Title

Large Language Model in Medical Information Extraction from Titles and Abstracts with Prompt Engineering Strategies: A Comparative Study of GPT-3.5 and GPT-4

Authors:

Yiyi Tang [a,b,#], Ziyan Xiao [b,c,#], Xue Li [a,c,d,#], Qingpeng Zhang [c,e], Esther W Chan [c,d], Ian CK Wong [c,d], Research Data Collaboration Task Force [h]

[a] Department of Medicine, School of Clinical Medicine, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

[b] Department of Statistics and Actuarial Science, Faculty of Science, The University of Hong Kong, Hong Kong SAR, China

[c] Department of Pharmacology and Pharmacy, Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

[d] Laboratory of Data Discovery for Health ($D^2$4H), Hong Kong Science Park, Hong Kong SAR, China

[e] Musketeers Foundation Institute of Data Science, The University of Hong Kong, Hong Kong SAR, China

[h] Li Ka Shing Faculty of Medicine, The University of Hong Kong, Hong Kong SAR, China

Co-first author with equal contribution
* Corresponding author

Xue Li
Assistant Professor
Department of Medicine and Department of Pharmacology & Pharmacy
LKS Faculty of Medicine, The University of Hong Kong
PB306, 3/F, Professional Block, Queen Mary Hospital
102 Pok Fu Lam Road, Hong Kong
Tel: +852 2255 3319
Email: sxueli@hku.hk

Co-author:

Yiyi Tang: yiyitang@connect.hku.hk

Ziyan Xiao: xiaozy@connect.hku.hk

Qingpeng Zhang: qpzhang@hku.hk

Esther W Chan: ewchan@hku.hk

Ian CK Wong: wongick@hku.hk

Author contributions:
Study concept and design: YT, ZX, XL, EWC, QZ, ICKW.
Information collecting and screening: YT, ZX.
Data extraction, analysis, and cross-checking: YT, ZX.
Drafting of the manuscript: XL, YT, ZX.
Data interpretation: all authors.
Critical revision of the manuscript of significant intellectual contribution: all authors.
Funding acquisition: XL
Study supervision: XL, QZ.

Abstract

Background: Large language models (LLMs) have significantly enhanced the Natural Language Processing (NLP), offering significant potential in facilitating medical literature review. However, the accuracy, stability and prompt strategies associated with LLMs in extracting complex medical information have not been adequately investigated. Our study assessed the capabilities of GPT-3.5 and GPT-4.0 in extracting or summarizing seven crucial medical information items from the title and abstract of research papers. We also validated the impact of prompt engineering strategies and the effectiveness of evaluating metrics.

Methodology: We adopted a stratified sampling method to select 100 papers from the teaching schools and departments in the LKS Faculty of Medicine, University of Hong Kong, published between 2015 and 2023. GPT-3.5 and GPT-4.0 were instructed to extract seven pieces of information, including study design, sample size, data source, patient, intervention, comparison, and outcomes. The experiment incorporated three prompt engineering strategies: persona, chain-of-thought and few-shot prompting. We employed three metrics to assess the alignment between the GPT output and the ground truth: BERTScore, ROUGE-1 and a self-developed GPT-4.0 evaluator. Finally, we evaluated and compared the proportion of correct answers among different GPT versions and prompt engineering strategies.

Results: GPT demonstrated robust capabilities in accurately extracting medical information from titles and abstracts. The average accuracy of GPT-4.0, when paired with the optimal prompt engineering strategy, ranged from 0.688 to 0.964 among the seven items, with sample size achieving the highest score and intervention yielding the lowest. GPT version was shown to be a statistically significant factor in model performance, but prompt engineering strategies did not exhibit cumulative effects on model performance. Additionally, our results showed that the GPT-4.0 evaluator outperformed the ROUGE-1 and BERTScore in assessing the alignment of information (Accuracy: GPT-4.0 Evaluator: 0.9714, ROUGE-1: 0.9429, BERTScore: 0.8714).

Conclusion: Our result confirms the effectiveness of LLMs in extracting medical information, suggesting their potential as efficient tools for literature review. We recommend utilizing an advanced version of LLMs to enhance the model performance, while prompt engineering strategies should be tailored to the specific tasks. Additionally, LLMs show promise as an evaluation tool to assess the model performance related to complex information processing.

Introduction

Large language models (LLM), including the GPT series, have emerged as a promising tool to revolutionize many domains in medicine [1-2]. LLMs distinguished themselves from traditional natural language processing models by their ability to generate responses that align with users' expectations [3], without requiring dedicated fine-tuning on specialized tasks [4]. Medical evidence summarization is one of these areas where GPT is promising to improve the traditional process of extracting information from the vast amount of medical research papers [5-7].

Medical literature screening is one of the major application domains of automatic medical information summarization and extraction. Before the advent of ChatGPT, one prominent approach to streamlining the screening process involved recommending articles based on relevance, thereby facilitating the prioritization of manual screening or providing suggestions for inclusion and exclusion [8]. Numerous software tools have been developed to realize these functions, such as Rayyan [9], RobotReviewer [10], and SWIFT-Review [11]. These tools utilized machine learning techniques related to Natural Language Processing (NLP) [12], and the Supportive Vector Machine (SVM) was the prevailing algorithm [13]. Research has also demonstrated the cost efficiency of employing these automated tools with text-mining-based single screening, reducing workload by over 60% compared to alternative methods [14]. However, the reliability and transparency of applying artificial intelligence to literature review have long been a concern. At the current stage, multiple studies have highlighted the necessity of incorporating manual screening while leveraging the existing tools [12-13]. The updated PRISMA guidance for reporting systematic reviews also expressed concerns about the erroneous exclusion of relevant studies using machine learning tools while recognizing its role in priority screening [15].

LLMs exhibit the potential to expand functionalities and enhance accuracy beyond existing tools. A recent study conducted by Matsui et al. (2023) has demonstrated that, with an appropriate prompt setting, GPT-3.5 can achieve exceptional sensitivity in literature screening, close to human evaluators, and even surpass them when confronted with a massive volume of articles.[16] LLMs can also be applied to less-explored steps of medical literature review, such as the risk of bias assessment and data extraction.[17] In another experiment, Hill et al. (2023) showcased the accurate performance of Microsoft Bing AI in retrieving information and constructing study characteristics tables from full-text PDF files.[18] Additionally, Shaib et al. (2023) attempted to synthesise medical evidence from multiple documents, although the current results are less optimistic than summarisation from single documents.[5]

Despite the promising potential of LLMs in the literature review, there remain a need for comprehensive empirical research addressing common concerns surrounding LLMs, including their evidence level and consistency [19]. The ability to extract complex information from research papers is essential in validating the reliability of LLMs for assisting medical literature review. Furthermore, while prompt engineering was recommended as a useful strategy to increase the performance of ChatGPT [20], its role in the context of medical literature review has yet to be thoroughly investigated.

In this study, we aim to design an experiment to answer the above questions. The research objectives of this study can be summarised as follows.

1. To implement and assess the capability of large language models in extracting critical information from titles and abstracts of medical research papers.
2. To compare GPT-3.5 and GPT-4 performance in the aforementioned task.
3. To validate the effects of prompt engineering strategies on the performance improvement of LLMs in the aforementioned tasks.

