UTRECHT UNIVERSITY

MASTER THESIS

Filling in evidence tables with LLMs: A proof-of-concept study

Carlos Vidal-Perez

Student number: 5182913

MSc Applied Data Science

Supervisors: Rens van de Schoot and Tim Christen

Examiners: Rens van de Schoot and Duco Veen

Academic year: 2023-24

Abstract:

The development of medical guidelines is essential but often labor-intensive and time-consuming. This study explores the potential of leveraging Large Language Models (LLMs), specifically GPT-4o, to automate the creation of evidence tables from Randomized Controlled Trials (RCTs). We designed and iteratively refined a prompt to optimize the data extraction from the studies. Then, we evaluated the model's performance column by column by comparing the numbers and texts extracted by the model with those manually extracted by healthcare experts. Numeric data was assessed by comparing the set of numbers extracted by each, while textual data was analyzed using 5 similarity measures: TF-IDF, Jaccard, BERT, Sentence-BERT, and spaCy. The results demonstrated that GPT-4o effectively extracted and summarized key elements of the studies, showcasing its potential to streamline the development of medical guidelines. This approach promises to reduce the workload of healthcare professionals, improve efficiency and ensure that patient care is based on the most current and comprehensive data available.

Contents

1 Introduction

In the rapidly evolving field of healthcare, the development of medical guidelines is a critical process that ensures clinical practices are based on the best available evidence [Grimshaw and Russell, 1993]. These guidelines help standardize care, improve patient outcomes, and optimize resource utilization. However, the current method of developing these guidelines is both labor-intensive and time-consuming since it involves manually finding the relevant studies and then extracting, synthesizing and summarizing the information from them. This process is not only prone to human fatigue, but also increasingly unsustainable given the exponential growth of scientific literature [Ghasemi et al., 2022].

The creation of a medical guideline begins with a question posed by healthcare professionals [Guyatt et al., 2011]. This question addresses a specific clinical issue, such as the best treatment approach for a particular condition. Once the question is defined, a comprehensive search for relevant studies is conducted. This search involves screening a vast number of papers to identify those that are pertinent to the question at hand. Machine learning projects have already been used to enhance this screening process, helping to efficiently sift through large volumes of literature to identify potentially relevant studies faster [Van De Schoot et al., 2021, Kebede et al., 2023].

Following the screening process, the next crucial step is to extract and summarize the information from these relevant studies [Guyatt et al., 2011]. This is where evidence tables come into play. They are structured summaries that distill the key elements of each study, including study design, population characteristics, interventions, outcomes and statistical findings. These tables provide a clear, concise and organized representation of the papers, making it easier for guideline developers to compare and contrast the findings from different studies. The accuracy and comprehensiveness of these tables are essential, as they form the foundation for the recommendations that will be made in the final guideline. However, the current manual approach to creating them is highly time-consuming and requires significant expertise [Nussbaumer-Streit et al., 2021].

Nevertheless, recent advancements in artificial intelligence (AI) and natural language processing (NLP) present a promising solution to this challenge. Large Language Models (LLMs), such as OpenAI's ChatGPT, have shown significant potential in understanding and generating human-like text [Achiam et al., 2023]. These models are capable of performing complex language tasks, including summarization, question-answering and text generation, making them suitable candidates for automating the creation of evidence tables from healthcare studies. Previous studies have shown the effectiveness of LLMs in assisting with systematic reviews [Alshami et al., 2023] and the extraction of data from research articles [Dagdelen et al., 2024, Polak and Morgan, 2024]. Specifically, [Gartlehner et al., 2024] demonstrated the promising potential of LLMs to extract information from medical studies for evidence synthesis, further supporting their potential use for filling in evidence tables.

This paper explored the potential of leveraging LLMs to automate the creation of evidence tables for Randomized Controlled Trials (RCTs). RCTs are considered the gold standard in clinical research [Devereaux and Yusuf, 2003] because they randomly assign participants to either the intervention group or the control group, thereby minimizing bias and providing the most rigorous evidence on the efficacy of treatments [Barton, 2000]. The primary goal is to assess the efficiency and reliability of this automation, thereby opening the possibility for healthcare professionals to reduce the time spent on data extraction tasks. By doing this, we can ensure that the most up-to-date and comprehensive data is used in formulating medical guidelines, ultimately leading to improved patient outcomes.

To achieve this, we used prompt engineering to optimize LLMs for this specific task. Prompt engineering is the process of designing and refining the inputs (prompts) given to a language model to elicit the most accurate and relevant responses. This involves carefully crafting the wording and structure of the prompts, providing context, specific rules to follow and detailed explanations of the information to extract [White et al., 2023, Wang et al., 2023]. Guiding the model's output by iteratively refining these prompts helped us understand the potential and limitations of LLMs in evidence synthesis and serve as a proof-of-concept for the feasibility of using AI to streamline the development of medical guidelines.

In what follows, we first describe the methodology, including the data used and the stages of the study. Then, we present the results of this analysis after implementing the methodology. Finally, we discuss the findings, noting limitations and suggesting future research directions. All scripts and publicly available data are available on GitHub at https://github.com/Carlos-Vi/ADS-thesis.

2 Methodology

2.1 Data

The data used to validate the results of this research comprised 15 RCT medical papers in PDF format. These papers were accompanied by their corresponding manually created evidence tables in Word format, which summarized each study. The studies selected for this research were randomly chosen from the medical guidelines created by the Kennisinstituut of the Federatie Medisch Specialisten (Knowledge Institute of the Federation of Medical Specialists), accessible at www.richtlijnendatabase.nl. They covered treatments for asthma, chronic obstructive pulmonary disease (COPD), acute kidney injury, bronchiolitis and recommendations for cardiac rehabilitation. The evidence tables were publicly available on the same webpage of the Institute.

These 15 papers were randomly divided into a training set of 2 papers [Gaudry et al., 2021, Modaressi et al., 2012] for the first part of the research and a test set of 13 papers [Bashir et al., 2018, Bremner et al., 2018, Djamin et al., 2019, Djamin et al., 2020, Faten et al., 2014, Flores-González et al., 2015, Geng et al., 2020, Jaquet-Pilloud et al., 2020, Raeisi et al., 2019, Risom et al., 2020, Ruangsomboon et al., 2021, Skjerven et al., 2013, Uysalol et al., 2017] for the final stage.

To enable the LLM to process the PDF studies, it was necessary to first convert them into plain text (.txt) files and then compress them to reduce their number of tokens by removing unnecessary spaces such as tabs, double spaces or white lines. This format transformation was performed using the *'fitz'* (PyMuPDF) and *'camelot'* Python libraries, , while the compression was handled by the *'re'* library. Additionally, the human curated Word tables had to be converted to text format using the *'docx'* library, which was necessary for the subsequent evaluation of the results. The code used for this preprocessing step can be found at https://github.com/Carlos-Vi/ADS-thesis/tree/main/Code/Pre-processing.

2.2 Study design

The experiment was structured into three main stages (see Figure 1). 1) Designing an initial version of the prompt to extract the desired information. 2) Iteratively refining it using ChatGPT-4o with the training dataset to develop an improved final version of the prompt. 3) Extracting the evidence tables of the test dataset through the GPT-4o API and assessing the quality of the results.

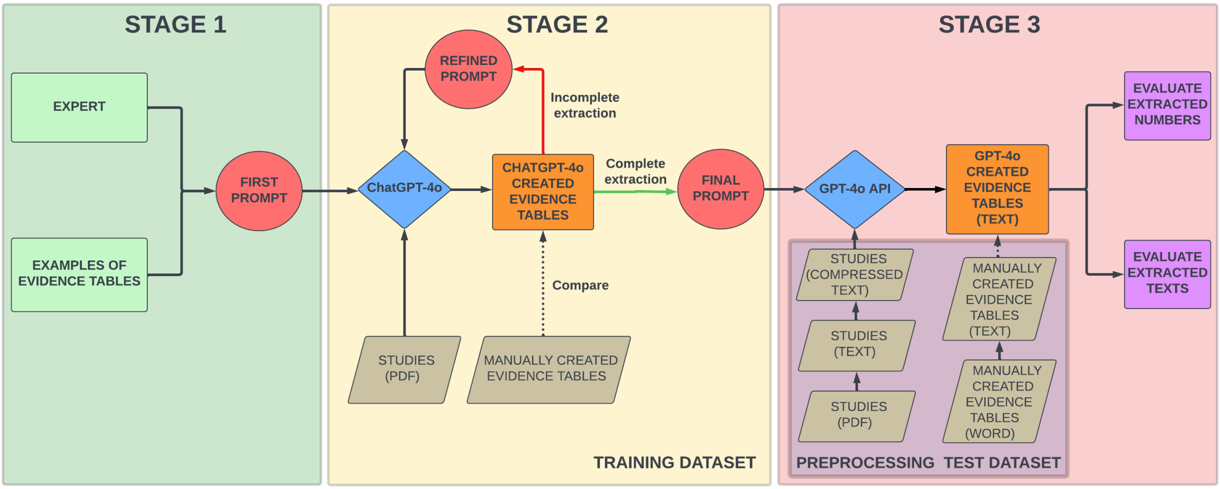
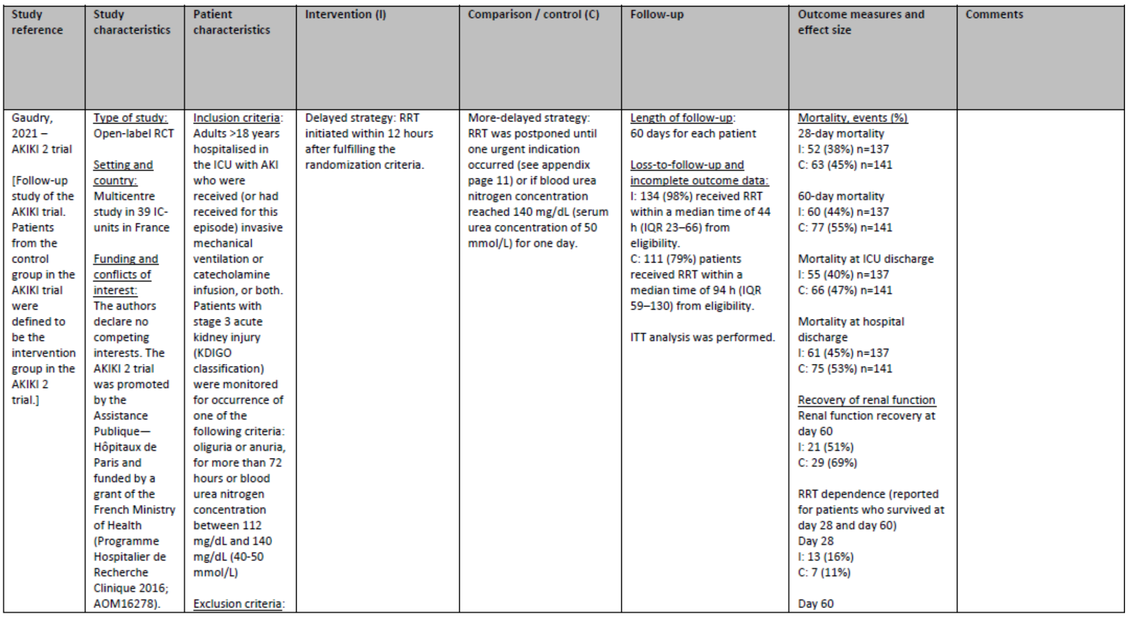

