UNIVERSITÀ
DI PAVIA

UNIVERSITA' DEGLI STUDI DI PAVIA

DIPARTIMENTO DI STUDI UMANISTICI

CORSO DI LAUREA MAGISTRALE IN LINGUISTICA TEORICA, APPLICATA E DELLE LINGUE MODERNE

AI LITERACY NEL PUBBLICO ADULTO ITALIANO: UN'INDAGINE EMPIRICA E LA PROGETTAZIONE DI UN CHATBOT RAG COME STRUMENTO DIVULGATIVO DI ALFABETIZZAZIONE.

RELATRICE
Prof.ssa Chiara Zanchi

CORRELATRICI
Prof.ssa Claudia Roberta Combei
Prof.ssa Malvina Nissim

Tesi di Laurea Magistrale di
Gaia Eleonora Di Raimondo
Matricola n 558972

Anno accademico 2025/2026

1. Introduzione

L'uso dei sistemi di intelligenza artificiale generativa si è esteso rapidamente ben oltre la cerchia degli specialisti: chatbot basati su *Large Language Model* (LLM) come ChatGPT, Gemini o Copilot sono oggi impiegati da un pubblico ampio e variegato, per compiti che spaziano dalla ricerca di informazioni alla stesura di testi, e si estendono sempre più alla pratica di professionisti – medici (Blease et al., 2025), avvocati e commercialisti (Thomson Reuters Institute, 2025) – in ambiti decisionali ad alto rischio; a questa diffusione non corrisponde, però, una comprensione adeguata del meccanismo che ne governa le risposte. Da questo scarto tra diffusione e consapevolezza derivano rischi documentati dalla letteratura: la produzione – e la conseguente diffusione – di informazioni errate ma formalmente coerenti – le cosiddette allucinazioni –, la riproduzione negli output di bias presenti nei dati di addestramento, la compiacenza verso le aspettative dell'utente e una conseguente fiducia non calibrata sull'affidabilità effettiva del sistema (Ferrario et al., 2020; cfr. Capitolo 2).

È in questo scenario che si inserisce il concetto di AI *literacy*, letteralmente "alfabetizzazione" sul tema dell'Intelligenza Artificiale. L'etichetta "AI literacy" si riferisce all'insieme delle competenze epistemiche, critiche, pratiche ed etiche che consentono di comprendere, usare e valutare i sistemi di IA in modo consapevole (Long e Magerko, 2020; Ng et al., 2021; Laupichler et al., 2023). Le ricerche empiriche disponibili rilevano livelli di AI literacy generalmente contenuti tra gli utenti non specialisti (Wang et al., 2023; Rheu & Cho, 2025); nel panorama italiano, Savoldi et al. (2025) documentano una diffusione capillare degli strumenti generativi – con un tasso di utilizzo dell'80,4% in un campione di oltre 1.500 partecipanti – a fronte di un livello di alfabetizzazione critica ancora limitato. Un filone parallelo di studi, inoltre, ha esplorato il potenziale formativo dei chatbot: Wollny et al. (2021), in una rassegna sistematica, ne identificano i possibili ruoli come ausiliari, tutori e facilitatori dell'apprendimento, indicando nell'adattività e nella personalizzazione le direzioni di sviluppo più promettenti. Sul versante tecnico, il paradigma Retrieval-Augmented Generation (RAG), introdotto da Lewis et al. (2020), ha fornito un'architettura che consente di ancorare le risposte dei modelli a una base di conoscenza (*knowledge base*) verificata, riducendo il rischio di allucinazioni negli output e aumentando la trasparenza dell'informazione prodotta.

In questo quadro teorico e di ricerca si inserisce la presente tesi, che persegue due obiettivi complementari. Il primo è descrittivo: indagare e provare a quantificare, attraverso

un questionario, il livello di AI literacy in un campione di adulti italiani non specialisti, con attenzione alla comprensione del funzionamento dei sistemi, alla consapevolezza dei rischi e ai comportamenti d'uso dichiarati. Il secondo obiettivo, conseguenza diretta del primo, è progettuale: sviluppare e sottoporre a una prima valutazione esplorativa un chatbot educativo basato su architettura RAG, concepito come possibile strumento per rispondere ai bisogni formativi rilevati. Rispetto alla letteratura esistente, il lavoro combina un'indagine diagnostica con la progettazione di uno strumento educativo costruito con lo scopo di intervenire sulle criticità documentate, rivolgendosi a una popolazione – adulti non specialisti nel contesto italiano – relativamente poco indagata in questa prospettiva. La domanda centrale può essere così formulata: in quale misura un campione di adulti italiani non specialisti dispone delle competenze che costituiscono l'AI literacy, e in quale modo uno strumento educativo come quello qui progettato può contribuire ad affrontare le lacune documentate?

La tesi è organizzata in cinque capitoli. A questa Introduzione, segue il Capitolo 2, che costruisce il quadro teorico di riferimento: definisce i sistemi di IA e i Large Language Model, illustra i principali rischi connessi all'uso dei sistemi generativi (allucinazioni, bias, fiducia, *privacy*, *copyright*, *sycophancy*) e ricostruisce il dibattito sull'AI literacy, includendo una rassegna dello stato dell'arte della ricerca. Il Capitolo 3, successivamente, descrive la metodologia adottata nelle sue tre componenti: il questionario sull'AI literacy, la progettazione del chatbot RAG e il questionario per la valutazione del prototipo. Il Capitolo 4 presenta e interpreta i risultati, integrando l'analisi descrittiva dei dati del questionario con la discussione teorica e con i risultati del test pilota del chatbot. Il Capitolo 5, infine, raccoglie le conclusioni, illustrando i limiti della ricerca e le possibili direzioni di sviluppo futuro. Chiude il lavoro un'appendice, contenente i questionari somministrati e le tabelle complete dei dati.

2. Intelligenza Artificiale, modelli linguistici e AI Literacy: un quadro teorico

2.1 Intelligenza Artificiale (IA): una definizione

L'Intelligenza Artificiale (IA) può essere considerata una delle più rilevanti innovazioni del nostro tempo: nell'ultimo decennio, infatti, ha conosciuto una crescita significativa, affermandosi come una disciplina ampia, il cui impatto supera i confini della ricerca specialistica e coinvolge un pubblico sempre più vasto e sempre meno specializzato.

Definire, però, cosa sia l'IA, tracciandone i confini netti, è un'operazione complessa. Non esiste infatti una definizione univoca e stabile del termine, poiché esso si riferisce a un insieme eterogeneo di approcci, tecnologie e ambiti di ricerca.

L'etichetta "intelligenza artificiale" venne coniata da John McCarthy, in occasione della Conferenza di Dartmouth, tenutasi ad Hanover, New Hampshire nel 1956; in quella occasione, McCarthy definì l'IA come "*la scienza di creare macchine intelligenti*". Fin dalle sue origini, dunque, l'IA è stata concepita come un campo volto a riprodurre o emulare capacità tipicamente umane. Anche definizioni più recenti pongono l'accento su questa relazione uomo-macchina. Ad esempio, la voce "Intelligenza Artificiale" proposta da Treccani, basata su Somalvico (1987), dopo averne sottolineato la natura multidisciplinare, descrive l'IA come *"quella disciplina appartenente all'informatica, che studia i fondamenti teorici, le metodologie e le tecniche che consentono di progettare sistemi hardware e sistemi di programmi software capaci di fornire all'elaboratore elettronico prestazioni che, a un osservatore comune, sembrerebbero essere di pertinenza esclusiva dell'intelligenza umana"*.

La difficoltà nel fornire una definizione univoca e netta va connessa al fatto che l'etichetta "intelligenza artificiale" non si riferisce a una specifica tecnologia, ma a una vera e propria scienza, una disciplina, che include strumenti tecnologici di natura diversa (Proietti, 2024). In questo senso, "intelligenza artificiale" è considerabile un "termine ombrello" (Ruschemeier, 2023), capace di includere, al suo interno, più ambiti di ricerca, ognuno con i propri metodi, le proprie caratteristiche e le proprie applicazioni.

Anche sul piano istituzionale le definizioni di "intelligenza artificiale" adottate sono volutamente aperte e riflettono la natura mutevole e transitoria della disciplina, subordinata al progresso tecnologico. Nel documento *Recommendation on the Ethics of Artificial Intelligence* (2021), con cui UNESCO promuove, a livello globale, lo sviluppo e l'uso etico dei sistemi di IA, l'IA viene descritta come un insieme di *"sistemi che hanno la capacità di elaborare dati e informazioni in un modo che ricorda il comportamento intelligente e*

tipicamente comprende aspetti di ragionamento, apprendimento, percezione, previsione, pianificazione, controllo".

Nel complesso, le diverse definizioni convergono su un punto centrale: l'intelligenza artificiale non è una singola tecnologia, ma costituisce un campo multidisciplinare che si colloca all'intersezione tra informatica e discipline che studiano l'intelligenza umana, come psicologia e neuroscienze; l'obiettivo della disciplina è quello di emulare, attraverso le macchine, facoltà proprie dell'intelligenza umana.

Tra i pionieri nel campo, Marvin Minsky, fondatore storico della disciplina, ha discusso la complessità della questione in due opere di riferimento: *Computation: Finite and Infinite Machines* (1967), in cui vengono tracciati i fondamenti matematici sulla scia della teoria della computabilità, e *The Society of Mind* (1986), in cui propone una teoria della mente come società di agenti specializzati, che cooperano per produrre comportamenti intelligenti. A Minsky si deve anche una delle intuizioni più feconde sulla difficoltà di costruire macchine intelligenti, condensata nella formula "easy things are hard": i compiti che gli essere umani svolgono spontaneamente e senza sforzo cosciente (compiti quotidiani legati alla percezione e alla mobilità) risultano fra i più complessi da realizzare in un sistema artificiale, mentre operazioni che per gli esseri umani risultano più ostiche, per le quali sono necessari tempo, addestramento e concentrazione (risolvere problemi matematici complessi o giocare a scacchi) si sono rivelate computazionalmente più trattabili. L'osservazione, nota anche come paradosso di Moravec, sottolinea che la difficoltà di un compito per un sistema artificiale è indipendente dalla difficoltà che esso pone all'intelligenza umana e suggerisce che molte intuizioni sull'intelligenza siano fondate sul tipo di compiti che riconosciamo come intellettuali, anziché sulla complessità effettiva dei processi cognitivi che essi implicano. Nella sintesi di Moravec (1988), "siamo tutti prodigiosi campioni olimpici nelle aree percettive e motorie, così bravi che facciamo sembrare facile ciò che è difficile"[1] (Moravec, 1988, cit. in Mitchell, 2021; trad. mia); il ragionamento astratto, invece, è "lo strato più sottile del pensiero umano"[2] (ibid.; trad. mia), reso possibile da un'intelligenza sensomotoria molto più antica e potente, ma inconscia.

[1] "We are all prodigious olympians in perceptual and motor areas, so good that we make the difficult look easy." (Moravec, 1988, cit. in Mitchell, 2021).

[2] "The deliberate process we call reasoning is, I believe, the thinnest veneer of human thought [...]" (Moravec, 1988, cit. in Mitchell, 2021).

Storicamente, la percezione del fenomeno IA ha conosciuto fasi di entusiasmo (*AI springs*) e fasi di disillusione (*AI winters*), soprattutto a causa di quattro fallacie ricorrenti nel dibattito sulla disciplina (Mitchell, 2021). Tra queste, la convinzione che i progressi su compiti specifici si collochino su un continuum con l'intelligenza generale (*first-step fallacy*); l'idea – già contraddetta da Minsky (1986) e Moravec (1988) – che ciò che è facile per gli esseri umani sia facile anche per le macchine e viceversa ("*easy things are easy and hard things are hard*"); il ricorso a mnemoniche linguistiche fuorvianti che antropomorfizzano le capacità dei sistemi[3]; l'assunzione che l'intelligenza possa essere "disincarnata", ricondotta unicamente al cervello, ignorando il ruolo – fondamentale per l'intelligenza reale – del corpo, del contesto e dell'interazione nei processi cognitivi[4]. Per Mitchell (2021), queste fallacie alimentano aspettative non corrispondenti alle capacità effettive dei sistemi di IA e producono cicli di sopravvalutazione e disillusione difficili da interrompere senza una riflessione più rigorosa sulla natura dell'intelligenza stessa, che rendono complesso valutare il reale progresso dell'IA.

Per comprendere la natura dei moderni sistemi di IA, è necessario distinguere tra i due principali paradigmi che hanno dominato la disciplina. Seguendo la distinzione proposta da Riguzzi (2021) e ripresa in studi successivi (Proietti, 2024), l'IA si è sviluppata seguendo un approccio simbolico (noto come GOFAI, *Good Old-Fashioned Artificial Intelligence*), prevalente fino agli anni '80, basato sulla manipolazione di simboli e regole logiche esplicite comprensibili per l'uomo. In questo modello, il sistema opera seguendo uno schema rigido: percezione e acquisizione dei dati, elaborazione tramite un motore di inferenza (*inference engine*), istruzioni da cui derivano i processi di ragionamento, e azione finale. Tali sistemi sono definiti "trasparenti" o *glass box*, poiché il percorso logico che conduce a un risultato è interamente tracciabile e spiegabile.

A questo paradigma si contrappone l'approccio sub-simbolico, che si ispira al funzionamento biologico del cervello umano e non rappresenta la conoscenza in forma linguistica o logica diretta. È in questo ambito che si collocano le reti neurali e il Machine Learning, in cui il sistema non riceve regole esplicite, ma apprende autonomamente regolarità a partire dai dati. Con l'incremento della potenza di calcolo e l'aumento degli strati delle reti neurali, si è giunti allo sviluppo del Deep Learning, il motore tecnologico che ha permesso i recenti successi

[3] Termini come comprendere, imparare, pensare usati in senso analogico, ma percepiti come letterali.

[4] In dialogo con la tradizione dell'*embodied cognition* (cognizione incarnata), secondo cui le facoltà cognitive umane sono radicate nell'esperienza corporea e ambientale e non sono riducibili a operazioni puramente simboliche, eseguite dal solo cervello (Lakoff e Johnson, 1999; Varela et al., 1991).

nella visione artificiale e nel trattamento del linguaggio naturale (NLP). L'approccio sub-simbolico, vantaggioso rispetto al precedente, manca però di uno dei principali punti di forza dei sistemi simbolici: la comprensibilità. A differenza dei sistemi simbolici, i modelli sub-simbolici sono descritti come sistemi a "scatola nera" (*black box)*. Il loro processo decisionale, infatti, non è direttamente accessibile né facilmente ricostruibile da un osservatore umano. Come conseguenza diretta, anche le decisioni prodotte risultano difficili da interpretare e da spiegare in modo trasparente.

È proprio in questo solco che si inserisce lo sviluppo dei Large Language Model (LLM), una categoria di modelli di Deep Learning che rappresenta il punto d'arrivo di anni di progressi nell'elaborazione del linguaggio naturale (NLP) e consente ai computer, attraverso il Machine Learning, di riconoscere, comprendere e generare testo e parlato (Ježek e Sprugnoli, 2023). Una definizione più articolata dei LLM sarà proposta nel paragrafo successivo, dedicato alle loro caratteristiche e al loro funzionamento.

2.1.1 I Large Language Model

Il linguaggio, inteso come capacità distintiva degli esseri umani di comunicare e di esprimersi, si configura come un elemento profondamente radicato nell'esperienza umana, tanto da rendere difficilmente concepibile un'esistenza priva di esso. Gli strumenti di intelligenza artificiale oggi più comuni, che occupano il dibattito pubblico e si sono diffusi capillarmente nella vita quotidiana di un'utenza generalista sempre più ampia, sono quelli che permettono alle macchine di comunicare proprio attraverso l'emulazione del linguaggio umano.

Fin dal 1950, a partire dal Test di Turing, proposto originariamente come "gioco dell'imitazione", il linguaggio è stato inteso come misura "dell'intelligenza" delle macchine: Turing sosteneva, infatti, che la capacità di comunicare in modo indistinguibile da un essere umano fosse un criterio sufficiente per definire una macchina "intelligente" (Turing, 1950).

Oggi l'esplorazione dell'intelligenza linguistica delle macchine ha raggiunto livelli senza precedenti grazie al language modeling (LM), in particolare attraverso i *Large Language Models (LLM)*. Il language modeling è un approccio che mira a modellare la distribuzione di probabilità delle sequenze linguistiche, permettendo così di prevedere i token[5] successivi o

[5] Unità minime di testo elaborate dai modelli linguistici.

mancanti all'interno di un testo. I language model, quindi, risolvono compiti di *word prediction*: a partire da una sequenza di parole, il modello predice la parola mancante o successiva per cui la probabilità condizionata ha il valore più alto (Lenci et al., 2025).
Seguendo Lenci et al. (2025), sul piano formale, la probabilità di una sequenza linguistica viene scomposta attraverso la *chain rule* come prodotto delle probabilità condizionate di ciascuna parola, dato il contesto delle parole precedenti: la probabilità di una frase di n parole è il prodotto delle probabilità di ogni parola condizionata a tutte quelle che la precedono. Calcolare la probabilità di una parola dato l'intero contesto pregresso è complesso, poiché la maggior parte delle sequenze lunghe occorre raramente o addirittura non occorre mai nei dati di addestramento, anche in corpora di grandi dimensioni. Per ovviare al problema, sono stati creati i modelli di Markov, *language model* che calcolano la probabilità di una parola solo su un numero limitato di parole precedenti; questi modelli sono detti *n-gram model* (modelli a n-grammi). In una catena di Markov di ordine n, ogni evento dipende dalla sequenza di n eventi che lo precedono, quindi le probabilità del modello vengono stimate a partire dalle frequenze di n+1 grammi. Nonostante le catene di Markov rendano più semplice il calcolo della probabilità di una sequenza di parole, non risolvono il problema della sparsità dei dati linguistici: molte sequenze plausibili non compaiono mai nel corpus di addestramento, per questo, come conseguenza, il modello assegnerebbe loro probabilità nulla. Per ovviare a questo problema si ricorre a tecniche di *smoothing* (ad. es. lo *smoothing* di Laplace), che redistribuiscono parte della probabilità dagli n-grammi osservati a quelli non osservati, facendo in modo che ogni sequenza linguistica riceva una probabilità non nulla. Dopo essere stati addestrati su un insieme di dati di addestramento, detto *training set*, la performance dei *language model* viene testata e valutata su altri dati, diversi da quelli di addestramento, che costituiscono il *test set*. *Training set* e *test set* appartengono spesso allo stesso corpus e condividono spesso la stessa distribuzione di parole, ma è importante che rimangano distinti per garantire la valutazione della capacità del modello di generalizzare su nuovi dati quanto appreso in fase di addestramento.

I LLM si collocano come stadio più recente nell'evoluzione del LM, iniziata con i modelli statistici (SLM) negli anni '90, passata poi attraverso modelli neurali (NLM), introdotti a partire dal 2013, fino ai modelli linguistici pre-addestrati (PLM), affermatisi intorno al 2018. I LLM, punto d'arrivo di questa evoluzione, sono modelli linguistici pre-addestrati (PLM) di dimensioni significative, caratterizzati da un numero eccezionale di parametri, dell'ordine delle decine o delle centinaia di miliardi. L'etichetta stessa di "large

language model" è stata coniata proprio per distinguere questi modelli di grandi dimensioni dai PLM più piccoli (come BERT) (Zhao et al., 2023).

Il fondamento teorico che sta alla base del funzionamento dei LLM è l'ipotesi distribuzionale, teoria linguistica formulata negli anni '50 dal linguista americano Z. Harris (1954) e, parallelamente, in Inghilterra da J.R. Firth (1957). Secondo tale teoria, l'insieme dei contesti in cui una parola occorre (ovvero la sua distribuzione) ne rivela il significato: di conseguenza, termini che compaiono in contesti simili tendono ad avere significati simili.

Prima dell'avvento dell'architettura *Transformer*, l'elaborazione neurale del linguaggio si basava prevalentemente su reti ricorrenti (*Recurrent Neural Networks*, RNN) e sulle loro varianti, in particolare le reti *long-short memory* (LSTM), introdotte da Hochreiter e Schmidhuber (1997). Tali architetture processavano il testo in modo sequenziale, parola dopo parola. Ciò rendeva difficile catturare le dipendenze a lungo raggio tra elementi distanti del testo e impediva la parallelizzazione del calcolo, limitando l'applicabilità di un sistema di questo tipo su corpora di grandi dimensioni (Lenci et al, 2025)

Dal punto di vista architetturale, i LLM contemporanei sono reti neurali profonde basate, nella maggior parte dei casi, sull'architettura Transformer, che si basa a sua volta su meccanismi di attenzione, tecniche di *Machine Learning* che consentono al modello di calcolare i pesi di attenzione, che esprimono l'importanza relativa delle diverse componenti in una sequenza di input, senza doverli elaborare in modo sequenziale (Vaswani et al., 2017). I pesi sono calcolati a partire dal prodotto scalare (opportunamente normalizzato) tra due proiezioni vettoriali di ciascun token, denominate rispettivamente *query* e *key*. Una terza proiezione, il *value*, contiene l'informazione che viene effettivamente combinata sulla base dei pesi così ottenuti. Tali pesi di attenzione vengono poi utilizzati per modulare il contributo di ciascun elemento dell'input, amplificandone o attenuandone l'influenza[6]. Il meccanismo viene applicato in parallelo attraverso più "teste" di attenzione (*multi-head attention*), ciascuna delle quali apprende a focalizzarsi su tipi diversi di relazione tra *token* (sintattica, semantica, di lungo raggio); gli output delle singole teste vengono poi combinati per produrre la rappresentazione finale di ciascun elemento (Lenci et al., 2025). Dato che il meccanismo di attenzione, di per sé, non incorpora informazioni sull'ordine sequenziale dei *token* in input, il Transformer integra la posizione di ciascun token attraverso un sistema di codifica posizionale (*positional encoding*), che aggiunge agli *embedding* un'informazione sulla loro collocazione sequenziale, permettendo al modello di cogliere l'ordine delle parole nel contesto (Lenci et

[6] https://www.ibm.com/it-it/think/topics/attention-mechanism#774698768

al., 2025). A differenza dei modelli sequenziali, attraverso il meccanismo di *self-attention* descritto da Vaswani et al. (2017), il modello cattura dipendenze globali tra *token*, anche distanti tra loro, disambiguando il significato di un termine sulla base dell'intero contesto globale del testo. L'architettura Transformer rende inoltre possibile anche la parallelizzazione del calcolo, ottimizzando il processo di addestramento rispetto ai metodi precedenti e consentendo l'uso di corpora di grandi dimensioni (Lenci et al., 2025).

I LLM contemporanei vengono addestrati su corpora di enormi dimensioni, suddivisi preliminarmente in token attraverso il processo di tokenizzazione. Questo processo nei LLM non corrisponde alla segmentazione in parole, ma si tratta di una suddivisione in sotto-unità (*sub-word tokens*) ottenuti attraverso algoritmi come *Byte Pair Encoding* (BPE, usato dai modelli della famiglia GPT) o *WordPiece* (utilizzato dai modelli con architettura BERT), che bilanciano la dimensione del vocabolario e la copertura della varietà morfologica delle lingue di addestramento. Una stessa parola può essere dunque rappresentata da uno o più *token* a seconda della sua frequenza nel corpus: le parole più frequenti vengono trattate come *token* unitari e associati a singoli *embedding,* le parole meno frequenti sono invece scomposte in sotto-parole (Lenci et al., 2025). Una volta suddiviso il testo in *token*, ogni *token* viene rappresentato come vettore numerico, noto come *embedding*. Gli *embedding* modellano i *token* come vettori in uno spazio multidimensionale nel quale la vicinanza tra vettori riflette relazioni di somiglianza semantica e, in parte, anche regolarità sintattiche, consentendo al modello di catturare proprietà linguistiche più complesse rispetto agli approcci precedenti (Ježek e Sprugnoli, 2023). A differenza degli *embedding* statici (*static embedding*) prodotti da modelli precedenti come *word2vec* (Mikolov et al., 2013), in cui a ogni parola corrisponde una sola rappresentazione vettoriale a prescindere dal contesto in cui occorre, i modelli linguistici basati su Transformer producono rappresentazioni contestuali (*contextual embedding*). In questi modelli, infatti, la rappresentazione di un *token* è calcolata tenendo conto degli altri *token* della sequenza attraverso i meccanismi di attenzione; di conseguenza lo stesso *token* assume rappresentazioni vettoriali differenti a seconda della frase in cui appare, cambiando dinamicamente a seconda del contesto e permettendo al modello di disambiguare significati polisemici. In questo modo, i LLM sono in grado di produrre sequenze testuali coerenti e contestualmente appropriate, basandosi su una modellizzazione probabilistica delle relazioni tra le unità linguistiche. Il pre-training dei LLM avviene mediante addestramento auto-supervisionato (*self-supervised learning*), una tecnica di *machine learning* che non

richiede annotazione manuale esplicita[7], a metà tra apprendimento supervisionato e non supervisionato: l'apprendimento avviene direttamente dal testo attraverso compiti di predizione su sequenze corrotte o incomplete fornite in input, che il modello ricostruisce correttamente. L'architettura Transformer originale (Vaswani et al., 2017) prevede due moduli principali: un *encoder*, che codifica la sequenza di input in rappresentazioni vettoriali contestualizzate, e un *decoder* che genera la sequenza di *output* un *token* alla volta. I modelli *decoder* utilizzano solo il secondo modulo, gli *encoder* solo il primo, le architetture *encoder-decoder* entrambi. Nei modelli *decoder*, il cui esempio principale è costituito dalla famiglia GPT (Radford et al., 2019; Brown et al., 2020), l'addestramento avviene su un compito di *language modeling* unidirezionale: l'obiettivo è la predizione del *token* successivo calcolando la sua probabilità data l'intera sequenza di *token* precedenti; in questo modo questi modelli, detti per questo generativi, generano testi. Nei modelli *encoder*, come BERT (Devlin et al., 2019), l'obiettivo è la predizione di *token* mascherati attraverso il *masked language modeling* (MLM): porzioni della sequenza vengono nascoste e il modello impara a ricostruirle a partire dal contesto circostante. Esistono, inoltre, architetture *encoder-decoder* (come BART o T5) pensate per compiti di trasformazione da una sequenza a un'altra, particolarmente adatte alla traduzione automatica. I LLM – e i chatbot più diffusi ad essi connessi, come ChatGPT o Gemini – cuore dell'IA generativa, appartengono alla famiglia dei *decoder*. Al *pre-training* si affiancano poi fasi successive di adattamento, dette *post-training* (Tie et al., 2025): l'*instruction fine-tuning* o *fine-tuning* supervisionato e *Reinforcement Learning from Human Feedback* – RLHF, che modellano il comportamento conversazionale del sistema e ne calibrano gli output secondo criteri di utilità, sicurezza e tono adeguato (Ouyang et al., 2022). L'*instruction fine-tuning* è una tecnica attraverso cui il modello impara a rispondere a richieste specifiche e a eseguire compiti richiesti in maniera mirata, attraverso l'osservazione di coppie istruzioni-output sulla base della conoscenza assunta in fase di pre-training. Durante la fase di Reinforcement Learning from Human Feedback, le risposte del modello vengono valutate da annotatori umani per addestrare un *reward model* (modello di ricompensa) che guida l'ulteriore ottimizzazione del modello del sistema (Lenci et al., 2025). È in questa fase di allineamento che il modello apprende a rispondere in modo cortese, a rifiutare richieste considerate problematiche e a modulare il proprio registro comunicativo. Il RLHF introduce, però, esso stesso distorsioni sistematiche: gli annotatori tendono a premiare risposte cortesi e accomodanti, per cui il modello apprende a privilegiare la

[7] https://www.ibm.com/it-it/think/topics/large-language-models

soddisfazione percepita dell'utente rispetto alla correttezza fattuale, fenomeno noto come *sycophancy* (cfr. Sezione 2.2.6). Successivamente, l'output linguistico è generato applicando la conoscenza appresa durante la fase di addestramento. La generazione concreta degli output, inoltre, dipende da strategie di decodifica diverse, a partire dalle quali il modello seleziona il *token* successivo da generare dopo la sequenza input. La strategia di decodifica più comune è la *greedy decoding*, che consiste nel selezionare a ogni passo *token* con la probabilità più alta. Strategie di decodifica più articolate, come quelle di campionamento (*top-k sampling* o il *top-p sampling)*, introducono un grado controllato di variabilità nelle risposte: con il campionamento il modello estrae il *token* successivo in maniera casuale, tra un campione di *token* candidati. In questo modo, una stessa domanda riceve risposte non identiche, con conseguenze rilevanti sull'esperienza dell'utente.

Proprio a causa della natura matematica e vettoriale, il linguaggio prodotto in output da un LLM non è, quindi, il risultato di un atto comunicativo fondato su intenzioni e significati, ma l'esito di un calcolo probabilistico effettuato su rappresentazioni vettoriali di unità linguistiche.

