RoBIn Chatbot: Leveraging LLMs for Automated Risk of Bias Assessment in Clinical Studies

Abel Corrêa Dias[1], **Viviane Pereira Moreira**[1], **João Luiz Dihl Comba**[1]

[1]Instituto de Informática – Universidade Federal do Rio Grande do Sul (UFRGS)
Caixa Postal 15.064 – 91.501-970 – Porto Alegre – RS – Brazil

{abel.correa,viviane,comba}@inf.ufrgs.br

Abstract. *The Risk of Bias (RoB) assessment is an essential instrument for evaluating the reliability of clinical studies and identifying any systematic error that can occur. This task is traditionally performed by humans, and only a few works tried to automate it using machine learning. Recent advances in large language models (LLMs) have revolutionized natural language processing and information retrieval, allowing us to build applications that can chat with documents and perform the most diverse tasks. In this work, we propose RoBIn chatbot, an LLM application able to receive clinical studies as input and classify their RoB. RoBIn chatbot uses a model trained on data derived from the Cochrane Database of Systematic Reviews and is able to perform inference for six bias types. To prevent the LLM from generating misleading conclusions, it relies on retrieval-augmented generation on the submitted file to extract the piece of evidence and send it to a pretrained model responsible for performing RoB inference.*

1. Introduction

The number of scientific publications has challenged researchers and clinical practitioners over the years [Landhuis 2016]. The COVID-19 pandemic had a significant role in biomedical research, resulting in a surge of publications in a short amount of time. While this growth in scientific output is a global response to the pandemic, it creates challenges for physicians, practitioners, researchers, and scientists who must analyze and synthesize these works. The rush to publish results has increased the probability of errors, fraud, or retractions due to a lack of peer review or pressure for immediate results. Furthermore, commercial or political interests may influence the interpretation of data, leading to bias [Brainard 2020].

Despite these challenges, various initiatives and tools have been created to facilitate access and analysis of biomedical and healthcare publications, promote interdisciplinary collaboration, and communicate results to the broader public and decision-makers [Wang et al. 2020, Romanov and Shivade 2018].

The Risk of Bias (RoB) is one of the best indicators to assess the quality of clinical studies. It refers to any systematic deviations in the results that can be introduced at the design or analysis stages of the studies. In this context, this paper developed an LLM application to support the assessment of the RoB in clinical studies. The RoBIn chatbot demo has a video demonstrating the system[1] and source code available at GitHub[2]. The

[1]https://youtu.be/1R3ZrcBovh8
[2]https://github.com/phdabel/RoBIn-chatbot

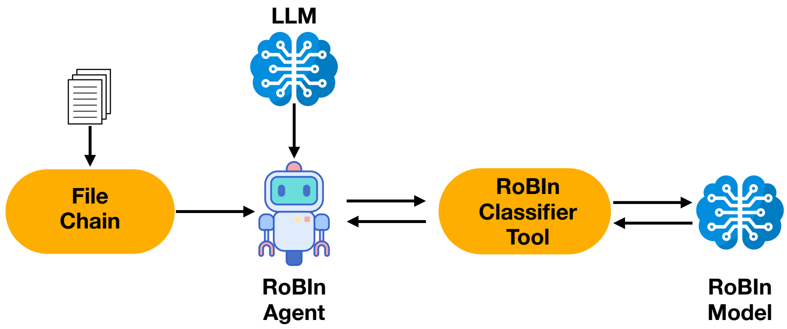

Figure 1. RoBIn Chatbot Architecture

GitHub repository has examples of text files that can be used to test the demo.

2. Clinical Trial Quality

The Cochrane Database of Systematic Reviews (CDSR)[3] is a healthcare database that features systematic reviews, protocols, editorials, and supplementary materials. Its systematic reviews offer in-depth analyses of studies, including tables that summarize references, evaluative judgments, and the supporting evidence for various types of bias assessments.

We consider the following bias types when interpreting clinical trials based on the guidelines of the Cochrane risk-of-bias tool for randomized trials [Higgins et al. 2019]: (i) **Selection bias** is due to methods used to assign patients to study treatment groups; (ii) **Performance bias** occurs when some of the participants or the staff are aware of the assigned treatment; (iii) **Detection bias** occurs in the measurement of the outcomes when assessors are aware of the assigned treatment; (iv) **Attrition bias** occurs as a result of patient withdrawals that affect a certain subset of the patients; (v) **Reporting bias** may occur when non-significant findings are ignored or omitted from the results. (vi) **Other bias** any bias type different from the previous ones.

