ROBoto2: An Interactive System and Dataset for LLM-assisted Clinical Trial Risk of Bias Assessment

Anthony Hevia[1*] **Sanjana Chintalapati**[1*] **Veronica Ka Wai Lai**[2]
Thanh Tam Nguyen[3] **Wai-Tat Wong**[4] **Terry Klassen**[5] **Lucy Lu Wang**[1]
[1]University of Washington [2]The Hospital for Sick Children [3]University of Bologna
[4]The Chinese University of Hong Kong [5]University of Saskatchewan
{hevia, lucylw}@uw.edu

Abstract

We present ROBoto2, an open-source, web-based platform for large language model (LLM)-assisted risk of bias (ROB) assessment of clinical trials. ROBoto2 streamlines the traditionally labor-intensive ROB v2 (ROB2) annotation process via an interactive interface that combines PDF parsing, retrieval-augmented LLM prompting, and human-in-the-loop review. Users can upload clinical trial reports, receive preliminary answers and supporting evidence for ROB2 signaling questions, and provide real-time feedback or corrections to system suggestions. ROBoto2 is publicly available at https://roboto2.vercel.app/, with code and data released to foster reproducibility and adoption. We construct and release a dataset of 521 pediatric clinical trial reports (8954 signaling questions with 1202 evidence passages), annotated using both manually and LLM-assisted methods, serving as a benchmark and enabling future research. Using this dataset, we benchmark ROB2 performance for 4 LLMs and provide an analysis into current model capabilities and ongoing challenges in automating this critical aspect of systematic review.[1]

1 Introduction

Clinical trials, especially when aggregated in systematic reviews, provide the highest quality of evidence for clinical care. While many steps in the systematic review pipeline have seen increasing automation (Marshall and Wallace, 2019; Khalil et al., 2021; Alshami et al., 2023), especially with the advent of LLMs and associated technology, assessing the quality of evidence in individual trials, specifically evaluating *risk of bias* (ROB), remains a critical and time-consuming bottleneck.

The Cochrane Risk of Bias tool version 2 (ROB2)[2] standardizes evaluation by asking 22 signaling questions over 5 domains and computing an overall judgment about risk of bias. However, applying ROB2 is time-consuming, taking trained reviewers 30+ minutes per clinical trial report. This limits scalability for large systematic reviews synthesizing hundreds or thousands of trials.

Previous systems such as RobotReviewer (Marshall et al., 2016) and others (Marshall et al., 2014) explored automating an earlier version of the ROB assessment (ROB) via supervised models, but practical, high-quality automation for ROB2 remains elusive. We therefore introduce ROBoto2, a web-based platform supporting human-AI collaborative ROB2 assessment. ROBoto2 integrates PDF parsing, within-document evidence retrieval, LLM prompting, and ROB2 logic to provide initial answers and rationales for each signaling question. Experts can accept, modify, or override suggestions, with feedback captured for future improvement. Using ROBoto2, our medical collaborators conducted ROB2 assessments on 521 pediatric clinical trials—245 via fully manual review and 276 using the LLM-assisted workflow—yielding a new dataset for benchmarking and research.

We evaluate retrieval methods and four LLMs (Llama-3.3-70B-Instruct, GPT-3.5-Turbo, GPT-4o, and Claude 3.5-Sonnet) on the 245 manual assessments subset, finding that LLMs remain overly conservative compared to human reviewers, frequently opting for high-risk or “No Information” judgments even when evidence is present. Larger context windows and more retrieved evidence somewhat mitigate these tendencies, but fully automated, accurate ROB2 assessment remains challenging.

To summarize, we contribute the following:

- We introduce the ROBoto2 system, a public web tool (code and API available) that supports a human-AI collaborative pipeline for clinical trial ROB2 assessment; the system integrates document preprocessing, passage retrieval, LLM prompting, and interactive expert review;

[*] denotes equal contribution

[1] Dataset and code at https://github.com/larchlab/ROBoto2

[2] https://methods.cochrane.org/bias/resources/rob-2-revised-cochrane-risk-bias-tool-randomized-trials

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 12–25
November 4-9, 2025 ©2025 Association for Computational Linguistics

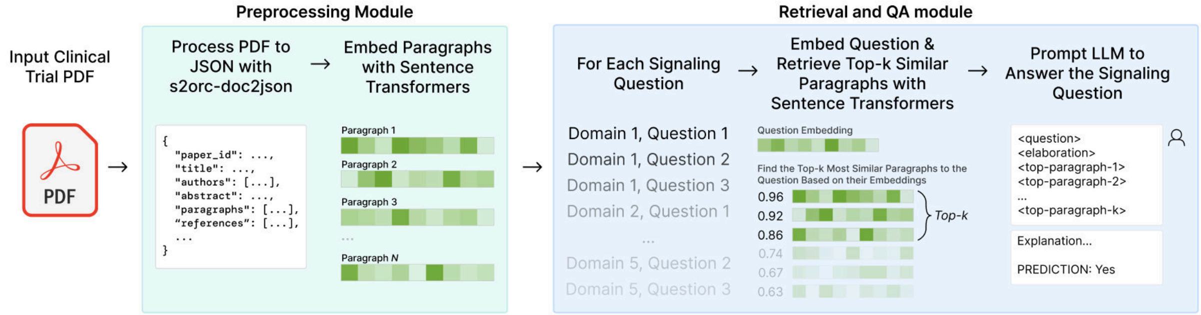


Figure 1: ROBOTO2 system pipeline. Given a clinical trial PDF as input, ROBOTO2 first preprocesses the document to extract and embed paragraphs. Then, a QA module iterates through all of the questions of the ROB2 assessment to identify evidence passages and prompt GPT3.5 to answer the question based on the retrieved evidence.

- We release a dataset of 521 ROB2 assessments (8954 questions; 1202 evidence passages), including both manual and LLM-assisted annotations by medical experts, conducted in the context of an ongoing, real-world systematic review of pediatric clinical trial literature;
- We benchmark retrieval strategies and 4 LLMs on this dataset, providing the first evaluation of LLM-assisted ROB2 assessment. Our analysis highlights current model limitations and directions for future improvement.

2 Related Work

Automating systematic review Prior work on automating systematic reviews have investigated ways to automate the retrieval of relevant papers on a review topic (Choong et al., 2014; Portenoy and West, 2020; van de Schoot et al., 2021), gauging the quality of clinical trials via risk of bias assessment (Marshall et al., 2014, 2016; Suster et al., 2021), extracting PICO (population, intervention, comparator, outcome) elements (Wallace et al., 2016; Nye et al., 2018; Jin and Szolovits, 2018; Hu et al., 2023), extracting numerical results (Yun et al., 2024; Naik et al., 2024), classifying the direction of evidence, also called evidence inference (Lehman et al., 2019; DeYoung et al., 2020), as well as synthesizing and summarizing results across different studies (Wallace et al., 2020; DeYoung et al., 2021; Wang et al., 2022; Sanchez-Graillet et al., 2022; Shaib et al., 2023). Our work builds upon this prior work, especially towards assessing the quality of trials via LLM-assisted risk of bias analysis, extending to v2 of the ROB tool.

ROB analysis The ROB assessment questionnaire from Higgins et al. (2011) and Sterne et al. (2019) can be used to determine the extent to which randomized control trials are at risk of bias. Suster et al. (2021) provide quality ratings for bodies of evidence, and found that some risk factors for quality have good accuracy when automatically assessed, while others do not due to data scarcity. RobotReviewer (Marshall et al., 2016) introduced a system that automatically assigns ROB categorizations to randomized control trials using a trained language model. We extend this work by (i) introducing a dataset corresponding to the newer and more reliable version of the ROB tool (ROB2) (Sterne et al., 2019), (ii) creating an annotation system geared towards supporting a researcher in the loop (Jardim et al., 2022), which leverages in-document retrieval and LLMs to answer signaling questions and identify rationales from the source articles, and (iii) conducting experiments and analysis demonstrating the performance and limitations of current LLMs in supporting this task.