Methodology

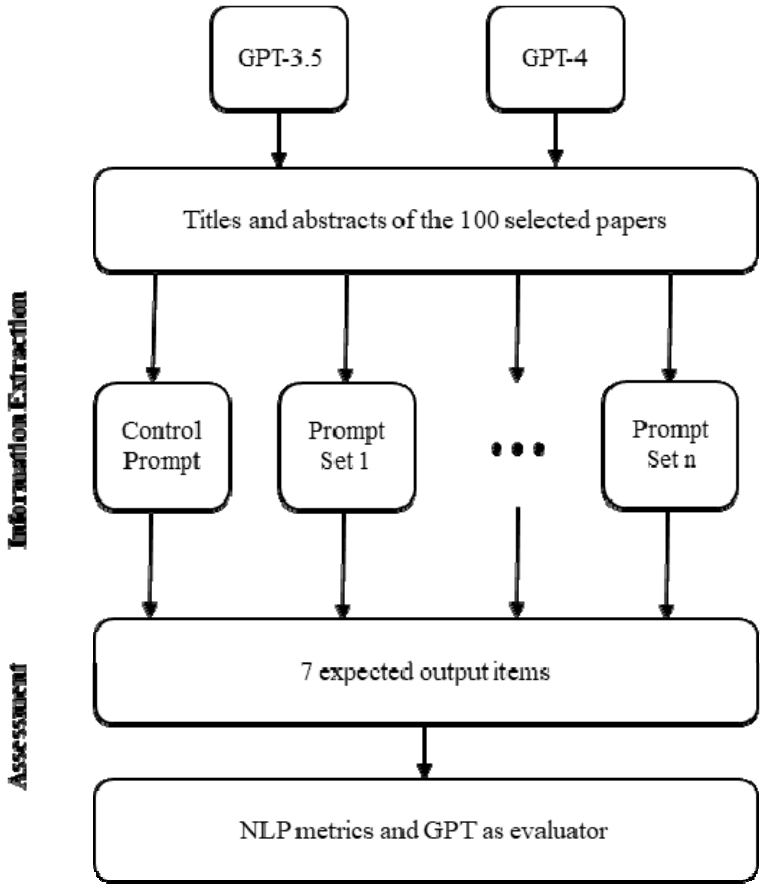
Study design

The scope of this study encompassed 100 research papers randomly selected from the publication pool of the Li Ka Shing Faculty of Medicine, University of Hong Kong. The chosen papers consisted solely of original research articles published between January 1, 2015, and December 31, 2023, with their titles and abstracts fully available on Scopus. To ensure representativeness, we adopted a stratified sampling method to randomly select papers from each department in proportion to the total number of publication records affiliated with that department. The paper's affiliation is the institution affiliated with the corresponding author. The departments and their related domains of the paper selected are presented in Supplementary Material 1.

Figure 1 presents the overall study design. All titles and abstracts were obtained from the Scopus online dataset and pre-processed to remove unreadable characters. Two researchers manually labeled the information to be extracted according to pre-defined criteria, to obtain the ground truth. To ensure accuracy, we employed cross-checking of their results to establish the ground truth.

Subsequently, the titles and abstracts were proceeded to GPTs to extract information. We implemented several prompt sets to compare the effectiveness of prompt engineering. The assessment of the information extraction performance was based on semantic similarity between GPT's output and ground truth, measured by several NLP metrics and a self-developed independent GPT evaluator. Finally, we perform a statistical analysis on the results.

Figure 1. Flowchart of overall study design;



To compare the performance of GPT-3.5 and GPT-4.0, our study conducts independent evaluations using the latest model versions at the time of study: **gpt-3.5-turbo-0125** and **gpt-4-0125-preview**, referred to as GPT-3.5 and GPT-4.0 in later script. These models

represent the most advanced version of their respective series and are provided by OpenAI through the API platform. Experiments will be executed using Python scripts to interact with the OpenAI API. Each model will receive prompts via individual API requests without the conversation history, maintaining the independence of each interaction and preventing prior context from influencing the model's performance. All experiments would be repeated five times to evaluate the performance stability. In total, there are 8,000 experiments.

This design aims to yield a fair and thorough comparison of the two models, highlighting their respective strengths and limitations in processing and analyzing medical research literature.

Information Extraction

Information extraction is a pivotal stage in a literature review. It not only facilitates the identification of related papers, but also has the potential to enhance the transparency of LLM's decision as an intermediate step in automatic literature screening. In this study, we identified seven important items in literature screening as representative samples, including study design, sample size, data source, and PICOs (Patient, Intervention, Comparison Outcomes). Their respective definitions are provided in Table 1.

Table 1. Definition of Information to extract (Duke University, 2019) [21]

Item	Definition
Study design	Type of study, such as randomized controlled trail, cohort study, case-control study and systematic review
Sample size	The number of participants involved in the study, and the basic characteristics of the participants.
Data source	Source of the experimental data, such as databases, previous studies or surveys
Patient	The patient involved in the experiment with some most important characteristics of patients
Intervention	Main intervention, exposure, or prognostic factor in the experiment
Comparison	Main alternative group being considered.
Outcomes	The outcome that the experiment trying to accomplish, measure, improve or affect.

We believe these elements are the basis of efficient and precise literature screening, providing researchers with a clear and standardized framework for evaluation. Particularly, the PICOs, as the gold standard for clinical study assessment, offer a systematic approach to identify relevant research questions and assess the quality of studies.

Validation on the effects of Prompt Engineering

Prompt engineering is an essential mechanism for optimizing the interaction with LLMs, serving to refine and enhance users' query in order to improve the performance on tasks. In this section, we will examine and identify the effect of several prompt engineering strategies discussed in current directions of research, including *Adopting a Persona*, *Chain of Thought* and *Few-shot Learning*.

Adopting a Persona [22-23] is often achieved by instructing the LLMs to adopt the role of an expert in the related field of research. *Chain of Thought* [24-25] asks the model to explain the reasoning or the rationale behind each step in its problem-solving process. Though our task may not involve complicated logical reasoning, we are interested in investigating whether incorporating requests for justification could lead to improved performance and greater transparency. *Few-shot Learning* [26-27] refers to the process in which we provide LLMs with expert output examples for similar tasks, which can serve as a guide for the model's responses.

We adopted the following approach for our study. We first established a standard prompt without any specialized engineering strategy to serve as a control. This prompt simply asked the LLM to perform the task without additional instruction or context. We then selected three prompt engineering strategies as mentioned previously. For each strategy, we crafted a series of prompts that incorporated the specific tactic. After that, we systematically removed one strategy at a time from the prompts, creating various ablated conditions for comparison against the baseline prompt and each other. For each prompt condition, we then evaluate the LLM's performance using several metrics.

The table below outlines the specific prompts that have been designed for each of these prompt engineering strategies.

Table 2 Prompt Setting for information extraction

Group	Prompt
Control	# Context [a] `You will be provided with titles and abstracts of medical papers, and your task is to parse it into structured data, including Study Design, Data Source, Sample Size, Patient, Intervention, Comparison and Outcomes, and separate them by semicolon.` # Input [insert paper title and abstract] # Instruction `Please read, extract and concisely report the following key details from the abstract:` `Study Design: What type of methodology was employed in the study?`

	Sample Size: How many participants were included in the study? Data Source: Where was the data for this study sourced from? Based on the Study Design, if the paper is a review paper OR a laboratory study, please marks Patient, Intervention, Comparison and Outcomes all as NA. Else, answer the following PICO question: - Patients: Who is the study's targeted patient or population group? - Intervention: What is the key intervention that the study assesses? - Comparison: Is there a comparison group or control used, and what does it consist of? - Outcomes: What outcomes are being measured to determine the intervention's success? Answer "NA" if any of the item is not mentioned in the abstract. # Output Please [b1] output the structured data separated by semicolon, such as: [b2] Study Design: [output]; Data Source: [output]; Sample Size: [output]; Patient: [output]; Intervention: [output]; Comparison: [output]; Outcomes: [output]; [c]
Strategy 1: Persona	# Inserted at [a] Imagine you are an expert in research methodology. Your role is essential in supporting a team of researchers by meticulously extracting critical information from medical paper abstracts. You have been trained to identify and collate specific elements that are crucial for the team's meta-analysis and database entry tasks.
Strategy 2: Chain of thought	# Inserted at [b1] present a concise reasoning for each step you take, and how you arrive at the final structured data. Also, please # Inserted at [b2] Reasoning: [output]; # Inserted at [b3] Reasoning: The abstract explicitly indicates that the study is a retrospective cohort study. The sample size is explicitly mentioned, consisting of three distinct groups with their respective counts. The data source is not explicitly named, so we mark it with NA. Since this is a cohort study (an epidemiological study) instead of a review paper or a laboratory study, we proceed with identifying the PICO elements. The patient population is women with PCOS, PCO, and age-matched controls undergoing IVF. The intervention is the IVF treatment itself. The comparison is made between the women with PCOS, those with PCO, and the age-matched controls. The outcomes being measured include various obstetric complications and outcomes such as GDM, GHT, PET, IUGR, gestation at delivery, baby's Apgar

