Utrecht University

The application of LLM prompt engineering to optimize title screening

MSc Applied Data Science

Thesis Project

Student name: Giulia Migliore

Student number: 2593122

First examiner: Rens van de Schoot

Second examiner: Beth Grandfield

2023-2024

Table of Contents

Abstract

Screening papers for systematic reviews is a resource-intensive and time-consuming process. This study aims to reduce the necessary resources by automating the title screening phase using Large Language Models (LLMs). Initially, prompt engineering is employed to identify the optimal prompt for the LLM. Subsequently, the performance of the LLM is evaluated against simpler machine learning models to determine its effectiveness in excluding irrelevant papers without false exclusions. The findings indicate that the Large Language Model outperforms simpler machine learning models in the title screening phase, accurately excluding 60% of the papers with only one false exclusion. These promising results suggest that LLMs can assist researchers in the title screening phase, significantly reducing time and costs while maintaining high screening quality.

1 Introduction

Conducting a systematic review is an extensive process that requires researchers to meticulously screen thousands of papers to identify those relevant to their review question, ensuring that no potentially relevant studies are overlooked (Lefebvre et al., 2008; Sampson et al., 2011; Z. Wang et al., 2020). After deciding on the review question, researchers must systematically search for all relevant literature. They then proceed through two main phases: title-abstract screening, where irrelevant papers are excluded, and a full-text review of the remaining papers. These two steps ultimately narrow down the selection to the most pertinent studies for the review (Calderon Martinez et al., 2023).

Mateen et al. (2013) demonstrate that a titles-first screening strategy can be highly effective for systematic reviews. In their pilot study the titles-first approach discarded 86% of irrelevant citations early, reducing the need to read their abstracts. Nonetheless, the final list of papers remained the same between both methods, with the titles-first strategy achieving 100% recall, indicating perfect performance in identifying every relevant citation. This would change the process from two phases to three phases, involving an initial screening of titles, followed by a screening of the abstracts, and finally an examination of the full text.

The inclusion and exclusion criteria for titles are quite straightforward. For example, according to Meline (2006), these criteria generally belong to one or more of the following categories: study population, type of intervention, outcome variables, time frame, cultural and linguistic scope, and methodological quality. These clear-cut criteria suggest that the initial title screening phase can be handled by an automated system. Automating the title screening phase would decrease the number of papers requiring further review through their abstracts, therefore reducing time and costs of the overall screening process.

Large Language Models (LLMs) could be the ideal automated assistants for screening titles. LLMs are computational models capable of decoding natural language and solving tasks like text generation, question answering, machine translation, summarization, and many more. These models are trained on a large amount of text data, whereupon the weights of the billions of parameters were determined (Zhao et al., 2023).

Guo et al. (2024) explored the automation of title-abstract screening using LLMs. They evaluated the performance of GPT on 24,000 papers by comparing its labels to those given by human reviewers, achieving an accuracy of 0.91, with a recall of 0.91 for excluded papers and of 0.76 for included papers.

Accuracy is calculated as the ratio of the number of correct predictions to the total number of predictions made (B. Liu & Udell, 2020). An accuracy of 0.91 reveals that the model correctly predicted 91% of the total papers, indicating a high level performance.

Recall measures the proportion of actual positives that are correctly identified by the model (Buckland & Gey, 1994). A recall of 0.91 for excluded papers and of 0.76 for included papers highlights that the model correctly identified 91% of the excluded papers and 76% of the included papers.

Syriani et al. (2023) found that LLMs perform comparably to traditional machine learning methods used for automating systematic review activities, with the advantage that LLMs achieve this without requiring additional training. In their study, they compare GPT 3.5 Turbo with Linear Regression, Random Forest, Complement Naive Bayes, C-Support Vector Classifiers, and a Random Classifier.

Moreover, they emphasize that an effective screening classifier should aim to minimize the omission of relevant articles, thereby maximizing recall, and enhance reviewer efficiency by eliminating as many irrelevant articles as possible, thus maximizing the Negative Predictive Value (NPV). NPV measures the classifier's ability to exclude only the articles that should be excluded, similarly to precision but for negative outcomes.

Similarly Huotala et al. (2024) conducted an experiment with a smaller sample of 20 papers, using GPT-3.5 and GPT-4 to label the papers by their titles and abstracts through various prompting techniques like Zero-shot, Few-shot, and Chain-of-Thought. Their results showed that specific prompt combinations allowed LLMs to match human screeners' performance, with newer models outperforming older ones.

Prompt engineering involves strategically guiding a generative AI model to elicit specific answers and behaviors. Since different prompts produce varying outputs, refining the prompt quality can significantly improve the model's responses, making them more tailored to the specific question asked (P. Liu et al., 2023).

According to Chen et al. (2023), there are several straightforward adjustments that can significantly improve a prompt's quality.

Comprehensive and specific instructions. If the instructions are overly simplistic, the model has too many options available and the response will lack precision (Yang et al., 2019). The same applies to vague instructions, which will elicit more general outputs (Chen et al., 2023).

Role-prompting. Assigning the model with a specific role to interpret will lead to more accurate responses (Zhang, Z. et al., 2023).

Zero-shot, **one-shot**, **few-shots**. Zero-shot prompting refers to providing no examples in the input prompt. In contrast, one-shot prompting allows the model to learn from a single example included in the input prompt (Chen et al., 2023). Few-shot prompting, as described by Logan IV et al. (2021), involves supplying the model with several examples from which to learn. The choice between one-shot and few-shot prompting usually depends on how complex the task is and how capable the model is.

Temperature and Top-p. Temperature and top-p are LLMs' parameters that determine the randomness of the output. Lower temperatures result in less diverse, thus more deterministic outputs (Tunstall et al., 2022), while top-p selects only the tokens whose combined probabilities reach the selected value of top-p (Holtzman et al., 2020). For instance, setting top-p to 0.5 implies that the model will select the tokens with the highest probabilities until their total probability equals 50%.

However, for complex tasks, simple methods are inadequate and there is the necessity to use advanced prompting techniques.

Chain of Thought (**CoT**). Chain of thought prompting facilitates the model to perform complex reasoning by incorporating intermediate reasoning steps in the input prompt. It shows efficacy on logical and reasoning tasks (Lewkowycz et al., 2022; Wu et al., 2023; Z. Zhang et al., 2022).

Tree of Thoughts (**ToT**). This method enables the hierarchical organization of prompts (Yao, S. et al., 2024). This approach makes it simpler for the model to determine which part of the prompt to address first or which task takes precedence over the others.

While some work has been done to use LLMs for automating the screening process in systematic reviews, no prior research has specifically targeted the automation of the title screening phase alone. By providing the LLMs with inclusion and exclusion criteria and then inputting the paper titles, LLMs could automatically exclude clearly irrelevant papers, resulting in fewer papers needing abstract screening.

This paper has two primary objectives: first, to compare various prompt techniques and identify the one that elicits the most effective results from the LLM; second, to evaluate the performance of the LLM against simpler machine learning models to determine if the LLM more effectively filters out irrelevant papers.
The tested models should exclude a significant portion of papers, while ensuring that no relevant paper is excluded. False inclusions are acceptable at this point, as this is merely the initial stage. As mentioned earlier, all papers identified as relevant will undergo a secondary review that includes reading their abstracts (Mateen et al., 2013).
If this method proves effective, it can serve as an initial stage to reduce the number of papers that need to be screened manually.

The subsequent sections begin by examining the data and design of the current simulation study. Following this, the analytic strategy is outlined, detailing the evaluation methods for the models' performances. Lastly, a discussion is provided on which model is more effective in saving time and resources during the preliminary phase of title screening for systematic reviews.

2 Methodology

The code used to generate the results of this project, along with the specific versions of the packages and links to the models used, is available in the following GitHub repository (Migliore, 2024).

2.1 Data

The dataset containing the titles and the human labels is a subset of a larger dataset from the FORAS project (Van De Schoot et al., 2023). The dataset includes meta-data of papers potentially eligible for a systematic review on the trajectories of PTSD after traumatic events. The subset used for the current study consists of the records used for the two calibration sessions. Calibration Session 1 involved two screeners collaboratively reviewing 100 old records, 100 new records, and 100 randomly selected records, to refine the inclusion criteria. In Calibration Session 2, the same two screeners independently reviewed 100 old records, 100 new records, and 100 randomly selected records. During the second session, inter-rater reliability (IRR) was calculated, and any disagreements were discussed and resolved with the assistance of a third screener. The first batch of 300 records (Calibration set 1) is used in the current study to optimize the prompts and the second batch (Calibration set 2) is used for validation purposes. The subset is not yet publicly available.