3 Background

We measure risk of bias of randomized trials using the Cochrane ROB2 tool.[3] The ROB2 tool assesses risk along five domains that can introduce bias into the results of a randomized trial:

D1: Randomization process
D2: Deviations from intended interventions
D3: Missing outcome data
D4: Measurement of the outcome
D5: Selection of the reported result

Each domain consists of 3-7 signaling questions, which help gather information and contribute to the final risk classification. For example, this D2 question assesses bias due to unblinded treatment

[3] ROB2 replaces its predecessor ROB after a formal evaluation identified areas for improvement (Sterne et al., 2019). ROB2 includes questions measuring newly identified ways that bias arise in randomized trials.

assignment: "Were participants aware of their assigned intervention during the trial?" There are five response options for each signaling question: (1) Yes; (2) Probably yes; (3) Probably no; (4) No; and (5) No information. All questions in App. A.

The ROB2 assessment is hierarchical. Signaling question responses for each domain first contribute to domain-level judgments for risk of bias, then domain-level judgments provide the basis for an overall risk of bias judgment. The tool provides flowcharts for computing the risk of each domain based on the answers to signaling questions (e.g., Figure 3 in App. A) as well as for computing overall risk. Domain-level and overall risk are assessed as either low risk, some concerns, or high risk.

In ROBOTO2, we model the ROB2 assessment as a document-level question-answering (QA) task. We use each signaling question as a query to retrieve relevant evidence passages from the trial report, then generate an answer based on the retrieved evidence. Answers are validated by a user who is conducting the assessment. The final risk assessment is produced by implementing the flowchart logic provided by the ROB2 tool.

4 ROBOTO2 System Pipeline

Figure 1 shows the ROBOTO2 pipeline. A user uploads a PDF of a clinical trial report. We (i) preprocess it to extract paragraphs of text; (ii) embed each paragraph using a document embedding model and index them for within-document retrieval; and then, for each signaling question from the ROB2 assessment tool, we (iii) embed the signaling question, retrieve the top-k similar paragraphs from the paper, and prompt an LLM to answer the question using the top-k paragraphs as context. We also experiment with providing all passages of text (full paper) as input for models with large input context windows. Details follow.

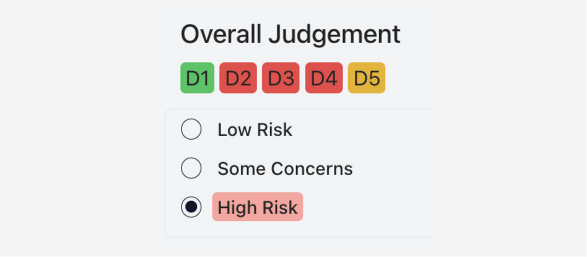
Preprocessing PDFs We convert each PDF into standardized JSON format using the S2ORC-doc2json library (Lo et al., 2020).[4] The output JSON contains a list of paragraphs in the paper, their section headers, and metadata elements such as the paper's title, authors, and abstract.

Embedding paragraphs for retrieval We compute embeddings for each paragraph in the uploaded paper using Sentence-Transformers all-MiniLM-L6-v2 (Reimers and Gurevych, 2019) and construct a key-value store for retrieval. Evaluation of the retriever and alternate methods is described in App. B. For each signaling question, we embed the question text using the same model and use cosine similarity to identify the top-k paragraphs to use as context for the QA module.

Answering signaling questions We then prompt an LLM to answer each ROB2 signaling question using an instruction prompt based on the ROB2 questionnaire and with the retrieved evidence paragraphs as context. Prompt templates and a complete example are given in App. C.

Collecting user feedback When conducting assessments with ROBOTO2, users can modify and provide feedback on all aspects of the assessment. While we show the top-3 retrieved paragraphs by default, users can add further paragraphs by selecting from the JSON parse. They can also provide feedback on the accuracy of retrieved passages via up- or downvotes, and modify the model-predicted answers and rationales (called "Explanation" in ROBOTO2). ROBOTO2 retains the original LLM responses and rationales, as well as the versions confirmed or edited by expert users. We include these user modifications as part of our dataset.

Domain-level and overall judgments We implement the logic provided by the ROB2 flowcharts (e.g., Figure 3 in App. A) to produce domain-level and overall risk of bias judgments. We visualize these at the end of each domain section and as a summative visualization when users complete the ROB2 assessment, e.g., three high risk domain-level judgments yields a "High Risk" overall rating visualized as follows:



