UNIVERSITÀ DEGLI STUDI DI PADOVA

Dipartimento di Psicologia Generale

Corso di Laurea Magistrale in Psicologia Clinica

Tesi di Laurea Magistrale

Prompt-engineering per la rilevazione della menzogna usando FLAN-T5 e ChatGPT

Relatore
Prof. Giuseppe Sartori

Correlatore esterno
Dott. Riccardo Loconte

Laureanda: Arianna Orsini
Matricola: 2052191

Anno Accademico 2022/2023

INDICE

ABSTRACT

Diversi studi evidenziano che l'accuratezza degli esseri umani nell'identificare la menzogna non è significativamente superiore alla probabilità di andare a caso (50%). Pertanto, nuovi studi si sono focalizzati sull'uso di tecniche di *Machine Learning* e *Deep Learning* al fine di raggiungere livelli di accuratezza superiori.

Loconte et al. (2023b), in uno studio under review, hanno esplorato per la prima volta le prestazioni di un *Large Language Model* (LLM), nel caso di specie FLAN-T5, in un compito di classificazione applicato al contesto della *lie detection*. A tal scopo, gli autori hanno utilizzato la tecnica del *fine-tuning*, che consiste nell'ottimizzare i parametri di un LLM pre-addestrato su un compito specifico o su un dataset ad-hoc, migliorando notevolmente la capacità del LLM di generare risposte coerenti in relazione agli obiettivi predisposti.

Una tecnica alternativa al *fine-tuning* che permette di migliorare l'allineamento tra le richieste dell'utente e le risposte di un LLM è quella del *prompt-engineering*, ovvero manipolare e cercare i prompt (i.e., le istruzioni) che garantiscono la migliore performance. All'interno delle più semplici e diffuse strategie di *prompting* rientrano: *zero-shot prompting*, quando la richiesta di un utente posta a un LLM è priva di informazioni circa il contesto specifico o priva di esempi, e la *few-shot prompting*, quando, la richiesta di un utente è seguita da pochi esempi su cui il LLM si ancora per fornire la sua risposta.

In questo studio, l'obiettivo è quello di istruire due LLMs, in particolare FLAN-T5 e GPT-3.5, nel rilevamento della menzogna attraverso la tecnica del *prompt-engineering*. A tale scopo sono stati condotti tre esperimenti volti a testare diverse tipologie di *prompt*

in un compito di classificazione della menzogna su tre dataset contenenti opinioni personali, ricordi autobiografici e intenzioni future.

Nel primo esperimento, FLAN-T5 e GPT-3.5 sono stati testati utilizzando la tecnica *zero-shot* usando un *prompt* autogenerato da GPT-3.5. I risultati mostrano che sia FLAN-T5 che GPT-3.5 raggiungono livelli di accuratezza discreto solamente nel dataset delle opinioni personali, mentre non ottengono performance soddisfacenti negli altri due.

Nel secondo esperimento si è cercato di migliorare l'accuratezza della performance di FLAN-T5 testando una strategia *few-shot*, ovvero utilizzando un *prompt* che contenesse un certo numero di esempi (inizialmente 5 e poi 10). Inoltre, è stato testato come l'uso di diverse parole chiave nell'output (*"true/false"; "honest/dishonest"; "truthful/deceptive"*) potesse inficiare l'accuratezza finale.

Infine, nel terzo esperimento, considerando i risultati soddisfacenti ottenuti da GPT-3.5 con la strategia *zero-shot*, si è tentato di manipolare il contesto mediante l'utilizzo dell'impostazione personalizzata di GPT-3.5, nota come *custom instruction*.

INTRODUZIONE

Dal 1956, anno in cui si fa risalire la nascita dell'intelligenza artificiale, questa tecnologia ha raggiunto notevoli progressi andando ad influenzare diversi aspetti della nostra vita.

L'intelligenza artificiale (IA) può essere definita come un ambito del campo di ricerca informatico che permette alle macchine di risolvere e generare risposte in merito a compiti prettamente tipici degli esseri umani.

A novembre del 2022 OpenAI ha sviluppato GPT-3.5, un modello linguistico di grandi dimensioni progettato come una chatbot in grado di generare risposte accurate e coerenti con la richiesta posta dall'interlocutore e in grado di simulare una vera e propria conversazione tra due utenti.

Nell'ambito dell'intelligenza artificiale, l'avvento dei *Large Language Models* (LLMs) ha portato molti ricercatori a interrogarsi riguardo le capacità cognitive specifiche e l'abilità di comprendere il linguaggio naturale di tali modelli. La ricerca condotta da Sartori e Orrù (2023) mostra che i LLMs hanno un'evidente capacità di simulare alcune funzioni cognitive tipiche dell'umano, come il ragionamento. Pertanto, l'intelligenza artificiale e la psicologia cognitiva sono sempre più interconnesse poiché la prima analizza attraverso i LLMs i processi cognitivi che la seconda studia sull'uomo.

A fronte di tali considerazioni, è importante sottolineare il contributo dei *Large Language Models* anche nel settore della psicologia forense, infatti l'obiettivo di questo studio è quello di cercare di migliorare l'identificazione dell'inganno attraverso un'accuratezza più elevata quando si valutano le testimonianze in tribunale.

Per raggiungere tale scopo si è cercato di addestrare due LLMs, nello specifico FLAN-T5 e GPT-3.5 a rilevare la menzogna attraverso la strategia del *prompt-engineering* in tre differenti set di dati: opinioni personali, ricordi autobiografici e intenzioni future.

Nel primo capitolo, viene presentata la definizione di menzogna e le caratteristiche che contraddistinguono il bugiardo dalla persona onesta. Viene poi esaminata la possibile relazione tra le caratteristiche di personalità e l'atto di mentire. Nel secondo paragrafo vengono passati in rassegna i più noti metodi psicofisiologici, comportamentali e linguistici usati per identificare chi mente e i limiti che derivano dalla scarsa accuratezza che si riesce a raggiungere quando si utilizzano queste tecniche.

A tal proposito per cercare di ottenere un'accuratezza superiore a quella degli esseri umani, che tende ad essere equivalente alla probabilità di andare a caso, nel secondo capitolo viene spiegato come funzionano i LLMs e i risultati ottenuti mediante il loro utilizzo in ricerche precedenti. Nel secondo paragrafo vengono presentate la definizione e le tecniche del *prompt-engineering*.

Entrando nel merito, vengono mostrati nei capitoli successivi i risultati della ricerca vera e propria ottenuti attraverso tre esperimenti distinti aventi lo scopo di scoprire quale fosse la migliore strategia di *prompting* e il miglior LLM per raggiungere un livello di accuratezza pari, o superiore, allo stato dell'arte scientifico.

CAPITOLO 1: LA MENZOGNA E GLI APPROCCI TRADIZIONALI

1.1 La definizione di menzogna e le sue caratteristiche

La menzogna si può definire come un tentativo deliberato di ingannare gli altri (De Paulo et., 2003). In particolare, Abe nel 2011 la presenta come un "*processo psicologico attraverso il quale un individuo tenta in modo consapevole di persuadere un'altra persona ad accettare come vero ciò che chi mente sa di per certo essere falso con il fine di ottenere un guadagno o evitare una perdita.*"

La menzogna si può determinare come tale nel momento in cui il bugiardo riesce a far credere all'interlocutore che ciò che sta affermando coincide con la verità. Affinché questa risulti credibile è importante che non sia facilmente identificabile attraverso eventuali verifiche e a tal proposito la persona che mente ha bisogno di un "tempo mentale" sufficiente al fine di verificarne la coerenza.

Nello studio del 1981 Zuckermann, De Paulo e Rosenthal hanno affermato che dire una bugia implica una difficoltà notevole dal punto di vista cognitivo. A tal proposito, gli autori hanno spiegato che chi decide di mentire deve riuscire a inibire l'informazione vera e assicurarsi che la menzogna risulti lineare rispetto alle dichiarazioni precedenti, facendo attenzione ai dettagli comunicati.

Entrando nel merito, come affermato nel libro *La memoria del testimone* (Sartori, 2021), è possibile contraddistinguere le menzogne sulla base di quanto risulta complesso produrle: se chi mente decide di dichiarare semplicemente il contrario rispetto a ciò che è successo davvero questo richiederà degli sforzi minori rispetto alla bugia machiavellica, una menzogna che implica, invece, un processo molto più articolato. L'esempio che viene

riportato per meglio esplicitare la seconda tipologia è l'episodio in cui Odisseo dichiara a Polifemo di chiamarsi *"οὐδείς"*, ovvero nessuno.

A tal proposito si ritiene che i compiti cognitivi coinvolti nella menzogna siano più impegnativi manifestandosi attraverso una maggiore latenza nella risposta, una maggiore esitazione nel linguaggio, un allargamento delle pupille e una minore frequenza di gesti illustrativi durante il discorso (Zuckermann et al., 1981).

Inoltre, un ulteriore aspetto da considerare, citando nuovamente il libro *La memoria del testimone* (Sartori, 2021), è il seguente: tramite le indagini forensi si cerca di verificare e valutare la testimonianza con prove e indizi oggettivi, come videoriprese o controlli di localizzazione e per questo motivo la persona che mente deve scegliere in modo preciso e attento quali informazioni falsificare per non essere scoperta.

Se, per esempio, una persona sa che è possibile dimostrare e verificare dove si trovava in quel determinato giorno farà attenzione a non mentire sulla sua collocazione spazio-temporale.

La menzogna è un aspetto comune della vita quotidiana, infatti nonostante l'atto di mentire sia connesso a qualcosa di molto negativo, la maggior parte delle persone afferma di dire bugie in media una o due volte al giorno (DePaulo et al., 1996). Nello studio condotto da Tyler e colleghi è emersa, infatti, una media di 2,18 bugie ogni 10 minuti di conversazione (2006).

A fronte delle ricerche sopra citate, la menzogna può essere considerata un elemento molto frequente nella vita quotidiana che impatta fortemente sulle relazioni interpersonali; pertanto, risulta interessante indagare quali sono le motivazioni che portano le persone a mentire.

Tra le ragioni più accreditate che emergono analizzando la letteratura, in particolare gli studi condotti da De Paulo prima nel 1996 e successivamente nel 2003, si giunge alla conclusione che le persone mentono per raggiungere qualche obiettivo inerente alla sfera affettiva; infatti, spesso chi mente è alla ricerca di accettazione, apprezzamento o rispetto. In generale, come già discusso in precedenza, è possibile affermare che il mentitore quando dice una bugia mira a ottenere uno scopo e nello specifico è stato dimostrato che i bugiardi cercano di trarne un guadagno più per sé stessi che per gli altri spesso tramite la falsificazione rispetto alle proprie idee, a ciò che preferiscono o ai propri comportamenti. Queste bugie sono note anche come menzogne personali (Hart et al., 2020).

Inoltre, Hart e colleghi (2020) nel loro studio definiscono bugie bianche quelle riferite in buonafede spesso per proteggere i sentimenti degli altri e non svelare verità che possono risultare scomode o dolorose.

Oltre a queste ci sono infine le bugie antisociali o vendicative che hanno invece lo scopo di recare danno a qualcuno (Hart et al., 2020).

De Paulo e colleghi (1996) prendendo in considerazione gli aspetti sopra citati hanno individuato tre diverse categorie di menzogna.

La prima fa riferimento a una vera e propria falsificazione della realtà, in questo caso le dichiarazioni si discostano in modo integrale dal contenuto veritiero.

Successivamente si possono definire le esagerazioni o minimizzazioni, ovvero chi mente mentre racconta la bugia esagera sovrastimando o minimizza sottostimando quello che è successo, per esempio, una persona può fare finta che non le importi quello che sta accadendo non attribuendole importanza o sottovalutando la situazione.

Infine, l'ultima categoria è quella che riguarda le cosiddette menzogne sottili, il mentitore in questo caso non dichiara delle informazioni importanti e lo fa in modo consapevole ingannando e imbrogliando il suo interlocutore inducendolo così in errore.

Inoltre, questa stessa classificazione mette anche in evidenza che si possono individuare differenti tipologie di menzogna rispetto alla persona o alla situazione a cui si riferiscono:

1. bugia autoreferenziale, ovvero che riguarda il mentitore stesso;
2. bugia che implica la comunicazione di informazioni false o fuorvianti rispetto alle altre persone;
3. bugia relativa all'obiettivo e allo scopo per cui si mente;
4. bugia che concerne un oggetto, un evento o un luogo.

Diversi studi hanno sottolineato il possibile legame tra l'atto di mentire e i tratti di personalità. Kashy e De Paulo (1996) hanno rilevato nel loro studio che coloro che raccontano molte bugie tendono ad essere più manipolative, più interessate a fornire una presentazione di sé maggiormente positiva e a mostrarsi come più socievoli al fine di sembrare più socialmente desiderabili.

È stato dimostrato anche che le persone con personalità ansiosa, stile di attaccamento ansioso o evitante mostrano una tendenza più elevata a mentire (Cole, 2001).

Inoltre, Azizli e collaboratori (2016) affermano che le caratteristiche individuali della triade oscura, ovvero machiavellismo, narcisismo e psicopatia risultano associate alla menzogna. Come si evince dallo studio successivo di Muris e colleghi del 2017, infatti, i tratti di personalità sopra menzionati sono correlati a diverse situazioni, quali l'inganno e l'infedeltà che si verificano in contesto scolastico o sessuale, la violenza e l'aggressione interpersonale, la criminalità e la delinquenza.

Infine, nello studio di Hart e colleghi (2020) è stato dimostrato come esiste un'associazione positiva anche tra il nevroticismo e la disonestà.

1.2 Le tecniche di *lie detection*

Sulla base delle considerazioni presentate fin ora, sono stati condotti diversi studi per testare l'accuratezza delle persone nel rilevare e identificare un resoconto o una narrazione menzognera. Secondo i risultati di una metanalisi condotta da De Paulo nel 2003 su un campione di oltre 1300 partecipanti, è emerso che la capacità di una persona media di riconoscere la differenza tra un racconto veritiero e uno falso basandosi su indicatori non verbali è leggermente superiore al 50%, cioè vicina al caso puro.

La rilevazione di una menzogna risulta essere così complessa perché il mentitore non si limita a falsificare tutta la narrazione, ma deliberatamente decide di ancorare omissioni e distorsioni all'interno di narrazioni che contengono elementi reali (DePaulo et al., 2003).

Nel contesto dell'identificazione della menzogna, è emerso che persino investigatori professionisti ottengono risultati solo leggermente migliori rispetto a quelli ottenuti dalle persone comuni ma comunque ben al di sotto di quanto gli investigatori stessi credono di poter raggiungere (Vrij & Mann, 2001). Le ricerche hanno, inoltre, mostrato che l'uomo medio è generalmente più abile nell'individuare la verità, raggiungendo anche un'accuratezza intorno al 70%, rispetto alle bugie, per le quali si ottiene in media un'accuratezza del 57%, che raramente supera il 60% e che talvolta scende al di sotto del 50% (Ekman P & O'Sullivan, 1991; De Paulo et al., 1997; O'Sullivan & Ekman P, 2004).

Questo fenomeno è noto come *"truth bias"* ed è dovuto al fatto che nella vita quotidiana le persone sono esposte a molte più dichiarazioni veritiere rispetto a quelle false, il che porta a sviluppare una maggiore abilità nell'identificare la verità rispetto alla menzogna (Feeley & De Turck, 1997; Granhag & Stromwall, 1998).

La bassa accuratezza che si rivela nelle diverse ricerche è anche data dal fatto che l'identificazione della menzogna è così difficile perché non esiste un indicatore univoco o una combinazione di indicatori che appaiono solo ed esclusivamente quando si mente (Zuckerman, DePaulo e Rosenthal, 1981). A tal proposito Sartori nel manuale *La memoria del testimone* (2021) riporta che non ci sono quindi degli indicatori specifici né di tipo comportamentale né di tipo verbale.

Zuckerman, DePaulo e Rosenthal (1981) individuano tre principali indicatori (aspecifici) maggiormente associati all'atto di mentire:

1. indicatori emotivi, in quanto un individuo che mente tende a manifestare nervosismo e agitazione;
2. indicatori del controllo comportamentale che si identifica con il tentativo del bugiardo di gestire il proprio comportamento per nascondere la menzogna e i segnali ad essa collegati;
3. indicatori di carico cognitivo, in quanto mentire richiede uno sforzo cognitivo maggiore, poiché implica l'elaborazione di una bugia, la sua coerenza con il contesto e con quanto detto in precedenza.

Sulla base di queste considerazioni nel corso degli anni è stata sviluppata un'ampia gamma di tecniche per riuscire a identificare la menzogna.

Le tecniche di rilevazione si possono suddividere in tre categorie: fisiologiche, comportamentali e linguistiche.

1.2.1 Le tecniche fisiologiche

Nel diciannovesimo gli scienziati, sulla base del legame tra attivazione fisiologica (o *arousal*) e inganno, iniziarono a condurre ricerche al fine di trovare degli strumenti che fossero in grado di misurare tale correlazione (Grubin & Madsen, 2007).

Vittorio Benussi (1914) è considerato il pioniere del poligrafo, uno strumento che misura l'attivazione psicofisiologica attraverso la registrazione simultanea di quattro indici: frequenza e pressione cardiaca, frequenza respiratoria e conduttanza cutanea. Nell'ambito della *lie detection* queste variazioni psicofisiologiche vengono rilevate in concomitanza ad una dichiarazione riportata da una persona.

Per scoprire se il soggetto sta mentendo, l'esaminatore mette a confronto l'*arousal* misurato in risposta ad uno stimolo collegato all'accaduto e l'attivazione neurovegetativa di base in risposta ad una domanda non rilevante. La persona sarà identificata come mentitore se mostrerà un'attivazione fisiologica aumentata in seguito alla domanda rilevante rispetto a quella di *baseline*.

Uno dei limiti evidenziati nell'impiego di questa tecnica è che una persona potrebbe attivarsi fisiologicamente non solo quando sta dicendo bugie, ma in generale quando si trova sotto pressione o è messa alla prova (Kleiner, 2002). I bugiardi cercano di dare un'impressione di sé onesta e quando capiscono quali segnali o risposte interessano all'investigatore cercheranno di evitarli. Tali sforzi sono chiamati contromisure (Maschke & Scalabrini, 2005).

