Enhancing Systematic Literature Review with Advanced AI: A Study on LLM-Based Screening

Dan Li, Leihong Wu, Svitlana Shpyleva, Ting Li, Joshua Xu*

National Center for Toxicological Research, U.S. Food and Drug Administration, Jefferson, AR, USA

Joshua.xu@fda.hhs.gov

Abstract

This study explores the use of Large Language Models (LLMs) to automate and enhance systematic literature reviews, significantly reducing screening time while improving categorization accuracy across various disciplines.

Systematic literature reviews are essential for synthesizing knowledge and informing decisions across various fields. However, the manual screening of extensive literature is a labor-intensive and time-consuming process. This study explores the use of Large Language Models (LLMs) to automate literature screening, aiming to accelerate the identification of relevant studies, improve the accuracy of categorization, and provide useful information to assist the review processes. We evaluate the performance of leading LLMs, including GPT-3.5, GPT-4, and Anthropic's Claude2, in automating the classification of publications. A pre-labeled dataset serves as the basis for assessing the effectiveness of these models. Our analysis includes the evaluation of N-shot learning, the utilization of chain-of-thought reasoning, and the measurement of sensitivity and specificity, along with an exploration of factors influencing accuracy. The results demonstrate the promising capabilities of LLMs in systematic literature review, particularly in terms of reproducibility and sensitivity, with moderate specificity. Providing several examples of abstracts, labels, and reasoning explanations significantly enhances the models' categorization performance. Furthermore, our study identified key factors affecting prediction outcomes, such as keyword selection, prompt formatting, training data balance, and variations in reasoning explanations. The application of advanced LLMs for efficient screening of extensive literature databases offers a transformative approach to systematic reviews. This automation not only reduces the time and effort required for manual review but also provides accurate and reliable information, enhancing decision-making processes. The potential of LLMs extends across various disciplines, demonstrating their versatility and impact in managing large volumes of textual data.

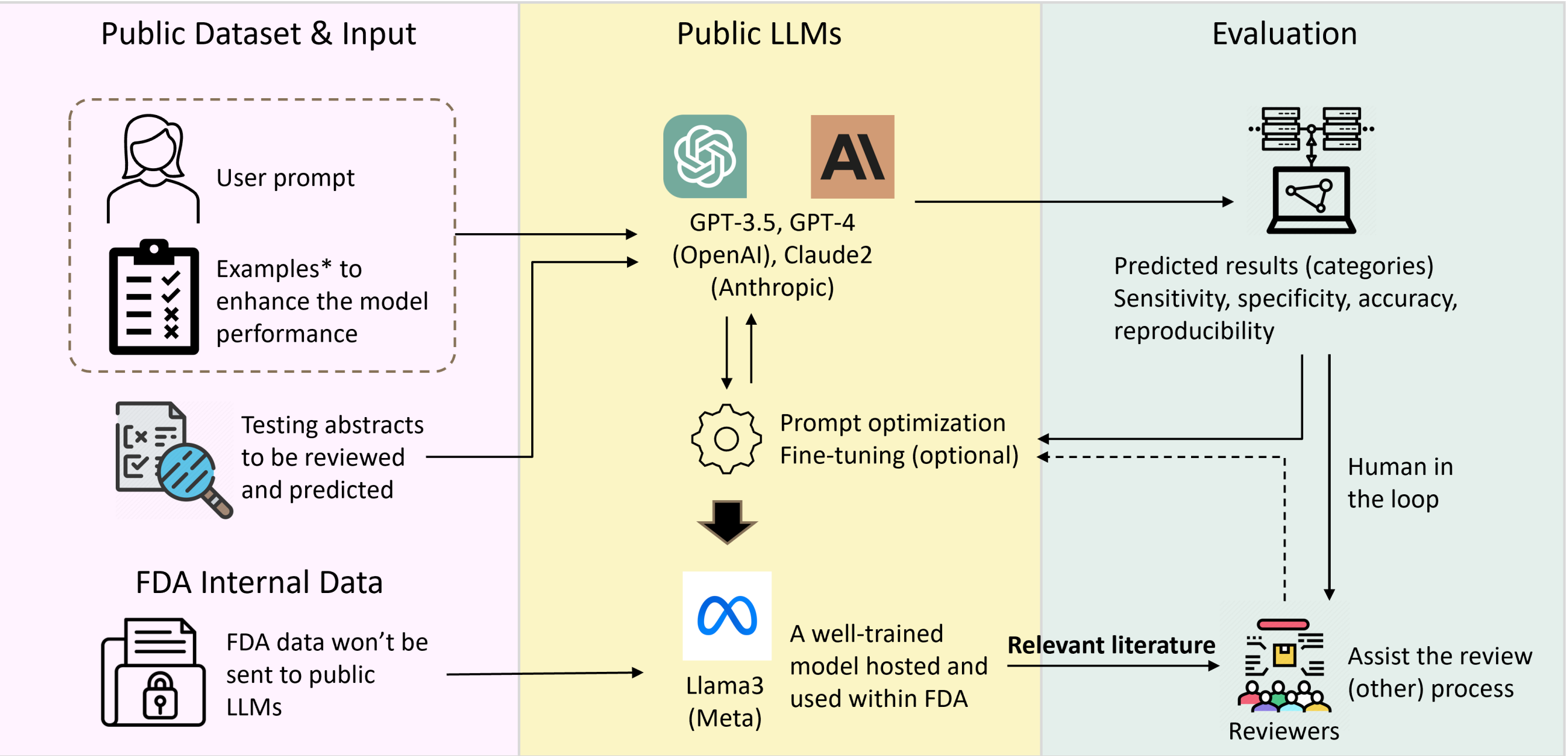
Public Dataset & Input
User prompt
Examples* to enhance the model performance
Testing abstracts to be reviewed and predicted
FDA Internal Data
FDA data won't be sent to public LLMs
Public LLMs
GPT-3.5, GPT-4 (OpenAI), Claude2 (Anthropic)
Prompt optimization
Fine-tuning (optional)
Llama3 (Meta)
A well-trained model hosted and used within FDA
Relevant literature
Evaluation
Predicted results (categories) Sensitivity, specificity, accuracy, reproducibility
Human in the loop
Assist the review (other) process
Reviewers

