Utilizing ChatGPT to select literature for meta-analysis shows workload reduction while maintaining a similar recall level as manual curation

Xiangming Cai, PhD[1]*#, Yuanming Geng, MS[2,3]*, Yiming Du, PhD[4]*, Bart Westerman, PhD[5], Duolao Wang, PhD[6]#, Chiyuan Ma, MD[2,3,7,8,9]#, Juan J. Garcia Vallejo, PhD[1]

1 Department of Molecular Cell Biology & Immunology, Amsterdam Infection & Immunity Institute and Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

2 Department of Neurosurgery, Jinling Hospital, Nanjing, China.

3 Department of Neurosurgery, Affiliated Jingling Hospital, Nanjing Medical University, Nanjing, China.

4 Department of System Engineering and Engineering Management, The Chinese University of Hong Kong, Hong Kong, China

5 Department of Neurosurgery, Cancer Center Amsterdam, Brain tumor center Amsterdam, Amsterdam UMC location Vrije Universiteit Amsterdam, Amsterdam, The Netherlands.

6 Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom.

7 School of Medicine, Southeast University, Nanjing, China.

8 Department of Neurosurgery, Affiliated Jinling Hospital, Medical School of Nanjing University, Nanjing, China.

9 Department of Neurosurgery, Jinling Hospital, the First School of Clinical Medicine, Southern Medical University, Nanjing, China.

*XC, YG, and YD contributed equally and share the co-first authorship

#Co-corresponding authors

Correspondence to:

Prof. Chiyuan Ma: Department of Neurosurgery, Affiliated Jinling Hospital, Medical School of Nanjing University, Nanjing, China. Email: machiyuan_nju@126.com

Prof. Duolao Wang: Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom. Email: duolao.wang@lstmed.ac.uk

Xiangming Cai: Department of Molecular Cell Biology & Immunology, Amsterdam Infection & Immunity Institute and Cancer Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, Amsterdam, The Netherlands. Email: x.cai@amsterdamumc.nl

Author Contributions

XC and CM conceived the idea. XC designed the study. XC, YG, and YD developed the methodology, acquired, and analyzed the data. XC, BW, DW, and JJGV were involved in

interpretation of data. XC drafted the manuscript and all authors edited the manuscript. All authors had full access to the raw data in the study. XC and YG accessed and verified the data. XC and JJGV oversaw conduction of the study. All authors contributed to the article and approved the submitted version. All authors had final responsibility for the decision to submit for publication.

Abstract

Background: Large language models (LLMs) like ChatGPT showed great potential in aiding medical research. A heavy workload in filtering records is needed during the research process of evidence-based medicine, especially meta-analysis. However, no study tried to use LLMs to help screen records in meta-analysis. In this research, we aimed to explore the possibility of incorporating ChatGPT to facilitate the screening step based on the title and abstract of records during meta-analysis.

Methods: To assess our strategy, we selected three meta-analyses from the literature, together with a glioma meta-analysis embedded in the study, as additional validation. For the automatic selection of records from curated meta-analyses, a four-step strategy called LARS was developed, consisting of (1) criteria selection and single-prompt (prompt with one criterion) creation, (2) best combination identification, (3) combined-prompt (prompt with one or more criteria) creation, and (4) request sending and answer summary. We evaluated the robustness of the response from ChatGPT with repeated requests. Recall, workload reduction, precision, and F1 score were calculated to assess the performance of LARS.

Findings: ChatGPT showed a stable response for repeated requests (robustness score: 0·747 – 0·996). A variable performance was found between different single-prompts with a mean recall of 0·841. Based on these single-prompts, we were able to find combinations with performance better than the pre-set threshold. Finally, with a best combination of criteria identified, LARS showed a 39·5% workload reduction on average with a recall greater than 0·9. In the glioma meta-analysis, we found no prognostic effect of CD8+ TIL on overall survival, progress-free survival, and survival time after immunotherapy.

Interpretation: We show here the groundbreaking finding that automatic selection of literature for meta-analysis is possible with ChatGPT. We provide it here as a pipeline, LARS, which showed a great workload reduction while maintaining a pre-set recall.

Funding: China Scholarship Council.

Keywords: ChatGPT; Meta-analysis; Glioma; Immunotherapy; T cells

Introduction

Our understanding of diseases advanced rapidly during the last decades. However, the translation from bench to bedside is lagging.[1] Evidence-based medicine (EBM), especially meta-analysis, facilitated the application of novel therapies into clinics. However, the processes of conducting meta-analysis are time-consuming and workload intensive.[2]

Artificial intelligence (AI) is becoming ubiquitous in medicine.[1] AI-based solutions are developed to reduce human efforts spent on EBM with promising performance.[3] Human annotators are needed to train the AI model on dataset.[4,5] AI models could provide predicted probability for all records based on “similarity” between them. However, researchers still need to screen all of them.

Recent releases of large language models (LLMs) like ChatGPT have huge implications on medical research.[6-8] LLMs showed impressive potential in generating answers following user's instruction. However, few studies have evaluated its application in aiding EBM and review writing. Shaib *et al*. utilized ChatGPT (text-davinci-003) to synthesize medical evidence.[9] Shuai *et al.* explored its effectiveness in generating Boolean queries for literature search.[10] However, no study has investigated its application in compensating or substituting human effort spent on filtering records during meta-analysis, which is a severe issue because of the exponentially increased number of primary literature and systemic reviews.[11]

Glioma is a common tumor of the central nervous system. Glioblastoma (GBM) is the most aggressive subtype of it with a five-year survival of 7·2 %.[12] Immunotherapy has become a promising approach in the treatment of it. However, no improved prognosis was reached yet in phase III immunotherapy trials of glioma.[13] Highly intertumoral heterogeneity was one of the reasons for this outcome. Hence, identifying patients who will respond to immunotherapy has become a core task in the research of GBM. In non-CNS malignancy, CD8+ tumor infiltrating lymphocytes (TILs) within tumor has been associated with response to immunotherapy and prognosis.[13] However, no consensus was reached regarding the value of CD8+ TILs in predicting glioma patients' response to immunotherapy.[14]

In the current study, we explore the possibility of using ChatGPT to aid automatic selection of literature records (based on their title and abstract) for meta-analysis by developing a pipeline named "LARS" (Literature Records Screener). Additionally, we investigated the prognostic predictive role of CD8+ TILs in immunotherapy treated glioma patients in the glioma meta-analysis, which was embedded in the study as a validation dataset.

Methods

Screen pipeline incorporating ChatGPT: LARS

In general, the workflow of meta-analysis has the following steps: (1) define research question; (2) select literature databases and design search strategy; (3) screen records based on their titles and abstracts; (4) screen records based on full text of records; (5) extract and synthesis data. In the present study, we focused on incorporating ChatGPT into the third step of this workflow.

To do that, we designed the four-step pipeline, LARS (**Fig 4**). First, users need to select criteria (some suitable criteria from filtering criteria of meta-analysis) and create a prompt for each criterion (single-prompt; **Table 1**). Second, users need to evaluate these single-prompts using a few records and then select the best combination of single-prompts. Third, users need to choose a prompt strategy and merge single-prompts in the best combination to make a combined prompt (combined-prompt; **Supplementary File 1**) in accordance with the selected prompt strategy. Finally, the combined-prompt, together with the title and abstract of each record, will be submitted to ChatGPT as chat completion. The decisions about whether a record meets the user's criteria will then be extracted from returned answers. In this study, we evaluated both GPT-3.5 (gpt-3.5-turbo-0301) and GPT-4 (gpt-4-0314) using the API (Application Programming Interface) provided by OpenAI. In practice, LARS could be performed in batches using Python.

Selection of validation meta-analyses

To cover broad medical fields, we selected three high-quality meta-analyses as validation datasets, which focused on inflammatory bowel diseases (IBD),[15] diabetes mellitus (DM),[16] and sarcopenia,[17] respectively (**Table 2**). These meta-analyses provided clear search strategies for Medline/PubMed database and complete list of records that remained after screening based on their titles and abstracts. Thanks to this, we were able to repeat their literature search in Medline/PubMed and match record list to obtain the correct answer that whether these identified records could pass the screening step in a real-world practice (**Table 2; Supplementary File 2**). On top of these published meta-analyses, we conducted a new meta-analysis about glioma in this research (see **Supplementary Methods**). In doing so, we can evaluate the performance of ChatGPT in a first-hand practice.