	scores, and NICU admissions;
Strategy 3: Few-shot Prompting	# Inserted at [c] Here is an example for your reference: # Input Title: Obstetric outcomes in women with polycystic ovary syndrome and isolated polycystic ovaries undergoing in vitro fertilization: a retrospective cohort analysis Abstract: Objective: This retrospective cohort study evaluated the obstetric outcomes in women with polycystic ovary syndrome (PCOS) and isolated polycystic ovaries (PCO) undergoing in vitro fertilization (IVF) treatment. Methods: We studied 104 women with PCOS, 184 with PCO and 576 age-matched controls undergoing the first IVF treatment cycle between 2002 and 2009. Obstetric outcomes and complications including gestational diabetes (GDM), gestational hypertension (GHT), gestational proteinuric hypertension (PET), intrauterine growth restriction (IUGR), gestation at delivery, baby's Apgar scores and admission to the neonatal intensive care unit (NICU) were reviewed. Results: Among the 864 patients undergoing IVF treatment, there were 253 live births in total (25 live births in the PCOS group, 54 in the PCO group and 174 in the control group). The prevalence of obstetric complications (GDM, GHT, PET and IUGR) and the obstetric outcomes (gestation at delivery, birth weight, Apgar scores and NICU admissions) were comparable among the three groups. Adjustments for age and multiple pregnancies were made using multiple logistic regression and we found no statistically significant difference among the three groups. Conclusion: Patients with PCO±PCOS do not have more adverse obstetric outcomes when compared with non-PCO patients undergoing IVF treatment. © 2014 Informa UK Ltd. All rights reserved: reproduction in whole or part not permitted. # Output [b3] Study Design: Retrospective cohort study; Sample Size: 864; Data Source: NA; Population: Women with polycystic ovary syndrome (PCOS) and isolated polycystic ovaries (PCO); Intervention: In vitro fertilization (IVF) treatment; Comparison: Age-matched controls; Outcomes: Obstetric complications (GDM, GHT, PET and IUGR) and the obstetric outcomes (gestation at delivery, birth weight, Apgar scores and NICU admissions);

Evaluation

To evaluate the accuracy of the generated outcomes, we employed the established automatic metrics in NLP, including ROUGE-1[28] and BERTScore[29]. These metrics were specifically designed to measure the quality of generated text compared to the reference text produced by human. ROUGE-N, a metric based on n-gram analysis, examined the overlap of common

words and phrases between the two summaries. On the other hand, BERTScore uses contextual embeddings from a pre-trained large language model to derive a similarity score between the generated and reference texts. Unlike N-gram (ROUGE-1) method that relies on exact matches, BERTScore can account for semantic similarities at the word and sentence level. For both metrics, we utilize the F1-score – which is the harmonic mean of the precision and recall scores – as our final standard for analysis. The score ranges from 0 to 1, and is calculated as

$$\text{F1-score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where

$$\text{Precision}_{\text{ROUGE}} = \frac{\text{1-gram Hyp.} \cap \text{1-gram Ref.}}{|\text{Hyp.}|}$$

$$\text{Recall}_{\text{ROUGE}} = \frac{\text{1-gram Hyp.} \cap \text{1-gram Ref.}}{|\text{Ref.}|}$$

And

$$\text{Precision}_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} \left(\mathrm{x}_i^{\mathrm{T}} \hat{x}_j \right)\}$$

$$\text{Recall}_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} \left(\mathrm{x}_i^{\mathrm{T}} \hat{x}_j \right)$$

in which x_i represents the reference token and $\hat{x}_j$ represents the candidate token (hypothesis).

Noticeably, recent research papers highlighted the inherent challenges in assessing the responses of LLM using traditional automatic metrics in NLP. Consequentially, we also designed an independent evaluation mechanism in the form of a separate and specifically tailored GPT algorithm to evaluate whether the generated responses correspond with the ground truth. To improve the stability of GPT's measurement, the temperature of this machine was set to 0, and no history of previous output was incorporated as input. The detailed prompt is outlined below.

Prompt of evaluation
Input `Hypothesis: {h}` `Reference: {r}` # Query `You will be given a hypothesis and a reference that represent the {element} element extracted from a medical paper. Your task is to compare the semantic similarity between the hypothesis and the reference. The similarity score should be between 0 and 1, where 0 means no similarity and 1 means the highest similarity`

```
# Output
Please output your similarity score in the format: The similarity score is {your_score}
```

To ensure the validity of the evaluators, a cross-evaluation was performed on the element extracted by two independent researchers. Specifically, an accordance dataset is first produced by manually comparing the results for each label pair in the 100 papers, in which a score of 1 is assigned if the labels from both researchers matched (indicating agreement), and a score of 0 if they differed (indicating disagreement). Subsequently, each evaluator would compute the similarity for each pair of labels, which will then be compared against the true accordance data to assess the consistency.

During this process, we also calculated threshold values for the metric score produced by the evaluators for each element category, in order to define what constitutes an acceptable level of agreement. Specifically, we iterate over the potential threshold value from 0 to 1 with stepsize of 0.01, and assign a True prediction for metric scores above the threshold, False for scores below. Then, we determine which threshold yields the highest accuracy rate of F1-score across all comparisons between the evaluators and the accordance ground truth, and select it as the eventual standard.

To assess the overall performance of the models, we employed this predefined threshold to calculate the accuracy, sensitivity and specificity in GPT's information extraction results.
We defined the accuracy rate $p_{correct}$ as the proportion of GPT's outputs that align with the ground truth in the five repetitive trials. It is calculated separately across the 100 papers as follows

$$\overline{p}_{correct} = \frac{1}{100}\sum_{i=1}^{100}\frac{\sum_{t=1}^{5}\max_{s\in S}(0,\ (s_{t,i} - \text{threshold}_s))}{5}$$

where $s_{t,i}$ is the metric score for the i^{th} paper in trial t, and threshold_s is the threshold calculated for the specific element nature. The average $p_{correct}$ was employed to horizontally compare the GPT models and prompt engineering strategies.

Given the risk of hallucination (producing information not grounded in the source material) and the possibility that not all elements of interest are present in a given abstract, we further define the following standard for a deeper analysis in sensitivity and specificity:

- True Positive (TP): An element that is both present in the abstract as labeled in groundtruth and correctly extracted by ChatGPT.
- True Negative (TN): An element that is neither present in the abstract nor falsely identified by ChatGPT.
- False Positive (FP): An element that ChatGPT incorrectly reports as present in the abstract (hallucination).
- False Negative (FN): An element that is present in the abstract but is missed or wrongly

extracted by ChatGPT.

And sensitivity and specificity is calculated as

$$\text{sensitivity} = \frac{TP}{TP + FN}$$

$$\text{specificity} = \frac{TN}{TN + FP}$$

which measures the proportion of actual positives and negatives that are correctly identified by the GPT, respectively.

Statistical Methods

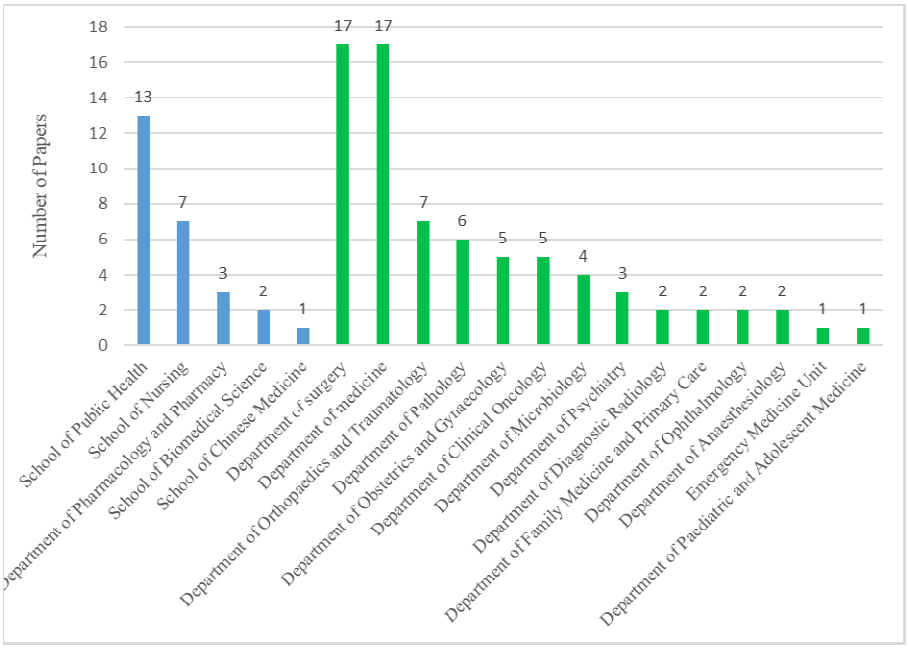
For each extracted item evaluated by one metric, a 2-way Analysis of Variance (ANOVA) model was used to analyze the impact of two factors, GPT versions and prompt engineering strategies. We summarized all p-values across items and evaluators in one table, to analyze the significance of GPT model and prompts effects on the performance. The statistical analysis was performed using the python package statsmodels (version 0.14.1) [30]. All significance level was set as 0.05, with all necessary assumptions for ANOVA, including normality and homogeneity of variances, being assessed and satisfied.