Figure 1. Study Design. Public LLMs and datasets were used to evaluate literature screening performance and explore influencing factors. Optimized prompts and a localized LLM will be applied to FDA literature screening to assist in the review processes.

Results

Study Design

In our previous study [1], we curated a dataset consisting of research abstracts, systematically categorizing them as either relevant or irrelevant to the topic of chlorine safety. From this dataset, we randomly selected a subset of 30 abstracts, 15 relevant and 15 irrelevant. Based on these training abstracts and their labels, we prepared multiple sets of reasoning explanations, employing both human experts and LLMs to facilitate comparative analyses. These example sets with reasoning explanations served as valuable background information for LLMs to refer. We also conducted tests using a variety of prompts, recognizing their crucial role in shaping the performance of LLMs, to observe and compare the outcomes and overall effectiveness for the literature screening task.

In this study, our primary objective was to assess and enhance the performance of LLMs in classifying research abstracts as relevant or irrelevant to our chlorine safety study. We specifically investigated the effects of the number of learning shots, different prompts, and various reasoning explanations on the model's performance.

The findings from this investigation will ultimately help optimize an AI-enhanced systematic literature review pipeline, incorporating a localized LLM capable of handling FDA internal data to assist in review and other related processes.

Impact of Numbers of Learning Shots

To evaluate the effect of different numbers of learning shots, we provided GPT-3.5 with various sets of examples, including abstracts, labels, and human-generated reasoning explanations, which facilitated chain-of-thought processes. For instance, in a 5-shot test, the model was provided with 5 relevant and 5 irrelevant abstracts, along with associated reasoning explanations in the prompts. We assessed metrics such as reproducibility, sensitivity, specificity, and precision across 0, 1, 3, 5, and 10 learning shots. The reasoning explanations were created by a human expert who reviewed the abstracts, explaining their relevance or irrelevance. Notably, a higher number of learning shots correlated with slightly improved performance.

