# Measuring Risk of Bias in Biomedical Reports: The RoBBR Benchmark

**Jianyou Wang*** **Weili Cao*** **Longtian Bao** **Youze Zheng** **Gil Pasternak**
**Kaicheng Wang** **Xiaoyue Wang** **Ramamohan Paturi** **Leon Bergen**
Laboratory for Emerging Intelligence
University of California, San Diego
{jiw101, w2cao, rpaturi, lbergen}@ucsd.edu

## Abstract

Systems that answer questions by reviewing the scientific literature are becoming increasingly feasible. To draw reliable conclusions, these systems should take into account the quality of available evidence from different studies, placing more weight on studies that use a valid methodology. We present a benchmark for measuring the methodological strength of biomedical papers, drawing on the risk-of-bias framework used for systematic reviews. Derived from over 500 biomedical studies, the three benchmark tasks encompass expert reviewers' judgments of studies' research methodologies, including the assessments of risk of bias within these studies. The benchmark contains a human-validated annotation pipeline for fine-grained alignment of reviewers' judgments with research paper sentences. Our analyses show that large language models' reasoning and retrieval capabilities impact their effectiveness with risk-of-bias assessment. The dataset is available at https://github.com/RoBBR-Benchmark/RoBBR.

## 1 Introduction

Systems that automatically answer questions by reviewing the scientific literature are becoming increasingly feasible. These systems, such as "Deep Research" offerings from OpenAI, Google, and Anthropic, have the potential to provide scientists with on-demand access to knowledge that is synthesized from across the literature. This can help scientists understand what is already known about topics outside of their research focus, and help clinicians and practitioners keep up to date with best practices.

When assessing what is known about a field, not all studies should be weighed equally (Boutron et al., 2023). Studies with stronger methodologies contribute more to a body of evidence than those with weaker methodologies. By weighing studies appropriately, systems can increase the reliability of their summaries and recommendations (Turner et al., 2009).

In biomedical research, there is a large body of work which investigates the factors that decrease the validity of a study's methods, and best practices for addressing these issues (Boutron et al., 2023; Welton et al., 2009). For instance, reporting bias occurs when there are systematic differences in how outcomes are reported or disclosed between the groups that are compared (e.g., the results of the treatment group are more frequently or favorably published than those in the placebo group). This bias can be mitigated by registering the study in advance and committing to publish all results (Kotz and West, 2022).

We introduce the RoBBR benchmark for evaluating the methodological strength of biomedical studies, which is referred to as their risk-of-bias levels in the context of this paper. RoBBR's main task involves labeling these risk-of-bias levels. Importantly, we also introduce two novel subtasks: support sentence retrieval (SSR) and support judgment selection (SJS). The SSR subtask is designed to test a model's ability to identify support sentences — specific sentences within an entire paper that signal potential biases. We created the SSR dataset using our annotation pipeline. The SJS subtask evaluates a model's capacity to synthesize this retrieved information through reasoning to form a well-supported judgment. Through our in-depth analysis, we find that the abilities to reason and, especially, to retrieve information are important milestones in improving models' capacity to assess risk-of-bias levels.

## 2 Background

We provide more background on systematic reviews and the risk-of-bias guidelines.

A systematic review is a method in evidence-

*Equal contribution



*Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3221–3249
November 4-9, 2025 

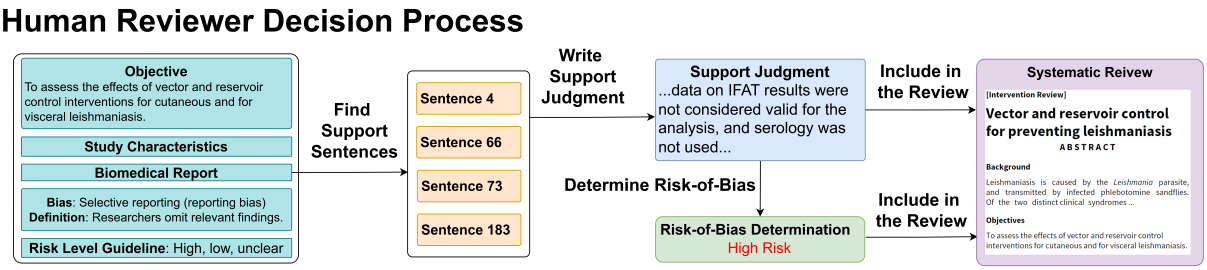


Figure 1: Reviewer finds Sentence 4, 66, 73, 183. The reviewer writes a support judgment based on these sentences. The support judgment indicates this biomedical study has a high risk for selective reporting bias. Both support judgment and the high risk rating are included in the systematic review.

based medicine that synthesizes evidence from multiple studies to answer important clinical questions. They help researchers and healthcare professionals make informed decisions based on the best available evidence (Ahn and Kang, 2018; Lasserson et al., 2023; Deeks et al., 2023).

Risk-of-bias assessment involves evaluating each study to determine the likelihood that its results may be biased (Higgins et al., 2023; Practice and of Care , EPOC; Sterne et al., 2014). Studies with a high risk of bias may overestimate or underestimate the true treatment effect, leading to inaccurate conclusions in the systematic review.

RoBBR follows the risk-of-bias guideline developed by Cochrane (Sterne et al., 2019; Higgins et al., 2023, 2011; Practice and of Care , EPOC; Sterne et al., 2014), a global independent network that conducts systematic reviews of healthcare interventions. Cochrane's risk-of-bias guideline is widely recognized as a standard for evaluating the quality and reliability of research studies. It serves as an important resource for national health agencies in various countries, and are used to formulate national and international guidelines on healthcare practices (Viswanathan et al., 2012; Alderson and Tan, 2011; NHMRC, 2019; Bunn et al., 2015).

During reviewer's assessment of a study's risk of bias, two review authors independently evaluate the risk of bias using the Cochrane tool and resolve disagreements through discussion. They record their rationale for each bias assessment. These rationales are referred to as "support judgments".

## 3 Related Work

### Risk-of-Bias Labeling

Previous research focuses on creating datasets and training models to label risk-of-bias levels (Lai et al., 2024a; Hasan et al., 2024; Wang et al., 2022b; Marshall et al., 2015). Traditionally, machine learning models including support vector machines (Marshall et al., 2014, 2017; Pereira et al., 2020), convolutional neural networks (Marshall et al., 2017; Zhang et al., 2016) and logistic regression (Millard et al., 2016; Marshall et al., 2020) are used for label prediction. More recently, some work employs transformers (Dias et al., 2025) and LLMs to determine risk-of-bias levels (Lai et al., 2024b; Pitre et al., 2023; Šuster et al., 2024). While RoBBR's main task is similar to existing work, its novel contribution is the introduction of two new subtasks: support sentence retrieval (SSR) and support judgment selection (SJS). These subtasks identify crucial intermediate steps in risk-of-bias labeling and respectively measure the information retrieval and reasoning capabilities of LLMs when they assess risk-of-bias.

### Support Sentences in Risk-of-Bias Labeling

Previous studies (Marshall et al., 2016; Dias et al., 2025) attempt to extract supporting sentences from reviewers' support judgments using direct quotes (Marshall et al., 2016) or semantic embeddings (Dias et al., 2025). However, these techniques falter because support judgments are free-form and frequently omit direct quotes, while semantic embeddings are imprecise and incomprehensive (See Table 2). This makes sentences extracted by such methods unsuitable for a rigorous benchmark like our Support Sentence Retrieval (SSR) subtask, which tests an LLM's capability in retrieving support sentences. To address this, in Section 4.3, our annotation pipeline aligns all aspects (such as paraphrases, synthesized and summarized information) from support judgment to sentences in biomedical studies, which would more reliably and comprehensively extract support sentences that are used to form the novel SSR subtask.

