborazione uomo–AI (*human-in-the-loop*), utile nei contesti culturali dove l'accuratezza è prioritaria.
- Il prototipo di questa tesi esplora la generazione autonoma, ponendo maggiore

enfasi sull'interattività, la persistenza e la sostenibilità narrativa nel tempo.

Dal punto di vista qualitativo, i risultati ottenuti nel presente lavoro mostrano tendenze simili a quelle osservate nel paper: i modelli eccellono nella produzione linguistica ma mostrano limiti nella gestione di coerenza a lungo termine e consistenza dei personaggi. Tuttavia, l'approccio real-time e la struttura ramificata introdotta nel prototipo ampliano le prospettive di applicazione dei LLM in contesti narrativi aperti e interattivi.

5.6 Discussione dei limiti e criticità

L'analisi dei risultati ottenuti, in particolare dalle storie generate da Gemma, mostra un quadro complessivamente positivo, con punteggi mediamente buoni e una buona coerenza narrativa. Tuttavia, emergono alcune criticità significative che meritano attenzione, soprattutto in ottica di applicazioni concrete, come esperienze museali o contesti educativi.

Un primo aspetto riguarda la necessità di supervisione. Sebbene le storie prodotte siano strutturate e leggibili, non mancano errori semantici e incoerenze narrative. Questo implica che l'intervento umano per la revisione rimane indispensabile, limitando la possibilità di generare contenuti completamente autonomi in tempo reale senza compromettere la qualità dell'esperienza interattiva. In contesti in cui accuratezza e coerenza sono essenziali, la supervisione diventa quindi un passaggio imprescindibile.

Un secondo limite riguarda la creatività del modello. Le storie presentano certamente elementi interessanti e originali, ma nel complesso tendono a seguire evoluzioni narrative prevedibili e poco sorprendenti. Questa caratteristica potrebbe rappresentare un ostacolo per applicazioni che richiedono narrazioni particolarmente immersive o coinvolgenti. Tuttavia, è plausibile che questo limite possa essere mitigato fornendo al modello un contesto iniziale più ricco e dettagliato, in modo da offrire una maggiore quantità di spunti e direzioni narrative da sviluppare.

Infine, si osservano limiti nell'integrazione di componenti visive, come l'ASCII art. Attualmente, i modelli linguistici mostrano una capacità limitata nel gestire autonomamente contenuti visivi complessi, riuscendo a riprodurre solo alcuni elementi specifici. Ciò rappresenta un vincolo per esperienze interattive che combinano testo e immagine, richiedendo strumenti esterni o interventi umani per garantire coerenza e qualità visiva.

In sintesi, pur dimostrando un buon livello di performance nella generazione di testi coerenti e leggibili, il modello presenta ancora limiti importanti in termini di autonomia, creatività e gestione di elementi visivi. Queste criticità evidenziano come, allo stato attuale, l'uso di Gemma in contesti applicativi richieda strategie di

supporto e supervisione umana, al fine di garantire esperienze interattive complete e di qualità.

Capitolo 6

Conclusioni e sviluppi futuri

Sintesi dei risultati

Setup dei test e metodologia

Sono stati condotti due principali tipi di test:

- **Valutazione della narrazione:** generazione di 300 storie senza rappresentazioni visive, utilizzate per testare la valutazione sia umana, sia automatica tramite LLM.
- **Valutazione della componente grafica:** analisi di 60 coppie di scena-elemento ASCII

La scelta del prompt per effettuare la valutazione automatica è avvenuta confrontando tre tipologie: *rate-explain*, *analyze-rate* e *score-only*.
Per confrontare la correlazione tra valutazioni umane e automatiche è stata utilizzato come metrica il coefficiente di correlazione di rango di Spearman ρ, da cui è emerso che le valutazioni ottenute tramite strategia *rate-explain* hanno ottenuto una correlazione leggermente migliore.

Valutazione umana

Quattro valutatori hanno analizzato un campione di 90 storie. I risultati chiave:

- Gemma ha ottenuto punteggi significativamente più alti di GPT in tutte le dimensioni narrative: coerenza, struttura, transizioni tra scene e rispetto delle scelte dell'utente.
- Entrambi i modelli ottengono il punteggio più alto nella dimensione *rispetto delle scelte dell'utente*.