From the Calibration set 1, two new datasets are created containing only the relevant columns. One dataset includes the title labels and contains the following columns: `MID` (the paper identifier), `title` (the title of each scientific paper), `title_eligible` (the label given by a human expert screener based solely on the title). The other dataset includes the abstract labels and contains the following columns: `MID`, `title`, `TI-AB_final_label` (the label given by a human expert screener based on the paper's title and abstract).

There are 2 missing values in the `title_eligible` column and in the `TI-AB_final_label` column, resulting in 298 records.

The labels in the `title_eligible` column are originally stored as Y (relevant) or N (irrelevant), however, they are converted to 1 and 0, respectively. The labels in the `TI-AB_final_label` column are already binarized.

Among the screened papers, 179 are labeled as relevant and 119 as irrelevant based on the title. There is no significant class imbalance, therefore it is not necessary to employ any balancing technique.

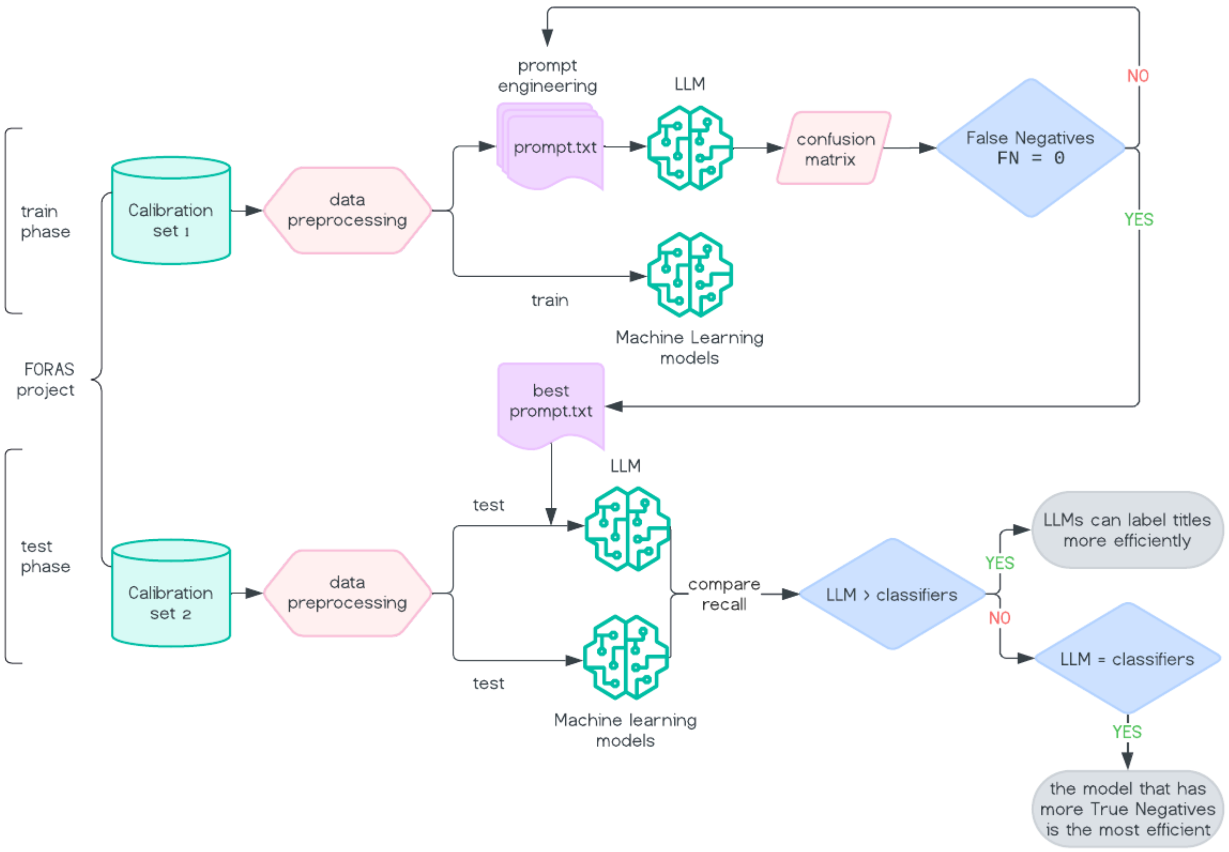
The procedure implemented for Calibration Set 2 mirrors that of Calibration Set 1.
Two new datasets are created for validation purposes: one containing labels based on titles, and the other containing labels based on abstracts. The column names remain consistent with those used in Calibration Set 1, as well as the total number of records, which is 300. However, there are 5 missing entries, resulting in 295 valid records. Finally, the `title_eligible` column is binarized.

2.2 Design

The entire design of the current study is illustrated in Figure 1. The study is divided into two phases: a training phase and a testing phase. For the training phase, the Calibration Set 1 is utilized. The dataset is first preprocessed, as detailed in Chapter 2.1. Next, the preprocessed dataset is used to conduct multiple trials on the LLM with different prompts to identify the optimal prompt. The ideal prompt should result in zero false negatives and maintain a high true negative rate when comparing the LLM's labels to the human labels. To validate the efficacy of the LLM method, the preprocessed Calibration Set 1 is also used to train simpler machine learning models, ensuring that the LLM produces more reliable labels compared to less complex methods. If the LLM does not outperform the simpler classifiers, it would not be justifiable to use the significant computational resources required for a LLM.

Once the optimal prompt is identified, the testing phase begins. The Calibration Set 2 is used to evaluate the machine learning models and to test the LLM with the selected optimal prompt. The performance of the LLM is then compared to that of the machine learning models to determine whether the LLM can screen titles more effectively than traditional methods.

Figure 1. *Design of the simulation*



2.3 Training phase

2.3.1 Machine Learning models

This section is inspired by the work of Syriani et al. (2023), who compared a Large Language Model (GPT-3.5 Turbo) with simpler machine learning models to understand which model is more convenient for title-abstract screening.

The models trained for the comparison include Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machine, all implemented using version 1.5.0 of the `scikit-learn` library (Pedregosa et al., 2011).

A tf-idf vectorizer (Sparck Jones, 1972) is employed to convert the text in the `title` column into numerical vectors before inputting it into the simple classifiers.

2.3.2 Large Language Model

The LLM used for this project is Meta’s Llama 3. It is an open-source pre-trained language model available in the 8B or 70B parameters version (AI@Meta, 2024). The model was released in April 2024, reaching state-of-the-art performance. Since the 70B version takes longer to process inputs and deliver outputs, the 8B version is utilized for the current study.

The Ollama API is used to run the model. Ollama supports Nvidia GPUs with compute capability 5.0+.

Vectorization is not necessary for Llama, since the model already vectorizes the text by itself.

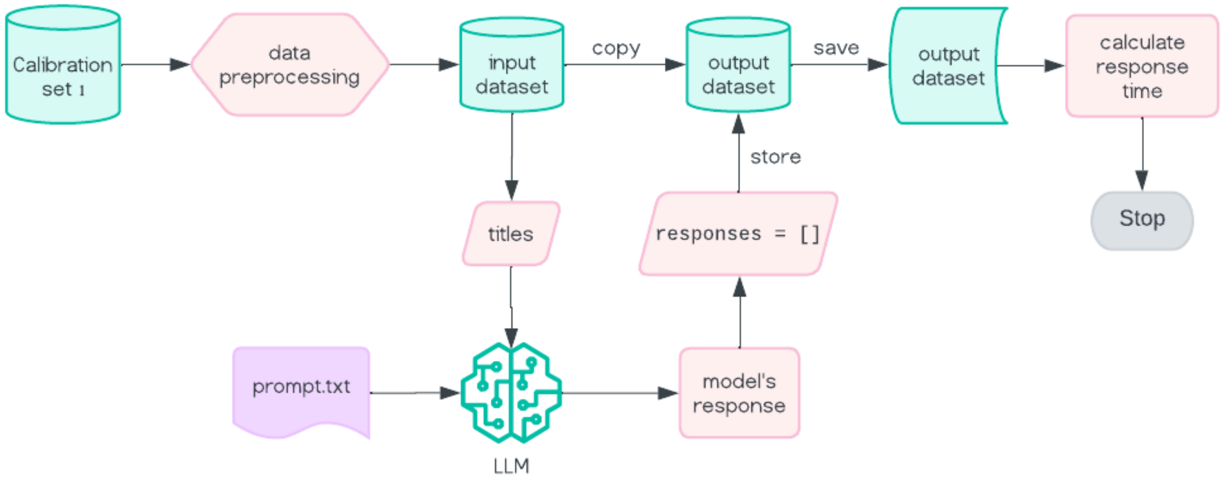
2.3.2.1 Prompt optimization

A pipeline is created to test the different prompts (Fig. 2). First, the text file with the prompt is loaded. Then, the prompt is run on the LLM for each title in the dataset, and the model's response is added to the output dataset. For reproducible results, the model’s temperature hyperparameter is set to 0 and the seed to 42.