### LLM Extraction from Systematic Reviews

In recent years, LLMs have been used to extract

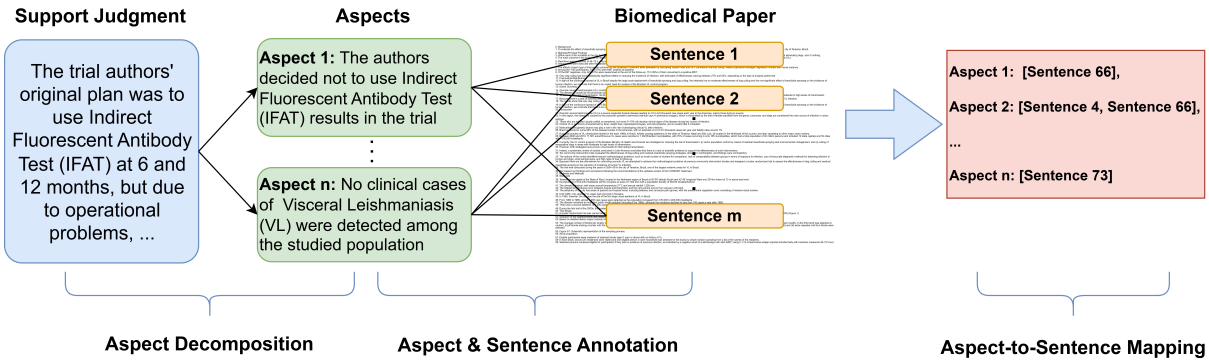


Figure 2: Given a support judgment, our annotation pipeline produces the Aspect-to-Sentence Mapping, a many-to-many relationship. The mapping on the right-side shows Sentence 66 covers both Aspect 1 and 2, so if a model retrieves Sentence 66, it should not retrieve Sentence 4 to avoid redundancy.

data from systematic reviews, such as experiment evidence (Sun et al., 2024; Gartlehner et al., 2023), summaries (Wang et al., 2022a; Wallace et al., 2020), quality of evidence (Suster et al., 2023) and PICO (Tang et al., 2024). In our work, aligning the human support judgment to various sentences in biomedical studies is a more complex and difficult "data extraction" task, as noted by Dias et al., which necessitates the aspect decomposition and aspect & sentence annotation components in our annotation pipeline.

## 4 Benchmark Development

### 4.1 Benchmark Statistics and Data Sources

| Task | Cochrane Train | Cochrane Test | Non-Cochrane Test |
|---|---|---|---|
| Retrieval (SSR) | n=235 | n=313 | N/A |
| Selection (SJS) | n=346 | n=465 | N/A |
| Risk-of-Bias (Main) | n=774 | n=906 | n=2,489 |

Table 1: RoBBR Benchmark Statistics.

There are two prevailing risk-of-bias assessment guidelines, RoB1 (Higgins and Green, 2011) and the updated RoB2 (Sterne et al., 2019). As of 2025, many recently published and updated systematic reviews still use RoB1, for example, (Miao et al., 2025; Mehrholz et al., 2025). Since the RoB2 guidelines do not require support judgments that are necessary for our subtasks, our subtasks only include systematic reviews that follow RoB1. For the main task, we include systematic reviews that follow either set of guidelines.

Since there are hundreds of bias names and definitions, such as *"incomplete adverse event reporting (reporting bias)"* and *"selective outcome reporting (reporting bias)"*, for clarity of presentation, we manually group them into 6 primary categories: "selection", "attrition", "performance", "detection", "reporting" and "deviation" (unique to RoB2) using two approaches. In some cases, bias names already include broader category labels (e.g., "random sequence generation (selection bias)"). When this happens, we follow these broader category labels. When bias names lack such broader category labels, we manually identify an equivalent bias name that has a category label and use this label. For example, the bias name "bias arising from the randomization process" is equivalent to "random sequence generation (selection bias)," so we use "selection bias" as its category label. A bias name belongs to multiple categories when these categories are explicitly included in the bias name. For example, a bias called "blinding for adverse event (performance and detection bias)" is categorized into both the "performance bias" category and the "detection bias" category. See Table 12 for examples of how bias names are categorized.

During evaluation, models would see the actual bias name and definition used in systematic reviews. Bias definitions are sourced from systematic reviews themselves and official guidelines like (Higgins and Green, 2011; Practice and of Care , EPOC; Sterne et al., 2014).

For data diversity, we include both systematic reviews that are published in Cochrane (denoted as Cochrane Train/Test) and that are published elsewhere (denoted as Non-Cochrane Test[1]). Non-Cochrane reviews do not have support judgments and cannot form our subtasks.

[1]Empirically, Llama3 fine-tuned on Cochrane train set also perform well on Non-Cochrane test set, so we did not create a separate Non-Cochrane train set.

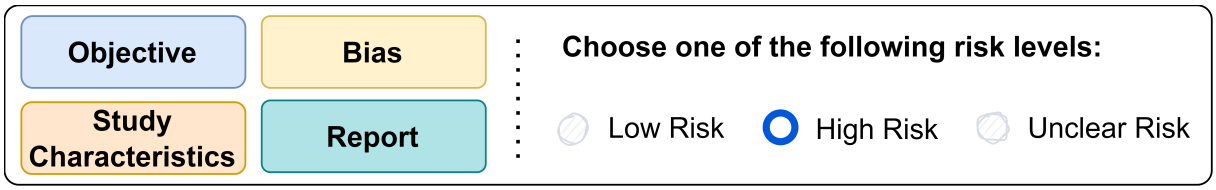


Figure 3: Main Task: Risk-of-Bias Determination. The goal of the main task is to provide an assessment of the paper's risk of specific biases.

RoBBR includes 58 Cochrane reviews that assess 204 papers, and 496 Non-Cochrane reviews that assess 496 papers. Appendix A.3 show details on token length, bias and label distributions. See Table 1 for statistics of the three tasks in our benchmark.

### 4.2 Main Task: Risk-of-Bias Determination

The main task evaluates a model's ability to label the risk-of-bias of a biomedical study as high, low, or unclear/some concern. The input to the model is exactly everything that a human expert would see, including the biomedical paper, study characteristics (i.e. PICO), the objective/topic of the systematic review, one specific bias name and definition, and risk level definition. See the top-left panel of Figure 1 for a visualization of input. The same input would be used for the two subtasks as described in section 4.4, 4.5. To visualize the task, see Figure 3.

The decision-making process of risk-of-bias determination is highly complex and involves multiple intermediate steps: retrieve support sentences from the biomedical paper, synthesize evidence from support sentences into a support judgment, and follow the guidelines and definitions of bias and risk level to give a final determination.

### 4.3 Automatic Annotation Pipeline

Support judgments contain a wealth of information, including evidence, data, reasoning, and commentaries written by expert reviewers. The goal of support judgment is to explain why reviewers would give this risk-of-bias rating. Therefore, support judgments mirror and shed light on the decision-making process of the reviewer.

It is difficult to find all support sentences in the study that form the basis of a support judgment. As shown in Figure 2, support judgments can have many aspects. While some aspects are direct quotes from the study, many other aspects are paraphrases from the study, or deeper analyses that synthesize information from multiple parts of the study.

To reverse engineer a support judgment, we find that no lexicon-based or embedding models can find all support sentences accurately (see Section 4.3.1 and Table 2). To address this, our proposed annotation pipeline has two main components: Aspect Decomposition and Aspect & Sentence Annotation. See Figure 2 for a visualization. GPT-4 (gpt-4-0125) is used by our annotation pipeline. All prompts used in our annotation pipeline are in Appendix D.