Prompt strategy design

We designed prompts (**Table 1**; **Supplementary File 1**) with the guidance from OpenAI (https://platform.openai.com/docs/guides/gpt-best-practices). However, the high flexibility of prompt and the "black box" nature of ChatGPT made it impossible to design a "best" prompt. In this study, we designed three distinct types of prompt strategies to help create better combined-prompt (**Fig 1 and 4; Supplementary File 1**). For the "single criterion" prompt (prompt strategy 1), we simply maintain these single-prompts in the best combination. ChatGPT will respond to each single-prompt and determine whether a record meets each criterion or not. After receiving answers from ChatGPT, users need to summarize answers for each single-prompt and make a final decision for each record. In this study, as long as there is one answer that is "No", the final decision for a record is "No". Otherwise, the final decision will be "Yes". For the "instruction prompt" (prompt strategy 3) and "chain of thought prompt" (prompt strategy 2), the best combination of single-prompts was merged into one combined-prompt (**Fig 1 and 4; Supplementary File 1**). Users expect a final judgement from ChatGPT directly.

Evaluation of the classification performance of single-prompt

We (XC and YG) manually labeled correct answers of each single-prompt within 100 randomly selected records (about 10 positive records and 90 negative records) for each validation meta-analysis. Here, records were called "positive" records if they were remained after the screening step based on their titles and abstracts. Otherwise, they were called "negative" records. With these 100-reords datasets, we evaluated the performance of ChatGPT and a random classifier regarding single-prompts.

Evaluation of single-prompt combination and identification of best combination

Before conducting any evaluation, the "best" combination of single-prompts was unknown. In other words, how many single-prompts and which single-prompts should be selected for combined-prompt creation? To address this question, we evaluated all possible combinations of designed single-prompts. Among these combinations, we selected the best combination, which has a recall ≥ 0.9 and the best workload reduction.

Glioma meta-analysis

This meta-analysis was conducted and reported following the Preferred Reporting Items for Systematic Review and Meta-analysis (**Supplementary File 3**).[18] The protocol of the current meta-analysis was registered on PROSPERO (CRD42023425790). Detailed methods of this glioma meta-analysis could be found in **Supplementary File 4**.

Statistical analysis

Because of the nature of LLMs, the generated answer from LLMs varies each time, even with exactly identical input. So, we assessed the robustness score of each single-prompt with repeated requests (see **Supplementary Methods**).

The performance of ChatGPT was assessed with precision, recall, F1 score, and workload reduction metrics. The workload reduction indicator was defined as:

$$workload\ reduction = n_{records\ excluded\ by\ model} / n_{all\ records}$$

Where n is the number of records. The workload reduction indicator varies between 0 and 1, where 0 indicates none work was reduced and 1 signifies that all work was reduced. For meta-analysis, recall is the most significant indicator, followed with workload reduction, F1, and precision. Throughout the study, we placed greater emphasis on recall and workload reduction as the primary performance metrics. Also, a random classifier was used as baseline reference (see **Supplementary Methods**).

Results

A case showing the request of a single-prompt and its response from ChatGPT is presented in **Figure 1A**. The work flow of this study is showed in **Figure 1B**. And the schematic illustration of LARS is presented in **Figure 4**. In this research, we selected 4-5 criteria from each meta-analysis (**Table 1**; **Supplementary File 1**).

ChatGPT returns show high robustness

The robustness score for each single-prompt was computed using GPT-3.5 (**Fig 2A; Supplementary Fig 1**). In general, the returns were stable, with a robustness score ranging from 0·747 to 0·996. The "Species", "Disease", and "Research type" single-prompts were evaluated in all four meta-analyses. Despite the robustness scores of "Species" single-prompts were slightly lower (from 0·747 to 0·837), they were still good.

Single-prompts exhibit distinct performance

The performance of each single-prompt based on GPT-3.5 or GPT-4 was assessed (**Table 3; Supplementary Table 1**). Overall, the majority of prompts had better performance with GPT than a random classifier. The mean recall for GPT was 0·841, with 69·4% single-prompts having a recall higher than 0·8. The GPT-3.5 (mean recall: 0·867) and GPT-4 (mean recall: 0·815) had similar and good recalls. Surprisingly, the recalls could be quite different between these two versions of GPT, even for the same single-prompt, *e.g.*, the "Control" single-prompt from sarcopenia meta-analysis (GPT-3.5: 0·838; GPT-4: 0·235; **Supplementary Table 1**) and the "Protein related" single-prompt from IBD meta-analysis (GPT-3.5: 0·897; GPT-4: 0·483;

Supplementary Table 1).

Different single-prompts also exhibited distinct recalls. Most single-prompts performed well like the "Research type" prompt from glioma meta-analysis (GPT-3.5: 0·966; GPT-4: 0·989; **Table 3**). However, few single-prompts demonstrated low recalls, *e.g.*, the "Disease_dm" prompt from IBD meta-analysis (GPT-3.5: 0·554; GPT-4: 0·770; **Supplementary Table 1**).

The best combination of single-prompts is identified by evaluating the performance of all possible combinations

All combinations of single-prompts were shown in the form of UpSet plots (**Fig 3 and Supplementary Fig 2-4**). As expected, when the number of single-prompts increases, the recall tends to decrease, while workload reduction and precision increase. In general, most combinations presented superior performance compared to a random classifier. To our surprise, it's not uncommon to find a combination with three single-prompts having a recall of 0·9 or higher, although these cases are all based on GPT-4 (**Fig 3F**; **Supplementary Fig 2F; Supplementary Fig 4F**).

Based on the preset threshold, we identified the best combination with the highest workload reduction from combinations, which have a recall greater than 0·9. However, in the DM meta-analysis using GPT-3.5, there was only one combination with a recall ≥ 0·9, which only included one single-prompt. Because we wanted to evaluate the performance of prompt strategy 2 and 3, which were specifically tailored for combinations involving multiple single-prompts, we selected another combination ("Research type" and "Disease_p") instead as a sub-best combination for following analyses.

Three prompt strategies show similar performance

Full combination (including all designed single-prompts) and best combination were both evaluated with three prompt strategies (**Table 4; Supplementary File 1**). Obviously, the best combinations had ideal and much better recalls than full combinations and random classifier. The best combinations demonstrated remarkable recalls ranged from 0·900 to 1·000. The corresponding workload reductions varied from 0·122 to 0·640, with an average of 0·395. The sub-best combination from DM meta-analysis also showed good performance, with recalls ranging from 0·778 to 1·000 and workload reductions varying from 0·280 to 0·460.

These three prompt strategies showed comparable levels of performance (**Supplementary Fig 5A-D**), regarding all four metrics. GPT-3.5 and GPT-4 had similar recalls, precisions, and F1 scores (**Supplementary Fig 5E-G**). However, the workload reduction were slightly higher when using GPT-4 than GPT-3.5 (GPT-3.5 vs. GPT-4, medium: 0·303 vs. 0·345; ANOVA test, P = 0·037; **Supplementary Fig 5H**).

Glioma meta-analysis

Literature search

In our glioma meta-analysis, a total of 8550 records were identified after the duplicates were removed (**Fig 5**). In the screening step based on the titles and abstracts of records, 8278 records were filtered out, and 272 records were remained for full text screening. Among these

retained records, 264 records were identified in PubMed. After full text screening, nine records[14,19-26] with 187 participants were included in the current meta-analysis (**Table 5; Supplementary File 5).**

Study characteristics and quality assessment

All research included was carried out on GBM patients (**Table 5**). Most of them are from Phase I or Phase II clinical trials. The immunotherapies utilized in these studies included peptide vaccination,[23,25,26] oncolytic viral vectors therapy,[24] dendritic cell (DC) vaccination,[14,21,22] immune checkpoint inhibitors (ICI),[20] and CAR-NK.[19] Detailed information of these participants was listed in **Supplementary File 5**.

The quality of seven studies was assessed using JBI Critical Appraisal Checklist (**Supplementary Table 2**). Lack of reporting on consecutive/complete inclusion of participants was the most common reason for potential bias. Two studies were evaluated with ROBINS-I tool (**Supplementary Table 2**). One study[26] was judged to be subject to severe risk of bias due to the possibility of potential confounding.

Main results of meta-analysis

Five studies with 143 patients were included to explore the impact of CD8+TIL infiltration on overall survival (OS). However, no statistically significant effect was observed (**Table 6; Supplementary Fig 6**). Similar results were also identified for progress-free survival (PFS) and survival time after immunotherapy (STAI) (**Table 6**). Since there were only two studies reporting treatment response, we didn't synthesis them. In the peptide vaccination research conducted by Narita *et al*.,[25] a clear trend was observed, where all 3 patients with stable disease (SD) exhibited high CD8+TIL infiltration. And the other 5 patients with disease progression (PD) had low CD8+TIL infiltration (**Supplementary Fig 7A**). In another study using CAR-NK,[19] a similar result was also found, suggesting that patients with high CD8+TIL infiltration tended to have an SD response, although no statistical significance was detected (**Supplementary Fig 7B**).