The newly labeled dataset is then saved as an excel file and the time needed to produce the response is generated in tokens per second (token/s).

The process is repeated for each attempted prompt.

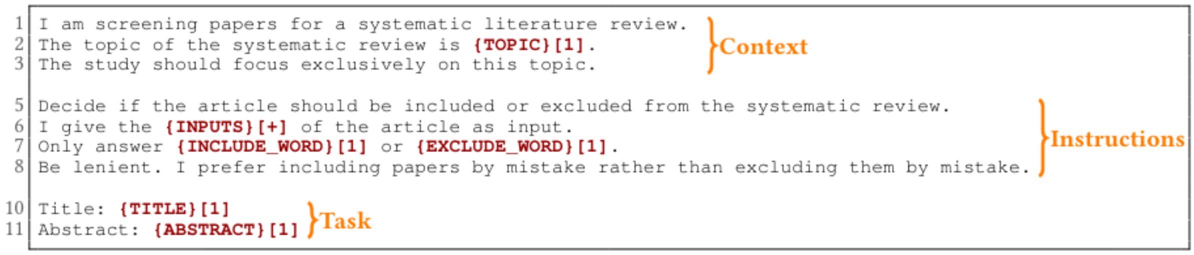
Figure 2. *Diagram of the function used to obtain labels from Llama 3 for each prompt*



Once the workflow pipeline for testing the prompts has been defined, the next step is to perform prompt engineering. All the attempted prompts can be found in the Appendix A of the current paper and have been created manually.

First, we make use of a basic template utilized by Syriani et al. in 2023, who divided their prompt into three sections: 'Context,' 'Instructions,' and 'Task'.

Figure 3. *Basic prompt template used for the current study (Syriani et al., 2023)*



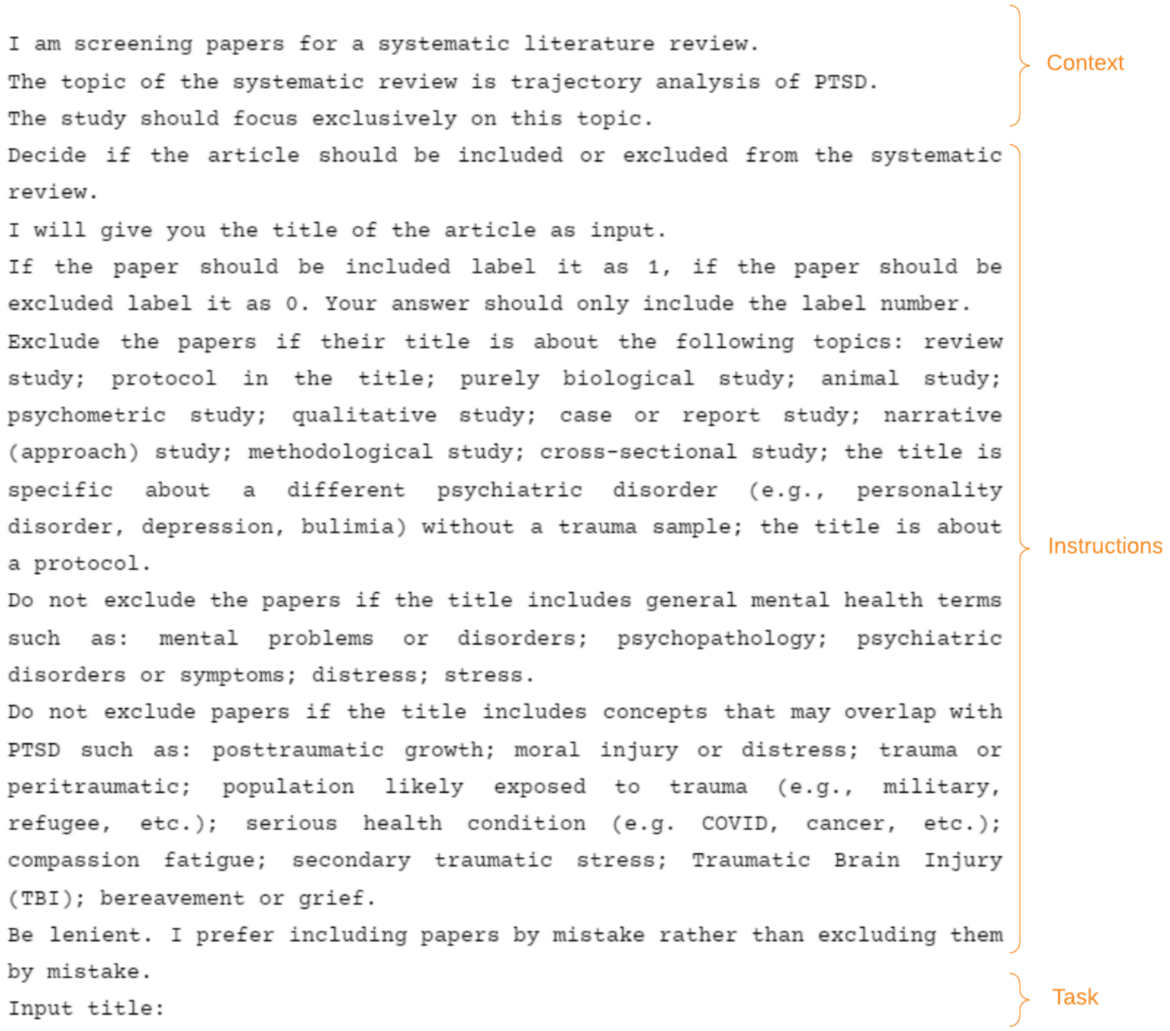
```
I am screening papers for a systematic literature review.
The topic of the systematic review is {TOPIC}[1].
The study should focus exclusively on this topic.

Decide if the article should be included or excluded from the systematic review.
I give the {INPUTS}[+] of the article as input.
Only answer {INCLUDE_WORD}[1] or {EXCLUDE_WORD}[1].
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.

Title: {TITLE}[1]
Abstract: {ABSTRACT}[1]
```

Based on the example in Figure 3, the initial prompt is structured as follows, with the inclusion and exclusion criteria (Van De Schoot et al., 2023) incorporated into the instructions section.

Figure 4. *First prompt used for the current simulation study*



```
I am screening papers for a systematic literature review.
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress.
Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Be lenient. I prefer including papers by mistake rather than excluding them by mistake.
Input title:
```

The initial attempt in Figure 4 yields poor results.

Following the recommendations outlined by Chen et al. (2023), the prompts are incrementally improved until optimal results are achieved. The most significant improvements are observed under the following conditions.

1. The request for leniency is included in both the prompt and the `system` parameter (see Appendix A, prompt 3.2).
2. The model is instructed to follow the directions in a specific sequence to establish priorities (see Appendix A, prompt 10).
3. Two-shots prompting (see Appendix A, prompt 12).
4. Step-by-step reasoning (see Appendix A, prompt 15).
5. Two-shots prompting + step-by-step reasoning (see Appendix A, prompt 17).

2.3.3 Analytic strategy

To determine the effectiveness of LLMs in accurately filtering out irrelevant papers, the quality of the prompts must be evaluated to identify the prompt that elicits the best labeling choices from the LLM. The LLM must label conservatively, ensuring no relevant titles are excluded, while correctly excluding a significant portion of papers. To evaluate the prompts two metrics are employed: confusion matrices and recall. The confusion matrix (Ting, 2011) displays the number of false positives and false negatives predicted by the model. If the number of false negatives in the confusion matrix is zero, the optimal prompt has been identified, allowing the transition to the testing phase. Recall (Buckland & Gey, 1994) is a valuable metric when the cost of false negatives is high, even if it results in an increased number of false positives. The desired outcome is a recall of 1.00, which indicates the absence of false negatives, ensuring that no relevant title is incorrectly labeled as irrelevant.

Prompts with a recall of 1.00, indicating zero false negatives in the confusion matrix, are compared. Among these, the prompt with the highest true negative rate is identified as the optimal prompt.