Results

Paper selection and Data source

Figure 2 represents the distribution of paper affiliation as a result of stratified sampling. The collected dataset covered the six teaching schools in HKU LKS Faculty of Medicine, Public Health, Nursing, Pharmacology and Pharmacy, Biomedical Sciences, Chinese Medicine and Clinical Medicine, and it included at least one paper for each school. Among them, the School of Clinical Medicine took 74% of the papers, and thus the affiliation of these papers was traced down to clinical departments, represented in green in Figure 2.

Figure 2. Paper affiliation distribution



The selected papers also provided comprehensive coverage across study types, sample size, publication year, and abstract length. The labeled ground truth indicated that the dataset consisted of 22 retrospective studies, 13 laboratory studies, 10 prospective studies, 7 case reports, 5 reviews, 4 randomized controlled trials, and other types of study design. The sample size ranged from 0 to 229428, with seven papers analyzing over 10,000 samples. Publication time was evenly distributed from 2015 to 2023. The average length of the inputted abstract was 1719.54 characters, with a standard deviation of 427.85.

Evaluator Performance

The process of selecting the optimal threshold is presented in supplementary material 2.

We randomly select the content and metric scores from three evaluators of 10 paper * 7 elements from different combination of models (GPT-3.5, GPT-4.0), prompt types and trials, and manually marked down the accordance between extracted information and ground truth as a test set. Table 3 presented are performance metrics for three different evaluators to assess the quality of semantic similarity rating, and they are compared based on their accuracy, precision, and recall.

Table 3. Accuracy, precision and recall of evaluators.

	BERT	**ROUGE-1**	**ChatGPT-4.0**
Accuracy	0.8714	0.9429	0.9714
Precision	0.8983	0.9643	0.9821
Recall	0.9464	0.9643	0.9821

The ChatGPT Evaluator outperforms the other two evaluators in all three metrics. It is the most

accurate and provides the highest precision and recall. The BERTScore Evaluator has the lowest accuracy and precision but maintains a high recall. The lower precision could imply that while BERT can identify many of the true positives, it may also mistakenly label some distinct elements as similar. Its relatively lower accuracy suggests that it doesn't perform as consistently across all cases compared to the other evaluators.

The high recall rates across all evaluators suggest that they are generally good at identifying relevant, semantically similar elements. In contrast, the variation in precision and accuracy indicates differences in their ability to exclude non-similar elements and correctly label the full range of elements, respectively.

Overall Performance

In our experiment, GPTs achieve considerable accuracy in extracting information from papers across medical disciplines. Measured by the BERTScore, GPT-4.0 achieves over 80% correctness in five out of the seven items, and GPT-3.5 achieves over 80% in three items. Supplementary Material 3 includes a comprehensive table summarizing the average proportions of correctness, covering all 7 items under 8 prompt settings, generated by GPT-3.5 and GPT-4.0 and measured by the three different metrics. For clarity, Table 4 selectively presented each item's optimal performance and corresponding prompt engineering strategy. The results also compare the optimal performance between GPT-3.5 and GPT4.0.

Table 4. Optimal performance for each item and the corresponding prompt engineering strategies. Mean and standard deviation across the five trials.

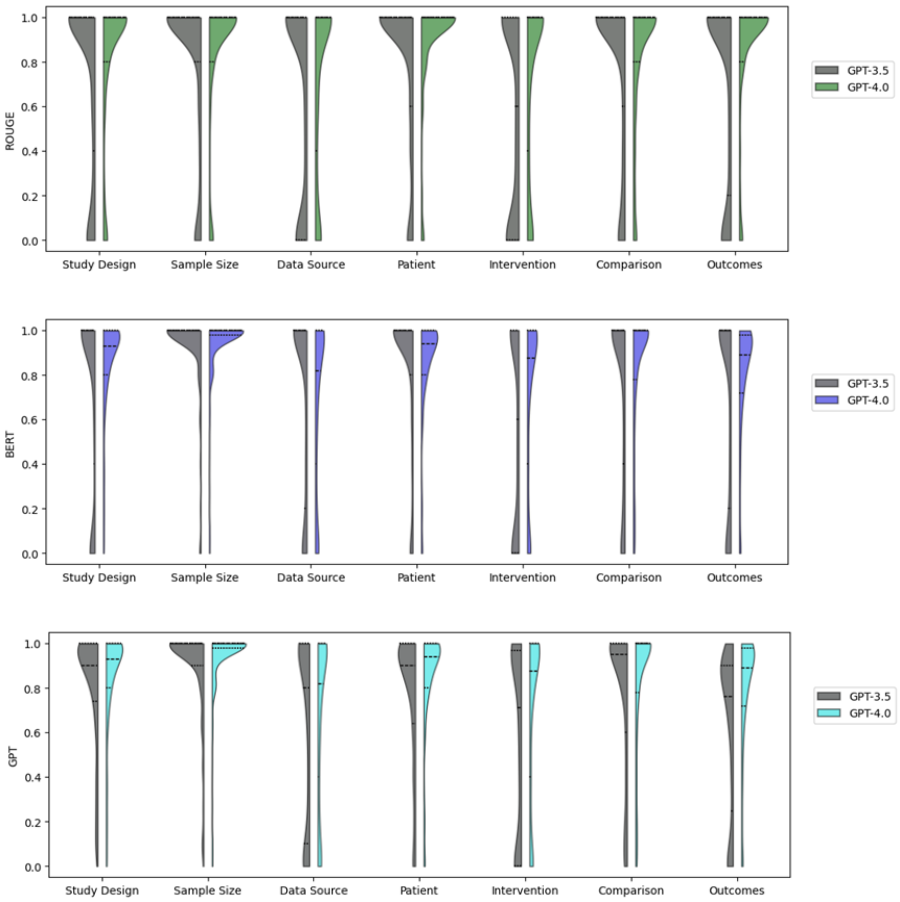
Items	GPT model	$\overline{p}_{correct}$ (BERT)	$\overline{p}_{correct}$ (GPT)	$\overline{p}_{correct}$ (ROUGE)	Optimal Prompt Engineering Strategies
Study Design	GPT-3.5	0.798 (0.4)	0.862(0.86)	0.798(0.4)	Alpha(BERT/ROUGE/ GPT)
	GPT-4.0	0.796(0.34)	0.86(0.86)	0.796(0.34)	Beta+Gamma(BERT/ROUGE), Alpha(GPT)
Sample Size	GPT-3.5	0.992(0.08)	0.955(0.95)	0.992(0.08)	Alpha+Gamma(BERT/ ROUGE), Alpha(GPT)
	GPT-4.0	0.964(0.16)	0.97(0.97)	0.964(0.16)	Gamma(BERT/ROUGE), Alpha(GPT)
Data Source	GPT-3.5	0.804(0.38)	0.716(0.72)	0.804(0.38)	Alpha(BERT/ROUGE/ GPT)
	GPT-4.0	0.852(0.32)	0.773(0.77)	0.852(0.32)	Gamma(BERT/ROUGE /GPT)
Patient	GPT-3.5	0.844(0.33)	0.8(0.8)	0.844(0.33)	Alpha(BERT/ROUGE/ GPT)
	GPT-4.0	0.924(0.2	0.864(0.86)	0.924(0.24)	Control(BERT/ROUGE