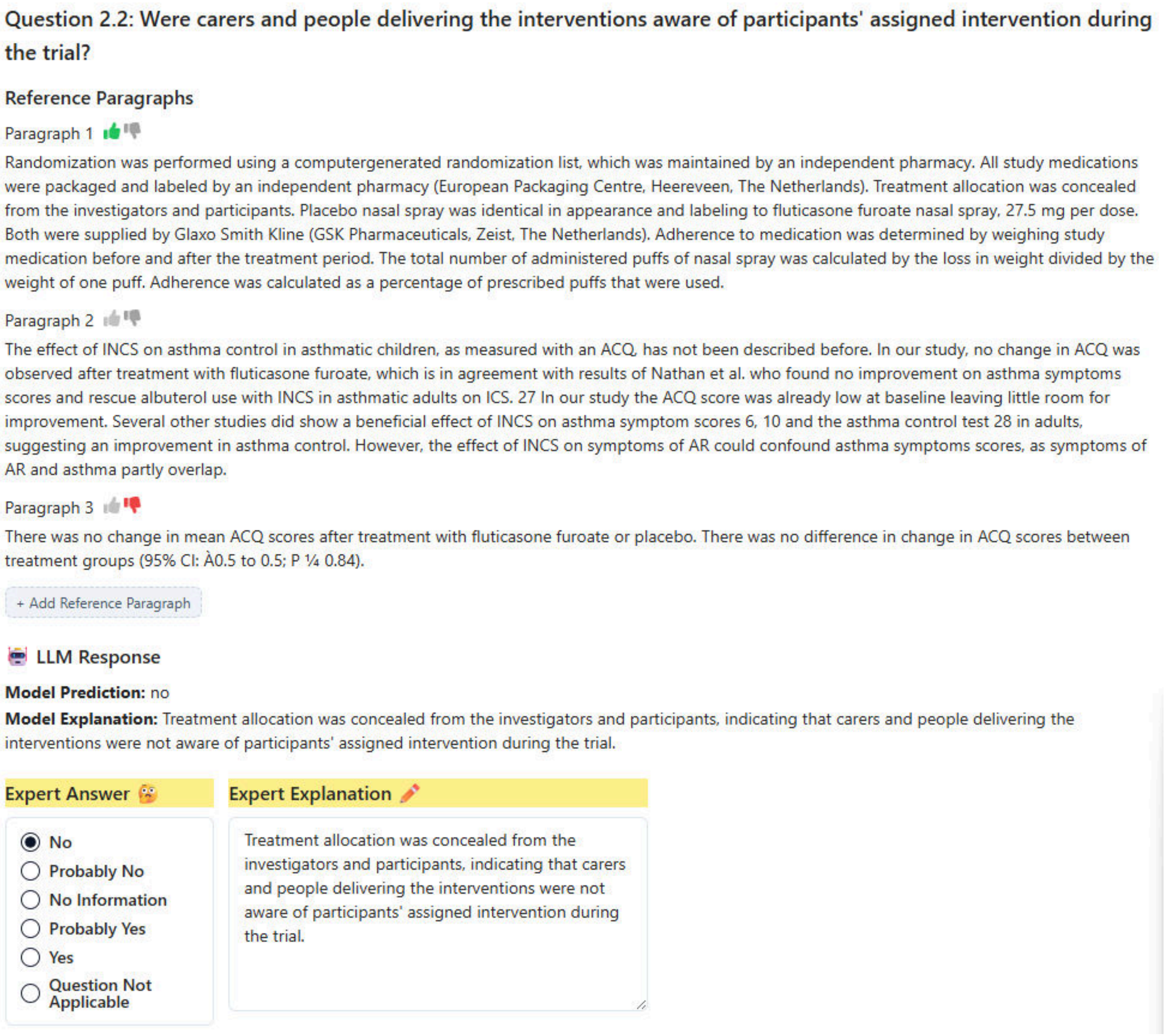
Overall Judgement

D1 D2 D3 D4 D5

○ Low Risk
○ Some Concerns
● High Risk

System implementation Part of the ROBOTO2 workflow is shown in Figure 2 (web app at https://roboto2.vercel.app/). The web interface is written in React and Typescript. It leverages Transformers.js for client-side embedding and retrieval, and a back-end API in Python and FastAPI for document parsing and calls to LLM services.

[4] https://github.com/allenai/s2orc-doc2json



Question 2.2: Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?

Reference Paragraphs

Paragraph 1 👍👎

Randomization was performed using a computergenerated randomization list, which was maintained by an independent pharmacy. All study medications were packaged and labeled by an independent pharmacy (European Packaging Centre, Heereveen, The Netherlands). Treatment allocation was concealed from the investigators and participants. Placebo nasal spray was identical in appearance and labeling to fluticasone furoate nasal spray, 27.5 mg per dose. Both were supplied by Glaxo Smith Kline (GSK Pharmaceuticals, Zeist, The Netherlands). Adherence to medication was determined by weighing study medication before and after the treatment period. The total number of administered puffs of nasal spray was calculated by the loss in weight divided by the weight of one puff. Adherence was calculated as a percentage of prescribed puffs that were used.

Paragraph 2 👍👎

The effect of INCS on asthma control in asthmatic children, as measured with an ACQ, has not been described before. In our study, no change in ACQ was observed after treatment with fluticasone furoate, which is in agreement with results of Nathan et al. who found no improvement on asthma symptoms scores and rescue albuterol use with INCS in asthmatic adults on ICS. 27 In our study the ACQ score was already low at baseline leaving little room for improvement. Several other studies did show a beneficial effect of INCS on asthma symptom scores 6, 10 and the asthma control test 28 in adults, suggesting an improvement in asthma control. However, the effect of INCS on symptoms of AR could confound asthma symptoms scores, as symptoms of AR and asthma partly overlap.

Paragraph 3 👍👎

There was no change in mean ACQ scores after treatment with fluticasone furoate or placebo. There was no difference in change in ACQ scores between treatment groups (95% CI: À0.5 to 0.5; P ¼ 0.84).

LLM Response

Model Prediction: no

Model Explanation: Treatment allocation was concealed from the investigators and participants, indicating that carers and people delivering the interventions were not aware of participants' assigned intervention during the trial.

Expert Answer

● No
○ Probably No
○ No Information
○ Probably Yes
○ Yes
○ Question Not Applicable

Expert Explanation

Treatment allocation was concealed from the investigators and participants, indicating that carers and people delivering the interventions were not aware of participants' assigned intervention during the trial.

Figure 2: Screenshot of ROBOTO2 assisting with a question from Domain 2. The user can modify the model-provided answer and explanation and rate reference paragraphs.

5 Annotated Evaluation Dataset

Our dataset consists of 521 ROB2 assessments (245 completed using the Cochrane ROB2 Excel tool and 276 with LLM assistance via ROBOTO2). These ROB2 assessments were conducted as part of an independent research project aiming to systematically evaluate the risk of bias of all child health clinical trials; we re-purpose the data in this work to explore the role and feasibility of LLMs in supporting this aspect of systematic review.

Annotation procedure An initial corpus of child health clinical trial reports was constructed by searching the Cochrane Central Register of Controlled Trials, filtering for pediatric clinical trials based on the procedures described in Boluyt et al. (2008), and identifying 2334 matching clinical trial reports published 1991-2020. We sampled trial reports from this corpus for annotation.

For a subset of 245 reports, a group of expert raters completed assessments manually using the Cochrane tool, an Excel sheet with macros implementing the logic of the ROB2 assessment. To support judgments, annotators identified evidence passages manually from paper PDFs for a subset of questions and copied these into the Excel sheet. For each clinical trial, the data consists of the paper PDF for the trial report, as well as judgments for each signaling question, evidence passages extracted from the paper for a subset of questions, domain-level judgments, and the overall risk assessment score. Five expert annotators participated in annotations, and all annotators have graduate degrees in public health, epidemiology, medical sciences, or clinical practice, as well as experience conducting systematic reviews. This set of 245

	Low risk	Some concerns	High risk
Domain 1	234	243	44
Domain 2	287	171	63
Domain 3	450	35	35
Domain 4	406	60	54
Domain 5	332	272	34
Paper-level	64	301	156

Table 1: Distribution of domain- and paper-level risk of bias judgments in our dataset.

papers, annotated using the current gold-standard ROB2 review process, is used for all LLM evaluation and comparison reported in this paper.

An additional 276 papers were annotated separately by two of the five annotators using ROBOTO2, with LLM assistance. The version of ROBOTO2 used for annotations (collected during early 2024) used retrieval and GPT-3.5 (gpt-3.5-turbo-0125) as the answer model.[5] Because this sample of 276 papers was annotated with LLM assistance, we withheld them from the final evaluation of ROBOTO2 as described in Section 6, but include them in the published dataset to support future work.

Dataset statistics The distribution of domain-level and overall risk of bias judgments in the full dataset are provided in Table 1. Distributions of answered signaling questions and evidence paragraphs in the manually-annotated subset are shown at the top of Table 2.

Inter-rater reliability To assess inter-rater reliability, 20 papers (totaling 440 signaling questions) are independently annotated by two annotators using ROBOTO2. We aggregate answers into the following classes: Yes/Probably Yes, No/Probably No, No Information, and N/A, when a question is skipped by ROB logic. Four-class Cohen's Kappa is 0.40, indicating fair to moderate agreement. This is consistent with prior work showing slight to moderate agreement (Fleiss' Kappa of 0.45 at the domain level) among experienced raters (Minozzi et al., 2020, 2021); it reflects well-documented challenges of applying the complex ROB2 tool (Nejadghaderi et al., 2024), as reviewers can differ in their interpretation of ambiguous scenarios and their thresholds for assigning risk levels. Further commentary in App. D.

6 Experimental Settings

We evaluate 4 models: GPT-3.5-Turbo, GPT-4o, Claude 3.5-Sonnet, and Llama-3.3-70B-Instruct. All models receive the same prompt and inputs (App. C). For all experiments, we aggregate labels and outputs into three classes: Yes/Probably Yes (Y/PY), No/Probably No (N/PN), and No Information (NI), and report micro-F1 at each domain level along with micro- and macro-averages across all signaling questions (Table 2).