2.4 Validation phase

Once the optimal prompt is found, the final step is to evaluate the performance of the machine learning models and the LLM. This is done by using the `classification_report` function from the `sklearn.metrics` module (Scikit-learn, 2024), which provides various metrics, such as precision, f1-score, support, and accuracy, but most importantly recall. In fact, as outlined in Chapter 2.3.3, recall is the metric used to compare the models and determine which one is more effective at screening titles.

The simple classifiers and the LLM are evaluated using identical methods. For each model, a confusion matrix is computed alongside the recall score to assess the highest exclusion rate and the lowest false negative rate.

Finally, a matrix is created to compile the outputs of all models, facilitating an easy comparison of their performances.

2.4.1 Machine Learning model

To ensure validity, a Random Classifier is first run to confirm that the other classifiers outperform random guessing, as they would otherwise be unsuitable for comparison. The Logistic Regression, Random Forest, Naive Bayes, and Support Vector Machine are then tested with the Calibration set 2.

2.4.2 Large Language Model

Llama3 8B is also tested with the Calibration set 2 using the prompt that led to 0 false negatives and the highest true negative rate.

3 Results

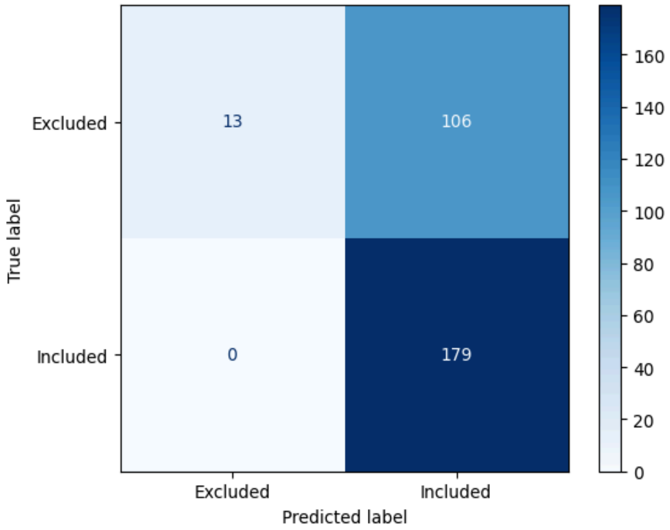
3.1 Training phase

During this phase, the Calibration Set 1 is utilized to train the models and to compare their generated labels with the human labels.

3.1.1 Prompt optimization

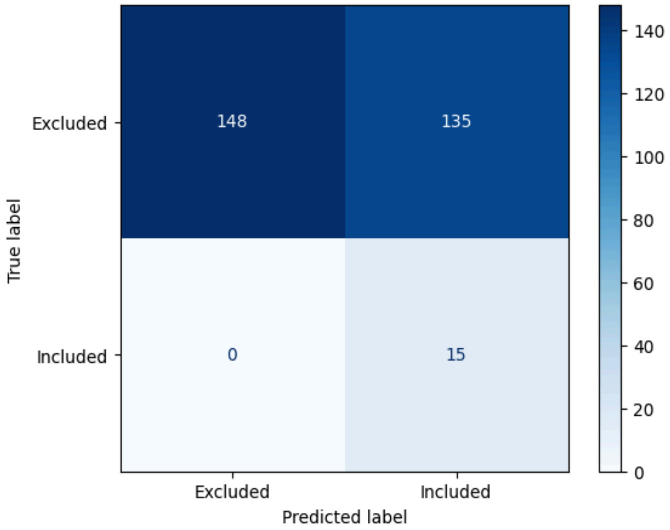
For each prompt, a confusion matrix is generated to compare the model's labels with the human title labels. Recall is subsequently calculated. After multiple attempts employing various prompt strategies, the optimal prompt is identified. Specifically, Prompt 17 (see Appendix A, prompt 17) yields 0 false negatives, resulting in a recall of 1.0. Figure 5 illustrates the confusion matrix for prompt 17.

Figure 5. *Confusion matrix of prompt 17*



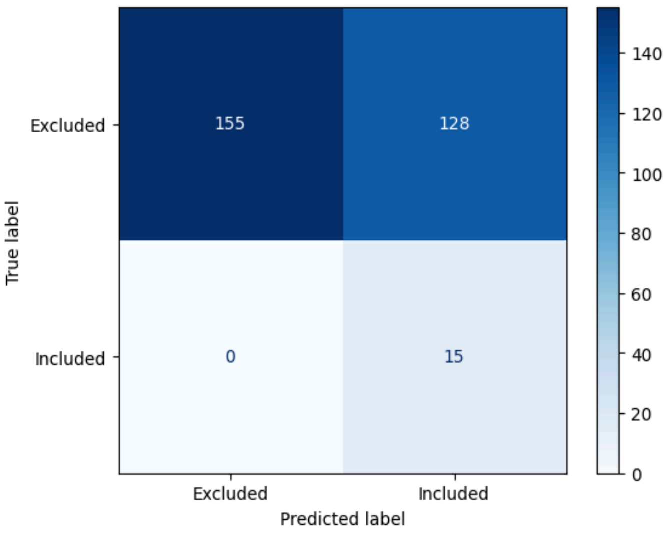
As shown in Figure 5, 21 out of 298 papers are excluded, indicating that the model excludes approximately 7% of the papers. While this result is promising, further analysis is conducted to evaluate the potential for the model to exclude more papers, thereby increasing time and cost savings for researchers. Consequently, the model's results from all previous prompts are re-evaluated, comparing them not against human labels for titles but against the human labels assigned after abstract screening. This approach helps determine whether the model is excluding relevant papers or those that would have been excluded later on during the abstract screening phase. Remarkably, prompt 1 (see Appendix A, prompt 1) already results in 0 false negatives, achieving a recall of 1.0, as illustrated in Figure 6.

Figure 6. *Confusion matrix of prompt 1, comparing the model's labels based on titles with the human labels based on the abstracts*



However, the prompt that achieves the highest number of excluded papers (true negatives) while maintaining 0 false negatives is Prompt 3.1 (see Appendix A, prompt 3.1). As illustrated in Figure 7, prompt 3.1 excludes 155 out of 298 papers, corresponding to a correct exclusion rate of 52%. This efficiency allows the model to significantly reduce the number of papers researchers need to review, resulting in substantial time and cost savings.

Figure 7. *Confusion matrix of prompt 3.1, comparing the model's labels based on titles with the human labels based on the abstracts*



3.2 Validation phase

During the validation phase, the Calibration Set 2 is utilized to test the models and to compare their generated labels with the human labels.

3.2.1 Machine Learning models

The labels made by the machine learning models are compared to the human labels based on titles. The confusion matrices for the classifiers described below are displayed in Figure 8.

Random Classifier

The Random Classifier returns a recall of 0.5180, with 67 false negatives and 86 true negatives, which is essentially equivalent to random performance. In comparison, all classifiers evaluated below exhibit superior performance relative to the Random Classifier.

Logistic Regression

The Logistic Regression already demonstrates a strong performance, with a recall of 0.9928, 1 false negative and 60 true negatives.

Random Forest

The Random Forest classifier returns a recall of 0.9281, with 10 false negatives and 93 true negatives.