**Aspect Decomposition:**
To reduce the complexity of the reverse engineering process, we decompose the support judgment into distinct, non-overlapping pieces of information, each focusing on a specific aspect of the original support judgment. This is accomplished by careful prompt engineering with GPT-4. We manually validated that all decomposed aspects from all support judgments are of high quality.

**Aspect & Sentence Annotation:**
To further reduce complexity and instability introduced by long context input to LLMs, we only ask GPT-4 to determine if one sentence from the study covers one aspect.

**Definition 1.** *A sentence $s_j$ in a paper **covers** an aspect $f_i$ if it satisfies these two criteria:*
*1: The content of the sentence conveys the majority of the aspect's information.*
*2: Any information about the aspect not directly stated in the sentence can be reasonably inferred from the surrounding context.*

We aggregate these annotations into the **Aspect-to-Sentence Mapping:** For each (study, support judgment) triplet, a sentence is mapped to several aspects (e.g., 0, 1, 2, or more). Most sentences are not mapped to any aspect because they do not cover them. A few sentences can even be mapped to more than one aspect because aspect-to-sentence mapping is a many-to-many relationship.

The total cost of our annotation pipeline via OpenAI API is around $1,000 with aspect decomposition and initial filtering costing less than $100.

| Metrics | Human & Human | Human & GPT | p-value |
|---|---|---|---|
| Exact Accuracy | $99.4 \pm 0.2$ | $99.3 \pm 0.2$ | 0.31 |
| F1 Binary | $73.4 \pm 6.3$ | $71.7 \pm 5.6$ | 0.55 |
| Cohen's $\kappa$ | $73.1 \pm 6.3$ | $71.4 \pm 5.6$ | 0.54 |
| Spearman's $\rho$ | $73.8 \pm 6.0$ | $71.7 \pm 5.6$ | 0.43 |
| **Metrics** | **Human & Human** | **Human & Embedding** | **p-value** |
| Exact Accuracy | $99.4 \pm 0.2$ | $98.5 \pm 0.2$ | $< 10e{-}3$ |
| F1 Binary | $73.4 \pm 6.3$ | $30.5 \pm 8.1$ | $< 10e{-}3$ |
| Cohen's $\kappa$ | $73.1 \pm 6.3$ | $29.7 \pm 8.2$ | $< 10e{-}3$ |
| Spearman's $\rho$ | $73.8 \pm 6.0$ | $29.8 \pm 8.3$ | $< 10e{-}3$ |

Table 2: Inter-annotator agreement. **Top:** comparison between two human teams and GPT-4. **Bottom:** comparison between two human teams and OpenAI-v3-large as a post-hoc analysis.

Next, we evaluate the quality of GPT-4 annotation.

#### 4.3.1 Evaluating Annotation Quality

We hypothesize a sentence matches an aspect is straightforward enough for GPT-4 to achieve human-level performance with a specialized prompt. Following the protocol introduced by (Wang et al., 2023), we optimized the instruction prompt for GPT-4 on a development set (see Appendix D). We randomly sampled 50 papers, with each assigned a single aspect. Across all 50 papers, this resulted in a total of 13,575 (aspect, sentence) pairs. Four graduate student were tasked with annotating these pairs. The annotators were divided into two teams, with each team consisting of two annotators. The annotators' training for annotation is described in Appendix B.1. Each annotator performed the annotation tasks individually given a pre-defined annotation guideline in Appendix B. Annotators from the same team then collaborated to resolve differences and eliminate mistakes.

Each team produced a set of annotation results. GPT-4-0125 annotated the same 13k (aspect, sentence) triplets. We calculated Exact Accuracy, F1-binary, Spearman Correlation, and Kappa Coefficient between the two human teams and between each human team and GPT-4. Table 2 shows human-human correlations in the low-to-mid 70s, indicating reasonable agreement, despite the fact that finding sentences that cover aspects is non-trivial due to different annotators' potentially varied interpretations of key terms in **Definition 1** such as *"majority"* and *"reasonably inferred"*. For context on inter-annotator agreement in similar aspect-based tasks, Wang et al. reported 54% $\rho$ (three-way annotation) and Mysore et al. found near 40% pre-adjudication $\rho$ (four-way annotation).

Table 2 shows human-GPT4 and human-human correlations are comparable. A post-hoc analysis reveals that powerful embedding models like OpenAI-v3-large (OpenAI, 2024) have significantly lower correlation with humans. For the primary human-GPT4 comparison, the p-value (p = 0.3) from our bootstrapped test fails to reject the null hypothesis (no difference in correlation). In contrast, we can reject the null hypothesis for the embedding model which has clear and detectable differences from human annotators.

### 4.4 Support Sentence Retrieval (SSR)

With our annotation pipeline, we transform a support judgment into a special information retrieval task that evaluates if a model can identify and retrieve these support sentences from the full text of the paper. Since these support sentences form the basis of the support judgment, if a model can locate these support sentences, it is one step closer to reaching a good support judgment and classifying the correct risk-of-bias level.

Figure/table captions are split into sentences as text. Tables are included and turned into markdown format and treated as individual sentences. Figures are excluded since SSR is designed to evaluate textual models.

We have obtained the **Aspect-to-Sentence Mapping**. From this mapping, there exists the smallest optimal integer $K$ such that there are $K$ sentences that can cover all aspects in a support judgment. Therefore, when a model is only allowed to retrieve $K$ sentences from a study, we measure the percentage of aspects that the model's retrieved set of sentences can cover. This metric is denoted as **Aspect Recall Ratio @ Optimal**.

Formally, let $\{s_1, \ldots, s_k\}$ be the set of retrieved sentences, and $\{f_1, \ldots, f_m\}$ be the set of aspects. We define the indicator function $\mathcal{S}(f_i, s_j)$ as:

$$\mathcal{S}(f_i, s_j) = \begin{cases} 1 & \text{if } s_j \text{ covers } f_i \\ 0 & \text{otherwise} \end{cases}$$

The **Aspect Recall Ratio @ Optimal** is defined as

$$\frac{\sum_{i=1}^{m} \mathbf{1}\left\{\left(\sum_{j=1}^{K} \mathcal{S}(f_i, s_j)\right) \geq 1\right\}}{m} \quad (1)$$

Not only does it evaluate if a model can find support sentences, but it also evaluates if a model can control information redundancy from these sentences. Figure 5 in Appendix C visualizes SSR.

| Model | Avg | Bias Type Selection n = 157 | Attrition n = 46 | Performance n = 54 | Detection n = 61 | Reporting n = 14 |
|---|---|---|---|---|---|---|
| OpenAI-v3 | 22.7 | 40.1 | 6.3 | 26.7 | 27.4 | 13.1 |
| GritLM-7B | 18.9 | 35.8 | 7.0 | 15.2 | 26.8 | 9.5 |
| GPT-4o | **47.5** | 60.5 | **42.4** | 41.5 | **50.1** | **43.0** |
| Sonnet-3.5 | 39.2 | 55.4 | 30.7 | 37.0 | 37.2 | 35.5 |
| Llama-3.1-70B | 45.6 | 61.2 | 34.2 | 45.2 | **50.1** | 37.4 |
| Llama-3-8B | 22.7 | 49.4 | 14.5 | 22.1 | 20.8 | 6.5 |
| Llama-3-8B Fine-tuned | 40.8 | **69.0** | 24.8 | **47.3** | 48.6 | 14.3 |