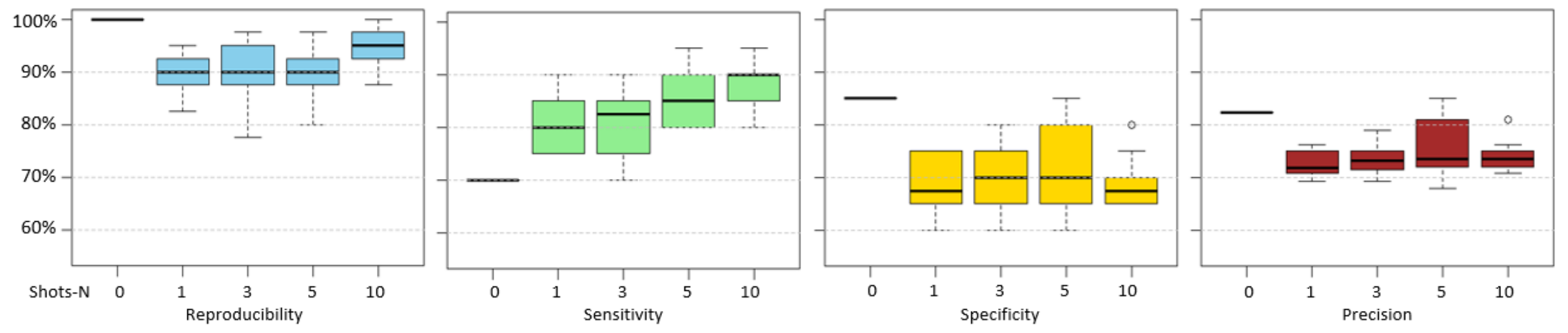

Figure 2. Impact of Numbers of Learning Shots. The comparison of various numbers of shots was conducted across different metrics, including reproducibility, sensitivity, specificity, and precision. The reasoning explanations were generated by a single human expert. The reproducibility was calculated between two runs (different combinations of examples provided to models) as the number of identical predictions divided by the total abstracts were predicted.

Disclaimer: This poster reflects the views of the author and does not necessarily reflect those of the U.S. Food and Drug Administration. Any mention of commercial products is for clarification only and is not intended as approval, endorsement, or recommendation.

Reference: [1] Wu L, Chen S, Guo L, Shpyleva S, Harris K, et al. Development of benchmark datasets for text mining and sentiment analysis to accelerate regulatory literature review. Regulatory Toxicology and Pharmacology. 2023;137:105287

Impact of Different Reasoning Sets

To evaluate the impact of diverse reasoning sets generated by AI models and human experts, we presented example abstracts with pre-defined labels to GPT-3.5, GPT-4, Claude2, and two human experts. We tasked them to read the abstracts and explain the reasons for their relevance or irrelevance based on the true labels in study [1]. We applied a 5-shot learning approach to assess the performance of GPT-3.5. Incorporating reasoning explanations resulted in only a marginal increase in the measured metrics. Meanwhile, no significant improvements were found when compared to scenarios with no reasoning provided (only abstracts and labels).