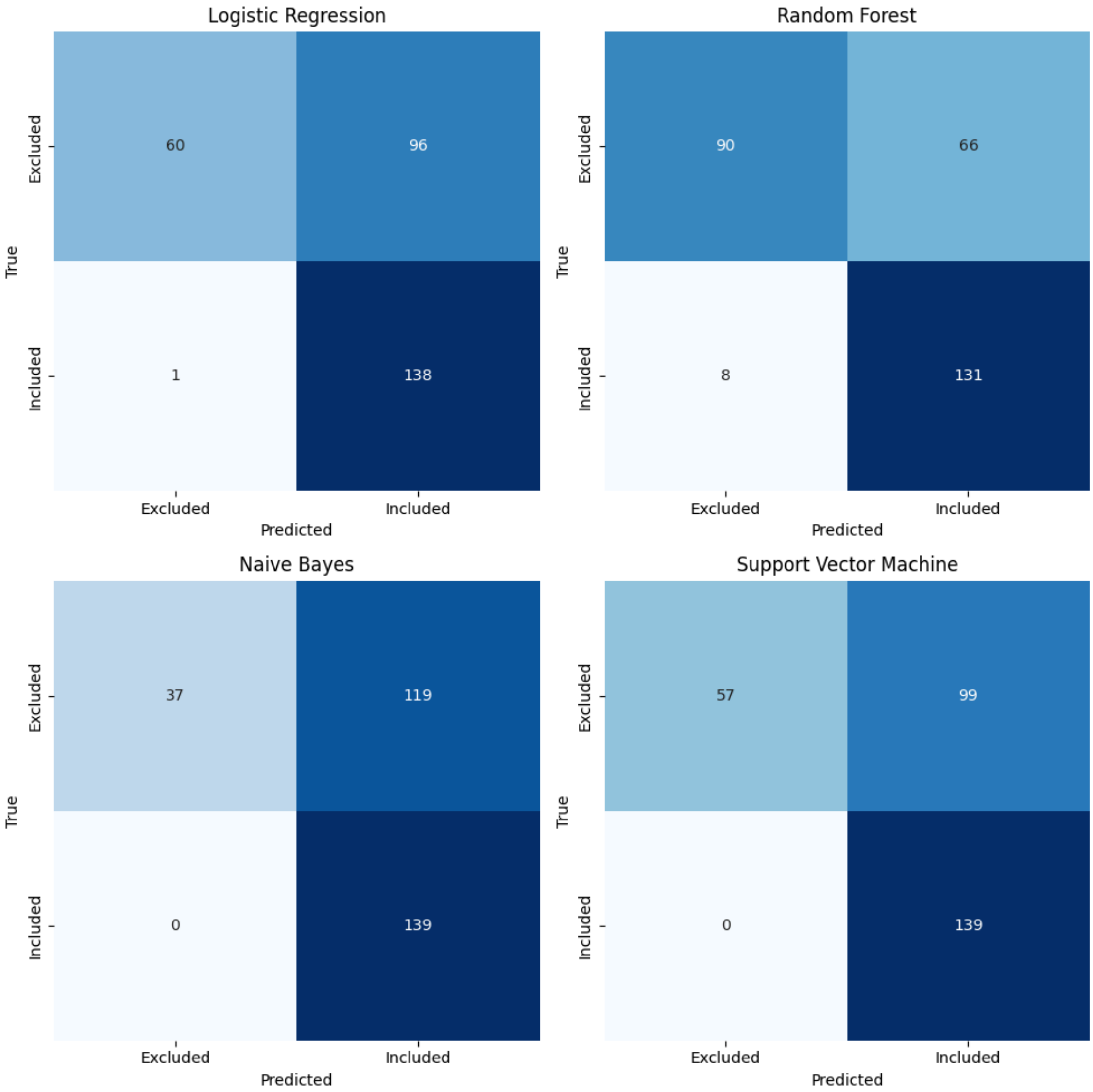
Naive Bayes

The Naive Bayes classifier reaches a recall of 1.0, with 0 false negatives and 37 true negatives.

Support Vector Machine (SVM)

The Support Vector Machine achieves a recall of 1.0, resulting in 0 false negatives. However, the SVM excludes more papers compared to the Naive Bayes classifier. Specifically, the SVM excludes 57 out of 295 papers, making it the most effective machine learning model among those tested.

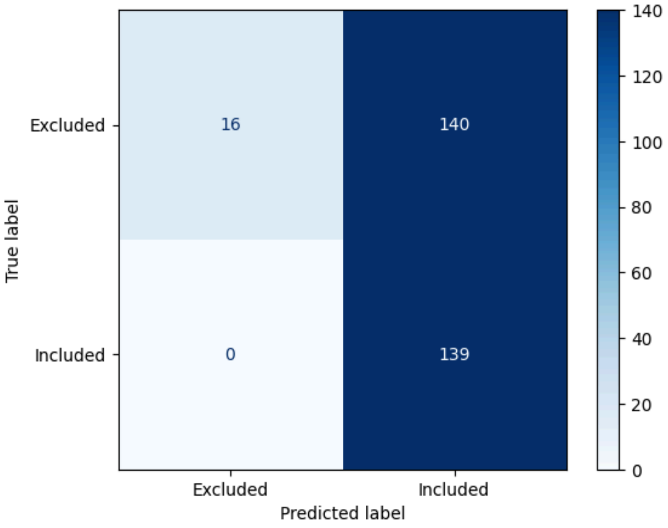
Figure 8. *Confusion matrices of the machine learning models in the test phase*



3.2.2 Large Language Model

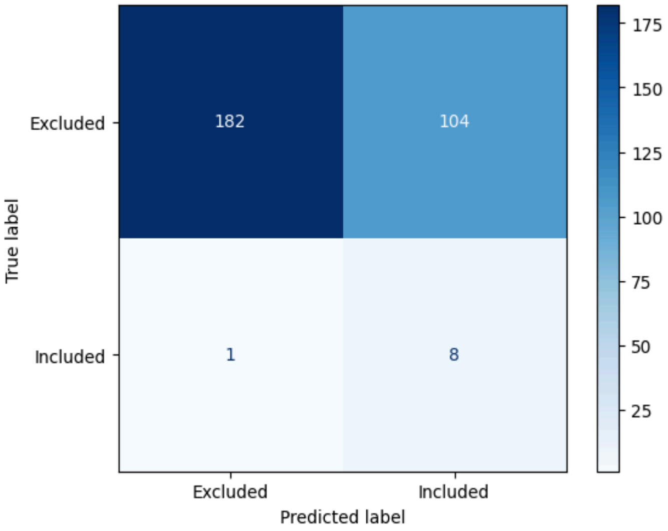
Prompt 17 is evaluated using the human labels based on titles, yielding 0 false negatives and achieving a recall of 1.0. Consistently with the training phase, the number of true negatives remains low, with a count of 16 (Fig. 9).

Figure 9. *Confusion matrix of prompt 17*



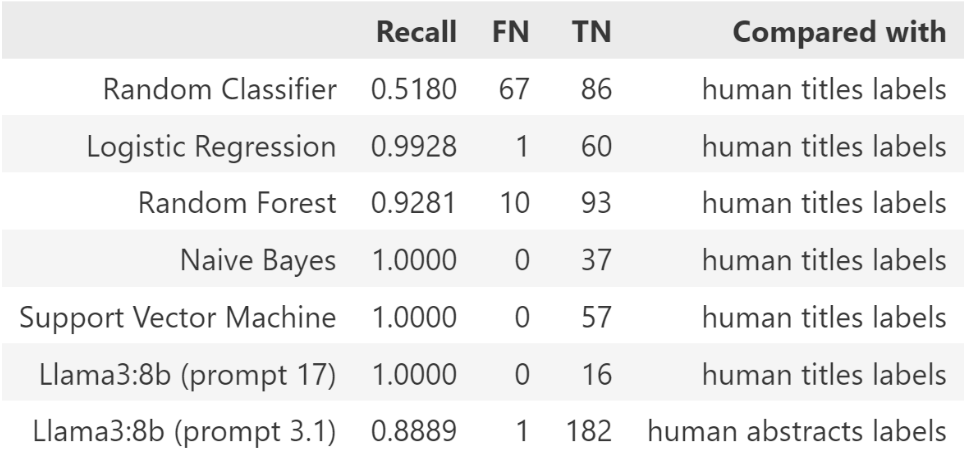
Prompt 3.1 is evaluated due to its higher true negative rate and its apparent lack of exclusion of relevant papers when compared to human labeling based on abstract screening. The recall for prompt 3.1 is 0.8889, with 1 false negative and 182 true negatives (Fig. 10). Utilizing this prompt, Llama3 8B can exclude approximately 62% of the papers, although it misses one relevant paper.

Figure 10. *Confusion matrix of prompt 3.1 compared to the human labels based on abstracts*



Finally, a matrix is created to compile the recall, the false negatives, and the true positives of all models tested during the validation phase, facilitating an easy comparison of their performances (Fig. 11).

Figure 11. *Matrix displaying the recall, False Negatives (FN), and True Negatives (TN) for each model tested in the validation phase*



	Recall	FN	TN	Compared with
Random Classifier	0.5180	67	86	human titles labels
Logistic Regression	0.9928	1	60	human titles labels
Random Forest	0.9281	10	93	human titles labels
Naive Bayes	1.0000	0	37	human titles labels
Support Vector Machine	1.0000	0	57	human titles labels
Llama3:8b (prompt 17)	1.0000	0	16	human titles labels
Llama3:8b (prompt 3.1)	0.8889	1	182	human abstracts labels

The Naive Bayes classifier, the Support Vector Machine, and Llama3 8B with prompt 17 each achieve a recall of 1.0. However, these models differ in the number of papers they correctly exclude: Naive Bayes excludes 37 papers, the SVM excludes 57 papers, and Llama3 8B excludes 16 papers. Based on the number of true negatives, the Support Vector Machine is the most effective model for screening papers based on titles for systematic reviews, as it can automatically exclude over 19% of the papers without missing any relevant paper, making it the safest choice.