Subgroup analysis

Comprehensive subgroup analyses were performed on several critical indicators (**Table 6; Supplementary Fig 6**). In general, most subgroup analyses detected no effect of CD8+TIL infiltration on OS, PFS, and STAI. The subgroup analysis on peptide vaccination included two studies with 14 patients, and results showed that patients with high CD8+TIL infiltration tended to have a lower OS (hazard ratio (HR) = 6·57; 95% confidence interval (CI): 1·12, 38·58; P = 0·037). In the subgroup analyses on male (HR = 12·34; 95% CI: 1·09, 139·18; P = 0·042) and recurrence (HR = 0·08; 95% CI: 0·01, 0·79; P = 0·030). it should be noted that although results showed important effect on PFS, each subgroup only included a single study. No obvious publication bias was detected in these analyses with significant results (**Supplementary Fig 6).**

Sensitivity analysis

Sensitive analysis was conducted using another method to classify high and low CD8 TIL group based on medium value (**Supplementary Table 3; Supplementary Fig 8)**. When compared to results obtained based on the first classifying method, similar negative results were found, that most analyses presented no significant impact of CD8+TIL infiltration on prognosis. However, the previously identified effect on subgroup analysis disappeared.

Discussion

In the current research, we developed LARS and proved that it can greatly reduce the workload while maintaining an ideal recall during the screening step based on the titles and abstracts of records for meta-analysis. A new glioma meta-analysis was conducted and served as one of the validation datasets, alongside three previously published meta-analyses. In this glioma meta-analysis, we investigated the prognostic predictive value of infiltrating CD8+ TIL in glioma patients treated with immunotherapy.

The mechanism employed by previous AI model and LLMs are quite different. Previous AI models utilized in systemic reviews used active learning to select the training dataset and returned all records ordered by a "similarity".[5] However, LLMs were trained to predict text that follows the input text. By doing so, LLMs could directly answer questions and return whether an input record meet provided criteria or not. As LLMs are pre-trained on large-scale datasets, extra training is unnecessary when applying it on a new meta-analysis (although fine-tuning is possible). However, for previous AI models, a training dataset is required for every new meta-analysis. What's more, because of the same reason, users don't need to worry about imbalanced data problem[5] when using LARS.

Another benefit of LARS is that it could be adapted to other LLMs, except ChatGPT, since most LLMs work in a similar way. We are positive to believe that a well-performed prompt in ChatGPT could also be used for other LLMs, though further research is needed to verify this idea.

The hallucination issue with LLMs, that LLMs make up fake information and describe it like it's real, was emphasized by many researchers.[27,28] For LARS, users need to provide the titles and abstracts of records to ChatGPT, rather than having ChatGPT search for it. In this way, LARS avoids the hallucination problem. Nonetheless, we did observe instances where ChatGPT made false casual inferences. For example, ChatGPT might give a reason supporting a record meeting one criterion, but followed by a opposite judgement. Similar false conclusion may occur when users ask ChatGPT to summarize a final judgement, *e.g.*, "The publication meets criterion 1, but not criterion 2. So, the publication meets all your criteria". Despite false judgement happening sometimes, LARS showed ideal performance in the current research.

Surprisingly, in this study, GPT-4 didn't present an overwhelmingly better performance, compared to GPT-3.5. GPT-4 may be more accurate than GPT-3.5 (**Table 3**). However, recall is much more important than precision in the context of literature screening for meta-analysis. When evaluating the performance of three prompt strategies, GPT-4 and GPT-3.5 shared similar performance across all measures, except workload reduction, where GPT-4 showed a slightly better performance (GPT-3.5 vs. GPT-4, medium: 0·303 vs. 0·345; ANOVA test, P = 0·037; **Supplementary Fig 5H**). Based on these findings, we concluded that neither GPT-3.5 nor GPT-4 was overwhelmingly superior to the other one in this particular context.

Selection of criteria for single-prompts creation is a key step in LARS. Potential criteria

should be derived from the inclusion and exclusion criteria of the designed meta-analysis. However, in some cases, researchers need to extract information from subgroup analysis, which may not be presented in the title and abstract of a record, *e.g.*, materials used in surgery.[29] Criteria related to such information are not suitable for prompt creation. The criteria related to "Species", "Disease", and "Research type" are more likely to be adequately judged using only the title and abstract of record. In fact, majority of the best combinations identified in the current research were based on these three criteria. Users are recommended to try them first when using LARS.

To apply LARS, users are required to manually label a few records for single-prompts, so that the best combination can be identified. Based on our experiences, each single-prompt needs around 10 positive and 10 negative records to be well evaluated. Considering overlaps between these records for single-prompts, researchers need to label about 20 to 100 records for five single-prompts. Once an application based on LARS is developed, it would be much easier to do this labeling.

In this research, we tried three prompt strategies, including a "chain of thought prompt" (prompt strategy 2) designed following the OpenAI's guidelines. Surprisingly, all three prompt strategies showed comparable performance (**Table 4; Supplementary Fig 5A-D**). Indeed, the "chain of thought prompt" takes more time for ChatGPT to response and answer in a more organized format. However, this improvement didn't translate into enhanced performance in LARS. A possible reason is that the two other "less-structured" strategies already sufficiently guided ChatGPT. However, due to the "black box" nature of ChatGPT, we can't really explain the phenomenon. Users are recommended to select any one of them.

In our research, we didn't use metrics like Work Saved over Sampling (WSS) and Average Time to Discover (ATD),[5] which were commonly used to evaluate previous AI. This is because LARS works in a completely different way. Within LARS, ChatGPT will directly answer whether to include or exclude a record, instead of returning a probability for it.

Glioma Meta-analysis

Whether the CD8+ TIL could be a prognostic predictive indicator for glioma patients treated with immunotherapy remains unclear. Narita *et al*. found that recurrent GBM patients with a TIL count of $\geq$ 87 at baseline had a prolonged PFS after treated with peptide vaccination.[25] In the research conducted by Hsu *et al.*, they discovered that a higher estimated TIL count of GBM, prior to DC vaccination, predicted better OS and PFS.[21] In the current meta-analysis, we found no prognostic predictive value of CD8+ TIL infiltration level in all population (**Table 6; Supplementary Fig 6**). A possible explanation is that CD8+ TILs in GBM, prior to immunotherapy, may be exhausted and do not contribute to the treatment effect afterword. Another reason is the small sample size used in the current analysis. However, currently, there is no sufficient data to validate this hypothesis. Further research with well-designed experiments is needed to explore it.

In the present research, we didn't include single-cell RNA sequencing data to deduce the level of CD8+ TILs, because there was little such research meets the data requirement of our meta-analysis. Also, sequencing studies measured the level of CD8+ TILs at RNA level. However, these nine included studies measured CD8 at protein level. To avoid this heterogeneity, we didn't include sequencing studies in the meta-analysis.

Conclusion

We show here that it's possible to have automatic selection of records for meta-analysis with ChatGPT by developing a pipeline named LARS. In the glioma meta-analysis, we found no prognostic predictive value of the CD8+ TILs infiltration at baseline for glioma patients treated with immunotherapy.

Acknowledgments

We would like to thank Marlous van den Braber and Konrad Reichel for their assistance in accessing ChatGPT. We would like to also thank OpenAI for sharing ChatGPT with the research field. This study was funded by the China Scholarship Council (CSC; grant no. 202206090022).

Declaration of interests

All authors declare no competing interests.

Data sharing

The original code used in this paper are available in Github (https://github.com/xiangmingcai/LARS). All responses from ChatGPT can be found in **Supplementary File 2**. Any additional information required is available from the corresponding author upon request.

Table

Table 1 Representative prompt with single criterion (single-prompt)

Table 2 Summary of meta-analyses included as validation datasets for LARS

Table 3 Performance of single-prompts from glioma meta-analysis using GPT-3.5, GPT-4, and random classifier

Table 4 Performance of three prompt strategies with the best combination and the full combination

Table 5 Summary of studies included in glioma meta-analysis

Table 6 Pooled results of the effect of CD8+ TIL infiltration on prognosis in all population and subgroups

Figure legends

Figure 1 The research flow of this study

A representative case showing a request containing a single-prompt and the response from ChatGPT (**A**). The schematic illustrations of the research flow (**B**). Single-prompt represents a prompt with only one criterion. Combined-prompt stands for the prompt with more than one criterion. Color of labels: single-prompt (blue), combined-prompt and prompt strategy (orange), answer and decision (yellow), and true outcome of validation datasets (green).