		4)			/GPT)
Intervention	GPT-3.5	0.568(0.41)	0.595(0.6)	0.568(0.41)	Alpha+Beta+Gamma(BERT/ROUGE), Control (GPT)
	GPT-4.0	0.688(0.41)	0.722(0.72)	0.688(0.41)	Gamma(BERT/ROUGE/GPT)
Comparison	GPT-3.5	0.764(0.39)	0.809(0.81)	0.67(0.46)	Alpha+Gamma(BERT/ROUGE), Alpha(GPT)
	GPT-4.0	0.804(0.36)	0.845(0.85)	0.804(0.36)	Gamma(BERT/ROUGE), Alpha(GPT)
Outcomes	GPT-3.5	0.706(0.37)	0.638(0.64)	0.706(0.37)	Alpha+Beta(BERT/ROUGE), Alpha(GPT)
	GPT-4.0	0.864(0.31)	0.825(0.83)	0.864(0.31)	Alpha(BERT/ROUGE/GPT)

The performance of GPT in extracting information across seven items can be categorized into three distinct levels of complexity. The first level encompasses questions where a direct answer can typically be found in the raw text. The sample size is an example of this level, and both GPT-3.5 and GPT-4.0 achieve accuracy levels exceeding 0.95 in extracting sample size. The second level pertains to questions requiring understanding and summarization skills to extract answers. Most extracted items, including study design, data source, patient, comparison, and outcomes, belong to this category. Table 3 shows that GPT-3.5 achieves optimal performance from 0.7 to 0.8 for these items and GPT-4.0 from 0.8 to 0.9. Finally, intervention represents the third level, which demands a high level of understanding and domain expertise to discern the correct answer accurately from potentially misleading information. In this regard, GPT-3.5 performed under 0.6 while GPT-4.0 demonstrated accuracy around 0.7.

Figure 3 employs a violin plot to illustrate the distribution of model performance. Noticeably, the plots reveal bimodal distributions, with performance clustering at high and low accuracy extremes, representing that the GPT models are either all correct or all incorrect in their extractions. This observation demonstrates the stability of GPT's performance in such information extraction tasks. Despite the common issue of the randomness of generative language models, these results demonstrated that GPTs are controllable and consistent in evidence summarization.

Figure 3. Violin plots of the performance distribution of GPT-3.5 and GPT-4.0 on each item to extracted. Y label represents the metrics and the dashed lines inside violins represent the quartiles.



Performance of GPT-3.5 and GPT-4.0

The GPT version is a statistically significant factor influencing model performance. As presented in Table 5, the ANOVA analysis reveals that 20 out of the 21 p-values assessing the impact of GPT are significantly lower than 0.05. The only exception of p-value is associated with the item Sample Size measured by BERTScore.

Table 5. Summary of p-values in ANOVA analysis. For example, the first cell represents the p-value corresponding to the factor "GPT versions", utilizing study design data evaluated by the ROUGE metric as input data for the ANOVA analysis.

Evaluator	Factor	Study Design	Sample Size	Data Source	Patient	Inter-vention	Comparison	Outcomes
ROUGE	GPT	**<0.0001**	**0.0069**	**0.0107**	**<0.0001**	**<0.0001**	**0.0086**	**<0.0001**
	Prompt	0.0721	**0.0005**	**0.0000**	0.4689	0.9667	**0.0022**	0.9982
	Interaction	**0.0053**	**0.0134**	0.9352	0.7269	0.9860	0.6483	0.9986
BERTScore	GPT	**0.0210**	0.8896	**<0.0001**	**<0.0001**	**<0.0001**	**0.0001**	**<0.0001**
	Prompt	0.3537	**<0.0001**	**0.0000**	0.9949	0.9470	**0.0007**	0.5658
	Interaction	0.0554	**<0.0001**	0.9218	0.9934	0.9647	0.7438	0.2389

	GPT	**0.0012**	**<0.0001**	**0.0011**	**<0.0001**	**<0.0001**	**0.0002**	**<0.0001**
GPT	Prompt	**0.0026**	**<0.0001**	**<0.0001**	0.1556	0.8368	**0.0434**	0.1503
	Interaction	**0.0012**	**<0.0001**	0.9431	0.8998	0.9806	0.6406	0.3165

In pair-wise comparison, GPT-4.0 tends to surpass GPT-3.5, particularly in complex tasks where direct answers are in the raw text. Table 4 demonstrated the significant edges of GPT-4.0 over GPT-3.5, characterized by both enhanced proportion of correctness and a reduced standard deviation, on Data Source (GPT-4.0: $\overline{p}_{correct}$ = 0.852; GPT-3.5, SD=0.32: $\overline{p}_{correct}$=0.852, SD=0.38), Patient (GPT-4.0: $\overline{p}_{correct}$ = 0.924, SD=0.24; GPT-3.5: $\overline{p}_{correct}$=0.844, SD=0.33), Intervention(GPT-4.0: $\overline{p}_{correct}$ = 0.688, SD=0.41; GPT-3.5: $\overline{p}_{correct}$ =0.568, SD=0.41), Comparison (GPT-4.0: $\overline{p}_{correct}$= 0.804, SD=0.36; GPT-3.5: $\overline{p}_{correct}$=0.764, SD=0.39), and Outcomes (GPT-4.0: $\overline{p}_{correct}$= 0.864, SD=0.31; GPT-3.5: $\overline{p}_{correct}$=0.706, SD=0.37). However, it is worth noting that GPT-4.0 may not necessarily outperform its predecessors. For example, GPT-4.0 does not reach the same level as GPT-3.5 on Sample Size (GPT-4.0: $\overline{p}_{correct}$= 0.964, SD=0.16; GPT-3.5: $\overline{p}_{correct}$=0.992, SD=0.08), even though the extraction is relatively simple and there tends to be a finite numeric answer in the abstract.

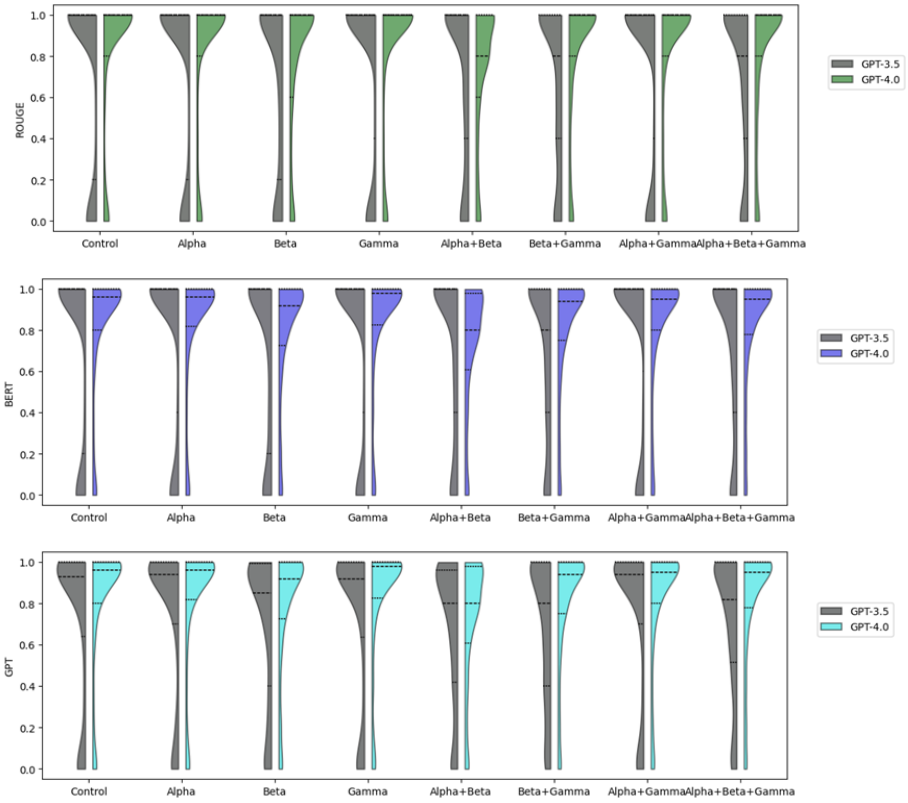
Figure 3 visually compares the performance distribution of GPT-3.5 and GPT-4.0 on the seven items. Noticeably, it can be found that the performance improvement of GPT-4.0 is brought by shrinking tails near 0, which means the occurrence of all false values. This can verify the fact that the enhanced ability of GPT-4.0 in extracting the expected information.

