Usare i Large Language Model per l'analisi del testo narrativo: strategie di prompt engineering per il riconoscimento del discorso indiretto libero nella narrativa italiana 1830-1930

Aurora Argenzio, Fabio Ciotti, Anna Chiara Corradino
Università di Roma -Tor Vergata, Italia
argenzioaurora@gmail.com; fabio.ciotti@uniroma2.it; annachiara.corradino1@gmail.com

ABSTRACT (ITALIANO)

Questo studio presenta i risultati di un esperimento di prompt engineering condotto nell'ambito del progetto "Leggere il romanzo italiano a distanza (1830-1930) (RIND)". Il progetto intende utilizzare metodi computazionali per rivalutare la periodizzazione tradizionale della letteratura italiana, concentrandosi in particolare sul fenomeno del discorso indiretto libero (DIL) come marcatore di cambiamento stilistico e narrativo del modernismo italiano. RIND ha predisposto un corpus di 1.000 testi letterari, di cui 500 romanzi italiani originali e 500 traduzioni. Per il presente esperimento è stato selezionato un sottoinsieme di 100 testi, bilanciando anno di pubblicazione e lunghezza dei testi. La metodologia ha combinato l'annotazione manuale di circa 3.000 frasi con test sistematici su modelli GPT-4 e Claude 3.5 Sonnet. L'esperimento ha utilizzato strategie di prompting, confrontando approcci zero-shot e *few-shot*, (con e senza spiegazioni preliminari del fenomeno linguistico, sia in italiano che in inglese). I risultati mostrano che i prompt few-shot ottengono le migliori prestazioni nell'identificazione del DIL. Abbiamo anche rilevato che la lingua dei prompt nell'analisi dei testi originali italiani non produce oscillazioni significative nei risultati sull'identificazione di DIL. La lunghezza ottimale del testo per l'analisi si è stabilizzata tra i 1.800 e i 5.000 caratteri. La tokenizzazione per *sentence* si è rivelata l'approccio più efficace per rilevare i marcatori linguistici e sintattici del DIL. Nel 50% dei casi di errore, le interpretazioni alternative proposte dai modelli erano comunque contestualmente valide, mostrando la complessità e le capacità dell'analisi letteraria automatizzata. Lo studio intende fornire un'ipotesi iniziale di framework metodologico per l'analisi di tecniche narrative attraverso modelli linguistici. I risultati suggeriscono l'importanza di elaborare una tassonomia chiara, ma flessibile, dei prompt utilizzabili per l'analisi dei testi.
Parole chiave: prompt engineering; discorso indiretto libero; analisi narrativa computazionale; Large Language Models; Romanzo Italiano

ABSTRACT (ENGLISH)

Using Large Language Models for Narrative Text Analysis: Prompt Engineering Strategies for Identifying Free Indirect Speech in Italian Fiction (1830–1930)

This abstract presents the findings of a prompt engineering experiment conducted as part of the "Reading the Italian Novel at a Distance (1830–1930) (RIND)" project. The project applies computational methods to reassess traditional periodisations of Italian literature, focusing on free indirect speech (FIS) as a marker of stylistic and narrative shifts in Italian Modernism. RIND includes a corpus of 1,000 texts—500 original Italian novels and 500 translations. For this specific experiment, a subset of 100 texts was selected, balanced by publication year and length. The methodology combined manual annotation of 3,000 sentences with systematic tests on GPT-4 and Claude 3.5 Sonnet using zero-shot, one-shot, and few-shot prompting strategies, with and without prior explanations of the phenomenon in Italian and English. Few-shot prompts performed best in identifying FIS, and the language of the prompts did not significantly affect results. Text lengths between 1,800 and 5,000 characters proved optimal for analysis, with sentence tokenization yielding the most accurate detection of FIS's linguistic and syntactic markers. In 50% of errors, the models' alternative interpretations were still contextually valid, highlighting the complexity and potential of automated literary analysis. This study emphasizes the need for a clear yet adaptable taxonomy of prompts to enhance text analysis.
Keywords: prompt engineering; Free Indirect Speech; LLMs; Italian Novel; computational narratology.

