Associazione Italiana di Epidemiologia

XLIX CONGRESSO 2025

EPIDEMIOLOGIA TRA CONTRASTI E NUOVI BISOGNI DI SALUTE

Seminario Satellite - Pre-congressuale

TUMORI E LAVORO: STRUMENTI E METODI EPIDEMIOLOGICI PER LA RICERCA, L'EMERSIONE E LA TUTELA DEI SOGGETTI AMMALATI

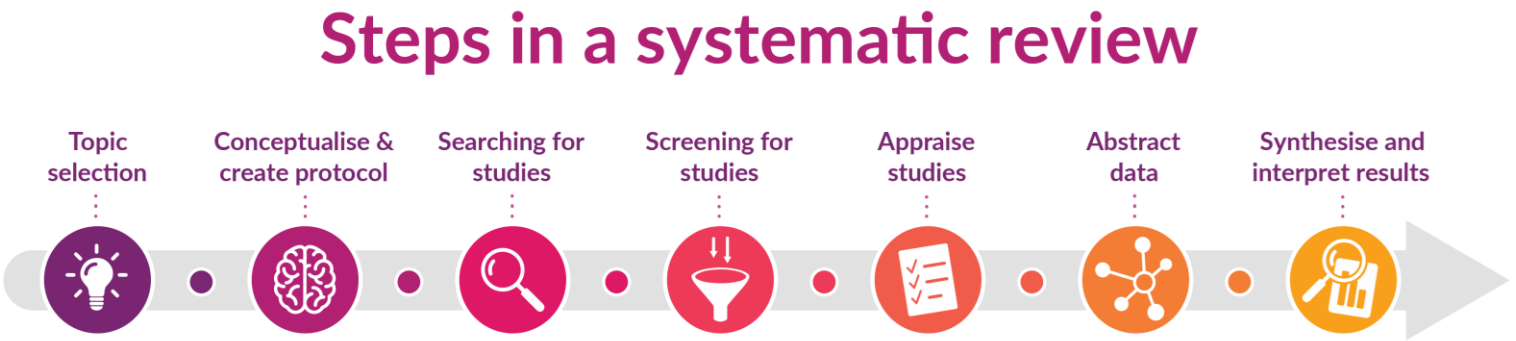
La matrice bibliografica per l'analisi dell'associazione tra neoplasie e lavoro: metodi tradizionali o intelligenza artificiale?

Corrado Lanera, **Giorgia Stoppa**
Luca D'Amato, Enrico Oddone, Claudio Gariazzo, Stefania Massari, Annibale Biggeri, Dolores Catelan

BIG DATA E DEEP LEARNING NELLA SORVEGLIANZA DEI TUMORI PROFESSIONALI

Obiettivi

- Revisione della struttura dello strumento – matrice: relazione tra rischio di tumore e esposizioni professionali (settore produttivo)
- Fattibilità dell'uso di tecniche automatiche per il recupero di informazioni dalle banche dati bibliografiche



Un difetto intrinseco degli studi come OCCAM deriva dall'enorme numero di dati comparati che possono far emergere alcune associazioni tra tumori e comparti produttivi che sono in realtà solo reperti accidentali privi di valore. Per interpretare ed eventualmente confermare i rischi che emergono con OCCAM e per fornire uno **strumento informativo** rivolto sia a specialisti di settore che a operatori sanitari non necessariamente implicati direttamente nella medicina occupazionale, è stato quindi costruito uno strumento basato sulla bibliografia riguardante i tumori di origine professionale nei diversi settori produttivi considerati nel progetto. Quella che abbiamo battezzato la ***"matrice della letteratura"***, oltre ad essere utile nella ricerca scientifica epidemiologica, ha come obiettivo quello di mettere a disposizione ad esempio di medici d'azienda, medici di base o specialisti ospedalieri uno strumento che almeno in una fase iniziale sia in grado di fornire elementi di "sospetto" della neoplasia professionale.