STAGE 1
EXPERT
EXAMPLES OF EVIDENCE TABLES
FIRST PROMPT
STAGE 2
REFINED PROMPT
Incomplete extraction
ChatGPT-4o
CHATGPT-4o CREATED EVIDENCE TABLES
Complete extraction
FINAL PROMPT
Compare
STUDIES (PDF)
MANUALLY CREATED EVIDENCE TABLES
TRAINING DATASET
STAGE 3
GPT-4o API
GPT-4o CREATED EVIDENCE TABLES (TEXT)
EVALUATE EXTRACTED NUMBERS
EVALUATE EXTRACTED TEXTS
STUDIES (COMPRESSED TEXT)
STUDIES (TEXT)
STUDIES (PDF)
MANUALLY CREATED EVIDENCE TABLES (TEXT)
MANUALLY CREATED EVIDENCE TABLES (WORD)
PREPROCESSING TEST DATASET

Figure 1: Study design

2.2.1 Stage 1

The evidence tables used, specific to the Knowledge Institute of the Federation of Medical Specialists (KI-FMN), always contained eight columns: “Study Reference”, “Study Characteristics”, “Patient Characteristics”, “Intervention”, “Comparison/Control”, “Follow-up”, “Outcome Measures and Effect Size” and “Comments.” Initially, we created a generic prompt aimed at asking the essential questions necessary to gather information for each column. The first version of this prompt (Appendix B) was designed through a conversation with an expert, Tim Christen, about the type of information he sought for each column when filling in an evidence table (Appendix A). Additionally, we reviewed existing tables to better identify the types of information they contained (an example of them can be seen in Figure 2).



Study reference	Study characteristics	Patient characteristics	Intervention (I)	Comparison / control (C)	Follow-up	Outcome measures and effect size	Comments
Gaudry, 2021 – AKIKI 2 trial [Follow-up study of the AKIKI trial. Patients from the control group in the AKIKI trial were defined to be the intervention group in the AKIKI 2 trial.]	Type of study: Open-label RCT Setting and country: Multicentre study in 39 IC-units in France Funding and conflicts of interest: The authors declare no competing interests. The AKIKI 2 trial was promoted by the Assistance Publique—Hôpitaux de Paris and funded by a grant of the French Ministry of Health (Programme Hospitalier de Recherche Clinique 2016; AOM16278).	Inclusion criteria: Adults >18 years hospitalised in the ICU with AKI who were received (or had received for this episode) invasive mechanical ventilation or catecholamine infusion, or both. Patients with stage 3 acute kidney injury (KDIGO classification) were monitored for occurrence of one of the following criteria: oliguria or anuria, for more than 72 hours or blood urea nitrogen concentration between 112 mg/dL and 140 mg/dL (40-50 mmol/L) Exclusion criteria:	Delayed strategy: RRT initiated within 12 hours after fulfilling the randomization criteria.	More-delayed strategy: RRT was postponed until one urgent indication occurred (see appendix page 11) or if blood urea nitrogen concentration reached 140 mg/dL (serum urea concentration of 50 mmol/L) for one day.	Length of follow-up: 60 days for each patient Loss-to-follow-up and incomplete outcome data: I: 134 (98%) received RRT within a median time of 44 h (IQR 23–66) from eligibility. C: 111 (79%) patients received RRT within a median time of 94 h (IQR 59–130) from eligibility. ITT analysis was performed.	Mortality, events (%) 28-day mortality I: 52 (38%) n=137 C: 63 (45%) n=141 60-day mortality I: 60 (44%) n=137 C: 77 (55%) n=141 Mortality at ICU discharge I: 55 (40%) n=137 C: 66 (47%) n=141 Mortality at hospital discharge I: 61 (45%) n=137 C: 75 (53%) n=141 Recovery of renal function Renal function recovery at day 60 I: 21 (51%) C: 29 (69%) RRT dependence (reported for patients who survived at day 28 and day 60) Day 28 I: 13 (16%) C: 7 (11%) Day 60	

Figure 2: Example of evidence table (not complete)

The prompt began by providing the language model with an explanation of the project. We specified the context, the format of the input and the desire format of the output when extracting the information. Following this, we tried to formulate clear and precise questions for each of the columns explaining exactly what we needed. We can see the beginning of this prompt here.

> *You are going to help healthcare experts develop medical guidelines. We are providing you with a PDF containing a Randomized Control Trial (RCT). The output must be structured into 8 clearly differentiated sections, labeled from one to eight with the following names and extracting the following information:*
>
> 1. *Study reference: Write the name of the first author of the paper and the year of publication separated by a comma.*

Figure 3: Beginning of the first version of the prompt

2.2.2 Stage 2

The next step involved using ChatGPT-4o to test and improve the first version of the prompt. This phase consisted of running the prompt through the chatbot interface to observe how the model extracted the required information from the papers in our training dataset. We opted not to use the API at this stage to avoid costs, as the chatbot interface was freely accessible once we had paid for the premium version. By examining the extracted information and qualitatively comparing it with the manually curated evidence tables, we iteratively refined the prompt. This process continued until we considered that the model's output contained all the information present in the manually curated evidence tables. A detailed explanation of this process and an example of it are provided in the 2.3 Prompt Optimization section. The final result of this iterative process can be found in Appendix C.

2.2.3 Stage 3

Finally, we transitioned to using the GPT-4o API. Although this required a payment per token used, it allowed us to write a script to automatically process all the papers in the test dataset and adjust the model's hyperparameters to increase reproducibility. Specifically, we set the temperature to 0 and the top_p parameter to 1 to minimize randomness in the model's responses.

At this stage, we evaluated the quality of the results by comparing the model's output with the manually curated evidence tables column by column. This process was divided into two parts: comparing the numbers and comparing the texts. A detailed explanation of this approach is provided in the 2.4 Analytic Strategy section.

2.3 Prompt optimization

To refine the prompt from its first version to the final one, we followed an iterative process that involved comparing the information extracted by the model with the information manually extracted by the experts.

First, we ran the initial version of the prompt through the model and examined the output. We compared this output with the manually curated evidence tables to identify any discrepancies, noting columns where the model's extraction was incomplete or erroneous. If the model consistently failed to extract certain types of information, we modified it to explicitly request that information, making targeted adjustments to improve its clarity and specificity.

For example, initially, the model wasn't extracting all the prognostic factors in the "Patient Characteristics" column. To address this, we modified this part of the prompt to specifically state that we wanted all the characteristics that each group had at the beginning of the experiment.

First version:

3.Patient characteristics: Explain the inclusion and exclusion criteria for participating in the study. Then, provide the total number of participants at the beginning of the experiment separated into the intervention and control groups. Additionally, include any other characteristic (e.g., age, sex, etc.) for each group, divided into the intervention and control groups. Finally, decide if the groups were comparable at the beginning of the experiment based on this information.

Final version:

3.Patient characteristics:

- *Explain the inclusion and exclusion criteria for participating in the study.*
- *Provide the total number of participants at the beginning of the experiment, separated into the intervention and control groups.*
- *List any other characteristics (e.g., age, sex, coexisting conditions, biological characteristics, etc.) you can find of the participants at the beginning of the study. Divide it into the intervention and control groups. Don't omit any numerical information even if it doesn't look relevant.*
- *Decide if the groups were comparable at the beginning of the experiment based on this information.*

Figure 4: Example of prompt optimization

After making these adjustments, we tested the revised prompt by running it through the model again and comparing the new output with the manually curated one. This iterative process of testing, identifying flaws and refining the prompt continued until the model consistently extracted all the information that the experts did, even if the model provided extra or unnecessary information. The final version of this prompt can be found in Appendix C.

2.4 Analytic strategy

To evaluate the results, we decided to use only six columns: “Study Characteristics”, “Patient Characteristics”, “Intervention”, “Comparison/Control”, ”Follow-up” and “Outcome Measures and Effect Size.” We excluded the ”Study Reference” column because it only included metadata. Additionally, the ”Comments” column was not used due to the significant variations in how it could be filled in, making systematic comparison very difficult. Our evaluation followed two approaches: comparing numbers and comparing texts. The scripts used can be found at https://github.com/Carlos-Vi/ADS-thesis/tree/main/Code.

2.4.1 Comparing numbers

For each column, we extracted all the numbers from the manually curated evidence tables and from the tables generated by the model. We compared the extracted numbers using the following criteria:

1. **True Positive (TP)**: If a number was present in both the table extracted by the model and the one created by the human, we added 1 point to the TP variable.
2. **False Negative (FN)**:If a number was present in the table created by the human but not in the one extracted by the model, we added 1 point to the FN variable.
3. **False Positive (FP)**: If a number was present in the table extracted by the model but not in the one created by the human, we added 1 point to the FP variable.

Using these values, we defined the precision and the recall for these columns with the following formulas:

$$\text{Precision} = \frac{TP}{TP + FP} \times 100; \qquad \text{Recall} = \frac{TP}{TP + FN} \times 100 \tag{2.1}$$

After calculating the precision and recall for each column of each individual study, we defined the total precision and recall for each column of our entire test dataset. To do this, we aggregated the counts of true positives (TP), false positives (FP) and false negatives (FN) from each study and calculated the global metrics

$$\text{Total precision} = \frac{\sum_i TP_i}{\sum_i (TP_i + FP_i)} \times 100; \qquad \text{Total recall} = \frac{\sum_i TP_i}{\sum_i (TP_i + FN_i)} \times 100 \tag{2.2}$$

where i indexes over all studies included in the test dataset.

Finally, we calculated the amount of numbers extracted by human experts and by the model for each study and each column. We then computed the average across all studies and its uncertainty (SEM) to establish a global metric.

2.4.2 Comparing texts

For each column of the evidence tables, we assessed the similarity between information extracted by human experts and that produced by the LLM using five distinct methods: TF-IDF, Jaccard, BERT, Sentence-BERT and the spaCy library. The Jaccard similarity measures the overlap between sets of words from each text, calculating the ratio of their intersection to their union. The other methods (TF-IDF, BERT, Sentence-BERT, and spaCy) each generate vector representations of the texts using different approaches. We then use cosine similarity to compare these vectors to determine the degree of similarity between the texts. Following this, we calculated the average similarity for each method across all studies and its uncertainties (SEM).

To summarize the results, we used the total recall to represent the number comparison and the Sentence-BERT similarity to represent the text comparison. The total recall of numbers extracted was chosen because the model consistently extracted more numbers than human annotators, suggesting that humans may not have captured all potentially useful numerical data for each column. The Sentence-BERT similarity was selected because it provided intermediate values among the five text comparison methods, offering what we thought was a balanced perspective on text similarity.

3 Results

After performing the analysis, we can find in Figure 5 the summary of the results.