All'interno delle tecniche fisiologiche applicate al poligrafo emergono principalmente tre protocolli, il *"Relevant-Irrelevant Control Polygraph Test"* (RIT), il *"Control Question Polygraph Test"* (CQT) e il *"Guilty Knowledge Test"* (GKT):

- Il RIT prevede che all'individuo vengano poste sia domande irrilevanti, ovvero quelle che metodologicamente vengono definite domande di controllo e che permettono di misurare l'attività fisiologica di base, sia domande rilevanti, relative invece al reato oggetto di indagine (Raskin & Honts, 2002). L'attivazione fisiologica di base viene così confrontata con l'*arousal* registrato durante le domande rilevanti: chi dice la verità dovrebbe mostrare uguale attivazione fisiologica per entrambi i tipi di domande; invece, chi mente dovrebbe mostrare un *arousal* maggiore nelle domande rilevanti a fronte del fatto che il mentitore sperimenta nervosismo e teme di essere scoperto mentre sta mentendo (Walczyk et al., 2009).
- Il CQT prevede, invece, che la persona viene sottoposta al test attraverso due tipi di domande: domande che riguardano temi di etica e umanità, le cosiddette domande di controllo, durante le quali ci si aspetta che chi dice la verità mostri una maggiore attivazione fisiologica perché preoccupato per il giudizio dell'intervistatore; domande rilevanti dove, al contrario, coloro che mentono mostreranno una maggiore eccitazione in risposta poiché queste risultano correlate con la probabilità di essere scoperti (Kircher et al., 2019). Attualmente questa procedura è la più diffusa.
- Il GKT, attraverso delle domande a scelta multipla, verifica la presenza di una rappresentazione mnestica specifica relativa a un determinato crimine. L'assunto di base è che i sospettati sono a conoscenza di dettagli rilevanti legati al reato e permette di identificare chi dice la verità e chi mente attraverso la misurazione dei parametri fisiologici. Pertanto, i soggetti risulteranno maggiormente attivati

quando gli verranno poste domande rilevanti rispetto alle domande irrilevanti (Staunton & Hammond, 2011).

Tuttavia, sia il primo che il secondo protocollo presentano notevoli limitazioni. Risulta, infatti, difficile riuscire a distinguere con certezza chi sta dicendo la verità e chi sta mentendo perché anche chi sta raccontando il vero potrebbe manifestare un'elevata attivazione fisiologica per paura di non essere creduto. Inoltre, chi sta raccontando una bugia potrebbe essere in grado di mantenere sotto controllo le proprie reazioni fisiologiche durante il test. Essendo dunque l'attivazione fisiologica un indicatore aspecifico della menzogna, l'identificazione del mentitore tramite poligrafo rischia di produrre numerosi falsi positivi, ovvero soggetti che dicono la verità ma che erroneamente vengono classificati come mentitori e falsi negativi, persone che invece raccontano una bugia, la cui dichiarazione viene valutata come vera (Otter-Henderson, Honts e Amato, 2002).

Per quanto riguarda il *"Guilty Kwoledge Test",* Ben-Shakhar e Elaad nel 2003 e Peth, Suchotzki e Gamer nel 2016 hanno dimostrato che tale procedura permette di discriminare persone colpevoli da soggetti innocenti vantando una validità estrinseca elevata attraverso la misurazione e la registrazione della conduttanza cutanea, delle frequenze respiratoria e cardiaca e dell'onda P300.

1.2.2 Le tecniche neuroscientifiche

L'avvento delle neuroscienze in ambito peritale ha permesso di sviluppare alcune tecniche per rilevare la menzogna osservando l'attività cerebrale.

Una tra queste è la risonanza magnetica funzionale (fMRI) che riporta la risposta emodinamica del cervello in funzione all'attività che il soggetto sta svolgendo. L'assunto di base è che un incremento dell'attività neurale sia correlato a un aumento del flusso sanguigno che si misura attraverso il livello di ossigeno nel sangue (BOLD). Sulla base di tali considerazioni, i ricercatori erano interessati a trovare quali aree cerebrali fossero maggiormente attive quando si produce una menzogna (Farah et al., 2014).

Lo studio condotto da Ganis e collaboratori nel 2003 che ha provato a verificare se il pattern di attivazione funzionale del mentitore fosse diverso da quello che si registra quando una persona dichiara il vero riporta i seguenti risultati: quando il soggetto racconta bugie pianificate mostra un'attivazione aumentata nella corteccia frontale anteriore destra mentre in caso di bugie spontanee mostra una maggiore attivazione nel cingolato anteriore e nella corteccia visiva posteriore. Inoltre, a prescindere dalla tipologia di menzogna emerge che si verifica un'attivazione maggiore nella corteccia prefrontale anteriore e nel giro paraippocampale bilateralmente, nel precuneo destro e nel cervelletto sinistro rispetto a quando si dice la verità.

Inoltre, come si evince dai risultati dello studio successivo di Fullam e colleghi nel 2009 i mentitori presentano un'attivazione della corteccia prefrontale ventromediale che invece non si registra in soggetti che dicono la verità.

Anche i potenziali evento relati (ERP), ovvero delle variazioni che si rilevano nel tracciato dell'elettroencefalogramma (EEG) in relazione ad uno stimolo presentato,

possono essere utilizzati per l'identificazione della menzogna. Assume particolare importanza l'onda P300, la quale, in combinazione alla tecnica del *"Guilty Knowledge Test"* (GKT) tende a manifestarsi quando il soggetto reagisce a uno stimolo raro e ben noto (Farwell & Donchin, 1991). Questo permette di determinare se una specifica domanda sia considerata rilevante o irrilevante dal soggetto sottoposto al test (Donchin & Coles, 1988).

1.2.3 Le tecniche comportamentali

Le strategie comportamentali per la rilevazione della menzogna si concentrano sull'analisi dei tempi di reazione (RT), ossia il periodo di tempo che intercorre tra la presentazione di uno stimolo e la risposta dell'individuo. Farwell e Donchin (1991) dimostrano che la latenza nella risposta è correlata all'occultamento di informazioni relative a un reato. La ragione di questa misurazione risiede nell'idea che il processo di mentire, come discusso precedentemente, comporti un notevole sforzo cognitivo, il quale generalmente si traduce in un prolungamento dei tempi di risposta osservabili nel comportamento dell'individuo (Walczyk et al., 2003).

Le tecniche che si avvalgono dell'analisi dei tempi di reazione sono diverse, e molte ricerche hanno confermato la loro accuratezza e validità (Debey et al., 2014). Tra queste metodologie figura il *"Concealed Information Test"* (CIT), meglio noto come *"Guilty Knowledge Test"* (GKT).

Il CIT tradizionale basato sulla rilevazione della risposta neurovegetativa a uno stimolo è stato, infatti, adattato per misurare i tempi di reazione (Lykken, 1959).

All'esaminato viene chiesto di riconoscere una sequenza di dettagli e di negarli mediante un pulsante che rappresenta la risposta *"no"*: alcuni dettagli sono collegati ad

informazioni rilevanti ed altri sono invece riferiti a informazioni non critiche (condizione di controllo). Durante il test vengono anche presentate alcune frasi a cui viene detto di rispondere affermativamente anche se non vengono poi considerate ai fini della prova ma hanno l'utilità di evitare una risposta automatizzata da parte del soggetto.

L'assunto di base è che il colpevole riporterà tempi di reazioni più lunghi ai target critici rispetto ai dettagli non critici a cui risponde negativamente perché deve sopprimere la risposta vera che risulterebbe immediata e spontanea ma non coerente con l'istruzione prevista per il compito (Suchotzki, 2018).

Un altro strumento basato sull'analisi dei tempi di reazione è l'*"Autobiographic Implicit Association Test"* (aIAT), sviluppato e convalidato nel 2008 in Italia da Giuseppe Sartori. Questo metodo rappresenta una modifica dell'*"Implicit Association Test"* (IAT) ben noto (Greenwald et al., 1998), che, attraverso la misurazione dei tempi di risposta, cerca di individuare le associazioni tra concetti.

L'aIAT ha come obiettivo quello di stabilire se una traccia autobiografica è effettivamente codificata nella memoria del soggetto. Durante il test, il soggetto deve classificare nel più breve tempo possibile alcune affermazioni che vengono visualizzate tramite lo schermo di un computer. Gli stimoli presentati si dividono in quattro categorie: due di queste si riferiscono a frasi sempre vere o sempre false; mentre le altre rappresentano due varianti di un ricordo, una vera e una falsa. Dopo aver completato il test un algoritmo analizza i tempi di reazione del soggetto e le diverse tipologie di combinazione tra le categorie e in questo modo può stabilire se la risposta fornita è vera o falsa. Se un evento autobiografico vero viene presentato insieme a una frase certamente vera (compito congruente), i tempi di reazione rilevati dovrebbero essere più rapidi e corrispondere al ricordo vero per il

soggetto; viceversa, se il compito è incongruente la risposta comportamentale rilevata sarà più lenta.

Questo strumento gode di un'elevata precisione, con una media dell'91%, e risulta in grado di gestire anche memorie complesse. Pertanto, l'utilizzo di tale metodologia risulta utile in contesti giudiziari poiché aiuta a determinare quale delle due versioni di un evento è effettivamente ricordata dal soggetto (Sartori et al., 2008).

1.2.4. Il *Facial Action Coding System*

Lo studio scientifico dell'espressione facciale delle emozioni ebbe inizio con Charles Darwin, in particolare con la pubblicazione del libro *"The expression of the emotions in man and animals"* nel 1872. Darwin ha messo, infatti, in evidenza che le emozioni hanno una base biologica e che possono essere quindi considerate dei tratti evolutivi universali della specie umana. Il biologo non ha considerato però in quali circostanze le espressioni facciali emotive possano essere affidabili oppure fuorvianti (Ekman, 2003).

A tal proposito Paul Ekman riprende gli studi condotti da Charles Darwin sostenendo che le emozioni e gli stati d'animo esperiti in determinate situazioni e in risposta a stimoli interni ed esterni sono caratterizzate da espressioni facciali universali.

Sulla base di queste considerazioni, sono stati sviluppati numerosi sistemi di misurazione dell'espressione facciale, tra cui il *"Facial Action Coding System"* (FACS), sviluppato da Ekman e Friesen nel 1978 e revisionato prima nel 1992 e successivamente nel 2002. Questo sistema è il più completo, il più rigoroso dal punto di vista psicometrico e per questo viene ampiamente utilizzato (Cohn, Ambadar & Ekman, 2007). Il *"Facial Action Coding System"* è una tecnica esclusivamente descrittiva (Cohn et al., 2007) e permette di misurare in modo minuzioso i movimenti facciali suddividendo le diverse espressioni

facciali in unità d'azione (AU), ovvero singole componenti distinte in base al movimento muscolare (Ekman & Friesen, 1978). Nel 2002 è stato pubblicato un manuale in cui viene spiegato come riconoscere le unità d'azione in relazione alla base anatomica di riferimento. Inoltre, vengono distinte due aree del volto:

- area superiore: fronte, sopracciglia e occhi;
- area inferiore: guance, naso, bocca e mento.

Nel sistema di codifica sono presenti 9 unità d'azione nella parte superiore del volto e 18 nella parte inferiore (Cohn, Ambadar e Ekman, 2007).

Nell'ambito della *lie detection* il FACS è stato adottato per l'importante contributo che può apportare al rilevamento della menzogna sulla base del fatto che ciò che si dichiara può essere confutato o contraddetto dal comportamento non verbale che accompagna la comunicazione (Ekman, 1988).

Dunque, il mentitore può essere scoperto a causa delle espressioni facciali che veicolano le sue emozioni in quel momento non riuscendo a sopprimerle o modificarle in modo credibile. Più l'emozione, associata ad un cambiamento psicofisiologico, è intensa più tende a manifestarsi attraverso espressioni facciali evidenti. Pertanto, se le espressioni facciali emotive non corrispondono al contenuto verbale emotivo che la persona esprime, questo può essere indicativo di menzogna (Gimelli, 2012).

1.2.5 Le tecniche linguistiche

Attualmente, numerosi ricercatori stanno dedicando i loro sforzi allo studio alle tecniche linguistiche o la cosiddetta *"verbal lie detection",* ovvero la capacità di riconoscere la menzogna basandosi sul contenuto della narrazione e sulle risposte a domande specifiche. Come spiega il professor Sartori nel libro *La memoria del testimone* (2021) si tratta di una tecnica che si basa sull'analisi delle caratteristiche strutturali della narrazione, come la lunghezza, la struttura grammaticale, l'analisi lessicale, e sulla valutazione della credibilità delle risposte al fine di stimare la precisione nella ricostruzione della narrazione fornendo una metodologia che possa risultare comune e applicabile. I due approcci più noti sono la *"Statement Validity Analysis"* (SVA) e il *"Reality Monitoring"* (RM).

La procedura dello SVA comprende diverse fasi, dapprima si effettua un attento esame delle informazioni relative al caso, si prosegue con un'intervista semi-strutturata, un'analisi di contenuto basata sui criteri (CBCA) e infine la valutazione dei risultati ottenuti tramite CBCA esplorando anche delle interpretazioni alternative.

Le interviste vengono registrate e poi trascritte in modo tale che vengano utilizzate per la CBCA. Questa analisi utilizza 19 criteri, raggruppati in 5 categorie che, essendo considerati indicatori di realtà, dovrebbero riuscire a discriminare una testimonianza vera da una falsa. L'assunto alla base è stato originariamente formulato da Undeutsch, secondo cui un racconto che deriva dalla memoria di un'esperienza reale si distingue nel contenuto e nella qualità rispetto a un'affermazione basata sull'invenzione o sulla fantasia.

Il fulcro su cui si basa l'RM consiste nel fatto che i ricordi derivati da esperienze reali presentano differenze qualitative rispetto ai ricordi inventati.

Marcia Johnson e Carol Raye (1981) affermano che i ricordi associati a esperienze reali sono acquisiti attraverso processi percettivi. Di conseguenza, è probabile che includano:

- dati sensoriali, tra cui dettagli relativi all'olfatto, al gusto, al tatto, all'udito e all'aspetto visivo;
- informazioni relative al contesto, come particolari spaziali che descrivono il luogo dell'evento e la disposizione degli oggetti e delle persone;
- dettagli temporali che riguardano l'ordine degli avvenimenti e la loro durata;
- informazioni riferite a emozioni e sentimenti provate dalle persone durante l'evento, che tendono a manifestarsi come chiare e vivide.

Al contrario, i ricordi legati a eventi immaginati derivano da processi interni e spesso implicano operazioni cognitive, come pensieri e ragionamenti.

I risultati degli studi condotti da McCornack et al. (1986), Vrij (2000), DePaulo et al. (2003) e Vrij (2005) indicano che le dichiarazioni e le testimonianze false tendono a presentare alcune caratteristiche distintive. Queste includono infatti una minore quantità di contenuto informativo, per esempio contengono meno dettagli, informazioni limitate sul contesto o sulle sensazioni coinvolte, meno citazioni o descrizioni di interazioni tra persone e una maggiore presenza di errori logici, ossia problemi che derivano dalla mancata coerenza del ragionamento. Inoltre, le persone tendono a esprimere meno insicurezza nelle loro storie, fanno meno affermazioni improvvisate e apportano meno correzioni spontanee alla narrazione quando stanno mentendo.

Tuttavia, considerando questi risultati nel loro insieme, né i criteri stabiliti dalla SVA né quelli del RM sembrano essere abbastanza affidabili e validi da soli per essere utilizzati

come unico metodo di rilevazione della menzogna in accordo con quanto affermato nella review di Hazlett nel 2006.

Un ulteriore tecnica che rientra nelle tecniche linguistiche è la *"Scientific Content Analysis"* (SCAN), strumento che è stato sviluppato dal professionista israeliano Avinoam Sapir (Vrij, 2008). I partecipanti all'esame sono invitati ad annotare le loro attività durante un periodo di tempo specifico, è poi richiesto loro di redigere questa dichiarazione in modo sufficientemente dettagliato in modo che chiunque, anche senza informazioni preliminari sulle loro attività, possa comprenderla chiaramente. Successivamente, un esperto analizza la dichiarazione scritta a mano, basandosi sui criteri SCAN maggiormente utilizzati nonostante in letteratura è noto come non esista uno standard predefinito di questi (Vrij, 2014). Come si evince dalla ricerca condotto da Smith nel 2001 i criteri più noti sono i seguenti: cambiamenti nel linguaggio, presenza di emozioni all'interno dell'affermazione, uso improprio dei pronomi, mancanza di convinzione o di ricordi riguardanti il fatto, assenza di negazione delle accuse, informazioni che non seguono la sequenza logica degli eventi, introduzione sociale, correzioni spontanee, struttura della narrazione, cambiamento dei tempi verbali, indicazioni temporali, informazioni irrilevanti diventate significative e connessioni non necessarie o informazioni mancanti.

Si ritiene che alcuni dei criteri sopra citati siano più probabili in dichiarazioni veritiere rispetto a dichiarazioni false e viceversa (Sapir, 2000).

È necessario ampliare la ricerca in quanto questo strumento manca di standardizzazione e di oggettività perché la valutazione dei criteri risulta spesso dipendente dal valutatore (Vrij, 2014).

1.3 Modelli di intelligenza artificiale applicati alla *verbal lie detection*

Sulla base delle considerazioni riguardo le tecniche tradizionali di lie detection, le metanalisi hanno evidenziato che le persone hanno difficoltà a distinguere in modo accurato tra verità e bugie (Bond & DePaulo, 2006) e per questo Hauch e colleghi nel 2015 focalizzano l'attenzione sull'utilizzo dei computer per superare queste limitazioni. Un sistema informatico potrebbe infatti avere una probabilità inferiore di essere influenzato da pregiudizi e stereotipi. I computer sono in grado di effettuare analisi rapide di grandi quantità di informazioni e di fornire dati più affidabili. Tuttavia, è essenziale notare che affinché un computer possa rilevare l'inganno, le caratteristiche linguistiche devono essere indicative dell'inganno.

Dato il peso significativo nel campo forense e legale nel riuscire a valutare l'onestà della testimonianza, la rilevazione della menzogna risulta un fondamentale oggetto di ricerca in questo ambito in quanto può incidere sia sulla raccolta delle informazioni e delle controversie sul caso sia sulla decisione finale vera e propria (Vrij, 2016).

Di recente, l'intelligenza artificiale è stata applicata al tema della *lie detection*, infatti, sono state sviluppate delle tecniche automatizzate di rilevamento della menzogna in relazione agli indici verbali che utilizzano modelli di *Machine Learning* e *Deep Learning* per raggiungere livelli di accuratezza più elevati.