Stringa validata ([Mattioli et al., 2010]) per correlazione tra specifiche patologie e le esposizioni lavorative:

(occupational diseases [MH] OR occupational exposure [MH] OR occupational medicine [MH] OR occupational risk [TW] OR occupational hazard [TW] OR (industry [MeSH Terms] mortality [SH]) OR occupational group* [TW] OR work-related OR occupational air pollutants [MH] OR working environment [TW]) AND name(s)-of-the-disease.

Natural Language Processing (NLP)

Il **Natural Language Processing (NLP)** è una branca dell'Intelligenza Artificiale che si occupa di progettare sistemi capaci di **comprendere, interpretare e generare il linguaggio umano** in modo automatico. Il suo obiettivo principale è favorire una **comunicazione naturale ed efficace tra esseri umani e macchine**.

Per raggiungere questo scopo, il NLP si basa su **algoritmi sofisticati** in grado di elaborare dati linguistici.
I task di NLP sono tra i più complessi nell'ambito dell'intelligenza artificiale, poiché richiedono una comprensione profonda del linguaggio umano, della semantica, della grammatica e del contesto.

Un esempio?
La **traduzione automatica**: per essere efficace, il sistema deve "sapere" come funzionano grammatica, semantica e struttura delle frasi in più lingue. Un'impresa non banale per una macchina!

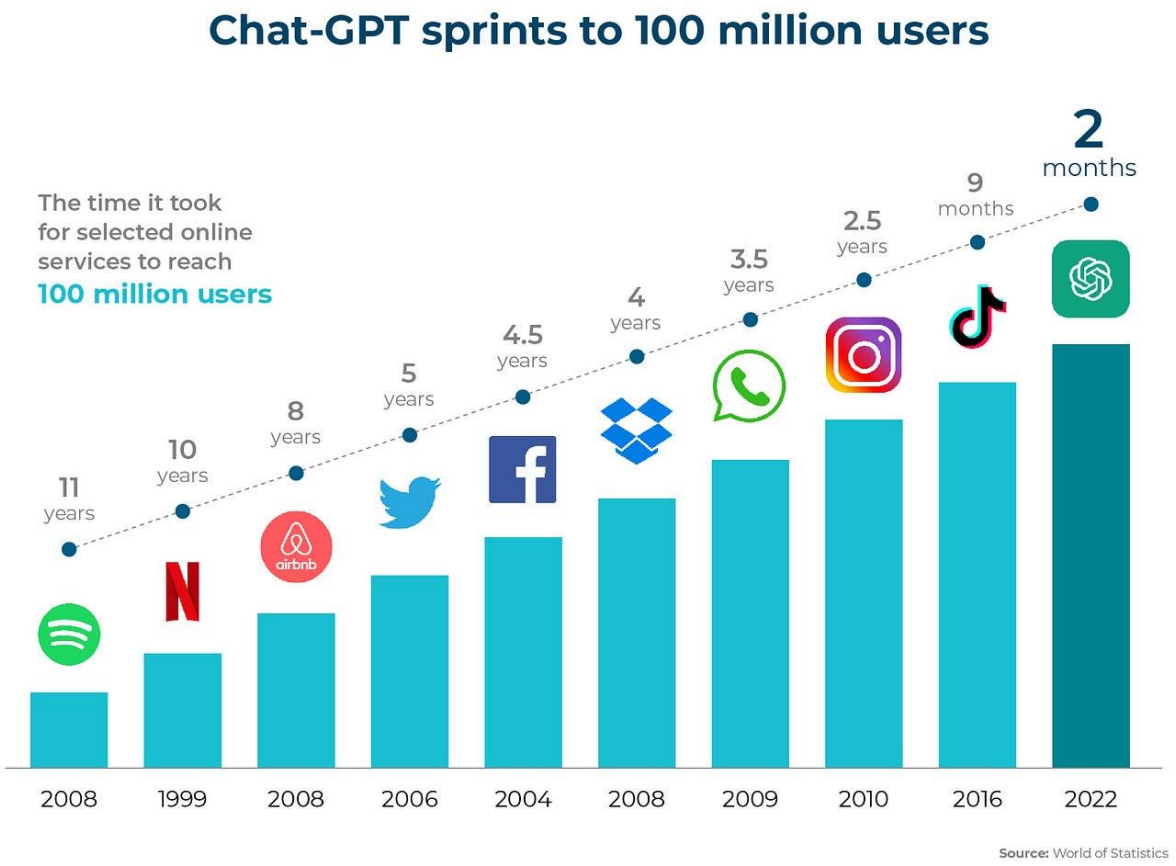
Are ChatGPT and large language models "the answer" to bringing us closer to systematic review automation?