Within-document retrieval We evaluate two retrieval methods: BM25 (Robertson et al., 1994) and paragraph embeddings using Sentence-Transformers (Reimers and Gurevych, 2019). Each signaling question has a max of one gold evidence passage in the dataset; we report recall@k for k=1,3,5,10 for all retrieval methods (Table 4). Detailed results and evaluation of the retrieval methods can be found in App B. In all cases, we use the questions from the ROB2 assessment as the query, and paragraphs from the clinical trial paper as the documents to retrieve. In the publicly available version of ROBOTO2, all-MiniLM-L6-v2 and k=3 were selected as these settings achieved competitive performance at low cost.

Prompting LLMs for QA We evaluate all models in a zero-shot setting with oracle evidence (providing the human-labeled evidence passage), as well as with the top-k retrieved evidence (with k=1,3,5), and the full paper setting for models with sufficient input context window sizes (all but GPT-3.5-Turbo). In the ROB2 assessment, each signaling question includes elaboration text that expands on when each answer should be chosen for that question; we provide this elaboration in the instructions for all prompting settings (example in App. C). We also conduct several experiments with in-context learning (Brown et al., 2020), which suggested minimal gains from the zero-shot setting; these results are reported in App E.

7 Results & Discussion

Results for all experimental settings are provided in Table 2. We analyze the model results and user statistics collected during ROBOTO2 annotations below (full statistics in App. G).

Room for improvement The best performing model (Claude 3.5-Sonnet with the full paper as context) achieved a micro-F1 of 0.71, highlighting considerable room for improvement. All evaluated

[5] In experiments, other LLMs and full-text context demonstrate better performance, but these were reasonable configurations at the time of annotation. Our public web interface supports the use of alternate LLMs.

Model	Retrieval	D1	D2	D3	D4	D5	Micro-Avg	Macro-Avg
n-oracle	-	197	124	37	73	11	-	-
n-total	-	750	1278	598	1027	750	-	-
Baseline w/ oracle evidence paragraphs								
Llama-3.3-70B-Instruct	Oracle	0.67	0.55	0.35	0.81	0.45	0.62	0.67
GPT-3.5-Turbo	Oracle	0.81	0.66	0.67	0.82	0.57	0.61	0.71
GPT-4o	Oracle	0.75	0.78	0.68	0.92	0.54	0.64	0.73
Claude 3.5-Sonnet	Oracle	0.75	0.81	0.66	0.92	0.59	0.66	0.75
Retrieved evidence paragraphs								
Llama-3.3-70B-Instruct	k=1	0.83	0.61	0.42	0.80	0.38	0.49	0.61
GPT-3.5-Turbo	k=1	0.81	0.73	0.69	0.74	0.66	0.58	0.73
GPT-4o	k=1	0.68	0.65	0.58	0.81	0.75	0.55	0.69
Claude 3.5-Sonnet	k=1	0.69	0.68	0.66	0.87	0.76	0.60	0.73
Llama-3.3-70B-Instruct	k=3	0.87	0.69	0.66	0.81	0.35	0.55	0.68
GPT-3.5-Turbo	k=3	0.82	0.69	0.68	0.73	0.59	0.55	0.70
GPT-4o	k=3	0.75	0.72	0.73	0.82	0.72	0.60	0.75
Claude 3.5-Sonnet	k=3	0.75	0.75	0.76	0.87	**0.77**	0.65	0.78
Llama-3.3-70B-Instruct	k=5	0.87	0.72	0.70	0.81	0.28	0.55	0.69
GPT-3.5-Turbo	k=5	**0.82**	0.69	0.64	0.71	0.56	0.53	0.68
GPT-4o	k=5	0.78	0.74	0.73	0.83	0.69	0.62	0.75
Claude 3.5-Sonnet	k=5	0.78	0.79	0.78	0.86	**0.77**	0.67	0.80
Full paper as input								
Llama-3.3-70B-Instruct	Full Paper	0.88	0.81	0.79	0.79	0.32	0.61	0.72
GPT-4o	Full Paper	0.80	0.82	0.77	0.85	0.66	0.66	0.78
Claude 3.5-Sonnet	Full Paper	0.81	**0.84**	**0.80**	**0.88**	**0.77**	**0.71**	**0.82**

Table 2: For all settings, we report micro-averaged domain-level F1 along with micro- and macro-averaged F1 across all signaling questions. The n-oracle is the number of instances where annotators identified an evidence passage, while n-total is the total number of signaling questions answered for that domain.

models achieve strong results in D1, where questions are more likely to be answerable based on text in the trial reports. Performance in D2 and D3 is mixed, as these may require interpreting numerical data (e.g., calculating attrition rates from recruitment and result numbers). D5 is similarly challenging for models as these questions may require knowledge of external clinical resources and guidelines.

Increasing context generally improves performance for most models while reducing accuracy for GPT-3.5-Turbo. In some cases, models with retrieval can surpass oracle retrieval performance, likely due to incomplete evidence labeling in our dataset, indicating that relevant information exists beyond annotator-selected passages. Few-shot prompting does not appear to outperform zero-shot prompting in our experimental results (App E).

Limited utility for fully-automated ROB assessment Model performance cannot be substituted for human judgment in ROB assessments, and we recommend that humans remain in the loop. Conservative question-level model judgments compound to conservative domain- and paper-level judgments, where the fully-automated pipeline judges most papers as having "some concerns" or "high risk" even when human raters did not. Human raters assessed 47 of 276 trials as high risk while the LLM-only pipeline assessed 101 as high risk. Error analysis (App. F) reveals that the strongest models tend to over-select "No Information," which may reflect cautiousness gained from safety and alignment training.

ROBOTO2 supports human review and editing We compute detailed metrics for the 276 ROB2 assessments annotated using ROBOTO2, including the number of times annotators accept the model's answers and explanations directly versus change them, and the number of retrieved evidence passages marked as good (offering evidence to support an answer) versus bad (irrelevant). Annotators provide their own answer (42.4%) and edit rationales (28.7%) around half the time, rather than use the answer (57.6%) and rationales (71.3%) provided by the model (Table 7). For evidence passages, 615 total up/downvotes are collected (out of 3370 retrieved passages), of which 78.0% are positive feedback. We provide all feedback in our dataset to support future model development. Detailed statistics can be found in App. G.

8 Conclusion

Assessing the quality of clinical trials is an important step to weighing their evidence in clinical decision-making. To support this, we introduce the ROBoto2 system to assist researchers in conducting risk of bias assessment for clinical trials with LLM support, along with associated code for running the web interface. We also release a dataset of 521 complete ROB analyses (8954 signaling questions with 1202 evidence passages) of child health clinical trial reports. We hope our system and dataset will promote better LLM applications for risk of bias assessment, and that access to this assisted annotation tool can enable quicker completion of ROB assessments and reduce the labor and costs around systematic literature reviews.

9 Limitations

Viewing model outputs could potentially bias annotations ROBoto2 is designed to expose all intermediate and final model outputs, and allows expert annotators to change any part of the model output. While we can compute the number of changes made, we cannot guarantee that seeing model outputs does not influence annotator responses. Prior work has shown that human annotators may demonstrate anchoring bias when exposed to LLM assistance during annotation (Choi et al., 2024), leading to discrepancies in downstream label distributions. We leave the measurement of this bias in the ROB setting to future work.

Dataset imbalance Though it reflects real-world ROB2 assessments, our dataset is unbalanced. The majority of papers are assessed as having some concerns, with fewer papers of low or high risk. This likely biases the evaluation of our system, similar to what was observed by Suster et al. (2021). Related, some signaling questions have very sparse annotations (especially those that depend on cascading logic) or are biased in terms of answer distribution (almost always one of the answer labels).