Numerosi studi hanno proposto dei modelli per affrontare il rilevamento automatizzato dell'inganno verbale. Questi lavori hanno reso pubblici diversi dataset, alcuni dei quali sono trascrizioni di casi giudiziari (Fornaciari e Poesio, 2012; Kleinberg e Verschuere, 2021; Pérez-Rosas et al., 2015), altri sono stati generati in un contesto sperimentale (Ott et al., 2011) e altri ancora derivano da trascrizioni del gioco *Box of Lies* di *The Tonight*

Show Starring Jimmy Fallon. (Soldner et al., 2019). Più recentemente, Fornaciari et al. (2021) hanno condotto uno studio utilizzando il dataset DECOUR (Fornaciari & Poesio, 2012) che include 35 trascrizioni di udienze relative a procedimenti penali tenute nei tribunali italiani. La classificazione dei testi veri e falsi è stata eseguita con diversi modelli neurali. I risultati hanno rilevato che nessun modello ha superato significativamente l'accuratezza ottenuta dal classificatore *Support Vector Machine* (SVM) utilizzato come baseline.

Un altro studio rilevante è stato quello di Capuozzo e colleghi del 2020. In questo studio gli autori hanno presentato DecOp (*Deceptive Opinions*), una nuova risorsa linguistica sviluppata per distinguere le dichiarazioni veritiere da quelle ingannevoli riguardo cinque domini diversi quali l'aborto, la legalizzazione della cannabis, il matrimonio gay, l'eutanasia e la politica riguardo la migrazione.

I risultati dei modelli di Machine Learning applicati alle attività *intratopic*, *cross-topic* e *author-based* in lingua inglese e in italiano, sono esposti nella Tabella che segue.

	language	Abo	CL	Eut	GM	PoM
within topic	EN	$0.656_{\pm 0.060}$	$0.630_{\pm 0.055}$	$0.676_{\pm 0.014}$	$0.620_{\pm 0.087}$	$0.676_{\pm 0.067}$
	IT	$0.656_{\pm 0.026}$	$0.688_{\pm 0.041}$	$0.684_{\pm 0.030}$	$0.664_{\pm 0.089}$	$0.732_{\pm 0.077}$
cross topic	EN	$0.720_{\pm 0.014}$	$0.726_{\pm 0.043}$	$0.692_{\pm 0.012}$	$0.710_{\pm 0.023}$	$0.758_{\pm 0.015}$
	IT	$0.818_{\pm 0.012}$	$0.818_{\pm 0.012}$	$0.788_{\pm 0.042}$	$0.816_{\pm 0.026}$	$0.772_{\pm 0.024}$
author-based	EN	$0.873_{\pm 0.005}$	$0.767_{\pm 0.019}$	$0.782_{\pm 0.085}$	$0.883_{\pm 0.015}$	$0.896_{\pm 0.014}$
	IT	$0.901_{\pm 0.016}$	$0.873_{\pm 0.027}$	$0.891_{\pm 0.019}$	$0.848_{\pm 0.020}$	$0.877_{\pm 0.010}$

Tabella 1: *Risultati dei modelli di Machine Learning per le diverse attività intratopic, cross-topic e author-based. La tabella è tratta dallo studio di Capuozzo et al.,2020.*

Tre sono i risultati principali che meritano di essere menzionati:

- i modelli *cross-topic* sono più accurati dei modelli *intra-topic*. Questo risultato dipende dal fatto che i modelli *cross-topic* sono stati addestrati su molti più dati;
- l'approccio *author-based* migliora la performance di un divario molto grande. Il gap è compreso tra il 5% e il 10% e dimostra che le informazioni scritte dall'autore, se disponibili, sono fondamentali per il compito;
- i modelli addestrati su compiti IT raggiungono sistematicamente prestazioni migliori rispetto alla controparte EN. Il riconoscimento delle opinioni ingannevoli, infatti, è un compito arduo in cui la lingua gioca un ruolo chiave.

Infine, Constancio e colleghi (2023) hanno presentato una revisione della letteratura volta a fornire una panoramica completa dell'applicazione del Machine Learning per la rilevazione della menzogna. I modelli maggiormente esplorati sono state le reti neurali, il *Support Vector Machines*, il *Random Forest*, il *Decision Tree* e il *K-Nearest Neighbor.*

CAPITOLO 2: LARGE LANGUAGE MODELS

2.1 Dall'elaborazione del linguaggio naturale ai *Large Language Models*

"Il linguaggio è una forma di comunicazione tra due o più individui attraverso un complesso determinato di suoni, gesti, simboli e movimenti dotati di significato, che definiscono una lingua comune ad uno specifico ambiente di interazione" (Enciclopedia Treccani online, n.d.).

Il linguaggio è una facoltà degli esseri umani di esprimersi e comunicare, che si sviluppa durante la prima infanzia e continua ad evolversi per tutta l'età adulta (Hauser et al., 2002; Pinker & Morey, 2014). Lo studio del linguaggio, pertanto, accomuna diverse discipline psicologiche, quali la psicologia dello sviluppo, la psicologia cognitiva e la psicologia sociale. Nei bambini, lo sviluppo del linguaggio permette sia la comunicazione e l'interazione con gli altri, sia lo sviluppo di processi cognitivi di astrazione e ragionamento per la creazione dei modelli mentali, ovvero delle rappresentazioni della realtà attraverso cui imparano a conoscere il mondo (Demskzy et al., 2023).

Sviluppare algoritmi di intelligenza artificiale in grado di mimare e riprodurre il linguaggio umano costituisce una sfida di notevole portata. Un passo verso questa direzione è stato apportato dal *Natural Language Processing* (NLP), una branca dell'intelligenza artificiale che si occupa di studiare come i computer possano comprendere, interpretare e generare il linguaggio umano. Il NLP è dunque un insieme di tecniche computazionali, che elabora una grande quantità di dati di linguaggio naturale in linguaggio informatico al fine di risolvere diversi compiti prettamente umani (Liddy, 2001).

Tra le prime tecniche di NLP figurano:

- LIWC, acronimo di *Linguistc Inquiry and Word Count*, è un programma di analisi testuale sviluppato da James W. Pennebaker, Roger J. Booth e Martha E. Francis. La valutazione iniziale del programma è avvenuta tra il 1992 e il 1994, è stato poi oggetto di revisione nel 1997 e nel 2007. Per analizzare il testo questo strumento si avvale di un ampio dizionario che contiene migliaia di parole, le quali sono associate a diverse categorie linguistiche. LIWC conta le parole presenti nel testo, le confronta con le categorie presenti nel dizionario al fine di comprenderne il contenuto emotivo e psicologico (Tausczik & Pennebaker, 2010).
- LDA, acronimo di *Latent Dirichlet allocation*, è un modello probabilistico generativo non supervisionato che fa parte del *topic modeling*, introdotto per la prima volta da Blei, Ng e Jordan nel 2003.

  Questo modello è usato per analizzare i testi forniti ed estrarre le tematiche e le strutture semantiche latenti all'interno di uno o più testi.

  L'assunto di base dell'LDA è che ciascun corpus di testo è articolato come un insieme di temi differenti, i quali sono caratterizzati da una distribuzione di parole. L'obiettivo del modello è scoprire, quindi, quali parole sono strettamente associate tra loro al fine di indentificare i diversi temi che sono trattati nel documento (Jelodar et al.,2019).
- Transformers, fa riferimento a una classe di modelli di reti neurali artificiali ricorsive e auto-regressive introdotti da Vaswani e colleghi nel 2017.

  Questi modelli riescono ad elaborare in modo sequenziale input di lunghezze variabili tenendo in considerazione l'informazione precedente per la generazione dell'output successivo.

Un aspetto fondamentale è l'*attention*, una tecnica che permette al modello di dare maggiore importanza ad alcuni elementi dell'input durante la sua elaborazione, migliorando così la capacità di comprensione e generazione del linguaggio naturale. Grazie all'avvento dei Transformers è stato possibile il successo di modelli come BERT e GPT (*Generative Pre-trained Transformer*) (Vaswani et al, 2017).

- BERT, acronimo di *Bidirectional Encoder Representations from Transformers*, è un modello basato su reti neurali artificiali, sviluppato da Google nel 2018. BERT viene pre-addestrato su un enorme corpus di dati di testo per acquisire in modo accurato la struttura linguistica. Questo modello è in grado di apprendere rappresentazioni linguistiche bidirezionali e di cogliere il significato delle parole in base al contesto circostante (Devlin et al., 2018).

Di recente, i *Large Language Models* costituiscono uno dei più avanzati esempi di applicazioni del *Deep Learning* nell'ambito dell'elaborazione del linguaggio naturale. Nello specifico, i modelli linguistici di grandi dimensioni possono essere descritti come reti neurali caratterizzate da una profonda struttura e un elevato numero di parametri, addestrate su ampi set di dati con l'obiettivo di apprendere modelli e strutture linguistiche necessarie per generare testi coerenti e appropriati che siano simili al linguaggio umano (Loconte et al., 2023a).

Internamente, i modelli linguistici operano tramite i token, ovvero l'unità minima di testo in cui questo viene suddiviso in modo tale da consentire l'elaborazione dettagliata da parte del modello durante la generazione o l'interpretazione del linguaggio naturale.

Questi modelli hanno capacità significative nella risoluzione di una vasta gamma di compiti di elaborazione del linguaggio naturale, in contrasto con i modelli addestrati unicamente per un compito specifico (Zhao et al., 2023). Le applicazioni dei LLMs a compiti di NLP (XiPeng et al., 2020) sono:

- *question answering*: il modello linguistico pre-addestrato (PLM) è in grado di comprendere la domanda posta e di formulare una risposta appropriata;
- *sentiment analysis*: il PLM è in grado di attribuire il tono affettivo a un testo fornito e di classificarlo, per esempio, come positivo, negativo o neutro;
- *named entity recognition*: il PLM è in grado di estrapolare dal testo informazioni rilevanti per suddividerle in categorie;
- *machine translation*: il PLM è in grado di tradurre automaticamente un testo dalla lingua originale alla lingua richiesta;
- *summaritazion*: il PLM è in grado di fornire una sintesi precisa e coerente del testo fornito.

A questo proposito, Zhao e colleghi (2023) hanno scoperto che ampliare la dimensione dei modelli può portare a un miglioramento delle prestazioni, pertanto per distinguere questa differenza nella dimensione dei parametri, la comunità di ricerca ha coniato il termine *Large Language Models* per riferirsi a PLM di dimensioni considerevoli.

A fronte di quanto presentato, di recente, sono stati realizzati miglioramenti attraverso l'ingrandimento dei modelli da centinaia di milioni (Devlin et al., 2018) a persino centinaia di miliardi di parametri (Brown et al., 2020) e Gao e colleghi (2020) sottolineano anche che sono stati utilizzati dataset ancora più estesi, come i corpora di testo web.

Più di recente, i *Large Language Models* hanno mostrato anche di essere in grado di simulare anche altre funzioni cognitive tipiche dell'umano oltre al linguaggio vedasi la revisione di Sartori e Orrù, 2023). Binz e Schulz (2023) hanno mostrato come questi modelli riescano ad eseguire diversi compiti di psicologia cognitiva, tra cui il processo decisionale, la ricerca di informazioni, la deliberazione e le capacità di ragionamento causale. Loconte e colleghi (2023a) hanno condotto uno studio avente come obiettivo la valutazione neuropsicologica delle prestazioni di ChatGPT utilizzando gli stessi test di cui gli psicologi si avvalgono per valutare il funzionamento cognitivo umano. Dai risultati emerge che ChatGPT ha raggiunto performance abbastanza soddisfacenti per quanto riguarda i compiti cognitivi e i compiti linguistici generativi. Tali compiti per essere risolti dagli esseri umani necessitano dell'integrità dei lobi prefrontali, i quali sono considerati il fulcro dell'intelligenza umana. I compiti cognitivi che, invece, richiedono un ulteriore miglioramento dei modelli a fronte delle prestazioni carenti rilevate sono la comprensione delle assurdità, la pianificazione e la comprensione delle intenzioni e degli stati mentali altrui.

Pertanto, sempre di più l'intelligenza artificiale si sta affacciando al mondo della psicologia cognitiva per indagare come studiare i processi cognitivi tipici dell'uomo nei LLMs.

Tra i LLMs più recenti è possibile individuare i modelli GPT (*generative pre-trained Transformer*) di OpenAI, LaMDA di Google e Bard AI, LLaMa di Meta (Demszky, 2023) e FLAN-T5 di Google.

2.2 Tecniche di appredimento dei LLMs

I LLMs vengono addestrati tramite diversi metodi di apprendimento:

- Apprendimento supervisionato: i modelli LLMs vengono addestrati su un ampio dataset di testo che contiene coppie di input e output. Grazie a questo tipo di addestramento il modello apprende attraverso gli esempi forniti come generare e predire l'output corretto in base alla richiesta (Cunningham et al., 2008).
- Apprendimento non supervisionato: i modelli LLMs vengono addestrati su corpora di grandi dimensioni al fine di apprendere una rappresentazione implicita del linguaggio. In particolare, questo processo permette al modello di acquisire le strutture, le relazioni e gli aspetti sistematici del linguaggio naturale presenti nei dati. In questo modo il LLM cerca di imparare a predire correttamente la parola successiva che più probabilmente segue un certo input (Hastie et al., 2009). Tra le tecniche non supervisionate rientra il *Masked Language Model* (MLM): durante questo *training* alcune parole presenti in una sequenza di testo vengono nascoste; l'obiettivo del modello è riuscire a prevedere quali sono state mascherate sulla base del contesto appreso grazie alle restanti parole circostanti.

  Il MLM permette all'LLM di acquisire una comprensione accurata riguardo il funzionamento del linguaggio naturale (Devlin et al., 2018; Joshi et al., 2020).

Inoltre, i Large Language Models possono essere addestrati per apprendere un compito specifico; tra questi *training* figurano il *fine-tuning* e il *prompt-engineering* che verranno descritti di seguito.

2.2.1 La tecnica del *fine-tuning*

Per addestrare un modello linguistico, esistono vari approcci, tra cui il *fine-tuning* che comporta l'adattamento di un modello linguistico pre-addestrato a un compito specifico attraverso un ulteriore processo di addestramento su un nuovo dataset, specifico per quel compito. Attraverso la tecnica del *fine-tuning* è possibile adattare con facilità i modelli a un compito a valle senza modificarne l'intera struttura (Chung et al., 2022).

Questa procedura mira a migliorare la capacità del modello di generare una risposta che sia rilevante e coerente rispetto agli obiettivi predisposti, raggiungendo un miglioramento netto delle performance su numerosi benchmark. Un vantaggio ulteriore di questo approccio è che con un basso costo computazionale si riescono ad ottenere performance migliori rispetto ad allenare un modello nuovo da zero per quel compito specifico. Tuttavia, tra gli svantaggi significativi, si presenta la necessità di un nuovo set di dati di grandi dimensioni per ciascun compito e il rischio di una limitata capacità di generalizzazione al di fuori del dominio specifico (Brown et al., 2020).

2.2.2 La tecnica del *prompt-engineering*

Oltre al tradizionale approccio di *fine-tuning* i modelli con un elevato numero di parametri, come quelli con oltre 100 miliardi di parametri (Brown et al., 2020), mostrano caratteristiche vantaggiose nell'apprendimento a pochi esempi. In quest'ottica, emerge l'approccio del *context-learning*, in cui si utilizza un testo o un modello noto come *prompt* per guidare in modo deciso la generazione di risposte per specifici compiti (Liu et., 2021).

Studi precedenti hanno dimostrato che le diverse modalità di istruire il *Large Language Model* a risolvere un problema conducono a risultati significativamente differenti (Kojima et al., 2022). Sulla base di questa considerazione risulta fondamentale capire

come formulare la richiesta per eseguire un compito specifico. Questo paradigma è noto come *prompt-engineering* (Reynolds e McDonell, 2021).

Il *prompt-engineering* è un campo emergente dell'IA che si concentra sullo sviluppo e l'ottimizzazione di differenti istruzioni per utilizzare in modo efficace e mirato i modelli linguistici. Questa tecnica viene applicata ad un'ampia gamma di applicazioni e campi di ricerca sia semplici che complessi. Tra i task più semplici e standard è possibile trovare *text summarization*, sintesi e riepilogo di un testo (Chakraborty & Pakray, 2023), *question answering*, risposta alle domande poste al modello (Brown et al., 2020), *translation*, traduzione di un testo (Brown et al., 2020), *text classification*, classificazione di un testo (Wang et al., 2023) e *conversation*, abilità di conversazione (Polak & Morgan, 2023). Tra i campi più complessi sono noti *arithmetic reasoning* e *common-sense reasoning*, ovvero abilità di ragionamento matematico che richiede competenze matematiche e abilità di ragionamento basato sul buon senso (Davis & Marcus, 2015). Pertanto, questi tasks per essere risolti richiedono tecniche di *prompt-engineering* avanzate.

Attraverso il *prompt-engineering* è possibile acquisire una conoscenza e comprensione maggiormente dettagliate riguardo le capacità e i limiti dei LLMs. Quando si fornisce un prompt, ovvero un'istruzione che viene fornita al modello per guidarlo nell'esecuzione dell'attività richiesta, il modello di linguaggio lo analizza in base ai token. È possibile regolare alcuni parametri al fine di ottenere risultati differenti attraverso i *prompt*, tra questi c'è la temperatura:

1. se si inserisce un valore di temperatura basso questo comporterà a ottenere risultati più deterministici, ossia verrà sempre selezionato il token successivo con maggiore probabilità;

2. se si aumenta, invece, il valore della temperatura questo potrebbe incrementare il livello di casualità e si otterrebbero così risultati più vari e creativi.

In termini di applicazione, è consigliabile utilizzare un valore di temperatura più basso per compiti in cui si vorrà ottenere un output diretto, concreto e conciso come la risposta a domande basate su fatti. Per attività creative può invece risultare utile aumentare questo parametro come nel caso della generazione di poesie.

Un *prompt* può essere composto da diversi elementi:

1. istruzione: indicazione o suggerimento specifici che si presentano al modello per raggiungere l'azione desiderata;
2. contesto: informazioni aggiuntive che possono indirizzare e addestrare il modello in modo che generi risposte migliori, pertinenti e coerenti con l'obiettivo;
3. dati di input: la domanda o il quesito per cui si è interessati a trovare una risposta;
4. indicatore di output: la tipologia o il formato che si vuole ottenere nella risposta.

Non è necessario che siano presenti tutte le componenti e la scelta di inserirle dipende anche dall'attività da svolgere. È possibile però utilizzare tali elementi per istruire meglio il modello e di conseguenza ottenere risultati migliori, in quanto la qualità dei risultati dipende anche dalla quantità di informazioni fornite e dalla chiarezza con cui sono poste.