A differenza dei modelli precedenti, inoltre, i LLM non si limitano alla modellazione o generazione del linguaggio, ma mostrano capacità avanzate di risoluzione di compiti generali, grazie alla loro scala e alle proprietà emergenti che ne derivano (Wei et al., 2022). Le abilità emergenti sono capacità di eseguire compiti senza uno specifico addestramento, non presenti nei modelli di scala ridotta, ma che appaiono quando il modello supera una certa soglia dimensionale (spesso indicata sopra i 10 miliardi di parametri), in correlazione anche alla qualità e alla quantità dei dati su cui è addestrato. Queste capacità sono connesse al cosiddetto *in-context learning,* ovvero la capacità di svolgere nuovi compiti sulla base di istruzioni o di pochi esempi forniti nel prompt, senza ulteriore addestramento; a seconda che il prompt contenga o meno esempi del compito, si distingue tra *few-shot* e *zero-shot in-context learning*. Quest'ultima modalità, in cui il modello esegue il compito sulla base della sola istruzione, è la più comune nelle interazioni con i chatbot generativi accessibili al pubblico. Tra le capacità emergenti, figurano anche forme di ragionamento logico apparente, che imitano il comportamento di un agente razionale, pur restando esiti di un calcolo probabilistico (Zhao et al., 2023). La nozione stessa di "proprietà emergente", comunque, è stata messa in discussione: Schaeffer et al. (2023) sostengono che molte capacità descritte come emergenti siano in realtà artefatti delle metriche di valutazione discrete adottate dalla letteratura, e che a metriche continue il miglioramento risulti più graduale e prevedibile.

I LLM sono accessibili al pubblico tramite interfacce conversazionali, chatbot open-domain, come ChatGPT di OpenAI oppure Gemini di Google, che simulano la conversazione umana con un utente finale.

Comprendere il funzionamento dei LLM, e di conseguenza dei chatbot che su questi si basano, è cruciale per evitare un fraintendimento che ricorre sia nell'uso quotidiano sia in parte del dibattito pubblico: trattare la fluidità sintattica e lessicale degli output come prova di una reale comprensione del significato da parte del modello. In termini linguistici, un LLM è un potente generatore di sequenze plausibili, addestrato a ottimizzare la coerenza locale e globale del testo; non è però un soggetto che intende, sa o crede qualcosa. In altri termini, non è un soggetto "intelligente" in senso umano. Quando un utente pone una domanda e riceve una risposta ben formata, la scorrevolezza del testo può indurre a proiettare sul sistema capacità cognitive che esso non possiede. Questa proiezione ha una lunga storia nelle interazioni uomo-macchina: già con ELIZA, il programma sviluppato da Joseph Weizenbaum (1966), era stata osservata la tendenza degli utenti ad attribuire comprensione e partecipazione emotiva a un sistema artificiale, che si limitava a riformulare in domanda gli enunciati ricevuti[8]. Nel complesso, l'interazione con i sistemi di IA espone gli utenti, soprattutto i non esperti, a vari fattori di rischio, che possono riguardare la possibile produzione di informazioni inesatte (allucinazioni), la presenza di bias nei dati e, di conseguenza, negli output, la tendenza degli utenti a sviluppare una fiducia eccessiva nei sistemi e le implicazioni legate alla gestione dei dati personali.

2.2 Rischi nell'interazione con i sistemi di IA generativa

La fluidità degli output prodotti dai LLM, coerenti nel tono e nella sintassi e tematicamente pertinenti, costituisce, paradossalmente, una delle loro caratteristiche più rischiose. Come precedentemente illustrato, questa fluidità è il risultato di un calcolo probabilistico su rappresentazioni vettoriali, non di un processo di comprensione fondato su intenzioni o conoscenze verificate. Ne consegue che un sistema di IA generativa può produrre testo stilisticamente corretto e convincente anche quando il contenuto è inesatto, distorto o privo di fondamento. La questione, più che l'affidabilità tecnica di questi strumenti, riguarda la ricezione degli output da parte degli utenti: un ruolo cruciale è giocato dalla loro valutazione critica, fondamentale per evitare una fiducia indebita ed eccessiva nella sola qualità espressiva dell'output. Quattrociocchi et al. (2026), segnalano come la radice teorica di queste fragilità

[8] Il cosiddetto "effetto Eliza" (Glesaz, 2023).

risieda in una confusione concettuale ricorrente nel dibattito pubblico: l'identificazione – e quindi la confusione – tra prestazioni su singoli task valutate tramite benchmark e intelligenza generale. I LLM, infatti, dimostrano prestazioni elevate su task specifici, e fluidità trasversale a domini diversi, ma il successo a livello del singolo task, di per sé, non è prova di una intelligenza generalizzata. L'apparente fluidità degli output, inoltre, può essere scambiata per evidenza di comprensione, generando aspettative non corrispondenti alle effettive capacità del sistema.

2.2.1 Allucinazioni

Il rischio epistemico più documentato è quello delle cosiddette allucinazioni (*hallucination*), fenomeno per cui un LLM genera informazioni plausibili e corrette nella forma, ma imprecise, non fedeli o prive di senso. Il termine, mutuato per analogia dalla psicologia clinica, fa originariamente riferimento a una percezione non reale che sembra, però, del tutto reale (Ji et al., 2023). Maleki, Padmanabhan e Dutta (2024) hanno condotto una revisione sistematica della letteratura su quattordici database accademici che ha portato, oltre che alla mappatura di 333 definizioni del fenomeno, anche all'identificazione di termini alternativi proposti per designarlo senza ricorrere alla metafora clinica. L'uso dell'etichetta "allucinazione" per indicare la generazione di materiale non veritiero da parte dei LLM e la conseguente metafora clinica, infatti, risultano impropri innanzitutto a causa della "umanizzazione" che ne sottende: i sistemi di IA, infatti, non possiedono percezioni sensoriali e i loro errori non emergono dall'assenza di stimoli, come nelle allucinazioni psicologiche, bensì dall'architettura statistica del modello e dai dati di addestramento. Inoltre, la metafora rafforza lo stigma legato alle malattie mentali, associando problematiche negative negli output generati dall'IA a specifiche condizioni cliniche psichiatriche (Østergaard e Nielbo, 2023). Tra le alternative proposte, in letteratura figurano "*confabulation*", "*stochastic parroting*", "*fabrication*" e "*falsification*". Sotto l'etichetta generica di "allucinazione", se ne distinguono due categorie: quelle intrinseche e quelle estrinseche. Nelle prime l'output generato contraddice il contenuto della fonte; nelle seconde, invece, il contenuto dell'output è indipendente dalla fonte, che non può contraddirlo né supportarlo (Ji et al., 2023).

Le allucinazioni nei modelli di generazione del linguaggio naturale possono dipendere da diversi fattori, riconducibili principalmente a tre livelli: i dati di addestramento, l'architettura e le dinamiche di funzionamento dei modelli e i processi di generazione dell'output. In primo luogo, i dataset utilizzati per l'addestramento possono presentare discrepanze o incoerenze tra

la fonte (input) e il riferimento (target), dovute sia a errori nella raccolta automatizzata dei dati (dove il testo di riferimento contiene informazioni extra non presenti nella fonte) sia alla natura stessa di alcuni compiti, come il *chit-chat*, che incoraggiano il modello a essere creativo a discapito della fedeltà ai fatti. Il modello, inoltre, potrebbe non comprendere correttamente le relazioni tra le parti dell'input, creando correlazioni errate (Ji et al., 2023). Infine, i processi di generazione, in particolare tecniche utilizzate per rendere il linguaggio più vario (come il campionamento top-k; Holtzman et al., 2019) introducono una casualità che aumenta la probabilità di generare contenuti non verificati.

Va poi considerato che i modelli di grandi dimensioni tendono a incorporare enormi quantità di informazioni durante il pre-addestramento. Quando generano un testo, quindi, possono dare priorità a questa "conoscenza interna" (parametrica) rispetto ai dati specifici forniti nell'input, creando come conseguenza informazioni aggiuntive o contraddittorie (Ji et al., 2023).

In generale, i modelli più grandi, pur mostrando capacità prestazionali elevate, possono presentare maggiori criticità in termini di controllo e veridicità degli output, risultando meno affidabili e veritieri. L'aumento della scala, infatti, rappresenta un'arma a doppio taglio: da un lato migliora la capacità generativa, dall'altro può amplificare la difficoltà nel garantire accuratezza e affidabilità.

Le conseguenze di questo fenomeno sono rilevanti sul piano sociale: output allucinati, infatti, possono indurre gli utenti a utilizzare informazioni false in modo inconsapevole, generando disinformazione accidentale (Lin et al., 2022). Allo stesso tempo, questi output possono essere usati in modo improprio anche deliberatamente, con scopi malevoli, per generare contenuti plausibili ma falsi e ingannevoli (Tamkin et al., 2021). Ciò rappresenta un ostacolo significativo all'applicazione dei LLM in ambiti ad alta criticità, come quello medico o giuridico, in cui l'accuratezza e l'affidabilità delle informazioni sono requisiti imprescindibili (Lin et al., 2022). Una rilettura più ampia del fenomeno è stata recentemente proposta dal gruppo di ricerca di Walter Quattrociocchi (Università degli studi di Roma "La Sapienza"), che configura le allucinazioni come un caso specifico di un problema epistemico più generale. Loru et al. (2025a), in uno studio pubblicato su PNAS, confrontano sei LLM con i sistemi di rating esperti NewsGuard e Media Bias/Fact Check e con valutazioni umane raccolte attraverso un esperimento controllato, sottoponendo modelli e partecipanti non esperti alla stessa procedura strutturata di selezione dei criteri, recupero dei contenuti e produzione di giustificazioni. Pur osservando un sostanziale allineamento dell'output finale, gli autori rilevano differenze sistematiche nei criteri osservabili che guidano le valutazioni: i modelli si appoggiano in misura prevalente ad associazioni lessicali e priors statistici, anziché

a processi di ragionamento contestuale, mostrando inoltre una sistematica asimmetria di orientamento politico nei giudizi di affidabilità – le fonti di area conservatrice vengono associate in misura sproporzionata a descrittori di inaffidabilità – e una tendenza a confondere la forma linguistica con l'affidabilità epistemica. Gli autori introducono il concetto di *epistemia* per descrivere questa illusione di conoscenza che emerge quando la plausibilità superficiale si sostituisce alla verifica: una condizione in cui un testo appare informativo, accurato e ben argomentato senza che vi sia, dietro la sua produzione, alcun processo di valutazione normativa. In questo contesto, le allucinazioni non sono "errori", ma il risultato strutturale di un sistema che massimizza la probabilità del *token* successivo, senza vincoli di verità o riferimento al mondo reale. Quattrociocchi et al. (2025) sviluppano teoricamente questa nozione, descrivendo i LLM non come agenti epistemici ma come sistemi di completamento stocastico di pattern, formalmente assimilabili a percorsi su grafi multidimensionali di transizioni linguistiche piuttosto che a sistemi che formano credenze o modelli del mondo. Emerge, dunque, uno scollamento tra la competenza linguistica dei LLM, estremamente alta, e il loro controllo epistemico, totalmente assente. Gli autori individuano sette "linee di faglia" epistemologiche tra cognizione umana e artificiale, riguardanti il *grounding* (l'ancoraggio alla realtà extralinguistica), il *parsing* semantico, l'esperienza, la motivazione, il ragionamento causale, la metacognizione e la valutazione di valore. L'epistemia rappresenta, in questa prospettiva, una condizione strutturale del rapporto contemporaneo tra utenti e sistemi generativi: la sensazione di sapere senza il lavoro epistemico del giudicare, con conseguenze rilevanti sul piano della valutazione, della governance e dell'alfabetizzazione critica[9]. La distinzione conserva una rilevanza specificamente linguistica, in quanto chiarisce che la fluidità di un output è una proprietà superficiale del testo, non un indicatore della sua fondatezza.

La divergenza tra produzione umana e produzione generativa lascia, inoltre, tracce statistiche misurabili. Su un primo livello, quello dell'articolazione della conoscenza pubblica, Hadad et al. (2026a) propongono un confronto sistematico tra Wikipedia e Grokipedia[10] e documentano una forma di mediazione selettiva: le pagine più popolari vengono riprodotte testualmente, mentre quelle controverse o ad alta densità di riferimenti vengono riscritte, con spostamenti localizzati nel tono valutativo. Anche quando la struttura narrativa rimane invariata, la

[9] Quattrociocchi et al. (2025) fanno esplicito riferimento alla "epistemic literacy" come una competenza distinta dal pensiero critico tradizionale, focalizzata sulla capacità di riconoscere quando un output è frutto di completamento statistico invece che di una valutazione dei fatti.

[10] L'enciclopedia generativa associate al modello Grok.

mediazione generativa altera selettivamente il modo in cui la conoscenza pubblica viene presentata. Su un piano di natura più strettamente statistica, Hadad et al. (2026b), in uno studio fondato sull'analisi della comprimibilità dei testi, mostrano che il linguaggio prodotto dai LLM si presenta strutturalmente più regolare rispetto a quello prodotto da esseri umani, coerentemente con una concentrazione degli output entro pattern statistici ricorrenti. Questa "firma statistica" è osservabile direttamente sul testo di superficie, senza accesso ai meccanismi interni del modello, e suggerisce che i sistemi generativi tendano a riconfigurare la produzione testuale verso forme più prevedibili e meno variabili rispetto a quella umana. Il dato, sul piano linguistico, segnala una possibile riduzione della varietà espressiva degli ambienti informativi mediati da IA generativa, con conseguenze che si estendono oltre il singolo output e investono la diversità lessicale e sintattica della comunicazione pubblica.

2.2.2 Bias

Un secondo ordine di rischi riguarda i bias, pregiudizi sistematici incorporati nei modelli durante la fase di addestramento[11]. L'addestramento dei LLM avviene su corpora di testo di enormi dimensioni, estratti prevalentemente dal web, dai quali i modelli apprendono regolarità statistiche poi riprodotte negli output. Tali dati riflettono, inevitabilmente, le asimmetrie, gli stereotipi e le distorsioni umane che il modello apprende e ripropone. La letteratura recente, sintetizzata in una rassegna sistematica condotta da Gallegos et al. (2024), ha articolato il fenomeno lungo tre dimensioni principali – metriche di valutazione, dataset di benchmark e tecniche di mitigazione – e distingue tra bias osservabili a livello di embedding, di probabilità di output e di testo generato. Gli output dei LLM, infatti, possono proporre come neutre associazioni stereotipate tra genere e professione, rappresentazioni distorte di minoranze etniche o culturali, squilibri nella copertura di determinate lingue o prospettive geografiche. La composizione demografica degli utenti di internet, infatti – sovrarappresentativa di popolazioni giovani, maschili, anglofone e di paesi occidentali ad alto reddito – si riflette strutturalmente nei corpora di addestramento: piattaforme come Reddit e Common Crawl, utilizzate per addestrare modelli di grande scala, rispecchiano non la diversità dell'esperienza umana, ma una visione egemonica di chi produce contenuti digitali accessibili (Bender et al., 2021). Se è vero, infatti, che i LLM agiscono come "*stochastic parrots*" (Bender et al., 2021), limitandosi a riprodurre pattern linguistici appresi dai dati di addestramento, è anche vero che tali pattern riflettono visioni del mondo e credenze umane

[11] https://www.ibm.com/it-it/think/topics/ai-bias

sedimentate nei dati. In questo senso, l'astrazione di un'umanità monolitica risulta riduttiva e non riflette la pluralità di prospettive e culture che caratterizzano i contesti reali (Atari et al., 2023). In particolare, gli output dei LLM su temi legati a moralità, etica e politica riproducono e simulano la mentalità e la psicologia delle popolazioni cosiddette WEIRD (Henrich et al., 2010), ovvero quelle occidentali (*Western*), scolarizzate (*Educated*), industrializzate (*Industrialized*), ricche (*Rich*) e democratiche (*Democratic*), escludendo o marginalizzando, di fatto, prospettive non occidentali, non industrializzate o non anglofone. Secondo l'ITU (*International Telecommunication Union*), nel 2025[12], una quota rilevante della popolazione mondiale, circa un quarto, infatti, risulta offline e l'accesso ad internet continua ad essere strettamente legato allo sviluppo economico: i paesi ad alto reddito si avvicinano a un uso quasi universale[13] di internet, che resta una prospettiva lontana nei paesi a basso reddito[14]. Tra quanti vi accedono, solo una minoranza produce contenuti nelle lingue dominanti dei corpora di pre-training; numerose lingue, dunque, non sono rappresentate in modo proporzionato alla numerosità delle popolazioni che le parlano. Gli output di un LLM in italiano, o in qualsiasi altra lingua diversa dall'inglese, non riflettono soltanto la distribuzione interna della lingua bersaglio, ma anche l'orientamento culturale del materiale anglofono che domina la fase di addestramento, mediato dai meccanismi di trasferimento crosslinguistico, tipici dei modelli multilingue. Gli output dei LLM, dunque, riflettono lo sbilanciamento culturale dei corpora di addestramento (Bender et al., 2021).

Caliskan et al. (2017), in uno studio che fornisce una delle prime dimostrazioni quantitative del modo in cui i bias umani vengono ereditati dalle rappresentazioni linguistiche apprese dai modelli, mostrano che tali rappresentazioni vettoriali replicano sistematicamente le associazioni stereotipiche documentate nella letteratura psicologica: nomi propri statisticamente associati a popolazioni di origine europeo-americana risultano più vicini al lessico della piacevolezza rispetto ai nomi associati a popolazioni afroamericane; termini relativi alle scienze e alla matematica si collocano più prossimi al lessico maschile, mentre quelli relativi alle arti risultano più prossimi a quello femminile; associazioni stereotipate simili si osservano sull'asse carriera/famiglia. Il quadro è stato recentemente confermato e ampliato da Bai et al. (2025): adattando la logica dell'*Implicit Association Test* al comportamento osservabile dei chatbot, gli autori introducono due misure psicologicamente

[12] Internet use - Statistics

[13] 94% della popolazione nei paesi ad alto reddito.

[14] 23% della popolazione nei paesi a basso reddito.

fondate (*LLM Word Association Test* e *LLM Relative Decision Test*) e documentano la persistenza di bias stereotipici pervasivi in otto modelli *value-aligned*, su quattro categorie sociali (razza, genere, religione, salute) e ventuno stereotipi specifici, anche quando gli stessi modelli risultano formalmente non-biased ai benchmark di valutazione standard. Il risultato suggerisce che le procedure di allineamento ai valori umani riducono la manifestazione esplicita dei bias ma non ne rimuovono la radice statistica. In una prospettiva applicata, Salinas et al. (2025) hanno condotto un audit sistematico di GPT-4 e di altri modelli commerciali in scenari decisionali di vita quotidiana, dalla negoziazione per l'acquisto di un'automobile alla previsione di esiti elettorali. Sottoponendo i modelli a quarantadue template di prompt che variavano solo per il nome del soggetto coinvolto, gli autori osservano che i nomi statisticamente associati a minoranze etniche e a donne ricevono consigli sistematicamente meno favorevoli, con i nomi associati a donne afroamericane che ottengono in media gli esiti peggiori. La presenza di ancore numeriche (come stime di valore esterne) nel prompt riesce a contrastare in parte le distorsioni; al contrario, dettagli qualitativi sulla persona possono in alcuni casi accentuarle. Lo studio sottolinea come i bias possano emergere in maniera occulta tramite associazioni implicite anche in modelli che, esplicitamente interrogati su categorie sensibili, rifiutano di rispondere o restituiscono output formalmente neutrali. Le procedure di debiasing (fine-tuning, prompt di sistema, filtri di sicurezza) mitigano alcuni effetti più visibili, senza però rimuovere lo sbilanciamento strutturale e, in alcuni casi, possono sostituire un bias con un altro, riducendo ulteriormente la diversità delle prospettive rappresentate. La difficoltà di un intervento generalizzabile sui bias emerge con particolare chiarezza nello studio di Ma et al. (2025), che applicano tecniche di *pruning* (rimozione selettiva di neuroni e di teste di attenzione) per ridurre i bias razziali in contesti specifici. I risultati mostrano che il *pruning* può ridurre i bias in modo efficace nel dominio in cui viene applicato, ma la mitigazione perde rapidamente efficacia quando si tenta di generalizzarla ad altri contesti: una strategia mirata a rimuovere il bias razziale nelle decisioni finanziarie risulta scarsamente efficace nelle decisioni commerciali. L'analisi degli autori suggerisce che i bias razziali siano solo parzialmente rappresentati nel modello come concetto generale, mentre la maggior parte di essi è codificata in modo *context-specific*, ossia in relazione allo specifico contesto applicativo in cui il modello è utilizzato: di conseguenza, strategie di mitigazione di tipo "*one-size-fits-all*" – concepite come soluzioni uniformi e generalizzabili indipendentemente dal contesto – risultano strutturalmente inadeguate, e la responsabilità giuridica per gli effetti discriminatori va in larga parte attribuita a chi impiega il modello in un determinato contesto applicativo, oltre che a chi lo sviluppa.

Accanto ai bias di natura sociale, esistono altre categorie di distorsioni sistematiche che caratterizzano il funzionamento dei LLM, di natura cognitiva e architetturale. Sul piano cognitivo, Knipper et al. (2025) hanno condotto una valutazione su larga scala di otto bias cognitivi ampiamente documentati nella letteratura psicologica (*anchoring, availability, confirmation, framing, interpretation, overattribution, prospect theory e representativeness*) su 45 LLM e oltre 2,8 milioni di risposte generate, riscontrando comportamenti coerenti con i bias attesi nel 17,8-57,3% dei casi. Modelli di dimensioni maggiori (oltre 32 miliardi di parametri) riducono la suscettibilità ai bias nel 39,5% dei casi, e una maggiore specificità del prompt riduce ulteriormente i bias fino al 14,9%, sebbene non in tutti i casi. Dal punto di vista architetturale, invece, è stata documentata l'esistenza di un bias di posizione strutturale nei transformer, per cui il modello tende a prestare maggiore attenzione a determinate regioni dell'input, in particolare alla sua parte iniziale: il fenomeno noto come "*lost in the middle*" comporta che le informazioni collocate al centro di un contesto lungo vengano elaborate con minore accuratezza rispetto a quelle posizionate all'inizio o alla fine (Wu et al., 2025). Si tratta di un bias non riconducibile ai dati di addestramento ma all'architettura stessa del meccanismo di attenzione e al *masking* causale, e ha conseguenze pratiche per l'utente non esperto: la posizione in cui un'informazione viene fornita al sistema può influenzare la qualità della risposta in modo non trasparente. Ashery et al. (2025) esplorano un ulteriore aspetto: in popolazioni di agenti basati su LLM messi in interazione tramite linguaggio naturale, possono emergere spontaneamente convenzioni sociali condivise, analogamente a quanto osservato nelle dinamiche di coordinazione tra esseri umani. Gli autori rilevano l'emergere di bias collettivi forti anche quando i singoli agenti, presi individualmente, non manifestano preferenze sistematiche: il bias non risiede nel singolo modello ma nasce dinamicamente nell'interazione tra modelli. L'esperimento mostra inoltre che minoranze "antagoniste" sufficientemente coordinate possono imporre convenzioni alternative al gruppo, con dinamiche che ricordano quelle studiate nei processi di dinamiche collettive nelle popolazioni umane. Il risultato apre una dimensione nuova del problema: in scenari in cui sistemi conversazionali interagiscono tra loro o con grandi platee di utenti, la formazione e la stabilizzazione di convenzioni linguistiche e culturali può sfuggire al controllo degli sviluppatori a livello dei singoli sistemi. Un problema analogo di controllo riguarda l'uso dei LLM come proxy di soggetti umani reali. In questo caso, il rischio riguarda la produzione di rappresentazioni amplificate e distorte dei comportamenti osservati. In un esperimento, a partire da un corpus di circa ventuno milioni di interazioni su X relative alle elezioni presidenziali statunitensi del 2024, è stata costruita una popolazione simulata di 1186 agenti

corrispondenti ad altrettanti utenti empirici. Confrontando le risposte generate da tre famiglie di modelli (Gemini, Mistral, DeepSeek) con quelle effettivamente prodotte dagli utenti umani, gli autori documentano un fenomeno che chiamano "generation exaggeration" (Nudo et al., 2026): un'amplificazione sistematica dei tratti salienti degli utenti simulati oltre i loro valori empirici. I modelli, quindi, non necessariamente restituiscono una replica fedele dei comportamenti osservati, ma li ricostruiscono secondo dinamiche di ottimizzazione interna che producono distorsioni strutturali rispetto ai dati di partenza, con effetti di polarizzazione, stilizzazione e tossicità linguistica superiori ai valori umani di riferimento. Il dato ha implicazioni dirette per l'uso dei LLM come proxy del comportamento umano in contesti di moderazione dei contenuti, simulazione deliberativa o modellazione delle politiche, perché segnala che la fedeltà rappresentativa del sistema è subordinata alle sue dinamiche generative interne. Una manifestazione specifica del fenomeno si osserva nelle valutazioni di credibilità delle fonti informative. Loru et al. (2025b), confrontando Gemini 1.5 Flash, GPT-4o mini e LLaMA 3.1 con i sistemi di rating NewsGuard e Media Bias Fact Check, mostrano che i giudizi di affidabilità prodotti dai LLM correlano sistematicamente con specifici marker lessicali, suggerendo che le valutazioni siano guidate più da associazioni statistiche apprese che da una ricostruzione strutturata del processo di valutazione. Il dato suggerisce che, anche quando i LLM appaiono allineati con criteri esperti di rating, il loro processo decisionale interno resta governato da associazioni statistiche distribuite sul lessico, piuttosto che da un ragionamento contestuale sul contenuto delle fonti.

La rilevanza di questi rischi emerge soprattutto nella loro capacità di contribuire, nel tempo, alla normalizzazione e alla riproduzione di rappresentazioni sociali stereotipate e distorte. Nel momento in cui una società inizia ad attribuire ai sistemi automatizzati un'aura di neutralità tecnica, infatti, le decisioni discriminatorie che questi producono rischiano di essere percepite come più giustificabili rispetto a quelle umane, proprio perché filtrate da un'apparenza di oggettività algoritmica (Acemoglu, 2021). La posta in gioco, insomma, è la naturalizzazione, su scala di massa, di rappresentazioni sociali distorte, in un contesto in cui il sistema interlocutore è percepito come autorevole, neutro e affidabile. Si tratta di un rischio etico sistemico, difficilmente individuabile a livello del singolo output e per questo particolarmente insidioso: un utente che non conosce i meccanismi di addestramento di un LLM, né le diverse forme di bias che possono manifestarsi negli output (sociali, cognitive, architetturali, collettive) non ha strumenti per comprendere che una risposta possa essere influenzata da bias strutturali piuttosto che riflettere una valutazione neutrale della realtà.

2.2.3 Fiducia

Un terzo rischio, di natura cognitivo-interpretativa, è quello della fiducia eccessiva negli output prodotti dai sistemi di IA generativa. Come già osservato a proposito dell'effetto ELIZA, la comunicazione in linguaggio naturale provoca nei lettori processi di attribuzione di intenzione e comprensione: la fluidità del testo porta a proiettare sul sistema caratteristiche cognitive che esso non possiede. Ferrario et al. (2020) propongono un modello a più livelli per analizzare la fiducia nelle interazioni uomo-IA, distinguendo tra una fiducia semplice, fondata sulla dipendenza cognitiva dall'agente in assenza di meccanismi di controllo, e una fiducia riflessiva, che richiede una valutazione esplicita delle ragioni per cui un sistema merita fiducia. La natura di *black box* dei sistemi IA e la loro opacità rendono strutturalmente difficile il passaggio dalla prima forma alla seconda: non potendo verificare il funzionamento interno del sistema, l'utente è portato a colmare questa lacuna con proiezioni antropomorfiche, cedendo a una forma di fiducia semplice che lo espone in misura maggiore alle dinamiche di *automation bias*, una distorsione cognitiva per cui le persone privilegiano i suggerimenti provenienti da sistemi decisionali automatizzati, tendendo a fidarsi eccessivamente di tali sistemi (Parasuraman e Manzey, 2010). La combinazione tra un testo fluente, una risposta rapida e l'autorevolezza percepita della tecnologia crea le condizioni per un'accettazione acritica degli output. Questa dinamica è osservabile anche nel contesto italiano: Savoldi et al. (2025) documentano che gran parte degli utenti utilizzano i chatbot GenAI come fonte primaria di informazione, anche per questioni mediche o di supporto emotivo, pur dichiarando di non essere certi della loro affidabilità. Il paradosso segnalato da Rheu e Cho (2025), che complica ulteriormente il quadro, è che una maggiore conoscenza dei sistemi di IA non si traduce automaticamente in comportamenti più cauti: la consapevolezza dei rischi non garantisce da sola la capacità di gestirli.

Le evidenze sperimentali raccolte negli ultimi anni hanno permesso di verificare empiricamente le dinamiche descritte e di indicarne i meccanismi linguistici. Kim et al. (2024) hanno condotto un esperimento pre-registrato su 404 partecipanti chiamati a rispondere a quesiti di area medica con il supporto di un sistema di ricerca arricchito da un LLM. Nella manipolazione sperimentale gli output del modello variavano per la sola formulazione linguistica dell'incertezza: in alcune condizioni il sistema esprimeva il proprio dubbio in prima persona ("I'm not sure, but…"), in altre attraverso una formula in prospettiva impersonale ("It's not clear, but…"). I risultati mostrano che le espressioni di incertezza in prima persona riducono in misura significativa l'*over-reliance* – l'affidamento eccessivo –

sugli output errati e migliorano l'accuratezza delle risposte fornite dai partecipanti, mentre le formulazioni impersonali producono effetti più deboli e statisticamente non significativi. Il dato è particolarmente interessante dal punto di vista linguistico: non è la presenza di un avverbio di dubbio in sé a modificare il comportamento, ma la sua configurazione pragmatica, l'attribuzione esplicita dell'incertezza a un soggetto enunciatore riconoscibile, a rendere l'utente più cauto.