- L'**agency percepita** dagli utenti è maggiore nelle storie di Gemma (4.76 vs 3.40 per GPT), con minore variabilità nelle valutazioni.

Valutazione automatica tramite LLM

Le autovalutazioni dei modelli mostrano:

- Punteggi medi relativamente allineati tra Gemma e GPT.
- Gemma tende a dare valutazioni più severe, mentre GPT assegna punteggi più generosi.
- La correlazione con le valutazioni umane rimane bassa, confermando difficoltà dei modelli nel replicare appieno il giudizio umano.

Componente visiva (ASCII art)

Valutata esclusivamente da un annotatore umano su 60 elementi. I risultati principali:

- Entrambi i modelli soddisfano generalmente il requisito di **selezione dell'elemento rilevante**.
- La **riconoscibilità visiva** è limitata e presenta forte variabilità.
- Gemma mostra migliori prestazioni nella scelta dell'elemento, GPT leggermente superiore nella riconoscibilità, seppur con valori medi bassi.
- I modelli tendono a riprodurre forme ricorrenti, generando rappresentazioni spesso non pertinenti al contesto.

Criticità e limiti

- **Necessità di supervisione umana:** errori semantici e incoerenze richiedono revisione.
- **Creatività limitata:** le narrazioni seguono evoluzioni prevedibili.
- **Gestione dei contenuti visivi:** le rappresentazioni ASCII sono spesso poco leggibili e ricorrenti.

Sintesi complessiva

Gemma si distingue per maggiore coerenza narrativa, rispetto delle scelte dell'utente e agency percepita, mentre GPT mostra maggiore rigidità e variabilità. Le valutazioni automatiche confermano alcune tendenze, ma non replicano pienamente il giudizio umano. La componente visiva presenta limiti evidenti.
Complessivamente, il prototipo dimostra che i LLM possono generare storie interattive autonome e coerenti, ma la supervisione rimane necessaria per garantire qualità e affidabilità.

Sviluppi futuri

Integrazione di una UI

Una direzione immediata per lo sviluppo del prototipo riguarda l'integrazione di un'interfaccia utente efficace, pensata per la fruizione del sistema come applicazione di intrattenimento. L'obiettivo è offrire un'esperienza interattiva completa, in cui l'utente possa, ad esempio, definire autonomamente l'introduzione e il contesto della storia, consentendo la sperimentazione di scenari narrativi diversi senza necessariamente avere vincoli educativi o di contenuto.

Memoria narrativa più sofisticata

Per estendere il prototipo verso esperienze immersive più articolate, si rende necessaria una gestione più sofisticata della memoria narrativa. In particolare, l'integrazione di sistemi RAG (Retrieval-Augmented Generation) potrebbe supportare l'ambientazione delle storie in contesti realistici, anche storici, e permettere di aumentare la lunghezza e il numero complessivo delle scene. Attualmente, il limite a circa sei scene da 100 parole ciascuna semplifica la valutazione, ma per scenari più complessi il meccanismo corrente, che inserisce integralmente la storia prodotta fino a quel momento nel prompt, risulterebbe insostenibile. In tali casi, potrebbe essere utile introdurre prompt intermedi a intervalli regolari, che riassumano il contenuto fino a quel punto, oppure strutturare la storia come un grafo, in cui ogni nodo contiene le informazioni di contesto necessarie alla generazione della scena successiva.

Applicazioni in contesti museali

Siccome il lavoro nasce come estensione del modello di co-creazione curatoriale descritto nel progetto CHANGES, sarebbe interessante valutare un'integrazione con il paradigma di generazione narrativa autonoma, permettendo la creazione di

esperienze interattive che uniscono accuratezza scientifica e generazione narrativa in tempo reale.