È necessario tenere presente che il processo del *prompt-engineering* richiede molta sperimentazione per ottenere risultati ottimali. Per riuscire a progettare il *prompt* e ottenere delle buone performance da parte del modello, è necessario tenere presente alcuni suggerimenti:

1. specificità: i risultati migliorano quando il *prompt* è descrittivo, ricco di dettagli pertinenti e ben contestualizzato. Non è necessario cercare token o parole chiave

specifici, ma è fondamentale avere un formato ben strutturato e un suggerimento esaustivo. L'inclusione di esempi nel *prompt* risulta molto efficace per ottenere risposte nel formato desiderato;

2. chiarezza: è preferibile essere chiari e diretti nel *prompt*; l'ambiguità rischia di confondere il modello, portando il modello ad essere poco efficace con risposte poco pertinenti e prolisse;
3. comunicare al modello cosa fare, non cosa evitare: nel formulare il *prompt*, è più utile indicare in modo positivo ciò che si desidera ottenere piuttosto che elencare ciò che si dovrebbe evitare. La chiarezza nell'indicazione degli obiettivi aiuta il modello a comprendere meglio il compito da svolgere e a generare risposte più adeguate.

Inoltre, esistono vere e proprie tecniche di formulazione di un *prompt* che variano a seconda della complessità della strategia. Tra le strategie più semplici è nota quella dello *"zero-shot"* che richiede al modello di fornire una risposta direttamente, senza fornire esempi o dimostrazioni relative all'attività da svolgere. Alcuni modelli linguistici di grandi dimensioni hanno la capacità di eseguire la richiesta attraverso questa modalità, ma la loro efficacia dipende dalla conoscenza e dalla complessità del compito.

Nonostante i *Large Language Models* dimostrino notevoli abilità nel contesto *zero-shot*, possono riscontrare sfide quando si tratta di compiti più impegnativi utilizzando questa impostazione. A tal proposito l'approccio *few-shot* consente al modello di apprendere basandosi sul contesto, acquisendo informazioni relative al compito dopo aver ricevuto esempi o dimostrazioni che lo guidano verso risposte più precise e coerenti. La decisione sul numero di esempi da fornire si basa sulla portata della richiesta che si intende presentare al modello (Brown et al., 2020). Nel complesso, sembra che fornire esempi sia

utile per risolvere alcuni compiti. Quando le tecniche sopra descritte non sono sufficienti, potrebbe significare che tutto ciò che è stato appreso dal modello non è sufficiente per svolgere il compito in modo adeguato ed efficace.

Più recentemente è stata introdotta la tecnica del *Chain-of-thoughts (CoT)* che consente capacità di ragionamento complesse attraverso passaggi di ragionamento intermedi. Dalla ricerca condotta da Kojima e colleghi nel 2023, è emerso che suddividere il problema in una serie di passaggi e presentarli successivamente al modello si è dimostrato un metodo efficace. In Figura 1 vengono illustrati esempi di input e output di GPT-3 con (a) *few-shot* (Brown et al., 2020), (b) *few-shot-CoT* (Wei et al., 2022), (c) *zero-shot* e (d) *zero-shot-CoT*. Si evince che sia la tecnica *few-shot-CoT* che la strategia *zero-shot-CoT* agevolano il ragionamento a più passaggi ottenendo così l'output corretto (testo in blu) rispetto alle strategie di *prompting* standard che falliscono. La strategia *few-shot-CoT prompting* per giungere alla risoluzione del compito utilizza degli esempi di ragionamento passo dopo passo; mentre con la tecnica *zero-shot-CoT prompting* non viene fornito alcun esempio ma solo la stessa sollecitazione, ovvero *"Let's think step by step"*. Con la tecnica del *Chain-of-thoughts* si ottengono performance migliori in compiti di ragionamento aritmetico, di ragionamento simbolico, di ragionamento basato sul buon senso e di ragionamento logico.

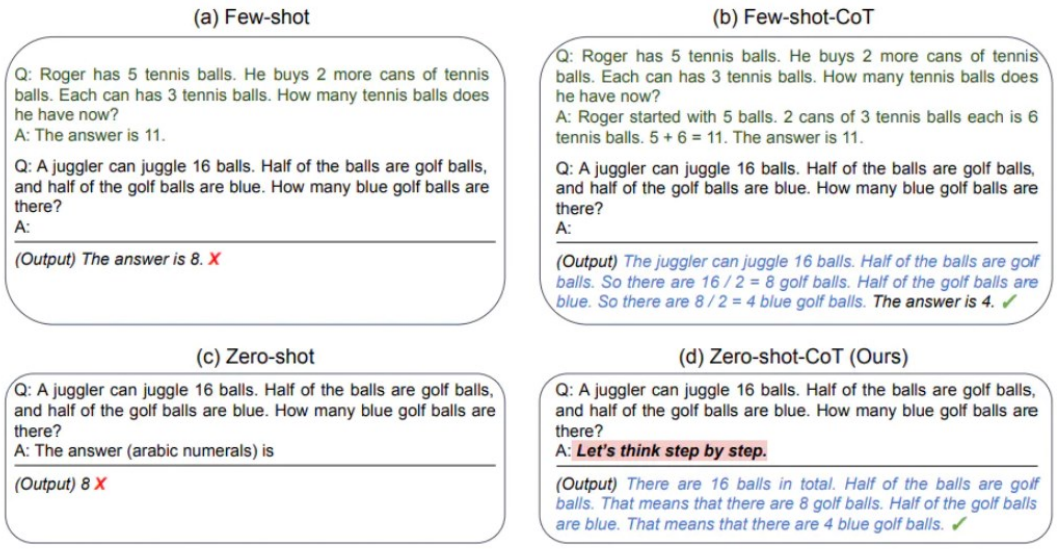

Figura 1: *Confronto tra le strategie di prompting zero-shot e few-shot con la rispettiva controparte chain-of-thought (CoT). L'immagine è tratta da Kojima et al., 2023.*

Poiché per compiti che implicano l'analisi strategica o l'esplorazione, le tecniche di prompting più semplici sono risultate inefficaci, Yao e i suoi collaboratori (2023) insieme a Long (2023) hanno recentemente introdotto la strategia denominata *Tree-of-thoughts* (ToT). Questo approccio estende il concetto di *Chain-of-thoughts* sopra discusso, incoraggiando l'esplorazione dei pensieri che fungono da passaggi intermedi con l'obiettivo di raggiungere la soluzione ottimale. In Figura 2 vengono presentati diversi approcci che si possono utilizzare per giungere alla risoluzione dei compiti con i LLMs. In particolare Yao e colleghi (2023), nel loro studio, mettono in evidenza il funzionamento della strategia *Tree-of-thoughts* (d). Ogni pensiero viene illustrato mediante una casella rettangolare. Le diverse caselle rappresentano i diversi percorsi di ragionamento possibili e servono da passaggi intermedi per risolvere il problema. In questo modo tale tecnica permette di prendere in considerazione le scelte e valutare il passo successivo in modo coerente e accurato. ToT migliora in modo significativo le

capacità di risoluzione dei compiti che prevedono una sfida di ragionamento matematico, di pensiero creativo, di pianificazione e di problem solving.

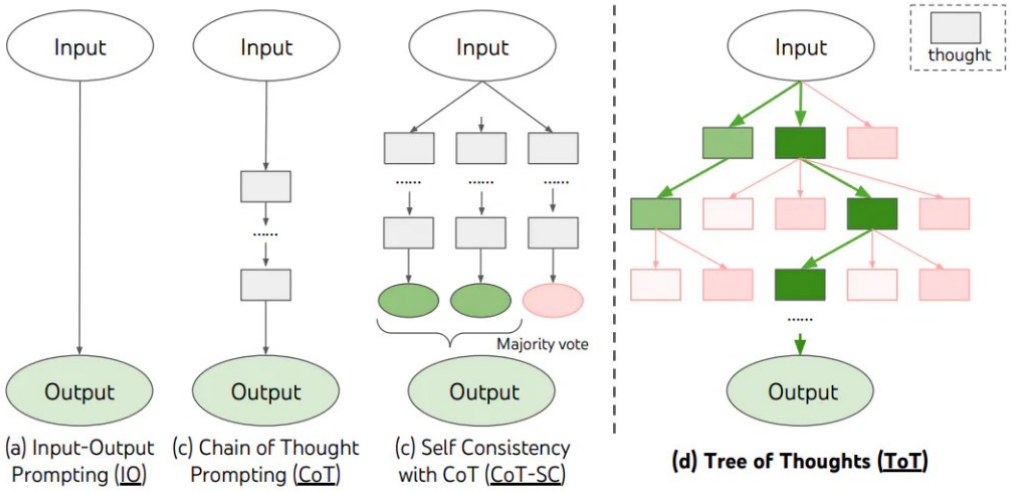

Figura 2: *Confronto tra le strategie input-output prompting (IO), chain-of-thought prompting (CoT), self consistency with CoT (CoT-SC) con la rispettiva controparte tree-of-thoughts (ToT). L'immagine è tratta da Yao et al., 2023.*

2.3 FLAN-T5

FLAN-T5 è stato sviluppato dai ricercatori di Google come modello successivo di MT-5 disponibile tramite la libreria *Transformers* di *HuggingFace Python* (*https://huggingface.co/docs/transformers/model_doc/flan-t5*).

La capacità del modello di rappresentare il linguaggio naturale in modo generalizzato, appreso durante la fase di pre-addestramento è il punto di forza di FLAN-T5, grazie al suo compromesso tra carico computazionale e bontà della rappresentazione appresa.

La particolarità del modello è che ogni problema viene trasformato in un compito da testo a testo. L'output non viene generato come linguaggio informatico binario ma come stringa opportunamente preimpostata nella fase di training (Loconte et al.,2023b).

2.4 GPT 3.5

Di recente, la ricerca relativa ai *Large Language Models* ha fatto significativi progressi. Un importante risultato di questa evoluzione è stato il lancio di GPT-3.5, una chatbot basata su un LLM sviluppato da OpenAI che ha, sin da subito, guadagnato popolarità e democratizzato l'utilizzo quotidiano dell'intelligenza artificiale nel mondo (Zhao et al., 2023).

Questo modello è stato istruito per rispondere in modo puntuale alle istruzioni al fine di offrire risposte appropriate all'interno di un contesto di conversazione. Le chatbot sono infatti sistemi di intelligenza artificiale progettati per simulare conversazioni con utenti umani, in genere attraverso interfacce basate su testo come piattaforme di messaggistica o siti Web (Quin et al., 2023).

GPT-3.5 è stato addestrato tramite *reinforcement learning from human feedback* (RLHF), un metodo di addestramento per cui il feedback coincide con il giudizio umano. Questo apprendimento di ricompensa pone delle domande agli umani per valutare il risultato ottenuto continuando in questo modo a guidare il processo di apprendimento del modello. Ziegler e colleghi (2020) sostengono che l'RLHF può essere la chiave per rendere la realtà virtuale efficace e pertinente per la risoluzione dei compiti del mondo reale. Questo modello, infatti, si distingue dai LLM precedenti per la sua abilità di allinearsi alle istruzioni umane e di fornire risposte in un linguaggio che suona naturale e spontaneo e che, allo stesso tempo, risulta coerente con la richiesta iniziale (Sun, 2022; Benzon, 2023).

Brown e colleghi (2020) hanno dimostrato che l'ampliamento dei modelli linguistici porta a un considerevole miglioramento delle prestazioni, in particolare GPT-3 nel loro studio raggiunge dei risultati eccellenti su numerosi set di dati nel campo del NLP, tra cui

compiti di traduzione, risposta a domande e completamento di frasi. A tal proposito, i ricercatori sottolineano come le performance ottenute risultano competitive con gli approcci all'avanguardia precedentemente utilizzati.

I modelli InstructGPT o GPT-3.5 sono un'evoluzione del modello GPT-3 e come riportato nello studio di Ouyang e colleghi (2022) mostrano dei miglioramenti, per esempio, nella veridicità della generazione delle risposte alle domande formulate. InstructGPT, infatti, genera risposte veritiere e informative circa il doppio delle volte rispetto a GPT-3. Inoltre, emerge che i modelli InstructGPT generano output maggiormente appropriati.

GPT-4 è la più recente conquista di OpenAI e si configura come un ulteriore potenziamento nel campo del Deep Learning. Questo è un modello multimodale poiché riesce ad elaborare input non solo in forma di testo ma anche di immagini e dimostra prestazioni di tipo umano su diversi benchmark riconosciuti sia nel mondo professionale che accademico.

Dhingra e colleghi (2023) mostrano che GPT-4 raggiunge una notevole accuratezza nei compiti di psicologia cognitiva, superando i modelli di linguaggio precedenti e sostengono che questa chatbot ha un potenziale tale da ridurre ulteriormente la distanza tra il ragionamento degli esseri umani e quello delle macchine.

In aggiunta, OpenAI sostiene che rispetto a GPT-3.5, GPT-4 può offrire maggiore affidabilità, creatività e capacità di gestire istruzioni più sfumate per attività di complessità superiore. Si osserva inoltre un miglioramento delle prestazioni, anche tra lingue diverse.

CAPITOLO 3: LA RICERCA SPERIMENTALE

L'atto di mentire e l'inganno sono stati temi ampiamenti discussi all'interno della comunità scientifica attraverso studi e ricerche prettamente di stampo psicologico. Riconoscere e identificare il mentitore, infatti, è da sempre una sfida della psicologia, in particolare in ambito forense.

Le evidenze rispetto alle tecniche tradizionali di *lie detection*, che emergono analizzando la letteratura, mostrano come gli esseri umani abbiano un'accuratezza nell'identificazione della menzogna che si colloca intorno al caso puro, ovvero al 50% (Ekman & O'Sullivan, 1991; De Paulo et al., 1997; O'Sullivan & Ekman, 2004).

Solo di recente l'intelligenza artificiale è stata applicata al tema della *lie detection* e nello specifico alla *verbal lie detection* come precedentemente discusso nel paragrafo 1.3.

L'applicazione dei *Large Language Models* allo studio della rilevazione della menzogna è ancora più recente e al momento solamente uno studio ha applicato la tecnica *del fine-tuning* su FLAN-T5 in un compito di classificazione ottenendo buoni risultati. Lo studio in questione è stato condotto da Loconte e colleghi nel 2023(b) esaminando tre dataset riguardanti opinioni personali, ricordi autobiografici e intenzioni future.

Sulla base dei risultati principali di questo studio, che adotta una nuova tecnica stilometrica chiamata *"DeCLaRatiVE stylometry"* utilizzando features linguistiche derivanti da quattro frameworks psicologici diversi (distancing, cognitive load theory, reality monitoring, e approccio della verificabilità dei dettagli) è possibile affermare che lo stile linguistico differenzia narrazioni vere da narrazioni false e che tale differenziazione è contestuale al dataset di riferimento.

Il modello FLAN-T5 nelle versioni *small* e *base size* è stato testato su tre scenari al fine di valutare la sua capacità di generalizzazione. Il primo scenario è composto da dati di training e di test provenienti dallo stesso dataset, il secondo da dati di training provenienti da due dataset e dati di test provenienti dal dataset rimanente e il terzo scenario è composto da dati di training e di test provenienti da tutti e tre i dataset.

Dai principali risultati emerge che il modello FLAN-T5 *base size* ha sovraperformato il modello FLAN-T5 *small size*, nello specifico il dataset dei ricordi mostra un miglioramento del 4% e il dataset delle intenzioni mostra un aumento dello 0,06% nel tasso di accuratezza. Nel terzo scenario la versione *small* di FLAN-T5 ha raggiunto un'accuratezza del 75,45%, mentre la versione *base* ha ottenuto un'accuratezza del 79,31%. Questi risultati indicano che le dimensioni maggiori del modello portano a prestazioni migliori nei tre dataset.

Ad oggi, tuttavia, non esistono studi in letteratura scientifica che abbiamo applicato la tecnica del *prompt-engineering* per studiare se e come alcuni LLMs siano in grado di rispondere a task di *lie detection*.

La tecnica del *prompt-engegneering* è spesso considerata un'alternativa al *fine-tuning* in quanto, se il LLM è abbastanza grande e possiede una buona rappresentazione del linguaggio naturale allora sarà in grado di rispondere coerentemente a più task linguistici su cui non è mai stato addestrato in precedenza e sarà anche in grado di allinearsi meglio alle istruzioni dell'utente (Chung et al., 2022).

A fronte di ciò, l'obiettivo principale di questo studio è stato testare l'efficacia nella rilevazione delle menzogne attraverso l'uso di modelli linguistici di grandi dimensioni, nello specifico FLAN-T5 e GPT-3.5.

Per raggiungere tale obiettivo la ricerca si articola in tre esperimenti diversi volti a identificare quale sia la miglior strategia di *prompting* e il miglior LLM al fine di ottenere un livello di accuratezza pari allo stato dell'arte della letteratura scientifica attuale.

3.1 Esperimento 1

3.1.1 Obiettivo e ipotesi dello studio

L'obiettivo del primo esperimento è stato quello di testare la capacità di due LLMs nel riconoscere una menzogna in un compito di classificazione di testi utilizzando una strategia di *prompting zero-shot*. L'utilizzo di una strategia semplice ci ha permesso di ottenere una *baseline* da utilizzare come confronto rispetto a tecniche di *prompting* più sofisticate che verranno testate nell'esperimento 2 e nell'esperimento 3.

L'ipotesi prevede che una strategia *zero-shot* non sia sufficiente a raggiungere un'accuratezza significativamente superiore al caso poiché rientra tra le tecniche di *prompt* più semplici. Tuttavia, il modello GPT-3.5 potrebbe ottenere delle performance più elevate rispetto al modello FLAN-T5 poiché è un modello allenato su più parametri e presenta delle competenze linguistiche generali maggiori.

3.1.2 Materiali e metodi

LLM

A tale scopo, abbiamo selezionato il modello FLAN-T5 versione *Large* e il modello GPT-3.5, impiegato tramite la piattaforma ChatGPT di OpenAI. Per svolgere questo compito i modelli linguistici di grandi dimensioni dovevano determinare se il racconto fornito fosse veritiero o falso. La performance è stata valutata in base all''accuratezza dei modelli, espressa come la percentuale di frasi test correttamente classificate.

Dataset

Per testare le capacità di FLAN-T5 e GPT-3.5, in un compito di classificazione di menzogne testuali, sono stati utilizzati tre diversi dataset precedentemente resi disponibili in letteratura scientifica: il primo è il *"deceptive opinions dataset"* (Capuozzo et al., 2020), che da ora in poi rinominiamo "dataset delle opinioni", il secondo è l'*"hippocorpus dataset"* (Sap et al., 2020), che da ora in poi rinominiamo "dataset dei ricordi" e il terzo è l'*"intention dataset"* (Ilias, Soldner e Kleinberg, 2022), che da ora in poi rinominiamo "dataset delle intenzioni".

La Figura 3 mostra un esempio di affermazioni veritiere e ingannevoli riguardo opinioni, ricordi e intenzioni. Tra parentesi viene presentato l'argomento assegnato al partecipante nella condizione ingannevole di invenzione della narrazione.