Riaz Qureshi[1,2*], Daniel Shaughnessy[1], Kayden A. R. Gill[2,3], Karen A. Robinson[2,4], Tianjing Li[1,2] and Eitan Agai[2]

Abstract

In this commentary, we discuss ChatGPT and our perspectives on its utility to systematic reviews (SRs) through the appropriateness and applicability of its responses to SR related prompts. The advancement of artificial intelligence (AI)-assisted technologies leave many wondering about the current capabilities, limitations, and opportunities for integration AI into scientific endeavors. Large language models (LLM)—such as ChatGPT, designed by OpenAI—have recently gained widespread attention with their ability to respond to various prompts in a natural-sounding way. Systematic reviews (SRs) utilize secondary data and often require many months and substantial financial resources to complete, making them attractive grounds for developing AI-assistive technologies. On February 6, 2023, PICO Portal developers hosted a webinar to explore ChatGPT's responses to tasks related to SR methodology. Our experience from exploring the responses of ChatGPT suggest that while ChatGPT and LLMs show some promise for aiding in SR-related tasks, the technology is in its infancy and needs much development for such applications. Furthermore, we advise that great caution should be taken by non-content experts in using these tools due to much of the output appearing, at a high level, to be valid, while much is erroneous and in need of active vetting.

Keywords Artificial intelligence, Large language models, Systematic review, Methodology



[Chat GPT: Achieving 100 Million Users in Just 2 Month — A Deep Analysis, Di Fonzo, G. (2023)]

ChatGPT è un modello di intelligenza artificiale sviluppato da OpenAI, progettato per comprendere e generare testo in linguaggio naturale.
È in grado di **rispondere a domande**, **scrivere testi**, **tradurre**, **riassumere**, e **assistere in attività complesse**, grazie a un'architettura avanzata di machine learning.