Figure 2 The responses of GPT-3.5 show high robustness in classifying records

(**A**) The bar plot shows the robustness score of each single-prompt from the four meta-analyses included. (**B**) The stack bar plot shows the answers of repeated requests sent to GPT-3.5 with single-prompts from glioma meta-analysis. N, no; Y, yes; NS, not sure.

Figure 3 Best combination of single-prompts was identified by evaluating the performance of all combinations of single-prompts from glioma meta-analysis

In the Upset plot, the bar chart above represents the evaluation metrics. The dotted line at the bottom presents the single-prompts included in the corresponding combination. Precision (**A**, **E**), recall (**B**, **F**), F1 score (**C**, **G**), and workload reduction (**D**, **H**) are presented for GPT-3.5 (**A-D**) and GPT-4 (**E-F**), respectively. Best combination is marked with a triangle.

Figure 4 Schematic illustration of the LARS pipeline

Figure 5 PRISMA flow diagram

Supplementary materials

Supplementary Figure 1 The responses of GPT-3.5 show high robustness in classifying records

The stack bar plots show the answers of repeated requests sent to GPT-3.5 with single-prompts from inflammatory bowel diseases (**A**), diabetes mellitus (**B**), and sarcopenia (**C**) meta-analyses, respectively. N, no; Y, yes; NS, not sure.

Supplementary Figure 2 Performance of all combinations of single-prompts from inflammatory bowel diseases meta-analysis

Precision (**A**, **E**), recall (**B**, **F**), F1 score (**C**, **G**), and workload reduction (**D**, **H**) are presented for GPT-3.5 (**A-D**) and GPT-4 (**E-F**), respectively. Best combination is marked with a triangle.

Supplementary Figure 3 Performance of all combinations of single-prompts from diabetes mellitus meta-analysis

Precision (**A**, **E**), recall (**B**, **F**), F1 score (**C**, **G**), and workload reduction (**D**, **H**) are presented for GPT-3.5 (**A-D**) and GPT-4 (**E-F**), respectively. Best combination is marked with a red triangle. Sub-best combination is marked with a yellow triangle.

Supplementary Figure 4 Performance of all combinations of single-prompts from sarcopenia meta-analysis

Precision (**A**, **E**), recall (**B**, **F**), F1 score (**C**, **G**), and workload reduction (**D**, **H**) are presented for GPT-3.5 (**A-D**) and GPT-4 (**E-F**), respectively. Best combination is marked with a triangle.

Supplementary Figure 5 Comparison of the performance of best combinations

Comparison of the performance between three prompt strategies, regarding precision (**A**), recall (**B**), F1 score (**C**), and workload reduction (**D**). Comparison of the performance between GPT-3.5 and GPT-4, regarding precision (**E**), recall (**F**), F1 score (**G**), and workload reduction (**H**).

Supplementary Figure 6 Pooled results of glioma meta-analysis

Forrest plots and funnel plots in all population and subgroups with statistical significance.

Supplementary Figure 7 Results of treatment responses

Treatment responses results from studies conducted by (**A**) Narita *et al.* and (**B**) Micheal *et al.*.

Supplementary Figure 8 Sensitive analysis of glioma meta-analysis

Forrest plots and funnel plots in all population and subgroups with statistical significance.

Supplementary Table 1 Performance of single-prompts from inflammatory bowel diseases, diabetes mellitus, and sarcopenia meta-analyses using GPT-3.5, GPT-4, and random classifier

Supplementary Table 2 Quality assessment of studies included in glioma meta-analysis

Supplementary Table 3 Pooled results of the effect of CD8+ TIL infiltration on prognosis in all population and subgroups from sensitive analysis

Supplementary File 1 The content of three prompt strategies using full combination

Supplementary File 2 Validation datasets from four meta-analyses

Supplementary File 3 PRISMA report checklist

Supplementary File 4 Supplementary Methods

Supplementary File 5 Detailed information of participants included in glioma meta-analysis

Supplementary File 6 Literature search strategies of the glioma meta-analysis

Reference

1. Subbiah V. The next generation of evidence-based medicine. *Nat Med* 2023; **29**: 49-58.
2. Abdelkader W, Navarro T, Parrish R, et al. A Deep Learning Approach to Refine the Identification of High-Quality Clinical Research Articles From the Biomedical Literature: Protocol for Algorithm Development and Validation. *JMIR Res Protoc* 2021; **10**: e29398.

3. Tercero-Hidalgo JR, Khan KS, Bueno-Cavanillas A, et al. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. *J Clin Epidemiol* 2022; **148**: 124-34.
4. Gates A, Gates M, Sebastianski M, et al. The semi-automation of title and abstract screening: a retrospective exploration of ways to leverage Abstrackr's relevance predictions in systematic and rapid reviews. *BMC Med Res Methodol* 2020; **20**: 139.
5. Ferdinands G, Schram R, de Bruin J, et al. Performance of active learning models for screening prioritization in systematic reviews: a simulation study into the Average Time to Discover relevant records. *Syst Rev* 2023; **12**: 100.
6. van Dis EAM, Bollen J, Zuidema W, van Rooij R, Bockting CL. ChatGPT: five priorities for research. *Nature* 2023; **614**: 224-26.
7. Li H, Moon JT, Purkayastha S, et al. Ethics of large language models in medicine and medical research. *Lancet Digit Health* 2023; **5**: e333-e35.
8. Ali SR, Dobbs TD, Hutchings HA, Whitaker IS. Using ChatGPT to write patient clinic letters. *Lancet Digit Health* 2023; **5**: e179-e81.
9. [preprint] Shaib C, Li M, Joseph S, et al. Summarizing, Simplifying, and Synthesizing Medical Evidence Using GPT-3 (with Varying Success). *ArXiv* 2023; **abs/2305.06299**.
10. [preprint] Wang S, Scells H, Koopman B, Zuccon G. Can ChatGPT Write a Good Boolean Query for Systematic Review Literature Search? *ArXiv* 2023; **abs/2302.03495**.
11. Borah R, Brown AW, Capers PL, Kaiser KA. Analysis of the time and workers needed to conduct systematic reviews of medical interventions using data from the PROSPERO registry. *BMJ Open* 2017; **7**: e012545.
12. Wu W, Klockow JL, Zhang M, et al. Glioblastoma multiforme (GBM): An overview of current therapies and mechanisms of resistance. *Pharmacol Res* 2021; **171**: 105780.
13. Lynes JP, Nwankwo AK, Sur HP, et al. Biomarkers for immunotherapy for treatment of glioblastoma. *J Immunother Cancer* 2020; **8**.
14. Dejaegher J, Solie L, Hunin Z, et al. DNA methylation based glioblastoma subclassification is related to tumoral T-cell infiltration and patient survival. *Neuro Oncol* 2021; **23**: 240-50.
15. Talebi S, Zeraattalab-Motlagh S, Rahimlou M, et al. The Association between Total Protein, Animal Protein, and Animal Protein Sources with Risk of Inflammatory Bowel Diseases: A Systematic Review and Meta-Analysis of Cohort Studies. *Adv Nutr* 2023; **14**: 752-61.
16. Aune D, Schlesinger S, Mahamat-Saleh Y, et al. Diabetes mellitus, prediabetes and the risk of Parkinson's disease: a systematic review and meta-analysis of 15 cohort studies with 29.9 million participants and 86,345 cases. *Eur J Epidemiol* 2023; **38**: 591-604.
17. Beaudart C, Demonceau C, Reginster JY, et al. Sarcopenia and health-related quality of life: A systematic review and meta-analysis. *J Cachexia Sarcopenia Muscle* 2023; **14**: 1228-43.
18. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *Bmj* 2021; **372**: n71.
19. Burger MC, Forster MT, Romanski A, et al. Intracranial injection of NK cells engineered with a HER2-targeted chimeric antigen receptor in patients with recurrent glioblastoma. *Neuro Oncol* 2023.
20. Friedman JS, Jun T, Rashidipour O, et al. Using EGFR amplification to stratify recurrent glioblastoma treated with immune checkpoint inhibitors. *Cancer Immunol Immunother* 2023; **72**: 1893-901.
21. Hsu M, Sedighim S, Wang T, et al. TCR Sequencing Can Identify and Track Glioma-Infiltrating T Cells after DC Vaccination. *Cancer Immunol Res* 2016; **4**: 412-18.
22. Jan CI, Tsai WC, Harn HJ, et al. Predictors of Response to Autologous Dendritic Cell Therapy in Glioblastoma Multiforme. *Front Immunol* 2018; **9**: 727.
23. Keskin DB, Anandappa AJ, Sun J, et al. Neoantigen vaccine generates intratumoral T

cell responses in phase Ib glioblastoma trial. *Nature* 2019; **565**: 234-39.