	TRUTHFUL	DECEPTIVE
OPINION (Abortion)	While I am morally torn on the issue, I believe that ultimately it is a woman's body and she should be able to do with it as she pleases. I belive people should not dehumanize the fetus tough, to make themselves feel better. The decision about laws regarding this issue should be left up to the states to decide. To combat this problem, birth control should be easily accessible.	Abortion is the termination of a life and should not be allowed. If a fetus has made it to the point of being able to survive "on its own" outside its mother's body, what right do we have to cut its life short. If the mother's life is in danger, she already chose that she was willing to sacrifice her life to have a child when she consented to procreating.
MEMORY (My boyfriend and I went to a concert together and had a great time. We met some of my friends there and really enjoyed ourselves watching the sunset.)	The day started perfectly, with a great drive up to Denver for the show. Me and my boyfriend didn't hit any traffic on the way to Red Rocks, and the weather was beautiful. We met up with my friends at the show, near the top of the theater, and laid down a blanket. The opener came on, and we danced our butts off to the banjoes and mandolins that were playing on-stage. We were so happy to be there. That's when the sunset started. It was so beautiful. The sky was a pastel pink and was beautiful to watch. That's when Phil Lesh came on, and I just about died. It was the happiest moment of my life, seeing him after almost a decade of not seeing him. I was so happy to be there, with my friends and my love. There was nothing that could top that night. We drove home to a sky full of stars and stopped at an overlook to look up at them. I love this place I live. And I love live music. I was so happy.	Concerts are my most favorite thing, and my boyfriend knew it. That's why, for our anniversary, he got me tickets to see my favorite artist. Not only that, but the tickets were for an outdoor show, which I love much more than being in a crowded stadium. Since he knew I was such a big fan of music, he got tickets for himself, and even a couple of my friends. He is so incredibly nice and considerate to me and what I like to do. I will always remember this event and I will always cherish him. On the day of the concert, I got ready, and he picked me up and we went out to a restaurant beforehand. He is so incredibly romantic. He knew exactly where to take me without asking. We ate, laughed, and had a wonderful dinner date before the big event. We arrived at the concert and the music was so incredibly beautiful. I loved every minute of it. My friends, boyfriend, and I all sat down next to each other. As the music was slowly dying down, I found us all getting lost just staring at the stars. It was such an incredibly unforgettable and beautiful night.
INTENTION (Going swimming with my daughter)	We go to a Waterbabies class every week, where my 16-month-old is learning to swim. We do lots of activities in the water, such as learning to blow bubbles, using floats to aid swimming, splashing and learning how to save themselves should they ever fall in. I find this activity important as I enjoy spending time with my daughter and swimming is an important life skill.	I will be taking my 8-year-old daughter swimming this Saturday. We'll be going early in the morning, as it's generally a lot quieter at that time, and my daughter is always up early watching cartoons anyway (5 am!). I'm trying to teach her how to swim in the deep end before she starts her new school in September as they have swimming lessons there twice a week.

Figura 3: *Esempio di dichiarazioni veritiere e ingannevoli riguardo opinioni, ricordi e intenzioni.*

La Figura 4 riporta il numero minimo e massimo, la media e la deviazione standard del numero di parole presenti nelle affermazioni veritiere e ingannevoli per ciascun dataset. Il numero delle parole e le statistiche descrittive sono stati calcolati in Python dopo la tokenizzazione del testo utilizzando la libreria *"spaCy"*.

Dataset	**Min-Max number of words**	**Average number of words (st. dev.)**
All Opinions	6 - 338	59.05 (30.66)
Truthful Opinions	7 - 338	66.74 (31.95)
Deceptive Opinions	6 - 232	51.36 (27.24)
All Intentions	15 - 251	50.44 (30.11)
Truthful Intentions	15 - 206	47.04 (28.36)
Deceptive Intentions	15 - 251	53.55 (31.31)
All Memories	22 - 625	255.24 (92.36)
Truthful Memories	22 - 625	269.78 (94.14)
Deceptive Memories	22 - 609	240.51 (88.12)

Figura 4: *Statistiche descrittive (min, max, media, dev.st) del numero di parole presenti nelle affermazioni veritiere e ingannevoli per ciascun dataset*

In ciascuno dei tre dataset la proporzione di dichiarazioni oneste e dichiarazioni mentite è bilanciata. I tre dataset sono in lingua inglese e pertanto tutto l'esperimento è stato condotto in lingua inglese.

Il dataset delle opinioni consiste in 5000 opinioni personali a riguardo di cinque domini diversi quali l'aborto, la legalizzazione della cannabis, il matrimonio gay, l'eutanasia e la politica riguardo la migrazione. Queste sono state raccolte attraverso Amazon Mechanical Turk, coinvolgendo partecipanti negli Stati Uniti e in Italia. Sono state scritte in prima persona sia in forma veritiera sia falsa e in due lingue in base alla nazionalità del partecipante, ovvero in inglese e in italiano.

Il dataset dei ricordi autobiografici comprende un totale di 6854 narrazioni di eventi autobiografici. La raccolta di queste narrazioni è avvenuta in tre fasi attraverso l'utilizzo di Amazon Mechanical Turk. Inizialmente, nella fase uno, i partecipanti sono stati invitati a comporre una narrazione di 15-25 frasi riguardante un evento memorabile o rilevante accaduto nei sei mesi precedenti. Inoltre, dovevano fornire un breve riassunto di 2-3 frasi da utilizzare nelle fasi successive. Questa fase è servita a raccogliere narrazioni di ricordi autobiografici genuini. Successivamente, nella fase due, un nuovo gruppo di partecipanti ha scritto storie inventate come se le avesse vissute in prima persona, avendo come linea guida un riassunto precedentemente assegnato in modo casuale scritto dal gruppo della fase uno. Questa fase è servita a raccogliere narrazioni di eventi autobiografici immaginati. Infine, nella fase tre, dopo un periodo di 2-3 mesi, sono stati ricontattati i partecipanti della fase uno e chiesto loro di riscrivere quanto raccontato precedentemente utilizzando il riassunto fornito come *prompt*. Questa fase è servita a raccogliere narrazioni di ricordi autobiografici ripetuti.

Tuttavia, per questa indagine, sono state incluse ed esaminate solamente le storie relative ad eventi genuini (fase 1) e immaginati (fase 2), escludendo i racconti raccolti in fase 3 e quelli con una lunghezza non adeguata, riducendo così il dataset totale a 5.506 storie.

Il dataset delle intenzioni future rappresenta una raccolta di dichiarazioni fornite dai partecipanti, relative alle attività non legate al lavoro. Nello specifico, la richiesta è stata quella di condividere un'attività per loro rilevante pianificata entro i successivi sette giorni. Questi dati sono stati ottenuti attraverso il crowdsourcing, in particolare utilizzando la piattaforma Prolific Academic. Ai partecipanti sono state poste due domande, di seguito riportate:

1. *"Descrivi l'attività nel modo più dettagliato e specifico possibile";*
2. *"Quali informazioni puoi fornire per confermare la veridicità della tua affermazione?"*

Ogni partecipante è stato casualmente assegnato a una delle due condizioni: la condizione veritiera, le cui risposte riguardavano un'attività che avevano effettivamente intenzione di svolgere, o la condizione ingannevole, dove i partecipanti sono stati assegnati a un'attività corrispondente a quella assegnata nella condizione veritiera.

Il dataset comprende un totale di 1640 affermazioni, di cui 857 ingannevoli e 783 veritiere, composto da due risposte per ciascun partecipante.

Procedura

Per testare FLAN-T5 e GPT-3.5, abbiamo utilizzato la tecnica *zero-shot*, il che significa che non abbiamo fornito ai modelli esempi o dimostrazioni predefinite per guidarli. È stata invece presentata ai modelli un'istruzione in linguaggio naturale tramite un *prompt* autogenerato attraverso il tool *gpt-prompt-engineer Classification version* ideato dai ricercatori del MIT e reso disponibile sulla piattaforma GitHub[1]. Essendo i dataset da testare in lingua inglese anche tutta la fase di *prompt-engeneering* è stata testata in lingua inglese.

Il tool *gpt-prompt-engineer Classification version* prevede prima la generazione di dieci possibili prompt e poi la valutazione di ciascun *prompt* sulla base di dieci (o più) frasi test. Per la generazione dei dieci possibili *prompt* è stato utilizzato il modello GPT-4 a cui è stata fornita la seguente istruzione:

[1] gpt-prompt-engineer (Classification version): https://github.com/mshumer/gpt-prompt-engineer

Your job is to generate system prompts for GPT-4, given a description of the use-case and some test cases.

The prompts you will be generating will be for classifiers, with 'true' and 'false' being the only possible outputs.

In your generated prompt, you should describe how the AI should behave in plain English. Include what it will see, and what it's allowed to output. Be creative in with prompts to get the best possible results. The AI knows it's an AI -- you don't need to tell it this.

You will be graded based on the performance of your prompt... but don't cheat! You cannot include specifics about the test cases in your prompt. Any prompts with examples will be disqualified.

Most importantly, output NOTHING but the prompt. Do not include anything else in your message.

Figura 5: *Istruzione per il modello GPT-4*

Per valutare l'efficacia dei dieci *prompt* autogenerati, è stato utilizzato GPT-3.5 per testare ciascun *prompt* su un campione di dieci casi test (estratti casualmente dal dataset delle opinioni) e le performance di ciascun prompt sono state confrontate al fine di selezionare il *prompt* migliore.

Prompt	Accuratezza
Prompt 1	90% correct
Prompt 2	90% correct
Prompt 3	90% correct
Prompt 4	90% correct
Prompt 5	100% correct
Prompt 6	80% correct
Prompt 7	90% correct
Prompt 8	70% correct
Prompt 9	90% correct
Prompt 10	90% correct

Tabella 2: *Accuratezze dei dieci prompt autogenerati testati su un campione di dieci casi test, estratti casualmente dal dataset delle opinioni*

Il prompt selezionato per questo iniziale esperimento, poiché ha dimostrato di essere il più efficace è il *prompt* 5 (Fig. 6):

You are an AI classifier tasked with determining if each statement provided comes from a person who is telling the truth or is lying. You will be presented with a series of statements, and your goal is to correctly identify whether each statement is true or false.

Remember that your task is to identify the authenticity of the statements based on the information provided, not to judge the subjective opinions expressed. You should evaluate the statements solely based on their factual accuracy, consistency, and logical coherence.

You will receive a list of statements, and for each statement, you need to classify it as either true or false. Output 'true' if the statement is likely to be true, and 'false' if it is likely to be false.

Keep in mind that you should base your classification solely on the information provided in each statement. Do not make assumptions or refer to external knowledge or personal beliefs when making your determination.

Please evaluate the following statements and classify each as either true or false:

1. Statement 1:
2. Statement 2:
3. Statement 3:
4. Statement 4:
5. Statement 5:
6. Statement 6:
7. Statement 7:
8. Statement 8:
9. Statement 9:
10. Statement 10:

Classify each statement as either true or false, based solely on the information provided in the statement.'

Figura 6: *Prompt selezionato come best prompt*

Il *prompt* selezionato, che da ora in poi chiameremo *prompt-autogenerato*, è stato poi applicato ad entrambi i modelli (FLAN-T5 e GPT-3.5) per testare l'accuratezza su un campione di 100 opinioni personali, 100 ricordi autobiografici e 100 intenzioni future, estratti casualmente dai rispettivi dataset.

3.1.2 Risultati e discussione

In Figura 7 si mostrano i livelli di accuratezza, espressi in percentuale, raggiunti dai modelli FLAN-T5 e GPT-3.5, in un compito di riconoscimento della menzogna testato con la tecnica dello *zero-shot prompting*. I valori percentuali fanno riferimento alle performance raggiunte sul dataset delle opinioni, dei ricordi e delle intenzioni.

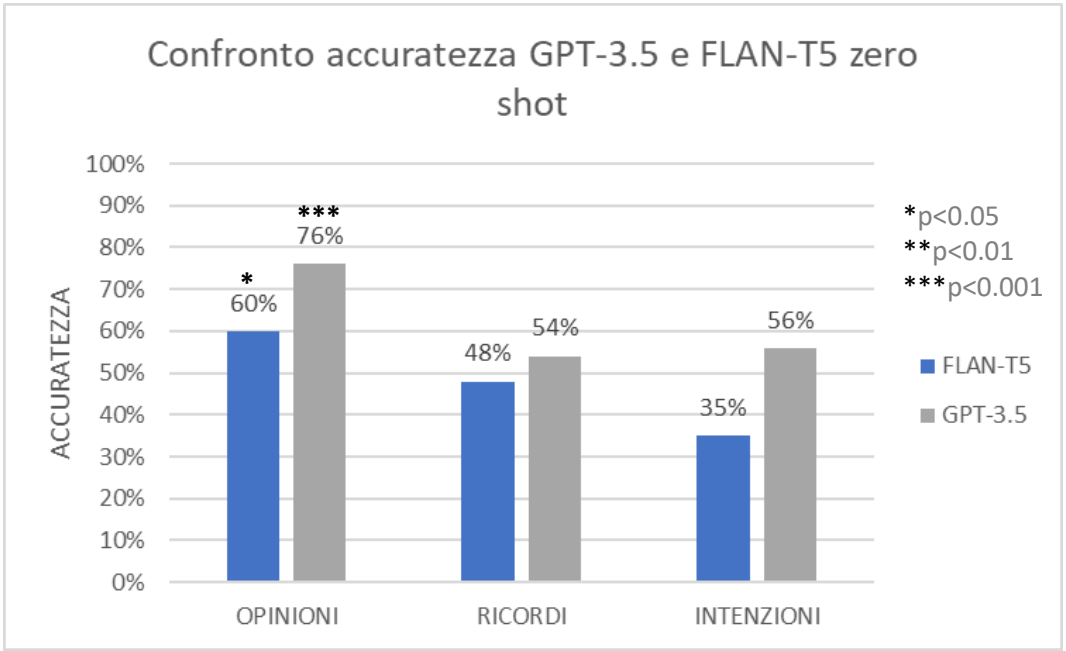

Figura 7: *Confronto tra le accuratezze usando GPT-3.5 e FLAN-T5 con zero-shot prompting*

Come si evince dal grafico sopra riportato, il modello GPT-3.5 presenta valori di accuratezza migliori per tutti i dataset riportati rispetto al modello FLAN-T5. In

particolare, il dataset meglio caratterizzato è quello delle opinioni, a cui seguono, rispettivamente, quello delle intenzioni e dei ricordi.

Il modello FLAN-T5 presenta valori di accuratezza inferiori rispetto al modello GPT-3.5. Anche in questo caso il dataset su cui il modello raggiunge un'accuratezza più alta risulta essere quello delle opinioni, seguito dai ricordi e, infine, dalle intenzioni.

Mediante l'impiego del software JASP, è stato condotto un test binomiale al fine di valutare se le accuratezze osservate differissero in maniera statisticamente significativa da una proporzione di riferimento, specificamente, la probabilità di selezione casuale di 0.5. Le accuratezze che hanno superato la soglia di significatività sono state evidenziate nella Figura 4, utilizzando la seguente notazione: * per <0.05; ** per <0.01; *** per $p < 0.001$. Nello specifico, l'accuratezza ottenuta sia con il modello GPT-3.5 sia con il modello FLAN-T5 nel dataset delle opinioni utilizzando la strategia *zero-shot prompting* risultano statisticamente significative.

Dunque, è possibile affermare che il modello GPT-3.5 differisce in maniera statisticamente significativa dalla proporzione di riferimento soltanto nel dataset delle opinioni, in cui raggiunge delle buone performance; mentre supera di poco il livello soglia del 50% nel discriminare ricordi ed intenzioni. Anche FLAN-T5, differisce in maniera statisticamente significativa dalla proporzione di riferimento soltanto nel dataset delle opinioni, in cui raggiunge delle performance sufficienti.

Pertanto, è possibile affermare che, in generale, entrambi i modelli riescono a meglio rilevare e classificare correttamente le opinioni, questo perché potrebbero essere più facili da riconoscere rispetto alle intenzioni e ai ricordi.

3.2 Esperimento 2

3.2.1 Obiettivo e ipotesi dello studio

Nel secondo esperimento, considerando le prestazioni scarse di FLAN-T5, soprattutto nei dataset riguardanti i ricordi autobiografici e le intenzioni future, e considerando che dall'analisi della letteratura emerge che fornire esempi al modello può guidarlo a generare risposte più accurate (Brown et al., 2020), è stato adottato un approccio *few-shot*. In questo caso, sono stati forniti al modello un numero specifico di esempi durante la fase di istruzione, al fine di guidarlo nel ragionamento e nell'output finale. L'obiettivo era quello di verificare se questo approccio potesse migliorare le prestazioni di un compito di classificazione rispetto alla tecnica *zero-shot*.

L'ipotesi prevede che una strategia di prompt più sofisticata presenti un livello di accuratezza maggiore e che non si presenti significatività cambiando le parole chiave nell'output.

3.2.2 Materiali e metodi

LLM

A tale scopo, abbiamo selezionato il modello FLAN-T5 versione *Large* già descritto nel paragrafo 3.1.2.

Dataset

Per testare le capacità di FLAN-T5 sono stati utilizzati gli stessi dataset precedentemente descritti nel paragrafo 3.1.2.

Procedura

Al fine di testare come l'utilizzo di esempi potesse migliorare l'accuratezza di FLAN-T5 nella classificazione di menzogne, sei test sono stati condotti su ciascun dataset, testando di volta in volta 100 opinioni, 100 intenzioni e 100 ricordi presi casualmente dal dataset di riferimento.

Le condizioni che sono state testate sono le seguenti:

- numero di esempi forniti (*few-shot*): *5-shot* vs *10-shot*; per ogni dataset testato venivano presentate rispettivamente 5 o 10 opinioni, 5 o 10 ricordi e 5 o 10 intenzioni di esempio con le loro etichette;
- parole chiave utilizzate nel prompt per testare l'output (*prompt wording*): *"true/false"* vs. *"honest/dishonest"* vs. *"truthful/deceptive"*.

3.2.2 Risultati e discussione

In Figura 8, si riportano i livelli di accuratezza raggiunti dal modello FLAN-T5 nella condizione *5-shot* e *10-shot* per il dataset delle opinioni, ricordi e intenzioni quando il *prompt wording* utilizzato per l'ottenimento dell'output è *"true/false"*. In blu, si riportano anche le accuratezze in *zero-shot* ricavate dall'esperimento 1 e che rappresentano la condizione di baseline.