1. INTRODUZIONE

L'introduzione dei modelli linguistici di grandi dimensioni (Large Language Model, o LLM) generativi, oltre a costituire un sostanziale avanzamento nel campo dell'Intelligenza Artificiale, rappresenta un'occasione di profonda innovazione per gli studi letterari computazionali. In linea generale i modelli generativi, a

differenza dei sistemi di reti neurali discriminativi, producono nuovi dati sulla base della distribuzione di probabilità dei dati di input. Nei modelli linguistici questa modalità operativa si traduce nella capacità di produrre sequenze linguistiche sintatticamente e semanticamente coerenti con una sequenza testuale di input, detta *prompt*. Con opportune tecniche di training, gli LLM si sono dimostrati in grado di rispondere in modo appropriato a domande o istruzioni, rivelando una imprevista capacità di adattarsi a task complessi non esplicitamente previsti nella fase di addestramento, soprattutto se la formulazione dei prompt viene progettata in modo adeguato (Ciotti, 2023). Di conseguenza negli studi sulle proprietà degli LLM ha assunto rilevanza la definizione di tecniche di progettazione dei prompt, o *prompt engineering* (Mollick, 2023; Wei et al., 2023).
Questo lavoro intende presentare i risultati di una sperimentazione nell'uso degli LLM per condurre sofisticate analisi di testi narrativi, spesso eccedenti le possibilità offerte delle tecniche di machine learning tradizionali (Byszuk et al., 2020), concentrandosi in particolare sulle strategie di *prompt engineering* e *prompt implementation*. La nostra sperimentazione si colloca nell'ambito del progetto di ricerca "Leggere il romanzo italiano a distanza (1830-1930)" (RIND) finanziato dal programma PRIN2022 del MUR. L'obiettivo generale del progetto è quello di utilizzare metodi computazionali e quantitativi per rivalutare e validare sulla base di evidenze empiriche la tradizionale periodizzazione della letteratura italiana dal 1800 ai primi anni del 1900. L'arco temporale scelto abbraccia l'emergere del romanzo in Italia, il dominio del romanzo storico e realista e l'ascesa del Modernismo. A questo fine sono state individuate due tipi di features testuali che la storiografia e la critica letteraria hanno considerato fondamentali per tracciare i cambiamenti diacronici nelle strategie narrative e dunque nei periodi letterari: la prima è la presenza caratterizzante del discorso/pensiero riportato (DR) e del discorso indiretto libero (DIL), una classe di fenomeni morfosintattici che gli studi indicano come strategie discorsive volte a far emergere la coscienza e il pensiero dei personaggi riducendo la mediazione del narratore; la seconda riguarda l'ambientazione narrativa e i ruoli assunti dagli attori umani all'interno dei testi, con l'obiettivo di studiare il riflesso dell'evoluzione sociale nelle transizioni tra i periodi letterari. In questa sede ci concentriamo sulle analisi condotte sul primo insieme di fenomeni, illustrando i risultati preliminari emersi da numerosi test di prompt engineering condotti sistematicamente sui campioni testuali scelti – sulla base di preciso bilanciamento - nell'ambito del corpus di riferimento del progetto. Gli LLM adottati nei test sono i modelli più capaci e performanti attualmente disponibili (sebbene gli esperimenti siano stati condotti anche su altri modelli): Open AI GPT-4o e Claude 3.5 Sonnet.

2. STATO DELL'ARTE

La periodizzazione storico-critica delle correnti letterarie è da sempre oggetto di accese controversie. René Wellek scriveva che "the concept of period is certainly one of the main instruments of historical knowledge" (Wellek, 1956: 268), tuttavia, la cultura opera come un sistema dinamico, rendendo problematica l'imposizione di rigidi confini di periodizzazione temporale. Le sfide sono ulteriormente complicate dai limiti epistemologici del tradizionale close-reading, che per lungo tempo ha foraggiato i tentativi di periodizzazione degli studiosi di letteratura. Negli ultimi decenni, l'emergere di metodologie quantitative ha arricchito gli studi letterari, portando all'adozione di approcci come il distant-reading, introdotto da Franco Moretti (2013; Ciotti, 2022a), che sfidano le consuete periodizzazioni letterarie, come dimostrato da Moretti (2005), Jockers (2013: cap. 6), Piper (2018: cap. 4), Underwood (2019) per la narrativa inglese, Jannidis e Lauer (2014) per quella tedesca e Ciotti (2022b) in un esperimento sulla letteratura italiana. Nell'ambito specifico della storia letteraria italiana, critici e storici di spicco (Castellana, 2010; Donnarumma, 2012; Luperini, 2018; Tortora, 2018; Cangiano, 2018) hanno tradizionalmente, sebbene variamente, datato l'emergere del Modernismo italiano all'inizio del XX secolo. Tra le molte caratteristiche che segnano questa transizione, il discorso indiretto libero (DIL) spicca come particolarmente significativo; Il DIL è uno dei tipi di Discorso Riportato (DR) il cui scopo è quello di fondere la voce del narratore con quella del personaggio: esso è un tratto distintivo della letteratura europea di inizio Novecento, che riflette una crescente attenzione alla profondità psicologica e all'interiorità dei personaggi narrati. Tuttavia, la presenza di DIL che precedono la canonica periodizzazione letteraria mette in discussione l'idea di una netta divisione tra Verismo e Modernismo, suggerendo un'evoluzione stilistica più graduale che può essere computata proprio attraverso l'analisi della sua presenza in testi che precedono l'inizio del Novecento. Basandosi su intuizioni teoriche (Calaresu, 2000; Sullet-Nylander et al., 2014) sul DIL e sulla distinzione dei diversi DR, la nostra sperimentazione ha inteso esplorare le capacità dei modelli linguistici di individuare le occorrenze di DIL in un corpus di romanzi e racconti italiani. Dopo una fase di test preliminari condotti su GPT-4o e su Claude 3.5 Sonnet, su porzioni di testo variabili tra i