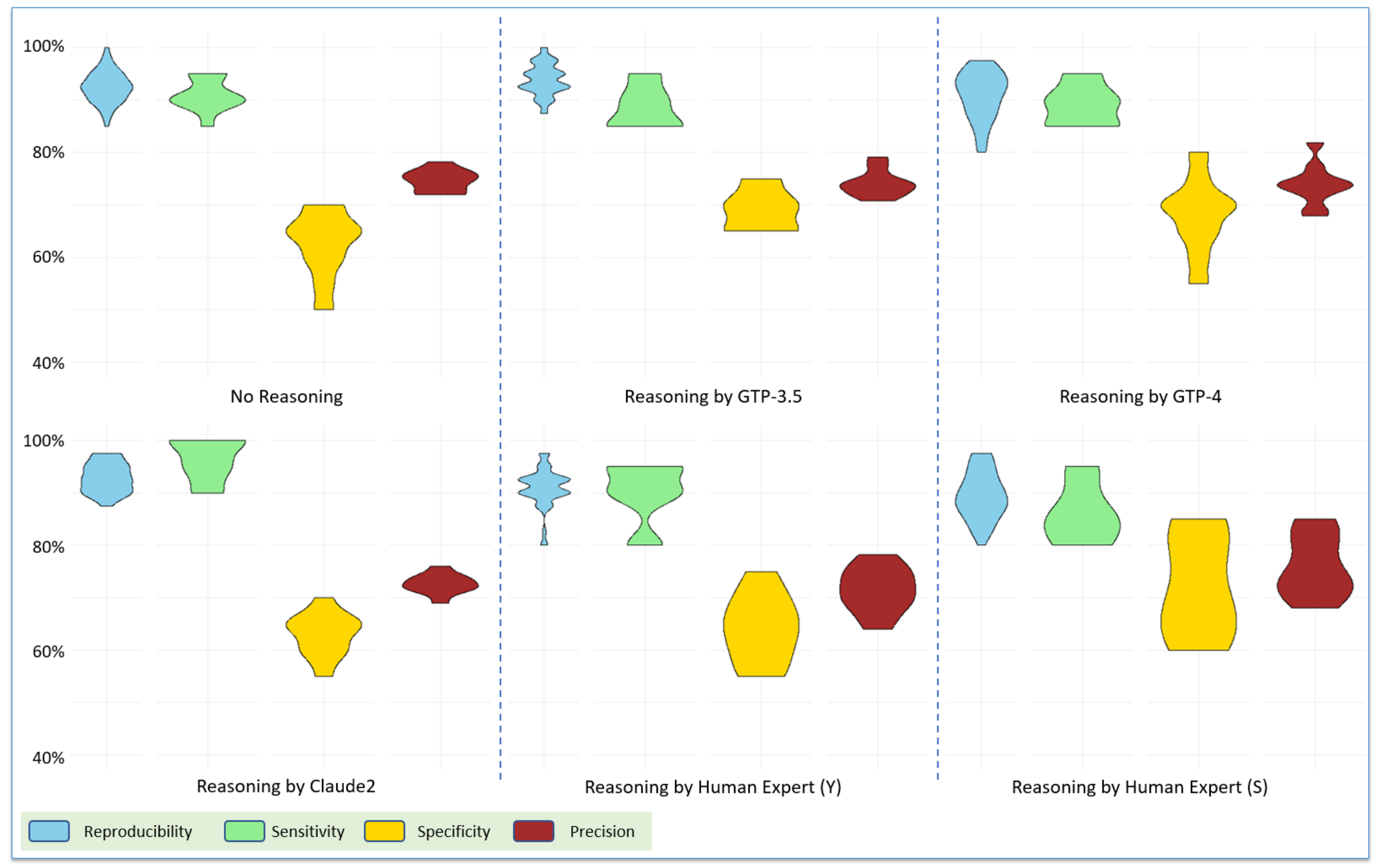

Figure 3. Comparison of different reasoning sets by human experts and LLMs. GPT-3.5 was used for the prediction using 5-shot learning.

Unexpectedly, the introduction of human-generated reasoning did not yield a discernible enhancement; instead, we observed a broader range in metrics, indicating a potential decrease in performance compared to AI-generated reasoning. These findings underscore the nuanced dynamics within the impact of reasoning explanations of model performance.

Impact of Prompts

After establishing the roles of the models and providing background information with examples, we tested various prompts to measure performance differences. These questions below were the only varying parts of the prompts.

Both GPT-3.5 and GPT-4 were used to predict the relevance of the abstracts using identical prompts. The results showed that prompt #3 achieved the best performance for GPT-3.5, while prompt #1 was the best-performing for GPT-4.

Prompt1: *With this new abstract, please first predict whether the abstract is relevant to my chlorine safety study or at least contains some information indicating that reviewing the full paper would be beneficial for my study. Provide me with the label. Then, describe your reasoning or thought process.*

Prompt2: *With this new abstract, please begin by predicting whether the new abstract is **directly** relevant to my chlorine safety study or not and provide me with the label. Then describe your reasoning or thought process.*

Prompt3: *Read this abstract, decide whether it is a study related to the risk of chlorine (or its products) safety. Please provide me labels.*

Prompt4: *Read this abstract, decide whether it reports the toxicity of chlorine gas or other compounds releasing chlorine in its application.*

Prompt5: *Read this abstract, decide whether it reports the safety of chlorine or other compounds releasing chlorine in its application.*

Table 1. The classification performances of GPT-3.5 and GPT-4 based on distinct prompts

Prompt	Predicting Model	Sensitivity	Specificity	Precision	F1-score	Accuracy
#1	GPT-3.5	95%	45%	63.3%	76%	70%
#2	GPT-3.5	95%	50%	66.1%	78.8%	72.5%
		97.5%		72.5%		
#4	GPT-3.5	87.5%	65%	72%	80%	77.5%
#5	GPT-3.5	97.5%	55%	67.9%	79.2%	75%
#1	GPT-4	95%	62.5%	71.7%	82.6%	80%
#2	GPT-4	90%	65%	73.1%	81.7%	78.8%
#3	GPT-4	90%	65%	72%	80%	77.5%
#4	GPT-4	60%	85%	80.6%	69.4%	72.5%
#5	GPT-4	80%	70%	72.7%	76.7%	75%

Impact of Different Reasoning Sets

Instead of providing reviewers a simple Yes or No answer regarding the literature relevance analysis, additional information, such as the reasoning or addressing specific questions reviewers may have, is more useful for further decision-making. Working together with some human exports, we designed 15 questions for LLMs to answers. For example:

- Does the abstract discuss chlorine or its byproducts?
- Does the abstract mention health effects cause by chlorine exposure?
- Do the study results have implications for human health?