Table 3: Support Sentence Retrieval (SSR). Evaluated on Cochrane Test (SSR). Metric is Aspect Recall Ratio @ Optimal. Llama-3-8B fine-tuned on Cochrane Train (SSR). Fine-tuning details are included in Appendix E.

| Model | Avg | Bias Type Selection n = 130 | Attrition n = 98 | Performance n = 87 | Detection n = 82 | Reporting n = 80 |
|---|---|---|---|---|---|---|
| GPT-4o | 47.2 | 58.5 | 60.2 | 48.3 | 42.7 | 26.3 |
| Sonnet-3.5 | **59.9** | **73.1** | **73.5** | **50.6** | **51.2** | **51.3** |
| Llama-3.1-70B | 53.2 | 66.2 | 62.2 | 44.8 | 46.3 | 46.3 |
| Llama-3-8B | 26.5 | 26.9 | 34.7 | 24.1 | 22.0 | 25.0 |
| Llama-3-8B Fine-tuned | 29.6 | 40.8 | 28.6 | 24.1 | 19.5 | 35.0 |

Table 4: Support Judgment Selection (SJS). Evaluated on Cochrane Test (SJS). Metric is Accuracy. Llama-3-8B fine-tuned on Cochrane Train (SJS). Fine-tuning details are included in Appendix E.

| Model | Avg | Bias Type Selection n = 933 | Attrition n = 629 | Performance n = 309 | Detection n = 645 | Reporting n = 594 | Deviation n = 331 |
|---|---|---|---|---|---|---|---|
| GPT-4o | **42.1** | 50.5 | 36.5 | **54.7** | 43.7 | **34.4** | 32.7 |
| Sonnet-3.5 | 41.9 | **52.8** | **37.1** | 50.6 | 43.3 | 33.6 | **34.2** |
| Llama-3.1-70B | 38.8 | 48.1 | 31.7 | 48.0 | **44.4** | 31.6 | 29.0 |
| Llama-3-8B | 30.1 | 36.4 | 32.4 | 39.1 | 37.2 | 19.8 | 15.4 |
| Llama-3-8B Fine-tuned | 36.3 | 49.5 | 33.3 | 42.4 | 40.1 | 27.0 | 25.7 |
| LR Trained | N/A | 26.0 | 21.7 | 27.9 | 28.6 | 21.8 | N/A |
| SVM Trained | N/A | 30.7 | 24.2 | 23.3 | 25.6 | 23.5 | N/A |

Table 5: Main Task, Risk-of-Bias Determination. Evaluated on Cochrane Test (Main) + Non-Cochrane Test (Main). Metric is Macro-F1. Llama-3-8B, logistic regression, and SVM are fine-tuned/trained on Cochrane Train (Main). Fine-tuning details are included in Appendix E.

### 4.5 Support Judgment Selection (SJS)

Simply retrieving support sentences is insufficient for a risk-of-bias determination. A model needs the ability to interpret and synthesize information from these retrieved support sentences in order to generate a good support judgment that will lead to the correct determination of risk-of-bias.

Following the convention of using multiple-choice questions as a proxy for generative tasks (Rein et al., 2023), we propose the support judgment selection subtask, an MCQ task, where the model selects the correct support judgment from a mix of three synthetically generated support judgments, and three human-written options derived from other papers' support judgments for the same exact bias. These three human-written options sometimes proved to be surprisingly hard distractors because their lack of direct paper quotes made them into generic, non-paper-specific statements. A deep logical understanding of the paper's content rather than superficial semantic matching is required to look past them and find the correct one.

We prompt GPT-4 to generate the three synthetic options by imitating support judgments from other papers concerning the same exact bias name. These options are tailored to be paper-specific while maintaining the underlying reasoning. Empirically, these six distractor options are often misleading. Table 4 shows the best LLM only achieves 60% accuracy. In order to ensure that all incorrect options were actually incorrect, all 906 datapoints were checked by human, and 465 datapoints were kept. See Figure 6 in Appendix C to visualize SJS.

## 5 Experiments

### 5.1 Evaluated Models

We evaluate various LLMs, GPT-4o-2024-05-13 (OpenAI, 2024a), Sonnet3.5-20240620 (Anthropic, 2024), and Llama-3.1-70B (Dubey et al., 2024). We evaluate and fine-tune light-weight Llama-3-8B (Dubey et al., 2024). We also evaluate two recent embedding models GritLM 7B (Muennighoff et al., 2024) and OpenAI-embedding-v3-large (OpenAI, 2024). Note the GPT-4o model is different from the GPT-4-0125 model used in our annotation pipeline.

For the main task, we further evaluate three traditional ML models: RobotReviewer (Marshall et al., 2016), Support Vector Machine (SVM) and Logistic Regression (LR) (Dias et al., 2025).

For all generative language models, we use chain-of-thought as a prompting strategy to stabilize the model performance. We develop our

prompts on train sets. For reproducibility, we use a common generation setting of temperature 0, top-p 1, frequency penalty 0, and presence penalty 0.

### 5.2 Evaluation Metrics

Some bias names may fall into multiple categories, causing the total datapoint count across categories to be slightly higher than the test set size. Model performance is reported per category and as an average across all categories.

The SSR subtask uses **Aspect Recall Ratio @ Optimal** defined in Equation 1 as its metric. It measures the percentage of aspects covered by the retrieved sentences. Unlike traditional metrics (e.g., recall, precision), it penalizes redundancy by rewarding models that retrieve the minimum number of sentences (typically 1-2) needed to cover all aspects for a given data point. For instance, if one sentence covers multiple aspects, retrieving additional sentences that cover already addressed aspects is discouraged.

## 6 Benchmarking Results and Analyses

### 6.1 LLMs Exhibit Distinct Strengths in Retrieval and Reasoning

While GPT-4o and Sonnet-3.5, with 42.1% and 41.9% average Macro-F1, are comparable in the main task Table 5, a closer inspection of their performances on the two subtask Tables 3, 4 reveals interesting differences. In support sentence retrieval (SSR), GPT-4o has 47.5% Aspect Recall Ratio, while Sonnet-3.5 only has 39.2%. This shows that Sonnet-3.5 is worse at retrieving the best support sentences that are the bases of human expert-level support judgments. However, in support judgment selection (SJS), Sonnet-3.5 outperforms other GPT-4o with 59.9% over 47.2% accuracy. This shows that Sonnet-3.5 can select better support judgments, which involves reasoning and deliberating which option best aligns with the paper's contents and the bias and risk-level definitions.

In other words, GPT-4o's advantage over Sonnet-3.5 is information retrieval, while Sonnet-3.5's advantage is information synthesis with reasoning. We provide an example in Appendix G comparing reasoning traces of the two models to clearly illustrate Sonnet-3.5's advantage in reasoning.

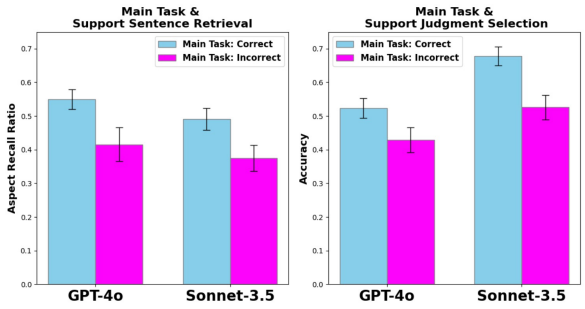


Figure 4: There is a positive correlation between the main task and SSR & SJS subtask performance.

| Retrieval | Reasoning | Main Task Result |
|---|---|---|
| Sonnet-3.5 | GPT-4o | 31.5 ± 5.0 |
| GPT-4o | GPT-4o | 40.3 ± 6.9 |
| Ground Truth | GPT-4o | 46.1 ± 7.5 |
| Ground Truth | Sonnet-3.5 | 49.4 ± 8.0 |

Table 6: Two stage pipeline that first retrieves and then reason based on the retrieved results (full paper is not shown). Main Task, Risk-of-Bias Determination. Evaluated on the 313 datapoints in Cochrane Test (SSR), all of which have risk-of-bias labels. Metric is Macro F1 score (%) averaged across bias categories. The error bar represents the 95% confidence interval.