24. Markert JM, Liechty PG, Wang W, et al. Phase Ib trial of mutant herpes simplex virus G207 inoculated pre-and post-tumor resection for recurrent GBM. *Mol Ther* 2009; **17**: 199-207.

25. Narita Y, Okita Y, Arakawa Y. Evaluation of the efficacy and safety of TAS0313 in adults with recurrent glioblastoma. *Cancer Immunol Immunother* 2022; **71**: 2703-15.

26. Steiner HH, Bonsanto MM, Beckhove P, et al. Antitumor vaccination of patients with glioblastoma multiforme: a pilot study to assess feasibility, safety, and clinical benefit. *J Clin Oncol* 2004; **22**: 4272-81.

27. Jin Q, Leaman R, Lu Z. Retrieve, Summarize, and Verify: How Will ChatGPT Affect Information Seeking from the Medical Literature? *J Am Soc Nephrol* 2023.

28. Stokel-Walker C, Van Noorden R. What ChatGPT and generative AI mean for science. *Nature* 2023; **614**: 214-16.

29. Cai X, Yang J, Zhu J, et al. Reconstruction strategies for intraoperative CSF leak in endoscopic endonasal skull base surgery: systematic review and meta-analysis. *Br J Neurosurg* 2022; **36**: 436-46.

Table 1 Representative prompt with single criterion (single-prompt)

single-prompt name	single-prompt content
Species	I want you to act as a helpful assistant. I will give you title and abstract of a publication and you will reply whether it meets our criteria or not. I want you to only reply with yes, no, or not sure, and followed with reasons. The criteria is: studies that use human as primary research subject.
Disease	.... The criteria is: studies that involve patients with glioma, glioblastoma, astrocytoma, oligodendroglioma.
Research type	.... The criteria is: studies that are prospective or retrospective cohort study, case-control study. Of note, these research types doesn't meet the criteria: cross-sectional study, randomized controlled trial, review, protocol or others.
Age	.... The criteria is: studies that involve adult patients (at least 18 years old).
Protein related	.... The criteria is: The title and abstract must mention that the study is related to the consumption of protein (e.g., total dairy, milk, meat, fish, poultry, process meat, and egg).

Table 2 Summary of meta-analyses included as validation datasets for LARS

First author	Field	Publication year	Journal	Original research			Our repetition (validation datasets)	
				All identified records	Identified records from Medline/PubMed	Records preserved in title and abstract screen step	Identified records from Medline/PubMed	Records preserved in title and abstract screen step (Matched)
Cai X.	Glioma	Current research		8550	6020	272	1360	264
Talebi S.	Inflammatory Bowel Diseases	2023	Adv Nutr	2755	1285	51	1284	45
Aune D.	Diabetes Mellitus	2023	Eur J Epidemiol	5320	1040	216	1039	124
Beaudart C.	Sarcopenia	2023	J Cachexia Sarcopenia Muscle	2293	NA	188	1293	122

Table 3 Performance of single-prompts from glioma meta-analysis using GPT-3.5, GPT-4, and random classifier

Glioma single-prompt	**Model**	**Precision**	**Recall**	**F1**	**Workload reduction**
Species	GPT-35	0.587	0.786	0.672	0.250
	GPT-4	0.791	0.607	0.687	0.570
	Random classifier	0.557	0.494	0.523	0.504
Disease	GPT-35	0.989	0.905	0.945	0.130
	GPT-4	1.000	1.000	1.000	0.050
	Random classifier	0.950	0.506	0.659	0.494
Treatment	GPT-35	0.530	0.917	0.672	0.170
	GPT-4	0.745	0.792	0.768	0.490
	Random classifier	0.485	0.504	0.494	0.501
Research type	GPT-35	0.915	0.966	0.940	0.060
	GPT-4	0.946	0.989	0.967	0.070
	Random classifier	0.893	0.495	0.635	0.506

Table 4 Performance of three prompt strategies with the best combination and the full combination

Glioma		**Best Combination**				**Full Combination**			
Model	**Prompt strategy**	**Precision**	**Recall**	**F1**	**Workload reduction**	**Precision**	**Recall**	**F1**	**Workload reduction**
GPT-3.5	Prompt strategy 1	0.257	0.913	0.401	0.310	0.296	0.830	0.436	0.456
	Prompt strategy 2	0.246	0.932	0.390	0.265	0.471	0.772	0.585	0.678
	Prompt strategy 3	0.217	0.981	0.356	0.122	0.358	0.852	0.504	0.538
Random classifier		0.194	0.499	0.280	0.502	0.194	0.499	0.280	0.502
GPT-4	Prompt strategy 1	0.263	1.000	0.417	0.620	0.400	0.800	0.533	0.800
	Prompt strategy 2	0.119	1.000	0.213	0.160	0.159	0.700	0.259	0.560
	Prompt strategy 3	0.125	1.000	0.222	0.200	0.233	0.700	0.350	0.700
Random classifier		0.100	0.495	0.166	0.504	0.100	0.495	0.166	0.504

Inflammatory Bowel Diseases		**Best Combination**				**Full Combination**			
Model	**Prompt strategy**	**Precision**	**Recall**	**F1**	**Workload reduction**	**Precision**	**Recall**	**F1**	**Workload reduction**
GPT-3.5	Prompt strategy 1	0.042	0.978	0.081	0.188	0.047	0.444	0.086	0.670
	Prompt strategy 2	0.045	0.956	0.086	0.252	0.180	0.533	0.270	0.896
	Prompt strategy 3	0.044	0.978	0.084	0.212	0.084	0.778	0.152	0.675
Random classifier		0.036	0.514	0.067	0.502	0.036	0.514	0.067	0.502
GPT-4	Prompt strategy 1	0.152	1.000	0.263	0.340	0.500	0.200	0.286	0.960
	Prompt strategy 2	0.136	0.900	0.237	0.340	1.000	0.200	0.333	0.980
	Prompt strategy 3	0.154	1.000	0.267	0.350	0.500	0.100	0.167	0.980
Random classifier		0.098	0.485	0.163	0.504	0.098	0.485	0.163	0.504

Diabetes Mellitus		**Sub-best Combination**				**Full Combination**			
Model	**Prompt strategy**	**Precision**	**Recall**	**F1**	**Workload reduction**	**Precision**	**Recall**	**F1**	**Workload reduction**
GPT-3.5	Prompt strategy 1	0.173	0.823	0.287	0.434	0.165	0.379	0.230	0.726
	Prompt strategy 2	0.171	0.806	0.282	0.436	0.520	0.106	0.176	0.976
	Prompt strategy 3	0.133	0.782	0.227	0.295	0.150	0.847	0.255	0.322
Random classifier		0.118	0.492	0.190	0.502	0.118	0.492	0.190	0.502
GPT-4	Prompt strategy 1	0.125	1.000	0.222	0.280	0.263	0.556	0.357	0.810
	Prompt strategy 2	0.119	0.889	0.211	0.330	0.222	0.444	0.296	0.820
	Prompt strategy 3	0.130	0.778	0.222	0.460	0.143	0.667	0.235	0.580
Random classifier		0.089	0.491	0.151	0.504	0.089	0.491	0.151	0.504

Sarcopenia		**Best Combination**				**Full Combination**			
Model	**Prompt strategy**	**Precision**	**Recall**	**F1**	**Workload reduction**	**Precision**	**Recall**	**F1**	**Workload reduction**
GPT-3.5	Prompt strategy 1	0.158	0.951	0.271	0.433	0.166	0.418	0.237	0.762

	Prompt strategy 2	0.166	0.975	0.284	0.445	0.286	0.148	0.195	0.951
	Prompt strategy 3	0.136	0.967	0.238	0.327	0.208	0.902	0.337	0.590
Random classifier		0.093	0.493	0.157	0.501	0.093	0.493	0.157	0.501
GPT-4	Prompt strategy 1	0.250	0.900	0.391	0.640	0.333	0.100	0.154	0.970
	Prompt strategy 2	0.250	0.900	0.391	0.640	0.500	0.600	0.545	0.880
	Prompt strategy 3	0.243	0.900	0.383	0.630	0.500	0.500	0.500	0.900
Random classifier		0.099	0.488	0.164	0.504	0.099	0.488	0.164	0.504