1.800 e i 5.000 caratteri, abbiamo deciso di valutare in modo sistematico varie strategie di prompt engineering (Bommasani et al., 2021; Liu et al., 2021) elaborate nella ricerca più recente sugli LLMs (Bach et al. 2022).
Il concetto di prompting è generalmente definito come il processo di istruzione di una IA generativa affinché esegua un compito ben preciso. Per "prompt engineering" (o prompt design) intendiamo qui "the activities of tailoring a prompt to a specific task and utilizing the possibilities of the respective generative AI model" (Böhmker et al., 2023: 559). Data l'assenza di modelli fissi per la progettazione dei prompt, la facilità di formularli attraverso il linguaggio naturale e soprattutto la velocità di cambiamento dei modelli utilizzati, la progettazione dei prompt è un processo soggetto a un notevole grado di arbitrarietà (Zamfirescu-Pereira et al., 2023). Recenti studi hanno fatto emergere l'importanza delle componenti singole dei prompt e il loro impatto sui risultati prodotti (Liu et al., 2021) e mostrato l'inefficacia di prompt molto lunghi e mancanti di concatenazione logico-consquenziale (Wu et al., 2022). Pertanto, la classificazione tassonomica dei prompt (come quelle di Böhmker et al., 2023; Kundisch et al., 2021; Nickerson et al., 2013) risulta un punto di partenza necessario e utile per ottenere risultati soddisfacenti da analisi narratologiche che comprendono porzioni di testo lunghe, intricate e spesso protette da copyright. I diversi approcci quantitativi e qualitativi di costruzione dei prompt, con particolare riguardo ai prompt zero, one- e few- shot (si veda *ultra*), hanno mostrato applicazioni adattabili a diverse tipologie di file, testo e contesto forniti (cf. anche Dang et al. 2022; Kojima et al. 2022).

3. IL CORPUS E LA METODOLOGIA

Il nostro esperimento, nella sua visione più generale, mira a indagare quantitativamente la presenza del DIL e di altre tracce di interiorità nei testi letterari italiani pubblicati tra il 1830 e il 1930. A tale fine è stato costituito un corpus cercando di conseguire la massima rappresentatività possibile, stante le intrinseche difficoltà nella definizione di campione rappresentativo per la tradizione letteraria (Bode, 2018; Moretti, 2017). Il corpus RIND è costituito da 1.000 romanzi, equamente suddivisi tra 500 italiani e 500 traduzioni in italiano. Il corpus è stato bilanciato in modo da garantire una rappresentazione completa della produzione letteraria italiana degli anni presi in analisi, considerando fattori quali il genere dell'autore, la provenienza geografica e l'appartenenza al canone. Per questo esperimento specifico, abbiamo escluso le traduzioni per concentrarci esclusivamente sulla prosa italiana originale. Ogni testo è stato preparato in formato solo testo con codifica UTF-8 per garantire uniformità e compatibilità con gli strumenti di calcolo. Dal corpus iniziale, abbiamo selezionato un sottoinsieme bilanciato (per anno e per genere) di testi per un'analisi dettagliata del DIL: 40 testi e per ogni testo abbiamo selezionato porzioni di circa 5000 caratteri a testo, estratti randomicamente da un programma Python, per poter sottoporre i diversi chunk all'analisi di GPT e di Claude attraverso differenti tipologie di prompt con l'obiettivo di identificare il DIL nei testi forniti.
Come fenomeno linguistico, i DIL si estendono oltre i confini della frase e spesso mancano di marcatori linguistici espliciti che ne segnalino inizio e fine, ponendo una sfida per l'analisi quantitativa classica del fenomeno. Gli approcci NLP tradizionali non riescono infatti a identificare sistematicamente i DIL nei testi letterari (Brunner et al., 2019; Tu et al., 2019; Taivalkoski-Shilov, 2019). Per affrontare queste difficoltà, abbiamo utilizzato i LLM in modo da rilevare e classificare i tipi di discorso riportato - diretto, indiretto e indiretto libero - sulla base dei verbi del discorso, dei marcatori pragmatici e degli indizi contestuali. In particolare, abbiamo testato i modelli Open AI GPT-4o e Claude 3.5 Sonnet per mappare quantitativamente l'evoluzione del DIL nei vari periodi letterari. Inizialmente abbiamo definito l'oggetto dell'analisi, identificato i gruppi di utenti target e specificato lo scopo della tassonomia. Abbiamo costruito diversi tipi di prompt basandoci sui criteri esposti qui nel paragrafo 4 e seguendo parzialmente la classificazione tassonomica e la metodologia di Böhmker et al. (2023). I test sono stati eseguiti sia su campioni di testo tokenizzati per frase e archiviati in formato tabellare (CSV), sia su porzioni di testo dei file in formato "solo testo" dei romanzi.
La fase iniziale del nostro esperimento si è concentrata sulla valutazione dell'accuratezza e delle capacità degli LLM nell'identificare il DIL all'interno del nostro sotto-corpus di riferimento. Per avere un raffronto verificabile, abbiamo taggato manualmente, evidenziando la presenza dei diversi DR, i 40 testi citati. Questo set di dati annotati ha fornito un'importante base di confronto per la valutazione delle prestazioni dei modelli e posto le basi per la seconda fase dell'esperimento, in cui si prevede di mettere a punto almeno due modelli più piccoli e di libero accesso e di confrontare i risultati con quelli dei grandi modelli testati finora. Data la variabilità della struttura e dello stile narrativo nel nostro corpus, abbiamo tentato diverse strategie di tokenizzazione per determinare l'unità più efficace per rilevare i DIL: a livello di frase,