### 6.2 Positive Correlation Between Main Task and Subtasks

We investigate the relationship between the main task and our subtasks. For an LLM (either GPT-4o or Sonnet 3.5), we categorize datapoints in the SSR subtask into two classes: datapoints whose risk-of-bias level is correctly predicted by the LLM and those incorrectly predicted. For each of the two classes, we calculate its average Aspect Recall Ratio @ Optimal. See the left part in Figure 4, we observe that the Aspect Recall Ratio is significantly higher in the correct class than the incorrect class, for both Sonnet-3.5 and GPT-4o, suggesting LLMs' performances on SSR are positively correlated with their performances on the main task.

Similarly, datapoints in the SJS subtask are divided into two classes. From the right part in Figure 4, we see that, for both Sonnet-3.5 and GPT-4o, their performances on SJS are positively correlated with their performances on the main task.

### 6.3 Retrieval and Reasoning Ablation

To investigate how the retrieval and reasoning abilities of LLMs affect their performance on the main task, we employ a two-stage retrieval and reasoning pipeline. In this setup, one LLM processes sentences retrieved by another LLM, rather than

accessing the original paper directly. The specific prompt used is detailed in Appendix I.3. This experiment was conducted on Cochrane Test (SSR), which comprises 313 datapoints, each associated with a support judgment and a risk-of-bias label.

We pair different models' retrieved sentences with different reasoning models to evaluate performance on the main task Risk-of-Bias Determination. Ground truth retrieved sentences are provided by our annotation pipeline.

As shown in Table 6, when results from different retrieval methods are paired with the same reasoning model (GPT-4o), ground-truth retrieval achieves significantly higher macro F1 score (46.1%) over Sonnet-3.5's retrieval (31.5%), confirming that retrieval quality impacts main task performance. When the retrieval quality is held constant (using ground-truth retrieval) and the reasoning model varied, there is some inconclusive evidence that suggests Sonnet-3.5 slightly outperforms GPT-4o. These ablations demonstrate that both retrieval and reasoning abilities are relevant for the Risk-of-Bias Determination task.

### 6.4 Retrieval Marginally Improves LLM on Non-Cochrane Test Set

So far we have focused on the Cochrane test set, which includes support judgments, unlike the Non-Cochrane test set. To investigate whether more accurate support sentences could enhance LLM performance on the Non-Cochrane test set (2489 datapoints), we used the o4-mini reasoning model (OpenAI, 2024b) to find such sentences within each paper, guided by the datapoint's actual risk level (see Appendix I.4 for the prompt). In this proof-of-concept setup, the LLM received the full paper and o4-mini-identified support sentences as highlights, which we compared against baselines with full paper but no highlights.

Table 7 reveals that both GPT-4o and Sonnet-3.5 showed slight performance improvements on the Non-Cochrane Test (Main) over their baselines. The limited gain is partly due to inaccuracies in retrieved sentences, as recovering support sentences without expert support judgments is challenging.

### 6.5 Challenges in Support Sentence Retrieval

Our findings establish the retrieval task SSR as a crucial intermediate milestone for improving LLM on the main task. Nevertheless, Table 3 indicates that in zero-shot setting, current models struggle significantly: modern embedding models

| Setting | Main Task |
|---|---|
| GPT-4o Full Paper | 42.5 ± 2.2 |
| GPT-4o Full Paper + Highlight Retrieval | 44.1 ± 2.2 |
| Sonnet-3.5 Full Paper | 40.3 ± 2.2 |
| Sonnet-3.5 Full Paper + Highlight Retrieval | 44.5 ± 2.2 |

Table 7: Main Task, Risk-of-Bias Determination. Evaluated on Non-Cochrane Test (Main). Metric is Macro F1 score (%) averaged across bias categories. The error bar represents the 95% confidence interval.

such as OpenAI-v3 and GritLM-7B achieve low recall ($\leq$22.7%), while commercial LLMs also underperform (recall $\leq$47.5%). Although fine-tuning LLama3-8B boosts its SSR recall from 22.7% to 40.8%, its retrieval quality remains unsatisfactory.

A significant limitation of embedding models is their lack of context awareness. Even recent contextual-aware models (Morris and Rush, 2024) cannot handle the long-range interactions across different sentences. In Appendix F, we examine a type of globally-connected datapoints in SSR potentially necessitates a more global understanding of the entire context. We leave the exploration of contextual embedding models and more accurate long-context LLMs for solving SSR as a future research direction.

### 6.6 Limitations of Traditional ML Models on RoBBR

| Model | Biases | | | | |
|---|---|---|---|---|---|
| | Avg | Allocate Conceal n=32 | Blind Outc n=19 | Blind Part n=18 | RanSeq Gen n=30 |
| RobotReviewer | 56.7 ± 8.4 | 75.0 | 39.1 | 43.8 | 68.9 |
| LR | 53.1 ± 9.7 | 71.9 | 50.4 | 51.8 | 38.4 |
| SVM | 44.8 ± 8.6 | 45.9 | 55.2 | 41.9 | 36.0 |
| GPT-4o | 65.6 ± 8.5 | 83.6 | 59.1 | 41.9 | 77.8 |
| Sonnet-3.5 | 67.5 ± 8.4 | 77.0 | 82.5 | 41.9 | 68.8 |

Table 8: Main Task, Risk-of-Bias Determination. Evaluated on a subset from the Cochrane Test (Main), comprising only the four RobotReviewer-assessable bias types: allocation concealment, blinding of outcome, blinding of participants, and random sequence generation. Only two judgment categories (low, and high/unclear) per RobotReviewer's specification. Metric is Macro-F1. The error bar represents the 95% confidence interval.

This study compares previous ML models with LLMs on four specific biases, the only ones assessable by RobotReviewer (Marshall et al., 2017) (an SVM/CNN ensemble) due to its inaccessible training code. We also trained logistic regression and SVM models on our Cochrane Train Set (Main)

(Dias et al., 2025). Table 8 shows inconclusive results : RobotReviewer, SVM, and LR perform marginally worse, if at all, than GPT-4o and Sonnet-3.5.

The four biases RobotReviewer assesses are considered straightforward because superficial keywords often directly indicate bias risk (e.g., "opaque envelope" for low allocation concealment bias; "randomness" or "random number generator" for low random sequence generation bias), suggesting significant keyword reliance. However, biases like selective reporting or deviation bias may necessitate a more holistic consideration of the entire biomedical paper.

Traditional ML models by definition are hard to generalize to new bias names and definitions. RoB1/RoB2 guidelines list hundreds of bias names, some with definitions re-interpreted in different reviews. we treat these as paper-specific biases, manually incorporating the paper's re-interpreted definition during LLM evaluation on RoBBR. Thus, RoBBR primarily evaluates models in zero-shot settings or fine-tunes models like Llama3-8B for generalization to unseen bias definitions via semantic understanding. Furthermore, most bert-based models' 512-token context limit is ill-suited for RoBBR's biomedical studies (average 8k tokens), unlike modern LLMs with longer context windows.