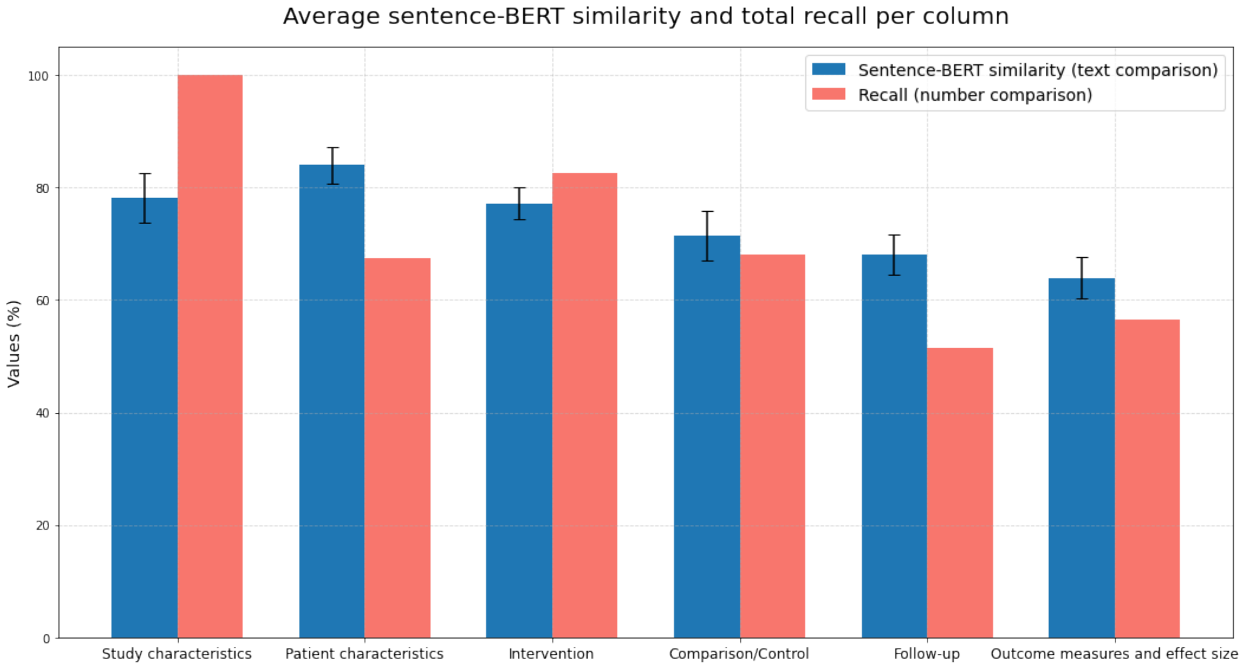


Figure 5: Average sentence-BERT similarity and total recall per column

The y-axis represents the total recall and the average Sentence-BERT similarity as percentages. The x-axis shows the names of the six evaluated columns, with bars corresponding to the results of the numeric (red) and text (blue) comparisons. No metric falls below 50%, with values going over 80% for text similarity in the ”Patient characteristics” column and for numbers total recall in the ”Study characteristics” and ”Intervention” columns. Except for the recall in the ”Follow-up” and ”Outcome measures and effect size” columns, all values are above 60%.

We examine now how these metrics varied per study and per comparison method.

3.1 Comparing numbers

This section focuses on the comparison of numeric data extracted from each column, disregarding the textual content. All the values used in this section can be found in Appendix D. Figure 6 illustrates the total precision and recall per column for the extracted numbers.

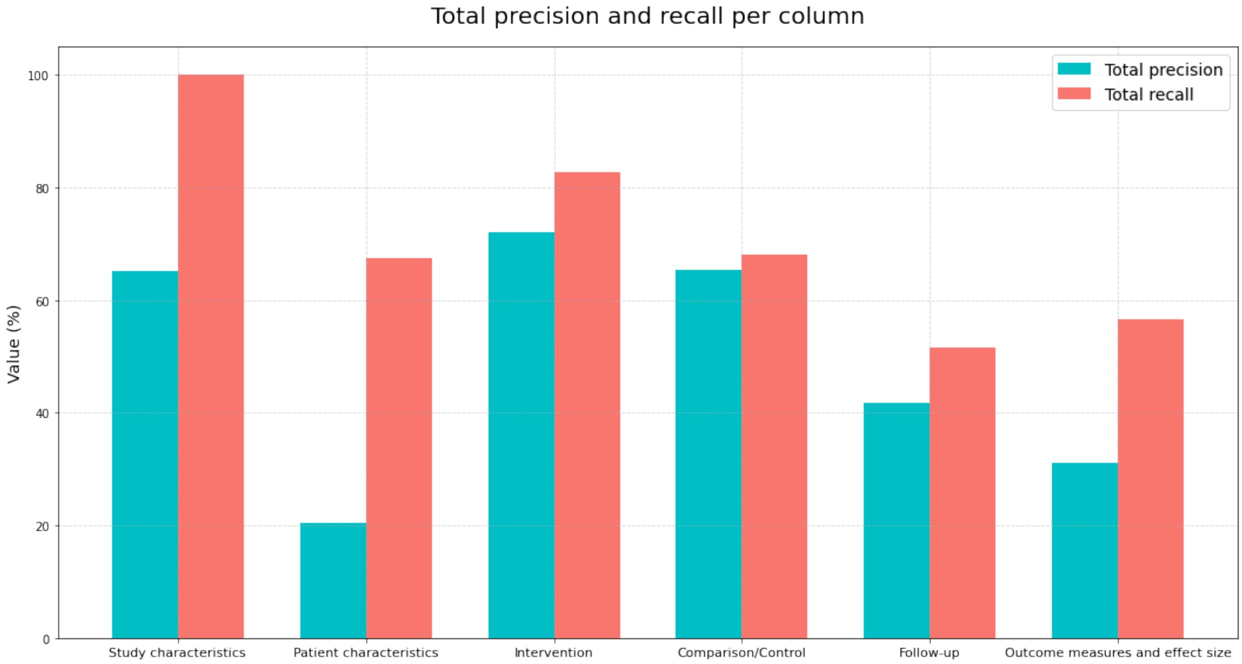
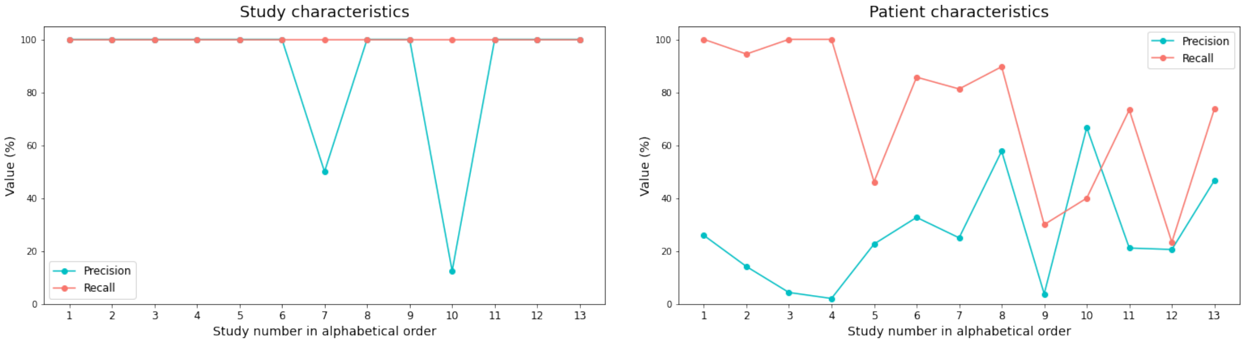


Figure 6: Total precision and recall per column

The results show that the recall is systematically higher than the precision. This outcome is expected, as we also saw that the model tended to extract more numbers than the humans. This tendency resulted in a higher count of false positives (FP) and a lower count of false negatives (FN), which reduced the precision and increased the recall. This higher recall indicates that the model was capturing a broader range of numerical data, even if it was including some irrelevant numbers.

To better understand the variability in precision and recall across individual studies, we present Figure 7, which provides a visualization of these metrics per study.



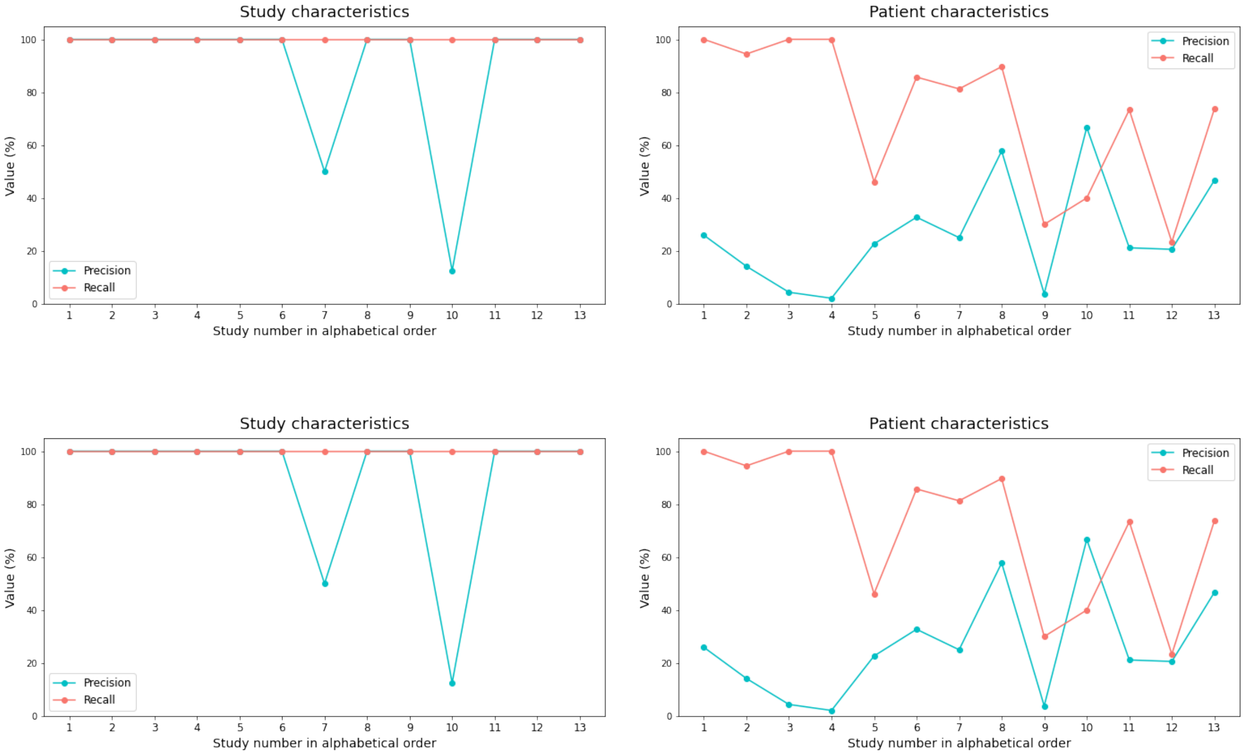


Figure 7: Precision and recall of extracted numbers per column and per study

This figure highlights the fluctuations in precision and recall among different studies, but showing the tendency of having higher recalls than precisions.

Next, we compare the total amount of numbers extracted by human experts and by the model. Figure 8 shows the average amount numbers extracted per column.