Among manually conducted reviews, annotator-provided evidence paragraphs are only available for a small portion of signaling questions, unbalanced across domains; D1 has the most signaling questions with evidence passages, while D5 has very few. Our retrieval methods as well as Oracle results are only reported on this biased subset, and may not accurately represent performance on sparsely annotated questions and domains.

Potential gains in quality or efficiency We hypothesize that ROBoto2 may either help to save time or improve the quality of ROB assessments. Qualitatively, the annotation team reported that ROBoto2 offers an opportunity to enhance evidence and rationale coverage, as the time savings on the ROB assessment itself were repurposed to judge and retrieve relevant evidence. However, the impact of such repurposing of effort on the quality of the resulting assessments was not explicitly measured in our system and should be confirmed and studied in future work.

Diversity of language models in experiments Our experiments do not include any reasoning models such as OpenAI's o3, Anthropic's Claude 4-Sonnet-Thinking, and DeepSeek-R1 (DeepSeek-AI, 2025). Future work could explore whether these models improve current performance in terms of both answer classification accuracy and generated rationales.

Acknowledgements

We thank the following individuals for producing parts of the ROB dataset we use for evaluation: Bernadette Zakher, Shannon Sim, Michele Dyson, Banke Oketola, and Aneet Saran, from the University of Victoria, University of Alberta, and University of Manitoba. This work is supported in part by the University of Washington Information School Strategic Research Fund and the University of Washington eScience Institute's Azure Cloud Credits for Research and Teaching.

References

Ahmad Alshami, Moustafa Elsayed, Eslam Ali, Abdelrahman E. E. Eltoukhy, and Tarek M. Zayed. 2023. Harnessing the power of chatgpt for automating systematic review process: Methodology, case study, limitations, and future directions. *Syst.*, 11:351.

Nicole Boluyt, Lisa Tjosvold, Carol Lefebvre, Terry P Klassen, and Martin Offringa. 2008. Usefulness of systematic review search strategies in finding child health systematic reviews in medline. *Archives of pediatrics & adolescent medicine*, 162 2:111–6.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Alexander S. Choi, Syeda Sabrina Akter, JP Singh, and Antonios Anastasopoulos. 2024. The LLM effect:

Are humans truly using LLMs, or are they being influenced by them instead? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22032–22054, Miami, Florida, USA. Association for Computational Linguistics.

Miew Keen Choong, Filippo Galgani, Adam G. Dunn, and Guy Tsafnat. 2014. Automatic evidence retrieval for systematic reviews. *Journal of Medical Internet Research*, 16.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MSˆ2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Julian PT Higgins, Douglas G Altman, Peter C Gotzsche, Peter Juni, David Moher, Andrew D Oxman, Jelena Savovic, Kenneth F Schulz, Laura Weeks, and Jonathan AC Sterne. 2011. The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343.

Yan Hu, Vipina Kuttichi Keloth, Kalpana Raja, Yong Chen, and Hua Xu. 2023. Towards precise pico extraction from abstracts of randomized controlled trials using a section-specific learning approach. *Bioinformatics*, 39.

Patricia Sofia Jacobsen Jardim, Christopher James Rose, Heather Melanie R Ames, J. Meneses Echavez, Stijn Van de Velde, and Ashley Elizabeth Muller. 2022. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Medical Research Methodology*, 22.

Di Jin and Peter Szolovits. 2018. Pico element detection in medical text via long short-term memory neural networks. In *Workshop on Biomedical Natural Language Processing*.

Hanan Khalil, Daniel Ameen, and A. Zarnegar. 2021. Tools to support the automation of systematic reviews: A scoping review. *Journal of clinical epidemiology*.

Eric Lehman, Jay DeYoung, Regina Barzilay, and Byron C. Wallace. 2019. Inferring which medical treatments work from reports of clinical trials. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3705–3717, Minneapolis, Minnesota. Association for Computational Linguistics.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Michael Kinney, and Daniel S. Weld. 2020. S2orc: The semantic scholar open research corpus. In *Annual Meeting of the Association for Computational Linguistics*.

Iain J Marshall, Joel Kuiper, and Byron C Wallace. 2016. Robotreviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201.

Iain James Marshall, Joël Kuiper, and Byron C. Wallace. 2014. Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, 19:1406–1412.

Iain James Marshall and Byron C. Wallace. 2019. Toward systematic review automation: a practical guide to using machine learning tools in research synthesis. *Systematic Reviews*, 8.

Silvia Minozzi, Michela Cinquini, Silvia Gianola, Marien González-Lorenzo, and Rita Banzi. 2020. The revised cochrane risk-of-bias tool for randomised trials (rob 2) showed low inter-rater reliability and challenges in its application. *Journal of clinical epidemiology*.

Silvia Minozzi, Kerry Dwan, Francesca Borrelli, and Graziella Filippini. 2021. Reliability of the revised cochrane risk-of-bias tool for randomised trials (rob2) improved with the use of implementation instruction. *Journal of clinical epidemiology*.

Aakanksha Naik, Bailey Kuehl, Erin Bransom, Doug Downey, and Tom Hope. 2024. CARE: Extracting experimental findings from clinical literature. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 4580–4596, Mexico City, Mexico. Association for Computational Linguistics.

Seyed Aria Nejadghaderi, Maryam Balibegloo, and Nima Rezaei. 2024. The cochrane risk of bias assessment tool 2 (rob 2) versus the original rob: A perspective on the pros and cons. *Health Science Reports*, 7.

Benjamin Nye, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

Jason Portenoy and Jevin D. West. 2020. Constructing and evaluating automated literature review systems. *Scientometrics*, 125:3233 – 3251.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. 1994. Okapi at trec-3. In *Text Retrieval Conference*.

Olivia Sanchez-Graillet, Christian Witte, Frank Grimm, Steffen Grautoff, Basil Ell, and Philipp Cimiano. 2022. Synthesizing evidence from clinical trials with dynamic interactive argument trees. *Journal of Biomedical Semantics*, 13.

Chantal Shaib, Millicent Li, Sebastian Joseph, Iain Marshall, Junyi Jessy Li, and Byron Wallace. 2023. Summarizing, simplifying, and synthesizing medical evidence using GPT-3 (with varying success). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1387–1407, Toronto, Canada. Association for Computational Linguistics.

Jonathan A. C. Sterne, Jelena Savović, Matthew J. Page, Roy G Elbers, Natalie S Blencowe, Isabelle Boutron, Christopher J Cates, Hung-Yuan Cheng, Mark S. Corbett, Sandra Eldridge, Jonathan R. Emberson, Miguel A. Hernán, Sally Hopewell, Asbjørn Hrõbjartsson, Daniela R. Junqueira, Peter Juni, Jamie J. Kirkham, Toby J Lasserson, Tianjing Li, Alexandra McAleenan, Barnaby C. Reeves, Sasha Shepperd, Ian Shrier, Lesley A Stewart, Kate Tilling, Ian R. White, Penny F. Whiting, and Julian P. T. Higgins. 2019. Rob 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366.

Simon Suster, Timothy Baldwin, Jey Han Lau, Antonio Jimeno Yepes, David Martinez Iraola, Yulia Otmakhova, and Karin M. Verspoor. 2021. Automating quality assessment of medical evidence in systematic reviews: Model development and validation study. *Journal of Medical Internet Research*, 25.

Rens van de Schoot, Jonathan de Bruin, Raoul Schram, Parisa Zahedi, Jan de Boer, Felix Weijdema, Bianca Kramer, Martijn Huijts, Maarten Hoogerwerf, Gerbrich Ferdinands, Albert Harkema, Joukje E Willemsen, Yongchao Ma, Qixiang Fang, Lars G Tummers, and Daniel L. Oberski. 2021. Asreview: Active learning for systematic reviews.