di paragrafo e di porzioni testo significative (intorno ai 5000 caratteri), utilizzando un programma di approssimazione Python per definire porzioni di testo comparabili. Dopo un test comparativo, la tokenizzazione per *sentence* (sebbene la definizione stessa di "sentence" sia complessa) è emersa come l'approccio più efficiente per catturare i marcatori linguistici e sintattici dei DIL che operano a livello di concatenazioni di frasi. Utilizzando il set di dati taggati manualmente come "verità di base", abbiamo poi testato e confrontato ChatGPT (in particolare lo strumento di analisi dei dati di OpenAI) e Claude 3.5 Sonnet. Il processo di test prevedeva l'inserimento di testo in questi modelli e la loro interrogazione per classificare le istanze di discorso diretto, indiretto e DIL in base a criteri linguistici predefiniti. Successivamente, i risultati dei modelli sono stati sistematicamente confrontati con i dati taggati manualmente per valutarne l'accuratezza e l'affidabilità nel riconoscimento dei DIL. Abbiamo costruito diversi tipi di prompt e utilizzando un approccio *empirical-to-conceptual* basandoci sui seguenti livelli di analisi e seguendo la classificazione tassonomica di Böhmker et al. 2023 per poter costruire una nuova tassonomia che intendiamo proporre per l'analisi del romanzo italiano tra Ottocento e Novecento:

a) Analisi comparativa dei prompt zero-shot (tab.1) e few-shot (tab.2): i.e. Richiesta di analisi di porzione di testo intorno ai 5000 caratteri senza e con esempi tratti dalla letteratura critica di base sul tema.
b) Analisi di precisione testuale e della progressione human-in-the-loop (tab. 3.) con richiesta di spiegazione preliminare del fenomeno linguistico DIL.
c) Analisi dei diversi esempi da fornire da GPT e Claude singolarmente, con il prompt di riferimento few-shot e human-in-the-loop.

Tab. 1. Zero-Shot Prompts

Tipo	Italiano	English
Base	Leggi e analizza il seguente testo e identifica eventuali passaggi in discorso indiretto libero [allega .csv/.txt]	Read and analyze the following text and identify any instances of free indirect speech: [attach .csv/.txt]
Dettagliato	Leggi il testo nel file CSV. Per ogni frase, valuta se contiene un discorso indiretto libero. Creare un nuovo file CSV che includa una colonna aggiuntiva etichettata "Discorso indiretto libero" con una voce "sì" o "no" per ogni frase, indicando la presenza o l'assenza di discorso indiretto libero. rinominare il file con il nome del file [allega .csv]	Read the text in the CSV file. For each sentence, evaluate if it contains free indirect speech. Create a new CSV file that includes an additional column labeled "Free Indirect Speech" with a "yes" or "no" entry for each sentence, indicating the presence or absence of free indirect speech. rename the file with the file name:[attach .csv]

Tab. 2. Few-Shot Prompts

Tipo	Italiano	English
Multi-esempio	Ecco tre esempi di DIL da diversi autori: [esempi]. Usando questi come riferimento, leggi e analizza il seguente testo: [attach -.csv/.txt]	Here are three examples of FIS from different authors: [examples]. Using these as reference, read and analyze the following text: [attach .csv/.txt]
Comparativo GPT	Leggi e riscrivi tutto il seguente testo indicando con "si" e "no" se è presente il discorso indiretto libero e sistemalo in una tabella [allega .txt]; Riconduci l'analisi di cui sopra alla luce di questi esempi: [esempi] [allega .txt]	Read and rewrite the whole of the following text, indicating with "yes" and "no" whether free indirect speech is present and arrange it in a table [attach .txt]; Reconduct the above analysis in the light of these examples: [examples].
Comparativo Claude	Individua il Discorso indiretto libero nel seguente testo, questi sono alcuni esempi: [esempi] [allega csv].	Identify Free Indirect Speech in the following text, these are some examples: [examples], attach csv.