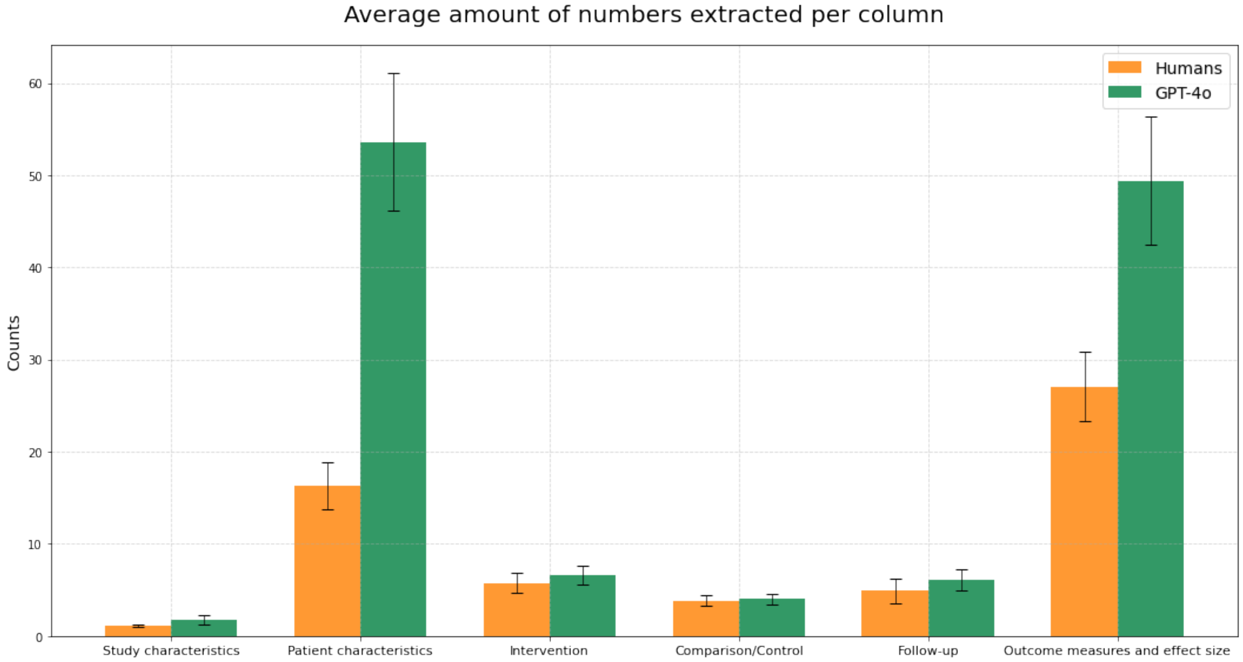
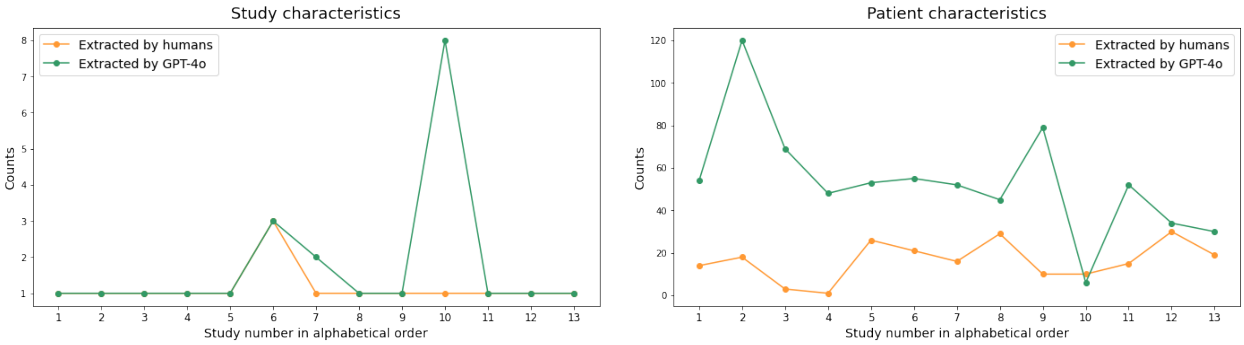


Figure 8: Average total amount of numbers extracted by humans and by GPT-4o per column

The LLM consistently extracted more numbers than human annotators. This observation is significant as it suggests the model's propensity to capture extensive numerical data, potentially including additional relevant details overlooked by human curators. The option that it is producing hallucinations cannot be disregarded yet, but manual comparison with the original studies don't point in this direction. The "Patient Characteristics" and "Outcome measures and effect size" columns are particularly noteworthy, containing significantly more numbers than the other columns. This was to be expected since these columns focus on numerical data related to patient characteristics and treatment outcomes, as can be seen in the human curated evidence tables. In contrast, the other columns, which contain fewer numerical values, are more narrative-driven and focus on qualitatively describing treatments or experimental conditions. Figure 9 details how these results vary across different studies.



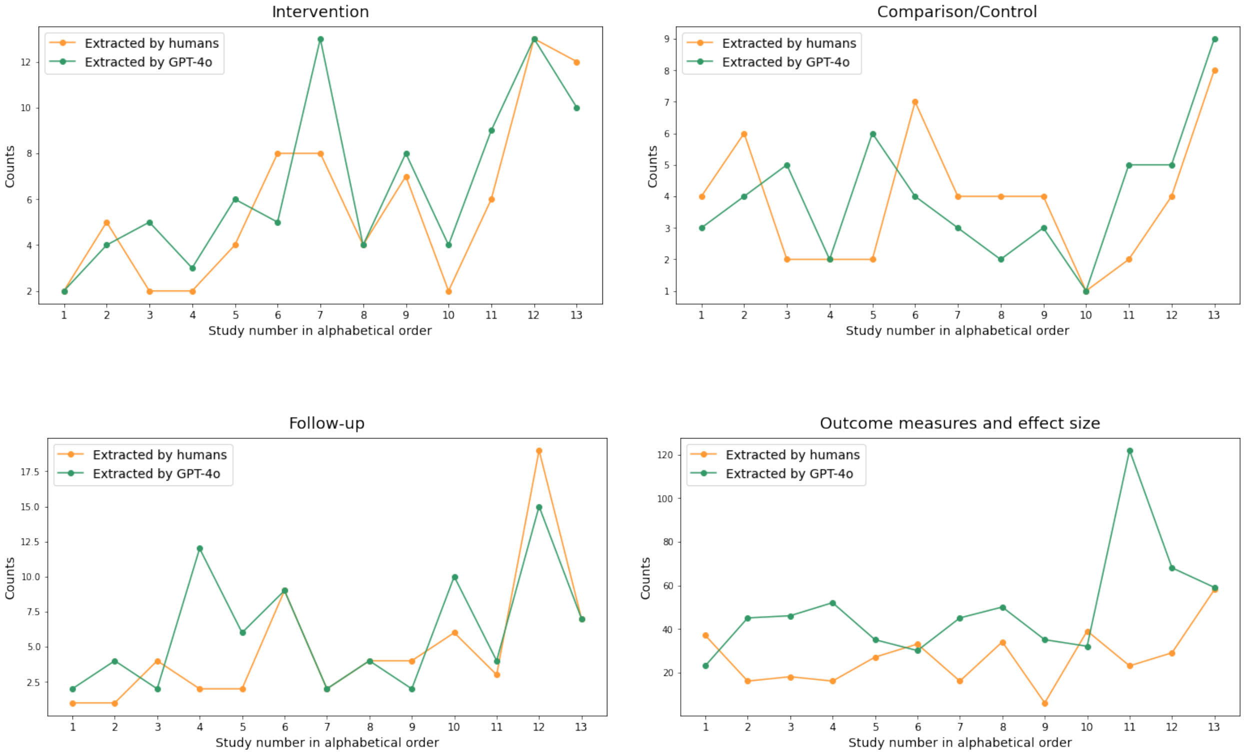


Figure 9: Amount of numbers extracted by humans and by GPT-4o per column and per study

In the ”Patient Characteristics” and ”Outcome measures and effect size” columns, which contained more numerical data, there was a greater consistency across studies in the higher amount of data extracted by GPT-4o with respect to humans compared to the other columns.

3.2 Comparing texts

This section compares the textual information extracted by human experts and by the LLM. All the values used in this section can be found in Appendix E. Figure 10 shows the average similarity per column using five different methods: TF-IDF, Jaccard, BERT, Sentence-BERT, and spaCy.

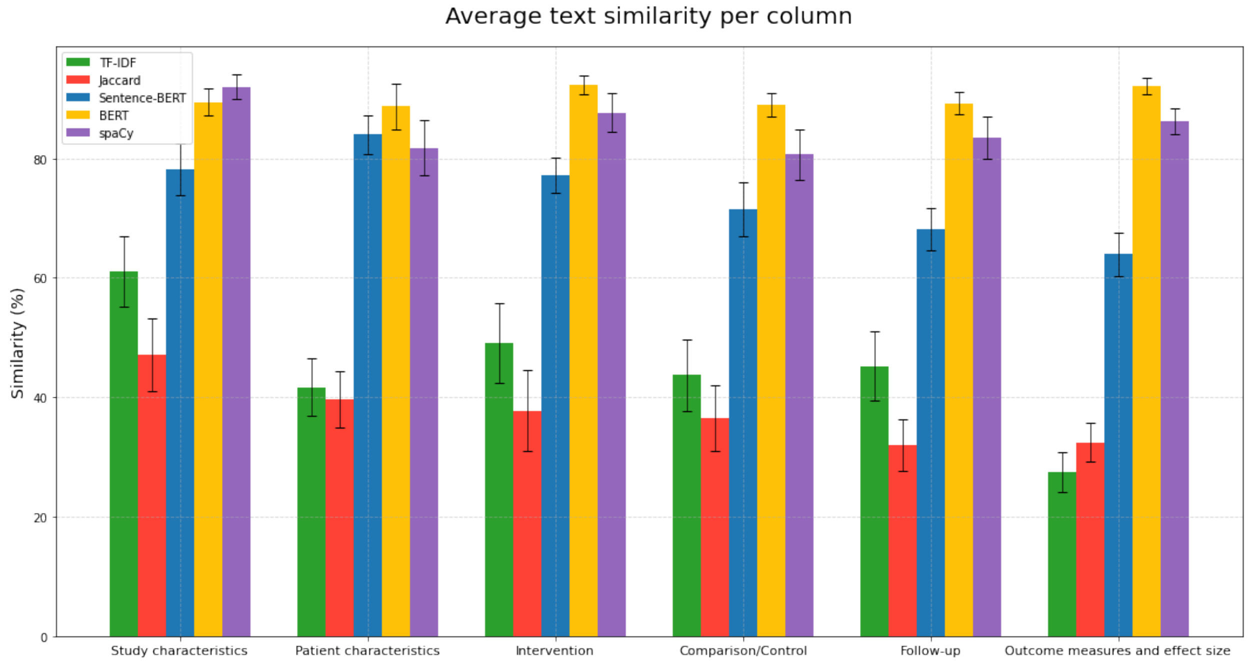
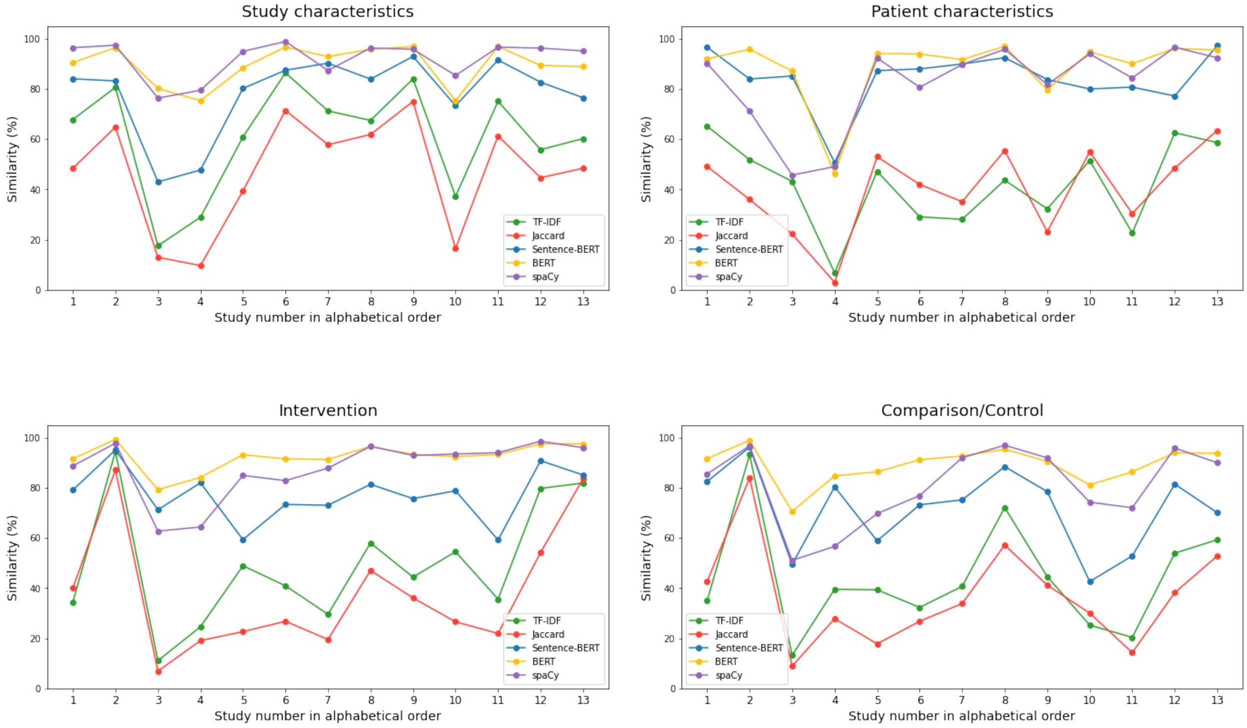