Table 5 Summary of studies included in glioma meta-analysis

Research	Design	Sample size	Disease	Age (year)	Gender*	Primary/Recurrence type	IDH mutation status**	MGMT methylation status***	Immunotherapy	Extension of resection****	Measure for CD8 TIL	OS	PFS	Survival data after immunotherapy	Treatment response
Steiner_2004	Phase II	7	GBM	30-57	4 M; 3 F	Primary	NA	NA	Peptide Vaccination	5 TE; 2 NA	CD8+ cells per mm2 tumor tissue	Yes	Yes	NA	NA
Markert_2009	Phase Ib	6	GBM	39.1-66	2 M; 4 F	Recurrence	WT	NA	Oncolytic viral vectors therapy	NA	%	NA	NA	Yes	NA
Hsu_2016	Phase I and phase II	15	GBM	NA	NA	6 Recurrence and 9 Primary	WT	3 M; 10 unM; 2 NA	DC Vaccination	NA	Estimated TIL percentage	Yes	Yes	NA	NA
Jan_2018	Phase II	27	GBM	27-68	12 M; 15 F	Primary	1 MUT; 24 WT	23 M, 4 unM	DC Vaccination	15 TE; 3 SE; 4 PE; 5 NA	CD8+ cell counts in 25 high-power field	Yes	Yes	NA	NA
Keskin_2019	Phase Ib	8	GBM	adult	NA	Primary	WT	unM	Peptide Vaccination	NA	CD8+ cells per mm2	Yes	Yes	NA	NA
Dejaegher_2021	Prospective cohort	93	GBM	36-70	61 M; 32 F	Primary	7 MUT; 3 NA; 83 WT	40 M; 38 unM; 15 NA	DC Vaccination	49 TE; 32 SE; 12 PE	CD8+ cell counts in 10 high power fields	Yes	NA	NA	NA
Narita_2022	PhaseI/II	10	GBM	adult	NA	Recurrence	NA	NA	Peptide Vaccination	NA	CD8+ count of ≥87 or <87 at baseline	NA	NA	NA	Yes
Friedman_2023	Retrospective cohort	12	GBM	adult	NA	5 Recurrence and 7 Primary	WT	NA	Immune Checkpoint Inhibitors	NA	%	NA	NA	Yes	NA
Michael_2023	Phase I	9	GBM	30-61	7 M; 2 F	Recurrence	WT	4 M; 5 unM	CAR-NK	NA	%	NA	Yes	Yes	Yes

NA, not available; GBM, glioblastoma; DC, dendritic cell; CAR-NK, chimeric antigen receptor NK-cell therapy; CD8 TIL, CD8+ tumor infiltrating lymphocytes

*Gender: M, Male; F, Female

**IDH mutation status: WT, wild type; MUT, mutated type

***MGMT methylation status: M, methylated; unM, unmethylated

****Extension of resection: TE, Total excision; SE, Subtotal excision; PE, Partial excision

Overall survival (OS)

Table 6 Pooled results of the effect of CD8+ TIL infiltration on prognosis in all population and subgroups

Subgroup	Study number	Patient number	HR	CI (lower)	CI (upper)	P	I^2	P (Q test)
All population	5	143	1.20	0.42	3.43	0.730	57.00 %	0.054
Gender								
Male	2	70	1.17	0.49	2.77	0.722	0.00 %	0.368
Female	1	13	1.07	0.22	5.15	0.933	NA	NA
IDH mutation status								
Wild type	4	124	0.76	0.29	1.94	0.562	52.10 %	0.100
Mutated type	0	0	NA	NA	NA	NA	NA	NA
MGMT methylation status								
unmethylated	2	44	2.27	0.45	11.58	0.323	0.00 %	0.443
methylated	2	59	1.30	0.50	3.35	0.591	0.00 %	0.359
Extension of resection								
Total excision	3	65	1.22	0.45	3.31	0.703	0.00 %	0.596
Subtotal excision	1	32	1.65	0.38	7.25	0.507	NA	NA
Partial excision	0	0	NA	NA	NA	NA	NA	NA
Primary/Recurrence type								
Primary	4	128	1.59	0.69	3.68	0.276	27.40 %	0.248
Recurrence	0	0	NA	NA	NA	NA	NA	NA
Immunotherapy								
DC Vaccination	3	129	0.73	0.28	1.91	0.524	54.10 %	0.113
Peptide Vaccination	**2**	**14**	**6.57**	**1.12**	**38.58**	**0.037**	**0.00 %**	**0.867**
Oncolytic viral vectors therapy	0	0	NA	NA	NA	NA	NA	NA
Immune Checkpoint Inhibitors	0	0	NA	NA	NA	NA	NA	NA
CAR-NK	0	0	NA	NA	NA	NA	NA	NA

Progress free survival (PFS)

Subgroup	Study number	Patient number	HR	CI (lower)	CI (upper)	P	I^2	P (Q test)
All population	5	61	0.70	0.22	2.18	0.536	58.30 %	0.048
Gender								
Male	**1**	**10**	**12.34**	**1.09**	**139.18**	**0.042**	**NA**	**NA**
Female	1	13	0.86	0.18	4.20	0.854	NA	NA
IDH mutation status								
Wild type	3	43	0.82	0.22	3.07	0.763	55.90 %	0.104
Mutated type	0	0	NA	NA	NA	NA	NA	NA
MGMT methylation status								
unmethylated	1	7	2.46	0.22	27.28	0.465	NA	NA
methylated	1	21	1.32	0.37	4.72	0.666	NA	NA

Extension of resection								
Total excision	2	17	0.84	0.23	3.10	0.791	0.00 %	0.845
Subtotal excision	0	0	NA	NA	NA	NA	NA	NA
Partial excision	0	0	NA	NA	NA	NA	NA	NA
Primary/Recurrence type								
Primary	3	37	1.66	0.70	3.94	0.250	0.00 %	0.937
Recurrence	**1**	**9**	**0.08**	**0.01**	**0.79**	**0.030**	**NA**	**NA**
Immunotherapy								
DC Vaccination	2	38	0.64	0.11	3.69	0.618	75.00 %	0.045
Peptide Vaccination	3	23	0.73	0.10	5.54	0.763	64.00 %	0.062
Oncolytic viral vectors therapy	0	0	NA	NA	NA	NA	NA	NA
Immune Checkpoint Inhibitors	0	0	NA	NA	NA	NA	NA	NA
CAR-NK	0	0	NA	NA	NA	NA	NA	NA

Survival time after immunotherapy (STAI)								
Subgroup	**Study number**	**Patient number**	**HR**	**CI (lower)**	**CI (upper)**	**P**	**I^2**	**P (Q test)**
All population	3	18	1.48	0.37	5.89	0.583	0.00 %	0.585
Gender								
Male	0	0	NA	NA	NA	NA	NA	NA
Female	1	4	2.45	0.15	39.72	0.529	NA	NA
IDH mutation status								
Wild type	3	18	1.48	0.37	5.89	0.583	0.00 %	0.585
Mutated type	0	0	NA	NA	NA	NA	NA	NA
MGMT methylation status								
unmethylated	0	0	NA	NA	NA	NA	NA	NA
methylated	0	0	NA	NA	NA	NA	NA	NA
Extension of resection								
Total excision	0	0	NA	NA	NA	NA	NA	NA
Subtotal excision	0	0	NA	NA	NA	NA	NA	NA
Partial excision	0	0	NA	NA	NA	NA	NA	NA
Primary/Recurrence type								
Primary	1	7	0.89	0.10	8.20	0.919	NA	NA
Recurrence	2	11	2.03	0.35	11.96	0.433	0.00 %	0.387
Immunotherapy								
DC Vaccination	0	0	NA	NA	NA	NA	NA	NA
Peptide Vaccination	0	0	NA	NA	NA	NA	NA	NA
Oncolytic viral vectors therapy	1	6	1.07	0.11	10.57	0.953	NA	NA
Immune Checkpoint Inhibitors	2	12	1.77	0.31	10.09	0.519	0.00 %	0.328
CAR-NK	0	0	NA	NA	NA	NA	NA	NA

HR, hazard ratio; CI, confidence interval; DC, dendritic cell; NA, not applicable.