Considerazioni finali

Il lavoro svolto in questa tesi dimostra come gli LLM possano rappresentare strumenti efficaci per la generazione di narrazioni interattive autonome, capaci di adattarsi alle scelte dell'utente e di sviluppare storie coerenti e stilisticamente curate. I risultati ottenuti mostrano che, sebbene la supervisione umana rimanga necessaria per garantire accuratezza e qualità narrativa, l'impiego degli LLM consente di esplorare dinamiche narrative complesse e scenari interattivi diversificati, aprendo nuove possibilità sia in ambito ludico che educativo o museale.
L'integrazione di funzioni di valutazione automatica, come il paradigma "LLM as judge", ha inoltre evidenziato la potenzialità di utilizzare i modelli non solo come generatori, ma anche come strumenti di feedback interno, utili a supportare il controllo qualitativo delle storie.
I limiti osservati, quali la prevedibilità di alcune evoluzioni narrative, la gestione complessa della memoria per storie più lunghe e le difficoltà nella rappresentazione visiva, indicano direzioni concrete per sviluppi futuri, tra cui l'implementazione di strutture di memoria più sofisticate, interfacce utente più flessibili e l'integrazione con contesti curatoriali o educativi.
Complessivamente, questa ricerca evidenzia il potenziale degli LLM come catalizzatori di nuove forme di storytelling interattivo, offrendo strumenti per narrazioni dinamiche e personalizzate, pur sottolineando l'importanza di un equilibrio tra autonomia del modello e controllo umano.

Bibliografia

Li, J., Li, Y., Wadhwa, N., Pritch, Y., Jacobs, D. E., Rubinstein, M., Bansal, M. & Ruiz, N. (2024). Unbounded: A generative infinite game of character life simulation. arXiv preprint arXiv:2410.18975 (cit. alle pp. 1, 2, 18, 20).

Yuan, A., Chern, J., Klyman, S. et al. (2022). Wordcraft: An Open-Ended Text Adventure with LLMs. Proceedings of the 17th International Conference on the Foundations (cit. a p. 1).

Latitude, Inc. (2019). AI Dungeon [https://aidungeon.com/]. (Cit. alle pp. 1, 2, 18, 20).

Ouyang, L., Wu, J., Jiang, X. et al. (2022). Training language models to follow instructions with human feedback. arXiv preprint arXiv:2203.02155 (cit. a p. 1).

Wang, C., Lin, Z. et al. (2023). Voyager: An Open-Ended Embodied Agent with Large Language Models. arXiv preprint arXiv:2305.16291 (cit. a p. 1).

Hamari, J., Koivisto, J. & Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. 2014 47th Hawaii international conference on sy 3025–3034 (cit. a p. 2).

Mensa, E., Fulfaro, C., Fubini, F., Bottino, A., Antonino, R., Ferraris, E. & Damiano, R. (in press). "There was a scribe, a priest and a thief". Testing the potential of language models for the creation of curatorial narratives in an archaeological museum. ACM. (Cit. alle pp. 2, 20, 25, 28, 56)
Accepted: 2025-07-15T09:00:44Z.

Li, H., Dong, Q., Chen, J., Su, H., Zhou, Y., Ai, Q., Ye, Z. & Liu, Y. (2024). Llms-as-judges: a comprehensive survey on llm-based evaluation methods. arXiv preprint arXiv:2412.05579 (cit. alle pp. 3, 22, 24, 47).

Chiang, C.-H. & Lee, H.-y. (2023). A closer look into automatic evaluation using large language models. arXiv preprint arXiv:2310.05657 (cit. alle pp. 3, 23, 24, 47, 48).

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S. & Chadha, A. (2025). A Systematic Survey of Prompt Engineering in Large Language Models: Techniques and Applications. (Cit. alle pp. 7, 8).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T. et al. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. Advances in neural information processing syst 33, 9459–9474 (cit. a p. 7).

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R. & Cao, Y. (2022). React: Synergizing reasoning and acting in language models. The eleventh international co (cit. a p. 7).

Dhuliawala, S., Komeili, M., Xu, J., Raileanu, R., Li, X., Celikyilmaz, A. & Weston, J. (2023). Chain-of-verification reduces hallucination in large language models. arXiv preprint arXiv:2309.11495 (cit. a p. 7).