Tab. 3. Prompt con Progressione Human-in-the-Loop

Fase	Italiano	English
Iniziale	Ciao! Spiegami cos'è il discorso indiretto libero	Hi, tell me what's Free Indirect Speech
Verifica	Individualo in questo csv e riscrivi un altro csv con il testo con una colonna si no in corrispondenza delle frasi con DIL	Analyse this file and produce a csv file with an additional column "Free Indirect Speech" with "yes/no" label
Raffinamento	Rivedi la tua analisi considerando anche [esempi]:	Review your analysis considering also [examples]:
Conclusione	Fornisci una nuova classificazione finale dei passaggi, indicando il grado di certezza per ciascuna identificazione:	Provide a final classification of passages, indicating the degree of certainty for each identification:

Questo approccio sistematico, che combina tokenizzazione a livello di frase, taggatura manuale e test comparativi dei prompt, ci ha permesso di costruire un framework metodologico solido, ma flessibile, per l'identificazione del DIL. La nostra esperienza dimostra come un'attenta progettazione delle fasi di analisi, dalla scelta dell'unità testuale fino alla costruzione dei prompt, sia fondamentale per ottenere risultati affidabili nello studio computazionale di fenomeni narrativi complessi.

4. RISULTATI DELL'ESPERIMENTO E DISCUSSIONE

L'analisi quantitativa della progressione metodologica, dalla baseline zero-shot fino all'implementazione few-shot, ha mostrato un incremento significativo nelle performance predittive sia di GPT sia di Claude, ma soprattutto di quest'ultimo. I prompt zero-shot hanno mostrato una tendenza alla sovra-classificazione, con tassi di non corrispondenza rispetto al tagging manuale del 33,7% per Claude S.3.5 e del 35,0% per GPT 4.o. I prompt few-shot hanno mostrato un miglioramento significativo nelle metriche di valutazione, con una riduzione della non corrispondenza: Claude è passato dal 33,7% (zero-shot) al 24,6% (few-shot), mentre GPT è passato dal 35,0% (zero-shot) al 34,3% (few-shot). In termini di corrispondenza con il tagging manuale, Claude S.3.5 ha raggiunto il 75,4% nella configurazione few-shot, mostrando il miglioramento più significativo tra i modelli testati. I grafici (Figura 1) mostrano chiaramente i cambiamenti dall'analisi da zero-shot a few-shot rispetto alla nostra baseline (il nostro tagging). Su un totale di 975 frasi analizzate con Claude in modalità zero-shot, 640 (66,3%) corrispondevano al tagging manuale, mentre con l'approccio few-shot, su 686 frasi analizzate, 517 (75,4%) mostravano corrispondenza. Per GPT 4.o, su 1108 frasi analizzate in modalità zero-shot, 720 (65,0%) corrispondevano al tagging manuale, mentre con l'approccio few-shot, su 1263 frasi, 830 (65,7%) mostravano corrispondenza.