Effects of Prompt Engineering Strategies

Prompt engineering strategies are likely to influence model performance positively, but their effectiveness is not guaranteed. As presented in Table 5, the ANOVA analysis reveals that the impact of the GPT prompt is statistically significant for three extracted items, Sample Size, Data Source, and Comparison, measured by all three evaluators. There needs to be more evidence for other items to prove the impact of prompt engineering strategies. The discordance of impact can also be verified in Table 4, where different extraction tasks exhibit a different preference for specific prompt engineering strategies to achieve optimal performance.

It is also noticeable that prompt engineering strategies may not have additive effects with each other. For example, in Figure 4, the combination Alpha+Beta does not perform as well as either Alpha or Beta. Combined strategies, such as Alpha+Beta+Gamma, may lead to inferior results compared to the simple one. The control set produces a considerably good performance, and it is the optimal strategy for GPT-4.0 to extract the patient information, as shown in Table 4.

Figure 4. Violin plots of the performance distribution of GPT-3.5 and GPT-4.0 using different prompt engineering strategies. Y label represents the metrics and the dashed lines inside violins represent the quartiles.



The effects of GPT versions and prompt engineering strategies will likely interact. In ANOVA analysis, the interaction between the GPT version and prompt engineering strategies is statistically significant based on the Sample Size extraction, as assessed by all three evaluators (ROUGE, $p < .001$; BERTScore, $p<.001$; GPT, $p<.001$). However, for other items, interaction may exist but needs more statistical strength. Table 4 demonstrates that the optimal prompt engineering strategy differs between GPT-3.5 and GPT-4.0 for each item. In general, GPT-3.5 tends to favour the Alpha strategy, persona (optimal strategy for Study Design, Data Source, Patient). In contrast, GPT-4.0 tends to prefer the Gamma strategy, with few-shot prompting (optimal strategy for sample size, data source, intervention, comparison).

Discission

Our study presents a rigorous evaluation of GPT-3.5 and GPT-4.0 in extracting medical information from titles and abstracts. To ensure comprehensive coverage across various medical domains, a stratified sampling method was adopted for paper selection from almost all affiliated medical schools and departments of a university. We employed multiple evaluators, repetitive trials, and experiments on prompt engineering strategies to enhance the integrity of results. Our findings demonstrated that GPTs can effectively extract or summarize information described in the abstracts. Notably, GPT-4.0 exhibits robust performance in providing thorough

answers and understanding and summarizing abstracts. However, there is still room for improvement in accurately discerning information that requires sophisticated understanding and domain expertise. When combined with appropriate prompt engineering strategies, the accuracy level achieves over 0.8 in extracting information related to study design, sample size, data source, patient, comparison and outcomes. Moreover, the improvement in efficiency is remarkable by reducing 8 to 10 hours of human labor to under 5 minutes (GPT-3.5) or 40 minutes (GPT-4.0). Our research pioneers the exploration of a new generation of Large Language Models in medical evidence summarization and offers potential applications in various scenarios. It provides empirical evidence to support the development of credible automatic tools for medical literature screening and review. With critical information extracted, automatic tools can strike a balance between efficiency and transparency.

The field of large language models is rapidly advancing. Our investigations reveal that the effect of the GPT version on the accuracy of information extraction is significant when comparing GPT-3.5 and GPT-4.0 (Table 5). In particular, GPT-4.0 presents a more robust performance in summarizing complex information that may not be readily apparent in the raw text, such as the PICOs. On the other hand, the drawback of GPT-4.0 compared to its predecessor is associated with the time and cost. According to the OpenAI website, by March 2024, the price of GPT-3.5 Turbo is one-twentieth of that of GPT-4.0 Turbo [31]. In our experiment, we found that the time required for GPT-3.5 to label 100 papers is approximately one-tenth of the time taken by GPT-4.0. This significant difference may be attributed to the rate limits imposed by the API, as noted on OpenAI's website. Specifically, the rate limit for GPT-4-turbo is 500 RPM (Requests Per Minute) for Tier 1 users, while GPT-3.5-turbo offers a higher rate limit of 3500 RPM [32].

Prompt engineering strategies play an essential role in enhancing LLMs' performance. This study find that the optimal prompt engineering strategies vary depending on the extraction tasks and GPT versions employed. Overall, two useful strategies are recommended to attempt: persona and few-shot prompting. Although, the chain of thought strategy might help guide multi-step tasks, it might not be effective in straightforward tasks like the information extraction in this study (Table 4). It is worth noting that the combination of prompt engineering strategies may not yield additive effects on the final results (Figure 4). Considering the cost associated with input tokens, it is recommended to use a conservative approach to employ prompt engineering strategy in prompt development.

Moreover, this study extensively examines and compares the performance of evaluators utilized in the experiment, including two well-established NLP metrics, ROUGE-1 and BERTScore, and one newly developed GPT evaluator. Overall, the three evaluators provide consistent performance evaluation across various extraction items and prompt engineering strategies (Figure 3 and Figure 4). However, each metric has its limitations. ROUGE-1, as a basic metric relying on common words, is susceptible to issues like spelling variations and abbreviations. For example, when the output is "cluster randomised controlled trial" while the ground truth is "cluster randomized controlled trial", ROUGE-1 assigns a score of 0.5, which tends to be labeled as a mismatch with the ground truth (record 17, Supplementary Material 5).

To overcome these limitations, BERTScore incorporates synonyms and offers a continuous scoring system. In this study, we employed a threshold to transfer from a continuous value to a binary label of a correct answer. While the threshold is optimized based on available data, it may not effectively decide on the alignment between the generated answer and ground truth. Take record 89 as an illustrative example (Supplementary Material 5). The labeled ground truth of record 89 is "survey," and the generated answer by GPT-4.0 is "school-based survey." In this case, the BERTScore output is 0.834 under the optimal threshold, resulting in a labeled mismatch with the ground truth.

Our study revealed an interesting observation regarding the potential of GPT as a promising and unique tool to assess the accuracy of generated text compared to the ground truth. Notably, GPT evaluators can leverage their pre-trained knowledge base to evaluate text based not only on lexical similarity but also on semantic similarity. This ability effectively addresses some significant limitations of existing NLP metrics. To illustrate, consider the extracted intervention for record 26 as an example (Supplementary Material 6). The labeled ground truth is "Pre-emptive use of proton pump inhibitors (PPI)," while the generated output of GPT-4.0 is "Pre-emptive PPI (intravenous esomeprazole followed by high-dose oral esomeprazole)." In this instance, the generated answer is not only correct but also superior to the ground truth, as it integrates the information from another helpful sentence: "The PPI group received intravenous esomeprazole 4 h before the EST and then every 12 h for 1 day, followed by high-dose oral esomeprazole for 10 days."[33]. However, both ROUGE-1 and BERTScore fail to label this correct answer due to a lack of overlapping words. In contrast, the GPT evaluator assigns a score of 0.85, acknowledging the reasonableness of the generated output. This outcome highlights the potential of the GPT evaluator that evaluating not merely the words but also meanings can offer a more authentic and accurate evaluation. We also found that GPT exhibits logical thought when evaluating the answer during development. For instance, it can distinguish the difference between the outcome of HbA1c and the drop of HbA1c. These properties allow GPT to be further developed into a powerful and systematic tool to evaluate the performance of complex information extraction.

However, it is important to acknowledge several limitations in this study. First, while covering a wide range of medical domains, the labeled ground truth represents the assessment level of human evaluators. It may not necessarily serve as the golden standard due to a lack of domain knowledge. As a result, the performance of GPTs could be underestimated. To address this, future research is encouraged to validate GPT's performance in one specific medical domain. When the targeted literature focuses on one area, domain knowledge can be provided as contextual information to enhance performance. Another limitation of this study is that we solely tested GPT from the abstracts. Given the exploding ability of LLMs in handling long text, figures, and tables, it is recommended that future researchers extend the GPT tools to operate on full text or PDF level. This expansion would extract more valuable information sources and open up broader possibilities for GPT to facilitate medical research.

Conclusion

In this study, we have developed a robust method based on GPT for extracting or summarizing information from the abstract of medical research papers. We conducted thorough experiments to systematically evaluate the effects of GPT versions and prompt engineering strategies on the performance of models. The evaluation was carried out utilizing both well-established NLP metrics and a newly developed GPT evaluator. Notably, the GPT evaluator demonstrated its effectiveness and advantage by leveraging its capability of semantic understanding. Our result validates the potential of GPT as a reliable, stable, and accurate tool for summarizing medical evidence, particularly when appropriate prompt settings are employed. We encourage further research and studies to continue refining and advancing this tool, unlocking the potential of the new generation of technology in medical research.