I limiti di un approccio puramente comunicativo al problema emergono con chiarezza nello studio di Bo et al. (2025), che valuta, tramite un esperimento randomizzato online, l'efficacia di tre diversi interventi finalizzati a modulare l'affidamento al sistema (*reliance)* applicati ai consigli generati da un LLM. Gli interventi testati riducono effettivamente l'*over-reliance* sugli output errati, ma non producono un miglioramento dell'affidamento appropriato, ovvero della capacità di affidarsi al sistema quando ha effettivamente ragione. Più sorprendente ancora, in determinate condizioni i partecipanti diventano più sicuri di sé dopo aver preso decisioni di affidamento sbagliate, segnalando una calibrazione cognitiva carente. Il risultato suggerisce che l'*over-reliance* non è un fenomeno superficiale risolvibile con avvertenze, etichette o disclaimer linguistici aggiunti agli output: si tratta di un meccanismo psicologico più profondo, in cui la fluidità del testo prodotto dal sistema interferisce con la capacità metacognitiva del lettore di valutare la sua stessa accuratezza. In termini di AI Literacy per l'utente, ciò implica che la formazione non possa limitarsi a insegnare il riconoscimento di errori del sistema, ma debba includere strumenti per monitorare e calibrare la propria fiducia e verificare autonomamente le informazioni. Una conseguenza ulteriore della fiducia eccessiva, indagata da Gerlich (2025), è il cosiddetto *cognitive offloading*, "scarico" cognitivo, ovvero la tendenza a delegare ai sistemi di IA quote crescenti dell'attività cognitiva (recupero dell'informazione, organizzazione del ragionamento, formulazione testuale), riducendo l'investimento cognitivo dell'utente. Uno studio empirico su un campione di adulti con questionari standardizzati di pensiero critico e misure di utilizzo dei sistemi generativi, documenta una correlazione negativa tra l'intensità d'uso di strumenti IA e i punteggi di pensiero critico, con il *cognitive offloading* che agisce da mediatore della relazione: gli utenti che ricorrono più frequentemente ai sistemi tendono a delegare loro le operazioni di valutazione e selezione delle informazioni, mostrando in test successivi una minore propensione al pensiero critico (Gerlich, 2025). Il problema, dunque, non è l'uso degli strumenti in sé, ma la modalità passiva con cui spesso vengono integrati nel lavoro intellettuale, in assenza di una sufficiente consapevolezza dei limiti e delle caratteristiche dei sistemi generativi. Celi (2026) propone una riflessione più ampia sul significato della delega

di giudizio ai sistemi generativi: in dialogo con la nozione di *epistemia* (Loru et al., 2025a; Quattrociocchi et al., 2025; cfr. Sezione 2.2.1), riconosce che i LLM producono fluidità linguistica senza accesso al mondo e dunque senza responsabilità epistemica, ma osserva che questa caratteristica non costituisce soltanto un deficit, bensì rivela la natura distribuita della conoscenza umana, che opera attraverso reti simboliche e validazioni collettive. La novità introdotta dai LLM, secondo questa lettura, supera la semplice distorsione della comprensione, arrivando alla separazione tra la generazione di strutture linguistiche significative e l'impegno epistemico verso la realtà: la comprensione, dunque, si esercita nella responsabilità per il significato all'interno di pratiche epistemiche condivise, non nella sola fluidità testuale. Pasquale (2026) articola criticamente questa delega cognitiva sull'uso dell'IA generativa nei contesti decisionali ad alta posta in gioco, in particolare in ambito giuridico. L'autore argomenta che la formulazione di una decisione giudiziaria non è soltanto un atto di trasmissione di informazione, ma il risultato di un coinvolgimento attivo del soggetto decisore con i fatti del caso e con il diritto applicabile: la simulazione del pensiero non è il pensiero, e affidare ai LLM la giustificazione di giudizi rischia di produrre testi che imitano le forme superficiali del ragionamento giuridico senza realizzare la comprensione situata che ne costituisce il fondamento di legittimità. Pasquale (2026) propone il riconoscimento di un "dovere non delegabile di pensare" (*non-delegable duty to think*) come limite normativo all'integrazione dei sistemi generativi nei processi decisionali in cui la responsabilità del giudizio sia un requisito costitutivo. Il principio, formulato in ambito giuridico, ha una portata più generale: la delega cognitiva ai sistemi generativi non può essere indiscriminata, ma deve essere modulata in funzione della natura del compito e del livello di responsabilità implicato.

2.2.4 Privacy dei dati personali

Un quarto ordine di rischi, di natura strutturale e peculiare dei sistemi conversazionali basati su LLM, riguarda le implicazioni per la privacy dei dati personali condivisi nelle interazioni con i chatbot. A differenza di altri strumenti digitali che raccolgono dati in modo passivo, i sistemi conversazionali acquisiscono informazioni in forma attiva e interattiva: i testi immessi dagli utenti, le domande poste, i contesti personali e sensibili condivisi nel corso della conversazione vengono elaborati e, in alcuni casi, possono essere utilizzati per affinare il modello o trasferiti a terze parti. Martin e Zimmermann (2024) osservano che i chatbot generativi occupano una posizione peculiare nel continuum tra tecnologie orientate alla

raccolta di dati e tecnologie orientate alla produzione di esiti: apprendendo dalle interazioni individuali, distillano informazioni di natura linguistica e contestuale la cui sensibilità non è sempre evidente all'utente nel momento della condivisione. La tradizionale analisi costi-benefici che gli individui compiono quando decidono di condividere dati personali, il cosiddetto *privacy calculus*, risulta sistematicamente distorta dall'interfaccia conversazionale, che tende a ridurre la salienza del momento di cessione dei dati attraverso meccanismi di coinvolgimento emotivo e di familiarità simulata: è più probabile che un utente riveli informazioni sensibili nel corso di una conversazione fluente con un agente apparentemente empatico che in risposta a un modulo di raccolta dati esplicito. Kakarala e Rongali (2025) sottolineano che la natura opaca dei sistemi IA rende ulteriormente difficile per gli utenti comprendere come i propri dati vengano trattati e per quali finalità, in assenza di meccanismi di trasparenza e *accountability* chiaramente accessibili. Il quadro normativo europeo e, in particolare, il Regolamento Generale sulla Protezione dei Dati (GDPR), impone obblighi di trasparenza, limitazione delle finalità e consenso informato che si applicano anche ai sistemi IA conversazionali; tuttavia, la distanza tra il *framework* normativo e la comprensione effettiva di questi diritti da parte degli utenti non esperti rimane ampia e rappresenta un terreno di ricerca ancora poco esplorato. Lo studio di Meng e Liu (2025), condotto nel quadro teorico del *privacy calculus* applicato ai chatbot generativi, dimostra che le dimensioni cognitiva, emotiva e comportamentale del coinvolgimento con ChatGPT influiscono positivamente sui benefici percepiti dell'interazione, e che questi benefici agiscono a loro volta come predittore significativo della disponibilità a rivelare informazioni personali: quanto più l'utente si sente coinvolto, gratificato o assistito dal sistema, tanto più ne sopravvaluta i vantaggi e ne sottostima i costi in termini di esposizione dei propri dati. Ciò è coerente con quanto osservato da Martin e Zimmermann a livello teorico, ma ne specifica il meccanismo psicologico: non è soltanto la presenza di un'interfaccia conversazionale a ridurre la salienza del momento della condivisione, ma il coinvolgimento che ne deriva, capace di alterare quantitativamente il bilancio rischi-benefici che la teoria classica del *privacy calculus* presuppone come razionale.

Tran et al. (2025) analizzano le norme di privacy emergenti intorno ai chatbot basati su LLM. Su un campione di 300 utenti statunitensi di ChatGPT, gli autori rilevano che l'82% percepisce le conversazioni con i chatbot come informazioni sensibili o altamente sensibili, più di quanto non considerino l'email o i social media. Tuttavia, e qui emerge un evidente *gap concern-behavior*, uno iato tra la preoccupazione e il comportamento: quasi la metà degli stessi utenti discute argomenti di salute con il sistema e oltre un terzo discute questioni di

natura finanziaria. L'analisi mostra inoltre che, tra i diversi fattori contestuali, soltanto quelli relativi alla trasmissione dei dati (consenso informato, anonimizzazione, rimozione delle informazioni di identificazione personale) influiscono significativamente sulla percezione di appropriatezza della condivisione, mentre l'identità del destinatario, la finalità d'uso e la natura del contenuto non producono effetti rilevanti. Gli utenti, quindi, tendono a scegliere cosa condividere e in che misura sulla base di elementi procedurali (come il consenso o l'anonimizzazione), trascurando invece il contesto comunicativo nel suo complesso (finalità dell'interazione, destinatario).

Il problema si riflette anche sulla percezione e sull'interpretazione dei sistemi conversazionali, soprattutto negli utenti non esperti. Kwesi et al. (2025), sulla base di ventuno interviste semi-strutturate a utenti adulti statunitensi che impiegano chatbot di uso generale a fini di supporto psicologico, rilevano due fraintendimenti ricorrenti; il primo è di natura normativa: gli intervistati ritenevano, sbagliando, che le conversazioni intrattenute con i chatbot fossero coperte dagli stessi vincoli di riservatezza che regolano il colloquio con un terapeuta con licenza[15], proiettando sul sistema una cornice deontologica che esso non possiede. Il secondo fraintendimento è di natura interpretativa: gli utenti tendono a scambiare l'empatia simulata dal modello per una forma di responsabilità professionale, attribuendo alle sue risposte un livello di affidabilità che il sistema non ha. Gli autori, inoltre, introducono il concetto di *intangible vulnerability* per descrivere il fenomeno per cui le informazioni di natura emotiva e psicologica vengono sistematicamente sottovalutate dagli utenti rispetto a dati considerati più tangibili, come quelli finanziari o di geolocalizzazione, pur essendo, dal punto di vista delle conseguenze a lungo termine, altrettanto - e talvolta addirittura più - sensibili. Il dato rinforza la rilevanza di una formazione all'AI literacy che non si limiti agli aspetti tecnici, ma includa una riflessione sulle cornici comunicative che gli utenti, anche inconsapevolmente, applicano ai sistemi conversazionali.

2.2.5 Copyright dei dati di addestramento e degli output

Un ulteriore livello di rischi, di natura giuridica e culturale, riguarda le implicazioni dell'uso dei sistemi di IA generativa per il diritto d'autore e, più in generale, per la circolazione dei contenuti coperti da copyright. A differenza dei rischi finora descritti, che riguardano in primo luogo l'esperienza diretta dell'utente nella ricezione e nella valutazione degli output, il problema del copyright coinvolge tre attori distinti: i creatori dei contenuti utilizzati per

[15] Nello specifico, dalle tutele della normativa statunitense HIPAA.

l'addestramento dei modelli, le aziende sviluppatrici dei sistemi e gli utenti finali che producono testi attraverso questi strumenti. La complessità della questione, che negli ultimi anni ha dato origine a numerose controversie internazionali, può essere ricondotta a tre livelli analitici distinti, identificati dalla letteratura giuridica come *input*, *model* e *output* (Quintais, 2025): l'ingestione di materiale protetto in fase di addestramento, la memorizzazione e potenziale riproduzione di tale materiale all'interno del modello, e infine l'autorialità giuridica dei contenuti generati attraverso il sistema.

Sul piano dell'input, i corpora utilizzati per addestrare i LLM contemporanei sono in larga parte estratti dal web tramite tecniche di *text and data mining* (TDM)[16] e includono volumi significativi di materiale protetto da diritto d'autore: libri, articoli accademici e giornalistici, contenuti culturali e creativi diffusi online. Il quadro normativo europeo, in particolare la Direttiva (UE) 2019/790 sul Copyright nel Mercato Unico Digitale (CDSMD), prevede un'eccezione TDM per finalità di ricerca scientifica (art. 3) e una più ampia eccezione per usi commerciali (art. 4), subordinata però alla possibilità per i titolari dei diritti di esercitare una riserva sull'uso delle proprie opere mediante il cosiddetto *opt-out*, così da impedirne l'impiego nell'addestramento di sistemi di intelligenza artificiale a finalità generale (Quintais, 2025).

Il Regolamento (UE) 2024/1689 sull'intelligenza artificiale (AI Act) si innesta su questo quadro normativo, introducendo, per i fornitori di modelli di IA *general-purpose*, obblighi di trasparenza sulla composizione dei dati di addestramento e l'obbligo di rispettare le riserve di diritti espresse dai titolari ai sensi della direttiva CDSMD (Quintais, 2025).

Come osservato in letteratura, tuttavia, i meccanismi di *opt-out* previsti risultano spesso difficilmente esercitabili nella pratica e l'effettiva applicabilità degli obblighi di trasparenza solleva ancora molte ambiguità interpretative.

Sul piano del modello, evidenze sperimentali recenti hanno mostrato che i LLM non si limitano ad apprendere regolarità statistiche dai dati di addestramento, ma in molti casi memorizzano e possono restituire *verbatim* porzioni consistenti del materiale ingerito. Carlini et al. (2021) dimostrano che è possibile estrarre da GPT-2, attraverso semplici interrogazioni in modalità *black-box*, centinaia di sequenze testuali memorizzate testualmente, comprensive di informazioni personali identificabili (nomi, indirizzi email, numeri di telefono), codice

[16] Processo che permette di trasformare grandi quantità di testo non strutturato in informazioni organizzate e analizzabili. Attraverso tecniche automatiche, consente di individuare parole chiave, schemi ricorrenti, tendenze, relazioni e significati nascosti all'interno dei testi (https://www.ibm.com/it-it/think/topics/text-mining).

sorgente e identificatori univoci. Lo studio rileva inoltre che modelli di dimensioni maggiori risultano più vulnerabili a forme di estrazione di questo tipo, anche per sequenze presenti una sola volta nel dataset di addestramento. Carlini et al. (2023) generalizzano il fenomeno individuando tre relazioni statistiche modellate su scala logaritmica che ne governano la portata: la memorizzazione cresce con la scala del modello, con il numero di duplicazioni di un esempio nel dataset e con la lunghezza del contesto fornito nel prompt. Nello stesso lavoro, gli autori stimano che il modello GPT-J da 6 miliardi di parametri abbia memorizzato almeno l'1% del proprio dataset di addestramento[17], una proporzione molto più alta di quanto precedentemente ipotizzato. Karamolegkou et al. (2023) applicano due strategie di interrogazione (*direct probing*[18] e *prefix probing*[19]) a sei famiglie di LLM, utilizzando come banco di prova bestseller letterari e problemi di programmazione protetti da copyright. I risultati confermano che i modelli sono in grado di riprodurre letteralmente porzioni significative di tali testi e che la capacità di riproduzione cresce con la dimensione del modello, in linea con quanto osservato da Carlini et al. (2021). Giuridicamente, il punto risiede nella differenza tra apprendimento (estrazione di informazione, che in molte giurisdizioni rientra nelle eccezioni di *fair use* o TDM) e ridistribuzione (riproduzione testuale, che invece configura tipicamente una violazione di copyright): un sistema in grado di restituire parola per parola un romanzo, non si limita ad averlo "letto" in fase di addestramento, ma può di fatto ridistribuirlo all'utente che lo interroghi nel modo opportuno, con il rischio di violare i diritti di proprietà intellettuale e sollevando interrogativi sull'eticità di un uso di questo tipo (Karamolegkou et al., 2023). Sul piano dell'output, il nodo della questione riguarda l'autorialità giuridica. Neubauer et al. (2026), in un'analisi comparata dei quadri giuridici di Regno Unito, Stati Uniti e Germania alla luce della Convenzione di Berna, osservano che le legislazioni vigenti, costruite intorno alla figura dell'autore umano, risultano sistematicamente inadeguate a rispondere a un sistema in grado di imitare l'espressione umana senza una coscienza autoriale. Si riscontrano ambiguità irrisolte su tre fronti principali: l'attribuzione dell'autorialità degli output (a chi appartiene un testo generato da un LLM a partire dal prompt di un utente?), il regime di licenza dei dati di addestramento (quali sono gli

[17] The Pile, un ampio corpus testuale open source (circa 800 GB), assemblato da EleutherAI a partire da fonti eterogenee, utilizzato per l'addestramento di modelli linguistici di grandi dimensioni (Gao et al., 2020).

[18] Interrogazione diretta, ad esempio: "Qual è la prima pagina di [TITOLO]?" (Karamolegkou et al., 2023).

[19] Indagare la capacità del modello di generare continuazioni coerenti a partire dai primi 50 *token* di un testo (Karamolegkou et al., 2023).

obblighi di compenso verso i creatori i cui contenuti sono stati utilizzati?) e la tutela dei diritti morali, in particolare la protezione dell'identità stilistica e della voce di un autore contro imitazioni generate algoritmicamente, tema sollevato pubblicamente da numerose figure del mondo artistico e culturale. Gli autori propongono un quadro globale di riforma articolato intorno a una definizione di "opera" sensibile alla dimensione semantica, a un sistema di licenza e remunerazione per i creatori i cui dati sono stati utilizzati in addestramento e a obblighi rafforzati di trasparenza, per garantire la tracciabilità e la responsabilità.

Per l'utente non esperto che produce testi con l'assistenza dei LLM (uno studente, un professionista, un creator di contenuti), le conseguenze pratiche di questo scenario sono concrete e si manifestano su due fronti. Il primo è il rischio di plagio non intenzionale: un output apparentemente originale può contenere riformulazioni o riproduzioni dirette di materiale protetto, esponendo l'utente a possibili contestazioni in contesti accademici o professionali (Cotton et al., 2024). Il secondo riguarda l'incertezza sull'autorialità e quindi sulla legittima rivendicazione dei propri output: in assenza di un quadro giuridico chiaro, lo statuto stesso dei testi prodotti con assistenza generativa rimane incerto. È legittimo, dunque, chiedersi: a chi appartiene un output prodotto mediante sistemi di IA generativa? Sul piano dell'AI literacy, ne deriva l'esigenza di una formazione che includa, accanto alla consapevolezza tecnica e critica del funzionamento dei sistemi, anche una conoscenza dei vincoli giuridici sull'uso degli strumenti generativi e degli obblighi deontologici di citazione, particolarmente rilevanti in ambiti come quello accademico o scientifico, in cui l'integrità del testo e la trasparenza sulle fonti rappresentano requisiti fondamentali della pratica professionale.

2.2.6 Sycophancy

Un altro rischio, strettamente connesso ai precedenti e di particolare rilevanza nel contesto delle interazioni con i chatbot generativi, è quello della *sycophancy*, traducibile in italiano come "adulazione", "compiacenza" o addirittura "servilismo" del modello nei confronti dell'utente. Il dizionario Cambridge definisce il termine *sycophancy* come "*il comportamento di qualcuno che elogia persone potenti o ricche in modo non sincero, di solito per ottenere da loro qualche vantaggio*" (trad. mia). Nell'ambito dei sistemi di IA generativa, si tratta della tendenza dei LLM a privilegiare l'allineamento con le opinioni, le aspettative o le preferenze espresse dall'interlocutore a discapito dell'accuratezza fattuale o della coerenza argomentativa: il modello, anziché correggere un'affermazione errata o sostenere un punto di

vista divergente, tende a confermare quanto detto dall'utente, eventualmente accompagnando la conferma con espressioni di apprezzamento o con una modulazione del proprio tono in direzione conciliante. Il fenomeno, già osservato a livello aneddotico fin dalla diffusione dei primi chatbot generativi, è stato progressivamente formalizzato in letteratura come comportamento sistematico e misurabile (Perez et al., 2022; Sharma et al., 2023), e rappresenta oggi uno dei principali ostacoli all'uso affidabile dei sistemi conversazionali in ambiti decisionali. Il meccanismo all'origine della *sycophancy* è in larga parte riconducibile alle fasi di adattamento successive al *pre-training*, in particolare al *Reinforcement Learning from Human Feedback*, già introdotto nella Sezione 2.1.2: durante l'allineamento, i modelli vengono ottimizzati sulla base delle preferenze espresse da valutatori umani, che tendono sistematicamente a premiare risposte cortesi, accomodanti e in linea con le proprie aspettative, anche quando queste contengono inesattezze (Sharma et al., 2023). La conseguenza è che il sistema apprende a "massimizzare" la soddisfazione percepita dell'utente piuttosto che la correttezza fattuale, sviluppando una tendenza strutturale all'accordo. Tale dinamica si intreccia con un secondo fenomeno, quello del cosiddetto *warmth training*, ovvero la pratica di addestrare i modelli a rispondere in modo empatico e amichevole. L'addestramento alla cordialità, infatti, produce un peggioramento sistematico delle prestazioni dei modelli, con tassi di errore superiori del 10-30% rispetto alle versioni originali: i modelli *warm* risultano più inclini a sostenere teorie del complotto, a fornire informazioni inesatte e a dare consigli medici scorretti (Ibrahim et al., 2026). L'effetto si amplifica quando l'utente esprime stati d'animo di tristezza o vulnerabilità: in tali condizioni, i modelli addestrati alla cordialità mostrano una probabilità di circa il 40% più alta di assecondare l'utente, confermando le sue convinzioni errate. Lo studio di Ibrahim et al. (2026) mette in luce una tensione tra empatia e accuratezza, che nello sviluppo recente dei sistemi conversazionali è stata spesso sottovalutata o trattata come secondaria. In particolare, i risultati suggeriscono che l'orientamento verso la simulazione affettiva, finalizzato a rendere l'interazione più coinvolgente e "umana", possa avere un impatto negativo sull'affidabilità degli output, configurandosi come un potenziale fattore di rischio sul piano epistemico.

Per quanto riguarda l'analisi empirica, Fanous et al. (2025) propongono il primo framework standardizzato per la misurazione del fenomeno, denominato SycEval, e lo applicano in modo comparativo a ChatGPT-4o, Claude-Sonnet e Gemini-1.5-Pro su due dataset, AMPS, di area

matematica, e MedQuad, di consulenza medica. I risultati documentano comportamenti tendenti alla *sycophancy* nel 58,19% dei casi complessivi, con tassi differenziati tra i modelli. Lo studio introduce inoltre una distinzione concettuale rilevante tra *sycophancy progressive*, in cui il cedimento del modello all'utente porta paradossalmente alla risposta corretta (43,52% dei casi), e *sycophancy regressive*, in cui invece il cedimento porta a una risposta scorretta (14,66%). Risulta rilevante che le confutazioni preventive fornite dall'utente, formulate quindi prima della risposta del modello, generano tassi di sycophancy più elevati rispetto a quelli formulati nel corso della conversazione (61,75% vs 56,52%, $p < 0,001$), con un effetto particolarmente marcato nei compiti computazionali. Una volta instaurata, la condotta *sycophantic* mostra infine una persistenza del 78,5% indipendentemente dal contesto e dal modello: si tratta, quindi, di un comportamento difficilmente reversibile attraverso semplici interventi conversazionali.

Una conseguenza diretta della *sycophancy* sul comportamento decisionale emerge da Li et al. (2026), che in uno studio sperimentale su 106 partecipanti confrontano un sistema di IA *non sycophantic* con tre varianti che incorporano rispettivamente accordo opinionistico (*opinion agreement*), lode diretta (*direct praise*) e auto-svalutazione (*self-deprecation*) del sistema, in due contesti decisionali a diverso livello di rischio: una predizione di compatibilità in uno scenario di *speed-dating* (a basso rischio) e una scelta di investimento in ETF (ad alto rischio). I risultati mostrano che le diverse forme di *sycophancy* producono effetti differenziati sui processi decisionali: l'accordo opinionistico tende a rinforzare le decisioni iniziali dell'utente, mentre l'auto-svalutazione del sistema ne aumenta la fiducia complessiva. Le interviste qualitative condotte a margine dell'esperimento rivelano un atteggiamento ambivalente dei partecipanti, che da un lato apprezzano il sostegno emotivo offerto dal sistema, dall'altro tendono a metterne in discussione l'obiettività quando la lode diventa eccessiva.

La *sycophancy*, inoltre, ha un impatto notevole anche sulla percezione dell'affidabilità del sistema. I sistemi a registro complimentoso che adattano la propria posizione a quella dell'utente, infatti, vengono percepiti come meno autentici e ricevono un livello di fiducia inferiore, mentre i sistemi a registro neutro adattivi vengono percepiti come più autentici e affidabili (Sun e Wang, 2026). Si tratta di un risultato che ha implicazioni rilevanti per il design dei sistemi conversazionali (Sun e Wang, 2026): si può ipotizzare che configurazioni

di *sycophancy* meno evidenti, in cui è assente la lode esplicita, producano un effetto di *over-trust* (e conseguente *over-reliance*) più subdolo, proprio perché non attivano nei lettori, che tendono quindi a fidarsi eccessivamente, meccanismi di sospetto; meno il sistema sembra adulatore, dunque, maggiore è il rischio che l'utente si fidi, risultando maggiormente vulnerabile a distorsioni informative. Questo meccanismo richiama quanto la linguistica ha osservato sul potere persuasivo dell'implicito: come mostra Lombardi Vallauri (2019), i contenuti veicolati in forma implicita risultano più efficaci proprio perché sfuggono al vaglio critico che il destinatario riserva alle asserzioni esplicite, venendo accolti senza essere sottoposti a valutazione. La compiacenza non dichiarata del modello agisce, in questo senso, come un implicito: orienta la fiducia dell'utente senza rendersi visibile. La sycophancy, dunque, si configura come un rischio trasversale, che si interseca con i fenomeni discussi nelle sezioni precedenti: amplifica gli effetti della fiducia eccessiva (sezione 2.2.3), può favorire la conferma e la diffusione di informazioni errate o allucinate (sezione 2.2.1) e introduce un'asimmetria sistematica nei processi di valutazione critica degli output. Per un utente non esperto, il rischio è particolarmente insidioso perché il segnale che permetterebbe di mettere in discussione una risposta, ovvero la sua plausibilità intrinseca, viene rinforzato proprio dalla forma compiacente con cui essa viene presentata. La fluidità, la cortesia e il tono accomodante non sono dunque, di per sé, indici di affidabilità del contenuto, ma possono in alcuni casi rappresentare segnali di una distorsione sistematica del comportamento del sistema indotta dalle stesse scelte di addestramento adottate dagli sviluppatori per migliorarne la qualità dell'esperienza d'uso.

I rischi qui descritti sono vari e insidiosi e non riconducibili a un'unica causa: le allucinazioni rappresentano errori tecnici con conseguenze epistemiche dirette, rionducibili al più ampio fenomeno dell'epistemia (Loru et al. 2025a; Quattrociocchi et al., 2025); i bias sono un problema sistemico con implicazioni etiche e sociali, legato alla struttura statistica del materiale linguistico utilizzato per l'addestramento e amplificato dallo sbilanciamento culturale dei corpora; la fiducia eccessiva è un rischio interpretativo che si situa al confine tra le caratteristiche del sistema e i meccanismi cognitivi dell'utente, come mostrano tanto gli studi sulla formulazione linguistica dell'incertezza quanto quelli sui limiti degli interventi di *reliance* e sul *cognitive offloading*; la *sycophancy* aggrava trasversalmente gli altri rischi, configurandosi come distorsione sistematica del comportamento del sistema, accomodante nei

confronti dell'utente. Le implicazioni per la privacy, invece, costituiscono un rischio strutturale aggravato dall'opacità dei processi di trattamento e dalla natura conversazionale dell'interfaccia, accentuato dalla distorsione del *privacy calculus*, dai fraintendimenti regolativi e dalla sottovalutazione sistematica delle informazioni di natura emotiva e psicologica; nell'ambito dei rischi legati alla privacy si inserisce la questione legata al copyright, che riguarda sia l'ingestione di materiale protetto in fase di addestramento, sia la riproduzione testuale negli output, sia l'incertezza sull'autorialità attribuita ai contenuti generati. Ad accomunare queste diverse categorie di rischi, la caratteristica seguente: il fatto di essere difficilmente rilevabili senza una conoscenza adeguata di base del funzionamento dei modelli, dei meccanismi di produzione degli output e del quadro normativo che ne regola l'uso. È questo il motivo per cui, in un contesto in cui i sistemi di IA sono sempre più diffusi e i rischi associati sempre più insidiosi, è fondamentale una vera e propria alfabetizzazione di base sul tema, soprattutto per non esperti: gli utenti necessitano non di un bagaglio di competenze tecniche, ma di una serie di strumenti concettuali che consentano loro di essere attori attivi nel rapporto coi sistemi di IA per mantenere un rapporto critico con essi, formulare giudizi fondati sulla loro affidabilità, di riconoscere le condizioni in cui i propri dati sono esposti e di non delegare acriticamente al sistema decisioni che richiedono valutazione autonoma.

2.3 Il concetto di AI Literacy

La crescente diffusione delle applicazioni di IA e dei LLM nella vita quotidiana rende necessario interrogarsi sulle competenze richieste agli utenti per comprendere e utilizzare tali sistemi in modo consapevole. Sistemi di IA sono infatti sempre più diffusi, accessibili a utenti eterogenei e adattabili a contesti differenti: anche se gli utenti possono non esserne del tutto consapevoli, la tecnologia IA è integrata in molti aspetti della vita quotidiana e le sue applicazioni linguistiche stanno trasformando le interazioni digitali. Senza conoscere adeguatamente queste tecnologie e il loro impatto, gli utenti rischiano di non avere gli strumenti per comprendere la nuova realtà digitale contemporanea, di non interpretarla in modo consapevole e, di conseguenza, di non riuscire ad adattarsi ad essa come soggetti autonomi. In questo scenario, si inserisce il concetto di AI *literacy*, letteralmente "alfabetizzazione" sul tema dell'Intelligenza Artificiale. Nella sua accezione originaria, il

termine *literacy* (alfabetizzazione) si riferisce alle capacità di leggere e scrivere (Long e Magerko 2020) . Quello di "alfabetizzazione" è un concetto centrale nell'educazione, cruciale per garantire uguaglianza sociale e opportunità di partecipazione attiva, sia a livello individuale sia collettivo (Kagitcibasi et al., 2005). Nel corso del Novecento, il concetto si è progressivamente ampliato fino a includere forme di alfabetizzazione specifiche (*digital literacy, scientific literacy, data literacy*) che rispondono alla crescente complessità degli ambienti informativi (Long e Magerko, 2020). In questa linea si colloca, nel mondo contemporaneo, l'AI *literacy*, un tipo di alfabetizzazione che si definisce in rapporto a un nuovo oggetto, i sistemi di intelligenza artificiale, e alle competenze necessarie per interagire con essi in modo consapevole.