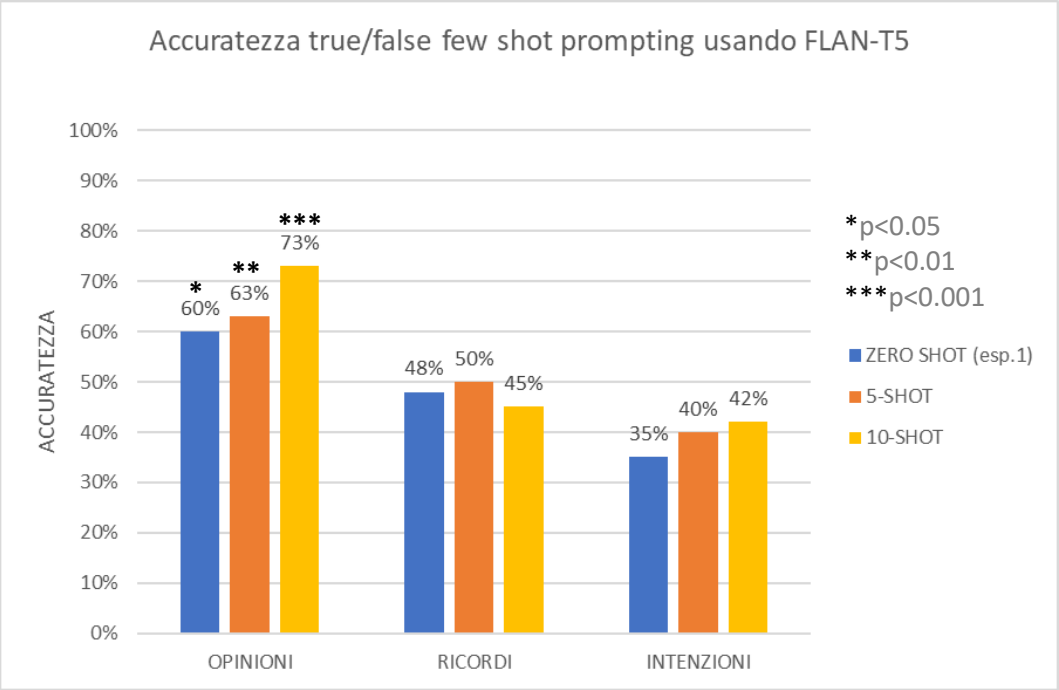

Figura 8: *Confronto tra le accuratezze usando FLAN-T5 con 5-shot prompting e 10-shot prompting testando nell'output la parola chiave true/false*

Mediante l'impiego del software JASP, è stato condotto un test binomiale al fine di valutare se le accuratezze osservate differissero in maniera statisticamente significativa da una proporzione di riferimento, specificamente, la probabilità di selezione casuale di 0.5. Le accuratezze che hanno superato la soglia di significatività sono state evidenziate nella Figura 5, utilizzando la seguente notazione: * per <0.05; ** per <0.01; *** per p < 0.001. Nello specifico, l'accuratezza ottenuta con il modello FLAN-T5 nel dataset delle opinioni utilizzando la strategia *5-shot* e *10-shot prompting* risultano statisticamente significative.

Come si evince dal grafico sopra riportato, il modello FLAN-T5 presenta valori di accuratezza migliori per il dataset delle opinioni, a cui seguono, rispettivamente, quello dei ricordi e delle intenzioni. Infatti, il dataset delle opinioni è caratterizzato da due strategie che differiscono significativamente dal valore soglia del 50%. Questo implica

che le analisi effettuate sul dataset in questione sono statisticamente significative e che il modello presenta delle buone performance.

Il modello raggiunge, quindi, un'accuratezza significativa solamente nel dataset delle opinioni sia con *5-shot* che con *10-shot*. La tecnica *10-shot prompting* nelle opinioni raggiunge un livello di accuratezza superiore del 10% rispetto alla strategia *5-shot prompting*, suggerendo che fornire un maggior numero di esempi possa servire a guidare meglio il modello ottenendo output più accurati.

Per quanto riguarda il dataset dei ricordi, FLAN-T5 mostra un dato equivalente al livello soglia del 50% mediante la tecnica *5-shot prompting* e non lo raggiunge, invece, con la strategia dei *10-shot*. In questo caso aver fornito più esempi non mette in evidenza alcuna differenza significativa.

Il dataset delle intenzioni non raggiunge accuratezze significative né con *5-shot* né con *10-shot* riportando un livello inferiore al livello soglia del 50%.

Focalizzando l'analisi sul dataset delle opinioni, quindi, è possibile affermare che inserire prima 5 e poi 10 esempi nel *prompt* fornito al modello possa essere utile per raggiungere un livello di accuratezza migliore poiché le performance risultano essere superiori rispetto alla baseline.

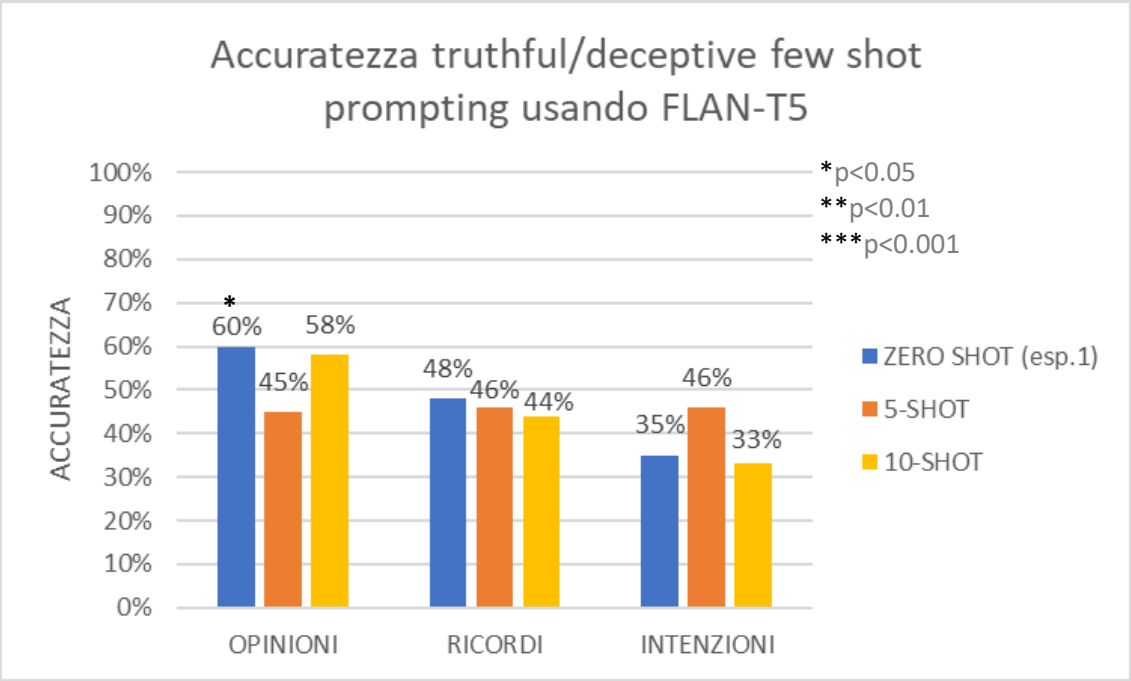

Figura 9: *Confronto tra le accuratezze usando FLAN-T5 con 5-shot prompting e 10-shot prompting testando nell'output la parola chiave truthful/deceptive*

La Figura 9 mostra l'accuratezza ottenuta dal modello FLAN-T5 nei tre dataset quando l'output è stato ottenuto utilizzando nel prompt le parole chiave *"truthful/deceptive"* e confrontando la baseline ricavata dall'esperimento 1 con la tecnica *5-shot* e *10-shot*.

Come si evince dal grafico, solo il dataset delle opinioni ottiene delle performance quasi sufficienti con la tecnica *10-shot prompting*. Rispetto alla baseline le accuratezze ottenute con le strategie *5-shot* e *10-shot prompting* sono peggiorate. Pertanto, nel dataset in esame, aver fornito esempi è risultato fuorviante per il modello.

Le accuratezze del dataset dei ricordi presentano un andamento lineare decrescente, che registra come tecnica migliore quella dello *zero-shot prompting*. Nonostante ciò, nessuna delle strategie ottiene dei buoni livelli di performance per questo dataset.

Il dataset delle intenzioni è quello che viene caratterizzato in modo peggiore, infatti, tutte le tecniche di *prompting* testate con la parola chiave *"truthful/deceptive"* registrano delle scarse performance; anche se, si nota un miglioramento con la tecnica *5-shot prompting* rispetto alla baseline.

Mediante l'impiego del software JASP, è stato condotto un test binomiale al fine di valutare se le accuratezze osservate differissero in maniera statisticamente significativa da una proporzione di riferimento, specificamente, la probabilità di selezione casuale di 0.5. Le accuratezze che hanno superato la soglia di significatività sono state evidenziate nella Figura 4, utilizzando la seguente notazione: * per <0.05; ** per <0.01; *** per p < 0.001. Nello specifico, nessuno dei dataset e nessuna strategia di *prompting* analizzati mostrano delle accuratezze statisticamente significative.

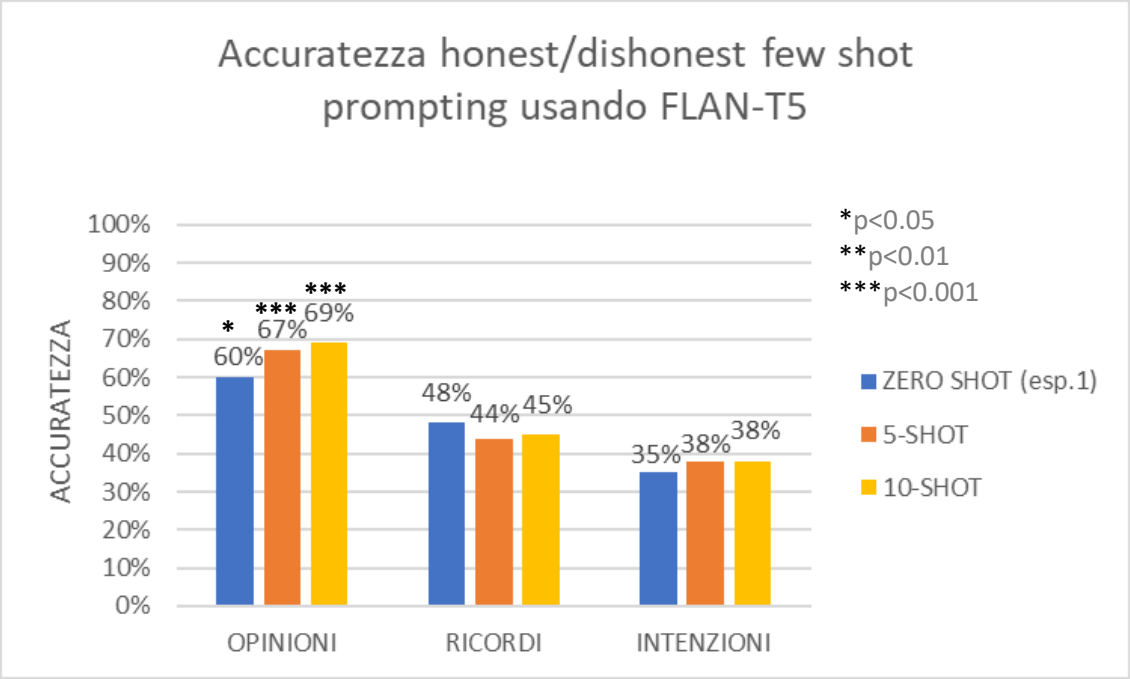

Figura 10: *Confronto tra le accuratezze usando FLAN-T5 con 5-shot prompting e 10-shot prompting testando nell'output la parola chiave honest/dishonest*

La Figura 10 mostra l'accuratezza ottenuta dal modello FLAN-T5 nei tre dataset quando l'output è stato ottenuto utilizzando nel prompt le parole chiave "*honest/dishonest*" e confrontando la baseline ricavata dall'esperimento 1 con la tecnica *5-shot* e *10-shot*.

Dal grafico sopra riportato si evince che la migliore accuratezza è stata ottenuta nel dataset delle opinioni, a cui seguono, rispettivamente, quello dei ricordi e delle intenzioni.

In particolare, per quanto riguarda il dataset delle opinioni, il modello ottiene delle buone performance per le strategie *5-shot* e *10-shot prompting*, registrando anche accuratezze che differiscono in maniera statisticamente significativa dalla proporzione di riferimento.

Per quanto riguarda il dataset dei ricordi, le tre tecniche presentano un andamento similare ma comunque non sufficiente per registrare delle buone performance o risultati statisticamente rilevanti.

Per il dataset delle intenzioni, l'accuratezza presenta i risultati peggiori, raggiungendo delle scarse performance sia con entrambe le tecniche di *prompting* testate.

Rispetto alla baseline l'accuratezza del dataset delle opinioni e delle intenzioni migliora, al contrario nel dataset dei ricordi il livello di accuratezza scende; pertanto, sembra che aver fornito degli esempi in questo caso possa essere risultato fuorviante per il modello.

Mediante l'impiego del software JASP, è stato condotto un test binomiale al fine di valutare se le accuratezze osservate differissero in maniera statisticamente significativa da una proporzione di riferimento, specificamente, la probabilità di selezione casuale di 0.5. Le accuratezze che hanno superato la soglia di significatività sono state evidenziate nella Figura 4, utilizzando la seguente notazione: * per <0.05; ** per <0.01; *** per $p < 0.001$. Nello specifico, le accuratezze ottenute con il modello FLAN-T5 nel dataset delle opinioni con le strategie *5-shot* e *10-shot* risultano statisticamente significative.

3.3 Esperimento 3

3.3.1 Obiettivo e ipotesi dello studio

Nel terzo esperimento, a fronte delle buone prestazioni ottenute da GPT-3.5 con la tecnica *zero shot*, si è voluto manipolare il contesto aggiungendo ulteriori informazioni tramite la funzione *custom instruction* disponibile in GPT-3.5. Questo esperimento è stato condotto per verificare se l'aggiunta di preferenze e requisiti specifici desiderati dall'utente potesse istruire in modo più efficace il modello e quindi generare output più accurati e coerenti con il contesto.

L'ipotesi prevede che modificando il contesto si rilevi una maggiore significatività con la versione *custom-instruction cognitive-load* rispetto alla versione *custom-instruction* base, in quanto i risultati di diversi studi condotti mostrano che le caratteristiche linguistiche associate alla teoria psicologica del carico cognitivo influenzano positivamente le previsioni del modello (Zhou et al., 2004; Solà-Sales et al., 2023; Loconte et al., 2023b).

3.3.2 Materiali e metodi

LLM

A tale scopo, abbiamo selezionato il modello GPT-3.5 già descritto nel paragrafo 3.1.2.

Dataset

Per testare le capacità di GPT-3.5 sono stati utilizzati gli stessi dataset precedentemente descritti nel paragrafo 3.1.2.

Procedura

L'opzione *custom instruction* è stata introdotta di recente da OpenAI per consentire a GPT-3.5 di considerare le preferenze dell'utente durante la generazione delle risposte. In pratica, questa funzione consente di personalizzare il comportamento della chatbot.

Per utilizzare questa funzione è necessario accedere al sito di ChatGPT[2] e, dalle impostazioni, selezionare l'opzione *custom instruction* come segue:

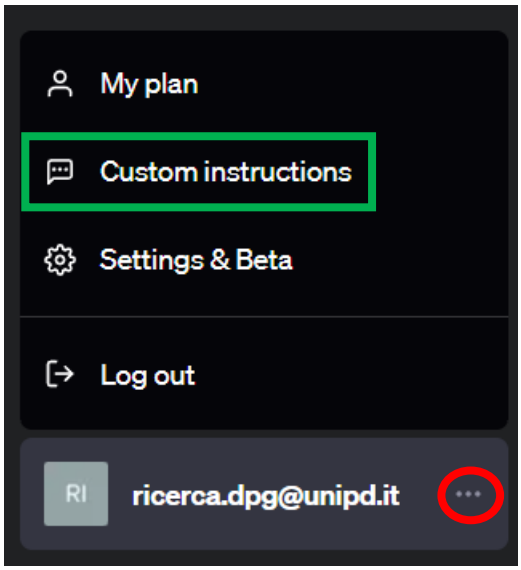

Figura 11 *Come accedere all'opzione custom instruction*

Le istruzioni personalizzate riguardano principalmente due domande:

1. nel primo riquadro, l'utente specifica cosa desidera che GPT-3.5 sappia su di lui al fine di ottenere risposte migliori;
2. nel secondo riquadro, l'utente indica come desidera che GPT-3.5 risponda, ciò può includere per esempio il formato di risposta, il tono e il contenuto della risposta.

[2] Sito ChatGPT: https://chat.openai.com/

Una volta attivata e configurata questa funzione, GPT-3.5 terrà conto di queste informazioni in ogni conversazione futura.

Sulla base di queste considerazioni, sono state identificate e testate due differenti versioni: *custom-instruction* base vs. *custom-instruction cognitive load*.

Per la versione custom-instruction base, nel primo riquadro sono state inserite le seguenti informazioni:

- professione: *"I'm a forensic neuropsychologist and a AI classifier tasked with determining if each statement provided comes from a person who is telling the truth or is lying"*;
- obiettivo: *"Identify correctly whether each statement is true or false and their authenticity based on the information provided, not to judge the subjective opinions expressed. You should evaluate the statements solely based on their factual accuracy, consistency, and logical coherence"*.

Nel secondo riquadro sono state riportate delle istruzioni in base al formato di risposta, in particolare *"Return 'true' if the statement is likely to be true and 'false' if it is likely to be false"*.