## 7 Conclusion

We present RoBBR, a benchmark for measuring models' ability to assess risk-of-bias in biomedical studies. RoBBR includes two novel subtasks that evaluate a model's retrieval and reasoning abilities. The support sentence retrieval subtask is created by our fully automatic and human-validated annotation pipeline for aligning support judgments to sentences in biomedical studies. Our analysis reveals the importance of retrieval and reasoning abilities and demonstrate their impact on the a model's ability to assess risk-of-bias.

## Limitations

While in this work all systematic reviews and biomedical studies are open-sourced (Creative Commons License or Public Domain), many other systematic reviews have licenses which restrict distributing their contents. Therefore, we share the entire codebase for creating the RoBBR benchmark and encourage researchers with access to non-open-source contents to create other versions of RoBBR.

In our evaluation of LLMs, while we have developed our prompts using RoBBR's train sets and investigated popular techniques such as chain-of-though reasoning etc, we acknowledge the possibility that better prompting techniques could lead to slight model improvements on RoBBR.

## Ethics Statement

All data in RoBBR come from open-sourced systematic reviews and biomedical studies covering the domain of evidence-based biomedicine. The authors make sure no personal information is included in RoBBR by manual inspection of all datapoints.

The RoBBR benchmark, which promotes the automation of risk-of-bias determination, might introduce over-reliance on commercial AI systems to evaluate a biomedical study's risk-of-bias.

The benchmark only includes systematic reviews and biomedical studies in English. Systems trained and evaluated on RoBBR might disadvantage the risk-of-bias determinations for non-English biomedical studies.

## References


Eunjin Ahn and Hyun Kang. 2018. Introduction to systematic review and meta-analysis. *Korean J. Anesthesiol.*, 71(2):103–112.

P Alderson and T Tan. 2011. The use of cochrane reviews in nice clinical guidelines.

Anthropic. 2024. Claude 3.5 sonnet. Accessed: 2024-07-14.

I Boutron, MJ Page, JPT Higgins, DG Altman, A Lundh, and A Hróbjartsson. 2023. Chapter 7: Considering bias and conflicts of interest among the included studies. In Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, and Welch VA, editors, *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane.

Frances Bunn, Daksha Trivedi, Phil Alderson, Laura Hamilton, Alice Martin, Emma Pinkney, and Steve Iliffe. 2015. The impact of cochrane reviews: a mixed-methods evaluation of outputs from cochrane review groups supported by the national institute for health research. *Health technology assessment (Winchester, England)*, 19(28):1—99, v—vi.

Donald C Comeau, Chih-Hsuan Wei, Rezarta Islamaj Doğan, and Zhiyong Lu. 2019. PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, 35(18):3533–3535.

Jonathan J Deeks, Julian PT Higgins, Douglas G Altman, Joanne E McKenzie, and Areti Angeliki Veroniki. 2023. Chapter 10: Analysing data and undertaking meta-analyses. *In: Higgins JPT*, 6.(4).

Abel Corrêa Dias, Viviane Pereira Moreira, and João Luiz Dihl Comba. 2025. Robin: A transformer-based model for risk of bias inference with machine reading comprehension. *Journal of Biomedical Informatics*, page 104819.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Gerald Gartlehner, Leila Kahwati, Rainer Hilscher, Ian Thomas, Shannon Kugley, Karen Crotty, Meera Viswanathan, Barbara Nussbaumer-Streit, Graham Booth, Nathaniel Erskine, Amanda Konet, and Robert Chew. 2023. Data extraction for evidence synthesis using a large language model: A proof-of-concept study. *Research Synthesis Methods*, n/a(n/a).

Bashar Hasan, Samer Saadi, Noora S Rajjoub, Moustafa Hegazi, Mohammad Al-Kordi, Farah Fleti, Magdoleen Farah, Irbaz B Riaz, Imon Banerjee, Zhen Wang, and Mohammad Hassan Murad. 2024. Integrating large language models in systematic reviews: a framework and case study using robins-i for risk of bias assessment. *BMJ Evidence-Based Medicine*.

J.P. Higgins and S. Green. 2011. *Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 (updated March 2011)*. Cochrane. URL: training.cochrane.org/handbook/archive/v5.1/.

JPT Higgins, J Savović, MJ Page, RG Elbers, and JAC Sterne. 2023. Chapter 8: Assessing risk of bias in a randomized trial. In Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, and Welch VA, editors, *Cochrane Handbook for Systematic Reviews of Interventions*. Cochrane.

Julian P T Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savovic, Kenneth F Schulz, Laura Weeks, Jonathan A C Sterne, Cochrane Bias Methods Group, and Cochrane Statistical Methods Group. 2011. The cochrane collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343(oct18 2):d5928.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alexander Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. 2023. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. *arXiv preprint arXiv:2305.02301*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Daniel Kotz and Robert West. 2022. Key concepts in clinical epidemiology: addressing and reporting sources of bias in randomized controlled trials. *J. Clin. Epidemiol.*, 143:197–201.

Honghao Lai, Long Ge, Mingyao Sun, Bei Pan, Jiajie Huang, Liangying Hou, Qiuyu Yang, Jiayi Liu, Jianing Liu, Ziying Ye, Danni Xia, Weilong Zhao, Xiaoman Wang, Ming Liu, Jhalok Ronjan Talukdar, Jinhui Tian, Kehu Yang, and Janne Estill. 2024a. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Netw. Open*, 7(5):e2412687.

Honghao Lai, Long Ge, Mingyao Sun, Bei Pan, Jiajie Huang, Liangying Hou, Qiuyu Yang, Jiayi Liu, Jianing Liu, Ziying Ye, et al. 2024b. Assessing the risk of bias in randomized clinical trials with large language models. *JAMA Network Open*, 7(5):e2412687–e2412687.

Toby J Lasserson, James Thomas, and Julian PT Higgins. 2023. Chapter 1: Starting a review. *In: Higgins JPT*, 6.(4).

Zhuowan Li, Cheng Li, Mingyang Zhang, Qiaozhu Mei, and Michael Bendersky. 2024. Retrieval augmented generation or long-context llms? a comprehensive study and hybrid approach. *arXiv preprint arXiv:2407.16833*.

Patrice Lopez. 2008–2024. Grobid. `https://github.com/kermitt2/grobid`.

Iain Marshall, Joël Kuiper, Edward Banner, and Byron C. Wallace. 2017. Automating biomedical evidence synthesis: RobotReviewer. In *Proceedings of ACL 2017, System Demonstrations*, pages 7–12, Vancouver, Canada. Association for Computational Linguistics.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2014. Automating risk of bias assessment for clinical trials. In *proceedings of the 5th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pages 88–95.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2015. Automating risk of bias assessment for clinical trials. *IEEE J. Biomed. Health Inform.*, 19(4):1406–1412.

Iain J Marshall, Joël Kuiper, and Byron C Wallace. 2016. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *J Am Med Inform Assoc*, 23(1):193–201.

Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. 2020. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912.

J Mehrholz, J Kugler, M Pohl, and B Elsner. 2025. Electromechanical-assisted training for walking after stroke. *Cochrane Database of Systematic Reviews*, 2025(5):CD006185. Art. No.: CD006185.

C Miao, Y Hu, G Bai, N Cheng, Y Cheng, and W Wang. 2025. Prophylactic abdominal drainage for pancreatic surgery. *Cochrane Database of Systematic Reviews*, 2025(5):CD010583. Art. No.: CD010583.

Louise AC Millard, Peter A Flach, and Julian PT Higgins. 2016. Machine learning to assist risk-of-bias assessments in systematic reviews. *International journal of epidemiology*, 45(1):266–277.