L'importanza dell'AI literacy si spiega in rapporto alla diffusione stessa dei sistemi generativi: se l'alfabetizzazione tradizionale è stata considerata, nel Novecento, una condizione necessaria per la partecipazione sociale e politica, oggi la capacità di comprendere e valutare criticamente i sistemi di IA va intesa come una competenza analoga, un prerequisito per orientarsi in un ambiente informativo sempre più mediato da agenti artificiali. In questo senso, la mancanza di AI literacy non costituisce soltanto un problema individuale, ma può contribuire ad accentuare forme di divario digitale e informativo, con effetti sulle opportunità educative, professionali e civiche di chi resta ai margini (Yi, 2021).

Il termine "AI literacy" risale, nella sua formulazione esplicita, al contributo di Kandlhofer et al. (2016), che lo introducono nel quadro di una proposta educativa estesa dalla scuola dell'infanzia all'università. A partire da quel primo riferimento, la ricerca sull'AI literacy ha conosciuto una rapida espansione, accompagnata da un proliferare di definizioni che riflettono la natura ancora aperta e dibattuta del costrutto. Secondo Long e Magerko (2020) l'AI literacy comprende "*un insieme di competenze che consentono agli individui di valutare criticamente le tecnologie di IA; comunicare e collaborare efficacemente con l'IA; e utilizzare l'IA come strumento online, a casa e sul posto di lavoro*" (trad. mia)[20]. Gli autori articolano il costrutto in diciassette competenze raggruppate in cinque macroaree, organizzate intorno alle domande fondamentali "Cos'è l'IA?", "Cosa può fare l'IA?", "Come funziona l'IA?", "Come dovrebbe essere usata l'IA?" e "Come la percepiscono le persone?". In una

[20] "We define AI literacy as a set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace." (Long e Magerko, 2020).

prospettiva più ampia, Aoun (2017), nel quadro di una riflessione sul futuro dell'istruzione superiore in un'epoca di automazione crescente, propone il modello della *humanics*, articolato in tre forme di alfabetizzazione complementari: la *data literacy*, la *technological literacy* (che include la comprensione dei principi di base dei sistemi di IA) e la *human literacy*, intesa come capacità di esercitare competenze tipicamente umane (creatività, empatia, ragionamento critico) in contesti pervasi dalla tecnologia. Altri contributi mettono in evidenza la natura multidimensionale dell'AI literacy. Wong et al. (2020), ad esempio, individuano tra i suoi elementi costitutivi la conoscenza dei concetti fondamentali, la comprensione delle applicazioni, la consapevolezza delle implicazioni etiche e dei rischi legati alla sicurezza. Yi (2021) amplia la prospettiva, sottolineando che l'AI *literacy* non è solo una competenza tecnica, ma comprende anche la capacità di riconoscere criticamente i cambiamenti culturali, permettendo così all'individuo di progettare la propria vita e imporsi come soggetto autonomo nell'era dell'IA. Pinski e Benlian (2023) la concettualizzano come una competenza socio-tecnica degli esseri umani, articolata in due componenti qualitativamente distinte (conoscenza esplicita ed esperienza pratica) e strutturata lungo cinque dimensioni: la conoscenza della tecnologia IA, la conoscenza degli attori umani coinvolti nell'interazione uomo-AI, la conoscenza delle fasi del processo IA (input, elaborazione, output), l'esperienza d'uso e l'esperienza di progettazione. Una concettualizzazione altrettanto influente è quella di Ng et al. (2021), che articolano l'AI *literacy* in quattro categorie ispirate alla tassonomia di Bloom: conoscere e comprendere l'IA (*know and understand*), usare e applicare l'IA (*use and apply*), valutare e creare con l'IA (*evaluate and create*) e riflettere sulle sue implicazioni etiche (*AI ethics*). Sul piano istituzionale, UNESCO (2024) ha recentemente pubblicato due *framework* di competenze sull'IA, uno rivolto agli studenti e uno agli insegnanti, che rappresentano un riferimento istituzionale condiviso per la progettazione di interventi educativi e di policy. I *framework* sono articolati in quattro dimensioni che includono un orientamento centrato sull'essere umano (*human-centred mindset*), l'attenzione esplicita all'etica dell'IA, la conoscenza delle tecniche e delle applicazioni dell'IA e la capacità di progettazione di sistemi.

Al di là delle differenze, le varie definizioni discusse consentono di riconoscere alcuni elementi ricorrenti: l'AI literacy va intesa come un costrutto multidimensionale, che include una componente concettuale (sapere cosa sono e come funzionano, in termini generali, i sistemi di IA), una componente critica (saperne valutare output, limiti e rischi), una componente pragmatica (saperli usare in modo efficace) e una componente etico-sociale (riconoscerne le implicazioni più ampie). Una simile articolazione in quattro componenti, pur

con etichette differenti, ricorre nei framework proposti da Ng et al. (2021) e da Wang et al. (2023), e converge con la sistematizzazione operata da UNESCO (2024). L'AI literacy, di conseguenza, non può essere ridotta a una semplice competenza tecnica, ma rappresenta una base imprescindibile per la valutazione dei sistemi IA, che va intesa come un insieme integrato di conoscenze, abilità e consapevolezze critiche, strettamente connesse agli obiettivi educativi legati all'uso responsabile dell'intelligenza artificiale.

Accanto alle definizioni teoriche, la ricerca si è concentrata anche sulla misurazione empirica del costrutto, attraverso la creazione di strumenti di valutazione in grado di quantificare la competenza degli utenti sul tema. Una revisione sistematica recente della letteratura accademica (Gounopoulos, 2025) ha identificato otto studi, tutti pubblicati tra il 2023 e il 2024, che propongono e testano strumenti di misurazione dell'AI Literacy, specificamente rivolti a un pubblico di non esperti, ovvero individui che utilizzano i sistemi di IA nella vita quotidiana o nel lavoro senza possedere competenze tecniche specializzate. La concentrazione di questa produzione in un arco temporale così recente è di per sé rivelatrice: il campo è ancora in fase di costruzione e la disponibilità di strumenti validati e condivisi resta limitata. Molti strumenti oggi disponibili, infatti, misurano soprattutto le competenze percepite degli utenti, ciò che gli utenti ritengono di sapere o di saper fare, più che la loro conoscenza effettiva.

Pinski e Benlian (2023) propongono uno strumento di valutazione testato su campione piccolo (50 persone), che si concentra più sul piano teorico che sulla validazione empirica. Laupichler et al. (2023), propongono invece una scala composta da 31 item, basata su ricerche precedenti e articolata in tre componenti chiave: comprensione tecnica dell'IA, applicazione pratica e valutazione critica. Questo strumento è empiricamente più robusto e psicometricamente solido. Anche Wang et al. (2023) propongono una scala di misurazione, validata su un campione adulto e articolata in quattro fattori: consapevolezza, uso, valutazione ed etica.

In generale, l'analisi comparativa degli strumenti svolta da Gounopoulos (2025) permette di isolare un nucleo di componenti ricorrenti, che si sovrappongono: tutti gli strumenti includono la comprensione di cosa sia l'IA e di come funzioni, la capacità di utilizzarla in contesti pratici, la valutazione critica degli output e delle implicazioni etiche e sociali del suo uso. Questi strumenti presentano, tuttavia, alcune criticità metodologiche. La maggior parte degli strumenti rileva competenze percepite piuttosto che conoscenze effettive, attraverso item di auto-valutazione che possono introdurre distorsioni significative. I campioni utilizzati per testare gli strumenti, inoltre, sono spesso ridotti e, di conseguenza,

scarsamente rappresentativi della popolazione generale. La dimensione etica, infine, tende spesso a rimanere marginale rispetto a quelle più tecniche e applicative. Più in generale, la tendenza degli strumenti a rilevare competenze percepite anziché effettive non è un problema specifico dell'AI literacy. La meta-analisi di Li e Zhang (2021), condotta su 97 campioni e oltre 68.000 partecipanti, rileva tra autovalutazione e prestazione linguistica effettiva una correlazione soltanto moderata (r = 0,47), che spiega circa un quinto della varianza nelle competenze misurate: l'auto-percezione si allinea in modo significativo ma imperfetto con la competenza reale. Gli stessi autori osservano, peraltro, che tale correlazione è perfino più alta di quelle riscontrate in altri ambiti – come la psicologia o l'istruzione superiore (r tra 0,30 e 0,40) – segno che la difficoltà di stimare accuratamente le proprie competenze è un fenomeno trasversale alle discipline. Il limite degli strumenti fondati sull'auto-percezione non è dunque specifico dell'AI literacy, ma ricorre ogni volta che si chiede a un soggetto di stimare le proprie competenze.

Tra gli strumenti più robusti sul piano psicometrico ed empirico, si distingue la scala SAIL4ALL (Soto-Sanfiel et al. 2024): uno strumento completo e robusto, articolato in 56 item raggruppati in quattro dimensioni ("Cos'è l'IA?", "Cosa può fare l'IA?", "Come funziona l'IA?", "Come dovrebbe essere usata l'IA?") e disponibile in due versioni, una con domande vero/falso e una con scala Likert a 5 punti.

Va sottolineato, inoltre, che la letteratura più recente suggerisce anche una prospettiva più critica sul costrutto di AI literacy. Rheu e Cho (2025), in uno studio condotto su studenti universitari statunitensi, documentano un paradosso significativo: tra i partecipanti, una più alta AI literacy autoriportata si associa a una minore frequenza di comportamenti di verifica degli output dei sistemi generativi. Il dato apre una questione rilevante sul rapporto tra competenza dichiarata e comportamento effettivo, mostrando come la sola consapevolezza dei rischi non si traduca automaticamente in pratiche d'uso più caute.

Nel complesso, la letteratura restituisce un campo in rapido sviluppo, ma ancora privo di standard condivisi: le definizioni di AI literacy proliferano, così come i tentativi di sviluppare strumenti di misurazione, che però non convergono verso un framework universalmente valido e i dati empirici su popolazioni adulte non esperte restano scarsi, soprattutto in contesti non anglofoni.

3. Metodologia

La ricerca si è articolata in due fasi: una prima fase di indagine empirica sul livello di AI literacy in un pubblico di non specialisti, condotta tramite un questionario proposto a un campione di adulti italiani; e una seconda fase, in cui è stato sviluppato un chatbot basato su architettura Retrieval-Augmented Generation (RAG), concepito come strumento di alfabetizzazione divulgativa rivolto allo stesso pubblico. La scelta di integrare i due strumenti – il questionario per la raccolta quantitativa e descrittiva dei dati e il chatbot come strumento conseguente di alfabetizzazione – risponde a un'esigenza ricorrente nella letteratura sul tema dell'AI literacy: la necessità di affiancare alla rilevazione dei livelli di competenza, e quindi a una semplice analisi descrittiva del fenomeno, interventi formativi mirati e accessibili (Long e Magerko, 2020; Pinski e Benlian, 2023; Gounopoulos, 2025). I due strumenti, infatti, sono concepiti come distinti, ma complementari: il questionario assolve a una funzione diagnostica, restituendo un quadro descrittivo non solo delle conoscenze, ma anche delle pratiche e delle percezioni del pubblico target; il chatbot, che nasce dalla letteratura sui rischi e sui fondamenti dei LLM e dal quadro fornita dal questionario, propone un modello di intervento divulgativo. L'idea di base è quella di offrire all'utente uno spazio interattivo, in cui poter porre domande sull'IA generativa e ricevere risposte fondate su una *knowledge base* controllata, riducendo la dipendenza da fonti non verificate o dai chatbot generalisti stessi, soggetti ai limiti documentati nel capitolo precedente. Nel seguente capitolo si descrivono il questionario sull'AI literacy (sezione 3.1) e la struttura del chatbot (sezione 3.2). A queste due componenti si aggiunge una fase di valutazione esplorativa del prototipo, condotta con un campione pilota di tre utenti attraverso un questionario di usabilità dedicato, descritto nella sezione 3.3. Entrambi i questionari – quello sull'AI literacy e quello di valutazione del chatbot – sono riportati integralmente in appendice. Il codice sorgente del prototipo è disponibile in una repository pubblica su GitHub (https://github.com/gaiaeleonoradiraimondo/Chatbot-AI-Literacy).

3.1 Il questionario sull'AI literacy

Il questionario si articola in sette sezioni per un totale di 40 item, distribuiti come illustrato di seguito. Le prime due sezioni rilevano informazioni sociodemografiche e di contesto (dati

demografici ed esperienza pregressa con i chatbot AI). Le sezioni successive misurano le tre dimensioni TUCAPA (Laupichler et al., 2023) integrate dalla dimensione etico-sociale derivata da Wang et al. (2023) e Ng et al. (2021); la sezione conclusiva verifica la conoscenza fattuale attraverso item vero/falso, secondo l'impostazione di SAIL4ALL (Soto-Sanfiel et al., 2024). Lo strumento è anonimo, non richiede l'identificazione del rispondente né la raccolta di dati personali identificativi ed è stato sviluppato secondo i criteri di trasparenza e trattamento dati previsti dal Regolamento (UE) 2016/679 (GDPR). Il questionario è stato somministrato online, attraverso Google Forms, scelta motivata dall'ampia accessibilità della piattaforma, dalla gratuità d'uso e dalla compatibilità con i requisiti di anonimato. Il tempo di compilazione stimato è di circa 10-15 minuti. Il questionario è stato diffuso attraverso una strategia di campionamento a valanga (*snowball sampling* – Goodman, 1961): il collegamento al modulo è stato condiviso attraverso la rete di contatti di chi scrive, tramite social network e applicazioni di messaggistica istantanea, con la richiesta esplicita di ulteriore condivisione. Questa tecnica di campionamento non probabilistico prevede che ciascun partecipante possa estendere la rete di partecipazione a contatti non direttamente raggiungibili dalla fonte. Il campione risultante non è rappresentativo della popolazione generale degli utenti italiani di chatbot IA; il profilo demografico dei rispondenti riflette le caratteristiche della rete di diffusione e costituisce un limite metodologico di cui si rende conto nell'analisi dei dati (cfr. Capitolo 4). La somministrazione si è svolta tra l'8 e il 22 maggio 2026 e ha portato alla raccolta di 92 questionari validi.

3.1.1 Il quadro teorico

Dalla letteratura discussa nella Sezione 2.3 emerge una convergenza sostanziale, pur nella diversità delle formulazioni, sull'idea che l'AI literacy vada concepita come un costrutto multidimensionale, che si articola su quattro direttrici: quella concettuale (comprensione di cosa sia e come funzioni l'IA), quella di valutazione critica (capacità di giudicare output, limiti, rischi e contesti), quella pragmatica (capacità di uso effettivo) e quella etico-sociale (consapevolezza delle implicazioni). A partire da questa concettualizzazione, condivisa da Ng et al. (2021), Wang et al. (2023) e ripresa anche dal framework UNESCO (2024), è stata articolata la struttura del questionario. Operativamente, lo strumento di base è la scala SNAIL di Laupichler et al. (2023), la cui struttura a tre fattori – il cosiddetto modello TUCAPA: Comprensione Tecnica ("*Technical Understanding*", TU), valutazione critica ("*Critical*

Appraisal", CA), applicazione pratica ("*Practical Application*", PA) – è risultata psicometricamente solida su un campione di non esperti. Sulla SNAIL si basano le sezioni C, D ed E del questionario. La dimensione etico-sociale, sottorappresentata nella SNAIL, è stata integrata come sezione autonoma (sezione F), sulla base degli item proposti da Wang et al. (2023) e Ng et al. (2021). Tenendo conto della criticità sollevata da Rheu e Cho (2025) sul cosiddetto paradosso dell'AI literacy – la possibile discrepanza tra competenza percepita e dichiarata dell'utente e conseguente comportamento effettivo – alle scale Likert di auto-valutazione (sezioni C-F) è stata affiancata una sezione di verifica della conoscenza fattuale in formato vero/falso (sezione G), ispirata al formato vero/falso previsto dalla scala SAIL4ALL (Soto-Sanfiel et al., 2024). La combinazione dei due formati consente di disporre, per ciascun rispondente, di una misura di competenza auto-percepita e di una misura di competenza effettiva, e di valutarne eventuali scostamenti.

Il questionario risultante, quindi, non costituisce la replica letterale di uno strumento esistente, ma un'integrazione adattata al caso specifico: target italiano, adulto (18-50+ anni), non specialista, con focus circoscritto ai chatbot basati su LLM. Gli item del questionario sono stati scritti in italiano e sono privi di tecnicismi specifici per garantire l'accessibilità a rispondenti privi di background tecnico. Nelle domande sono stati inseriti, inoltre, riferimenti concreti a chatbot di uso comune (ChatGPT, Gemini, Copilot) per ancorare le domande all'esperienza diretta dell'utente.

3.1.2 Sezione A – Dati socio-demografici

La sezione iniziale indaga il profilo socio-demografico del rispondente, articolato in quattro variabili: età (cinque fasce: 18-25, 26-35, 36-45, 46-50, 50+), genere (donna, uomo, non binario/altro, preferisco non rispondere), titolo di studio più elevato conseguito (articolato secondo i livelli del sistema di istruzione italiano: diploma di scuola secondaria di primo grado, diploma di scuola secondaria di secondo grado, laurea triennale, laurea magistrale o superiore) e ambito professionale attuale (dieci categorie più la voce "altro"). La granularità è stata scelta in modo da consentire confronti per sottogruppi (in particolare per fascia d'età e titoli di studio) senza appesantire la compilazione.

3.1.3 Sezione B – Esperienza con i chatbot

La seconda sezione rileva informazioni sull'esperienza diretta del rispondente con i chatbot, attraverso cinque item: l'avvenuto utilizzo di almeno un chatbot AI (sì/no), la frequenza d'uso (cinque livelli, da "mai" a "quotidianamente"), gli scopi d'uso (selezione multipla su undici categorie funzionali, dalla ricerca di informazioni alla programmazione, dal supporto emotivo alla traduzione, con possibilità di segnalare sia il non utilizzo, sia altri utilizzi specifici), un giudizio complessivo sulla propria familiarità con i chatbot e sulla propria fiducia nelle risposte generate da questi ultimi (per entrambi gli item, sei livelli possibili, da "nessuna" a "molto alta"). Lo scopo della sezione è fornire uno sfondo comportamentale e percettivo, utile a contestualizzare le risposte delle sezioni successive.

3.1.4 Sezioni C-F – Auto-valutazione delle competenze

Le sezioni C, D, E ed F rappresentano il nucleo del questionario, si compongono di ventiquattro item che misurano l'AI literacy auto-percepita lungo quattro dimensioni. Tutti gli item utilizzano una scala Likert a cinque punti (1 = "per niente d'accordo"; 2 = "poco d'accordo"; 3 = "abbastanza d'accordo"; 4 = "d'accordo"; 5 = "completamente d'accordo") e sono formulati come affermazioni in prima persona del tipo "sono in grado di…", su modello dell'impostazione della scala SNAIL (Laupichler et al., 2023). L'adozione di una scala Likert a cinque punti risponde a un compromesso tra granularità della misurazione e accessibilità per il rispondente: la letteratura sul disegno di scale di risposta, infatti, suggerisce che per popolazioni non abituate a strumenti psicometrici, scale più ampie tendono ad aumentare il carico cognitivo senza portare a un effettivo guadagno proporzionale in termini di informazione raccolta (Dawes, 2008).

La sezione C – *Comprensione del funzionamento dei sistemi AI* corrisponde alla dimensione di Comprensione Tecnica (modello TUCAPA di Laupichler et al., 2023) e si compone di sei item che indagano la capacità auto-percepita di spiegare gli aspetti principali del funzionamento dei chatbot LLM, senza entrare in dettagli eccessivamente tecnici: funzionamento dei chatbot, meccanismo statistico generativo, nozione di Large Language Model, fenomeno delle allucinazioni, differenza tra IA generativa e software tradizionali, dati di addestramento.

La sezione D – *Valutazione critica degli output* si rifà alla dimensione di Valutazione Critica (modello TUCAPA di Laupichler et al., 2023) e contiene sei item incentrati sulle capacità di riconoscimento dei limiti e di valutazione dell'affidabilità delle risposte: identificazione di informazioni inesatte, incomplete o inventate, valutazione dell'affidabilità prima dell'uso, identificazione dei rischi in ambiti decisionali sensibili, riconoscimento dei bias, distinzione tra fluidità linguistica apparente e accuratezza informativa, valutazione del contesto d'uso appropriato per un chatbot basato su IA generativa.

La sezione E – *Uso pratico dei chatbot* corrisponde alla dimensione di Applicazione Pratica (modello TUCAPA di Laupichler et al., 2023) ed è costituita da sette item che indagano le competenze d'uso operative dei rispondenti. Cinque item adottano il formato standard di capacità percepita ("sono in grado di…") e riguardano: formulazione di prompt efficaci, verifica delle informazioni attraverso altre fonti, capacità di scelta tra chatbot e altre fonti, riconoscimento delle situazioni in cui si entra in contatto con sistemi AI, capacità divulgativa nei confronti di altre persone. I restanti due item della sezione sono formulati in chiave comportamentale ("quando uso un chatbot per informazioni importanti, controllo altre fonti" e "utilizzo chatbot per chiedere consigli su argomenti intimi o personali"), per rilevare il comportamento effettivamente messo in atto. La compresenza dei due formati consente, in fase di analisi, di confrontare la capacità auto-percepita di verificare le informazioni con la pratica dichiarata di farlo, in linea con la riflessione di Rheu e Cho (2025) sul paradosso dell'AI literacy.

La sezione F – Consapevolezza etica e sociale, integrata sulla base di Wang et al. (2023) e Ng et al. (2021), si compone di cinque item che indagano la consapevolezza delle implicazioni etiche legate all'IA generativa: privacy dei dati condivisi, rischi etici dell'uso diffuso, responsabilità personale nell'uso dei contenuti generati, importanza di uno sviluppo e di un utilizzo responsabili, riconoscimento di situazioni in cui l'uso di un chatbot può essere rischioso o dannoso.

3.1.5 Sezione G – Verifica delle conoscenze

La sezione conclusiva introduce un cambio di registro metodologico: alle scale Likert di auto-valutazione, infatti, si sostituiscono sei item in formato vero/falso che hanno lo scopo di verificare la conoscenza fattuale del rispondente. La scelta, ispirata a SAIL4ALL (Soto-Sanfiel et al., 2024), risponde all'esigenza di affiancare a un'indagine della competenza

percepita una misura di competenza effettiva, in modo da rilevare eventuali discrepanze tra percezione e uso (Rheu e Cho, 2025).
Gli item di questa sezione coprono i nodi concettuali centrali emersi nel Capitolo 2: il meccanismo statistico di generazione, la riproduzione di bias appresi dai dati di addestramento, la possibilità che vengano prodotte risposte fluide ma false, il presunto accesso a informazioni aggiornate e verificate, i rischi per la privacy nella condivisione di dati sensibili, l'influenza del tono linguistico sulla fiducia dell'utente. Per ciascun item è definita una risposta corretta sulla base della letteratura scientifica di riferimento. Il numero di risposte corrette è considerato un indicatore di conoscenza effettiva sui chatbot LLM.

3.2 Il chatbot divulgativo per l'AI literacy

Accanto al questionario, la seconda componente della ricerca consiste nello sviluppo di un chatbot pensato come strumento di alfabetizzazione divulgativa sul tema. Lo scopo del chatbot è offrire al pubblico target uno spazio in cui indagare in modo interattivo i temi discussi nel Capitolo 2: il funzionamento dei LLM, i rischi associati al loro uso, le strategie per un'interazione consapevole. I paragrafi seguenti descrivono le motivazioni delle scelte progettuali (3.2.1), l'architettura tecnica del sistema (3.2.2), la costruzione della knowledge base (KB – 3.2.3), il design dell'interazione (3.2.4) e il protocollo di valutazione (3.2.5).

3.2.1 Premesse di progetto: perché un chatbot RAG in locale

La scelta di sviluppare un chatbot come strumento di alfabetizzazione risponde a esigenze di coerenza tra forma e contenuto. L'oggetto di ricerca, infatti, è proprio la competenza degli utenti rispetto ai chatbot LLM: utilizzare lo stesso medium per fini di alfabetizzazione consente la familiarizzazione dell'utente con la modalità di interazione tipica di quei sistemi, e al tempo stesso ne mostra in atto potenzialità e limiti, in un percorso che si potrebbe definire di metaconsapevolezza guidata. La forma del chatbot, inoltre, presenta un grado di accessibilità superiore rispetto a un manuale o a un sito informativo, consentendo all'utente di ottenere informazioni formulando domande e ricevendo risposte su argomenti specifici in modalità interattiva, senza dover navigare una struttura predefinita.

Il chatbot è stato progettato secondo un'architettura di tipo *Retrieval-Augmented Generation* (RAG). Questo approccio, introdotto da Lewis et al. (2020), affianca a un LLM un meccanismo di recupero di informazioni da una base di conoscenza esterna (*knowledge base* – KB): prima di generare la risposta, il sistema cerca all'interno della KB i passaggi più pertinenti rispetto alla domanda dell'utente e li fornisce al modello come contesto su cui fondare la risposta. L'architettura RAG, infatti, opera integrando due tipi distinti di conoscenza, la conoscenza parametrica e quella non parametrica (Lewis et al., 2020). La prima è quella interiorizzata nei pesi del modello durante la fase di addestramento e accessibile in modo implicito durante la generazione, la seconda è quella contenuta nella KB esterna e recuperata in modo esplicito al momento della query. La risposta finale nasce dall'integrazione tra le due: il modello mantiene la propria capacità di produrre linguaggio fluido e ben formato, ma vincola il contenuto a passaggi documentali verificabili. Rispetto a un chatbot basato sulla sola conoscenza parametrica, l'architettura RAG presenta tre vantaggi rilevanti per il caso in oggetto (Gao et al., 2023). In primo luogo riduce la frequenza di allucinazioni, ancorando la risposta a passaggi di documenti verificabili invece che alla semplice estrapolazione statistica del modello; inoltre, consente di tracciare la provenienza di ogni risposta, restituendo all'utente le fonti utilizzate; infine, permette di aggiornare o sostituire la KB senza eseguire nuovamente l'addestramento del modello, garantendo un controllo diretto non solo sugli output del chatbot, ma anche sulle fonti informative su cui essi si basano.

Un'ulteriore scelta di progetto riguarda l'esecuzione locale del sistema, tramite la piattaforma Ollama[21], in alternativa al ricorso a servizi commerciali accessibili via API (OpenAI, Anthropic, Google). La scelta è legata a motivazioni di natura diversa. Dal punto di vista tecnico, un sistema locale consente di mantenere il pieno controllo sui dati delle conversazioni, che non vengono trasmessi a server esterni. Inoltre, il fatto che il sistema sia basato su modelli *open-weight*[22], scaricabili e installabili in autonomia, ne garantisce l'autonomia di esecuzione e rende possibili modifiche (fine-tuning) da parte di altri ricercatori senza dipendere da servizi commerciali, dai loro termini d'uso o da costi variabili nel tempo.

[21]Ollama, https://ollama.com/ (ultima consultazione: giugno 2026).

[22] Con "open-weight" si indicano i modelli i cui pesi sono pubblicamente accessibili, in contrapposizione ai modelli "open-source" in senso stretto, in cui sono rilasciati anche codice e dati di addestramento (https://opensource.org/ai/open-weights – ultima consultazione: giugno 2026).

Tuttavia, va notato che la condivisione dei soli pesi, senza il codice e i dati di addestramento, limita la piena riproducibilità del processo di creazione del modello stesso[23].

3.2.2 Architettura del sistema

Il chatbot è stato sviluppato in linguaggio Python e si articola lungo una pipeline che integra cinque componenti principali: una knowledge base documentale, un modello di embedding, un archivio vettoriale *(vector store*), un LLM e un'interfaccia utente. La configurazione qui descritta rappresenta la versione finale del sistema, raggiunta dopo una fase preliminare di test interno che ha portato alla revisione di alcune scelte iniziali, come illustrato nei paragrafi seguenti.

3.2.2.1 La knowledge base

La base di conoscenza del chatbot è costituita da nove documenti scritti appositamente per il progetto e organizzati in due gruppi (sei documenti sono dedicati ad altrettanti rischi associati ai chatbot LLM e tre documenti sono, invece, complementari). I contenuti rielaborano la letteratura discussa nel Capitolo 2 in forma divulgativa, calibrata sul pubblico target. Una descrizione più estesa della struttura, della costruzione e dei criteri redazionali della KB seguirà nel paragrafo 3.2.3.