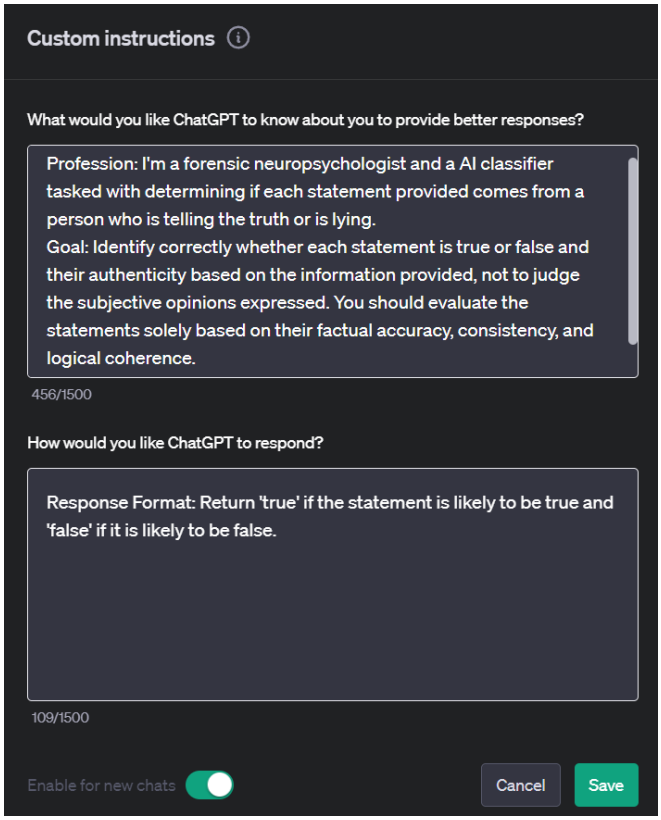

Figura 12: *versione base custom-instruction*

Per la *versione custom-instruction cognitive load*, a fronte di alcune ricerche precedenti (Zhou et al., 2004; Solà-Sales et al., 2023; Loconte et al., 2023b) che hanno messo in evidenza come possa sussistere un legame tra lunghezza, leggibilità e complessità del testo sulla base della teoria del carico cognitivo sono state inserite alcune informazioni che prendessero spunto da queste considerazioni. Successivamente vengono riportate le indicazioni scelte per il primo riquadro:

- professione: *"I'm a forensic neuropsychologist and a AI classifier tasked with determining if each statement provided comes from a person who is telling the truth or is lying"*;

- contesto/attuale linea di pensiero: *"According to Cognitive Load Theory, false statement are characterized by being shorter and less complex than true statement"*;
- obiettivo: *"Identify correctly whether each statement is true or false and their authenticity based on the metrics provided below. Those metrics should be interpeted along with the Cognitive Load Theory expressed in the Background"*.

Nel secondo riquadro vengono inserite le istruzioni rispetto a:

- formato di risposta: *"1. Compute the following metrics associated with the Cognitive Load Theory: 1. number of words; 2. number of sentences; 3. number of syllables; 4. average words per sentence; 5. average syllables per word; 6. Flesch-Kincaid Grade Level; 2 Based on the scores obtained, return 'true' if the statement is likely to be true and 'false' if it is likely to be false"*;
- tono: *"I need concise response"*.

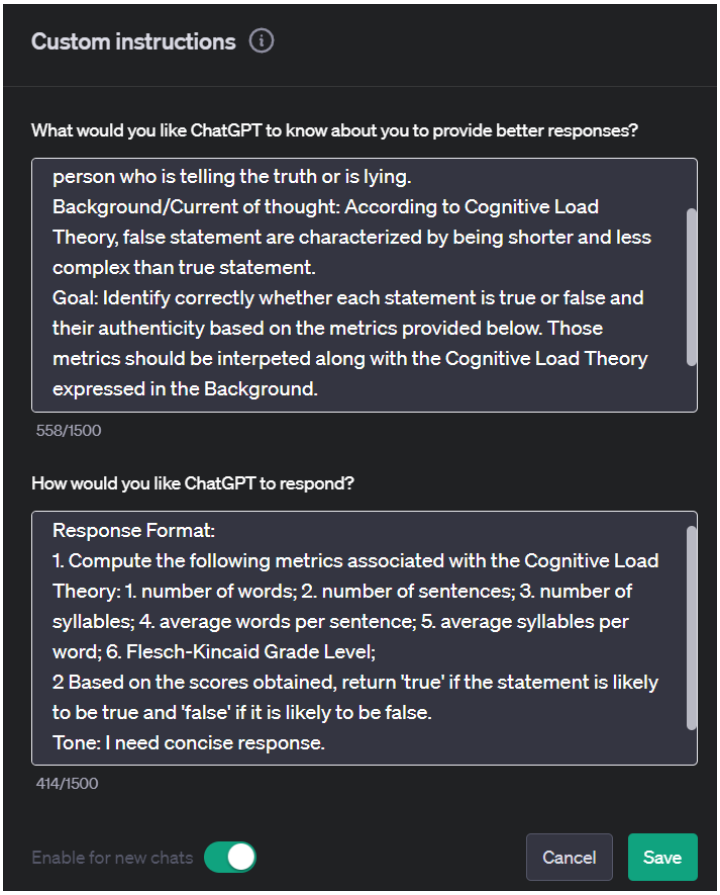

Figura 13: *versione custom-instruction cognitive load*

È stato chiesto al modello di calcolare, prima di fornire l'output finale, alcuni parametri, ovvero il numero di frasi, di parole, di sillabe, il numero medio di sillabe per parola e il livello di grado *Flesch-Kincaid*.

Il *Flesch-Kincaid Grade Level* è stato utilizzato per determinare la difficoltà di comprensione delle dichiarazioni. Questo parametro verifica la leggibilità e la complessità del testo in base alla lunghezza delle parole e della frase. I valori variano da 0 a 18 e l'indice può essere interpretato come gli anni di istruzione generalmente richiesti e necessari al fine di comprendere il testo.

Per entrambe le versioni *custom instruction* sono state selezionate le stesse 100 opinioni, gli stessi 100 ricordi e le stesse100 intenzioni testate nell'esperimento 1, per un confronto più pulito con il metodo *zero-shot* precedentemente testato su GPT-3.5.

3.3.2 Risultati e discussione

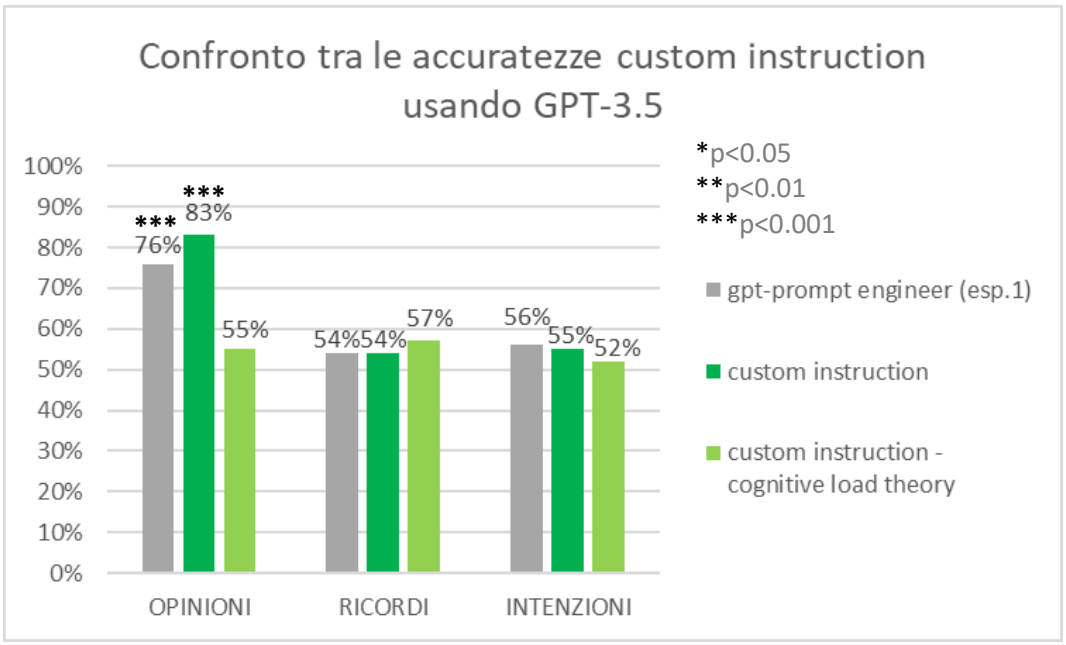

Figura 14: *Confronto tra le accuratezze usando GPT-3.5 con la versione custom-instruction base e la versione custom-instruction cognitive load*

I principali risultati ottenuti vengono mostrati nel grafico di Figura 14 che presenta, sull'asse delle ordinate, i livelli di accuratezza e, sull'asse delle ascisse, i dataset oggetto di analisi. Come si evince dalla legenda, in grigio sono rappresentate le accuratezze, espresse in percentuale, ottenute con l'LLM GPT-3.5 con la tecnica *zero-shot prompting*, riportate come baseline dell'esperimento; in verde scuro e in verde chiaro sono illustrate le accuratezze, espresse in percentuale, ottenute, rispettivamente, con la versione *custom-instruction* base e con la *versione custom-instruction cognitive load*.

In generale, come si evince dalla Figura 12, è possibile affermare che il dataset delle opinioni presenta performance migliori rispetto ai dataset dei ricordi e delle intenzioni.

In particolare, l'LLM GPT-3.5 con la tecnica *zero-shot prompting* e la versione *custom-instruction* presenta dei valori di accuratezza molto buoni. Ciò implica che i modelli specificati nel dataset in esame registrano delle performance molto buone. La performance si abbassano, invece, utilizzando il modello *custom-instruction cognitive load*.

La differenza tra i valori riportati potrebbe essere spiegata analizzando il *Flesch-Kincaid Grade Level*. Secondo tale metrica, la complessità di una frase può essere classificata secondo i range che seguono:

- 0-1: livello base per chi ha appena imparato a leggere libri;
- 1-5: molto facile da leggere;
- 5-11: livello medio;
- 11-18: livello per lettori esperti.

Dalle analisi relative ai range sopra riportati, è emerso quanto segue:

- 0-5: l'opinione viene giudicata poco complessa e, in accordo con quanto riportato dalla teoria del carico cognitivo, viene classificata come falsa;
- 5-11: l'opinione viene giudicata mediamente complessa, pertanto GPT-3.5 fatica a classificarla correttamente;
- 11-18: l'opinione viene giudicata complessa e, in accordo con quanto riportato dalla teoria del carico cognitivo, viene classificata come vera.

Dunque, le basse performance ottenute da GPT-3.5, nel caso di *custom instruction-cognitive load*, possono essere spiegate dalle opinioni che sono state classificate nel range 5-11.

I dataset dei ricordi e delle intenzioni presentano performance inferiori rispetto al dataset delle opinioni.

In particolare, analizzando il dataset dei ricordi, come si evince dalla figura 12, l'accuratezza del modello GPT-3.5 non ottiene delle performance soddisfacenti.

Infine, per quanto riguarda l'accuratezza ottenuta dal modello GPT-3.5 nel dataset delle intenzioni, come si evince dalla figura 12, peggiora sia nel caso della versione base *custom instruction* che della versione *custom instruction- cognitive load*, rispetto al livello di baseline.

Mediante l'impiego del software JASP, è stato condotto un test binomiale al fine di valutare se le accuratezze osservate differissero in maniera statisticamente significativa da una proporzione di riferimento, specificamente, la probabilità di selezione casuale di 0.5. Le accuratezze che hanno superato la soglia di significatività sono state evidenziate nella Figura 4, utilizzando la seguente notazione: * per <0.05; ** per <0.01; *** per $p < 0.001$. Nello specifico solo le accuratezze ottenute l'LLM GPT-3.5 con la tecnica *zero-shot prompting* e con la versione *custom-instruction* base nel dataset delle opinioni risultano statisticamente significative.

3.4 Confronto tra le procedure di *prompting* nei tre esperimenti

Al fine di confrontare quale fosse la strategia di *prompting* che ottiene l'accuratezza più elevata e significativa rispetto alle altre strategie utilizzando il modello FLAN-T5 e GPT-3.5 abbiamo condotto il test multinomiale.

È stato condotto un test multinomiale per verificare se tra le diverse tecniche di prompt testate, che in maniera significativa hanno raggiunto una accuratezza superiore al caso, ce ne fosse qualcuna che si distinguesse per l'accuratezza ottenuta. Il confronto tra le accuratezze riportate dalle varie tecniche di *prompting* è riportato in Figura 15. I risultati hanno indicato l'assenza di una differenza significativa tra le strategie di *prompting* ($\chi 2(5) = 3.520$, $p = 0.620$). I dettagli dei rapporti di probabilità per ogni categoria sono riportati nella Tabella 3.

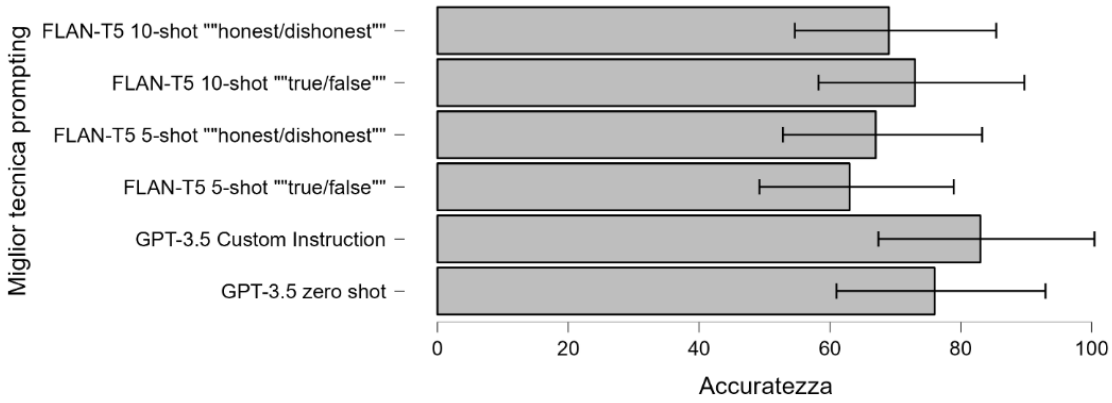

***Figura 15**: Confronto tra le accuratezze riportate dalle diverse tecniche di prompting mediante il test multinomiale*

Best prompting strategy	Observed	Expected: Multinomial	95% Confidence Interval	
			Lower	Upper
FLAN-T5 10-shot ""honest/dishonest""	69	71.833	54.629	85.418
FLAN-T5 10-shot ""true/false""	73	71.833	58.261	89.728
FLAN-T5 5-shot ""honest/dishonest""	67	71.833	52.819	83.256
FLAN-T5 5-shot ""true/false""	63	71.833	49.214	78.920
GPT-3.5 Custom Instruction	83	71.833	67.412	100.434
GPT-3.5 zero shot	76	71.833	60.996	92.950

Note. Confidence intervals are based on independent binomial distributions.

Tabella 3: *Rapporti di probabilità per ogni categoria*

CAPITOLO 4: DISCUSSIONE

4.1 Interpretazione dei risultati

L'obiettivo del presente elaborato è stato quello di istruire due LLMs, in particolare FLAN-T5 e GPT-3.5, nel rilevamento della menzogna attraverso la tecnica del *prompt-engineering.* A tale scopo sono stati condotti tre esperimenti volti a testare diverse tipologie di prompt in un compito di classificazione della menzogna su tre dataset contenenti opinioni personali (Capuozzo et al., 2020), ricordi autobiografici (Sap et al., 2020) e intenzioni future (Ilias et al., 2022).

Nel primo esperimento, FLAN-T5 e GPT-3.5 sono stati testati utilizzando la tecnica *zero-shot prompting*. L'ipotesi prevedeva che una strategia semplice non fosse sufficiente per ottenere performance significativamente superiori alla probabilità di andare a caso e che il modello GPT-3.5 ottenesse un'accuratezza maggiore rispetto al modello FLAN-T5 grazie alle competenze linguistiche maggiori. I risultati principali, come previsto, hanno mostrato che il modello GPT-3.5 ha ottenuto un'accuratezza migliore per tutti i dataset rispetto al modello FLAN-T5. Il modello GPT-3.5, come ipotizzato nell'esperimento 1, ha ottenuto delle performance più elevate rispetto al modello FLAN-T5 poiché è un modello allenato su più parametri e, pertanto, presenta delle competenze linguistiche generali maggiori.

Nel secondo esperimento si è cercato di migliorare l'accuratezza della performance di FLAN-T5 testando una strategia *few-shot*. L'ipotesi prevedeva che una strategia di prompt più sofisticata potesse raggiungere delle performance migliori e che testando l'uso di diverse parole chiave nell'output *("true/false"*; *"honest/dishonest"*; *"truthful/deceptive"*) l'accuratezza non ne fosse influenzata. I risultati principali hanno

mostrato che la parola chiave *"truthful/deceptive"* riporta l'accuratezza peggiore sia con la strategia *5-shot prompting* sia con quella *10-shot prompting*. Con la tecnica *5-shot prompting* è stata ottenuta l'accuratezza migliore nel dataset delle opinioni testando nell'output la parola chiave *"honest/dishonest"*. Con la tecnica *10-shot prompting* le performance migliori nel dataset delle opinioni sono state, invece, raggiunte testando nell'output la parola chiave *"true/false"*. Pertanto, questo dimostra che, a differenza delle aspettative, le diverse parole chiave influenzano l'accuratezza.

Nel terzo esperimento, considerando i risultati soddisfacenti ottenuti da GPT-3.5 con la strategia *zero-shot* del primo esperimento, si è tentato di manipolare il contesto mediante l'utilizzo dell'impostazione personalizzata di GPT-3.5, nota come *"custom instruction"*. L'ipotesi prevedeva che modificando il contesto si raggiungesse una maggiore significatività con la versione *custom-instruction cognitive-load* rispetto alla versione *custom-instruction* base. Dai risultati sopra menzionati si evince che le metriche linguistiche riferite alla teoria del carico cognitivo non hanno, in realtà, mostrato i risultati aspettati. In particolare, dalle performance ottenute si evince che siano risultate fuorvianti per il modello; infatti, le performance migliori sono state ottenute dalla versione *custom-instruction* base.

In generale, il dataset che è stato meglio riconosciuto da entrambi i modelli è stato quello delle opinioni poiché risultano più semplici e brevi rispetto ai ricordi e alle intenzioni che sono più articolati e, dunque, più difficili da identificare correttamente.

Dai risultati ottenuti mediante il test binomiale eseguito sui tre diversi esperimenti è possibile affermare che il dataset delle opinioni risulta statisticamente significativo con:

- la tecnica *zero-shot prompting* sia con il modello FLAN-T5 che con GPT-3.5;

- la tecnica *few-shot prompting* testando nell'output le parole chiave *"true/false"* e *"honest/dishonest"* usando FLAN-T5;
- la versione base dell'opzione *custom-instruction* usando il modello GPT-3.5.

Infine, è possibile sottolineare che dai risultati del test multinomiale tutte le strategie di *prompting* testate nei tre diversi esperimenti si equivalgono e che nessuna di queste risulta più statisticamente significativa rispetto alle altre.

4.2 Limiti e sviluppi futuri

La validità ecologica è il limite principale dello studio presentato. I risultati ottenuti mediante l'utilizzo dei *Large Language Models* in un contesto sperimentale potrebbero, infatti, non riflettere accuratamente i comportamenti di una persona in ambito forense.

Questa limitazione sottolinea la necessità di ricerche future per poter espandere l'applicabilità e la generalizzabilità dei modelli di *Machine* e *Deep Learning* addestrati al rilevamento della menzogna in contesti di vita reale. Pertanto, è necessario condurre nuove ricerche.