3. RoBIn Chatbot

The goal of the RoBIn chatbot is to perform *automatic* RoB evaluations based on LLMs that leverage the RoBIn model [Dias et al. 2025] trained to perform RoB assessments. This section provides an overview of its architecture, implementation details, and potential use cases.

Architecture RoBIn chatbot's core is built around a single ”ReACT” agent [Yao et al. 2023], created with the framework LangChain, that is configured with exactly two tools: a File chain, and a linear RoB classifier (as seen in Figure 1). The File chain process documents uploaded by the users, breaks the documents into chunks, embeds them on the fly with OllamaEmbeddings into a Chroma vector store, and uses a MultiQueryRetriever[4] to pull back the most relevant passages. It then constructs a “Context + Instructions” prompt and feeds it into the ReACT agent, which can invoke (if needed) the RoBIn classifier tool.

[3]CDSR: `https://www.cochranelibrary.com/cdsr/about-cdsr`
[4]`https://python.langchain.com/docs/how_to/MultiQueryRetriever/`

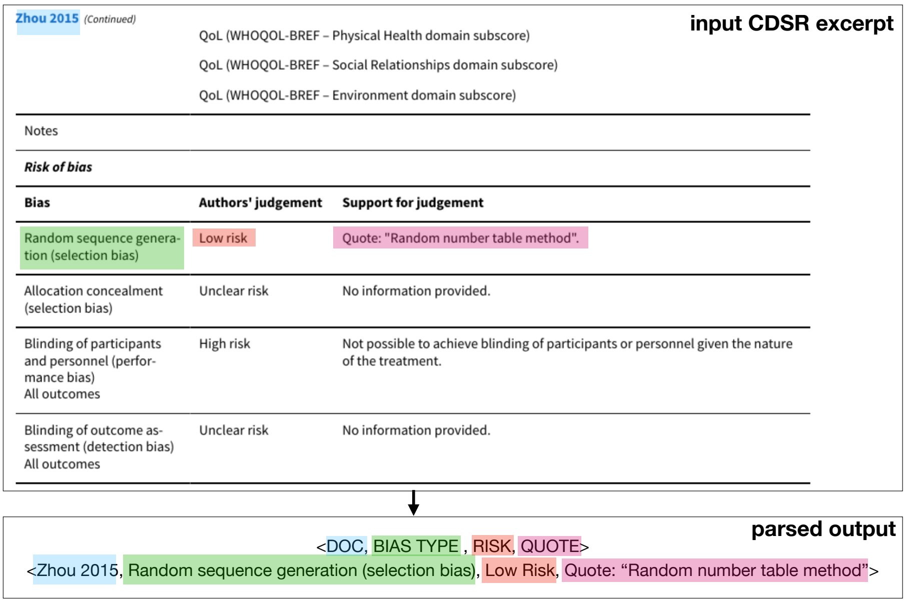

Figure 2. Parsing the systematic reviews to obtain instances for the dataset.

Auxiliary model The RoBIn classifier tool is an LLM wrapper for the RoBIn model [Dias et al. 2025], a pretrained model that labels inputs as “Low” or “High/Unclear” risk of bias. This approach enables the LLM to use an auxiliary specialized model to infer the RoB without fine-tuning. When the user needs to evaluate the RoB from a given excerpt or file, the agent redirects the call for the RoBIn model to obtain the classification output and provide an answer to the user. The accuracy of the assessment is attributed to the RoBIn model, which can be tuned at a lower cost.

Table 1 shows the results of the extractive and generative RoBIn models against machine learning approaches and some LLMs in terms of Macro-F1, Precision, and Recall. RoBIn^{Ext}, which is used as a tool for RoBIn Chatbot, has the highest Macro F1 and precision. The differences were statistically significant compared to the other models ($p < 0.05$). In contrast, the difference in recall between RoBIn^{Ext} and RoBIn^{Gen} is not statistically significant.

Table 1. Risk of Bias Inference Results: Models are listed in columns, with overall metric results in rows (best results in red).

Metric	RoBIn^{Gen}	RoBIn^{Ext}	SVM	LR	Llamma3.1 8B			Gemma2 9B			GPT-4o-Mini 8B		
					0-shot	1-shot	few-shot	0-shot	1-shot	few-shot	0-shot	1-shot	few-shot
F1	72.75	74.18	71.29	69.11	64.54	68.15	68.12	67.57	67.56	70.03	69.20	69.62	69.60
Prec.	73.21	75.49	70.84	68.96	65.06	67.75	68.43	67.18	68.04	71.16	68.83	69.17	69.14
Rec.	74.51	73.62	73.02	71.26	67.15	69.35	67.87	68.39	67.21	69.35	69.78	70.99	70.72

Dataset RoBIn was trained and validated using a public dataset, referred to **RoBIn dataset** [Dias et al. 2025], which was created by processing the RoB sections from CDSR reviews (as illustrated in Figure 2). This dataset captures information such as bias types, judgments, and supporting evidence from clinical studies, and is a valuable resource to enable future research on RoB assessment.