Byron C. Wallace, Joel Kuiper, Aakash Sharma, Mingxi Zhu, and Iain James Marshall. 2016. Extracting pico sentences from clinical trial reports using supervised distant supervision. *Journal of machine learning research (JMLR)*, 17.

Byron C. Wallace, Sayantani Saha, Frank Soboczenski, and Iain James Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with neural multi-document summarization. *AMIA Annual Symposium proceedings*, 2021:605–614.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022. Overview of MSLR2022: A shared task on multi-document summarization for literature reviews. In *Proceedings of the Third Workshop on Scholarly Document Processing*, pages 175–180, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Hye Sun Yun, David Pogrebitskiy, Iain James Marshall, and Byron C. Wallace. 2024. Automatically extracting numerical results from randomized controlled trials with large language models. *ArXiv*, abs/2405.01686.

A ROB2 Assessment Tool

The Signaling Questions for the Cochrane ROB2 Tool for Randomized Trials are given in Table 3. Note that some questions are *cascading*, and are only answered if previous questions in the domain are answered in a pre-specified way.

Figure 3 reproduces a flowchart from the ROB2 tool that indicates how signaling questions contribute to a domain-level judgment for Domain 4. Based on how these questions are answered, the domain-level judgment can be low risk, some concerns, or high risk. Flowcharts are also provided for the other four domains and are available at https://methods.cochrane.org/risk-bias-2.

B Retriever Evaluation

We experiment with a sparse retriever, BM25 (Robertson et al., 1994), and Sentence-Transformers (Reimers and Gurevych, 2019). We assess retriever performance by varying k, the number of passages retrieved and provided to the QA reader module. For models with large context windows, we also experiment with providing the entire paper as context.

All methods are validated using gold evidence paragraphs identified by the annotators in our dataset. Each signaling question has a maximum of one gold evidence passage in the dataset; we report recall@k for k=1,3,5,10 for all retrieval methods (Table 4).

For within-document retrieval, we find that S-BERT successfully retrieves the gold evidence passage at a higher rate than BM25 at comparable k (Table 4). However, prompting models with the full paper achieves the highest overall F1-scores (Table 2). ROBoto2 uses S-BERT for its balance between performance, speed, and enabling models with smaller context windows to be used in the web interface.

Domain	Question
Domain 1: Risk of bias arising from the randomization process	
1.1	Was the allocation sequence random?
1.2	Was the allocation sequence concealed until participants were enrolled and assigned to interventions?
1.3	Did baseline differences between intervention groups suggest a problem with the randomization process?
Domain 2: Risk of bias due to deviations from the intended interventions	
2.1	Were participants aware of their assigned intervention during the trial?
2.2	Were carers and people delivering the interventions aware of participants' assigned intervention during the trial?
2.3	If Y/PY/NI to 2.1 or 2.2: Were there deviations from the intended intervention that arose because of the trial context?
2.4	If Y/PY to 2.3: Were these deviations likely to have affected the outcome?
2.5	If Y/PY/NI to 2.4: Were these deviations from intended intervention balanced between groups?
2.6	Was an appropriate analysis used to estimate the effect of assignment to intervention?
2.7	If N/PN/NI to 2.6: Was there potential for a substantial impact (on the result) of the failure to analyse participants in the group to which they were randomized?
Domain 3: Risk of bias due to missing outcome data	
3.1	Were data for this outcome available for all or nearly all participants randomized?
3.2	If N/PN/NI to 3.1: Is there evidence that the result was not biased by missing outcome data?
3.3	If N/PN to 3.2: Could missingness in the outcome depend on its true value?
3.4	If Y/PY/NI to 3.3: Is it likely that missingness in the outcome depended on its true value?
Domain 4: Risk of bias in measurement of the outcome	
4.1	Was the method of measuring the outcome inappropriate?
4.2	Could measurement or ascertainment of the outcome have differed between intervention groups?
4.3	If N/PN/NI to 4.1 and 4.2: Were outcome assessors aware of the intervention received by study participants?
4.4	If Y/PY/NI to 4.3: Could assessment of the outcome have been influenced by knowledge of intervention received?
4.5	If Y/PY/NI to 4.4: Is it likely that assessment of the outcome was influenced by knowledge of intervention received?
Domain 5: Risk of bias in selection of the reported result	
5.1	Were the data that produced this result analysed in accordance with a pre-specified analysis plan that was finalized before unblinded outcome data were available for analysis?
5.2	Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible outcome measurements (e.g., scales, definitions, time points) within the outcome domain?
5.3	Is the numerical result being assessed likely to have been selected, on the basis of the results, from multiple eligible analyses of the data?

Table 3: Signaling Questions in the Cochrane Risk of Bias Tool for Randomized Trials (ROB2).

Model	R@1	R@3	R@5	R@10
BM25	0.140	0.272	0.367	0.533
S-BERT	0.268	0.455	0.519	0.678

Table 4: Recall@k for our tested retrieval methods.

C Prompting

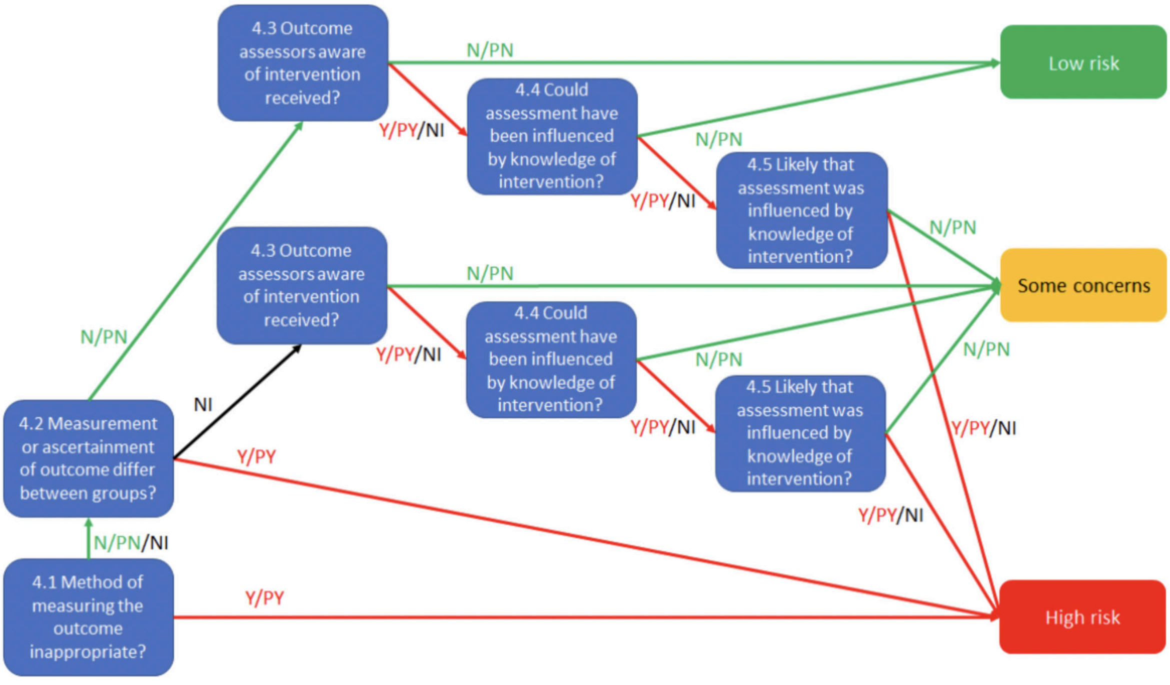
The prompt is formatted as follows:

```
<instruction>
<signaling_question>
<elaboration>
<retrieved_paragraph_1>
...
<retrieved_paragraph_k>
```

After an instruction to answer the question, we provide the signaling question itself from the ROB2 assessment, as well as additional elaboration text explaining all answer options and when they should be used. These elaborations are adapted from explanations given in the ROB2 tool, and we further augment them such that all possible answer options are represented—not all answers are represented in elaborations from the original ROB2 tool, which we found may bias models towards answers that were. Retrieved context paragraphs are then appended. We instruct the model to make a prediction and generate a rationale for its prediction.