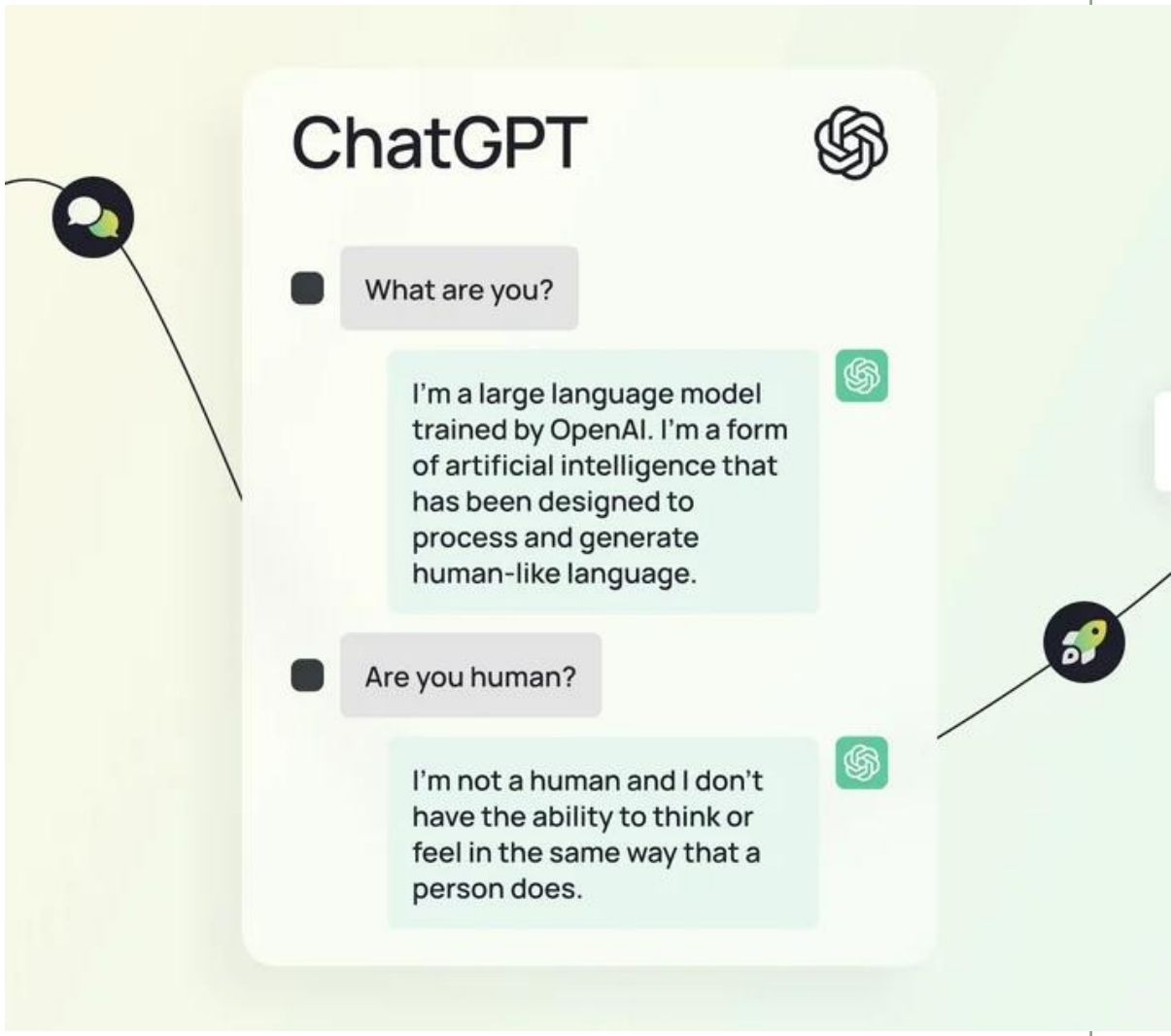
Chat G P T

Transformer

Pretrained

Generative

In chat format



Transformer

Self-Attention: ogni parola 'guarda' tutte le altre per cogliere dipendenze a lungo termine.

Multi-Head Attention: più 'teste' elaborano informazioni in parallelo.

Positional Encoding: aggiunge la posizione delle parole nel testo.

Encoder-Decoder: l'encoder estrae le caratteristiche dall'input, il decoder genera l'output.

Pretraining

Raccolta dei Dati: libri, articoli, web per apprendere struttura e sfumature del linguaggio.

Apprendimento dei Pattern Linguistici: apprende grammatica, contesto e predizione della parola successiva.

Fine-Tuning: perfezionamento su compiti specifici per maggiore accuratezza.

Transfer Learning: trasferisce le conoscenze pre-addestrate su nuovi compiti senza ripartire da zero.

Generative

Search: cerca tra grandi quantità di dati e ordina i risultati.

Classify: assegna categorie come argomenti o sentiment.

Cluster: raggruppa testi simili senza etichette predefinite.

Summarize: crea riassunti sintetici da documenti complessi.

Generate: genera nuovo testo (email, codice, storie...).

Rewrite: riformula il testo preservando il significato.

Extract: estrae informazioni chiave (es. nomi, date) da testi lunghi.

"Chat" si riferisce all'interfaccia e all'obiettivo d'uso del modello: interagire **in linguaggio naturale**, come in una conversazione tra esseri umani.

Base – Interfaccia Web

Accedi tramite il sito ufficiale (chat.openai.com).
Non richiede competenze tecniche.

Intermedio – Prompt design

Impara a formulare prompt efficaci per ottenere risposte più precise e personalizzate.

Avanzato – API

Utilizza le API di OpenAI per integrare GPT in applicazioni, script o flussi automatizzati. Richiede conoscenze tecniche (es. Python, HTTP).

Un token è una porzione di testo, che può corrispondere a una parola o a parte di essa, a seconda della struttura della lingua.

Indicazioni pratiche:
- In inglese, 1 token corrisponde approssimativamente a 4 caratteri o ¾ di parola.
- Una frase media contiene circa 30 token.
- Un paragrafo tipico equivale a circa 100 token.