Figure 10: Average text similarity per column using 5 different methods

Each method provided different results due to the complexity of comparing information within texts. Jaccard similarity produced the lowest values, while BERT generated the highest ones, covering a broad range of results. These varied outcomes highlight the challenges in text comparison and the importance of using multiple methods for a comprehensive analysis. Figure 11 provides a more detailed view, showing text similarity per study.



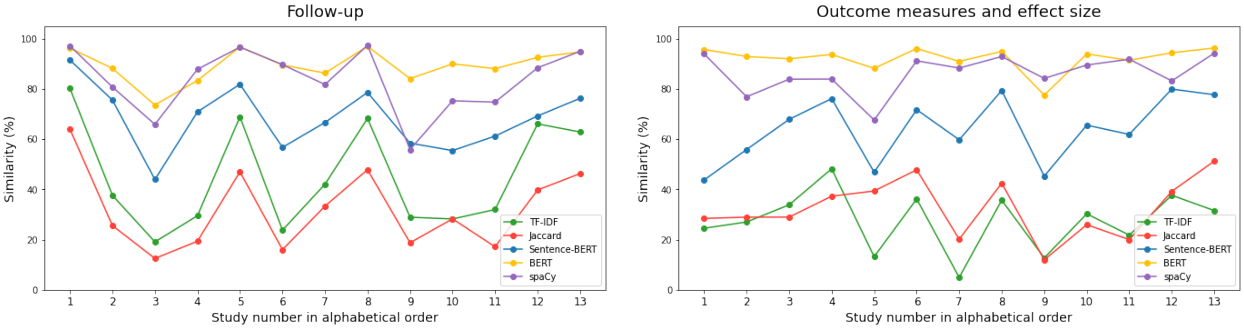


Figure 11: Text similarity per column and per study using 5 different methods

While each method yielded different results, they were consistent in the order of their similarity scores. This means that, for each given column and study, typically BERT is going to give the highest value, followed by spaCy, Sentence-BERT, TF-IDF, and Jaccard. Additionally, it is notable the significant variability in these scores among the different methods, ranging from 10% to 90% for the same texts in some cases.

For a manual comparison of the evidence tables generated by humans and by GPT-4o used in this section they can be found in text format at https://github.com/Carlos-Vi/ADS-thesis/tree/main/Evidence_tables_for_manually_comparison.

4 Discussion

Our proof-of-concept research highlights the significant potential of using LLMs to automate the process of filling in evidence tables from medical studies. We have demonstrated how GPT-4o can effectively extract both numerical data and qualitative results from scientific papers for the six categories explored: "Study Characteristics", "Patient Characteristics", "Intervention", "Comparison/Control", "Follow-up" and "Outcome Measures and Effect Size." This confirms that GPT-4o is capable of handling complex data extraction tasks traditionally performed by humans.

Our findings contribute to the evolving field of AI in healthcare, showcasing how LLMs like GPT-4o can be integrated into medical research workflows to enhance efficiency and accuracy. This study aligns with previous research, such as [Gartlehner et al., 2024], which demonstrated the promising potential of Claude 2 for semi-automating data extraction in systematic reviews. While Gartlehner et al. reported a high accuracy rate of 96.3% for Claude 2 in extracting specific data elements from Randomized Controlled Trials on plaque psoriasis, our study extends this by using GPT-4o, a more advanced LLM, across a broader range of medical topics, including asthma, COPD, acute kidney injury, bronchiolitis and cardiac rehabilitation. Additionally, our approach was more holistic, defining a generic prompt applicable to any medical study, aiming to extract all relevant information rather than just selected data points. Unlike Gartlehner et al., who manually compared 160 items for each study, we automated the evaluation process using scripts for numeric and text comparison metrics, which, while efficient, might compromise some level of accuracy.

Despite the promising results, our study has several limitations. Firstly, we focused exclusively on Randomized Controlled Trials (RCTs), excluding Observational Studies and Systematic Reviews, which have different characteristics and might require modified prompts. Secondly, the pre-processing of PDFs to make them compatible with the LLM is complex, potentially affecting the model's ability to accurately interpret the data due to formatting changes. Thirdly, due to API limitations on Tokens Per Minute (TPM), we had to compress the texts, which might have altered the data structure. Additionally, our evaluation approach had inherent constraints. While comparing numbers, we only assessed the overlap of sets without verifying the context of the numbers, not demonstrating whether the LLM's higher extraction rate was due to human oversight or model hallucinations. Furthermore, text comparison posed significant challenges due to the complexity of aligning semantic content across different texts, exacerbated by inherent inconsistencies in human annotations. We also noticed a systematic bias in the evaluation strategies since we included the names of the columns with their numeric positions (e.g., "2. Study characteristics," "3. Patient characteristics," etc.) in the comparisons. This meant that each column comparison had at least one number in common, the column number, and the same title, artificially boosting the similarity results. Lastly, the inherent stochasticity of LLMs, despite attempts to minimize it, remains a concern, as re-running the model might not yield identical results.

Future research should focus on refining evaluation methods to better gauge the accuracy and relevance of extracted information. Enhancing text comparison techniques will be crucial for accurately assessing semantic similarities between AI-extracted and human annotated data. Expanding the scope to include Observational Studies and Systematic Reviews will provide a more comprehensive evaluation. Additionally, improving pre-processing steps to maintain the integrity of tables and figures and exploring methods to handle API limitations, including higher budgets, without data loss are essential. Assessing the long-term stability of LLMs and their ability to produce consistent results over time will be vital for ensuring their reliability in evidence synthesis. Investigating the possibility of LLMs assessing their own accuracy could also offer valuable insights. Ultimately, while our final prompt for extracting evidence tables is likely improvable, we consider that developing an effective automatic evaluation strategy will be necessary for achieving significant advancements in this area.

In conclusion, while full automation of information extraction for evidence tables is not yet guaranteed, our study demonstrates that this approach holds considerable promise for the future. Continued refinement of these methods could significantly speed up the development of medical guidelines, reduce costs, save time and ensure that patient care is based on the most current and comprehensive data available. This advancement represents a meaningful step forward in leveraging AI to support and enhance the work of healthcare professionals.

Data/Scripts

All publicly available data and the Python scripts used can be found at https://github.com/Carlos-Vi/ADS-thesis. They were run using Python 3.9.6.

Generative AI statement

Generative AI, specifically ChatGPT-4o, has been used to correct grammatical errors in the text, improve its clarity, comment the code and assist with debugging and optimizing the Python scripts.

References

[Achiam et al., 2023] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). Gpt-4 technical report.

arXiv preprint arXiv:2303.08774.

[Alshami et al., 2023] Alshami, A., Elsayed, M., Ali, E., Eltoukhy, A. E. E., and Zayed, T. (2023). Harnessing the power of chatgpt for automating systematic review process: Methodology, case study, limitations, and future directions. *Systems*, 11(7).

[Barton, 2000] Barton, S. (2000). Which clinical studies provide the best evidence?: The best rct still trumps the best observational study.

[Bashir et al., 2018] Bashir, T., Reddy, K. V., Ahmed, K., and Shafi, S. (2018). Comparative study of 3% hypertonic saline nebulisation versus 0.9% normal saline nebulisation for treating acute bronchiolitis. *Journal of Clinical & Diagnostic Research*, 12(6).

[Bremner et al., 2018] Bremner, P. R., Birk, R., Brealey, N., Ismaila, A. S., Zhu, C.-Q., and Lipson, D. A. (2018). Single-inhaler fluticasone furoate/umeclidinium/vilanterol versus fluticasone furoate/vilanterol plus umeclidinium using two inhalers for chronic obstructive pulmonary disease: a randomized non-inferiority study. *Respiratory research*, 19:1–10.

[Dagdelen et al., 2024] Dagdelen, J., Dunn, A., Lee, S., Walker, N., Rosen, A. S., Ceder, G., Persson, K. A., and Jain, A. (2024). Structured information extraction from scientific text with large language models. *Nature Communications*, 15(1):1418.

[Devereaux and Yusuf, 2003] Devereaux, P. and Yusuf, S. (2003). The evolution of the randomized controlled trial and its role in evidence-based decision making. *Journal of internal medicine*, 254(2):105–113.

[Djamin et al., 2019] Djamin, R. S., Bafadhel, M., Uzun, S., Russell, R. E., Ermens, A. A., Kerstens, R., Aerts, J. G., Pavord, I. D., and van der Eerden, M. M. (2019). Blood eosinophil count and gold stage predict response to maintenance azithromycin treatment in copd patients with frequent exacerbations. *Respiratory Medicine*, 154:27–33.

[Djamin et al., 2020] Djamin, R. S., Talman, S., Schrauwen, E. J., von Wintersdorff, C. J., Wolffs, P. F., Savelkoul, P. H., Uzun, S., Kerstens, R., van der Eerden, M. M., and Kluytmans, J. A. (2020). Prevalence and abundance of selected genes conferring macrolide resistance genes in copd patients during maintenance treatment with azithromycin. *Antimicrobial Resistance & Infection Control*, 9:1–8.