John X Morris and Alexander M Rush. 2024. Contextual document embeddings. *arXiv preprint arXiv:2410.02525*.

Niklas Muennighoff, Hongjin Su, Liang Wang, Nan Yang, Furu Wei, Tao Yu, Amanpreet Singh, and Douwe Kiela. 2024. Generative representational instruction tuning. *arXiv preprint arXiv:2402.09906*.

Sheshera Mysore, Tim O'Gorman, Andrew McCallum, and Hamed Zamani. 2021. Csfcube-a test collection of computer science research articles for faceted query by example. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

NHMRC. 2019. Guidelines for guidelines: Assessing risk of bias. `https://nhmrc.gov.au/guidelinesforguidelines/develop/assessing-risk-bias`. Last published 29 August 2019.

OpenAI. 2024a. Hello gpt-4o. Accessed: 2024-05-22.

OpenAI. 2024b. Introducing o3 and o4 mini.

OpenAI. 2024. New embedding models and api updates. Accessed: 2024-05-19.

Ramon Gonçalves Pereira, Giulia Zanon Castro, Pamela Azevedo, Lucas Tôrres, Isabella Zuppo, Túlio Rocha, and Augusto Afonso Guerra. 2020. Mcrb: A multi-classifier tool for risk of bias assessment in a systematic review to produce health evidence to decision making. In *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 1–6. IEEE.

Tyler Pitre, Tanvir Jassal, Jhalok Ronjan Talukdar, Mahnoor Shahab, Michael Ling, and Dena Zeraatkar. 2023. Chatgpt for assessing risk of bias of randomized trials using the rob 2.0 tool: A methods study. *Medrxiv*, pages 2023–11.

Cochrane Effective Practice and Organisation of Care (EPOC). 2017. Suggested risk of bias criteria for epoc reviews.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

J.A.C. Sterne, J.P.T. Higgins, and B.C. on behalf of the development group for ACROBAT-NRSI Reeves. 2014. A cochrane risk of bias assessment tool: for non-randomized studies of interventions (acrobat-nrsi), version 1.0.0.

Jonathan A C Sterne, Jelena Savović, Matthew J Page, Rebecca G Elbers, Natalie S Blencowe, Isabelle Boutron, Christopher J Cates, Hsiu-Yin Cheng, Mhairi S Corbett, Sandra M Eldridge, Miguel A Hernán, Sally Hopewell, Asbjørn Hróbjartsson, Daniela R Junqueira, Peter Jüni, Jamie J Kirkham, Toby Lasserson, Tianjing Li, Alison 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:l4898.

Zhuanlan Sun, Ruilin Zhang, Suhail A. Doi, Luis Furuya-Kanamori, Tianqi Yu, Lifeng Lin, and Chang Xu. 2024. How good are large language models for automated data extraction from randomized trials? *medRxiv*.

S. Suster, T. Baldwin, J. Lau, A. Jimeno Yepes, D. Martinez Iraola, Y. Otmakhova, and K. Verspoor. 2023. Automating quality assessment of medical evidence in systematic reviews: Model development and validation study. *J Med Internet Res*, 25:e35568.

Simon Šuster, Timothy Baldwin, and Karin Verspoor. 2024. Zero-and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. *Research Synthesis Methods*, 15(6):988–1000.

Yiyi Tang, Ziyan Xiao, Xue Li, Qingpeng Zhang, Esther W Chan, and Ian CK Wong. 2024. Large language model in medical information extraction from titles and abstracts with prompt engineering strategies: A comparative study of gpt-3.5 and gpt-4. *medRxiv*.

Rebecca M Turner, David J Spiegelhalter, Gordon C S Smith, and Simon G Thompson. 2009. Bias modelling in evidence synthesis. *J. R. Stat. Soc. Ser. A Stat. Soc.*, 172(1):21–47.

Meera Viswanathan, Mohammed T. Ansari, Nancy D. Berkman, Sally Chang, Lisa Hartling, Melissa L. McPheeters, Pasqualina L. Santaguida, Tatyana Shamliyan, Khai Singh, Alexander Tsertsvadze, and J. Richard Treadwell. 2012. *Assessing the Risk of Bias of Individual Studies in Systematic Reviews of Health Care Interventions*. AHRQ Publication.

Byron C. Wallace, Sayantan Saha, Frank Soboczenski, and Iain J. Marshall. 2020. Generating (factual?) narrative summaries of rcts: Experiments with

neural multi-document summarization. *Preprint*, arXiv:2008.11293.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Doris-mae: Scientific document retrieval using multi-level aspect-based queries. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lucy Lu Wang, Jay DeYoung, and Byron Wallace. 2022a. 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.

Qianying Wang, Jing Liao, Mirella Lapata, and Malcolm Macleod. 2022b. Risk of bias assessment in preclinical literature using natural language processing. *Res. Synth. Methods*, 13(3):368–380.

N. J. Welton, A. E. Ades, J. B. Carlin, D. G. Altman, and J. A. C. Sterne. 2009. Models for potentially biased evidence in meta-analysis using empirically based priors. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1):119–136.

Guilherme L. Werneck, Carlos H. N. Costa, Fernando Aecio Amorim de Carvalho, Maria do Socorro Pires e Cruz, James H. Maguire, and Marcia C. Castro. 2014. Effectiveness of insecticide spraying and culling of dogs on the incidence of leishmania infantum infection in humans: A cluster randomized trial in teresina, brazil. *PLoS Neglected Tropical Diseases*, 8(10):e3172.

Ye Zhang, Iain Marshall, and Byron C Wallace. 2016. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2016, page 795.

# A Dataset

## A.1 Dataset License and Code License

The RoBBR dataset is made available under CC-BY-NC. A copy of the full license can be found at RoBBR License.

The code used in this paper is released under the MIT License. The MIT License is a permissive open-source license that allows for the free use, modification, and distribution of the code, as long as the original license is included with any derivative work. A copy of the full license can be found at RoBBR License.

## A.2 Dataset Hosting, Accessibility and Maintenance

The RoBBR dataset with its meta-data is released and can be accessed freely at RoBBR License. We commit to regularly maintain the dataset and codebase by incorporating user feedback. We will potentially introduce more features as part of future work in the next version of RoBBR. We confirm that the current version of RoBBR will always remain accessible at the same link.

## A.3 Dataset Statistics

See Tables 9, 10, 11 for detailed statistics of the RoBBR tasks.

| Statistic | min | avg | max |
|---|---|---|---|
| Tokens per report | 2,908 | 9,215 | 19,080 |
| Sentences | 87 | 218.6 | 400 |
| Covered Aspects | 1 | 1.9 | 7 |
| Optimal # of sentences | 1 | 1.4 | 5 |

Table 9: Statistics for the Cochrane Test Set SSR (313 data points).

| Bias Type | Cochrane Test SSR | Cochrane Test SJS | Cochr. + Non-Cochr. Test Main |
|---|---|---|---|
| Selection | 157 | 130 | 933 |
| Attrition | 46 | 98 | 629 |
| Performance | 54 | 87 | 309 |
| Detection | 61 | 82 | 645 |
| Reporting | 14 | 80 | 594 |
| Deviation | 0 | 0 | 331 |

Table 10: Test set category statistics

## A.4 Dataset Collection and Processing

We use BioC (Comeau et al., 2019) API to download meta-reviews and papers in PMC database.

| Statistic | Cochrane Train | Non-Cochr. Test | Cochrane Test |
|---|---|---|---|
| Points | 774 | 2 489 | 906 |
| Tokens per report: | | | |
| min | 1,991 | 1,488 | 2,274 |
| avg | 9,907 | 7,791 | 8,946 |
| max | 18,894 | 26,154 | 19,080 |
| Label distribution: | | | |
| Low | 387 | 1,465 | 562 |
| Unclear | 275 | 574 | 195 |
| High | 112 | 450 | 149 |

Table 11: Test-set statistics for **Risk-of-Bias Determination**. Unclear/Some concerns for Non-Cochrane test

We manually download the rest of the papers. We use GROBID (Lopez, 2008–2024) to parse papers from PDF to XML format. We use Stanza (Qi et al., 2020) to split paragraphs (including table and figure captions) into sentences.