In contrast, if the aim is to maximize the number of correctly excluded papers, Llama3 8B with prompt 3.1 is the best model, correctly excluding 182 papers with only one missed paper. Although the LLM makes isolated errors, it is important to consider whether these errors are comparable to those made by humans.

4 Discussion and Conclusions

The objective of the current study was to address the significant time and financial costs associated with screening papers for systematic reviews. Specifically, we investigated whether the title screening phase could be automated using Large Language Models (LLMs) and prompt engineering to optimize labeling decisions. Our findings indicate that LLMs can effectively automate title screening, outperforming simpler classifiers. This approach enables researchers to eliminate nearly 60% of papers at the preliminary stage, thereby reducing the number of papers requiring further screening.

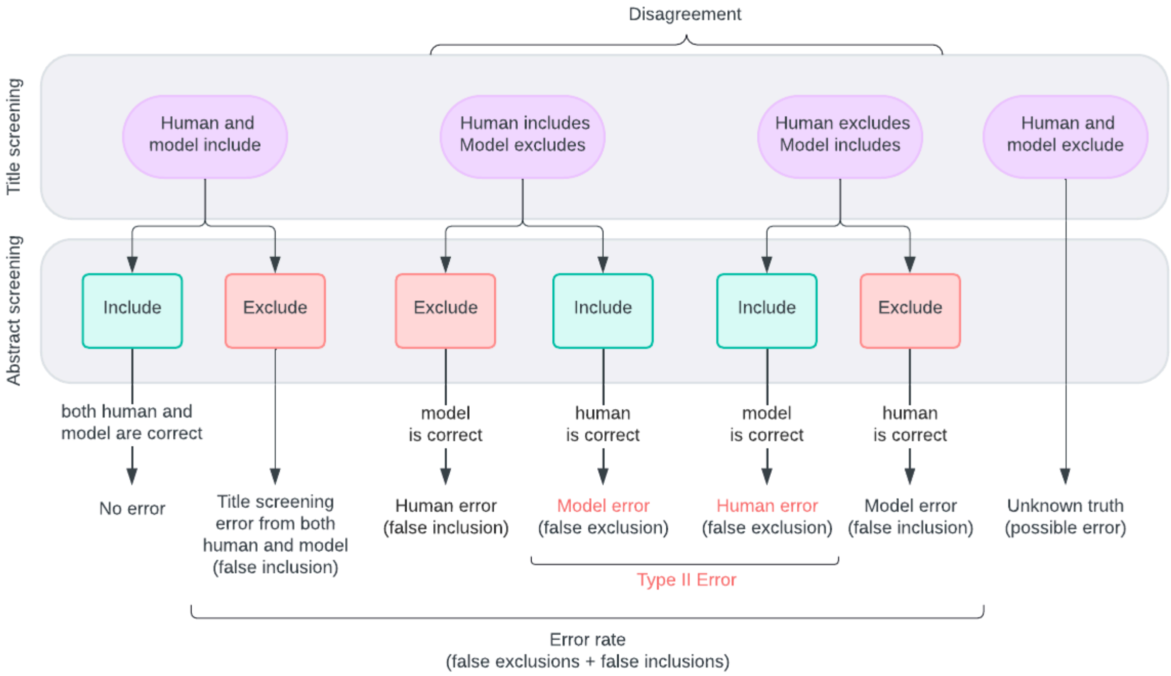
Prior research has examined the efficacy of Large Language Models in the title-abstract screening phase, demonstrating that LLMs can indeed reduce the time required for this process (Guo et al., 2024; Huotala et al., 2024; Syriani et al., 2023). Additionally, other studies indicate that an initial title-screening approach could potentially eliminate 50% of the papers at this stage (Mateen et al., 2013).

Wang et al. (2020) investigated human error rates, comprising both false exclusions and false inclusions, during the screening process of systematic reviews. Their study revealed that, depending on the question type, the average human error rate is 10.76%.

In the current study, calculating the model's error rate is not pertinent; instead, it is crucial to prioritize having zero false exclusions at the cost of more false inclusions. This approach aims to minimize the likelihood of type II errors, which involve erroneously excluding potentially relevant papers (Sedgwick, 2014). However, it is important to note that humans are also prone to type II errors, as illustrated in Figure 12.

In the current study, the model's labels were compared to human labels of Calibration set 1 assigned by multiple reviewers collaboratively (Van De Schoot et al., 2023), as detailed in Chapter 2.1. Consequently, the human error within Calibration set 1 is already minimized. Conversely, for Calibration Set 2, the two reviewers independently labeled the papers and encountered some disagreements: 3 out of 300 records were labeled differently by the reviewers. This suggests that one disagreement between the machine and the human (Fig. 10) falls within the human error rates for type II errors.

Figure 12. *Human and model errors during title screening in systematic reviews*


Disagreement
Title screening
Abstract screening
Human and model include
Human includes Model excludes
Human excludes Model includes
Human and model exclude
Include
Exclude
Exclude
Include
Include
Exclude
both human and model are correct
model is correct
human is correct
model is correct
human is correct
No error
Title screening error from both human and model (false inclusion)
Human error (false inclusion)
Model error (false exclusion)
Human error (false exclusion)
Model error (false inclusion)
Unknown truth (possible error)
Type II Error
Error rate (false exclusions + false inclusions)

Prompt engineering is crucial for eliciting the desired behaviors from Large Language Models. Notably, varying prompts lead to different levels of performance quality (Kocoń et al., 2023). In the current paper we outlined the best practices for generating high-quality prompts and described our successful outcomes. However, it is important to acknowledge that prompts can vary between users, indicating that the quality of the outputs is influenced by the individual creating the prompt.

To assess the quality of prompts, either subjective or objective evaluations can be utilized. Subjective evaluations typically involve human reviewers and, despite potential inconsistencies, often result in higher quality assessments. For the current project, a human-labeled dataset was used as a benchmark to assess the accuracy of the LLM generated labels. On the other hand, objective evaluations, such as BLEU (Papineni et al., 2001), ROUGE (Chin-Yew, 2004), METEOR (Banerjee & Lavie, 2005), and BERTScore (T. Zhang et al., 2020), provide more standardized measures but may lack precision due to their generalized nature (Chen et al., 2023).

Although the findings of this study are promising, certain limitations must be considered for future implementations.

First, the datasets used for training and testing in the current study are relatively small (300 papers each) compared to the thousands of papers typically encountered by researchers.

Future work should apply this procedure to larger datasets to evaluate the performance of the LLM with a more substantial volume of data.

Second, Llama3 8B was utilized in the current study due to its faster response time. However, Llama3 70B offers higher performance, while requiring more time to generate responses. Future research could investigate whether using Llama3 70B can reduce the error rate of the LLM.

Third, the classifiers used are machine learning models designed to minimize loss, aiming to reduce incorrect labels, including false positives (Q. Wang et al., 2022), which are not the focus of this study. In contrast, the Large Language Model can be instructed to adopt a more conservative approach, thereby maximizing recall. Consequently, the two models are optimized differently, resulting in the observed variation in performance. Future projects should aim to compare models that are optimized for the same objective.

Lastly, we conducted a comparative analysis between errors made by humans and those made by the machine. However, we are not aware of any existing papers reporting the type II error rates of humans in paper screening for systematic reviews. Therefore, there is no benchmark to determine whether the type II error rate of the LLM aligns with human error rates or if humans perform better. Future studies could concentrate on calculating type II error rates in human screeners, similarly to the work of Wang et al. in 2020.

The primary conclusion of this study is that Large Language Models can be effectively used by researchers during the title screening phase to reduce both time and financial costs. Through prompt engineering, LLMs can be tailored to operate conservatively, ensuring that no pertinent papers are inadvertently excluded, thereby enhancing their usability.

Finally, it is essential to consider the trustworthiness of Large Language Models, since they can produce non-factual information, and their reasoning process is often opaque. Consequently, reliance solely on these models should be approached with caution. For this reason, we suggest using them exclusively in the preliminary screening phase, rather than in the more critical stages of abstract and full-text screening, which necessitate greater human involvement.

Acknowledgments

We would like to express our sincere gratitude to the following individuals for their valuable contributions to this research.