Il presente studio è stato il primo a testare l'approccio del *prompt engineering* in un compito di classificazione per identificare la menzogna, pertanto sarebbe interessante:

- riprodurre gli esperimenti utilizzando nuovi dataset già noti in letteratura come DECOUR, acronimo di *DEception in COURt corpus*, costruito con l'obiettivo di addestrare i modelli a discriminare tra affermazioni sincere e ingannevoli in altri contesti da quelli già testati (Fornaciari & Poesio, 2012);
- condurre ulteriori ricerche testando le capacità di altri modelli di intelligenza artificiali come GPT-4 e Llama 2 sui dataset utilizzati per la presente ricerca;

- testare altre tecniche di prompt più sofisticate come il *Tree-of-thoughts*.

Inoltre, risulta interessante sottolineare che i modelli FLAN-T5 e GPT-3.5 utilizzati nel presente studio potrebbero essere paragonati a dei partecipanti in un esperimento volto a valutare la loro capacità di rilevamento della menzogna. L'approccio analogo viene adottato in uno studio di psicologia cognitiva condotto sugli esseri umani. In linea con i risultati emersi dallo studio di Capuozzo e colleghi (2020) i modelli di *Machine Learning* raggiungono performance superiori rispetto agli esseri umani, che in questo compito, ottengono un'accuratezza di circa il 58%. Sulla base di queste considerazioni essendo che i modelli superano le prestazioni degli esseri umani ottenendo risultati promettenti risulta importante continuare a condurre ricerche per apportare vantaggi e progressi significativi alla disciplina psicologica.

CAPITOLO 5: CONCLUSIONI

Lo scopo del presente studio è stato quello di istruire due LLMs, in particolare FLAN-T5 e GPT-3.5 in un compito di classificazione della menzogna utilizzando la tecnica del *prompt-engineering*. Sono stati, dunque, condotti tre esperimenti volti a testare diverse strategie di *prompt* su tre diversi dataset contenenti opinioni personali, ricordi autobiografici e intenzioni future. Nel primo esperimento vengono confrontate le accuratezze raggiunte con i modelli FLAN-T5 e GPT-3.5 utilizzando la tecnica *zero-shot prompting*. Nel secondo esperimento, viste le scarse performance ottenute dal modello FLAN-T5 nell'esperimento 1, è stata utilizzata la tecnica *few-shot*. Questo approccio aveva come scopo quello di valutare sei inserire nel *prompt* degli esempi (prima 5 e poi 10) potesse guidare meglio il modello a fornire risposte più accurate. Nello stesso esperimento sono state testate anche diverse parole chiave (*"true/false"*; *"truthful/deceptive"*; *"honest/dishonest"*) nell'output per verificare come queste influenzassero le performance del modello. Nel terzo esperimento, viste le performance soddisfacenti ottenute dal modello GPT-3.5 nell'esperimento 1, si è tentato di manipolare il contesto mediante le opzioni *custom-instruction* base e *custom-instruction cognitve-load*.

Dalle analisi principali è emerso che GPT-3.5 riporta delle performance superiori rispetto a FLAN-T5 poiché le dimensioni maggiori del modello presentano migliori capacità linguistiche. Inoltre, analizzando i dataset utilizzati, quello meglio caratterizzato risulta essere quello delle opinioni in tutti gli esperimenti. Ciò implica che le opinioni sono più facili da riconoscere correttamente rispetto ai ricordi e alle intenzioni. A seguito degli esperimenti condotti è possibile affermare che per quanta riguarda il modello FLAN-T5 le performance nel dataset delle opinioni migliorano con la tecnica *few-shot* testando

nell'output le parole chiave *"true/false"* e *"honest/dishonest"*. Per quanto riguarda il modello GPT-3.5 l'accuratezza nel dataset delle opinioni migliora con la versione *custom-instruction* base.

In conclusione, è possibile affermare che l'accuratezza nel dataset delle opinioni, ottenuta mediante i modelli utilizzati, è superiore rispetto a quella degli esseri umani.

RIFERIMENTI BIBLIOGRAFICI

Abe, N. (2011). How the Brain Shapes Deception: An Integrated Review of the Literature. *The Neuroscientist*, 17(5), 560–574.

Azizli, N., Atkinson, B. E., Baughman, H. M., Chin, K., Vernon, P. A., Harris, E., & Veselka, L. (2016). Lies and crimes: Dark triad, misconduct, and high-stakes deception. *Personality and Individual Differences*, 89, 34–39.

Ben-Shakhar G & Elaad E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: a meta-analytic review. *Journal of Applied Psychology*, 88, 131–151.

Benzon, W. L. (2023). Discursive Competence in ChatGPT, Part 1: Talking with Dragons. *SSRN Electronic Journal*.

Binz, M., & Schulz, E. (2023) Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences of the United States of America*, 120.

Bond, C. F., & DePaulo, B. M. (2006). Accuracy of deception judgments. *Personality and Social Psychology Review*, 10, 214-234.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, vol. 33.

Capuozzo, P., Lauriola, I., Strapparava, C., Aiolli, F., & Sartori, G. (2020). DecOp: A multilingual and multi-domain corpus for detecting deception in typed text. In *Proceedings of the 12th Language Resources and Evaluation Conference* (pp. 1423-1430).

Chakraborty, S., & Pakray, P. (2023). Abstractive Summarization Evaluation for Prompt Engineering. *International Visual Informatics Conference*.

Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W et al. (2022). Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Cohn, J. F., Ambadar, Z., & Ekman, P. (2007). Observer-based measurement of facial expression with the Facial Action Coding System. *The Handbook of Emotion Elicitation and Assessment*, 1(3), 203-221.

Cole, T. (2001). Lying to the one you love: The use of deception in romantic relationships. *Journal of Social and Personal Relationships*, 18(1), 107–129.

Constâncio, A. S., Tsunoda, D. F., Silva, H. de F. N., Silveira, J. M. da, & Carvalho, D. R. (2023). Deception detection with machine learning: A systematic review and statistical analysis. *PLoS ONE, 18, e0281323*.

Cunningham, P., Cord, M., Delany, S.J. (2008). Supervised Learning. In: Cord, M., Cunningham, P. (eds) Machine Learning Techniques for Multimedia. *Cognitive Technologies. Springer, Berlin, Heidelberg*.

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM 58, 9, 92–103*.

Debey, E., De Houwer, J., & Verschuere, B. (2014). Lying relies on the truth. *Cognition, 132(3), 324–334*.

DePaulo, B. M., Kashy, D. A., Kirkendol, S. E., Wyer, M. M., & Epstein, J. A. (1996). Lying in everyday life. *Journal of Personality and Social Psychology, 70, 979–995*.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., & Cooper, H. (2003). Cues to deception. *Psychological Bulletin, 129(1), 74-118*.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805.*

Dhingra, S., Singh, M., SB, V., Malviya, N., Gill, S. S. (2023). Mind meets machine: Unravelling GPT-4's cognitive psychology. *arXiv preprint arXiv: 2303.11436*.

Donchin, E., & Coles, M. G. H. (1988). Is the P300 component a manifestation of context updating? *Behavioral and Brain Sciences, 11(3), 357-374*.

Demszky, D., Yang, D., Yeager, D. S., Bryan, C. J., Clapper, M., Chandhok, S et al. (2023). Using large language models in psychology. *Nature Reviews Psychology*.

Ekman, P., & Friesen, W. V. (1978). Facial Action Coding System: A Technique for the Measurement of Facial Movement. *Consulting Psychologists Press, Palo Alto*.

Ekman, P. (1988). Lying and nonverbal behavior: Theoretical issues and new findings. *Journal of Nonverbal Behavior, 12, 163-175*.

Ekman P, O'Sullivan M. (1991). Who can catch a liar? *American Psychologist, 46(9), 913–920*.

Ekman, P. (2003). Darwin, deception, and facial expression. *Annals of the New York Academy of Sciences, 1000(1), 205-221*.

Ekman, P. (2011). *I volti della menzogna. Gli indizi dell'inganno nei rapporti interpersonali, negli affari, nella politica, nei tribunali*. Giunti.

Farah, M. J., Hutchinson, J. B., Phelps, E. A., Wagner, A.D. (2014). Functional MRI-based lie detection: scientific and societal challenges. *Nature Reviews Neuroscience 15, 123-131*.

Farwell, L. A., & Donchin, E. (1991). The Truth Will Out: Interrogative Polygraphy ("lie detection") with event-related brain potentials. *Psychophysiology, 28(5), 531-547*.

Feeley, T. H., & DeTurck, M. A. (1997). Perceptions of communications as seen by the actor and as seen by the observer: The case of lie detection. *In International Communication Association Annual Conference, Montreal, Canada*.

Fornaciari, T., & Poesio, M. (2012, May). DeCour: a corpus of DEceptive statements in Italian COURts. *In LREC (pp. 1585-1590)*.

Fornaciari, T., Bianchi, F., Poesio, M., & Hovy, D. (2021). BERTective: Language models and contextual information for deception detection. *In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. Association for Computational Linguistics*.

Fullam, R. S., McKie, S., & Dolan, M. C. (2009). Psychopathic traits and deception: functional magnetic resonance imaging study. *The British Journal of Psychiatry, 194(3), 229-235*.

Ganis, G., Kosslyn, S. M., Stose, S., Thompson, W. L., & Yurgelun-Todd, D. A. (2003). Neural correlates of different types of deception: an fMRI investigation. *Cerebral cortex, 13(8), 830-836*.

Gao, C. A., Howard, F. M., Markov, N. S., Dyer, E. C., Ramesh, S., Luo, Y., & Pearson, A. T. (2022). Comparing scientific abstracts generated by ChatGPT to original abstracts using an artificial intelligence output detector, plagiarism detector, and blinded human reviewers. *BioRxiv, 2022-12*.

Gimelli, C. (2012). La mente mente, il volto no. Gli indizi dell'inganno nei rapporti interpersonali, negli affari, nella politica, nei Tribunali. *Linguæ &-Rivista di lingue e culture moderne, 11(1-2), 147-155*.

Granhag, P. A., & Strömwall, L. A. (1998). Let's go over this again...": Effects of repeated interrogations on deception detection performance. *Paper presented at the eighth European Conference on Psychology and Law, Cracow, Poland*.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of personality and social psychology, 74(6), 1464*.

Grubin, D., & Madsen, L. (2005). Lie detection and the polygraph: A historical review. *The Journal of Forensic Psychiatry & Psychology, 16(2), 357-369*.

Hart, C. L., Lemon, R., Curtis, D. A., & Griffith, J. D. (2020). Personality traits associated with various forms of lying. *Psychological Studies, 65, 239-246*.

Hastie, T., Tibshirani, R., Friedman, J., Hastie, T., Tibshirani, R., & Friedman, J. (2009). Unsupervised learning. *The elements of statistical learning: Data mining, inference, and prediction, 485-585*.

Hauser, M. D., Chomsky, N., & Fitch, W. T. (2002). The faculty of language: what is it, who has it, and how did it evolve?. *science, 298(5598), 1569-1579*.

Hazlett, G. (2006). Research on detection of deception: What we know vs. what we think we know. *NDIC, Educing information interrogation: Science and art foundations for the future, 45-62*.

Ilias, L., Soldner, F., & Kleinberg, B. (2022). Explainable Verbal Deception Detection using Transformers. *arXiv preprint arXiv:2210.03080*.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78, 15169-15211*.

Johnson, M. K., & Raye, C. L. (1981). Reality monitoring. *Psychological review, 88(1), 67*.

Joshi, M., Chen, D., Liu, Y., Weld, D. S., Zettlemoyer, L., & Levy, O. (2020). Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics, 8, 64-77*.

Kashy, D. A., & DePaulo, B. M. (1996). Who lies?. *Journal of Personality and Social Psychology, 70(5), 1037*.

Kircher, J. C., & Raskin, D. C. (2019). Polygraph techniques: History, controversies, and prospects. *Psychology and social policy, 295-308*.

Kleinberg, B., & Verschuere, B. (2021). How humans impair automated deception detection performance. *Acta psychologica, 213, 103250*.

Kleiner, M. E. (2002). Handbook of polygraph testing. *Academic Press*.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y., & Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *Advances in neural information processing systems, 35, 22199-22213*.

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023a). Challenging ChatGPT'Intelligence'with Human Tools: A Neuropsychological Investigation on Prefrontal Functioning of a Large Language Model. *Intelligence*.

Long, J. (2023). Large Language Model Guided Tree-of-Thought. *arXiv preprint arXiv:2305.08291*.

Lykken, D. T. (1959). The GSR in the detection of guilt. *Journal of Applied Psychology, 43(6), 385*.

Maschke, G. W., & Scalabrini, G. J. (2005). The lie behind the lie detector. *Antipolygraph org*.

McCornack, S. A., & Parks, M. R. (1986). Deception detection and relationship development: The other side of trust. *Annals of the International Communication Association, 9(1), 377-389*.

Muris, P., Merckelbach, H., Otgaar, H., & Meijer, E. (2017). The malevolent side of human nature: A meta-analysis and critical review of the literature on the dark triad (narcissism, Machiavellianism, and psychopathy). *Perspectives on psychological science, 12(2), 183-204*.

O'Sullivan, M., & Ekman, P. (2004). 12 the wizards of deception detection. *The detection of deception in forensic contexts, 269*.

Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

Otter-Henderson, K. D., Honts, C. R., & Amato, S. L. (2002). Spontaneous countermeasures during polygraph examinations: An apparent exercise in futility. *Polygraph, 31(1), 9-13*.

Pérez-Rosas, V., & Mihalcea, R. (2015). Experiments in open domain deception detection. *In Proceedings of the 2015 conference on empirical methods in natural language processing (pp. 1120-1125)*.

Peth, J., Suchotzki, K., & Gamer, M. (2016). Influence of countermeasures on the validity of the Concealed Information Test. *Psychophysiology, 53(9), 1429-1440*.

Pinker, S., & Morey, A. (2014). The Language Instinct: How the Mind Creates Language (Unabridged edition). *Brilliance Audio*.

Polak, M. P., & Morgan, D. (2023). Extracting Accurate Materials Data from Research Papers with Conversational Language Models and Prompt Engineering--Example of ChatGPT. *arXiv preprint arXiv:2303.05352*.

Qin, C., Zhang, A., Zhang, Z., Chen, J., Yasunaga, M., & Yang, D. (2023). Is ChatGPT a general-purpose natural language processing task solver?. *arXiv preprint arXiv:2302.06476*.

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences, 63(10), 1872-1897*.

Raskin, D. C., & Honts, C. R. (2002). The comparison question test.

Reynolds, L., & McDonell, K. (2021, May). Prompt programming for large language models: Beyond the few-shot paradigm. *In Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-7)*.

Sap, M., Horvitz, E., Choi, Y., Smith, N. A., & Pennebaker, J. (2020, July). Recollection versus imagination: Exploring human memory and cognition via neural language models. *In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 1970-1978)*.

Sapir, A. (2000). The LSI course on scientific content analysis: Workshop book. *Phoenix, Arizona: Laboratory of Scientific Interrogation*.

Sartori, G. (2021). La memoria del testimone: Dati scientifici utili a magistrati, avvocati e consulenti. *Giuffrè Francis Lefebvre*.

Sartori, G., Agosta, S., Zogmaister, C., Ferrara, S. D., & Castiello, U. (2008). How to accurately detect autobiographical events. *Psychological science, 19(8), 772-780*.

Sartori, G., & Orrù, G. Large Language Models and Psychological Science. *Frontiers in Psychology, 14, 1279317*.

Solà-Sales, S., Alzetta, C., Moret-Tatay, C., & Dell'Orletta, F. (2023). Analysing Deception in Witness Memory through Linguistic Styles in Spontaneous Language. *Brain Sciences, 13(2), 317*.

Soldner, F., Pérez-Rosas, V., & Mihalcea, R. (2019, June). Box of lies: Multimodal deception detection in dialogues. *In Proceedings of the 2019 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 1768-1777).

Staunton, C., & Hammond, S. (2011). An investigation of the guilty knowledge test polygraph examination. *Journal of Criminal Psychology, 1(1), 1-14*.

Suchotzki, K. (2018). Challenges for the Application of Reaction Time–Based Deception Detection Methods. *In Detecting Concealed Information and Deception (pp. 243-268). Academic Press*.

Sun, F. (2022). ChatGPT, the start of a new era.

Tausczik, Y. R., & Pennebaker, J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of language and social psychology, 29(1), 24-54*.

Tyler, J. M., Feldman, R. S., & Reichert, A. (2006). The price of deceptive behavior: Disliking and lying to people who lie to us. *Journal of Experimental Social Psychology, 42(1), 69-77*.

Hauch, V., Blandón-Gitlin, I., Masip, J., & Sporer, S. L. (2015). Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and social psychology Review, 19(4), 307-342*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems, 30*.

Vrij, A. (2000). Detecting lies and deceit: The psychology of lying and implications for professional practice. *Wiley*.

Vrij, A., & Mann, S. (2001). Telling and detecting lies in a high-stake situation: The case of a convicted murderer. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 15(2), 187-203*.

Vrij, A. (2005). Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law, 11(1), 3*.

Vrij, A. (2008). Detecting lies and deceit: Pitfalls and opportunities. *John Wiley & Sons*.

Vrij, A. (2014). Verbal lie detection tools: Statement validity analysis, reality monitoring and scientific content analysis. *Detecting deception: Current challenges and cognitive approaches, 1-35*.

Walczyk, J. J., Mahoney, K. T., Doverspike, D., & Griffith-Ross, D. A. (2009). Cognitive lie detection: Response time and consistency of answers as cues to deception. *Journal of Business and Psychology, 24, 33-49*.

Walczyk, J. J., Roper, K. S., Seemann, E., & Humphrey, A. M. (2003). Cognitive mechanisms underlying lying to questions: Response time as a cue to deception. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 17(7), 755-774*.

Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., ... & Zhang, S. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems, 35, 24824-24837*.

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of thoughts: Deliberate problem solving with large language models. *arXiv preprint arXiv:2305.10601*.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., ... & Wen, J. R. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zhou, L., Burgoon, J. K., Nunamaker, J. F., & Twitchell, D. (2004). Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications. *Group decision and negotiation, 13, 81-106*.

Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., ... & Irving, G. (2019). Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Zuckerman, M., DePaulo, B. M., & Rosenthal, R. (1981). Verbal and nonverbal communication of deception. *In Advances in experimental social psychology (Vol. 14, pp. 1-59). Academic Press*.

SITOGRAFIA

https://openai.com/research/gpt-4

https://github.com/dair-ai/Prompt-Engineering-Guide#guides