[Faten et al., 2014] Faten, T., Sana, A., Imen, B. H., Samia, H., Ines, B., Bechir, Z., and Khadija, B. (2014). A randomized, controlled trial of nebulized 5% hypertonic saline and mixed 5% hypertonic saline with epinephrine in bronchiolitis. *La Tunisie medicale*, 92(11).

[Flores-González et al., 2015] Flores-González, J. C., Matamala-Morillo, M. A., Rodríguez-Campoy, P., Pérez-Guerrero, J. J., Serrano-Moyano, B., Comino-Vazquez, P., Palma-Zambrano, E., Bulo-Concellón, R., Santos-Sánchez, V., Lechuga-Sancho, A. M., et al. (2015). Epinephrine improves the efficacy of nebulized hypertonic saline in moderate bronchiolitis: a randomised clinical trial. *PLoS One*, 10(11):e0142847.

[Gartlehner et al., 2024] Gartlehner, G., Kahwati, L., Hilscher, R., Thomas, I., Kugley, S., Crotty, K., Viswanathan, M., Nussbaumer-Streit, B., Booth, G., Erskine, N., et al. (2024). Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*.

[Gaudry et al., 2021] Gaudry, S., Hajage, D., Martin-Lefevre, L., Lebbah, S., Louis, G., Moschietto, S., Titeca-Beauport, D., La Combe, B., Pons, B., de Prost, N., et al. (2021). Comparison of two delayed strategies for renal replacement therapy initiation for severe acute kidney injury (akiki 2): a multicentre, open-label, randomised, controlled trial. *The Lancet*, 397(10281):1293–1300.

[Geng et al., 2020] Geng, W., Batu, W., You, S., Tong, Z., and He, H. (2020). High-flow nasal cannula: a promising oxygen therapy for patients with severe bronchial asthma complicated with respiratory failure. *Canadian respiratory journal*, 2020(1):2301712.

[Ghasemi et al., 2022] Ghasemi, A., Mirmiran, P., Kashfi, K., and Bahadoran, Z. (2022). Scientific publishing in biomedicine: A brief history of scientific journals. *International Journal of Endocrinology and Metabolism*, 21(1):e131812.

[Grimshaw and Russell, 1993] Grimshaw, J. M. and Russell, I. T. (1993). Effect of clinical guidelines on medical practice: a systematic review of rigorous evaluations. *The Lancet*, 342(8883):1317–1322.

[Guyatt et al., 2011] Guyatt, G., Oxman, A. D., Akl, E. A., Kunz, R., Vist, G., Brozek, J., Norris, S., Falck-Ytter, Y., Glasziou, P., DeBeer, H., et al. (2011). Grade guidelines: 1. introduction—grade evidence profiles and summary of findings tables. *Journal of clinical epidemiology*, 64(4):383–394.

[Jaquet-Pilloud et al., 2020] Jaquet-Pilloud, R., Verga, M.-E., Russo, M., Gehri, M., and Pauchard, J.-Y. (2020). Nebulised hypertonic saline in moderate-to-severe bronchiolitis: a randomised clinical trial. *Archives of disease in childhood*, 105(3):236–240.

[Kebede et al., 2023] Kebede, M. M., Le Cornet, C., and Fortner, R. T. (2023). In-depth evaluation of machine learning methods for semi-automating article screening in a systematic review of mechanistic literature. *Research Synthesis Methods*, 14(2):156–172.

[Modaressi et al., 2012] Modaressi, M.-R., Asadian, A., Faghihinia, J., Arashpour, M., and Mousavinasab, F. (2012). Comparison of epinephrine to salbutamol in acute bronchiolitis. *Iranian journal of pediatrics*, 22(2):241.

[Nussbaumer-Streit et al., 2021] Nussbaumer-Streit, B., Ellen, M., Klerings, I., Sfetcu, R., Riva, N., Mahmić-Kaknjo, M., Poulentzas, G., Martinez, P., Baladia, E., Ziganshina, L., Marqués, M., Aguilar, L., Kassianos, A., Frampton, G., Silva, A., Affengruber, L., Spjker, R., Thomas, J., Berg, R., Kontogiani, M., Sousa, M., Kontogiorgis, C., and Gartlehner, G. (2021). Resource use during systematic review production varies widely: a scoping review. *Journal of Clinical Epidemiology*, 139:287–296.

[Polak and Morgan, 2024] Polak, M. P. and Morgan, D. (2024). Extracting accurate materials data from research papers with conversational language models and prompt engineering. *Nature Communications*, 15(1):1569.

[Raeisi et al., 2019] Raeisi, S., Fakharian, A., Ghorbani, F., Jamaati, H. R., and Mirenayat, M. S. (2019). Value and safety of high flow oxygenation in the treatment of inpatient asthma: a randomized, double-blind, pilot study. *Iranian Journal of Allergy, Asthma and Immunology*.

[Risom et al., 2020] Risom, S. S., Zwisler, A.-D., Sibilitz, K. L., Rasmussen, T. B., Taylor, R. S., Thygesen, L. C., Madsen, T. S., Svendsen, J. H., and Berg, S. K. (2020). Cardiac rehabilitation for patients treated for atrial fibrillation with ablation has long-term effects: 12-and 24-month follow-up results from the randomized copenheartrfa trial. *Archives of Physical Medicine and Rehabilitation*, 101(11):1877–1886.

[Ruangsomboon et al., 2021] Ruangsomboon, O., Limsuwat, C., Praphruetkit, N., Monsomboon, A., and Chakorn, T. (2021). Nasal high-flow oxygen versus conventional oxygen therapy for acute severe asthma patients: A pilot randomized controlled trial. *Academic Emergency Medicine*, 28(5):530–541.

[Skjerven et al., 2013] Skjerven, H. O., Hunderi, J. O. G., Brügmann-Pieper, S. K., Brun, A. C., Engen, H., Eskedal, L., Haavaldsen, M., Kvenshagen, B., Lunde, J., Rolfsjord, L. B., et al. (2013). Racemic adrenaline and inhalation strategies in acute bronchiolitis. *New England Journal of Medicine*, 368(24):2286–2293.

[Uysalol et al., 2017] Uysalol, M., Haşlak, F., Özünal, Z. G., Vehid, H., and Uzel, N. (2017). Rational drug use for acute bronchiolitis in emergency care. *The Turkish journal of pediatrics*, 59(2):155–161.

[Van De Schoot et al., 2021] Van De Schoot, R., De Bruin, J., Schram, R., Zahedi, P., De Boer, J., Weijdema, F., Kramer, B., Huijts, M., Hoogerwerf, M., Ferdinands, G., et al. (2021). An open source machine learning framework for efficient and transparent systematic reviews. *Nature machine intelligence*, 3(2):125–133.

[Wang et al., 2023] Wang, J., Shi, E., Yu, S., Wu, Z., Ma, C., Dai, H., Yang, Q., Kang, Y., Wu, J., Hu, H., et al. (2023). Prompt engineering for healthcare: Methodologies and applications. *arXiv preprint arXiv:2304.14670*.

[White et al., 2023] White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., and Schmidt, D. C. (2023). A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Appendix

A Desired information per column in evidence tables

1. Study reference
   - 1st author
   - Year of publication
2. Study characteristics
   - Type of study
   - Setting and country
   - Funding and conflicts of interest
3. Patient characteristics
   - Inclusion criteria
   - Exclusion criteria
   - Total number of participants in the intervention and control group at the beginning of the experiment

- Prognostic factors
- Are groups comparable at the baseline?

4. Intervention

   - Describe the treatment in the intervention

5. Comparison/Control

   - Describe the treatment in the control

6. Follow-up

   - Length of follow-up
   - Loss-to-follow-up
   - Incomplete outcome data

7. Outcome measures and effect size

   - Outcome measures

8. Comments

   - Confusing or contradictory information in the study

B First prompt

You are going to help healthcare experts develop medical guidelines. We are providing you with a PDF containing a Randomized Control Trial (RCT). The output must be structured into 8 clearly differentiated sections, labeled from one to eight with the following names and extracting the following information:

1. *Study reference: Write the name of the first author of the paper and the year of publication separated by a comma.*

2. *Study characteristics: Specify the type of study, its setting and country and its funding and conflicts of interest.*

3. *Patient characteristics: Explain the inclusion and exclusion criteria for participating in the study. Then, provide the total number of participants at the beginning of the experiment separated into the intervention and control groups. Additionally, include any other characteristic (e.g., age, sex, etc.) for each group, divided into the intervention and control groups. Finally, decide if the groups were comparable at the beginning of the experiment based on this information.*

4. *Intervention: Describe the intervention exactly as outlined in the article (treatment/procedure/test).*

5. *Comparison/Control: Describe the control exactly as outlined in the article (treatment/procedure/test).*

6. *Follow-up: State the length of the follow-up in days. Explain important events during the follow-up that affected the experiment, especially loss-to-follow-up (number, percentage and reason per study group) and incomplete outcome data (number, percentage and reason per study group) divided into the intervention and control groups.*

7. *Outcome measures and effect size: Provide all the results from the experiment divided into the intervention and control groups. These results can include mortality, recovery, time of illness, etc.*

8. *Comments: Add any information that might be contradictory or confusing in the study.*

Accuracy is crucial. Be as rigorous as possible when answering each section. If some information is missing, explicitly state that the information could not be found. Not finding something is okay. Quote directly from the study whenever possible. Provide not only the numerical values of the questions asked but also any other characteristics of this values (SD, IQR, etc.)

C Final prompt

You are going to help healthcare experts develop medical guidelines. We are providing you with a text containing a Randomized Controlled Trial (RCT) from a medical study. The output must be structured into 8 clearly differentiated sections, labeled from one to eight, with the following information:

1. *Study reference:*
   - *Write the name of the first author of the paper and the year of publication, separated by a comma.*

2. *Study characteristics:*
   - *Specify the "type of study" and its characteristics, the "setting and country" where it was conducted, and its "funding and conflicts of interest".*

3. *Patient characteristics:*
   - *Explain the inclusion and exclusion criteria for participating in the study.*
   - *Provide the total number of participants at the beginning of the experiment, separated into the intervention and control groups.*
   - *List any other characteristics (e.g., age, sex, coexisting conditions, biological characteristics, etc.) you can find of the participants at the beginning of the study. Divide it into the intervention and control groups. Don't omit any numerical information even if it doesn't look relevant.*
   - *Decide if the groups were comparable at the beginning of the experiment based on this information.*

4. *Intervention:*
   - *Describe the treatment and process followed by the intervention group exactly as outlined in the article. Write the answer directly without starting with phrases like "Intervention:"*

5. *Comparison/Control:*
   - *Describe the treatment and process followed by the control group exactly as outlined in the article. Write the answer directly without starting with phrases like "Control:"*

6. *Follow-up:*

- *State the length of the follow-up, meaning the time the patients were monitored after the intervention.*
- *State if there was "Loss-to-follow-up", including the number, percentage and reasons for participants who could not be tracked or reached for further data collection, divided into the intervention and control groups. This can happen because of the relocation of the participants, loss of contact, health problems, death or loss of interest for example.*
- *Explain if there was "Incomplete outcome data", including the number, percentage, and reasons for missing data, divided into the intervention and control groups. This can happen because of the loss of data due to administrative errors, incomplete measurements, not answering specific questions or dropping out of the study for example. Usually this information is not directly specified in the study, so you have to compare the number of participants at the beginning and at the end of the study.*

7. *Outcome measures and effect size:*

   - *Provide all the results from the experiment you can find, divided into the intervention and control groups. These results can include mortality, recovery, duration of the treatment, complications and any other information of the participants at the end or any stage of the experiment. Don't omit any numerical information present in the study even if it doesn't look relevant.*
   - *Include p-values and 95% Confidence Intervals if possible.*

8. *Comments:*

   - *Add any information that might be contradictory or confusing in the study.*

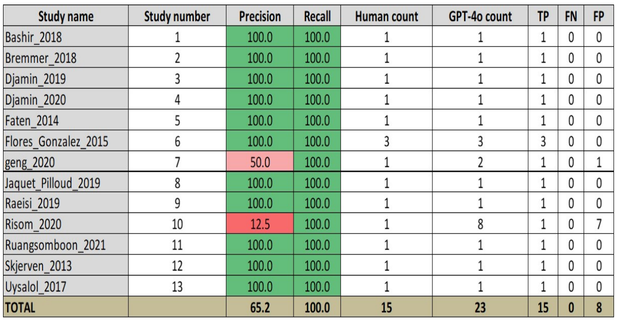
When providing the answers, do not use any special formatting such as italics or bold. To differentiate between sections, finish each one with the chain of characters "—". When listing the numerical values for the characteristics of the participants at the beginning or end of the study, use a separate line for each characteristic. Your answer will be saved in a plain text (.txt) file[1]*.*

IMPORTANT: Accuracy is crucial. Be as rigorous as possible when answering each section. Do not assume anything. If some information is missing, explicitly state that the information could not be found. Not finding something is acceptable. Quote directly from the study whenever possible. Provide not only the numerical values but also any other characteristics of these values (e.g., SD, IQR, etc.).