3.2.2.2 Il modello di embedding

Per consentire al sistema di confrontare semanticamente la domanda dell'utente con i passaggi della KB, ciascun documento viene preliminarmente trasformato in una sequenza di embedding (cfr. Sezione 2.1.2). Il calcolo è affidato a un modello dedicato, eseguito in locale tramite Ollama.

Il modello adottato è BGE-M3 (Chen et al., 2024), modello open-source multilingue progettato per oltre cento lingue, fra cui l'italiano, che produce vettori a 1024 dimensioni e

[23] https://opensource.org/ai/open-weights (ultima consultazione: giugno 2026).

supporta documenti di contesto esteso[24]. La scelta è ricaduta su un modello esplicitamente multilingue per garantire una qualità di retrieval adeguata su testi in italiano, lingua per la quale i modelli di embedding inglese-first risultano meno performanti (Chen et al., 2024). In una fase preliminare di sviluppo era stato adottato il modello nomic-embed-text (Nussbaum et al., 2024), addestrato prevalentemente su corpus in inglese, le cui prestazioni di retrieval sono risultate non adeguate sui contenuti italiani della KB.

3.2.2.3 L'archivio vettoriale

In linea con il pattern di indicizzazione descritto da Gao et al. (2023), la KB viene sottoposta a una fase di pre-elaborazione articolata in tre passaggi: i documenti vengono dapprima divisi in passaggi di dimensioni ridotte (chunking, cfr. Sezione 3.2.3 per i criteri di granularità adottati); ciascun chunk viene trasformato in un embedding tramite il modello descritto nel paragrafo precedente; gli embedding ottenuti vengono infine memorizzati in un archivio vettoriale (*vector store*), una base dati ottimizzata per la ricerca per similarità. Quando il sistema riceve una domanda dall'utente, anche la domanda viene trasformata in un embedding utilizzando lo stesso modello di codifica adottato per i chunk. La similarità tra il vettore della domanda e i vettori dei chunk viene calcolata e i passaggi i cui vettori risultano più vicini vengono restituiti come potenzialmente più rilevanti per la generazione della risposta. Per la presente implementazione si è utilizzato ChromaDB[25], archivio vettoriale open-source compatibile con il framework LangChain[26], scelto per la semplicità di configurazione e la possibilità di esecuzione interamente locale.

3.2.2.4 Il Large Language Model

La generazione effettiva delle risposte è affidata a un LLM. Per il prototipo descritto si è scelto LLaMA 3.1 (versione Instruct a 8 miliardi di parametri), modello open-weight rilasciato da Meta AI a luglio 2024 (Grattafiori et al., 2024), eseguito in locale tramite Ollama. Il modello, addestrato su corpora di portata multilingue, è stato sottoposto a una fase

[24] Il modello supporta sequenze di input fino a 8192 token (Chen et al., 2024).

[25] ChromaDB, https://www.trychroma.com (ultima consultazione: giugno 2026).

[26] LangChain, https://www.langchain.com (ultima consultazione: giugno 2026).

di post-training basata su Supervised Fine-Tuning (SFT) e Direct Preference Optimization (DPO)[27], che ne ha rafforzato la capacità di seguire istruzioni esplicite, caratteristica rilevante in un contesto come quello del presente lavoro, in cui l'aderenza ai vincoli formulati nel system prompt è centrale. La scelta di un modello di taglia contenuta risponde alla necessità di garantire la possibilità di esecuzione su un hardware non specialistico e all'esigenza di mantenere ridotti i tempi di risposta (Jiang et al., 2023). Un modello di taglia maggiore (come la versione a 70 o 405 miliardi di parametri della stessa famiglia Llama 3) avrebbe presumibilmente prodotto risposte più articolate e meglio aderenti alle istruzioni del system prompt, ma a costo di tempi di inferenza incompatibili con un uso interattivo su macchine standard (Grattafiori et al., 2024).

In una fase preliminare di sviluppo era stato utilizzato Mistral 7B Instruct (Jiang et al., 2023), modello open-weight di taglia comparabile rilasciato dall'azienda francese Mistral AI. A seguito di test interni in cui il modello ha mostrato limitata aderenza ai vincoli espressi nel system prompt e una tendenza a inserire nelle risposte contenuti tematicamente vicini ma non pertinenti alla domanda posta, si è scelto di sostituirlo con Llama 3.1 Instruct, che ha mostrato un comportamento più coerente ai vincoli previsti.

3.2.2.5 L'interfaccia utente

L'interazione con il chatbot avviene attraverso un'interfaccia web sviluppata con Streamlit[28], framework di Python pensato per il prototipaggio rapido di applicazioni con componenti di IA. L'interfaccia presenta un campo di input testuale per la domanda, l'area di visualizzazione della cronologia conversazionale e, per ciascuna risposta, un riquadro espandibile che mostra le fonti della KB utilizzate nella generazione.

3.2.2.6 Il flusso di una singola interazione

Il flusso di elaborazione di una singola domanda riprende lo schema *advanced RAG* descritto da Gao et al. (2023), articolato in più fasi di pre- e post-retrieval. La domanda dell'utente

[27] DPO (Rafailov et al., 2023) è stato adottato in alternativa ad algoritmi di RLHF più complessi per ragioni di stabilità e scalabilità (Grattafiori et al., 2024).

[28] Streamlit, https://streamlit.io (ultima consultazione: giugno 2026).

viene, innanzitutto, espansa in una serie di varianti tematiche (*query variants*), ottenute concatenando alla formulazione originale alcune parole-chiave legate ai principali temi della KB. Questa procedura, riconducibile alle tecniche di *query expansion* (Carpineto e Romano, 2012), risponde al cosiddetto *vocabulary problem*, per cui è possibile che una query formulata in poche parole non modelli in modo adeguato il bisogno informativo dell'utente e il disallineamento tra il linguaggio della domanda e quello dei documenti indicizzati comporti, di conseguenza, un calo della pertinenza del retrieval. L'espansione tramite parole-chiave tematiche aumenta la probabilità di recupero dei passaggi rilevanti anche quando la formulazione della domanda diverge a livello lessicale dalla terminologia dei documenti. Ciascuna variante viene trasformata in embedding e confrontata con l'archivio vettoriale. I passaggi recuperati vengono sottoposti a un riordinamento basato sulla sovrapposizione lessicale (*reranking*) e a un filtro che limita il numero di passaggi per documento, ispirato al criterio di *Maximal Marginal Relevance* (Carbonell e Goldstein, 1998), in modo da garantire la varietà delle fonti e ridurre la ridondanza informativa nel contesto trasmesso al modello generativo (Gao et al., 2023). Un meccanismo di controllo finale, denominato *gating* nel codice del sistema, verifica poi che almeno un passaggio selezionato contenga parole-chiave significative della domanda; in caso contrario, il sistema rifiuta di rispondere e invita l'utente a riformulare la richiesta. Questo meccanismo realizza in forma operativa il principio della *negative rejection* (Gao et al., 2023), ovvero la capacità del sistema di segnalare esplicitamente l'assenza di informazioni pertinenti, anziché produrre una risposta non grounded. La negative rejection è una caratteristica fondamentale dei sistemi RAG e contribuisce in modo diretto alla riduzione delle allucinazioni.

I passaggi superstiti vengono quindi assemblati in un blocco di contesto e trasmessi al modello generativo insieme a un system prompt che specifica i vincoli di risposta – aderenza esclusiva al contesto fornito, divieto di introdurre informazioni esterne, vincoli di registro e lunghezza. Le scelte specifiche di formulazione del system prompt sono discusse nel paragrafo 3.2.4. La risposta generata, infine, viene restituita all'utente attraverso l'interfaccia, accompagnata dalla lista dei documenti della KB fonte dei passaggi utilizzati – i soli titoli, con funzione di tracciabilità, senza accesso al contenuto integrale dei documenti che compongono la KB.

3.2.3 Costruzione della *knowledge base*

La knowledge base del chatbot divulgativo è stata costruita secondo criteri redazionali pensati per il pubblico target del progetto, adulti italiani di età compresa tra 18 e oltre 50 anni, non specialisti del settore. Si compone di nove documenti tematici in formato `.docx`, organizzati in due gruppi distinti per funzione. Il primo gruppo raccoglie sei documenti dedicati ai principali rischi associati all'uso dei chatbot basati su LLM, ciascuno dei quali rielabora in forma divulgativa la letteratura discussa nel Capitolo 2: `01_hallucinations` (meccanismo, esempi tipici, strategie di riconoscimento), `02_bias` (origine nei dati di addestramento, tipologie e implicazioni), `03_fiducia` (*over-reliance*, *automation bias*, calibrazione della fiducia), `04_privacy` (trattamento dei dati nelle interazioni, opzioni di opt-out, raccomandazioni operative), `05_copyright` (addestramento su materiale protetto, fenomeno della memorizzazione, quadro normativo europeo), `06_sycophancy` (tendenza all'assecondamento dell'utente, ruolo del post-training, riconoscimento e strategie di mitigazione). Il secondo gruppo, complementare, è costituito da tre documenti di natura trasversale: `07_basi_llm` (sintesi divulgativa del funzionamento dei LLM, dai *token* al pre-training, fino al post-training); `08_glossario` (voci brevi dei termini tecnici ricorrenti, da LLM a RAG, da embedding a allucinazione); `09_faq_trasversale` (quindici domande comuni che attraversano più temi, dall'uso quotidiano alla calibrazione della fiducia, dalla privacy alle decisioni in ambiti delicati). Stilisticamente tutti i documenti adottano un registro discorsivo amichevole, coerente con il pubblico target individuato per il questionario. La scelta linguistica privilegia il "tu" allocutivo informale e un lessico semplice e quotidiano, in cui i tecnicismi sono introdotti solo se necessari e definiti al primo uso. Ciascun documento è articolato in titoli e paragrafi tematici autonomi che affrontano ognuno un singolo aspetto del tema, di modo che i singoli paragrafi possano essere recuperati in isolamento dal sistema di retrieval senza perdita di coerenza.

Sul piano della granularità, ciascun chunk testuale è stato calibrato per avere una lunghezza compresa tra 430 e 540 caratteri. Questa scelta è funzionale al parametro `chunk_size` = 500 (espresso in caratteri) del text splitter[29] utilizzato nella fase di indicizzazione (cfr. Sezione 3.2.2) e si configura come una forma di *chunk optimization* (Gao et al., 2023), tecnica di Advanced RAG finalizzata a calibrare la granularità dei chunk – un parametro critico per la qualità del retrieval. Chunk di dimensioni lievemente inferiori al limite, infatti, garantiscono la conservazione del contesto semantico (*context retention*, Gao et al., 2023), permettendo al

[29] `RecursiveCharacterTextSplitter` della libreria LangChain (sottopacchetto `langchain_text_splitters`), https://www.langchain.com

sistema di recuperare ogni paragrafo come unità semantica coerente, senza che la spezzatura automatica introdotta dal text splitter ne comprometta la leggibilità o la coesione. Una calibrazione manuale della lunghezza dei chunk in fase di scrittura della knowledge base ha consentito di evitare l'intervento automatico del text splitter, che in caso di chunk eccessivamente lunghi avrebbe operato tagli non controllati. Inoltre, il text splitter è configurato con un parametro `chunk_overlap` di 80 caratteri, di modo tale che, nei rari punti in cui la spezzatura automatica intervenisse, i due chunk risultanti condividerebbero 80 caratteri al confine, garantendo continuità informativa tra segmenti contigui.
La KB, comunque, va intesa come una sintesi e non una trattazione esaustiva: la copertura tematica è circoscritta agli aspetti ritenuti centrali per un primo livello di alfabetizzazione di utenti privi di un background tecnico e teorico sul tema.

3.2.4 Il design dell'interazione

Il design dell'interazione tra l'utente e il chatbot è regolato dal system prompt: un blocco testuale che precede ogni chiamata al modello generativo e ne vincola il comportamento prima ancora che la domanda dell'utente venga elaborata. Nel paradigma RAG, il system prompt costituisce lo strumento principale attraverso cui chi progetta trasmette al modello i vincoli di aderenza al contesto recuperato e le specifiche stilistiche della risposta (Gao et al., 2023). La capacità del modello di rispettare queste istruzioni dipende dalla fase di post-training cui è stato sottoposto: nel caso di LLaMA 3.1 Instruct, il Supervised Fine-Tuning e la Direct Preference Optimization (cfr. Sezione 3.2.2) hanno reso il modello particolarmente ricettivo a istruzioni esplicite, rendendolo adatto a un uso in cui il system prompt svolge un ruolo centrale (Grattafiori et al., 2024). La scelta di formulare il prompt attraverso regole esplicite e numerate, anziché principi generali da elaborare autonomamente, è coerente con la taglia del modello adottato. Wei et al. (2022) mostrano che capacità di ragionamento più complesse emergono in modo affidabile solo a partire da determinate soglie di scala: un modello di otto miliardi di parametri beneficia di istruzioni operative formalizzate, che riducono l'ambiguità interpretativa e limitano i comportamenti non previsti.

Il system prompt finale è articolato in due blocchi. Il primo definisce cinque regole vincolanti sul contenuto e sul grounding della risposta. La prima regola impone l'uso esclusivo delle informazioni presenti nel contesto trasmesso – i chunk recuperati dalla knowledge base – con divieto esplicito di integrare conoscenza parametrica del modello. La

seconda regola vieta l'invenzione di nomi propri, riferimenti normativi, dati numerici e percentuali non presenti nei documenti: è stata formulata in risposta a un comportamento osservato empiricamente nelle sessioni di test, in cui il modello tendeva a integrare risposte corrette con dettagli plausibili ma inventati, un pattern riconducibile al fenomeno delle allucinazioni discusso nella Sezione 2.2.1. La terza regola definisce il comportamento del sistema in presenza di una domanda non coperta dalla knowledge base: in tal caso il modello è istruito a rispondere con la formula fissa "Su questo i documenti non dicono di più", evitando la generazione di risposte non ancorate ai documenti della KB. Questa scelta è coerente con la riflessione sull'*over-reliance* sviluppata nel Capitolo 2 (cfr. Sezione 2.2.3): un sistema che ammette esplicitamente i propri limiti contribuisce alla calibrazione della fiducia dell'utente, invece di produrre risposte fluide ma non fondate. La quarta regola impone la focalizzazione sulla domanda posta, scoraggiando le divagazioni verso temi adiacenti: una criticità emersa nelle fasi di test, in cui il modello tendeva a espandere la risposta verso chunk tematicamente vicini ma non pertinenti alla formulazione specifica della domanda. La quinta regola introduce un criterio di lunghezza adattiva – due-cinque frasi per domande chiuse, massimo otto per domande aperte – in risposta alla tendenza dei modelli addestrati con feedback umano a privilegiare risposte elaborate anche quando la domanda non lo richiederebbe (Ouyang et al., 2022).

Il secondo blocco del system prompt disciplina il registro linguistico delle risposte. Il modello è istruito a rivolgersi all'utente con il "tu", coerentemente con il registro divulgativo adottato nella knowledge base e con il pubblico target della ricerca. Sono escluse le formule di apertura convenzionali e una serie di costruzioni ricorrenti – "è importante", "è bene", "vale la pena", "in conclusione", "inoltre" – identificate come marker tipici di un registro artificialmente compiacente. Per le domande chiuse, il modello è istruito ad aprire la risposta direttamente con la risposta effettiva, evitando preamboli che ritardano l'informazione richiesta. Sono scoraggiate le strutture in elenchi puntati quando la prosa è sufficiente e le chiusure aforistiche che generalizzano oltre la domanda posta. La motivazione di queste scelte è legata al tema della fiducia: la fluidità formale di una risposta tende a generare nell'utente una percezione di credibilità non sempre giustificata dalla qualità informativa del contenuto (Kim et al., 2024; Ferrario et al., 2020), e i marker linguistici tipicamente associati agli output automatici possono interferire con la corretta calibrazione di tale fiducia.

Il system prompt descritto non è il risultato di un atto di progettazione unico, ma l'esito di un processo iterativo di osservazione e affinamento. Una versione iniziale, più sintetica, è stata progressivamente arricchita sulla base dei comportamenti osservati

empiricamente durante le sessioni di test interni. I due interventi più significativi hanno riguardato la tendenza del modello a integrare nella risposta dettagli inventati ma plausibili – che ha portato all'introduzione della regola 2 – e la tendenza a costruire risposte composte da informazioni provenienti da chunk tematicamente eterogenei – che ha portato all'introduzione della regola 4 e a meccanismi di filtraggio a livello di retrieval. Non si documenta qui l'intera sequenza delle versioni intermedie: l'iterazione tra formulazione del prompt e osservazione del comportamento del sistema costituisce, in questo contesto, una forma di calibrazione empirica del design dell'interazione, complementare alle scelte fatte a livello architetturale. Il passaggio dal modello Mistral 7B a LLaMA 3.1 (cfr. Sezione 3.2.2), inoltre, ha comportato una nuova fase di verifica: il comportamento del sistema rispetto alle regole del prompt si è confermato sostanzialmente stabile, senza richiedere modifiche strutturali al design dell'interazione.

3.2.5 Protocollo di valutazione del chatbot

La valutazione del chatbot è avvenuta tramite un test pilota che ha coinvolto tre partecipanti adulti, non professionisti del settore tecnologico, reclutati tramite contatti personali. Le fasce generazionali rappresentate – 21, 29 e 64 anni – corrispondono alle principali classi d'età rappresentate nel campione principale (18-25, 26-35, 50+). La selezione di fasce anagrafiche differenziate risponde all'esigenza di rilevare eventuali variazioni nell'esperienza d'uso riconducibili all'età, pur nella limitazione numerica propria di uno studio pilota. Følstad e Brandtzæg (2020), infatti, mostrano come l'età influenzi i driver dell'esperienza utente con i chatbot, con i partecipanti più giovani orientati verso attributi edonici (intrattenimento, novità) e quelli più anziani verso aspetti pragmatici (efficienza, assistenza). Includere fasce generazionali distinte consente quindi di rilevare, anche in forma esplorativa, variazioni di questo tipo.

Le sessioni si sono svolte in presenza. A ciascun partecipante è stato chiesto di interagire liberamente con il chatbot per una durata indicativa di 30 minuti, senza compiti predefiniti né domande guida. Questa scelta favorisce quella che Haugeland et al. (2022) definiscono conversazione *topic-led* (guidata dall'argomento) rispetto a quella *task-led* (guidata dal compito), avvicinando l'interazione alle condizioni reali d'uso. Permettere all'utente di esprimersi con le proprie parole senza domande guida consente inoltre, come

indicato da Følstad e Brandtzæg (2020), di far emergere gli attributi del sistema realmente salienti per l'utente, riducendo il bias dello sperimentatore, che, proponendo delle domande, potrebbe orientare artificialmente il comportamento dell'utente (Wollny et al., 2021). Prima e dopo la sessione, i partecipanti hanno compilato il questionario di valutazione descritto nella sezione 3.3.

3.3 Il questionario di valutazione del chatbot

La terza componente del disegno di ricerca riguarda la valutazione esplorativa del chatbot, condotta tramite un test pilota su un campione di tre utenti, selezionati tra adulti non professionisti del settore informatico e rappresentanti di tre fasce d'età diverse. I tre partecipanti avevano già compilato il questionario sull'AI *literacy* descritto nella Sezione 3.1: si tratta dunque di utenti già esposti ai temi dell'indagine, condizione di cui occorre tenere conto nell'interpretare le loro reazioni al prototipo. Gli utenti hanno interagito con il prototipo e compilato un questionario strutturato, con l'obiettivo di raccogliere osservazioni qualitative su usabilità, affidabilità delle risposte e qualità linguistica. Dato l'esiguo numero di partecipanti, i risultati non sono statisticamente generalizzabili; lo strumento ha piuttosto una funzione orientativa, utile a identificare criticità preliminari e a raccogliere osservazioni qualitative sull'esperienza d'uso.

3.3.1 Struttura del questionario

Il questionario è strutturato in due parti somministrate in sequenza: una sezione preliminare (Parte A), compilata prima dell'interazione con il chatbot, e una sezione valutativa (Parte B), compilata immediatamente dopo. Questa scansione pre/post consente di rilevare le aspettative iniziali dell'utente e di confrontarle con l'esperienza effettiva, mettendo in luce eventuali scostamenti tra percezione attesa e percezione osservata.

La Parte A raccoglie informazioni sulla frequenza d'uso di sistemi di intelligenza artificiale conversazionale, sul livello di fiducia iniziale nel chatbot da valutare e sulle aspettative in termini di utilità e facilità d'uso. La Parte B comprende quattordici item a scala Likert (1-5, da "per niente d'accordo" a "completamente d'accordo") distribuiti in cinque

sezioni tematiche, a cui sono state frapposte cinque domande aperte per raccogliere osservazioni qualitative in corrispondenza di ciascuna dimensione.

3.3.2 Dimensioni valutative e quadro teorico

Le dimensioni indagate nella Parte B sono state selezionate sulla base della letteratura consolidata sulla valutazione di sistemi conversazionali e chatbot educativi, con l'obiettivo di coprire tanto gli aspetti funzionali quanto quelli linguistici e relazionali dell'interazione. Gli item completi sono riportati nella Sezione 6.3.

Gli item dal B1 al B3 indagano in quale misura il chatbot è percepito come uno strumento utile per apprendere e per chiarire dubbi sull'IA. Gli item si basano sul costrutto di utilità percepita del *Technology Acceptance Model*[30] (Davis, 1989) e sui criteri di efficacia educativa proposti da Wollny et al. (2021) nella loro analisi sistematica degli agenti conversazionali in contesti di apprendimento.

La seconda sezione, che comprende gli item B4, B5 e B6, si concentra sulla dimensione della facilità d'uso. Questa dimensione rileva la naturalezza dell'interazione, la chiarezza dell'interfaccia testuale e l'assenza di difficoltà tecniche percepite. Gli item si ispirano alla System Usability Scale (Brooke, 1996) e al costrutto di facilità d'uso percepita del TAM (Davis, 1989). L'item B6, formulato in senso inverso, è contrassegnato come item rovesciato e richiede inversione del punteggio nella fase di analisi.

La sezione successiva, che comprende gli item dal B7 al B9**,** esplora il grado di fiducia dell'utente nella correttezza, nell'accuratezza e nella trasparenza delle risposte fornite dal chatbot. Il framework teorico di riferimento è quello di Nordheim et al. (2019), che applicano al contesto dei chatbot il modello classico della fiducia interpersonale di Mayer et al. (1995), che individua nella competenza (*ability*), nell'integrità (*integrity*) e nella benevolenza (*benevolence*) le dimensioni costitutive dell'affidabilità percepita (*trustworthiness*). Applicate a un sistema conversazionale, queste dimensioni riguardano rispettivamente la percezione che il chatbot risponda in modo accurato, che non manipoli l'utente e che agisca nel suo interesse.

[30] Il *Technology Acceptance Model* (TAM), proposto da Davis (1989), è un modello teorico sviluppato nell'ambito dei sistemi informativi per spiegare i fattori che determinano l'accettazione e l'adozione di una nuova tecnologia. Il modello individua due costrutti principali – l'utilità percepita (*perceived usefulness*) e la facilità d'uso percepita (*perceived ease of use*) – come predittori dell'intenzione d'uso e del comportamento effettivo nei confronti del sistema.

Gli item dal B10 al B12 indagano la dimensione della qualità linguistica. Questa dimensione valuta la chiarezza, la naturalezza e la qualità formale della lingua prodotta dal chatbot. Gli item si fondano sulle scale di qualità pragmatica e qualità edonica derivate dall'approccio AttrakDiff[31] (Haugeland et al., 2022) e sul framework di esperienza utente nei sistemi conversazionali proposto da Følstad e Brandtzæg (2020).

La sezione conclusiva raccoglie una valutazione globale del sistema in termini di soddisfazione d'uso e di intenzione a utilizzarlo nuovamente o a consigliarlo. Gli item si rifanno ai criteri di valutazione complessiva proposti da Wollny et al. (2021).

[31] Nel framework AttrakDiff, sviluppato originariamente da Hassenzahl et al. (2003) e ripreso e adattato ai sistemi conversazionali da Haugeland et al. (2022), la qualità pragmatica riguarda l'efficacia funzionale dell'interazione (quanto il sistema aiuta l'utente a raggiungere i propri obiettivi), mentre la qualità edonica riguarda la dimensione esperienziale e percettiva (piacevolezza, naturalezza, stimolazione).

4. Analisi dei dati

Questo capitolo riporta l'analisi descrittiva dei dati raccolti attraverso il questionario sull'AI literacy, somministrato a un campione di novantadue rispondenti (n = 92), tra l'8 e il 22 maggio 2026. L'analisi procede dalla descrizione del campione alla rilevazione delle conoscenze, delle competenze e dei comportamenti dichiarati, per concludersi con le analisi incrociate per età e titolo di studio e con una sintesi dei pattern principali. Trattandosi di un campione di convenienza raccolto tramite campionamento a valanga (cfr. Sezione 3.1), i risultati hanno valore esplorativo e non sono generalizzabili alla popolazione adulta italiana nel suo complesso. Per non appesantire la trattazione, nel testo si riportano i valori più rilevanti; le distribuzioni complete, item per item, sono raccolte in appendice (cfr. Sezione 6.1). All'analisi descrittiva fa seguito, nell'ultima parte del capitolo, una discussione che interpreta i risultati alla luce del quadro teorico delineato nel Capitolo 2.

4.1 Profilo del campione

Il campione è composto da novantadue persone adulte. La distribuzione per genere è fortemente sbilanciata: le donne rappresentano l'80,4% dei rispondenti (n = 74), contro il 18,5% di uomini (n = 17) e un'unica persona che si identifica come non binaria o di altro genere. Questo squilibrio impone cautela: ogni interpretazione che chiami in causa variabili genere-dipendenti risulterebbe poco fondata e non viene quindi sviluppata.

Sul piano anagrafico, il campione si concentra agli estremi della distribuzione. La fascia 18-25 anni è la più rappresentata (40,2%, n = 37), seguita dalla fascia 50 anni o più (27,2%, n = 25) e dalla 26-35 (23,9%, n = 22). Le fasce centrali sono quasi vuote: data la numerosità esigua, la fascia 36-45 (n = 3) non viene commentata in nessuna delle analisi successive, e i risultati della 46-50 (n = 5) sono riportati solo a titolo indicativo.

Il livello di istruzione è mediamente elevato: il 41,3% (n = 38) dei rispondenti possiede una laurea magistrale o un titolo superiore e quasi i due terzi del campione (65,2%, n = 60) hanno almeno una laurea triennale. Anche in questo caso la categoria meno numerosa, la licenza media (n = 3), non è interpretabile. Quanto all'ambito professionale, la componente

studentesca è la più ampia (28,3%, n = 26), seguita dal settore sanitario (23,9%, n = 22) e dall'istruzione e ricerca (13,0%, n = 12). La presenza marcata di sanità e istruzione non è priva di interesse: si tratta, infatti, di contesti in cui un uso poco consapevole dei sistemi generativi per compiti decisionali o informativi ha ricadute pratiche dirette. Gli ambiti tecnico-informatici sono invece minoritari (7,6%, n = 7), un dato che concorre a delineare un campione composto in prevalenza da utenti non specialisti.

4.2 Uso e atteggiamenti verso i chatbot

L'uso dei chatbot è ampiamente diffuso nel campione: il 92,4% (n = 85) dichiara di averne utilizzato almeno uno. Più della metà dei rispondenti (53,3%, n = 49) li impiega quotidianamente o regolarmente, mentre solo il 7,6% (n = 7) non li ha mai usati. L'uso generale è quindi non soltanto diffuso, ma frequente. A questa diffusione non corrisponde però un atteggiamento di fiducia incondizionata. La fiducia complessiva dichiarata si colloca in prevalenza su un livello medio (59,8%, n = 55), mentre solo il 7,6% (n = 7) dichiara una fiducia alta e nessun rispondente la dichiara molto alta in questa voce. Quasi un terzo del campione (28,3%, n = 26, sommando i livelli basso e molto basso) esprime una fiducia ridotta. Emerge così un primo dato rilevante: la maggioranza dei rispondenti usa i chatbot in modo frequente pur non riponendovi una fiducia piena, configurando una forma di uso pragmatico in condizione di incertezza.

Gli scopi d'uso, rilevati con una domanda a risposta multipla (261 menzioni complessive), confermano un impiego orientato all'informazione e allo studio: la ricerca di informazioni è lo scopo più citato (26,8% delle menzioni), seguita da studio e formazione (16,1%) e scrittura di testi (13,8%). Merita attenzione, per le sue implicazioni, la quota di chi dichiara di usare i chatbot per consulenza medica o legale: pur minoritaria (3,8%, pari a dieci menzioni), riguarda un ambito decisionale ad alto rischio, in cui la produzione di contenuti errati ma plausibili – discussa nel Capitolo 2 a proposito delle allucinazioni – può avere conseguenze concrete.

4.3 Conoscenze concettuali dichiarate

Il blocco concettuale del questionario raccoglie sei item che chiedono al rispondente di valutare la propria capacità di spiegare o descrivere il funzionamento dei sistemi generativi, su una scala da 1 a 5. Sono, significativamente, gli item con le medie più basse dell'intero questionario.

L'item con la media più bassa è la capacità di spiegare cosa sia un LLM (M = 2,14, SD = 1,44). La distribuzione è eloquente: il 48,9% (n = 45) sceglie il valore minimo (non sa spiegarlo per niente) e un ulteriore 21,7% (n = 20) sceglie il valore 2, per un totale del 70,6% (n = 65) collocato nella fascia bassa della scala. La conoscenza concettuale dei meccanismi alla base dei sistemi generativi appare dunque molto limitata. Le deviazioni standard elevate (comprese tra 1,35 e 1,47) segnalano inoltre una forte dispersione: una minoranza con competenza tecnica alta – riconducibile soprattutto all'ambito informatico – innalza la media, mentre il valore mediano sarebbe ancora più basso.