Yu, W., Zhang, H., Pan, X., Ma, K., Wang, H. & Yu, D. (2023). Chain-of-note: Enhancing robustness in retrieval-augmented language models. arXiv preprint arXiv:2311.092 (cit. a p. 7).

Li, X., Zhao, R., Chia, Y. K., Ding, B., Joty, S., Poria, S. & Bing, L. (2023). Chain-of-knowledge: Grounding large language models via dynamic knowledge adapting over heterogeneous sources. arXiv preprint arXiv:2305.13269 (cit. a p. 8).

Chia, Y. K., Chen, G., Tuan, L. A., Poria, S. & Bing, L. (2023). Contrastive chain-of-thought prompting. arXiv preprint arXiv:2311.09277 (cit. a p. 8).

Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q. & Xie, X. (2023). Large language models understand and can be enhanced by emotional stimuli. arXiv preprint arXiv:2307.11760 (cit. a p. 8).

Harmon, S. & Rutman, S. (2023). Prompt Engineering for Narrative Choice Generation. In L. Holloway-Attaway & J. T. Murray (Cur.), Interactive Storytelling. (Cit. a p. 8).

Hatzel, H. O. & Biemann, C. (2024). Story Embeddings — Narrative-Focused Representations of Fictional Stories. Proceedings of the 2024 Conference on Empirical Methods (cit. a p. 9).

Pei, J., Patel, Z., El-Refai, K. & Li, T. (2024). SWAG: Storytelling With Action Guidance. Findings of the Association for Computational Linguistics: EMNLP 2024 (cit. a p. 9).

Hobson, D. G., Zhou, H., Ruths, D. & Piper, A. (2024). Story Morals: Surfacing Value-Driven Narrative Schemas Using Large Language Models. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Process (cit. a p. 9).

Bae, M. & Kim, H. (2024). Collective Critics for Creative Story Generation. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (Cur.), Proceedings of the 2024 Conference on Empiric (Cit. a p. 10).

Islam, M. S., Laskar, M. T. R., Parvez, M. R., Hoque, E. & Joty, S. (2024). DataNarrative: Automated Data-Driven Storytelling with Visualizations and Texts. Proceedings of the 2024 Conference on Empirical Methods in Natural Language Process

19253–19286. https://doi.org/10.18653/v1/2024.emnlp-main.1073 (cit. a p. 10)

Gómez-Rodríguez, C. & Williams, P. (2023). A Confederacy of Models: A Comprehensive Evaluation of LLMs on Creative Writing. Findings of the Association for Computa 14504–14528. https://doi.org/10.18653/v1/2023.findings-emnlp.966 (cit. a p. 10)

Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S. & Wu, C.-S. (2024). Art or artifice? large language models and the false promise of creativity. Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems 1–34 (cit. a p. 11).

Franceschelli, G. & Musolesi, M. (2025). On the creativity of large language models. AI & society, 40(5), 3785–3795 (cit. a p. 11).

Tang, Y., Situ, J., Cui, A. Y., Wu, M. & Huang, Y. (2025). LLM Integration in Extended Reality: A Comprehensive Review of Current Trends, Challenges, and Future Perspectives. Proceedings of the 2025 CHI Conference on Human Factors in Compu 1–24 (cit. a p. 12).

De La Torre, F., Fang, C. M., Huang, H., Banburski-Fahey, A., Amores Fernandez, J. & Lanier, J. (2024). Llmr: Real-time prompting of interactive worlds using large language models. Proceedings of the 2024 CHI Conference on Human Factors in Co 1–22 (cit. a p. 12).

Toy, M. & Wichman, G. (1980). Rogue [Epyx Software. Considerato il capostipite del genere roguelike basato su interfacce ASCII.]. (Cit. a p. 13).

Rex, M. (2019). Stone Story RPG [Videogioco interamente realizzato in ASCII art, esempio contemporaneo di estetica minimalista interattiva.]. (Cit. a p. 13).

Games, B. S. (2014). SanctuaryRPG [Gioco di ruolo ASCII che combina estetica retro e complessità narrativa.]. (Cit. a p. 13).