Supplementary Material

Supplementary material 1 Summary table of selected papers
Supplementary material 2 Optimal threshold and the process of grid search
Supplementary material 3 Summary of model performance
Supplementary material 4 ANOVA tables
Supplementary material 5 Illustrative example of study design.
Supplementary material 6 Illustrative example of intervention

Reference

1. Li, J., Dada, A., Puladi, B., Kleesiek, J., & Egger, J. (2024). ChatGPT in healthcare: a taxonomy and systematic review. Computer Methods and Programs in Biomedicine, 108013.

2. Lim, Z. W., Pushpanathan, K., Yew, S. M. E., Lai, Y., Sun, C. H., Lam, J. S. H., ... & Tham, Y. C. (2023). Benchmarking large language models' performances for myopia care: a comparative analysis of ChatGPT-3.5, ChatGPT-4.0, and Google Bard. EBioMedicine, 95.

3. Shah NH, Entwistle D, Pfeffer MA. Creation and Adoption of Large Language Models in Medicine. JAMA. 2023;330(9):866-9.

4. Nori H, King N, McKinney SM, Carignan D, Horvitz E. Capabilities of gpt-4 on medical challenge problems. arXiv preprint arXiv:230313375. 2023.

5. Tang, L., Sun, Z., Idnay, B., Nestor, J. G., Soroush, A., Elias, P. A., ... & Peng, Y. (2023). Evaluating large language models on medical evidence summarization. npj Digital Medicine, 6(1), 158.

6. Shaib, C., Li, M. L., Joseph, S., Marshall, I. J., Li, J. J., & Wallace, B. C. (2023). Summarizing, simplifying, and synthesizing medical evidence using gpt-3

(with varying success). arXiv preprint arXiv:2305.06299.

7. Tian, S., Jin, Q., Yeganova, L., Lai, P. T., Zhu, Q., Chen, X., ... & Lu, Z. (2024). Opportunities and challenges for ChatGPT and large language models in biomedicine and health. Briefings in Bioinformatics, 25(1), bbad493.

8. O'Mara-Eves A, Thomas J, McNaught J, Miwa M, Ananiadou S. Using text mining for study identification in systematic reviews: a systematic review of current approaches. Systematic Reviews. 2015;4(1):5.

9. Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. Systematic reviews, 5, 1-10.

10. Marshall, I. J., Kuiper, J., & Wallace, B. C. (2016). RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. Journal of the American Medical Informatics Association, 23(1), 193-201.

11. Howard, B. E., Phillips, J., Miller, K., Tandon, A., Mav, D., Shah, M. R., ... & Thayer, K. (2016). SWIFT-Review: a text-mining workbench for systematic review. Systematic reviews, 5, 1-16.

12. Blaizot AA-O, Veettil SK, Saidoung P, Moreno-Garcia CF, Wiratunga N, Aceves-Martins MA-OX, et al. Using artificial intelligence methods for systematic review in health sciences: A systematic review. (1759-2887 (Electronic))

13. Feng Y, Liang S, Zhang Y, Chen S, Wang Q, Huang T, et al. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. (1527-974X (Electronic)).

14. Shemilt I, Khan N, Park S, Thomas J. Use of cost-effectiveness analysis to compare the efficiency of study identification methods in systematic reviews. Systematic Reviews. 2016;5(1):140.

15. Matthew JP, David M, Patrick MB, Isabelle B, Tammy CH, Cynthia DM, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.

16. Matsui KaU, Tomohiro and Aoki, Yumi and Maruki, Taku and Takeshima, Masahiro and Yoshikazu, Takaesu. Large Language Model Demonstrates Human-Comparable Sensitivity in Initial Screening of Systematic Reviews: A Semi-Automated Strategy Using GPT-3.5. SSRN:4520426. 2023.

17. Mahuli SA, Rai A, Mahuli AV, Kumar A. Application ChatGPT in conducting systematic reviews and meta-analyses. British Dental Journal. 2023;235(2):90-2.

18. Hill, J. E., Harris, C., & Clegg, A. (2023). Methods for using Bing's AI□powered search engine for data extraction for a systematic review. Research Synthesis Methods.

19. Gilbert S, Harvey H, Melvin T, Vollebregt E, Wicks P. Large language model AI chatbots require approval as medical devices. Nature Medicine. 2023;29(10):2396-8.

20. Chen, P., Huang, Z., Deng, Z., Li, T., Su, Y., Wang, H., ... & He, J. (2023). Enhancing Medical Task Performance in GPT-4V: A Comprehensive Study on Prompt Engineering Strategies. arXiv preprint arXiv:2312.04344.

21. Duke University. (2019). LibGuides: Evidence-Based Practice: PICO.

Duke.edu. https://guides.mclibrary.duke.edu/ebm/pico

22. Grabb, D. (2023). The impact of prompt engineering in large language model performance: a psychiatric example. Journal of Medical Artificial Intelligence, 6.

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

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

25. Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

26. Zhao, Z., Wallace, E., Feng, S., Klein, D., & Singh, S. (2021, July). Calibrate before use: Improving few-shot performance of language models. In International Conference on Machine Learning (pp. 12697-12706). PMLR.

27. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901.

28. Lin, C. Y. (2004, July). Rouge: A package for automatic evaluation of summaries. In Text summarization branches out (pp. 74-81).

29. Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2019). Bertscore: Evaluating text generation with bert. arXiv preprint arXiv:1904.09675.

30. Seabold, Skipper, and Josef Perktold. “statsmodels: Econometric and statistical modeling with python.” Proceedings of the 9th Python in Science Conference. 2010.

31. OpenAI. (2023). Pricing. Openai.com. https://openai.com/pricing

32. OpenAI. (2023). Rate Limits. Openai.com. https://platform.openai.com/docs/guides/rate-limits/usage-tiers?context=tier-one

33. Leung, W. K., But, D. Y., Wong, S. Y., Tong, T. S., Liu, K. S., Cheung, K. S., ... & Hung, I. F. (2018). Prevention of post□sphincterotomy bleeding by proton pump inhibitor: A randomized controlled trial. Journal of digestive diseases, 19(6), 369-376.

Figure 1. Flowchart of overall study design;