Figure legends

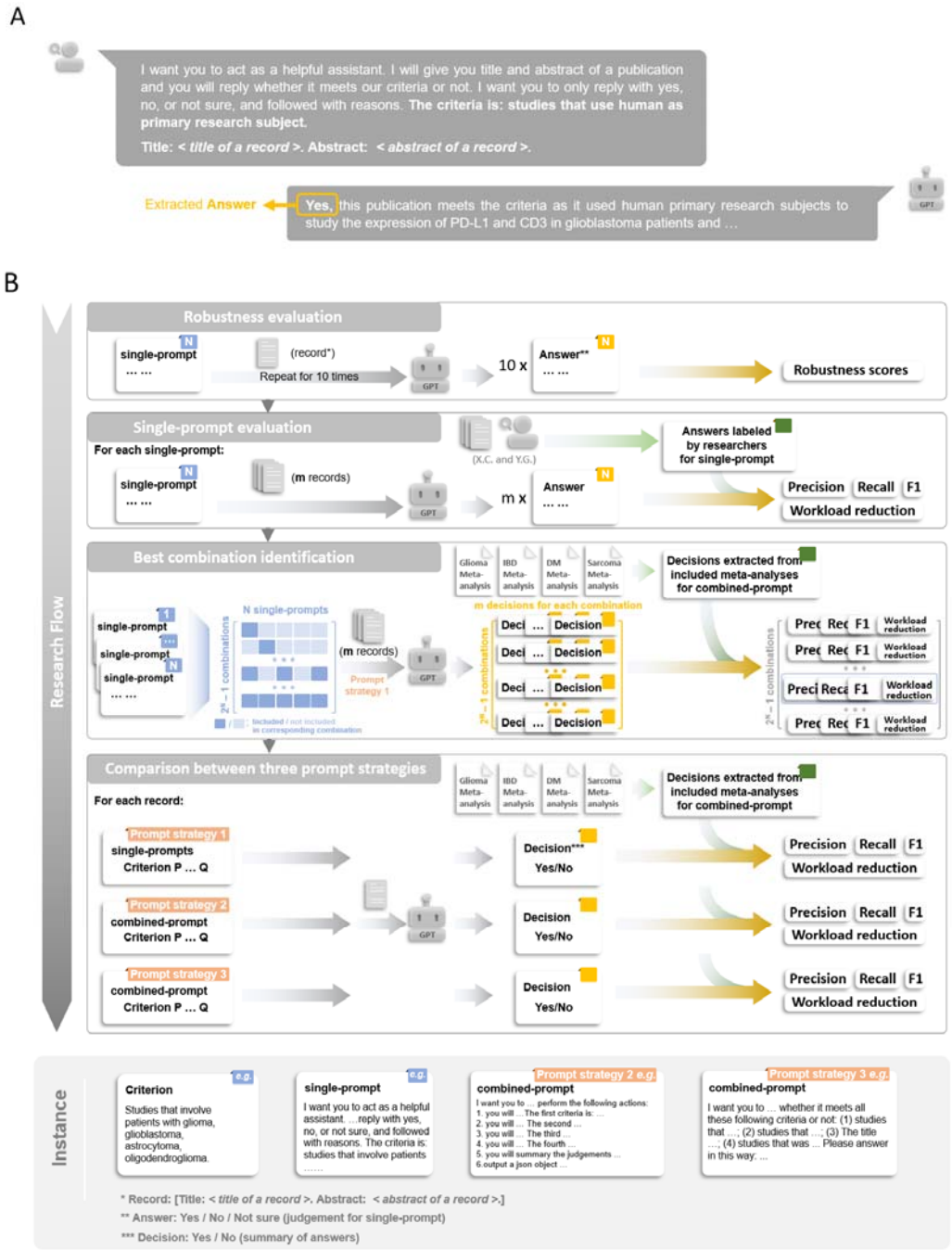

Figure 1 The research flow of this study

A representative case showing a request containing a single-prompt and the response from ChatGPT (**A**). The schematic illustrations of the research flow (**B**). Single-prompt represents a prompt with only one criterion. Combined-prompt stands for the prompt with more than one criterion. Color of labels: single-prompt (blue), combined-prompt and prompt strategy (orange), answer and decision (yellow), and true outcome of validation datasets (green).

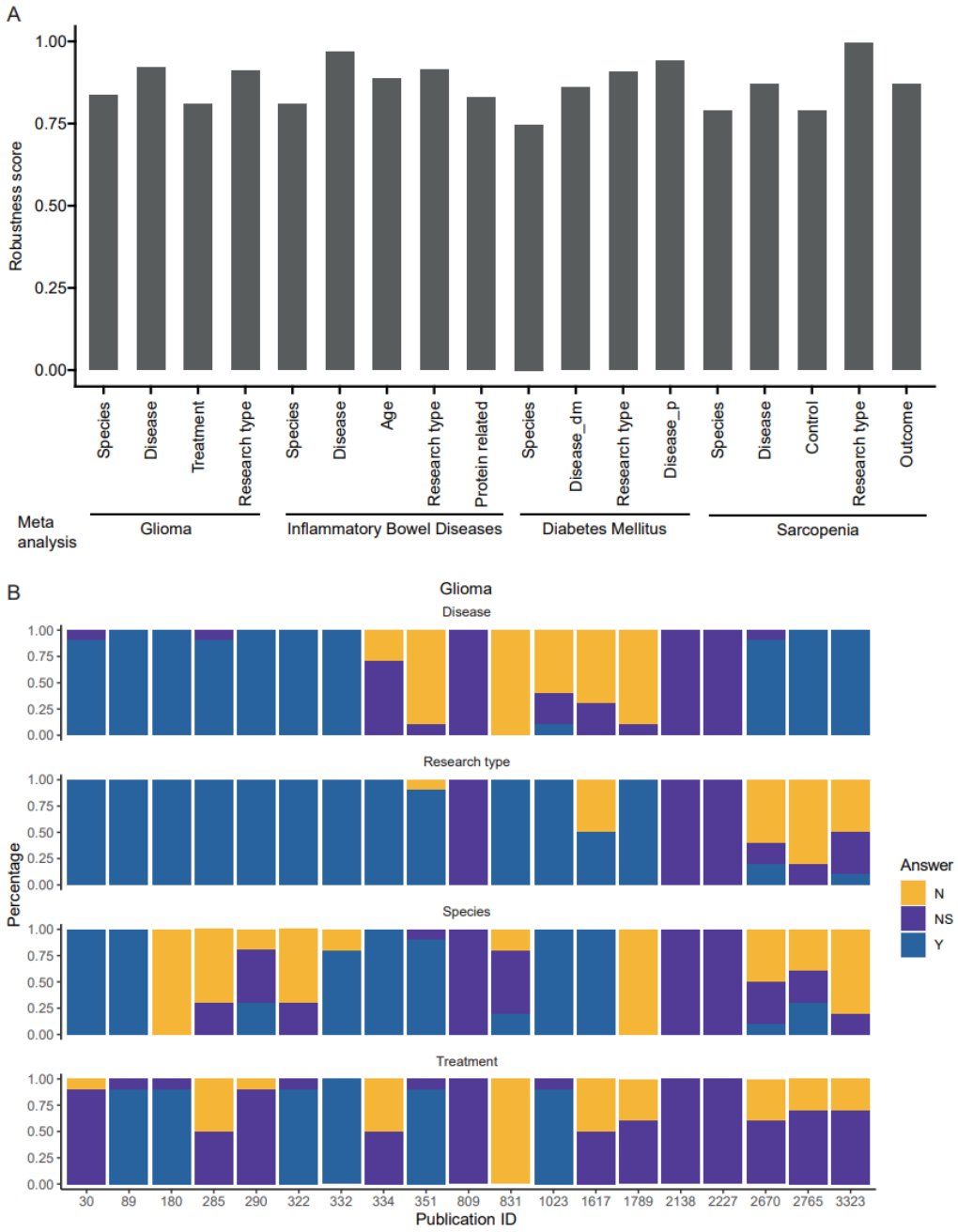

Figure 2 The responses of GPT-3.5 show high robustness in classifying records

(**A**) The bar plot shows the robustness score of each single-prompt from the four meta-analyses included. (**B**) The stack bar plot shows the answers of repeated requests sent to GPT-3.5 with single-prompts from glioma meta-analysis. N, no; Y, yes; NS, not sure.