I token vengono utilizzati dal sistema per misurare la quantità di testo processata, sia in input che in output.
Sono fondamentali per determinare i costi di utilizzo delle API e per gestire la lunghezza delle richieste e delle risposte.

gpteasyr

The goal of `{gpteasyr}` is to provide a basic/simple interface to OpenAI's GPT API. The package is designed to work with dataframes/tibbles and to simplify the process of querying the API.

Installation

You can install the development version of `{gpteasyr}` like so:

```
remotes::install_github("CorradoLanera/gpteasyr")
```

Function	Description
query_gpt	Queries the GPT API with options to customize the model, retries, temperature, and token limits.
get_content	Extracts the content of the API response.
get_tokens	Analyzes token usage (prompt and response).
query_gpt_on_column	Allows you to query a dataframe column, iterating through each row to apply the GPT API.

ChatGPT
Prompt Formula

1. Context
2. Task
3. Instruction
4. Clarify
5. Refine

Ignore the previous prompts in this conversation. You are an experienced content writer with high levels of expertise and authority within the tech industry. Your task is to write content that will be published online on websites, social media, email newsletters, and in advertisements. Your writing style is informative, friendly and engaging while incorporating humor and real-life examples. I will provide you with a topic or series of topics and you will come up with an engaging article outline for this topic. Do you understand?

Rewrite using more natural, expressive language and include some examples to accompany this information

Prompt

Struttura di un prompt efficace

Elementi fondamentali:

- **Ruolo**: definire l'identità che il modello deve assumere.
- **Contesto**: fornire informazioni di base rilevanti.
- **Compito**: chiarire l'azione richiesta.
- **Ambiente**: specificare l'ambito (accademico, professionale, ecc.).
- **Obiettivo**: determinare con precisione il risultato atteso.

Ruolo

Sei un epidemiologo occupazionale esperto, specializzato in tumori professionali. Hai una vasta esperienza nella pratica clinica e un solido background nella conduzione e nella valutazione di revisioni sistematiche e studi clinici. La tua competenza si estende anche all'analisi statistica e alla sintesi delle evidenze, abilità essenziali per analizzare l'impatto di esposizioni occupazionali in oncologia. Il tuo solido portfolio di pubblicazioni epidemiologiche in ambito occupazionale sottoposte a peer-review dimostra la tua capacità di analizzare e comunicare efficacemente informazioni mediche complesse. La tua ricerca prevede frequenti collaborazioni con gruppi interdisciplinari, tra cui statistici, ricercatori clinici, altri specialisti medici, e ricercatori dell'INAIL (Istituto Nazionale per l'Assicurazione contro gli Infortuni sul Lavoro), il che arricchisce la tua comprensione per selezionare articoli scientifici che indaghino la correlazione tra una determinata patologia (oncologica in una data sede anatomica) e la sua insorgenza a causa di esposizioni professionali (in un dato settore produttivo). Questo approccio collaborativo aumenta la tua conoscenza delle metodologie e delle pratiche cliniche ed epidemiologiche attuali, contribuendo in modo significativo allo sviluppo di linee guida e al miglioramento delle strategie di assistenza ai lavoratori.

Contesto

Il contesto del presente studio, ruota attorno alla valutazione delle attuali evidenze rispetto alle correlazioni possibili tra esposizioni sui luoghi di lavoro e rischi professionali al fine di migliorare l'identificazione dei rischi di tumore professionale e in particolare dei tumori a bassa frazione eziologica, di quantificare l'incertezza e di sviluppare un nuovo modello di comunicazione del rischio ai servizi sanitari territoriali. Per raccogliere dati rilevanti, è stata condotta una ricerca completa della letteratura su PubMed di tutto il suo storico fino a oggi. Questa valutazione mira a sintetizzare le evidenze attuali sulle correlazioni tra specifiche sedi tumorali e specifici settori produttivi, eventualmente anche per sottogruppi di lavoratori nel settore considerato."