4.4 Competenze critiche, pratiche ed etiche

Un secondo gruppo di item chiede non di spiegare, ma di valutare, riconoscere, decidere e fare. Le medie di questo blocco sono sistematicamente più alte di quelle del blocco concettuale, e si collocano in prevalenza tra 3,2 e 4,1.

Il dato più rilevante non è il singolo valore, ma il confronto fra i due blocchi. Le competenze pratiche e critiche (M tra 3,21 e 4,14) superano in modo sistematico le conoscenze concettuali (M tra 2,14 e 3,12), in ogni fascia d'età e per ogni livello di istruzione. I rispondenti, in altre parole, si percepiscono capaci di fare senza sapere spiegare perché. L'item con la media più alta dell'intero questionario è "decidere autonomamente quando usare un chatbot" (M = 4,14, SD = 1,10), che non richiede alcuna comprensione tecnica ma esprime un giudizio di opportunità; l'item più basso del blocco critico è "riconoscere quando i bias influenzano le risposte" (M = 3,21, SD = 1,20), che richiede invece la comprensione di un meccanismo. La competenza etica mostra valori intermedi (3,29-3,83), con i punteggi più alti sugli item più concreti (riconoscere un uso inappropriato) e i più bassi su quelli più astratti (le implicazioni per la privacy).

4.5 Comportamenti dichiarati

Due item rilevano comportamenti d'uso concreti. Il primo riguarda la verifica delle informazioni: alla domanda se, usando un chatbot per informazioni importanti, si controllino altre fonti, la distribuzione è fortemente spostata verso l'alto (M = 4,11, SD = 1,15). La metà esatta del campione (n = 46) risponde "sempre" e oltre tre quarti (76,1%, n = 70) si collocano ai valori 4 o 5; solo l'8,7% (n = 8) si colloca ai valori bassi. Il comportamento di verifica appare quindi molto diffuso, almeno a livello dichiarato. Trattandosi di un dato auto-riferito, è plausibile che la desiderabilità sociale ne abbia gonfiato il valore: la verifica delle fonti è un comportamento socialmente approvato, e il confronto con la fiducia solo "media" rilevata in precedenza suggerisce che chi dichiara di verificare comunque continui a fidarsi in misura apprezzabile.

Il secondo item riguarda l'uso dei chatbot per consigli su argomenti intimi e personali, e registra la media più bassa di tutto il questionario (M = 1,68, SD = 0,98). Il rifiuto di affidare ai chatbot temi sensibili è quasi universale e trasversale a età e titolo di studio, per questo motivo l'item non produce varianza utile nelle analisi incrociate.

4.6 Conoscenza fattuale: la sezione Vero/Falso

La sezione conclusiva del questionario propone sette affermazioni fattuali (item 25-31) su cui esprimersi con Vero o Falso. Sei item su sette registrano tassi di risposta corretta superiori all'80%: un risultato che segnala che la maggior parte delle affermazioni sono risultate probabilmente intuitive per i rispondenti. L'eccezione è però il dato più significativo dell'intera analisi.

L'item 25 è l'unico su cui la maggioranza del campione sbaglia: il 66,3% dei rispondenti (n = 61) crede erroneamente che ChatGPT comprenda il significato delle parole e recuperi informazioni da internet. L'affermazione descrive precisamente ciò che un modello linguistico non fa: il dato individua quindi un fraintendimento di fondo sul meccanismo di funzionamento dei sistemi generativi, diffuso in due terzi del campione. Questo risultato assume particolare interesse se messo a confronto con gli altri item della sezione. Il 98,9% dei rispondenti (n = 91) riconosce che la fiducia in una risposta non dipende solo dalla sua accuratezza ma anche dalla sua fluidità (item 30), e il 95,7% (n = 88) sa che un chatbot può

produrre testo fluido ma falso (item 27). Si delinea così una tensione interna rilevante: la quasi totalità del campione sa che un chatbot può sbagliare pur sembrando convincente, ma due terzi credono al tempo stesso che esso "comprenda" e "cerchi". Le due credenze sono difficilmente compatibili, eppure coesistono. Si può ipotizzare che i rispondenti abbiano acquisito una consapevolezza dei rischi a livello di esito – sono quindi consapevoli che chatbot può sbagliare – senza avere consapevolezza del meccanismo che spiega perché ciò accade.

4.7 Analisi incrociate per età e titolo di studio

Le analisi incrociate mettono in relazione le variabili demografiche con l'errata concezione di fondo (item 25), con la conoscenza concettuale dei LLM e con il comportamento di verifica delle fonti. Si ricorda che le celle con numerosità inferiore a cinque (fascia 36-45 e licenza media) non sono interpretabili e vengono omesse dal commento.

4.7.1 Età, titolo di studio e fraintendimento concettuale

La fascia 26-35 è l'unica in cui la maggioranza risponde correttamente (55%, 12 su 22). La fascia 50 o più registra l'84% di risposte errate (21 su 25), il valore più alto del campione. Di particolare interesse è il dato della fascia 18-25: pur essendo quella che usa i chatbot con maggiore frequenza, sbaglia più della 26-35 (68% per la fascia 18-25, 25 su 37, contro 45% per la fascia 26-35, 10 su 22). La familiarità d'uso non si traduce dunque in comprensione del meccanismo: conoscere uno strumento per averlo usato a lungo è cosa diversa dal comprenderne il funzionamento.

Anche rispetto al titolo di studio il risultato è controintuitivo e non lineare. Il tasso di fraintendimento più alto si registra fra i diplomati (79%, 23 su 29), seguiti dai laureati magistrali (68%, 26 su 38); la laurea triennale è invece il gruppo con la quota di errori più bassa (50%, 11 su 22). Il gradiente non è dunque monotòno, e il titolo di studio più elevato non corrisponde a una comprensione maggiore. Lo stesso andamento emerge dalla conoscenza concettuale auto-valutata: sull'item relativo agli LLM la triennale ($M = 2,73$, $SD = 1,63$) supera la magistrale ($M = 2,32$, $SD = 1,49$), mentre il diploma resta il valore più basso ($M = 1,52$, $SD = 0.93$). L'istruzione superiore non garantisce la comprensione del

funzionamento dei sistemi generativi, che appare indipendente dal livello di istruzione formale e riconducibile semmai a una conoscenza specifica e non scolastica del dominio.

4.7.2 Età, titolo di studio e comportamento di verifica

La fascia 50 o più è l'unica con una quota rilevante di non-verificatori (28%, 7 su 25) e presenta la media di verifica più bassa (M = 3,44, SD = 1,53). All'opposto, i 18-25 dichiarano una verifica altissima – l'84% (31 su 37) ai valori 4-5 della scala, nessuno ai valori più bassi – nonostante la loro bassa comprensione concettuale dei LLM: un comportamento che sembra rispondere a un'abitudine critica generale più che alla comprensione del sistema. Si può ipotizzare che vi contribuisca una più ampia competenza di alfabetizzazione digitale e mediatica, di cui le fasce più giovani dispongono probabilmente in misura maggiore.

A differenza di quanto osservato per il fraintendimento concettuale, qui il titolo di studio discrimina con nettezza: si osserva un gradiente lineare, dal diploma (66%, 19 su 29) alla laurea triennale (73%, 16 su 22) fino alla laurea magistrale (89%, 34 su 38, la quota di verifica più alta del campione). L'istruzione superiore sembra quindi sviluppare l'abitudine alla verifica delle fonti – una competenza trasversale – ma non la comprensione del meccanismo specifico dei modelli generativi (cfr. Sezione 4.7.1): le due dimensioni seguono logiche distinte.

4.7.3 La responsabilità del contenuto generato

Un ultimo dato incrociato merita attenzione. All'item 31 – secondo cui la responsabilità di un contenuto generato da un chatbot ricade sull'utente che lo utilizza (risposta corretta: Vero) – il 18,5% del campione (17 su 92) risponde Falso, negando tale responsabilità. Il dato è controintuitivo nella sua distribuzione anagrafica: sono i più giovani (18-35: 22-23%, ossia 8 su 37 e 5 su 22) a negare la responsabilità più dei 50 o più (12%, 3 su 25). Si può ipotizzare che chi usa i chatbot quotidianamente tenda a percepire l'output come testo della macchina, e non come testo proprio – una forma di dissociazione dall'autorialità che chi li usa meno sembra non sviluppare. Va però segnalato un limite dell'item, la cui formulazione recita: "Quando un utente usa un contenuto generato da un chatbot, la responsabilità di quel contenuto è dell'utente che l'ha generato". L'enunciato è ambiguo perché attribuisce l'azione

del “generare” prima al chatbot (“un contenuto generato da un chatbot”) e poi, con la subordinata “che l’ha generato”, all’utente: non è ben chiaro se “che l’ha generato” si riferisca all’utente o al chatbot nominato poco prima. Chi risponde “Falso” potrebbe quindi farlo non perché ritenga l’utente non responsabile, ma perché ha inteso la frase nel senso “il contenuto è generato dal chatbot, dunque la responsabilità non è dell'utente”. Non è perciò possibile distinguere con certezza una reale credenza sulla responsabilità da un fraintendimento della frase, e il dato va letto con la dovuta prudenza.

4.8 La valutazione del chatbot: il test pilota

L’ultima componente dell’analisi riguarda la valutazione esplorativa del chatbot educativo, condotta attraverso il test pilota descritto in 3.2.5. Vi hanno preso parte tre utenti adulti non esperti del settore tecnologico, appartenenti a tre fasce generazionali distinte – 21, 29 e 64 anni – che hanno interagito liberamente con il prototipo e compilato il questionario di valutazione (cfr. Sezione 3.3) prima e dopo la sessione. Data la numerosità (n = 3), i punteggi della scala Likert vanno intesi come semplice orientamento, privi di valore statistico; il dato primario è costituito dalle risposte aperte (cfr. Sezione 3.3.3). Nel seguito i partecipanti sono identificati in base all’età e alla frequenza d’uso dichiarata: U29 (uso quotidiano), U64 (uso raro) e U21 (uso regolare).

4.8.1 Punti di forza

Il primo elemento che emerge è la valutazione complessiva positiva. Tutti e tre i partecipanti dichiarano che consiglierebbero il chatbot (B14 = 5) e si dicono soddisfatti dell’esperienza (B13 = 4, 4 e 5). Soprattutto, l’obiettivo educativo del sistema risulta raggiunto: tutti indicano che il chatbot li ha aiutati a comprendere meglio almeno uno dei temi trattati (B1 = 4) e che le informazioni fornite sono state utili (B2 = 4). La conferma più diretta è qualitativa: U21 riferisce di aver ottenuto “approfondimenti sui chatbot che mi hanno aiutato a capire meglio il mondo dell’intelligenza artificiale”.

Un secondo punto di forza riguarda l’accessibilità. L’interazione è giudicata semplice e immediata (B4 = 5, 5 e 4) e il linguaggio chiaro (B10 = 5, 4 e 4). Il dato è significativo in rapporto al pubblico target: anche U64, l’utente più anziano e quello che dichiara di usare i

chatbot meno frequentemente, trova l'interazione immediata (B4 = 5). Anche la lunghezza delle risposte è percepita come adeguata (B12 = 4, 4 e 5), un riscontro che sembra convalidare la regola sulla brevità adottata nel system prompt (cfr. Sezione 3.2.4).

4.8.2 Limiti e criticità

La dimensione più debole è l'affidabilità percepita delle risposte. L'item sulle risposte "accurate e affidabili" (B7) riceve da tutti e tre il valore intermedio (3), il punteggio più basso e più omogeneo del questionario.Si nota inoltre che i partecipanti attribuiscono una fiducia leggermente maggiore alle fonti del sistema (B8 = 4) che all'accuratezza delle risposte in sé (B7 = 3): uno scarto che, pur minimo e da verificare su campioni più ampi, sembra indicare un atteggiamento prudente verso l'output.

Il rilievo qualitativamente più forte riguarda la rigidità del sistema sulle domande aperte, un punto su cui U29 insiste in due risposte distinte: descrive il chatbot come "troppo deterministico" e osserva che "non ha ampliato il pensiero alle domande aperte, ma ha cercato di dare una risposta netta"; nei suggerimenti conclusivi aggiunge che il sistema dovrebbe "ampliare i pensieri" anziché fornire risposte chiuse. La stessa criticità affiora, da un altro utente (U64), nella risposta alla domanda se un chatbot possa essere "immorale", giudicata pertinente ma non del tutto puntuale. Si tratta di una conseguenza diretta delle scelte di progettazione descritte nella Sezione 3.2.4 – istruzioni a rispondere solo alla domanda, andando dritto al punto. L'orientamento al *grounding* e al contenimento delle allucinazioni, sulle domande riflessive e astratte, si traduce in una resa percepita come rigida: le stesse scelte che rendono il sistema affidabile ne riducono la flessibilità. È la tensione di fondo principale messa in luce dal test.

Un secondo limite riguarda la robustezza del recupero rispetto alla formulazione della domanda. Due utenti su tre hanno dovuto riformulare la richiesta: U64 riferisce di aver dovuto "riformulare la domanda" e U21 racconta di una "domanda sulle allucinazioni" che "il chatbot non aveva capito". Quest'ultimo caso è particolarmente indicativo, poiché le allucinazioni sono un tema effettivamente coperto dalla KB: il meccanismo di espansione della *query* e il filtro di pertinenza (cfr. Sezione 3.2.4) non hanno intercettato una domanda rilevante. Ne emerge un margine di miglioramento sulla tolleranza alle variazioni lessicali nella

formulazione dell'utente.

Un terzo rilievo riguarda il registro. L'item sulla naturalezza del tono (B11) è il più basso del blocco linguistico (3, 4 e 3). U29 descrive un tono analogo a quello "di un qualsiasi altro chatbot, con lunghezza delle frasi sempre uguale", precisando, però, che "se ne accorgerebbe solo chi usa i chatbot quotidianamente"; U64 lamenta "poca empatia". Anche in questo caso si tratta di un compromesso voluto: la scelta di un registro asciutto e diretto, privo di formule di cortesia (cfr. Sezione 3.2.4) e pensato per contrastare la compiacenza artificiale del modello, viene percepita come uniformità o freddezza. Il rilievo di U64 sulla "poca empatia", inoltre, ha duplice valenza interpretativa. Oltre a confermare che la scelta di un registro non compiacente (cfr. Sezione 3.2.4) è stata colta, rivela un'aspettativa – che il chatbot esibisca calore affettivo – plasmata dai sistemi commerciali, in cui il calore (*warmth*) è uno stile cordiale ed emotivamente accomodante indotto in fase di addestramento e correlato alla *sycophancy* (cfr. Sezione 2.2.6; Ibrahim et al., 2026). Attribuire "empatia" a un sistema generativo, inoltre, costituisce una proiezione affine all'effetto ELIZA (cfr. Sezione 2.1.2): il modello non prova empatia ma ne produce il registro, così come non comprende il significato genera contenuti a cui gli utenti attribuiscono significato. La stessa confusione tra forma di superficie e stato sottostante che governa il fraintendimento centrale (cfr. Sezione 4.6) si manifesta qui in chiave affettiva.

Un ultimo gruppo di osservazioni riguarda i contenuti. Da un lato, la copertura della KB mostra i propri limiti, già dichiarati nella Sezione 3.2.3: la domanda culturalmente articolata di U29 – relativa alla possibilità che i sistemi penalizzino i gruppi meno rappresentati nei testi di addestramento – non ha trovato risposta soddisfacente. Dall'altro, e con implicazioni linguistiche dirette, U64 segnala che il chatbot ha introdotto il termine tecnico "*sycophancy*" senza spiegarlo. Per uno strumento divulgativo rivolto a non esperti, l'impiego di un tecnicismo – per di più in lingua inglese – non chiosato costituisce un limite di accessibilità che contraddice l'impostazione del registro (cfr. Sezione 3.2.4).

4.8.3 Differenze per profilo e sintesi

Pur con tutte le cautele imposte dalla numerosità minima, i tre profili si comportano in modo coerente con quanto ipotizzato nella Sezione 3.2.5 sulla scorta di Følstad e Brandtzæg (2020). L'utente con uso quotidiano (U29) è il più critico, in particolare sugli aspetti edonici e stilistici

– il tono, la rigidità – e ne è consapevole quando afferma, riferendosi a certe sfumature, che "se ne accorgerebbe solo chi usa i chatbot quotidianamente". L'utente con uso raro (U64) privilegia gli aspetti pragmatici e accoglie il sistema come "ottimo come inizio". L'utente con uso regolare (U21) è il più soddisfatto in assoluto. L'utilità pragmatica è dunque apprezzata in modo trasversale, mentre è la qualità edonica del sistema a essere percepita in modo differenziato a seconda dell'età e della frequenza d'uso. Nel complesso, il test pilota restituisce l'immagine di un prototipo valutato come utile, accessibile e raccomandabile, le cui principali aree di lavoro si concentrano su due compromessi di progettazione: la rigidità prodotta dall'impostazione anti-allucinazione e l'uniformità del registro prodotta dall'impostazione anti-compiacenza. Trattandosi di una valutazione esplorativa su tre utenti, questi rilievi non costituiscono una validazione conclusiva del sistema, ma un insieme di indicazioni per il suo affinamento, oltre che un primo riscontro empirico sul rapporto tra le scelte di *design* del chatbot e l'esperienza dei suoi utenti.

4.9 Discussione

Questa sezione conclusiva si propone di mettere in relazione i risultati presentati nelle sezioni precedenti con il quadro teorico costruito nel Capitolo 2, al fine di interpretarne il significato rispetto agli obiettivi della ricerca. L'analisi dei dati ha restituito un insieme di schemi coerenti che consentono di rispondere alle domande centrali della tesi: in quale misura un campione di adulti non specialisti dispone delle competenze che compongono l'AI literacy, e in quale modo uno strumento educativo come il chatbot sviluppato in questa ricerca può contribuire a colmare le lacune rilevate? La discussione è organizzata attorno ai principali schemi interpretativi emersi dall'analisi, integrati con una riflessione sul potenziale e sui limiti del chatbot come risposta ai bisogni emersi.

4.9.1 Un uso attivo ma epistemicamente fragile

Il primo dato rilevante emerso dall'analisi non riguarda la frequenza d'uso in sé, ma la combinazione tra frequenza d'uso e livello di comprensione tecnica. Il 53,3% del campione utilizza i chatbot quotidianamente o più volte a settimana, ma le medie del blocco concettuale

– che misura la capacità di spiegare il funzionamento dei sistemi IA – si collocano tra M=2,14 e M=3,12 su una scala a cinque punti. La quasi totalità del campione usa strumenti che non sa spiegare. Su scala nazionale, l'indagine di Savoldi et al. (2025) rileva tra gli adulti italiani un uso diffuso della GenAI – l'80,4% del campione (1.533 rispondenti) – a fronte di livelli di alfabetizzazione contenuti e di divari marcati: il quadro che emerge dai presenti dati si inscrive dunque in una tendenza più ampia. Tale condizione può inoltre essere estesa alla maggior parte delle tecnologie, utilizzate spesso senza che i propri utenti ne comprendano i principi di funzionamento. Ciò che distingue i sistemi di IA generativa è però la natura dell'output prodotto: testo in linguaggio naturale, formulato in prima persona, privo di indicatori formali che segnalino l'origine automatica o la possibilità di errore. Come discusso nel Capitolo 2, la fluidità linguistica degli output dei Large Language Model costituisce di per sé un fattore di rischio, poiché attiva nei lettori le stesse aspettative di attendibilità che si associano alla produzione umana intenzionale (cfr. Sezione 2.2.3). In questo senso, l'uso frequente senza comprensione tecnica non è equivalente all'uso, per esempio, di un'applicazione di navigazione GPS: il rischio cognitivo è strutturalmente diverso, perché dipende dalla forma linguistica dell'output.

Le definizioni di AI literacy più influenti nella letteratura recente convergono nel distinguere tra la capacità di utilizzare uno strumento IA e la capacità di comprenderne il funzionamento, valutarne criticamente l'output e ragionare sulle sue implicazioni sociali ed etiche. Long e Magerko (2020) identificano diciassette competenze distribuite su cinque aree, tra cui la comprensione del meccanismo sottostante occupa una posizione centrale. Ng et al. (2021) articolano l'AI literacy in quattro dimensioni – conoscere, usare, valutare, creare – in cui la conoscenza del funzionamento precede e fonda la capacità di valutazione critica. Il modello TUCAPA di Laupichler et al. (2023), su cui si fonda in parte il questionario di questa ricerca, distingue esplicitamente tra la dimensione della comprensione tecnica e quella dell'applicazione pratica. I risultati del campione mostrano uno sbilanciamento sistematico verso quest'ultima, con la prima che rimane largamente sottosviluppata.

4.9.2 Il paradosso della competenza percepita: valutazione senza comprensione

Allo scarto tra uso e comprensione appena descritto se ne affianca un secondo, interno alle competenze che i rispondenti si attribuiscono: le medie del blocco critico-pratico (M = 3,21-4,14) superano sistematicamente quelle del blocco concettuale (M = 2,14-3,12). È il pattern più robusto dell'intera analisi, e pone una questione interpretativa rilevante: come è possibile che utenti con una comprensione tecnica molto limitata dei LLM si percepiscano comunque in grado di valutare criticamente l'affidabilità di una risposta, identificare i rischi in contesti decisionali importanti o verificare le informazioni attraverso altre fonti?

Un'interpretazione plausibile chiama in causa il trasferimento di competenze critiche già sviluppate per altri media digitali. La valutazione dell'affidabilità di una fonte, il controllo incrociato delle informazioni, il riconoscimento della differenza tra contenuto attendibile e contenuto persuasivo sono competenze che appartengono al dominio più ampio della alfabetizzazione mediatica (*media literacy*) e che possono essere mobilitate anche nell'interazione con i chatbot, senza che sia necessaria una comprensione specifica del meccanismo generativo (Aufderheide, 1993). Questa ipotesi è coerente con il dato sui 18-25, che dichiarano comportamenti di verifica molto più frequenti dei 50 o più (83,8% vs 56,0% nella fascia 4-5) pur avendo una comprensione concettuale dei LLM analoga: la loro abitudine alla verifica sembra precedere l'incontro con i chatbot, non derivarne. Questa lettura trova un riscontro nella letteratura recente: Rheu e Cho (2025) osservano che le diverse dimensioni dell'AI literacy agiscono in modo opposto sul comportamento di verifica – la comprensione dei processi alla base dei modelli la favorisce, mentre la fiducia nelle proprie capacità d'uso e la conoscenza delle funzionalità la riducono, alimentando un affidamento di tipo euristico che innalza la credibilità percepita del sistema. La verifica dichiarata dai più giovani, che non hanno una comprensione dei processi superiore a quella delle altre fasce, difficilmente deriva dunque dalla loro familiarità con i chatbot, e rimanda piuttosto a una competenza critica maturata altrove.

Se questa interpretazione è corretta, ha implicazioni importanti per la progettazione di interventi educativi. Significa che una parte delle competenze necessarie per un uso critico dei chatbot è già presente nel patrimonio di molti utenti, almeno in forma tacita o trasferibile. L'obiettivo prioritario di un intervento di AI literacy non sarebbe allora costruire da zero una

competenza critica, ma fornire la base tecnica che ne consenta l'applicazione consapevole: la comprensione del meccanismo generativo come prerequisito per calibrare correttamente la fiducia nel sistema. È esattamente su questo punto che il fraintendimento centrale rilevato nel campione agisce come fattore di interferenza.

4.9.3 Profili differenziati: implicazioni per l'intervento educativo

Le analisi incrociate hanno restituito profili differenziati che meritano una riflessione specifica, perché suggeriscono la necessità di interventi educativi orientati in modo diverso a seconda del gruppo di riferimento.

La fascia 50 o più presenta la combinazione più critica: tasso di fraintendimento più alto (84,0%), punteggio concettuale più basso (M=1,60 sull'item che richiedeva di spiegare il funzionamento dei LLM), quota di non-verificatori più alta (28,0%), frequenza d'uso più bassa. Non si tratta di un profilo che ripone fiducia cieca in tali strumenti – la fiducia dichiarata non è la più alta del campione – ma di un profilo in cui l'assenza di strumenti concettuali rende difficile formulare un giudizio fondato sull'affidabilità del sistema. È il profilo che, sulla base dei dati, avrebbe più bisogno di un intervento educativo specifico. È anche, paradossalmente, quello meno facilmente raggiungibile attraverso uno strumento digitale come il chatbot, dato il minor uso dei sistemi IA conversazionali in questa fascia d'età. I giovani adulti nella fascia 18-25 presentano un profilo diverso, e per certi aspetti più rassicurante: la frequenza di verifica delle fonti è alta (83,8% nella fascia 4-5), il rifiuto di usare i chatbot per temi sensibili è pressoché universale. Eppure anche i 18-25 hanno una comprensione del meccanismo molto limitata: ottengono M=2,14 sull'item LLM – un valore basso, per quanto superiore a quello dei 50+ (M=1,60) – e in maggioranza (68%) condividono il fraintendimento centrale, ossia la convinzione che il chatbot comprenda il significato delle parole e recuperi informazioni da internet. La differenza tra le due fasce sta soprattutto nei comportamenti, non nella comprensione. Come discusso nella sezione precedente, questo suggerisce che l'abitudine alla verifica sia il prodotto di un'esposizione critica ad altri media digitali, non di una comprensione specifica dei LLM. Questo profilo è quello in cui il chatbot educativo ha maggiori probabilità di essere utilizzato e risulta potenzialmente efficace:

l'utente già abituato agli strumenti digitali, già dotato di istinto critico, può beneficiare di uno strumento che fornisce spiegazioni fondate sul meccanismo.

4.9.4 Il chatbot educativo come risposta ai bisogni rilevati

I risultati del questionario consentono di valutare, in modo più fondato, il potenziale e i limiti del chatbot educativo sviluppato in questa ricerca come strumento di AI literacy.

La criticità più rilevante emersa dall'analisi è il fraintendimento sul funzionamento dei LLM: il 66,3% del campione non conosce il meccanismo statistico-probabilistico che governa la generazione del testo. La KB del chatbot (cfr. Sezione 3.2.3) è stata costruita appositamente per rispondere a domande su questo meccanismo: il funzionamento dei transformer, il concetto di *token* e di distribuzione di probabilità, la differenza tra generazione testuale e comprensione semantica, le allucinazioni come conseguenza strutturale del processo generativo. L'architettura RAG garantisce che le risposte siano ancorate ai documenti della KB, riducendo il rischio che il chatbot stesso produca i fraintendimenti che è chiamato a correggere. La seconda lacuna rilevante riguarda la conoscenza dei bias. L'item sul riconoscimento dei bias è il più basso del blocco critico (M=3,21) ed è quello in cui la comprensione del meccanismo tecnico è più direttamente rilevante: per riconoscere che un modello può riprodurre i pregiudizi presenti nei suoi dati di addestramento, occorre sapere cosa sono i dati di addestramento e come influenzano il comportamento del modello. Anche questa tematica è coperta dalla KB del chatbot, che include documentazione specifica sui bias nei sistemi IA e sulle loro implicazioni per l'uso quotidiano.

Il potenziale del chatbot va però letto con cautela, a partire da due limiti strutturali. Il primo riguarda il raggiungimento del pubblico target: il chatbot intercetta prevalentemente chi già usa i chatbot con frequenza, che è anche il profilo demografico più giovane e con comportamenti critici già attivi. Il gruppo più vulnerabile – la fascia 50 o più – è anche quello che usa i chatbot meno frequentemente e che quindi ha meno probabilità di incontrare uno strumento come questo nel proprio percorso quotidiano. Questi risultati vanno però letti tenendo conto di un limite nella costruzione delle fasce d'età: la categoria "50 o più" è aperta e particolarmente ampia, e accorpa profili potenzialmente molto diversi – dal cinquantenne ancora attivo nel mondo del lavoro, e quindi più esposto all'uso quotidiano e professionale dei chatbot, all'utente più anziano, con minori occasioni di contatto con questi strumenti. È plausibile che il livello di AI literacy vari sensibilmente all'interno della fascia: il profilo qui

delineato va inteso, perciò, come una tendenza aggregata e non come un tratto uniforme di tutti i suoi membri. Un intervento educativo pensato per questa fascia, in ogni caso, richiederebbe canali e modalità diverse, in cui un chatbot di questo tipo – che da solo probabilmente non sarebbe efficace – può essere incluso, come parte di un percorso più ampio. Il secondo limite riguarda la natura dell'apprendimento mediato da un chatbot. Nella loro rassegna sistematica dei chatbot in contesti educativi, Wollny et al. (2021) distinguono i diversi ruoli che tali sistemi possono assumere – dall'ausilio alla consultazione fino al mentore o al tutor – e osservano come i sistemi più evoluti tendano verso l'adattività e la personalizzazione rispetto all'utente. Rispetto a questo spettro, il chatbot sviluppato in questa ricerca si colloca all'estremità più semplice: risponde a domande specifiche e consente un dialogo articolato, ma non prevede una valutazione adattiva delle competenze dell'utente né un percorso strutturato di apprendimento progressivo. È uno strumento di consultazione consapevole, non un tutor adattivo. Il suo contributo all'AI literacy è reale ma parziale: può correggere fraintendimenti puntuali, fornire spiegazioni fondate, promuovere un uso più informato dei chatbot; non può sostituire un percorso formativo strutturato, di cui però può essere parte integrante.