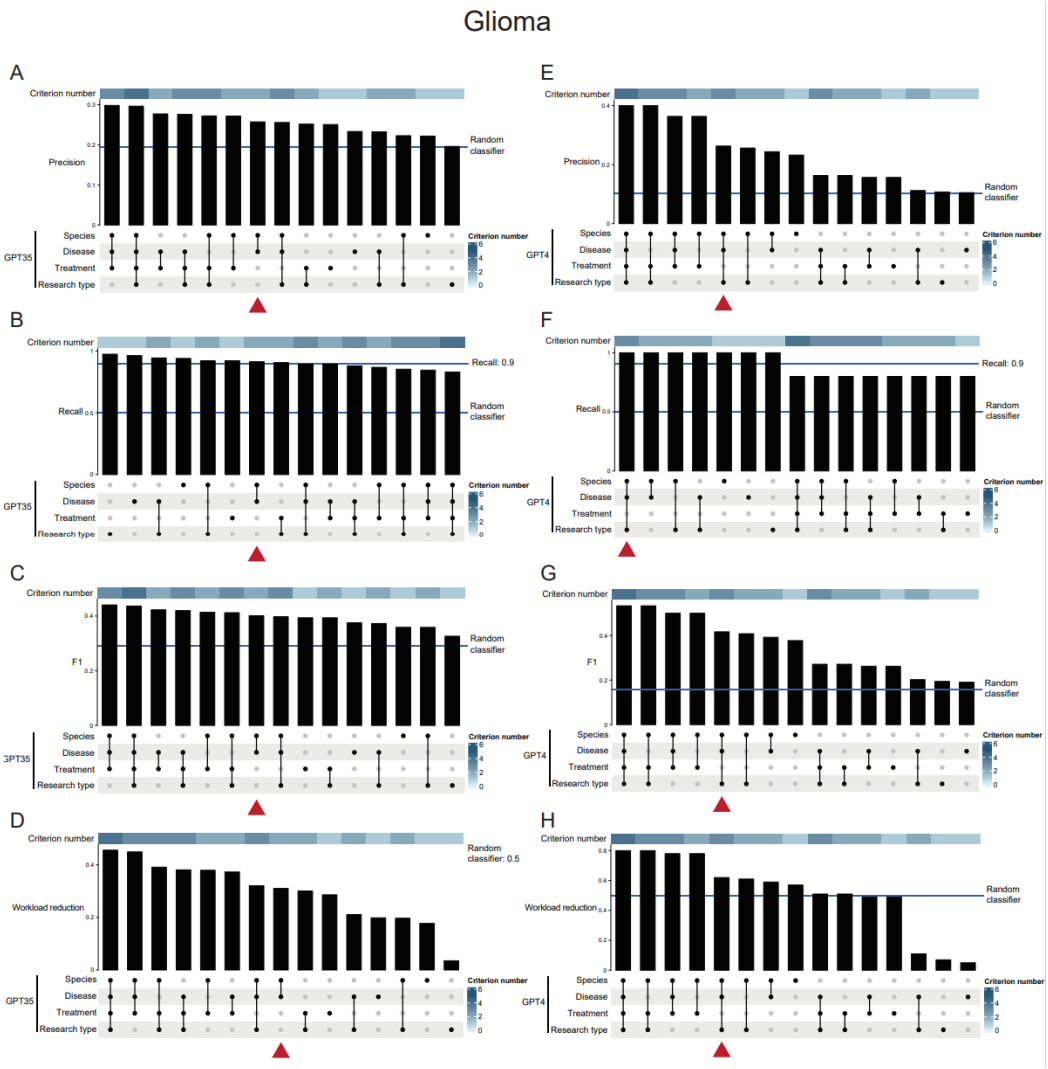

Figure 3 Best combination of single-prompts was identified by evaluating the performance of all combinations of single-prompts from glioma meta-analysis

In the Upset plot, the bar chart above represents the evaluation metrics. The dotted line at the bottom presents the single-prompts included in the corresponding combination. Precision (**A**, **E**), recall (**B**, **F**), F1 score (**C**, **G**), and workload reduction (**D**, **H**) are presented for GPT-3.5 (**A-D**) and GPT-4 (**E-F**), respectively. Best combination is marked with a triangle.

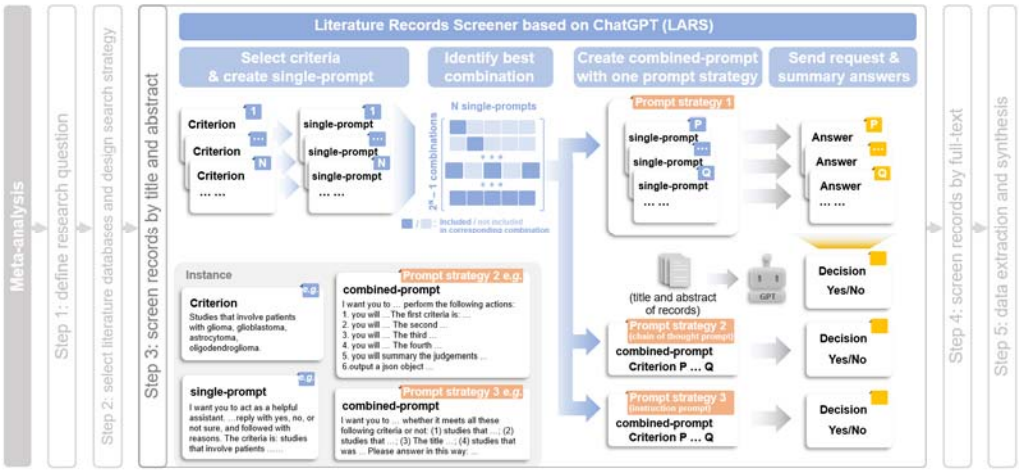

Figure 4 Schematic illustration of the LARS pipeline

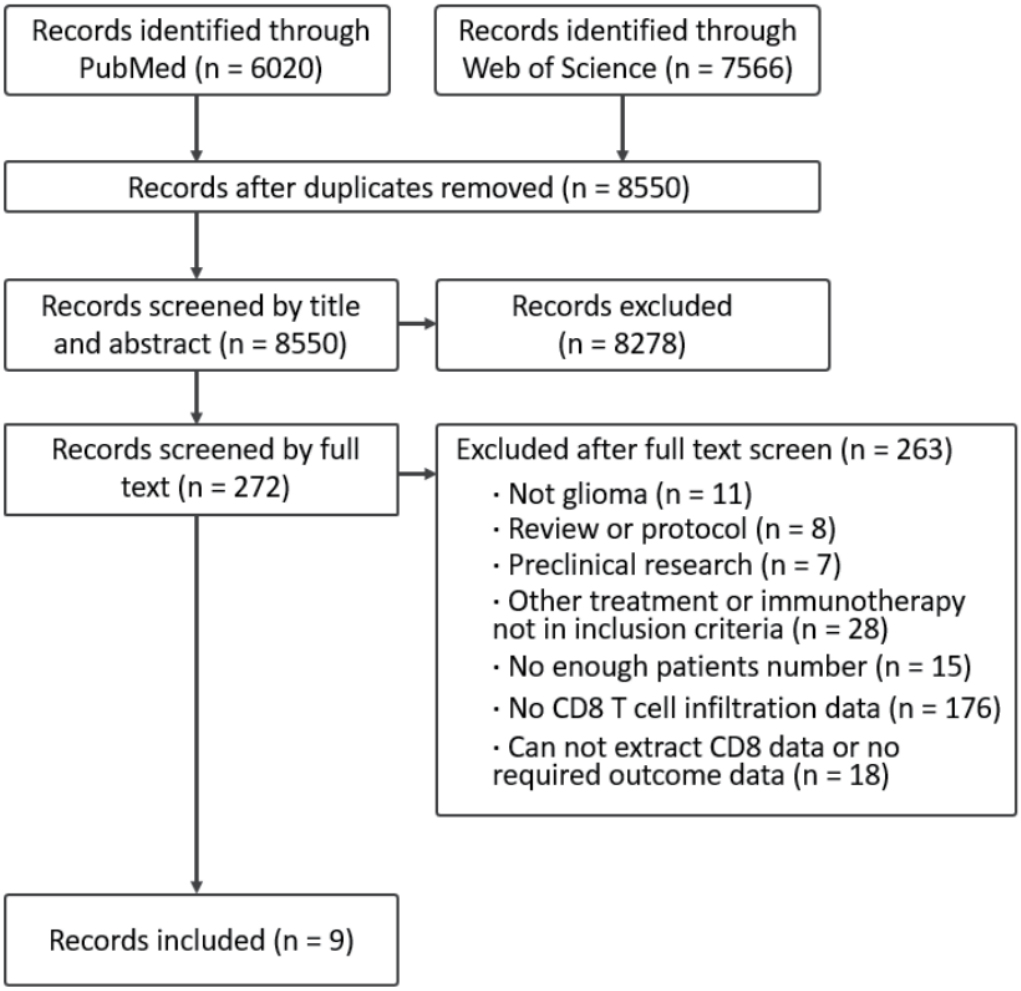

Figure 5 PRISMA flow diagram