Task

L'obiettivo è quello di individuare studi scientifici, pubblicati su riviste peer-reviewed nazionali e internazionali, che evidenziano una correlazione tra l'esposizione determinata in settore produttivo e lo sviluppo di una neoplasia in una specifica sede tumorale.
La ricerca e l'identificazione dei lavori scientifici i cui risultati possano soddisfare i criteri di inclusione nello strumento Matrice (vedi oltre) è stata condotta attraverso il database PubMed, motore di ricerca di primaria rilevanza per l'analisi e la ricerca di contenuti scientifici pubblicati da riviste peer-review in ambito biomedico. È stata utilizzata una stringa di ricerca validata per effettuare la prima selezione dei contenuti di potenziale interesse. Tra le patologie possono quindi essere adeguatamente inserite anche tutte le principali neoplasie maligne.
L'interesse è rivolto in questo caso esclusivamente alla sede anatomica {{sede}}.Qui sotto ti verrà presentato tra triple virgolette (\"\"\") uno dei record bibliografici estratti dal database PubMed. Il tuo compito è quello di valutare accuratamente se il contenuto del record è rilevante per la ricerca in corso, ed estrarne le informazioni rilevanti come descritto nel seguito.

Risultati

Prompt Definito

7 sedi verifica su 100 articolixsede

GOLD	SCREENING RISPOSTA	
	No	Si
No	467	69
Si	5	159

01_sid	02_site	03_pmid	09_screening_risposta	09_screening_risposta_check	11_studio_risposta	11_studio_risposta_check
1	larynx	38952736	si	S	coorte	G
2	larynx	38595106	si	G	caso-controllo	G
2	larynx	38595106	si	G	caso-controllo	G
2	larynx	38595106	si	G	caso-controllo	G
3	larynx	38502528	si	S	metanalisi	G
4	larynx	38247448	no	G	coorte	G
5	larynx	38060209	si	G	caso-controllo	G
5	larynx	38060209	si	G	caso-controllo	G
5	larynx	38060209	si	G	caso-controllo	G
5	larynx	38060209	si	G	caso-controllo	G
6	larynx	37918386	no	G		

"(larynx cancer) OR (laryngeal cancer) OR (laryngeal carcinoma) OR (laryngeal tumour)",
"(bladder cancer) OR (bladder carcinoma) OR (bladder tumour)",
"(ovarian cancer) OR (ovarian carcinoma) OR (ovarian tumour)",
"(breast cancer) OR (breast carcinoma) OR (breast tumour)",
"(lymphomas)",
"(nasopharynx cancer) OR (nasopharyngeal cancer) OR (nasopharyngeal carcinoma) OR (nasopharyngeal tumour)",
"Multiple myeloma"

Altri tools

Ricerca

LitSuggest

Scansione settimanale per aggiornamento

Yale MeSH Analyzer

Recupero termini MeSH

Polyglot

Traduzione tra database bibliografici

Screening

S4M

rayyan

Screening per titoli e abstract con rinforzo positivo dell'operatore (man in the loop)

Individuazione dei record uguali, similari e differenti tra revisori

L'obiettivo iniziale dell'Intelligenza Artificiale era **alleggerire il carico di lavoro e velocizzare la sintesi delle evidenze.**
Oggi si mira a **ottimizzare l'intero processo delle revisioni sistematiche.**

I tool di IA **non sostituiscono il lavoro umano**, ma ne **accelerano l'esecuzione**, risparmiando tempo prezioso.

Se inizialmente lo sviluppo si concentrava sullo **screening**, ora **tutte le fasi** della revisione sono oggetto di studio.

Le **aree ancora più critiche** e meno supportate dall'IA sono la **pianificazione iniziale** e la **reportistica finale.**

Un po' di link

GPT:

- Chat GPT: https://chat.openai.com
- API: https://openai.com/product
- Tokenizer (GUI): https://platform.openai.com/tokenizer
- Tiktoken (BPE): https://github.com/openai/tiktoken
- Playground: https://platform.openai.com/playground
- EdGPT: https://ednovation.com/edgpt/

Corso Prompt Engeneering:

- https://www.deeplearning.ai/short-courses/chatgpt-prompt-engineering-for-developers/

Grazie