Questa distinzione è coerente con i principi della RAG come architettura di grounding. Il sistema non genera contenuti *ex novo* ma recupera e sintetizza informazioni dalla *knowledge base*: questo garantisce affidabilità e tracciabilità, ma pone anche un limite alla profondità dell'interazione possibile. Un chatbot RAG risponde meglio alle domande fattuali che a quelle che richiedono ragionamento complesso o adattamento al contesto specifico dell'utente, come dimostrato anche nel test pilota (cfr. Sezione 4.8.2). Nel dominio dell'AI literacy, le domande fattuali – cosa è un LLM, come funziona un'allucinazione, cosa sono i dati di addestramento – rappresentano però esattamente il tipo di conoscenza che il campione di questa ricerca mostra di non possedere: il chatbot risponde a un bisogno reale e ben definito.

Un primo riscontro empirico in questa direzione proviene dal test pilota presentato in 4.8: pur nei limiti di un campione di tre utenti, la valutazione ha confermato l'utilità percepita e l'accessibilità dello strumento, segnalando al contempo come le scelte di progettazione orientate al grounding possano irrigidire le risposte sulle domande più aperte.

5. Conclusioni

Questa ricerca ha preso le mosse dallo scarto tra la rapida diffusione dei sistemi di intelligenza artificiale generativa presso il pubblico generalista e la comprensione, assai più limitata, del loro funzionamento. Per indagare questo scarto, il lavoro si è articolato in due direzioni principali: anzitutto rilevare, attraverso un questionario, il livello di AI literacy in un campione di adulti italiani non specialisti, e, come conseguenza, progettare un chatbot educativo basato su architettura Retrieval-Augmented Generation, pensato come possibile risposta ai bisogni emersi. Le pagine che seguono riprendono i risultati principali della ricerca, ne dichiarano i limiti e indicano alcune possibili direzioni di lavoro futuro.

L'indagine ha restituito l'immagine di un campione tecnologicamente attivo ma epistemicamente fragile. Il campione utilizza i chatbot in modo diffuso e frequente, ma con una comprensione del loro funzionamento molto limitata: a un uso quotidiano o regolare dichiarato da più della metà dei rispondenti corrispondono, sul blocco concettuale del questionario, le medie più basse dell'intero strumento. A questo disallineamento tra uso e comprensione se ne affianca un secondo, interno alle competenze auto-attribuite, in quanto i rispondenti si percepiscono più capaci di valutare e verificare gli output che di spiegare come essi siano prodotti. Si è ipotizzato che questa competenza critica derivi, almeno in parte, dal trasferimento di abilità già sviluppate per altri media digitali, più che da una comprensione specifica del meccanismo generativo.

Il dato singolo più significativo è il fraintendimento concettuale sul funzionamento dei modelli linguistici: due terzi del campione (66,3%) ritiene erroneamente che un sistema come ChatGPT comprenda il significato delle parole e recuperi informazioni da internet, mentre la quasi totalità riconosce, al contempo, che un chatbot può produrre testo fluido ma falso. La consapevolezza del rischio coesiste, dunque, con l'assenza di una conoscenza, anche basilare, del meccanismo che lo genera. Le analisi incrociate hanno inoltre delineato profili differenziati: la fascia più anziana riunisce gli indicatori più critici, mentre i più giovani, pur usando i chatbot con maggiore frequenza, non ne comprendono il funzionamento meglio degli altri e devono la loro maggiore propensione alla verifica a un'abitudine critica generale. Il titolo di studio, infine, non predice la comprensione del meccanismo in modo lineare, pur risultando associato all'abitudine alla verifica delle fonti.

Sul versante progettuale, il chatbot è stato concepito per rispondere proprio alle lacune rilevate: la sua KB copre il funzionamento dei modelli e i rischi – discussi nel Capitolo 2 –

associati al loro uso, e l'architettura RAG vincola le risposte a documenti verificati, riducendo il rischio che il sistema riproduca i fraintendimenti che dovrebbe correggere. Il test pilota, condotto su tre utenti di fasce d'età diverse, ha natura esclusivamente esplorativa e non costituisce una validazione del sistema, ma offre un primo riscontro qualitativo sull'esperienza d'uso. Entro questi limiti, il prototipo è stato giudicato utile, accessibile e raccomandabile anche dall'utente meno esperto, ma ha messo in luce due compromessi di progettazione: l'orientamento al *grounding*, che contiene le allucinazioni, irrigidisce le risposte alle domande più aperte; il registro asciutto, scelto per contrastare la compiacenza del modello, viene percepito come poco naturale.

I risultati dello studio, comunque, vanno letti alla luce di alcuni limiti, che condividono una radice comune: la natura auto-riferita del dato. Uno strumento di auto-valutazione non rileva ciò che una persona sa fare, ma ciò che ritiene di sapere e accetta di dichiarare; le medie delle sezioni Likert misurano dunque una percezione di competenza, non una competenza verificata. La distanza tra i due piani emerge da un confronto interno ai dati: il comportamento di verifica delle fonti è dichiarato altissimo (cfr. Sezione 4.5), eppure due terzi degli stessi rispondenti non possiedono un modello corretto del funzionamento dei sistemi che dichiarano di controllare (cfr. Sezione 4.6). Affermare di verificare sistematicamente le informazioni prodotte da uno strumento di cui si fraintende il funzionamento è probabilmente, almeno in parte, un effetto di desiderabilità sociale: il dato non è privo di valore, ma va letto come autorappresentazione, non come prova di una pratica effettiva.

Un secondo ordine di limiti riguarda la composizione del campione. Il campionamento a valanga produce per costruzione un campione di convenienza, non probabilistico e non rappresentativo della popolazione adulta italiana (cfr. Sezione 4.1): fortemente sbilanciato per genere, concentrato agli estremi della distribuzione anagrafica e nettamente più scolarizzato della media nazionale, con una quota di rispondenti con almeno una laurea triennale (65,2%) circa tripla di quella nazionale tra i 25-64enni in possesso di un titolo terziario (21,6%)[32]. A ciò si aggiunge un limite nella costruzione stessa della variabile età: le fasce adottate hanno ampiezze irregolari e la classe aperta "50 o più" accorpa profili molto eterogenei, da chi è ancora attivo professionalmente a chi ha smesso da tempo di usare strumenti digitali. Il dato secondo cui questa fascia condivide in larga misura il fraintendimento centrale (cfr. Sezione 4.7.1) va quindi interpretato con cautela, perché la variabile, così costruita, non consente di

[32] https://www.istat.it/wp-content/uploads/2025/05/Rapporto-Annuale-2025-integrale.pdf

distinguere l'effetto dell'età da quello di fattori connessi, come la posizione rispetto al mondo del lavoro. Un'ambiguità analoga riguarda l'item sull'ambito professionale, che mescola settori di attività e stati occupazionali: la categoria "non occupato", in particolare, rende invisibile la condizione del pensionamento, uno dei fattori più plausibili dietro la minore familiarità con questi strumenti. Anche la sezione finale del questionario, il cui scopo era fornire uno strumento di misura delle conoscenze fattuali del campione, ha mostrato dei limiti. Nella sezione Vero/Falso, infatti, sei item su sette sono stati risolti correttamente da oltre l'80% dei rispondenti: il formato dicotomico e la formulazione di alcune affermazioni le rendevano riconoscibili e risolvibili per via linguistica o pragmatica, a prescindere da una reale conoscenza del dominio, cosicché solo l'item sul fraintendimento centrale ha avuto un effettivo potere discriminante (cfr. Sezione 4.6). A un ultimo item, infine, va imputato un limite di formulazione più che di difficoltà: l'item 31, relativo alla responsabilità del contenuto generato, è sintatticamente ambiguo, al punto che non è possibile distinguere chi nega la responsabilità dell'utente da chi ha semplicemente frainteso la frase (cfr. Sezione 4.7.3). Sul piano dell'analisi, la ricerca si è fondata esclusivamente su statistiche descrittive – frequenze, percentuali, medie e tabelle incrociate – adeguate a descrivere il campione, ma non a stimare in modo inferenziale le relazioni tra le variabili né a controllarne i reciproci confondimenti. Un'analisi più robusta avrebbe potuto modellare statisticamente i dati – ad esempio con modelli a effetti misti – così da verificare, e non solo descrivere, l'effetto di variabili come l'età e il titolo di studio sulla comprensione e sui comportamenti dichiarati. Tale approfondimento costituisce una direzione di sviluppo per ricerche future.

Sul versante progettuale, anche il chatbot presenta limiti precisi. Il primo limite riguarda la sua stessa natura: è uno strumento di consultazione, non un tutor adattivo e risponde a domande puntuali non valutando le competenze dell'utente né proponendo un percorso di apprendimento strutturato (cfr. Sezione 4.9.4; Wollny et al., 2021). La sua *knowledge base* ha una copertura circoscritta, e la valutazione su tre soli utenti non consente conclusioni generalizzabili. Vi è infine un limite pratico, che riguarda il raggiungimento del pubblico: uno strumento digitale di questo tipo intercetta soprattutto chi già usa i chatbot, ossia il profilo più giovane e già dotato di strumenti critici, mentre la fascia che dai dati risulta più vulnerabile – gli adulti più anziani – è anche quella che meno probabilmente lo incontrerebbe.

A partire da questi limiti, si possono delineare le possibili direzioni future di sviluppo della ricerca. A livello progettuale, la knowledge base del chatbot educativo potrebbe essere ampliata e aggiornata per estenderne la copertura tematica, e il sistema potrebbe evolvere

verso forme di maggiore adattività, capaci di calibrare le risposte sul profilo dell'utente. Va riconosciuto, del resto, che il carattere interattivo del chatbot ne fa uno strumento potenzialmente utile sul piano didattico, poiché a differenza di un materiale statico, coinvolge attivamente l'utente, che formula domande proprie e riceve risposte calibrate sulle sue richieste, in una forma di apprendimento attivo di cui il test pilota ha restituito un primo riscontro positivo (cfr. Sezione 4.8.1). Più che come applicazione isolata, però, un chatbot di questo tipo sembra trovare la sua collocazione più efficace all'interno di un percorso educativo strutturato, di cui costituisca una componente e non l'intero intervento; per la fascia più anziana, in particolare, sarebbero necessari canali e modalità di mediazione diversi da quelli puramente digitali.

I limiti rilevati riguardanti il disegno di ricerca si traducono in indicazioni concrete per una revisione del questionario. L'età andrebbe rilevata come dato numerico aperto e ricodificata a posteriori in fasce mutuamente esclusive e teoricamente motivate, così da non comprimere in un'unica categoria età molto diverse; lo stato occupazionale e il settore di attività andrebbero rilevati con due domande distinte, isolando quest'ultimo come variabile di controllo e rendendo visibile la condizione del pensionamento. Gli item Vero/Falso più intuitivi potrebbero essere sostituiti da domande a scelta multipla con distrattori plausibili o da quesiti basati su scenari, che riducono la probabilità di risposta casuale e misurano il ragionamento applicato anziché una credenza isolata; l'item sulla responsabilità andrebbe riscritto disambiguando il soggetto dell'azione. Infine, le scale di auto-percezione andrebbero affiancate da almeno un compito concreto – per esempio la valutazione di un output reale – così da triangolare ciò che i rispondenti dichiarano con ciò che effettivamente fanno. Una replica su un campione più ampio e bilanciato, con quote minime per ciascuna fascia, e una valutazione del chatbot su un numero più ampio di partecipanti consentirebbe infine di superare la natura esplorativa di questo lavoro.

Nel complesso, il lavoro offre un quadro esplorativo del livello di AI *literacy* in un campione di adulti italiani non specialisti – un pubblico relativamente poco indagato – e un prototipo di strumento educativo costruito sui bisogni emersi durante la fase diagnostica. Il risultato più netto – la coesistenza di un uso diffuso e di una generale consapevolezza dei rischi abbinati a una comprensione assai limitata del meccanismo – suggerisce che un intervento di AI literacy rivolto a questo pubblico debba agire meno sulla costruzione di una competenza critica *ex novo* e più sulla comprensione del funzionamento dei sistemi, che di quella competenza costituisce il presupposto; a questo bisogno, circoscritto ma reale, lo strumento sviluppato in questa ricerca prova a rispondere. Nel suo insieme, il lavoro trova il

proprio senso in una prospettiva duplice: rendere visibile uno scarto di consapevolezza e mettere a disposizione uno strumento per ridurlo. La questione, però, è più ampia e va oltre il singolo prototipo. Man mano che i sistemi generativi entrano stabilmente nell'esperienza quotidiana – nello studio, nel lavoro, nell'accesso alle informazioni – ciò che è in gioco non è soltanto una competenza pratica, ma la comprensione del loro funzionamento, quanto basta per valutarne i prodotti. È questa comprensione la *conditio sine qua non* che distingue un utente capace di servirsi criticamente di tali strumenti da un fruitore che ne accetta passivamente i risultati: è in questo punto, più che nella sola dimestichezza pratica, che si misura il senso dell'AI literacy come competenza diffusa. Coltivarla non significa trasformare gli utenti in specialisti dotati di competenze tecniche, ma porre gli utenti nelle condizioni di un rapporto consapevole con tecnologie ormai pervasive.

6. Appendice

6.1 Questionario AI Literacy

Conoscenze, competenze e pratiche nell'uso dei chatbot basati su intelligenza artificiale

Gentile partecipante,
Grazie per aver scelto di contribuire alla nostra ricerca.
Lo studio ha come obiettivo indagare l'alfabetizzazione degli utenti relativa ai sistemi di intelligenza artificiale generativa, con particolare riferimento ai chatbot basati su grandi modelli linguistici (LLM), come ChatGPT, Gemini o Copilot.
Il questionario consiste di 7 parti, è anonimo e richiede circa 10-15 minuti. Non esistono risposte giuste o sbagliate: ci interessa la tua esperienza e la tua percezione reale.
Desideriamo informarti che i dati raccolti durante la tua partecipazione all'esperimento verranno immagazzinati, trattati e usati nel massimo rispetto della normativa sulla privacy (ai sensi dell'art. 13 GDPR, Regolamento UE 2016/679, tutela della privacy). La registrazione delle tue risposte fornite all'indagine non comporta in alcun modo la tua identificazione diretta.
I dati raccolti attraverso questo esperimento potranno essere usati per scopi di ricerca. Nello specifico, i tuoi dati anonimizzati potranno essere presentati in occasione di conferenze e convegni, potranno essere descritti in modo aggregato in articoli scientifici in riviste e in volumi.
Potrai decidere di ritirarti in qualsiasi momento, abbandonando questa pagina web oppure contattando i responsabili della ricerca agli indirizzi e-mail qui riportati: Gaia Eleonora Di Raimondo, gaiaeleonora.diraimondo01@universitadipavia.it
In caso di ulteriori domande o dubbi, siamo a tua disposizione.
Continuando con la compilazione del questionario, si acconsente esplicitamente alla partecipazione a questo studio e ai termini sopracitati.

SEZIONE A – Dati demografici

A1. Età:

☐ 18–25 ☐ 26–35 ☐ 36–45 ☐ 46–50 ☐ 50 o più

A2. Genere:

☐ Donna ☐ Uomo ☐ Non binario/altro ☐ Preferisco non rispondere

A3. Titolo di studio più elevato conseguito:

☐ Licenza media ☐ Diploma superiore ☐ Laurea triennale ☐ Laurea magistrale o superiore

A4. Ambito professionale attuale:

☐ Istruzione / ricerca ☐ Sanità ☐ Giuridico / amministrativo ☐ Commercio / finanza ☐ Industria / artigianato ☐ Comunicazione / media / cultura ☐ Informatica / tecnologia ☐ Studente ☐ Non occupato ☐ Altro

SEZIONE B – Esperienza con chatbot

Indica il tuo livello di utilizzo dei chatbot basati su intelligenza artificiale (es. ChatGPT, Gemini, Copilot, Claude o simili).

B1. Hai mai utilizzato un chatbot IA?

☐ Sì ☐ No

B2. Con quale frequenza lo utilizzi?

☐ Mai ☐ Raramente (qualche volta all'anno) ☐ Occasionalmente (qualche volta al mese) ☐ Regolarmente (più volte a settimana) ☐ Quotidianamente

B3. Per quali scopi lo utilizzi? (puoi selezionare più opzioni)

☐ Ricerca di informazioni ☐ Scrittura e redazione testi ☐ Studio e formazione

☐ Lavoro professionale ☐ Supporto emotivo / conversazione ☐ Traduzione

☐ Consulenza medica o legale ☐ Programmazione / codice ☐ Creatività / intrattenimento

☐ Non lo utilizzo ☐ Altro: _____

B4. Come giudichi complessivamente la tua familiarità con i chatbot IA?

☐ Nessuna ☐ Molto bassa ☐ Bassa ☐ Media ☐ Alta ☐ Molto alta

B5. Generalmente, ho fiducia nelle risposte generate dai chatbot IA.

☐ Nessuna ☐ Molto bassa ☐ Bassa ☐ Media ☐ Alta ☐ Molto alta

SEZIONE C – Comprensione del funzionamento dei sistemi IA

Le domande seguenti riguardano il grado di conoscenza del funzionamento dei chatbot e dei sistemi di IA generativa. Per ciascuna, indica il tuo accordo su una scala da 1 a 5, dove: 1 = Per niente d'accordo | 2 = Poco d'accordo | 3 = Abbastanza d'accordo | 4 = D'accordo | 5 = Completamente d'accordo

Sono in grado di:

1. spiegare, a grandi linee, come funziona un chatbot come ChatGPT o Gemini.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

2. spiegare cosa vuol dire che i sistemi IA come i chatbot elaborano il linguaggio in modo statistico

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

3. spiegare cosa si intende per "modello linguistico di grandi dimensioni" (LLM).

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

4. descrivere perché un chatbot IA può produrre informazioni errate o inventate (le cosiddette "allucinazioni").

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

5. spiegare la differenza tra IA generativa (come i chatbot) e altri tipi di software tradizionale.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

6. descrivere che cosa sono i dati di addestramento e quale ruolo hanno nello sviluppo di un sistema IA.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

SEZIONE D – Valutazione critica degli output dei chatbot

Le domande seguenti riguardano la tua capacità di valutare criticamente le risposte dei chatbot IA. Usa la stessa scala da 1 a 5.

7. Sono in grado di riconoscere quando un chatbot IA sta fornendo informazioni inesatte, incomplete o inventate.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

8. Sono in grado di valutare criticamente l'affidabilità di una risposta di un chatbot prima di utilizzarla.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

9. Sono in grado di identificare i rischi legati all'uso di chatbot IA per decisioni importanti (es. mediche, legali, finanziarie).

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

10. Sono in grado di riconoscere quando i pregiudizi (bias) presenti nei dati di addestramento possono influenzare le risposte di un chatbot.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

11. Sono in grado di distinguere tra un testo linguisticamente fluido e convincente e uno che contiene informazioni corrette e verificate.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

12. Sono in grado di valutare se una domanda o un compito specifico può essere affrontato in modo appropriato da un chatbot IA oppure no.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

SEZIONE E – Uso pratico dei chatbot IA

Le domande seguenti riguardano le tue pratiche d'uso dei chatbot IA. Usa la stessa scala da 1 a 5.

13. Sono in grado di formulare richieste (prompt) efficaci per ottenere risposte più utili e accurate da un chatbot.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

14. Sono in grado di verificare attraverso altre fonti le informazioni importanti fornite da un chatbot IA.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

15. Sono in grado di decidere autonomamente quando è opportuno usare un chatbot e quando invece è preferibile ricorrere ad altre fonti.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

16. Sono in grado di identificare situazioni della mia vita quotidiana o lavorativa in cui entro in contatto con sistemi IA.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

17. Sono in grado di spiegare a qualcuno le capacità e i limiti di un chatbot IA in modo comprensibile.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

18. Quando uso un chatbot per informazioni importanti, controllo altre fonti.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

19. Utilizzo chatbot per chiedere consigli su argomenti intimi e personali.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

SEZIONE F – Consapevolezza etica e sociale

Le domande seguenti riguardano le implicazioni etiche e sociali dei chatbot IA. Usa la stessa scala da 1 a 5.

Sono in grado di:

20. riconoscere le implicazioni per la privacy dei dati che condivido con un chatbot IA.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

21. descrivere i rischi etici legati all'uso diffuso dell'IA generativa nella società.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

22. riflettere sulla responsabilità personale nell'uso dei contenuti generati da un chatbot.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

23. spiegare perché è importante che i sistemi IA siano sviluppati e utilizzati in modo responsabile.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

24. riconoscere situazioni in cui l'uso di un chatbot potrebbe essere inappropriato o dannoso.

☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

SEZIONE G – Verifica delle conoscenze (Vero/Falso)

Indica se ritieni ciascuna affermazione vera o falsa.

25. Un chatbot IA come ChatGPT genera le sue risposte comprendendo il significato delle parole inserite dall'utente nell'input e recuperando informazioni su Internet per costruire risposte pertinenti.

☐ Vero	☐ Falso

26. I chatbot IA possono riprodurre anche i pregiudizi presenti nei dati di addestramento.

☐ Vero	☐ Falso

27. Un chatbot IA può produrre una risposta linguisticamente fluida, corretta e apparentemente autorevole anche quando il contenuto è falso o inventato.

☐ Vero	☐ Falso

28. I chatbot IA hanno sempre accesso alle informazioni più aggiornate disponibili su Internet e le verificano prima di rispondere.

☐ Vero	☐ Falso

29. Condividere informazioni personali sensibili (dati medici, documenti, password) con un chatbot IA non comporta alcun rischio per la privacy.

☐ Vero	☐ Falso

30. Il livello di fiducia che un utente ripone in un chatbot IA non dipende solo dall'effettiva accuratezza delle risposte, ma può essere influenzato dalla fluidità e dal tono del testo generato.

☐ Vero	☐ Falso

31. Quando un utente usa un contenuto generato da un chatbot, la responsabilità di quel contenuto è dell'utente che l'ha generato.

☐ Vero	☐ Falso

6.2 Distribuzione dei dati

Le tabelle seguenti riportano le distribuzioni complete, item per item, del questionario somministrato (n = 92). Salvo diversa indicazione, i valori percentuali sono calcolati su n = 92; eventuali totali diversi da 100% sono dovuti agli arrotondamenti.

Sezione A – Dati demografici

Tabella 1 – Distribuzione del campione per età (n = 92)

Fascia d'età	n	%
18–25	37	40,2
26–35	22	23,9
36–45	3	3,3
46–50	5	5,4
50 o più	25	27,2
Totale	**92**	**100**

Tabella 2 – Distribuzione per genere (n = 92)

Genere	n	%
Donna	74	80,4
Uomo	17	18,5
Non binario/altro	1	1,1
Preferisco non rispondere	0	0,0
Totale	**92**	**100**

Tabella 3 – Distribuzione per titolo di studio (n = 92)

Titolo di studio	n	%
Licenza media	3	3,3

Diploma superiore	29	31,5
Laurea triennale	22	23,9
Laurea magistrale o superiore	38	41,3
Totale	**92**	**100**

Tabella 4 – Distribuzione per ambito professionale (n = 92)

Ambito professionale	**n**	**%**
Studente	26	28,3
Sanità	22	23,9
Istruzione / ricerca	12	13,0
Non occupato	10	10,9
Informatica / tecnologia	7	7,6
Comunicazione / media / cultura	4	4,3
Commercio / finanza	4	4,3
Altro	4	4,3
Industria / artigianato	3	3,3
Giuridico / amministrativo	0	0,0
Totale	**92**	**100**

Sezione B – Esperienza con i chatbot IA

Tabella 5 – Utilizzo dei chatbot IA (n = 92)

Hai mai utilizzato un chatbot IA?	**n**	**%**
Sì	85	92,4
No	7	7,6
Totale	**92**	**100**

Tabella 6 – Frequenza d'uso (n = 92)

Frequenza d'uso	n	%
Mai	7	7,6
Raramente (qualche volta all'anno)	12	13,0
Occasionalmente (qualche volta al mese)	24	26,1
Regolarmente (più volte a settimana)	30	32,6
Quotidianamente	19	20,7
Totale	**92**	**100**

Tabella 7 – Scopi d'uso dichiarati

Domanda a risposta multipla; n menzioni = 261. Le percentuali sono calcolate sul totale delle menzioni.

Scopo	Menzioni	%
Ricerca di informazioni	70	26,8
Studio e formazione	42	16,1
Scrittura e redazione testi	36	13,8
Lavoro professionale	24	9,2
Traduzione	24	9,2
Programmazione / codice	17	6,5
Creatività / intrattenimento	17	6,5
Consulenza medica o legale	10	3,8
Supporto emotivo / conversazione	8	3,1
Non lo utilizzo	7	2,7
Altro	6	2,3
Totale menzioni	**261**	**100**

Tabella 8 – Familiarità percepita con i chatbot IA (n = 92)

Familiarità percepita	n	%
Nessuna	5	5,4
Molto bassa	6	6,5
Bassa	14	15,2
Media	44	47,8
Alta	17	18,5
Molto alta	6	6,5
Totale	**92**	**100**

Tabella 9 – Fiducia complessiva nelle risposte dei chatbot IA (n = 92)

Fiducia complessiva	n	%
Nessuna	4	4,3
Molto bassa	9	9,8
Bassa	17	18,5
Media	55	59,8
Alta	7	7,6
Molto alta	0	0,0
Totale	**92**	**100**

Sezione C – Comprensione del funzionamento dei sistemi IA

Tabella 10 – Comprensione del funzionamento dei sistemi IA (item 1-6; n = 92)

Item con formula introduttiva comune “Sono in grado di…”; scala 1-5 (1 = per niente d’accordo; 5 = completamente d’accordo). I valori nelle colonne 1-5 indicano il numero di

rispondenti per ciascun grado. M = media; DS = deviazione standard.

Item	1	2	3	4	5	M	DS
…spiegare, a grandi linee, come funziona un chatbot.	10	17	31	20	14	3,12	1,20
…spiegare cosa vuol dire che i chatbot elaborano il linguaggio in modo statistico-probabilistico.	18	24	19	17	14	2,84	1,35
…spiegare cosa si intende per “modello linguistico di grandi dimensioni” (LLM).	45	20	10	3	14	2,14	1,44
…descrivere perché un chatbot può produrre informazioni errate o inventate (allucinazioni).	25	24	14	16	13	2,65	1,40
…spiegare la differenza tra IA generativa e software tradizionale.	32	15	18	16	11	2,55	1,42
…descrivere che cosa sono i dati di addestramento e il loro ruolo.	34	17	17	9	15	2,50	1,47

Sezione D – Valutazione critica degli output dei chatbot

Tabella 11 – Valutazione critica degli output (item 7-12; n = 92)

Item con formula introduttiva comune “Sono in grado di…”; scala 1-5 (1 = per niente d’accordo; 5 = completamente d’accordo). I valori nelle colonne 1-5 indicano il numero di rispondenti per ciascun grado. M = media; DS = deviazione standard.

Item	1	2	3	4	5	M	DS
…riconoscere quando un chatbot fornisce informazioni inesatte, incomplete o inventate.	6	16	29	31	10	3,25	1,07
…valutare criticamente l’affidabilità di una risposta prima di utilizzarla.	4	10	26	35	17	3,55	1,05

…identificare i rischi dell'uso per decisioni importanti (mediche, legali, finanziarie).	6	3	18	21	44	4,02	1,18
…riconoscere quando i bias nei dati di addestramento influenzano le risposte.	12	11	27	30	12	3,21	1,20
…distinguere tra un testo fluido e convincente e uno con informazioni corrette e verificate.	6	13	25	31	17	3,43	1,14
…valutare se un compito può essere affrontato in modo appropriato da un chatbot.	6	10	27	32	17	3,48	1,11

Sezione E – Uso pratico dei chatbot IA

Tabella 12 – Uso pratico (item 13–17; n = 92)

Item con formula introduttiva comune "Sono in grado di…"; scala 1-5 (1 = per niente d'accordo; 5 = completamente d'accordo). I valori nelle colonne 1-5 indicano il numero di rispondenti per ciascun grado. M = media; DS = deviazione standard.

Item	1	2	3	4	5	M	DS
…formulare richieste (prompt) efficaci per ottenere risposte utili e accurate.	5	12	28	29	18	3,47	1,11
…verificare attraverso altre fonti le informazioni importanti fornite da un chatbot.	5	5	21	25	36	3,89	1,15
…decidere autonomamente quando usare un chatbot e quando ricorrere ad altre fonti.	5	3	11	28	45	4,14	1,10
…identificare situazioni quotidiane o lavorative in cui entro in contatto con sistemi IA.	5	4	27	33	23	3,71	1,06
…spiegare a qualcuno le capacità e i limiti di un chatbot in modo comprensibile.	8	15	28	22	19	3,32	1,22

Tabella 13 – Comportamenti dichiarati (item 18–19; n = 92)

Item formulati come comportamento; scala 1–5 (valori alti = comportamento più frequente). M = media; DS = deviazione standard.

Item	1	2	3	4	5	M	DS
Quando uso un chatbot per informazioni importanti, controllo altre fonti.	6	2	14	24	46	4,11	1,15
Utilizzo chatbot per chiedere consigli su argomenti intimi e personali.	53	21	14	2	2	1,68	0,95

Sezione F – Consapevolezza etica e sociale

Tabella 14 – Consapevolezza etica e sociale (item 20–24; n = 92)

Item con formula introduttiva comune "Sono in grado di…"; scala 1-5 (1 = per niente d'accordo; 5 = completamente d'accordo). I valori nelle colonne 1-5 indicano il numero di rispondenti per ciascun grado. M = media; DS = deviazione standard.