These questions helped extract more information from the contexts and better assist reviewers in their daily work. Interestingly, however, the individual questions did not contribute to the relevance prediction.

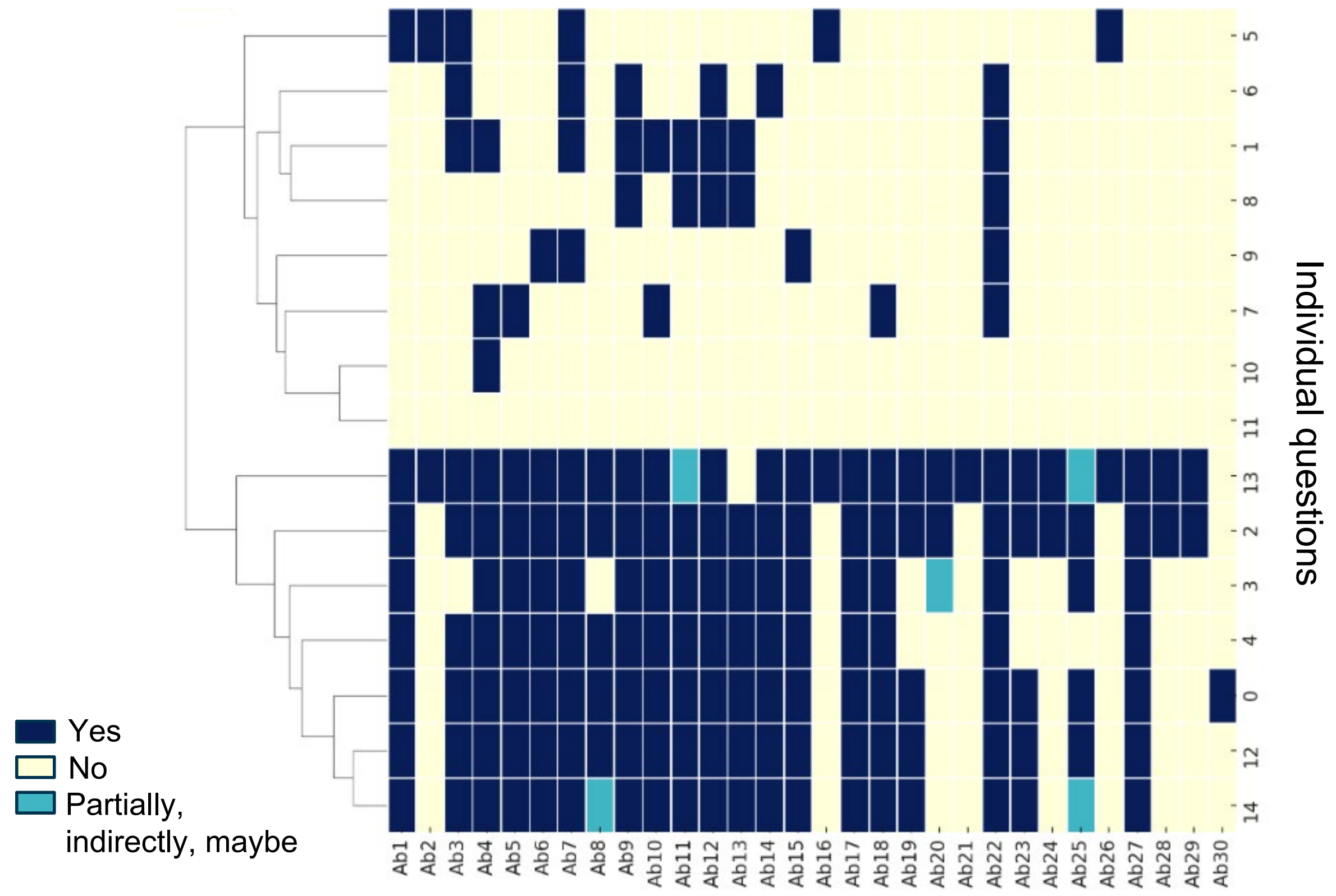

Figure 4. Hierarchical clustering of abstract relevance predictions using individual questions. Question #14 directly asked whether the abstract was relevant or irrelevant. Abstracts (Ab) 1-15 were relevant while 16-30 were irrelevant, according to the truth set.

Conclusion

Large Language Models show great potential in automating systematic literature reviews by improving efficiency, accuracy, and decision-making while reducing the time and effort required for manual screening and assisting the review process.

FDA | **SYMPOSIUM** | SCIENTIFIC COMPUTING + DIGITAL TRANSFORMATION | 2024