A full example prompt for signaling question 1 in Domain 1 is reproduced below:

```
You are an expert scientific researcher.
```


4.1 Method of measuring the outcome inappropriate?
4.2 Measurement or ascertainment of outcome differ between groups?
4.3 Outcome assessors aware of intervention received?
4.4 Could assessment have been influenced by knowledge of intervention?
4.5 Likely that assessment was influenced by knowledge of intervention?
4.3 Outcome assessors aware of intervention received?
4.4 Could assessment have been influenced by knowledge of intervention?
4.5 Likely that assessment was influenced by knowledge of intervention?
Low risk
Some concerns
High risk
N/PN
Y/PY/NI
N/PN/NI
Y/PY
NI

Algorithm for suggested judgement of risk of bias in measurement of the outcome

Figure 3: Flowchart for how answers to signaling questions contribute to a domain-level judgment for Domain 4 in the ROB2 tool. Reproduced from https://sites.google.com/site/riskofbiastool/welcome/rob-2-0-tool.

You will be given a passage from a scientific paper reporting on a randomized controlled trial along with a question and elaboration of the question. Your task is to return the answer to the question out of the following set of answers: "yes", "no", "probably yes", "probably no", "no information". You should use the given passage to answer the question.

Question: "Was the allocation sequence random?"

Elaboration: "Answer 'Yes' if a random component was used in the sequence generation process. Examples include computer-generated random numbers; reference to a random number table; coin tossing; shuffling cards or envelopes; throwing dice; or drawing lots. Minimization is generally implemented with a random element (at least when the scores are equal), so an allocation sequence that is generated using minimization should generally be considered to be random.

Answer 'No' if no random element was used in generating the allocation sequence or the sequence is predictable. Examples include alternation; methods based on dates (of birth or admission); patient record numbers; allocation decisions made by clinicians or participants; allocation based on the availability of the intervention; or any other systematic or haphazard method.

Answer 'No information' if the only information about randomization methods is a statement that the study is randomized.

In some situations a judgment may be made to answer 'Probably no' or 'Probably yes'. For example, in the context of a large trial run by an experienced clinical trials unit, absence of specific information about generation of the randomization sequence, in a paper published in a journal with rigorously enforced word count limits, is likely to result in a response of 'Probably yes' rather than 'No information'. Alternatively, if other (contemporary) trials by the same investigator team have clearly used non-random sequences, it might be reasonable to assume that the current study was done using similar methods."

Passage(s):
<retrieved_paragraph_1>
...
<retrieved_paragraph_k>

D Further Commentary on Inter-Rater Reliability Analysis

Two independent reviewers independently assessed risk of bias for 20 trials using the revised Cochrane ROB2 tool. These assessments were used to compute IAA as reported in the main paper. Following these annotations, the two reviewers conducted a consensus meeting to better understand discrepancies arising in their annotations. The discussion process involved revisiting the Cochrane Handbook and the official ROB2 guidance document (Higgins et al., 2011; Sterne et al., 2019) to ensure alignment with recommended best practices.

Notable discrepancies emerged in this meeting, classified into four main categories: disagreement at the signaling question level, disagreement at the domain level, differences in judgments between Yes and Probably Yes, and No and Probably No. One reviewer tended to adopt a more conservative approach and tended to opt for "some concerns" or "high risk" judgments whereas the other reviewer more frequently opted for "low risk" ratings when the available information appeared sufficient. This divergence was typical of what has been described in prior research on ROB2, which has noted that even experienced reviewers may differ in how they interpret the level of concern warranted by ambiguous or incomplete reporting (Minozzi et al., 2020).

Following this consensus meeting, the reviewers were able to reach full agreement across all signaling questions and domains. This calibration process is useful for achieving subsequent consistent application of the ROB2 assessment tool.

E Few-Shot Prompting Results

Results from few-shot prompting experiments are shown in Table 5. The few-shot prompt is created by sampling one example for each class from the gold label annotations, using the same prompt template in App C. The oracle paragraph is provided for each example, the elaboration for the signaling questions is removed (due to token constraints), and the answer is appended to the end of the prompt in the form `Answer:<Label>`. Few shot examples sampled are removed from the evaluation set for models. No substantial differences are observed between the zero- and few-shot settings when models are provided the same number of context passages.

F LLM Error Analysis

The 4 LLMs we evaluated exhibit different patterns of answers, though the evaluation metrics are comparable. Performance across models according to micro- and macro-averaged F1 across domains suggests similar performance, qualitative performance is very different between models. We plot error counts in Figure 4, showing true positives (TP), alongside each of the two types of false positives (FPs) and false negatives (FNs). Here, class 1 FP/FN errors are those considered to be less severe (e.g., Y/PY wrongly classified as NI is not as severe as Y/PY wrongly classified as N/PN). We also provide raw counts of these errors in Table 6.

As seen in the figure, GPT-4o is more likely than other LLMs to answer "No Information" (large number of FP in the first column) or "No/Probably No". Llama-3.3-70B-Instruct, Claude-3.5-Sonnet, and GPT-4o most often predicted "No Information" for signaling questions where the true label was No/Probably No. On the other hand, GPT-3.5-Turbo almost never abstains with a "No Information" prediction, leading to more false positive errors for the N/PN and Y/PY classes. Claude 3.5-Sonnet was the best performing model evaluated and has fairly comparable false positive rates across "No/Probably No" and "Yes/Probably Yes". Llama-3.3-70b-Instruct demonstrates a similar answer distribution to Claude 3.5-Sonnet, but with a consistent false positive rate across all 3 classes, but higher false negatives, with "No/Probably No" being the highest.

These observed behaviors for over-predicting "No Information" or "No/Probably No" could stem from safety mechanisms learned during model post-training, which might explain GPT-3.5-Turbo's extreme bias towards "No/Probably No" and "Yes/Probably Yes". Some domains have questions phrased in a way that requires interpreting numerical data (D2 & D3) or understanding current best practices in the field (D5); stronger models tend to abstain in these cases and select "No Information". However, in the context of ROB2 assessments, these cautious predictions lead to overconservative domain- and paper-level labels (high risk judgments) for LLM-supported assessments.

G ROBoto2 Usage Statistics

Acceptance rates of model answers and rationales versus user-corrected rates are provided in Table 7. Our annotation interface allows users to rate the 3

Model	Retrieval	D1	D2	D3	D4	D5	Micro-avg	Macro-avg
n-total	-	750	1278	598	1027	750	-	-
Zero-shot setting								
Llama-3.3-70B-Instruct	k=1	0.83	0.61	0.42	0.80	0.38	0.49	0.61
GPT-3.5-Turbo	k=1	0.81	0.73	0.69	0.74	0.66	0.58	0.73
GPT-4o	k=1	0.68	0.65	0.58	0.81	0.75	0.55	0.69
Claude 3.5-Sonnet	k=1	0.69	0.68	0.66	0.87	0.76	0.60	0.73
Few-shot setting								
Llama-3.3-70B-Instruct (FS)	k=1	0.83	0.61	0.43	0.76	0.38	0.49	0.61
GPT-3.5-Turbo (FS)	k=1	0.81	0.70	0.65	0.80	0.71	0.60	0.73
GPT-4o (FS)	k=1	0.70	0.64	0.60	0.83	0.70	0.55	0.69
Claude 3.5-Sonnet (FS)	k=1	0.69	0.68	0.65	0.87	0.77	0.60	0.73

Table 5: Micro-averaged domain-level F1 along with micro- and macro-averaged F1 across signaling questions for few shot retrieval. Zero-shot k=1 results from Table 2 are reproduced here for reference.