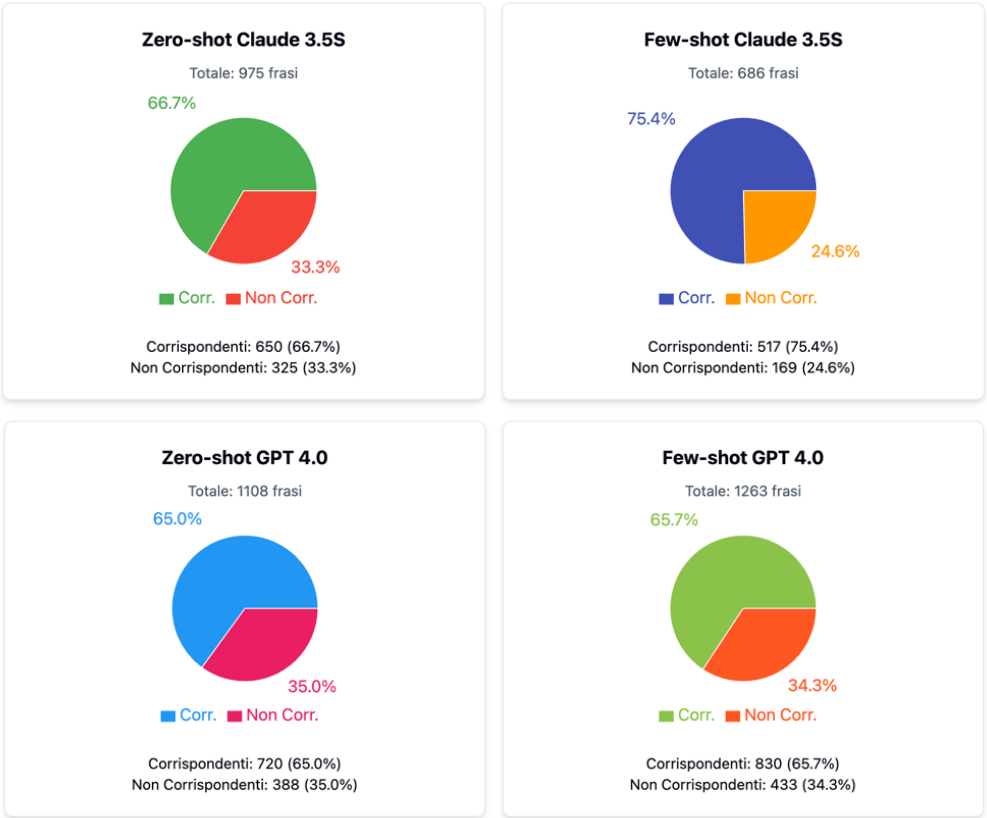

Figura 1 Risultati progressione Zero>Few-shot con GPT e Claude

In una seconda analisi dei dati (analisi in questo caso non automatizzata) abbiamo osservato che l'approccio few-shot, arricchito da una formalizzazione teorica preliminare del fenomeno linguistico target tramite esempi di varia natura ha permesso agli LLM di sviluppare rappresentazioni più granulari del DIL, ottimizzando la discriminazione tra le diverse tipologie di discorso riportato.

L'analisi contrastiva dei verbi di comando nelle due lingue target ha rivelato una correlazione positiva tra la coerenza linguistica prompt-testo e l'accuratezza predittiva (i.e. è preferibile utilizzare prompt in italiano su testi in italiano, sebbene la variazione non sia così sostanziale nei risultati, ma test aggiuntivi si richiedono per fornire questa affermazione di solidità scientifica). I risultati non sono qui riportati, ma sono stati eseguiti anche test con prompt strutturalmente più complessi, dimostrando che la variazione verbale abbia una ovvia ricaduta sui risultati forniti da GPT e Claude. In particolare si riportano qui due problematiche interessanti: nel caso di Claude l'utilizzo dei verbi "rewrite" e "riscrivi" pone problemi dal punto di vista di copyright (l'output di alcuni testi viene bloccato e si è rimandati alla copyright policy page) aggirabili parzialmente fornendo testi più brevi (intorno ai 1800 caratteri); mentre nel caso di GPT l'utilizzo del verbo "analyze" in coppia con i file CSV porta GPT a creare un programma di tokenizzazione Python per l'analisi del DIL secondo alcune regole arbitrarie e sistematiche che tuttavia eludono l'elemento contestuale; si è pertanto deciso, nel caso di GPT di fornire chunk di testo .txt per aggirare il problema e di far tokenizzare il testo per sentence direttamente a GPT (per questo si troverà nella tabella finale una variazione significativa nel numero di frasi analizzate tra Claude e GPT).
Il processo human-in-the-loop è stato fondamentale per l'ottimizzazione dei prompt. Attraverso più di 100 cicli di feedback, sono stati identificati e corretti pattern di errore ricorrenti, conducendo alla definizione di prompt sempre più robusti e *context-aware*. Tuttavia, l'utilizzo di questa metodologia ha anche dimostrato che il processo stesso sia utile per la formazione e creazione di prompt concretamente più performanti, sebbene non lo sia a livello di analisi vera e propria. Dalle nostre analisi, il coinvolgimento dialogico porta, infatti, sia GPT sia Claude a un alto livello di sovrainterpretazione del testo.
La finestra ottimale per l'analisi testuale si è attestata nell'intervallo 1.800-5.000 caratteri (per motivi di gestibilità del documento da parte delle IA e per motivi di copyright, come detto), con una degradazione non lineare dell'accuratezza oltre tale threshold. Nel 50% dei casi di classificazione scorretta, le interpretazioni alternative generate dagli LLM presentavano una validità ermeneutica significativa secondo la nostra annotazione manuale: è stato rilevato inoltre che nel 32% di casi l'annotazione positiva di DIL operata dagli LLM fosse giustificabile contestualmente e in parte superasse le sviste del tagging manuale, fornendo importanti spunti di riflessione sui labili confini del fenomeno preso in analisi e sulla necessità di un più ampio spettro di annotatori manuali. Le sfide più complesse riguardano l'identificazione dei marcatori DIL *low-salience* e la gestione dell'ambiguità interpretativa, aspetti che richiedono non solo un ulteriore raffinamento delle strategie di prompting, ma anche l'implementazione di tecniche di tokenizzazione semanticamente informate.