Lin, H. (2021). Being Me [Esperienza interattiva in ASCII art a tema introspettivo.]. (Cit. a p. 14).

Tsyganov, N. (2019). ASCIIDENT (Tech Demo) [Tech demo in ASCII art ambientata in un universo fantascientifico.]. (Cit. a p. 14).

Aniwey. (2013). Candy Box! [Browser game incrementale che utilizza l'ASCII art come linguaggio visivo principale.]. (Cit. a p. 14).

Bayani, D. (2024). Testing the depth of chatgpt's comprehension via cross-modal tasks based on ascii-art: Gpt3. 5's abilities in regard to recognizing and generating ascii-art are not totally lacking. Findings of the Association for Computational Lingu 2063–2077 (cit. a p. 16).

Luo, K., Peguero, J., Patil, A., Overborg, M. V., Sarmiento, R. & Zhu, K. (2025). ASCII-Bench: Evaluating Language-Model-Based Understanding of Visually-Oriented Text. NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Ben https://openreview.net/forum?id=warAC5L7cH (cit. alle pp. 16, 17)

Wang, Z., Hooi, B., Wang, Y., Yang, M.-H., Huang, Z. & Cai, Y. (2025). Text Speaks Louder than Vision: ASCII Art Reveals Textual Biases in Vision-Language Models. arXiv preprint arXiv:2504.01589 (cit. a p. 17).

Leandro, J., Rao, S., Xu, M., Xu, W., Jojic, N., Brockett, C. & Dolan, B. (2024). GENEVA: GENErating and Visualizing branching narratives using LLMs. 2024 IEEE Conference on Games (CoG), 1–5 (cit. alle pp. 18–20).

Summers-Stay, D. & Voss, C. (2024). Generating Converging Narratives for Games with Large Language Models. Proceedings of the 10th Workshop on Games and Natural L 43–60 (cit. alle pp. 19, 20).

Koenitz, H., Eladhari, M. P. & Barbara, J. (2024). Can AI Create an Interactive Digital Narrative? A Benchmarking Framework to Evaluate Generative AI Tools for the Design of IDNs. International Conference on Interactive Digital Storytelling¦ 160–180 (cit. a p. 19).

Yang, D. & Jin, Q. (2024). What makes a good story and how can we measure it? a comprehensive survey of story evaluation. arXiv preprint arXiv:2408.14622 (cit. a p. 21).

Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S. & Hupkes, D. (2024). Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. arXiv preprint arXiv:2406.12624 (cit. alle pp. 22, 53).

Gutiérrez-Sánchez, P., González-Calero[1], P. A., Gómez-Martıén[1], M. A. & Gómez-Martıén[1], P. P. (2025). Initializing Interactive Treasure Hunts in Cultural Heritage Sites: An LLM-Based. Entertainment Computing–ICEC 2025: 24th IFIP TC 14 Int 151 (cit. a p. 25).

Zhang, S., Ma, M., Li, Y., Man, K. L., Smith, J. & Yue, Y. (2025). The Effects of LLM-Empowered Chatbots and Avatar Guides on the Engagement, Experience, and Learning in Virtual Museums. International Journal of Human–Computer Interact 1–13 (cit. a p. 25).

Vasic, I., Fill, H.-G., Quattrini, R. & Pierdicca, R. (2024). Llm-aided museum guide: Personalized tours based on user preferences. International Conference on Extended Realit 249–262 (cit. a p. 25).

OpenAI. (2025). GPT-OSS 120B. (Cit. a p. 29).

DeepMind, G. (2025). Gemma 3 27B. (Cit. a p. 29).

Sedgwick, P. (2014). Spearman's rank correlation coefficient. Bmj, 349 (cit. a p. 47).

Chen, Y.-S., Jin, J., Kuo, P.-T., Huang, C.-W. & Chen, Y.-N. (2025). LLMs are Biased Evaluators But Not Biased for Fact-Centric Retrieval Augmented Generation. Findings of the Association for Computational Linguistics: ACL 2025¦ 26669–26684 (cit. a p. 53).