[1] This paragraph was added to facilitate the evaluation of the results.

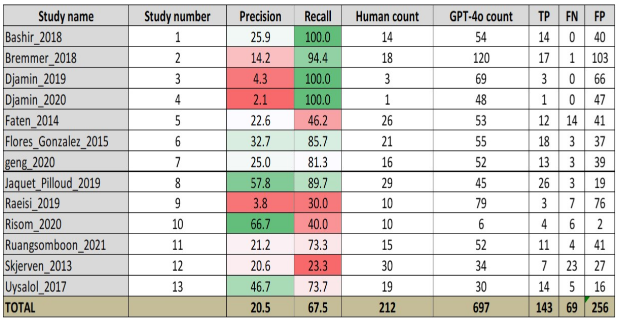
D Tables used in comparing numbers

In these tables, greener cells represent the highest values, while redder cells indicate the lowest within each column.



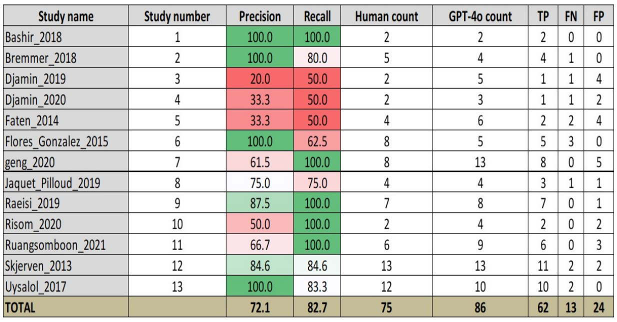
Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	100.0	100.0	1	1	1	0	0
Bremmer_2018	2	100.0	100.0	1	1	1	0	0
Djamin_2019	3	100.0	100.0	1	1	1	0	0
Djamin_2020	4	100.0	100.0	1	1	1	0	0
Faten_2014	5	100.0	100.0	1	1	1	0	0
Flores_Gonzalez_2015	6	100.0	100.0	3	3	3	0	0
geng_2020	7	50.0	100.0	1	2	1	0	1
Jaquet_Pilloud_2019	8	100.0	100.0	1	1	1	0	0
Raeisi_2019	9	100.0	100.0	1	1	1	0	0
Risom_2020	10	12.5	100.0	1	8	1	0	7
Ruangsomboon_2021	11	100.0	100.0	1	1	1	0	0
Skjerven_2013	12	100.0	100.0	1	1	1	0	0
Uysalol_2017	13	100.0	100.0	1	1	1	0	0
TOTAL		**65.2**	**100.0**	**15**	**23**	**15**	**0**	**8**

Table 1: Study characteristics



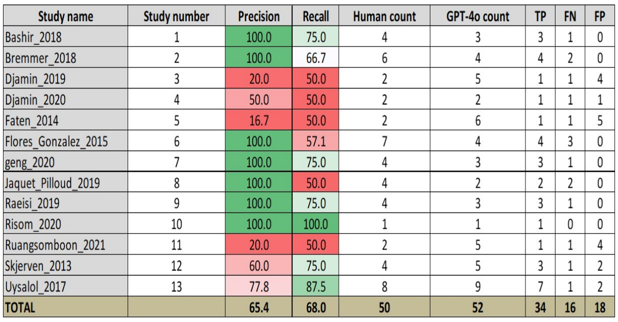
Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	25.9	100.0	14	54	14	0	40
Bremmer_2018	2	14.2	94.4	18	120	17	1	103
Djamin_2019	3	4.3	100.0	3	69	3	0	66
Djamin_2020	4	2.1	100.0	1	48	1	0	47
Faten_2014	5	22.6	46.2	26	53	12	14	41
Flores_Gonzalez_2015	6	32.7	85.7	21	55	18	3	37
geng_2020	7	25.0	81.3	16	52	13	3	39
Jaquet_Pilloud_2019	8	57.8	89.7	29	45	26	3	19
Raeisi_2019	9	3.8	30.0	10	79	3	7	76
Risom_2020	10	66.7	40.0	10	6	4	6	2
Ruangsomboon_2021	11	21.2	73.3	15	52	11	4	41
Skjerven_2013	12	20.6	23.3	30	34	7	23	27
Uysalol_2017	13	46.7	73.7	19	30	14	5	16
TOTAL		**20.5**	**67.5**	**212**	**697**	**143**	**69**	**256**

Table 2: Patient characteristics



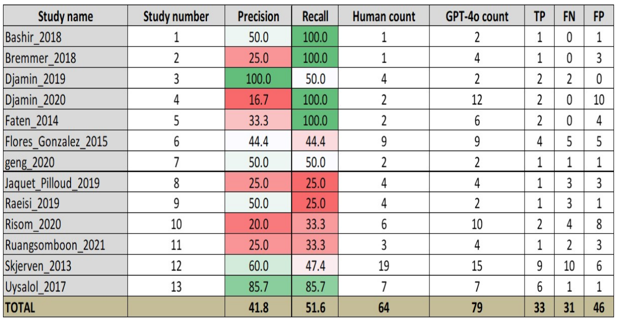
Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	100.0	100.0	2	2	2	0	0
Bremmer_2018	2	100.0	80.0	5	4	4	1	0
Djamin_2019	3	20.0	50.0	2	5	1	1	4
Djamin_2020	4	33.3	50.0	2	3	1	1	2
Faten_2014	5	33.3	50.0	4	6	2	2	4
Flores_Gonzalez_2015	6	100.0	62.5	8	5	5	3	0
geng_2020	7	61.5	100.0	8	13	8	0	5
Jaquet_Pilloud_2019	8	75.0	75.0	4	4	3	1	1
Raeisi_2019	9	87.5	100.0	7	8	7	0	1
Risom_2020	10	50.0	100.0	2	4	2	0	2
Ruangsomboon_2021	11	66.7	100.0	6	9	6	0	3
Skjerven_2013	12	84.6	84.6	13	13	11	2	2
Uysalol_2017	13	100.0	83.3	12	10	10	2	0
TOTAL		**72.1**	**82.7**	**75**	**86**	**62**	**13**	**24**

Table 3: Intervention



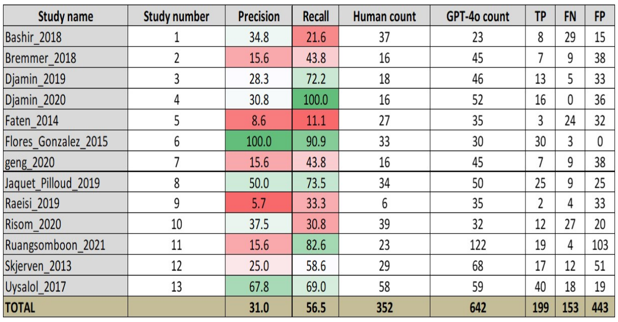
Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	100.0	75.0	4	3	3	1	0
Bremmer_2018	2	100.0	66.7	6	4	4	2	0
Djamin_2019	3	20.0	50.0	2	5	1	1	4
Djamin_2020	4	50.0	50.0	2	2	1	1	1
Faten_2014	5	16.7	50.0	2	6	1	1	5
Flores_Gonzalez_2015	6	100.0	57.1	7	4	4	3	0
geng_2020	7	100.0	75.0	4	3	3	1	0
Jaquet_Pilloud_2019	8	100.0	50.0	4	2	2	2	0
Raeisi_2019	9	100.0	75.0	4	3	3	1	0
Risom_2020	10	100.0	100.0	1	1	1	0	0
Ruangsomboon_2021	11	20.0	50.0	2	5	1	1	4
Skjerven_2013	12	60.0	75.0	4	5	3	1	2
Uysalol_2017	13	77.8	87.5	8	9	7	1	2
TOTAL		**65.4**	**68.0**	**50**	**52**	**34**	**16**	**18**

Table 4: Control/Comparison



Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	50.0	100.0	1	2	1	0	1
Bremmer_2018	2	25.0	100.0	1	4	1	0	3
Djamin_2019	3	100.0	50.0	4	2	2	2	0
Djamin_2020	4	16.7	100.0	2	12	2	0	10
Faten_2014	5	33.3	100.0	2	6	2	0	4
Flores_Gonzalez_2015	6	44.4	44.4	9	9	4	5	5
geng_2020	7	50.0	50.0	2	2	1	1	1
Jaquet_Pilloud_2019	8	25.0	25.0	4	4	1	3	3
Raeisi_2019	9	50.0	25.0	4	2	1	3	1
Risom_2020	10	20.0	33.3	6	10	2	4	8
Ruangsomboon_2021	11	25.0	33.3	3	4	1	2	3
Skjerven_2013	12	60.0	47.4	19	15	9	10	6
Uysalol_2017	13	85.7	85.7	7	7	6	1	1
TOTAL		**41.8**	**51.6**	**64**	**79**	**33**	**31**	**46**