Item	1	2	3	4	5	M	DS
…riconoscere le implicazioni per la privacy dei dati che condivido con un chatbot.	8	11	34	24	15	3,29	1,14
…descrivere i rischi etici legati all'uso diffuso dell'IA generativa nella società.	7	12	29	25	19	3,40	1,17
…riflettere sulla responsabilità personale nell'uso dei contenuti generati.	9	9	26	26	22	3,47	1,23
…spiegare perché è importante che i sistemi IA siano sviluppati e usati responsabilmente.	7	6	26	30	23	3,61	1,15
…riconoscere situazioni in cui l'uso di un chatbot potrebbe essere inappropriato o dannoso.	6	6	19	28	33	3,83	1,18

Sezione G – Verifica delle conoscenze (Vero/Falso)

Tabella 15 – Verifica delle conoscenze: item Vero/Falso (item 25–31; n = 92)

La risposta corretta è definita dalla letteratura scientifica. La percentuale di risposte corrette è calcolata su N = 92.

Affermazione	Vero	Falso	Corretta	% corrette
Un chatbot come ChatGPT genera le risposte comprendendo il significato delle parole e recuperando informazioni su Internet per costruire risposte pertinenti.	61	31	Falso	33,7%
I chatbot IA possono riprodurre anche i pregiudizi (bias) presenti nei dati di addestramento.	88	4	Vero	95,7%
Un chatbot può produrre una risposta fluida, corretta e apparentemente autorevole anche quando il contenuto è falso o inventato.	88	4	Vero	95,7%
I chatbot IA hanno sempre accesso alle informazioni più aggiornate su Internet e le verificano prima di rispondere.	16	76	Falso	82,6%
Condividere informazioni personali sensibili (dati medici, documenti, password) con un chatbot non comporta alcun rischio per la privacy.	6	86	Falso	93,5%
Il livello di fiducia in un chatbot non dipende solo dall'effettiva accuratezza delle risposte, ma può essere influenzato dalla fluidità e dal tono del testo.	91	1	Vero	98,9%
Quando un utente usa un contenuto generato da un chatbot, la responsabilità di quel contenuto è dell'utente che l'ha generato.	75	17	Vero	81,5%

Analisi incrociate per età e titolo di studio

Le celle con n < 5 (fascia 36-45 e licenza media) non sono interpretate nel testo; la fascia 46–50 (n = 5) è riportata a titolo indicativo. I valori sono comunque riportati per completezza.

Tabella 16 – Fraintendimento sul funzionamento (item 25) per fascia d'età

Fascia d'età	n	Corrette (n)	Corrette (%)
18–25	37	12	32
26–35	22	12	55
36–45	3	2	67
46–50	5	1	20
50 o più	25	4	16

Tabella 17 – Fraintendimento sul funzionamento (item 25) per titolo di studio

Titolo di studio	n	Corrette (n)	Corrette (%)
Licenza media	3	2	67
Diploma superiore	29	6	21
Laurea triennale	22	11	50
Laurea magistrale o superiore	38	12	32

Tabella 18 – Conoscenza concettuale LLM (item 3) per fascia d'età

Fascia d'età	n	M	DS
18–25	37	2,14	1,40
26–35	22	2,77	1,65
36–45	3	3,00	1,63
46–50	5	1,60	0,80
50 o più	25	1,60	1,06

Tabella 19 – Conoscenza concettuale LLM (item 3) per titolo di studio

Titolo di studio	n	M	DS
Licenza media	3	1,67	0,94
Diploma superiore	29	1,52	0,93
Laurea triennale	22	2,73	1,63
Laurea magistrale o superiore	38	2,32	1,49

Tabella 20 – Verifica delle fonti (item 18) per fascia d'età

Fascia d'età	n	M	DS
18–25	37	4,43	0,75
26–35	22	4,27	0,96
36–45	3	5,00	0,00
46–50	5	3,80	0,75
50 o più	25	3,44	1,53

Tabella 21 – Verifica delle fonti (item 18) per titolo di studio

Titolo di studio	n	M	DS
Licenza media	3	2,67	1,70
Diploma superiore	29	3,93	1,28
Laurea triennale	22	4,14	0,92
Laurea magistrale o superiore	38	4,34	0,98

Tabella 22 – Responsabilità del contenuto generato (item 31) per fascia d’età

Risposta corretta: “Vero”. La colonna “Falso” indica chi nega la responsabilità dell’utente.

Fascia d'età	n	“Vero” (%)	“Falso” (%)
18–25	37	78	22
26–35	22	77	23
36–45	3	67	33
46–50	5	100	0
50 o più	25	88	12

6.3 Questionario di valutazione

Chatbot educativo sull'AI *literacy: pilot test* – valutazione con utenti

Istruzioni generali

Il questionario si divide in due parti: una breve sezione da compilare prima dell'interazione con il chatbot (Parte A), e una sezione più estesa da compilare dopo l'interazione (Parte B). Non esistono risposte giuste o sbagliate. Le tue risposte serviranno esclusivamente a migliorare lo strumento e a documentare il processo di valutazione nell'ambito di una tesi universitaria. La compilazione è anonima.

Per gli item con scala numerica, segnare con una X o un cerchio il valore corrispondente alla propria opinione, dove 1 = Completamente in disaccordo e 5 = Completamente d'accordo, salvo diversa indicazione. Gli item contrassegnati con (R) hanno significato invertito.

Parte A – Prima dell'interazione

Compila questa sezione prima di iniziare a usare il chatbot.

A1. Indica la tua età

A2. Profilo d'uso

A1	**Con quale frequenza utilizzi chatbot IA (ChatGPT, Gemini, ecc.)?**
	○ Mai ○ Raramente (qualche volta l'anno) ○ Qualche volta al mese ○ Regolarmente (più volte a settimana) ○ Quotidianamente

A3. Fiducia iniziale nei chatbot

		1 = Completamente in disaccordo – 5 = Completamente d'accordo				
#	Affermazione	1	2	3	4	5
A2a	In generale, mi fido delle informazioni fornite da un chatbot IA.	○	○	○	○	○
A2b	Penso che un chatbot possa darmi informazioni accurate su un argomento che non conosco bene.	○	○	○	○	○

A4. Aspettative sull'utilità e sulla facilità d'uso

		1 = Completamente in disaccordo – 5 = Completamente d'accordo				
#	Affermazione	1	2	3	4	5
A3a	Penso che questo chatbot possa aiutarmi a capire meglio i rischi dell'uso dei chatbot IA.	○	○	○	○	○
A3b	Mi aspetto che interagire con questo chatbot sia semplice e immediato.	○	○	○	○	○

Parte B – Dopo l'interazione

Compila questa sezione subito dopo aver interagito con il chatbot.

Sezione 1 – Utilità percepita

Questa sezione valuta se l'interazione con il chatbot ha raggiunto il suo obiettivo educativo: aiutarti a capire meglio i rischi e il funzionamento dei chatbot IA.

		1 = Completamente in disaccordo —— 5 = Completamente d'accordo				
#	Affermazione	1	2	3	4	5
B1	Il chatbot mi ha aiutato a capire meglio almeno uno dei temi trattati (allucinazioni, bias, privacy, ecc.).	○	○	○	○	○
B2	Le informazioni fornite dal chatbot sono state utili per i miei scopi.	○	○	○	○	○

B3	Userei questo chatbot per approfondire altri aspetti legati all'intelligenza artificiale.	○	○	○	○	○

C’è qualcosa che il chatbot non è riuscito a spiegarti in modo soddisfacente? Se sì, cosa?

Sezione 2 – Facilità d’uso e usabilità

Questa sezione valuta quanto è stato semplice e fluido interagire con il chatbot.

		1 = Completamente in disaccordo – 5 = Completamente d’accordo				
#	**Affermazione**	**1**	**2**	**3**	**4**	**5**
B4	Interagire con questo chatbot è stato semplice e immediato.	○	○	○	○	○
B5	La conversazione con il chatbot è stata chiara e facile da seguire.	○	○	○	○	○
B6	Ho avuto difficoltà a ottenere dal chatbot le risposte che cercavo. (R) (R)	○	○	○	○	○

Hai incontrato difficoltà nell’interazione con il chatbot? Se sì, di che tipo?

Sezione 3 – Fiducia nella risposta

Questa sezione valuta quanto le risposte del chatbot ti sono sembrate accurate, fondate e degne di fiducia.

		1 = Completamente in disaccordo – 5 = Completamente d’accordo				
#	**Affermazione**	**1**	**2**	**3**	**4**	**5**
B7	Le risposte del chatbot mi sono sembrate accurate e affidabili.	○	○	○	○	○
B8	Ho avuto l’impressione che il chatbot fornisse informazioni basate su fonti attendibili.	○	○	○	○	○
B9	Mi sono fidato/a delle risposte del chatbot per capire meglio l’argomento trattato.	○	○	○	○	○

Ci sono state risposte del chatbot che ti hanno lasciato dubbi o che ti sono sembrate poco convincenti? Puoi descriverle?

Sezione 4 – Qualità linguistica

Questa sezione valuta il modo in cui il chatbot si esprime: chiarezza, naturalezza del linguaggio, lunghezza delle risposte.

		1 = Completamente in disaccordo – 5 = Completamente d'accordo				
#	**Affermazione**	**1**	**2**	**3**	**4**	**5**
B10	Il linguaggio usato dal chatbot era chiaro e comprensibile.	○	○	○	○	○
B11	Il tono e il registro del chatbot mi sono sembrati naturali, non artificiali o meccanici.	○	○	○	○	○
B12	Le risposte erano della lunghezza giusta: né troppo brevi né troppo lunghe.	○	○	○	○	○

Come descriveresti il modo in cui il chatbot si esprime? C'è qualcosa che cambieresti nel suo stile o nel registro delle risposte?

Sezione 5 — Valutazione complessiva

Questa sezione raccoglie una valutazione d'insieme dell'esperienza e l'intenzione di utilizzo futuro.

		1 = Completamente in disaccordo – 5 = Completamente d'accordo				
#	**Affermazione**	**1**	**2**	**3**	**4**	**5**
B13	Complessivamente, sono soddisfatto/a dell'esperienza con questo chatbot.	○	○	○	○	○
B14	Consiglierei questo chatbot a qualcuno che vuole capire meglio i rischi dell'IA.	○	○	○	○	○

Hai altri commenti, osservazioni o suggerimenti sull'esperienza complessiva con il chatbot?

BIBLIOGRAFIA

Acemoglu, D. (2021). Harms of AI. NBER Working Paper No. 29247. National Bureau of Economic Research.

Aoun, J. E. (2017). Robot-Proof: Higher education in the age of artificial intelligence. Cambridge, MA: MIT Press.

Ashery, A. F., Aiello, L. M., & Baronchelli, A. (2025). Emergent social conventions and collective bias in LLM populations. Science Advances, 11, eadu9368.

Atari, M., Xue, M., Park, P., Blasi, D., & Henrich, J. (2023). Which humans? PsyArXiv preprint.

Aufderheide, P. (1993). Media Literacy: A Report of the National Leadership Conference on Media Literacy. Washington, DC: Aspen Institute.

Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2025). Explicitly unbiased large language models still form biased associations. Proceedings of the National Academy of Sciences.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623).

Blease, C., Hagström, J., Garcia Sanchez, C., Kharko, A., McMillan, B., Gaab, J., Brulin, E., Locher, C., Hägglund, M., Riggare, S., & Mandl, K. D. (2025). General practitioners' adoption of generative artificial intelligence in clinical practice in the UK: An updated online survey. Digital Health, 11.

Bo, J. Y., Wan, S., & Anderson, A. (2025). To Rely or Not to Rely? Evaluating Interventions for Appropriate Reliance on LLMs. CHI 2025, arXiv:2412.15584.

Brooke, J. (1996). SUS: A 'quick and dirty' usability scale. In P. W. Jordan, B. Thomas, B. A. Weerdmeester, & I. L. McClelland (Eds.), Usability Evaluation in Industry (pp. 189-194). London: Taylor & Francis.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., et al. (2020). Language Models are Few-Shot Learners. Advances in Neural Information Processing Systems, 33, 1877-1901.

Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics derived automatically from language corpora contain human-like biases. Science, 356(6334), 183-186.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 335-336).

Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting Training Data from Large Language Models. In Proceedings of the 30th USENIX Security Symposium.

Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramèr, F., & Zhang, C. (2023). Quantifying Memorization Across Neural Language Models. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023).

Carpineto, C., & Romano, G. (2012). A survey of automatic query expansion in information retrieval. ACM Computing Surveys, 44(1), 1-50.

Celi, L. (2026). Understanding Is Not in the Text: Meaning, Interpretation, and Artificial Intelligence. SSRN Working Paper.

Chen, J., Xiao, S., Zhang, P., Luo, K., Lian, D., & Liu, Z. (2024). BGE M3-Embedding: Multi-Lingual, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation. arXiv:2402.03216.

Cotton, D. R. E., Cotton, P. A., & Shipway, J. R. (2024). Chatting and cheating: Ensuring academic integrity in the era of ChatGPT. Innovations in Education and Teaching International, 61(2), 228-239.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. MIS Quarterly, 13(3), 319-340.

Dawes, J. (2008). Do data characteristics change according to the number of scale points used? An experiment using 5-point, 7-point and 10-point scales. International Journal of Market Research, 50(1), 61-104.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT 2019 (pp. 4171-4186).

Direttiva (UE) 2019/790 del Parlamento europeo e del Consiglio del 17 aprile 2019 sul diritto d'autore e sui diritti connessi nel mercato unico digitale (CDSMD).

Fanous, A., Goldberg, J., Agarwal, A., Lin, J., Zhou, A., Xu, S., Bikia, V., Daneshjou, R., & Koyejo, S. (2025). SycEval: Evaluating LLM Sycophancy. In Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025).

Ferrario, A., Loi, M., & Viganò, E. (2020). In AI We Trust Incrementally: A Multi-layer Model of Trust to Analyze Human-Artificial Intelligence Interactions. Philosophy & Technology, 33, 523-539.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In Studies in Linguistic Analysis (pp. 1-32). Oxford: Blackwell.

Følstad, A., & Brandtzæg, P. B. (2020). Users' experiences with chatbots: findings from a questionnaire study. Quality and User Experience, 5(1), 3.

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024). Bias and Fairness in Large Language Models: A Survey. Computational Linguistics, 50(3).

Gao, L., Biderman, S., Black, S., Golding, L., Hoppe, T., Foster, C., … & Leahy, C. (2020). The Pile: An 800GB dataset of diverse text for language modeling. arXiv:2101.00027.

Gao, Y., Xiong, Y., Gao, X., Jia, K., Pan, J., Bi, Y., Dai, Y., Sun, J., Wang, M., & Wang, H. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Gerlich, M. (2025). AI Tools in Society: Impacts on Cognitive Offloading and the Future of Critical Thinking. Societies, 15(1), 6.

Goodman, L. A. (1961). Snowball sampling. The Annals of Mathematical Statistics, 32(1), 148-170.

Gounopoulos, E. (2025). A Systematic Review of the Definition and Assessment of AI Literacy for Non-Experts.

Grattafiori, A., et al. (2024). The Llama 3 Herd of Models. arXiv:2407.21783.

Hadad, O., Loru, E., Nudo, J., Bonetti, A., Cinelli, M., & Quattrociocchi, W. (2026a). Wikipedia and Grokipedia: A Comparison of Human and Generative Encyclopedias. arXiv:2602.05519.

Hadad, O., Loru, E., Nudo, J., Di Marco, N., Cinelli, M., & Quattrociocchi, W. (2026b). The Statistical Signature of LLMs. arXiv:2602.18152.

Harris, Z. S. (1954). Distributional structure. Word, 10(2-3), 146-162.

Haugeland, I. K. F., Følstad, A., Taylor, C., & Bjørkli, C. A. (2022). Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design. International Journal of Human-Computer Studies, 161, 102788.

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world? Behavioral and Brain Sciences, 33(2-3), 61-135. https://doi.org/10.1017/S0140525X0999152X

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735-1780.

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2019). The curious case of neural text degeneration. arXiv:1904.09751.

Ibrahim, L., Hafner, F. S., & Rocher, L. (2026). Training language models to be warm can reduce accuracy and increase sycophancy. Nature.

Istat (2025). Rapporto annuale 2025. La situazione del Paese. Roma: Istituto nazionale di statistica. https://www.istat.it/wp-content/uploads/2025/05/Rapporto-Annuale-2025-integrale.pdf

Ježek, E., & Sprugnoli, R. (2023). Linguistica computazionale. Introduzione all'analisi automatica dei testi. Bologna: Il Mulino.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., et al. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al. (2023). Mistral 7B. arXiv:2310.06825.

Kagitcibasi, C., Goksen, F., & Gulgoz, S. (2005). Functional Adult Literacy and Empowerment of Women: Impact of a Functional Literacy Program in Turkey. Journal of Adolescent & Adult Literacy, 48(6), 472-489. https://doi.org/10.1598/JAAL.48.6.3

Kakarala, M. K., & Rongali, S. K. (2025). Data Privacy and Security in AI. World Journal of Advanced Research and Reviews, 25(3), 2517-2523.

Kandlhofer, M., Steinbauer, G., Hirschmugl-Gaisch, S., & Huber, P. (2016, October). Artificial intelligence and computer science in education: From kindergarten to university. In 2016 IEEE Frontiers in Education Conference (FIE) (pp. 1-9). IEEE.

Karamolegkou, A., Li, J., Zhou, L., & Søgaard, A. (2023). Copyright Violations and Large Language Models. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023).

Kim, S. S. Y., Liao, Q. V., Vorvoreanu, M., Ballard, S., & Wortman Vaughan, J. (2024). «I'm Not Sure, But…»: Examining the Impact of Large Language Models' Uncertainty Expression on User Reliance and Trust. In Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), 822-835.

Knipper, R. A., Knipper, C. S., Zhang, K., Sims, V., Bowers, C., & Karmaker, S. (2025). The Bias is in the Details: An Assessment of Cognitive Bias in LLMs. arXiv:2509.22856.

Kwesi, J., Cao, J., Manchanda, R., & Emami-Naeini, P. (2025). Exploring User Security and Privacy Attitudes and Concerns Toward the Use of General-Purpose LLM Chatbots for Mental Health. In Proceedings of the 34th USENIX Security Symposium. arXiv:2507.10695.

Lakoff, G., & Johnson, M. (1999). Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books.

Laupichler, M. C., Aster, A., Haverkamp, N., & Raupach, T. (2023). Development of the "Scale for the assessment of non-experts' AI literacy"–An exploratory factor analysis. Computers in Human Behavior Reports, 12, 100338.

Lenci, A., Auriemma, S., & Miliani, M. (2025). Linguistica computazionale. Natural language processing e intelligenza artificiale. Milano: Hoepli.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459-9474.

Li, M., & Zhang, X. (2021). A meta-analysis of self-assessment and language performance in language testing and assessment. Language Testing, 38(2), 189-218.

Li, Z., Pan, J., Liu, Q., Xi, Y., Zhou, Y., Jin, Y., Mao, R., & Chen, P. (2026). Does Sycophancy Change Decisions? Effect of LLM Sycophancy on AI-Assisted Decision-Making. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26).

Lin, S., Hilton, J., & Evans, O. (2022, May). TruthfulQA: Measuring how models mimic human falsehoods. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 3214-3252).

Lombardi Vallauri, E. (2019). La lingua disonesta. Contenuti impliciti e strategie di persuasione. Bologna: Il Mulino.

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-16).

Loru, E., Nudo, J., Di Marco, N., Santirocchi, A., Atzeni, R., Cinelli, M., Cestari, V., Rossi-Arnaud, C., & Quattrociocchi, W. (2025a). The simulation of judgment in LLMs. Proceedings of the National Academy of Sciences.

Loru, E., Nudo, J., Di Marco, N., Cinelli, M., & Quattrociocchi, W. (2025b). Decoding AI Judgment: How LLMs Assess News Credibility and Bias. arXiv:2502.04426.

Ma, S., Salinas, A., Henderson, P., & Nyarko, J. (2025). Breaking Down Bias: On The Limits of Generalizable Pruning Strategies. arXiv:2502.07771.

Maleki, N., Padmanabhan, B., & Dutta, K. (2024, June). AI hallucinations: a misnomer worth clarifying. In 2024 IEEE Conference on Artificial Intelligence (CAI) (pp. 133-138). IEEE.

Martin, K. D., & Zimmermann, J. (2024). Artificial intelligence and its implications for data privacy. Current Opinion in Psychology, 58, 101829.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. Academy of Management Review, 20(3), 709-734.

McCarthy, J. (2007). What is Artificial Intelligence. Available at http://www-formal.stanford.edu/jmc/whatisai/

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence. Dartmouth College.

Meng, X., & Liu, J. (2025). «Talk to me, I'm secure»: investigating information disclosure to AI chatbots in the context of privacy calculus. Online Information Review.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781.

Minsky, M. (1967). Computation: Finite and Infinite Machines. Englewood Cliffs, NJ: Prentice-Hall.

Minsky, M. (1986). The Society of Mind. New York: Simon & Schuster.

Mitchell, M. (2021). Why AI is harder than we think. arXiv:2104.12871.

Moravec, H. (1988). Mind Children: The Future of Robot and Human Intelligence. Cambridge, MA: Harvard University Press.

Neubauer, A., Wynn, M., & Bown, R. (2026). AI, Authorship, Copyright, and Human Originality. Encyclopedia, 6(1), 9.

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. Computers and Education: Artificial Intelligence, 2, 100041.

Nordheim, C. B., Følstad, A., & Bjørkli, C. A. (2019). An initial model of trust in chatbots for customer service—findings from a questionnaire study. Interacting with Computers, 31(3), 317-335.

Nudo, J., Pandolfo, M. E., Loru, E., Samory, M., Cinelli, M., & Quattrociocchi, W. (2026). Generative exaggeration in LLM social agents: Consistency, bias, and toxicity. Online Social Networks and Media, 51, 100344.

Nussbaum, Z., Morris, J. X., Duderstadt, B., & Mulyar, A. (2024). Nomic Embed: Training a Reproducible Long Context Text Embedder. arXiv:2402.01613.

Østergaard, S. D., & Nielbo, K. L. (2023). False Responses From Artificial Intelligence Models Are Not Hallucinations. Schizophrenia Bulletin, 49(5), 1105-1107. https://doi.org/10.1093/schbul/sbad068

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., et al. (2022). Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems, 35, 27730-27744.

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. Human Factors, 52(3), 381-410. https://doi.org/10.1177/0018720810376055

Pasquale, F. (2026). The Non-Delegable Duty to Think: Judicial Legitimacy and the Limits of Generative AI. Cornell Law School Legal Studies Research Paper No. 26-03.

Perez, E., Ringer, S., Lukošiūtė, K., Nguyen, K., Chen, E., Heiner, S., et al. (2022). Discovering Language Model Behaviors with Model-Written Evaluations. arXiv:2212.09251.

Pinski, M., & Benlian, A. (2023). AI Literacy — Towards Measuring Human Competency in Artificial Intelligence. In Proceedings of the 56th Hawaii International Conference on System Sciences.

Proietti, G. (2024). Definire l'indefinibile? I sistemi di intelligenza artificiale alla ricerca di un inquadramento sistematico. Contratto e impresa, 3, 832-925.

Quattrociocchi, W., Capraro, V., & Perc, M. (2025). Epistemological Fault Lines Between Human and Artificial Intelligence. arXiv:2512.19466.

Quattrociocchi, W., Capraro, V., & Marcus, G. (2026). Statistical approximation is not general intelligence. Nature, 659, 792.

Quintais, J. P. (2025). Generative AI, copyright and the AI Act. Computer Law & Security Review, 56, 106107.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. OpenAI.

Regolamento (UE) 2024/1689 del Parlamento europeo e del Consiglio del 13 giugno 2024 che stabilisce regole armonizzate sull'intelligenza artificiale (AI Act).

Rheu, M., & Cho, J. (2025). The Trap of AI Literacy: The Paradoxical Relationships Between College Students' Use of LLMs, AI Literacy, and Fact-checking Behavior. CHI Conference on Human Factors in Computing Systems Extended Abstracts.

Riguzzi, F. (2021). Introduzione all'Intelligenza Artificiale. arXiv:1511.04352.

Ruschemeier, H. (2023). AI as a challenge for legal regulation – the scope of application of the artificial intelligence act proposal. ERA Forum, 23, 361-376.

Salinas, A., Haim, A., & Nyarko, J. (2025). What's in a Name? Auditing Large Language Models for Race and Gender Bias. arXiv:2402.14875.

Savoldi, B., Attanasio, G., Gorodetskaya, O., Marchiori Manerba, M., Bassignana, E., Casola, S., Negri, M., Caselli, T., Bentivogli, L., Ramponi, A., Muti, A., Balbo, N., & Nozza, D. (2025). Pratiche, Alfabetizzazione e Divari nell'IA Generativa: Un'analisi empirica nel contesto italiano. Fondazione Bruno Kessler et al.

Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? In Advances in Neural Information Processing Systems 36 (NeurIPS 2023).

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M., & Perez, E. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548.

Somalvico, M. (1987). Intelligenza artificiale. Scienza & vita nuova.

Soto-Sanfiel, M. T., Angulo-Brunet, A., & Lutz, S. (2024). The scale of artificial intelligence literacy for all (SAIL4ALL): assessing knowledge of artificial intelligence in all adult populations. Humanities and Social Sciences Communications, 12(1), 1-14.

Sun, Y., & Wang, T. (2026). Be Friendly, Not Friends: How LLM Sycophancy Shapes User Trust. In Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26).

Tamkin, A., Brundage, M., Clark, J., & Ganguli, D. (2021). Understanding the capabilities, limitations, and societal impact of large language models. arXiv:2102.02503.

Thomson Reuters Institute (2025). Generative AI in Professional Services Report. Toronto: Thomson Reuters.

Tie, G., Zhao, Z., Song, D., Wei, E., Zhou, R., Dai, Y,... e Gao, J. (2025). Large Language Models Post-Training: Surveying Techniques From Alignment to Reasoning. arXiv:2503.06072.

Tran, S., Lu, H., Slaughter, I., Herman, B., Dangol, A., Fu, Y., Chen, L., et al. (2025). Understanding Privacy Norms Around LLM-Based Chatbots: A Contextual Integrity Perspective. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES). arXiv:2508.06760.

Turing, A. M. (1950). Computing machinery and intelligence. Mind, LIX(236), 433-460.

UNESCO (2021). Recommendation on the Ethics of Artificial Intelligence. Paris: UNESCO.

UNESCO (2024). AI Competency Framework for Students; AI Competency Framework for Teachers. Paris: UNESCO.

Varela, F. J., Thompson, E., & Rosch, E. (1991). The Embodied Mind: Cognitive Science and Human Experience. Cambridge, MA: MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., et al. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

Wang, B., Rau, P. L. P., & Yuan, T. (2023). Measuring user competence in using artificial intelligence: validity and reliability of artificial intelligence literacy scale. *Behaviour & information technology*, *42*(9), 1324-1337.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022). Emergent Abilities of Large Language Models. Transactions on Machine Learning Research.

Weizenbaum, J. (1966). ELIZA — A computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich, J., Rittberger, M., & Drachsler, H. (2021). Are we there yet? A systematic literature review on chatbots in education. Frontiers in Artificial Intelligence, 4, 654924.

Wong, G., Ma, X., Dillenbourg, P., & Huan, J. (2020). Broadening artificial intelligence education in K-12: where to start? ACM Inroads, 11(1), 20-29.

Wu, X., Wang, Y., Jegelka, S., & Jadbabaie, A. (2025). On the Emergence of Position Bias in Transformers. In Proceedings of the 42nd International Conference on Machine Learning (ICML 2025).

Yi, Y. (2021). Establishing the concept of AI literacy. JAHR, 12(2), 353-368.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., et al. (2023). A survey of large language models. arXiv:2303.18223.

SITOGRAFIA

Chroma, sito ufficiale. https://www.trychroma.com (ultima consultazione: giugno 2026).

Di Raimondo, G. E. (2026). Chatbot-AI-Literacy [repository GitHub]. https://github.com/gaiaeleonoradiraimondo/Chatbot-AI-Literacy (ultima consultazione: giugno 2026).

IBM, «AI bias» (IBM Think). https://www.ibm.com/it-it/think/topics/ai-bias (ultima consultazione: giugno 2026).

IBM, «Attention mechanism» (IBM Think). https://www.ibm.com/it-it/think/topics/attention-mechanism (ultima consultazione: giugno 2026).

IBM, «Large language models» (IBM Think). https://www.ibm.com/it-it/think/topics/large-language-models (ultima consultazione: giugno 2026).

IBM, «Text mining» (IBM Think). https://www.ibm.com/it-it/think/topics/text-mining (ultima consultazione: giugno 2026).

International Telecommunication Union (ITU), «Facts and Figures 2025 – Internet use». https://www.itu.int/itu-d/reports/statistics/2025/10/15/ff25-internet-use/ (ultima consultazione: giugno 2026).

LangChain, sito ufficiale. https://www.langchain.com (ultima consultazione: giugno 2026).

Ollama, sito ufficiale. https://ollama.com (ultima consultazione: giugno 2026).

Open Source Initiative, «Open Weights». https://opensource.org/ai/open-weights (ultima consultazione: giugno 2026).

Streamlit, sito ufficiale. https://streamlit.io (ultima consultazione: giugno 2026).