First, we thank Jelle Teijema for his guidance and support throughout the course of this project. His insightful feedback and reflections on critical aspects of the research were instrumental in shaping this work.

Second, we thank Bruno Messina Coimbra for his assistance in crafting the step-by-step reasoning prompts. His detailed explanations of the reasoning behind specific labeling decisions in the human dataset were crucial in establishing a connection between the human and machine labeling processes.

Reproducibility statement

When using Llama3, hyperparameters such as the seed and a temperature set to zero were employed to achieve reproducible and deterministic outputs.
Moreover, the code used to generate the results of this project, along with the specific versions of the packages, following all the steps mentioned in Chapter 2, is available in the following GitHub repository (Migliore, 2024).

Generative AI Statement

GPT-4-o was utilized to review and correct the linguistic form of the paper. However, GPT-4-o was not used to generate knowledge, and thus, it does not contribute any intellectual property to this work.

References

AI@Meta. (2024). *Llama 3 Model Card*. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

Banerjee, S., & Lavie, A. (2005). METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the Acl Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, 65–72. https://aclanthology.org/W05-0909

Buckland, M., & Gey, F. (1994). The relationship between Recall and Precision. *Journal of the American Society for Information Science*, *45*(1), 12–19. https://doi.org/10.1002/(SICI)1097-4571(199401)45:1<12::AID-ASI2>3.0.CO;2-L

Calderon Martinez, E., Flores Valdés, J. R., Castillo, J. L., Castillo, J. V., Blanco Montecino, R. M., Morin Jimenez, J. E., Arriaga Escamilla, D., & Diarte, E. (2023). 10 Steps to Conduct a Systematic Review. *Cureus*. https://doi.org/10.7759/cureus.51422

Chen, B., Zhang, Z., Langrené, N., & Zhu, S. (2023). *Unleashing the potential of prompt engineering in Large Language Models: A comprehensive review* (arXiv:2310.14735). arXiv. http://arxiv.org/abs/2310.14735

Chin-Yew, L. (2004). *Rouge: A package for automatic evaluation of summaries.: Vol. Text Summarization Branches Out*. Association for Computational Linguistics. https://aclanthology.org/W04-1013

Guo, E., Gupta, M., Deng, J., Park, Y.-J., Paget, M., & Naugler, C. (2024). Automated Paper Screening for Clinical Reviews Using Large Language Models. *Journal of Medical Internet Research*, *26*, e48996. https://doi.org/10.2196/48996

Holtzman, A., Buys, J., Du, L., Forbes, M., & Choi, Y. (2020). *The Curious Case of Neural Text Degeneration* (arXiv:1904.09751). arXiv. http://arxiv.org/abs/1904.09751

Huotala, A., Kuutila, M., Ralph, P., & Mäntylä, M. (2024). *The Promise and Challenges of Using LLMs to Accelerate the Screening Process of Systematic Reviews* (arXiv:2404.15667). arXiv. http://arxiv.org/abs/2404.15667

Kocoń, J., Cichecki, I., Kaszyca, O., Kochanek, M., Szydło, D., Baran, J., Bielaniewicz, J., Gruza, M., Janz, A., Kanclerz, K., Kocoń, A., Koptyra, B., Mieleszczenko-Kowszewicz, W., Miłkowski, P., Oleksy, M., Piasecki, M., Radliński, Ł., Wojtasik, K., Woźniak, S., & Kazienko, P. (2023). ChatGPT: Jack of all trades, master of none. *Information Fusion*, *99*, 101861. https://doi.org/10.1016/j.inffus.2023.101861

Lefebvre, C., Manheimer, E., & Glanville, J. (2008). Searching for Studies. In J. P. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (1st ed., pp. 95–150). Wiley. https://doi.org/10.1002/9780470712184.ch6

Lewkowycz, A., Andreassen, A., Dohan, D., Dyer, E., Michalewski, H., Ramasesh, V., Slone, A., Anil, C., Schlag, I., Gutman-Solo, T., Wu, Y., Neyshabur, B., Gur-Ari, G., & Misra, V. (2022). *Solving Quantitative Reasoning Problems with Language Models* (arXiv:2206.14858). arXiv. http://arxiv.org/abs/2206.14858

Liu, B., & Udell, M. (2020). *Impact of Accuracy on Model Interpretations* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2011.09903

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*, *55*(9), 1–35. https://doi.org/10.1145/3560815

Logan IV, R. L., Balažević, I., Wallace, E., Petroni, F., Singh, S., & Riedel, S. (2021). *Cutting Down on Prompts and Parameters: Simple Few-Shot Learning with Language Models* (arXiv:2106.13353). arXiv. http://arxiv.org/abs/2106.13353

Mateen, F., Oh, Tergas, Bhayani, & Kamdar. (2013). Titles versus titles and abstracts for initial screening of articles for systematic reviews. *Clinical Epidemiology*, 89. https://doi.org/10.2147/CLEP.S43118

Meline, T. (2006). Selecting Studies for Systemic Review: Inclusion and Exclusion Criteria. *Contemporary Issues in Communication Science and Disorders*, *33*(Spring), 21–27. https://doi.org/10.1044/cicsd_33_S_21

Migliore, G. (2024). *The application of LLM prompt engineering to optimize title screening*. https://github.com/GiuliMigliore/LLMs-title-screening

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2001). BLEU: A method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, 311. https://doi.org/10.3115/1073083.1073135

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Sampson, M., Tetzlaff, J., & Urquhart, C. (2011). Precision of healthcare systematic review searches in a cross‐sectional sample. *Research Synthesis Methods*, *2*(2), 119–125. https://doi.org/10.1002/jrsm.42

Scikit-learn. (2024, May 13). Scikit-learn. *GitHub*. https://github.com/scikit-learn/scikit-learn/blob/main/sklearn/metrics/_classification.py

Sedgwick, P. (2014). Pitfalls of statistical hypothesis testing: Type I and type II errors. *BMJ*, *349*(jul03 1), g4287–g4287. https://doi.org/10.1136/bmj.g4287

Sparck Jones, K. (1972). A Statistical Interpretation of Term Specificity and its Application in Retrieval. *Journal of Documentation*, *28*(1), 11–21.

https://doi.org/10.1108/eb026526

Syriani, E., David, I., & Kumar, G. (2023). *Assessing the Ability of ChatGPT to Screen Articles for Systematic Reviews* (arXiv:2307.06464). arXiv. http://arxiv.org/abs/2307.06464

Ting, K. M. (2011). Confusion Matrix. In C. Sammut & G. I. Webb (Eds.), *Encyclopedia of Machine Learning* (pp. 209–209). Springer US. https://doi.org/10.1007/978-0-387-30164-8_157

Tunstall, L., Werra, L. von, & Wolf, T. (2022). *Natural language processing with transformers: Building language applications with Hugging Face* (First edition). O'Reilly Media.

Van De Schoot, R., Coimbra, B., Evenhuis, T., Lombaers, P., Van Zuiden, M., Grandfield, B., De Bruin, J., Teijema, J., De Bruin, L., Neeleman, R., & Jalsovec, E. (2023). Trajectories of PTSD Following Traumatic Events: A Systematic and Multi-database Review. *PROSPERO*.

Wang, Q., Ma, Y., Zhao, K., & Tian, Y. (2022). A Comprehensive Survey of Loss Functions in Machine Learning. *Annals of Data Science*, *9*(2), 187–212. https://doi.org/10.1007/s40745-020-00253-5

Wang, Z., Nayfeh, T., Tetzlaff, J., O'Blenis, P., & Murad, M. H. (2020). Error rates of human reviewers during abstract screening in systematic reviews. *PLOS ONE*, *15*(1), e0227742. https://doi.org/10.1371/journal.pone.0227742

Wu, S., Shen, E. M., Badrinath, C., Ma, J., & Lakkaraju, H. (2023). *Analyzing Chain-of-Thought Prompting in Large Language Models via Gradient-based Feature Attributions* (arXiv:2307.13339). arXiv. http://arxiv.org/abs/2307.13339

Yang, M., Qu, Q., Tu, W., Shen, Y., Zhao, Z., & Chen, X. (2019). Exploring Human-Like Reading Strategy for Abstractive Text Summarization. *Proceedings of the AAAI*

Conference on Artificial Intelligence, *33*(01), 7362–7369. https://doi.org/10.1609/aaai.v33i01.33017362

Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., & Narasimhan, K. (2024). Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, *36*.