Table 5: Follow-up

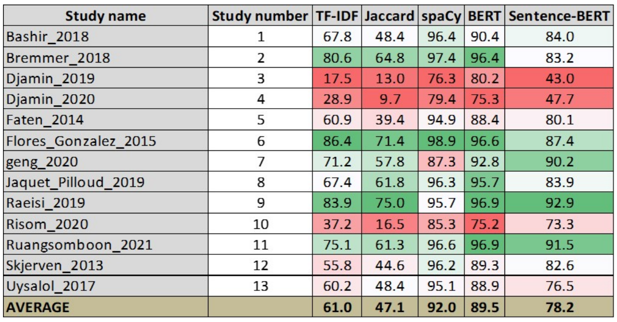


Study name	Study number	Precision	Recall	Human count	GPT-4o count	TP	FN	FP
Bashir_2018	1	34.8	21.6	37	23	8	29	15
Bremmer_2018	2	15.6	43.8	16	45	7	9	38
Djamin_2019	3	28.3	72.2	18	46	13	5	33
Djamin_2020	4	30.8	100.0	16	52	16	0	36
Faten_2014	5	8.6	11.1	27	35	3	24	32
Flores_Gonzalez_2015	6	100.0	90.9	33	30	30	3	0
geng_2020	7	15.6	43.8	16	45	7	9	38
Jaquet_Pilloud_2019	8	50.0	73.5	34	50	25	9	25
Raeisi_2019	9	5.7	33.3	6	35	2	4	33
Risom_2020	10	37.5	30.8	39	32	12	27	20
Ruangsomboon_2021	11	15.6	82.6	23	122	19	4	103
Skjerven_2013	12	25.0	58.6	29	68	17	12	51
Uysalol_2017	13	67.8	69.0	58	59	40	18	19
TOTAL		**31.0**	**56.5**	**352**	**642**	**199**	**153**	**443**

Table 6: Outcome measures and effect size

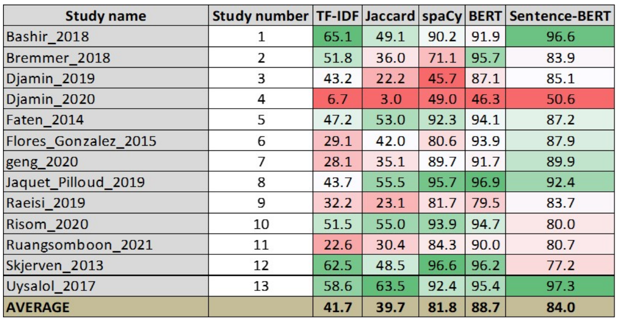
E Tables used in comparing texts

In these tables, greener cells represent the highest values, while redder cells indicate the lowest within each column.



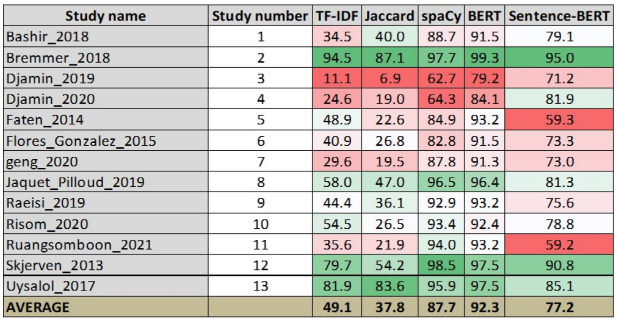
Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	67.8	48.4	96.4	90.4	84.0
Bremmer_2018	2	80.6	64.8	97.4	96.4	83.2
Djamin_2019	3	17.5	13.0	76.3	80.2	43.0
Djamin_2020	4	28.9	9.7	79.4	75.3	47.7
Faten_2014	5	60.9	39.4	94.9	88.4	80.1
Flores_Gonzalez_2015	6	86.4	71.4	98.9	96.6	87.4
geng_2020	7	71.2	57.8	87.3	92.8	90.2
Jaquet_Pilloud_2019	8	67.4	61.8	96.3	95.7	83.9
Raeisi_2019	9	83.9	75.0	95.7	96.9	92.9
Risom_2020	10	37.2	16.5	85.3	75.2	73.3
Ruangsomboon_2021	11	75.1	61.3	96.6	96.9	91.5
Skjerven_2013	12	55.8	44.6	96.2	89.3	82.6
Uysalol_2017	13	60.2	48.4	95.1	88.9	76.5
AVERAGE		**61.0**	**47.1**	**92.0**	**89.5**	**78.2**

Table 7: Study characteristics



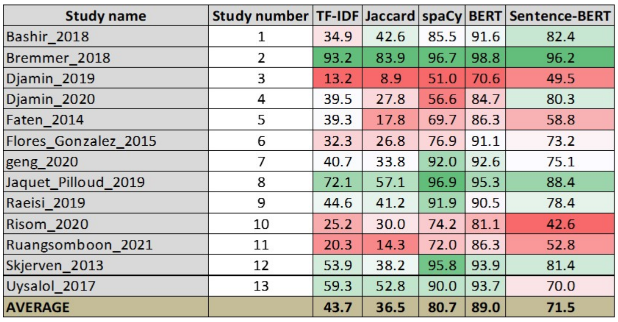
Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	65.1	49.1	90.2	91.9	96.6
Bremmer_2018	2	51.8	36.0	71.1	95.7	83.9
Djamin_2019	3	43.2	22.2	45.7	87.1	85.1
Djamin_2020	4	6.7	3.0	49.0	46.3	50.6
Faten_2014	5	47.2	53.0	92.3	94.1	87.2
Flores_Gonzalez_2015	6	29.1	42.0	80.6	93.9	87.9
geng_2020	7	28.1	35.1	89.7	91.7	89.9
Jaquet_Pilloud_2019	8	43.7	55.5	95.7	96.9	92.4
Raeisi_2019	9	32.2	23.1	81.7	79.5	83.7
Risom_2020	10	51.5	55.0	93.9	94.7	80.0
Ruangsomboon_2021	11	22.6	30.4	84.3	90.0	80.7
Skjerven_2013	12	62.5	48.5	96.6	96.2	77.2
Uysalol_2017	13	58.6	63.5	92.4	95.4	97.3
AVERAGE		**41.7**	**39.7**	**81.8**	**88.7**	**84.0**

Table 8: Patient characteristics



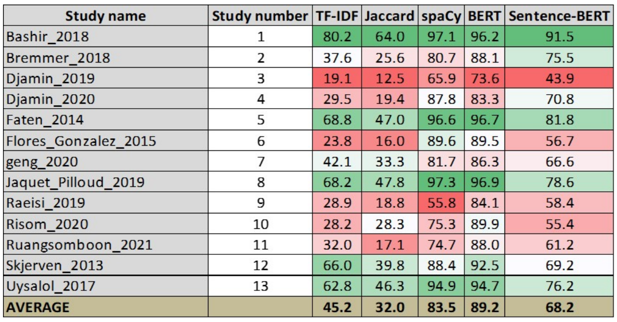
Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	34.5	40.0	88.7	91.5	79.1
Bremmer_2018	2	94.5	87.1	97.7	99.3	95.0
Djamin_2019	3	11.1	6.9	62.7	79.2	71.2
Djamin_2020	4	24.6	19.0	64.3	84.1	81.9
Faten_2014	5	48.9	22.6	84.9	93.2	59.3
Flores_Gonzalez_2015	6	40.9	26.8	82.8	91.5	73.3
geng_2020	7	29.6	19.5	87.8	91.3	73.0
Jaquet_Pilloud_2019	8	58.0	47.0	96.5	96.4	81.3
Raeisi_2019	9	44.4	36.1	92.9	93.2	75.6
Risom_2020	10	54.5	26.5	93.4	92.4	78.8
Ruangsomboon_2021	11	35.6	21.9	94.0	93.2	59.2
Skjerven_2013	12	79.7	54.2	98.5	97.5	90.8
Uysalol_2017	13	81.9	83.6	95.9	97.5	85.1
AVERAGE		**49.1**	**37.8**	**87.7**	**92.3**	**77.2**

Table 9: Intervention



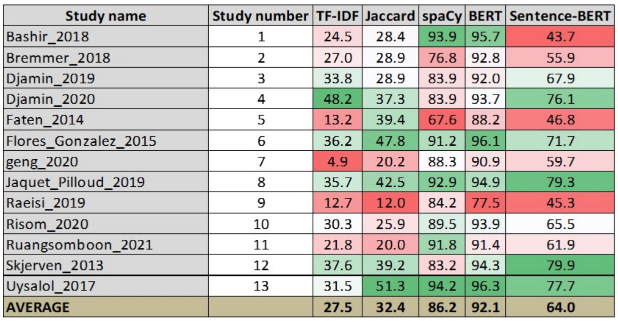
Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	34.9	42.6	85.5	91.6	82.4
Bremmer_2018	2	93.2	83.9	96.7	98.8	96.2
Djamin_2019	3	13.2	8.9	51.0	70.6	49.5
Djamin_2020	4	39.5	27.8	56.6	84.7	80.3
Faten_2014	5	39.3	17.8	69.7	86.3	58.8
Flores_Gonzalez_2015	6	32.3	26.8	76.9	91.1	73.2
geng_2020	7	40.7	33.8	92.0	92.6	75.1
Jaquet_Pilloud_2019	8	72.1	57.1	96.9	95.3	88.4
Raeisi_2019	9	44.6	41.2	91.9	90.5	78.4
Risom_2020	10	25.2	30.0	74.2	81.1	42.6
Ruangsomboon_2021	11	20.3	14.3	72.0	86.3	52.8
Skjerven_2013	12	53.9	38.2	95.8	93.9	81.4
Uysalol_2017	13	59.3	52.8	90.0	93.7	70.0
AVERAGE		**43.7**	**36.5**	**80.7**	**89.0**	**71.5**

Table 10: Control/Comparison



Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	80.2	64.0	97.1	96.2	91.5
Bremmer_2018	2	37.6	25.6	80.7	88.1	75.5
Djamin_2019	3	19.1	12.5	65.9	73.6	43.9
Djamin_2020	4	29.5	19.4	87.8	83.3	70.8
Faten_2014	5	68.8	47.0	96.6	96.7	81.8
Flores_Gonzalez_2015	6	23.8	16.0	89.6	89.5	56.7
geng_2020	7	42.1	33.3	81.7	86.3	66.6
Jaquet_Pilloud_2019	8	68.2	47.8	97.3	96.9	78.6
Raeisi_2019	9	28.9	18.8	55.8	84.1	58.4
Risom_2020	10	28.2	28.3	75.3	89.9	55.4
Ruangsomboon_2021	11	32.0	17.1	74.7	88.0	61.2
Skjerven_2013	12	66.0	39.8	88.4	92.5	69.2
Uysalol_2017	13	62.8	46.3	94.9	94.7	76.2
AVERAGE		**45.2**	**32.0**	**83.5**	**89.2**	**68.2**

Table 11: Follow-up



Study name	Study number	TF-IDF	Jaccard	spaCy	BERT	Sentence-BERT
Bashir_2018	1	24.5	28.4	93.9	95.7	43.7
Bremmer_2018	2	27.0	28.9	76.8	92.8	55.9
Djamin_2019	3	33.8	28.9	83.9	92.0	67.9
Djamin_2020	4	48.2	37.3	83.9	93.7	76.1
Faten_2014	5	13.2	39.4	67.6	88.2	46.8
Flores_Gonzalez_2015	6	36.2	47.8	91.2	96.1	71.7
geng_2020	7	4.9	20.2	88.3	90.9	59.7
Jaquet_Pilloud_2019	8	35.7	42.5	92.9	94.9	79.3
Raeisi_2019	9	12.7	12.0	84.2	77.5	45.3
Risom_2020	10	30.3	25.9	89.5	93.9	65.5
Ruangsomboon_2021	11	21.8	20.0	91.8	91.4	61.9
Skjerven_2013	12	37.6	39.2	83.2	94.3	79.9
Uysalol_2017	13	31.5	51.3	94.2	96.3	77.7
AVERAGE		**27.5**	**32.4**	**86.2**	**92.1**	**64.0**

Table 12: Outcome measures and effect size