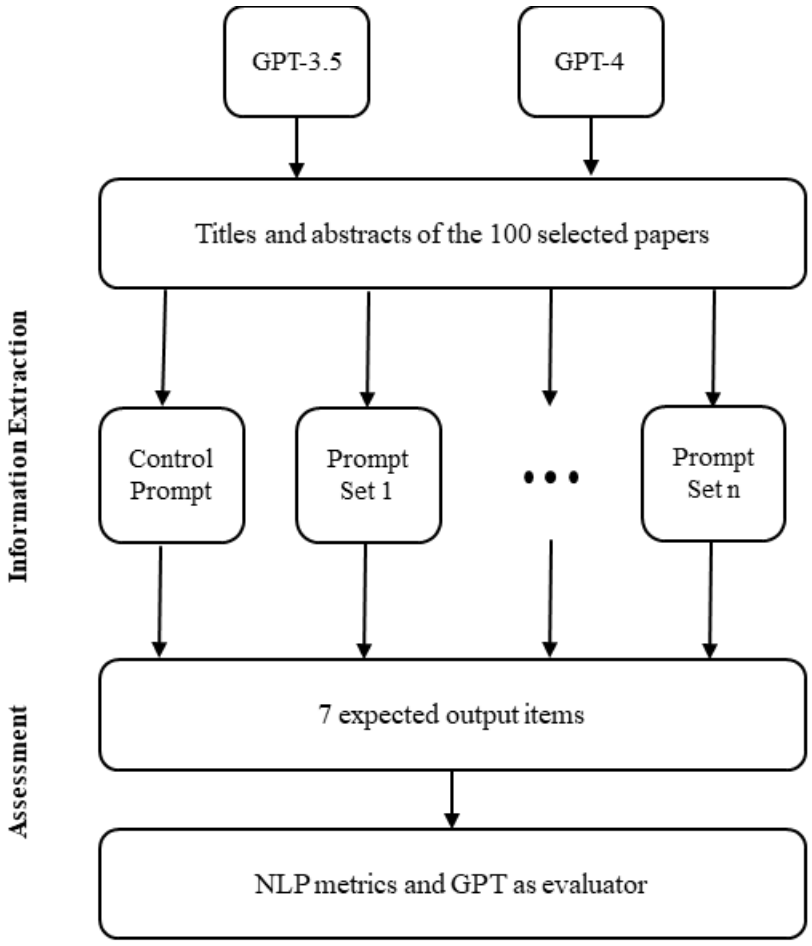


Figure 2. Paper affiliation distribution

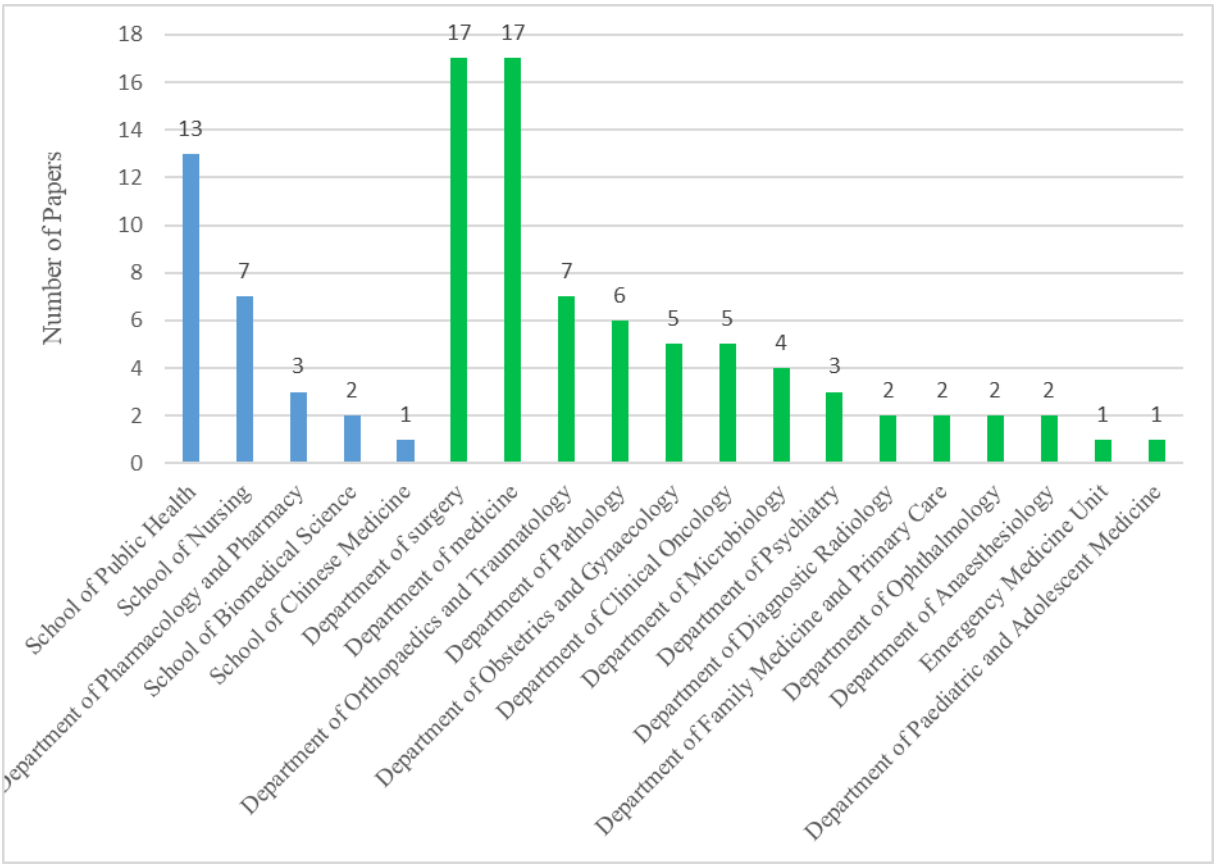


Figure 3. Violin plots of the performance distribution of GPT-3.5 and GPT-4.0 on each item to extracted. Y label represents the metrics and the dashed lines inside violins represent the quartiles.

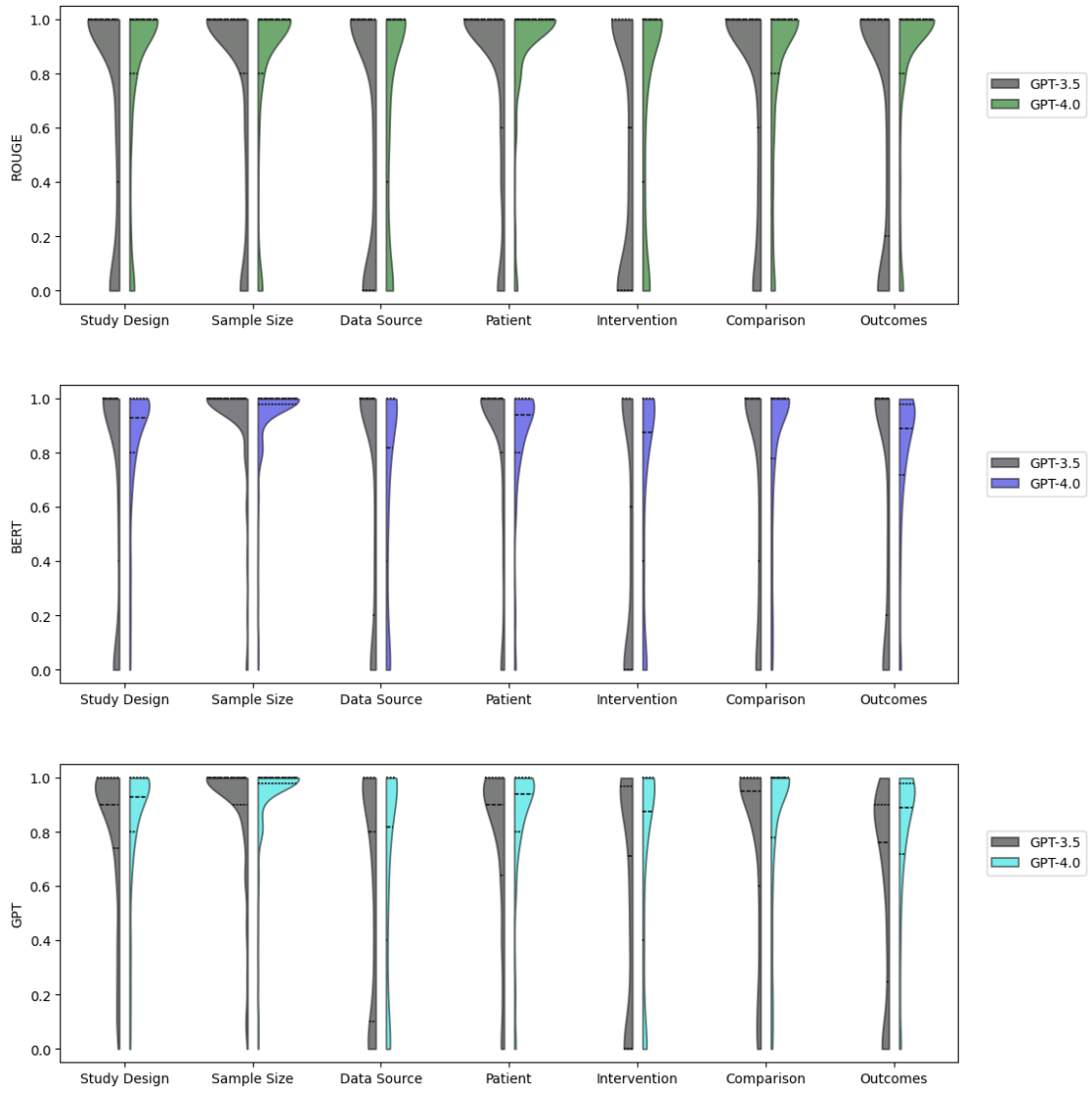


Figure 4. Violin plots of the performance distribution of GPT-3.5 and GPT-4.0 using different prompt engineering strategies. Y label represents the metrics and the dashed lines inside violins represent the quartiles.