5. CONCLUSIONI

I risultati dell'esperimento mostrano la capacità degli LLM di applicare strategie di classificazione su fenomeni intrinsecamente complessi e sfumati, soggetti a variabilità interpretativa anche da parte di esperti umani, aprendo spazi di sperimentazione senza precedenti per lo studio su vasta scala di testi complessi come il romanzo, come mostrano anche le ricerche di (Piper & Bagga, 2024) e di (Underwood, 2023) - ma anche il testo filosofico, si veda ad esempio (Mosca, 2024). Allo stesso tempo esso evidenzia come queste capacità analitiche siano soggette a una forte variabilità e dipendenza dal contesto (tecniche di prompt, tipo e dimensione dei frammenti di testo, basi teorico-critiche labili etc.); a questo si aggiunge alla ben nota mancanza di perspicuità del comportamento in fase di inferenza delle reti neurali di grandi dimensioni pone sostanziali problemi di validazione, affidabilità e accettabilità epistemica. In questa ottica, se ovviamente è necessario innalzare la soglia di attenzione nella valutazione critica dei risultati da parte dei ricercatori, è anche opportuno esplorare l'efficacia di modelli linguistici di medie o piccole dimensioni, soggetti ad opportune di tecniche di fine tuning verticali su dati di training di alta qualità. Nell'ambito del progetto RIND abbiamo iniziato a sondare queste possibilità, nei limiti concessi da una scala progettuale limitata come quella del programma PRIN. Considerata tuttavia la complessità e i costi per lo sviluppo e l'utilizzo di sistemi di AI generativa, consideriamo vieppiù opportuno promuovere sperimentazioni su frammenti piccoli, ma significativi dell'analisi, come nel caso delle analisi qui proposte sul prompt engineering.

BIBLIOGRAFIA

Bach, S. H., Sanh, V., Yong, Z. X., Webson, A., Raffel, C., Nayak, N. V., Sharma, A., Kim, T., Bari, M. S., Févry, T., Alyafeai, Z., Dey, M., Santilli, A., Sun, Z., Ben-David, S., Xu, C., Chhablani, G., Wang, H., Fries, J. A., ... Rush, A. M. (2022). Promptsource: An integrated development environment and repository for natural language prompts. arXiv, abs/2202.01279.

Bode, K. (2018). *A world of fiction: Digital collections and the future of literary history*. University of Michigan Press.

Böhmker, M., Greve, M., Kegel, F., Kolbe, L. M., & Beyer, P. E. (2024). Can (A)I have a word with you? A taxonomy on the design dimensions of AI prompts. Proceedings of the Annual Hawaii International Conference on System Sciences.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., & Liang, P. (2021). On the opportunities and risks of foundation models.

Brunner, A., Tu, N. D. T., Weimer, L., & Jannidis, F. (2019). Deep Learning for Free Indirect Representation. In Proceedings of the 15th Conference on Natural Language Processing (KONVENS 2019): Short Papers (pp. 241-245). Erlangen, Germany. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/9315

Byszuk, J., Woźniak, M., Kestemont, M., Leśniak, A., Łukasik, W., Šeļa, A., & Eder, M. (2020). Detecting Direct Speech in Multilingual Collection of 19th-century Novels. In R. Sprugnoli & M. Passarotti (A c. Di), *Proceedings of LT4HALA 2020—1st Workshop on Language Technologies for Historical and Ancient Languages* (pp. 100–104). European Language Resources Association (ELRA). https://aclanthology.org/2020.lt4hala-1.15/

Calaresu, E. (2000). Il discorso riportato. Una prospettiva testuale. Il Fiorino.

Cangiano, M. (2018). La nascita del modernismo italiano: Filosofie della crisi, storia e letteratura. 1903-1922. Quodlibet Studio.

Castellana, R. (2010). Realismo modernista. Un'idea del romanzo italiano (1915-1925). Italianistica, 39(1), 23-45.

Ciotti, F. (2022a). Una nuova svolta negli studi letterari: La convergenza tra computazione, cognizione ed evoluzione. In La narrazione come incontro (pp. 19-36). Firenze University Press. https://doi.org/10.36253/979-12-215-0045-5.04

Ciotti, F. (2022b). Computational approaches to literary periodization: An experiment in Italian narrative of 19th and 20th century. In Digital Humanities 2022. Conference Abstracts (pp. 181-183). DH2022 Local Organizing Committee.