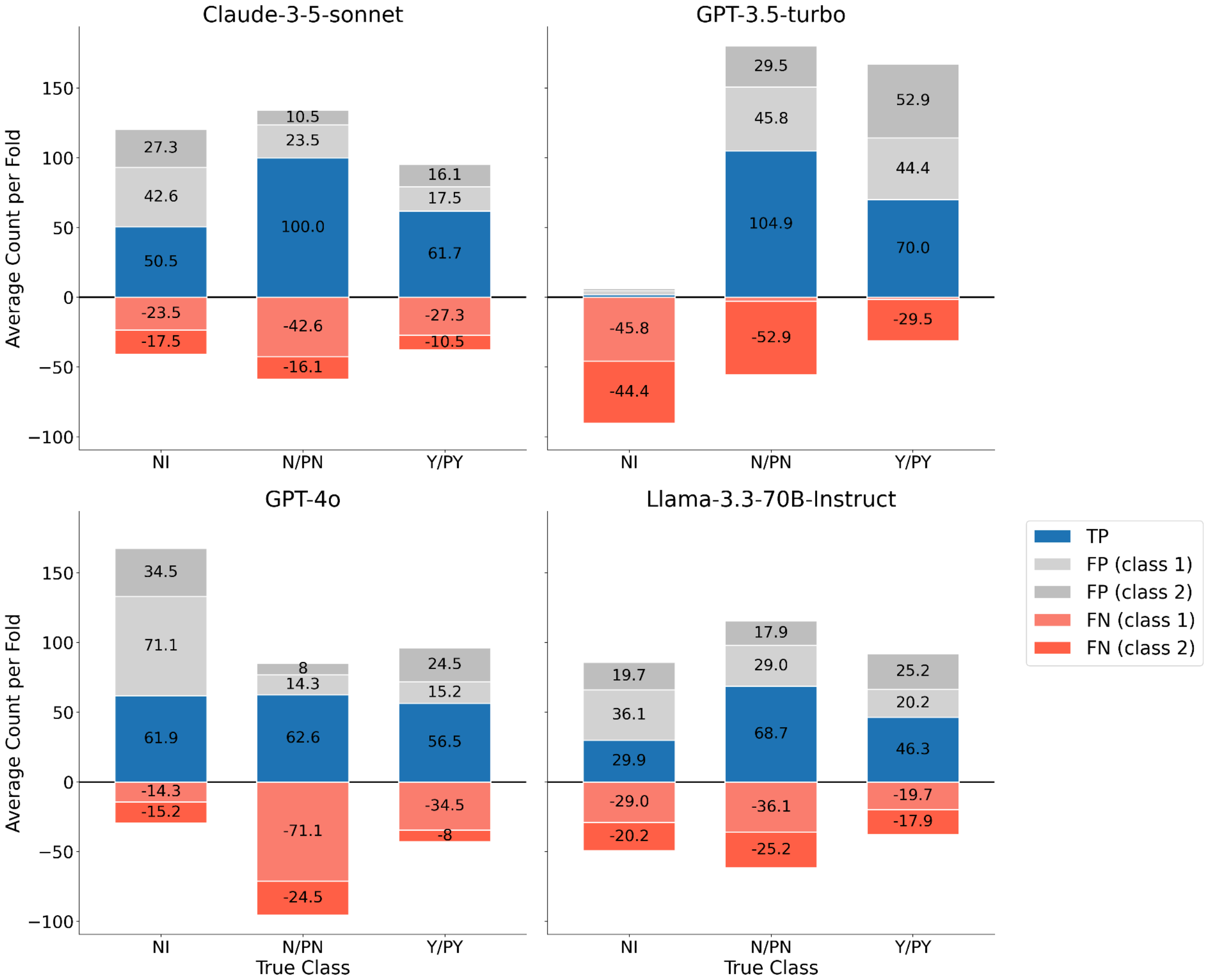


Figure 4: Stacked bar chart showcasing the aggregate true positive (TP) classifications versus false positive/negative (FP/FN) errors made by each model. FPs and FNs are each broken down into two classes, where class 1 (lighter color) are milder errors than class 2 (darker color) (e.g., misclassifying NI and N/PN or Y/PY is less severe than misclassifying N/PN as Y/PY or vice versa). Counts less than 3 have their numbers hidden for chart readability, and full counts are available in Table 6.

retrieved paragraphs with a good/bad rating and/or add their own paragraphs from the paper as context. Feedback statistics for retrieved evidence paragraphs are given in Table 8.

Model	Class	True Positives	FP (class 1)	FP (class 2)	FN (class 1)	FN (class 2)
Llama-3.3-70B-Instruct	NI	29.9	36.1	19.7	29.0	20.2
	N/PN	68.7	29.0	17.9	36.1	25.2
	Y/PY	46.3	20.2	25.2	19.7	17.9
GPT-3.5-turbo	NI	1.8	2.8	1.6	45.8	44.4
	N/PN	104.9	45.8	29.5	2.8	52.9
	Y/PY	70.0	44.4	52.9	1.6	29.5
GPT-4o	NI	61.9	71.1	34.5	14.3	15.2
	N/PN	62.6	14.3	8.3	71.1	24.5
	Y/PY	56.5	15.2	24.5	34.5	8.3
Claude-3.5-Sonnet	NI	50.5	42.6	27.3	23.5	17.5
	N/PN	100.0	23.5	10.5	42.6	16.1
	Y/PY	61.7	17.5	16.1	27.3	10.5

Table 6: Confusion-matrix summary for the LLMs when answering ROB2 signaling questions. For each response category Yes/Probably Yes (Y/PY), No/Probably No (N/PN), and No Information (NI), we list the number of true positives (TP) and the false positive (FP) and false negative (FN) counts accrued against the two alternative classes ("class 1" and "class 2"). Higher TP and lower FP/FN values reflect better agreement with the gold label. All values are normalized averages across run configurations.

	Predictions		**Rationales**	
Domain	Model (%)	Expert (%)	Model (%)	Expert (%)
1	377 (49.2%)	390 (50.8%)	430 (56.1%)	337 (43.9%)
2	853 (59.2%)	588 (40.8%)	1117 (77.5%)	324 (22.5%)
3	432 (65.2%)	231 (34.8%)	494 (74.5%)	169 (25.5%)
4	591 (64.6%)	325 (35.4%)	717 (78.3%)	199 (21.7%)
5	368 (48.2%)	396 (51.8%)	485 (63.5%)	279 (36.5%)
Total	2621 (57.6%)	1930 (42.4%)	3243 (71.3%)	1308 (28.7%)

Table 7: Counts and percentages of model-originated versus expert-corrected predictions and explanations across domains for ROB assessments completed using ROBOTO2.

Domain	Downvotes	Upvotes	User Added Paragraphs
1	43	207	74
2	40	120	84
3	12	48	24
4	22	64	112
5	18	41	62
Total	135	480	356

Table 8: Feedback on evidence passages provided by users by domain at the question level, for the ROBOTO2 subset, i.e., each number corresponds to the number of questions in that domain for which a user provided a downvote, an upvote, or added paragraph, as opposed to the total number of downvotes or upvotes etc.