Using ReAct prompting [Yao et al. 2023], we built the chatbot to generate reasoning traces as intermediate steps, representing the sequential planning and decision-making

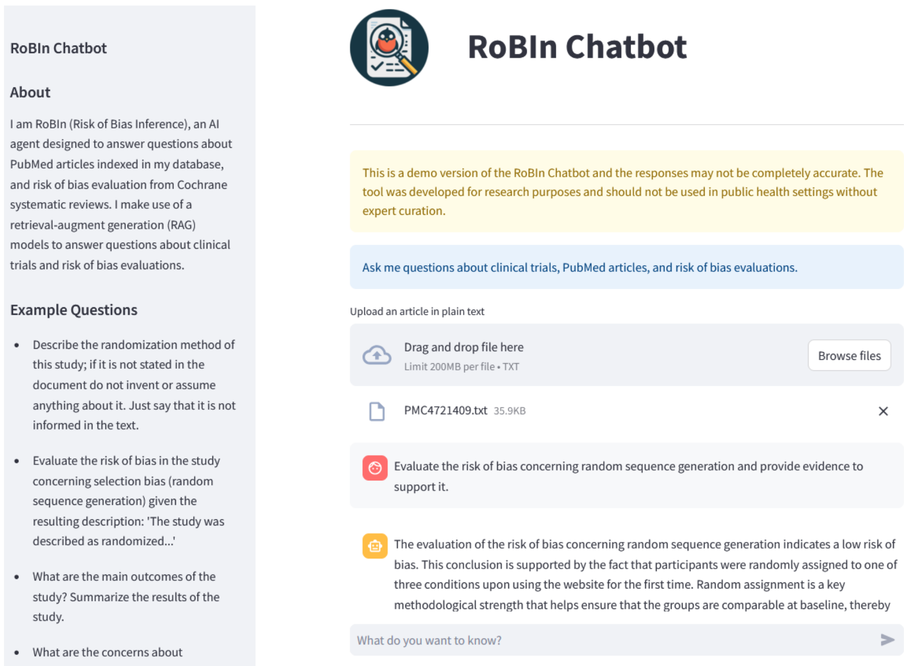

Figure 3. RoBIn chatbot.

steps the LLM takes before reaching its final answer. These intermediate steps interleave reasoning with action based on both internal logic and external feedback, and they are captured and displayed for transparency of the LLM decision-making process. The prototype allows users to submit files for RoB evaluation, summarization, and other tasks inherent to LLMs (Figure 3). The RoBIn chatbot is not tied to any specific LLM and can rely on GPT or Ollama models as backend LLMs.

Use Case: Let us consider a study about smoking cessation [Meyer et al. 2008], which is included in a Cochrane systematic review [Stead et al. 2013]. The input to the RoBIn chatbot should be a text file. We uploaded the file corresponding to the article and issued the following question: "Is this study free from selection bias? Provide evidence to support your answer." The reasoning traces are shown in the interface displaying agent actions and thoughts (*e.g.*,, "Thought: Do I need to use a tool? Yes. Action: RoBIn classifier. Action Input: Is this study free..."). In this case, the classifier indicates that the study has a high/unclear RoB for this bias type. The LLM answers: *"The study does not appear to be completely free from selection bias. While the authors attempted to minimize bias with a quasi-randomization procedure and found no significant baseline differences between groups, the fact that patients attending frequently are underrepresented in later study groups raises concerns. ..."* This result matches the RoB judgment performed by human evaluators in the CDSR for random sequence generation and allocation concealment.

For the same study, we inquired about the risk of attrition bias using the following question: "Is the incomplete outcome data addressed in this study? Evaluate the risk of bias concerning attrition bias and provide evidence to support it." This resulted in an incorrect high/unclear RoB label, whereas the correct label is low. The accuracy of the RoBIn model [Dias et al. 2025] varies across bias types, with an 0.83 AUC ROC.