Ciotti, F. (2023). Minerva e il pappagallo. IA generativa e modelli linguistici nel laboratorio dell'umanista digitale. *TESTO & SENSO*, *26*, 289–315. https://doi.org/10.58015/2036-2293/671

Dang, H., Mecke, L., Lehmann, F., Goller, S., & Buschek, D. (2022). How to prompt? Opportunities and challenges of zero- and few-shot learning for human-AI interaction in creative applications of generative models. http://arxiv.org/abs/2209.01390

Donnarumma, R. (2012). Tracciato del modernismo italiano. In R. Luperini & M. Tortora (Eds.), Sul modernismo italiano (pp. 13-38). Liguori Editore.

Jannidis, F., & Lauer, G. (2014). Burrows's Delta and Its Use in German Literary History. In M. Erlin & L. Tatlock (Eds.), Distant readings. Topologies of German culture in the long nineteenth century (pp. 29-54). Camden House.

Jockers, M. L. (2013). Macroanalysis: Digital methods and literary history. University of Illinois Press.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. http://arxiv.org/abs/2205.11916

Kundisch, D., Muntermann, J., Oberländer, A. M., Rau, D., Röglinger, M., Schoormann, T., & Szopinski, D. (2022). An update for taxonomy designers. *Business & Information Systems Engineering, 64*(4), 421-439.

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. http://arxiv.org/abs/2107.13586

Luperini, R. (2018). Il modernismo italiano esiste. In M. Tortora (Ed.), Modernismo e modernità. Liguori Editore.

Mollick, E. (2023, settembre 16). Working with AI: Two paths to prompting. *One Useful Thing*. https://www.oneusefulthing.org/p/working-with-ai-two-paths-to-prompting

Moretti, F. (2005). Graphs, maps, trees: Abstract models for a literary history. Verso.

Moretti, F. (2013). Distant reading. Verso.

Moretti, F. (A c. Di). (2017). *Canon/Archive: Studies in Quantitative Formalism*. N+1.

Mosca, F. (2024). Wittgensteinian Network and Wittgensteinian Oracle. Two humanistic-digital tools in and beyond lexicological enquiry. *Umanistica Digitale*, *8*(18), 145–173. https://doi.org/10.6092/issn.2532-8816/20549

Nickerson, R. C., Varshney, U., & Muntermann, J. (2013). A method for taxonomy development and its application in information systems. *European Journal of Information Systems, 22*(3), 336-359.

Piper, A. (2018). Enumerations: Data and literary study. The University of Chicago Press.

Piper, A., & Bagga, S. (2024). Using Large Language Models for Understanding Narrative Discourse. In Y. K. Lal, E. Clark, M. Iyyer, S. Chaturvedi, A. Brei, F. Brahman, & K. R. Chandu (A c. Di), *Proceedings of the The 6th Workshop on Narrative Understanding* (pp. 37–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.wnu-1.4

Sullet-Nylander, F., Roitman, M., Muñoz, J.-M. L., Marnette, S., & Rosier, L. (Eds.). (2014). Discours rapporté, genre(s) et médias. Stockholm University Press.

Taivalkoski-Shilov, K. (2019). Free Indirect Discourse: An Insurmountable Challenge for Literary MT Systems? Translation Spaces, 8(1), 9-29. https://doi.org/10.1075/ts.00002.tai

Tortora, M. (2018). Il modernismo italiano. Carocci.

Tu, N. D. T., Krug, M., & Brunner, A. (2019). Automatic Recognition of Direct Speech Without Quotation Marks: A Rule-Based Approach. In Proceedings of the Digital Humanities Conference (DHD). https://www.bibsonomy.org/bibtex/13c802571e843affc5bba0f30a334938e/krug

Underwood, T. (2019). Distant horizons: Digital evidence and literary change. The University of Chicago Press.

Underwood, T. (2023, marzo 19). Using GPT-4 to measure the passage of time in fiction. *The Stone and the Shell*. https://tedunderwood.com/2023/03/19/using-GPT-4-to-measure-the-passage-of-time-in-fiction/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., & Zhou, D. (2023). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models*. https://doi.org/10.48550/ARXIV.2201.11903

Wellek, R. (1956). Periods and movements in literary history. In R. Wellek & A. Warren, Theory of literature (pp. 263-283). Harcourt, Brace & World.

Wu, T., Terry, M., & Cai, C. J. (2022). AI chains: Transparent and controllable human-AI interaction by chaining large language model prompts. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1-22.

Zamfirescu-Pereira, J. D., Wong, R. Y., Hartmann, B., & Yang, Q. (2023). Why Johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1-21.