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q., & Artzi, Y. (2020). *BERTScore: Evaluating Text Generation with BERT* (arXiv:1904.09675). arXiv. http://arxiv.org/abs/1904.09675

Zhang, Z., Gao, J., Dhaliwal, R. S., & Li, T. J. J. (2023). Visar: A human-ai argumentative writing assistant with visual programming and rapid draft prototyping. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–30.

Zhang, Z., Zhang, A., Li, M., & Smola, A. (2022). *Automatic Chain of Thought Prompting in Large Language Models* (arXiv:2210.03493). arXiv. http://arxiv.org/abs/2210.03493

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2023). *A Survey of Large Language Models* (Version 13). arXiv. https://doi.org/10.48550/ARXIV.2303.18223

Appendix A. Prompts

This appendix contains all the prompts tested during the prompt engineering phase of the study, arranged in chronological order from the earliest to the most recent.

1. *“I am screening papers for a systematic literature review.*
   The topic of the systematic review is trajectory analysis of PTSD.
   The study should focus exclusively on this topic.
   Decide if the article should be included or excluded from the systematic review.
   I will give you the title of the article as input.
   If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
   Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
   Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress.
   Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
   Be lenient. I prefer including papers by mistake rather than excluding them by mistake.
   Input title: ”

2. *“I am screening papers for a systematic literature review.*
   The topic of the systematic review is trajectory analysis of PTSD.
   The study should focus exclusively on this topic.
   Decide if the article should be included or excluded from the systematic review.
   I will give you the title of the article as input.

If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

Step 1: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Step 2: Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.

Step 1 has the priority on step 2.

Be lenient. I prefer including papers by mistake rather than excluding them by mistake.

Input title: "

3. *"I am screening papers for a systematic literature review.*

   The topic of the systematic review is trajectory analysis of PTSD.

   The study should focus exclusively on this topic.

   Decide if the article should be included or excluded from the systematic review.

   I will give you the title of the article as input.

   If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

   Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

   Do not exclude the papers if the title includes general mental health *terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.*

If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake.

You want to have as little false negatives as possible.

Input title: "

3.1. `system = None`

3.2. `system = "You need to be lenient. You need to minimize the number of false negatives as much as possible."`

4. *"You are screening papers for a systematic literature review.*

The topic of the systematic review is trajectory analysis of PTSD.

The study should focus exclusively on this topic.

Decide if the article should be included or excluded from the systematic review.

I will give you the title of the article as input.

If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain

Injury (TBI); bereavement or grief.
BE CONSERVATIVE!!! I prefer including papers by mistake rather than excluding them by mistake.
Input title: “

5. *“You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Input title: “

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible."
```

6. *“You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.

I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Do not exclude the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Do not exclude papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title: "

6.1. system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible."

6.2. system = "You are a researcher conducting a systematic review. Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible."

7. *"You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.

I will give you the title of the article as input.

If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

Do not include the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress.

Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.

Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.

Input title: “

```
system = "Be conservative. If you find an ambiguous title, it is
better to include it: I prefer including papers by mistake rather
than excluding them by mistake. You want to have as little false
negatives as possible."
```

8. *“You are screening papers for a systematic literature review.*

   The topic of the systematic review is trajectory analysis of PTSD.

   The study should focus exclusively on this topic.

   Decide if the article should be included or excluded from the systematic review.

   I will give you the title of the article as input.

   If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

   Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric *disorders or symptoms; distress; stress.*

Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.

Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.

Input title: "

```
system = "Be conservative. If you find an ambiguous title, it is
better to include it: I prefer including papers by mistake rather
than excluding them by mistake. You want to have as little false
negatives as possible."
```

9. *"You are screening papers for a systematic literature review.*

The topic of the systematic review is trajectory analysis of PTSD.

The study should focus exclusively on this topic.

Decide if the article should be included or excluded from the systematic review.

I will give you the title of the article as input.

If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

Step 1: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain

Injury (TBI); bereavement or grief.
Step 2: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Step 1 has the priority over step 2.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title: "

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible."
```

10. *"You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
Step 1: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Step 2: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study;

cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.

Input title: "

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. Step 1 has the priority over step 2."
```

11. *"You are screening papers for a systematic literature review.*

The topic of the systematic review is trajectory analysis of PTSD.

The study should focus exclusively on this topic.

Decide if the article should be included or excluded from the systematic review.

I will give you the title of the article as input.

If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.

Inclusion criteria: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress.

Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.

Exclusion criteria: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.

Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than

excluding them by mistake. You want to have as little false negatives as possible.
Input title: "

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. The inclusion criteria have the priority over the exclusion criteria."
```

12. *"You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Your answer should only include the label number.
Step 1: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Step 2: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
I will give you two examples.
Example title: "Internet-based early intervention to prevent

posttraumatic stress disorder in injury patients: randomized controlled trial"
Example label: 1
Example title: "Time-course analysis of frontal gene expression profiles in the rat model of posttraumatic stress disorder and a comparison with the conditioned fear model"
Example label: 0
Here is your input title.
Input title: ‟

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. Step 1 has the priority over step 2."
```

13. You are screening papers for a systematic literature review.
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Solve this task using step by step reasoning and write down the label number at the end of your answer. The label number should be in between hashes. Example: "#1#".
Step 1: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Step 2: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to

include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title:

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. Step 1 has the priority over step 2."
```

14. *You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Solve this task using step by step reasoning and write down the label number at the end of your answer. The label number should be in between hashes. Example: "#1#".
Here there are the inclusion and exclusion criteria.
Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title:

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. Step 1 has the priority over step 2."
```

15. *You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Solve this task using step by step reasoning, based on the inclusion and exclusion criteria, and write down the label number at the end of your answer. The label number should be in between hashes. Example: "#1#".
Here there are the inclusion and exclusion criteria.
Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title:

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather
```

```
than excluding them by mistake. You want to have as little false
negatives as possible."
```

16. *You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Solve this task using step by step reasoning, based on the inclusion and exclusion criteria, and write down the label number at the end of your answer. The label number should be in between hashes. Example: "##1##".
Here there are the inclusion and exclusion criteria.
Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
Input title:

```
system = "Be conservative. If you find an ambiguous title, it is
better to include it: I prefer including papers by mistake rather
than excluding them by mistake. You want to have as little false
negatives as possible. Use step by step reasoning."
```

17. *You are screening papers for a systematic literature review.*
The topic of the systematic review is trajectory analysis of PTSD.
The study should focus exclusively on this topic.
Decide if the article should be included or excluded from the systematic review.
I will give you the title of the article as input.
If the paper should be included label it as 1, if the paper should be excluded label it as 0. Solve this task using step by step reasoning, based on the inclusion and exclusion criteria, and write down the label number at the end of your answer. The label number should be in between hashes, like this "#1#" or this "#0#".
Step 1: Include the papers if the title includes general mental health terms such as: mental problems or disorders; psychopathology; psychiatric disorders or symptoms; distress; stress. Include papers if the title includes concepts that may overlap with PTSD such as: posttraumatic growth; moral injury or distress; trauma or peritraumatic; population likely exposed to trauma (e.g., military, refugee, etc.); serious health condition (e.g. COVID, cancer, etc.); compassion fatigue; secondary traumatic stress; Traumatic Brain Injury (TBI); bereavement or grief.
Step 2: Exclude the papers if their title is about the following topics: review study; protocol in the title; purely biological study; animal study; psychometric study; qualitative study; case or report study; narrative (approach) study; methodological study; cross-sectional study; the title is specific about a different psychiatric disorder (e.g., personality disorder, depression, bulimia) without a trauma sample; the title is about a protocol.
Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible.
I will give you two examples.
Example title: "Internet-based early intervention to prevent posttraumatic stress disorder in injury patients: randomized controlled trial"
Example answer: "A 'controlled trial' is mentioned, which means it is a longitudinal study. Patients have an injury, which means a trauma is involved, so they might have a posttraumatic stress disorder. For these reasons maybe they did the analysis that we are looking for. That's why this paper should be included. Label: #1#"
Example title: "Time-course analysis of frontal gene expression

profiles in the rat model of posttraumatic stress disorder and a comparison with the conditioned fear model"
Example answer: "Rats are mentioned, which means it is an animal study, which is in the exclusion criteria. Moreover, this study sounds purely biological, which is also in the exclusion criteria. Therefore this title should be excluded. Label: #0#."
Here is your input title.
Input title:

```
system = "Be conservative. If you find an ambiguous title, it is better to include it: I prefer including papers by mistake rather than excluding them by mistake. You want to have as little false negatives as possible. Step 1 has the priority over step 2."
```