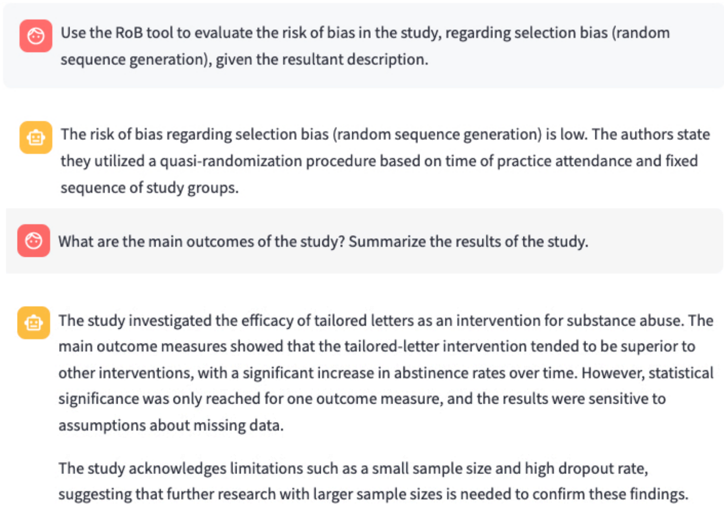

Figure 4. RoBIn chatbot interaction example.

Next, we modified the question as follows: "Is the incomplete outcome data addressed for all outcomes in this study? Evaluate the risk of bias concerning attrition bias for all outcomes and provide evidence to support it." This modification provided more context to the classification model, resulting in low RoB: *"The text states that the authors took steps to address incomplete outcome data by using several methods. They conducted regression imputation to handle missing data on covariates, and they also used a weighted estimating equation (WEE) model to account for sample selection bias. Additionally, they explored different assumptions about missing follow-up data, including complete case analysis, last value carried forward, and coding all missing observations as smokers. ..."*. Another example of interactions with the RoBIn chatbot are given in Figure 4.

Minimal Requirements: The RoBIn chatbot requires a system with at least an Intel Core i5 processor (2.9GHz) to ensure efficient task execution. A minimum of 16GB of RAM is recommended for continuous operation, particularly when managing LLMs and multiple user interactions. Additionally, a dedicated GPU, such as the NVIDIA GeForce RTX 4060 with 8GB of VRAM, is essential for acceleration, enabling real-time inference and optimal performance in conversational AI applications.

4. Conclusion

This paper introduced RoBIn Chatbot, an LLM-based application that relies on a Transformer-based model to perform RoB inference. As a future work, we intend to evaluate the results with biomedical experts and explore how their feedback can be used to improve the underlying RoB classification model.

The performance of the RoBIn chatbot relies on three key components: the file retrieval system (MultiQueryRetriever), the LLM, and the RoBIn classifier. Since the model input is derived from the retrieved file content, exceeding the context size limit can result in missing file context, leading to incorrect classifications. To mitigate this issue, the augmented generative retrieval strategy should be reviewed and improved to ensure that the bias classification model always receives relevant file context. Additionally, integrating RoBIn with existing systematic reviews, publications, and clinical trial databases could further enhance its ability to assess RoB with greater accuracy and contextual support.

Ethical considerations are crucial when applying AI in the biomedical domain. The proposed application serves as a research demonstration, and all results generated by the LLM must be reviewed and validated by experts in systematic reviews.

References

Brainard, J. (2020). Scientists are drowning in COVID-19 papers. Can new tools keep them afloat? Science.

Dias, A. C., Moreira, V. P., and Comba, J. L. D. (2025). RoBIn: A Transformer-based model for risk of bias inference with machine reading comprehension. *Journal of Biomedical Informatics*, 166:104819.

Higgins, J. P., Savović, J., Page, M., and Sterne, J. (2019). Revised Cochrane risk-of-bias tool for randomized trials (RoB 2).

Landhuis, E. (2016). Scientific literature: Information overload. *Nature*, 535:457–458.

Meyer, C., Ulbricht, S., Baumeister, S., Schumann, A., Rüge, J., Bischof, G., Rumpf, H., and John, U. (2008). Proactive interventions for smoking cessation in general medical practice: a quasi-randomized controlled trial to examine the efficacy of computer-tailored letters and physician-delivered brief advice. *Addiction*, pages 294–304.

Romanov, A. and Shivade, C. (2018). Lessons from Natural Language Inference in the Clinical Domain. *CoRR*, abs/1808.06752.

Stead, L., Buitrago, D., Preciado, N., Sanchez, G., Hartmann-Boyce, J., and Lancaster, T. (2013). Physician advice for smoking cessation (review). `https://www.cochranelibrary.com/cdsr/doi/10.1002/14651858.CD000165.pub4/epdf/full`.

Wang, L. L., Lo, K., Chandrasekhar, Y., et al. (2020). CORD-19: The COVID-19 Open Research Dataset. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*. Association for Computational Linguistics.

Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., and Cao, Y. (2023). React: Synergizing reasoning and acting in language models. `https://arxiv.org/abs/2210.03629`.