## A.5 RoBBR Structure

- SSR_test.json and SSR_dev.json: Test and dev set of Support Sentence Retrieval
  - paper_doi: The DOI of the paper.
  - bias: The bias to be considered.
  - PICO: PICO of a study in the paper, including Methods, Participants, Intervention, Outcome, and Notes.
  - objective: The meta-analysis objective.
  - paper_as_candidate_pool: A tuple of text elements from the paper. Each text element is a sentence, a section title, a table, or a figure caption.
  - aspects: A dictionary that maps aspect id to bias aspect.
  - aspect2sentence_indices: a mapping (i.e. dictionary) between aspect id and all sentence indices that independently are source of information for that aspect, as annotated by our pipeline.
  - sentence_index2aspects: a mapping (i.e. dictionary) between sentence index and all aspect ids that this sentence is the source of information of.
  - bias_retrieval_at_optimal_evaluation: This is a dictionary that contains the

necessary information for evaluating the model's performance on the task Support Sentence Retrieval @ Optimal.
    - optimal: A positive integer, which is the smallest number of sentences needed to cover the largest number of aspects.
    - one_selection_of_sentences: a list of sentence indices. The list size is the optimal number. The list of sentences cover the largest number of aspects. Note, there are potentially other lists of sentences that has the same size and also cover the largest number of aspects.
    - covered_aspects: the list of aspects that are covered. In this case, the list of aspects covered is list of all aspects.

- SJS_test.json and SJS_dev.json: Test and dev set of Support Judgment Selection
    - paper_doi: The DOI of the paper.
    - bias: The bias to be considered.
    - PICO: PICO of a study in the paper, including Methods, Participants, Intervention, Outcome, and Notes.
    - objective: The meta-analysis objective.
    - full_paper: The full paper content.
    - options: The seven options for the multiple choice.
    - label: The index of the correct option.

- Main_task_test.json and Main_task_dev.json: Test and dev set of the main task
    - paper_doi: The DOI of the paper.
    - bias: The bias to be considered.
    - PICO: PICO of a study in the paper, including Methods, Participants, Intervention, Outcome, and Notes.
    - objective: The meta-analysis objective.
    - full_paper: The full paper content.
    - label: One of [low,high,unclear], representing the risk level of the bias.

## B Annotation Guideline

Below, we show the annotation guideline for aspect mapping of 50 randomly sampled (aspect, full paper) pair. The four annotators form two teams of two persons, all seeing the same annotation guideline.

You have a total of 50 annotation task packets. Each task packet is a docx. file that contains the following information.

- An Aspect (one piece of important information/detail).
- The support judgment and the bias.
- The doi of the paper.
- An indexed list of text elements of the paper (a text element could be a sentence, tables, a figure caption, etc.)

You have to pledge the following conditions are met during annotation for each task packet.

- For words you are not familiar with and believe are important for comprehension, conduct the search to understand its meaning.
- For every text element in the list, you have to look at it and read it at least once.
- You cannot talk to the other annotator team about anything related to your task, including progress and insights.
- You should independently do the task first, and then consult with the other person in your team after completing the task.
- You have to take a mandatory 5 minute break after every 1 hour of performing annotation.
- You cannot exceed 8 hours of annotation per day.

Below is the recommended procedure for annotating each packet.

- Read and understand the aspect and the support judgment first. You need to understand the context of the aspect, i.e., the role of the aspect in the support judgment given the bias. You also need to understand why this aspect is important for judging the bias. You can consult with LLM to understand this aspect.
- Decide what details are important in the aspect. Geo, temporal, and numerical data are all important details.

- For each text element, you need to decide if a significant amount of important details can be implied from the text element.
  - When deciding the level of implication, consider how the text element can help judge the bias. You should not make complicated implications, i.e., can you see the aspect from the text element within 30 seconds? If not, then it is not a match.
  - Pay attention to acronyms, abbreviations, or different presentation formats of the same information.
  - Even if you find significant information that can be implied from the text element, you have to make sure the text element is in the same context as the aspect.
    - Same context typically refers to the same study or experiment, and for numerical results in Table, it means row and column must indicate the same setting.
    - Check if the text element refers to the same experiment as the aspect. Since different experiments could be in one paper. Maybe the text element does not refer to any experiment at all.
  - If you suspect that there might be a relationship between the text element and the aspect but do not understand the meaning of the text element, you can use LLM for help. However, you must justify the response from LLM, and you should not rely on the response from LLM.
- If a text element is a Table, refer to the actual Table in the pdf for better understanding. However, only consider the information in the text element.
- After you have independently finished all 50 tasks, you should talk to the other person in your team following these procedures:
  - Go through task 0-49.
  - Resolve your difference, check if you made a mistake, or if you missed something. If you made a conceptual error (e.g., fail to understand some terminology), you may have to quickly go through the paper again.
  - For sentences that you cannot resolve your difference after discussion, i.e., one person says yes and the other person says no, or if both people are unsure, you should include them in your final list of decisions.
- Ultimately, you and your teammate should collaboratively arrive at a consensus for each of the 50 tasks. Write the collective answers in the answer file provided.

  End of Annotation Guideline.

### B.1 Annotator Training

The annotation teams each comprised two graduate students. To ensure a thorough understanding of bias names and risk levels, all annotators were trained and briefed on their definitions prior to beginning annotation. A key part of their preparation involved an in-depth reading of five to ten full-length meta-analyses from Cochrane Train (Main), including their extensive supplementary materials (often over 100 pages), to gain familiarity with relevant terminology and details. Although our annotation guidelines allowed for the use of Large Language Models (LLMs) to facilitate understanding of unfamiliar subject matter, annotators infrequently consulted these tools. Their preference was to use search engines (with AI overview features disabled). On the rare occasions LLMs were employed, their function was strictly confined to assisting with the comprehension of unfamiliar technical details. Annotators were monitored and were expressly forbidden from soliciting LLM decisions or opinions regarding annotation choices. It is important to note that LLMs were infrequently utilized and only as a supplementary tool for assisting the understanding of unfamiliar subject matters.

## C Task Descriptions and Visualizations

We provide visual illustrations of the three tasks.

### C.1 Main Task: Risk-of-Bias Determination

The model is given the whole biomedical paper, study characteristics (i.e. participants, intervention methods, comparators, outcome, etc.), the objective/topic of the systematic review, one bias name and definition, and general guideline that explains

| Bias Name | Bias Category |
|---|---|
| random sequence generation (selection bias) | selection bias |
| bias arising from the randomization process | selection bias |
| bias due to deviations from intended intervention | deviation bias |
| bias due to missing outcome data | attrition bias |
| incomplete outcome data (attrition bias) all outcomes | attrition bias |
| similarity of other baseline characteristics (selection and performance bias) | selection bias / performance bias |
| bias in measurement of the outcome | detection bias |
| blinding of outcome assessment (detection bias) all outcomes | detection bias |
| blinding for adverse events (performance and detection bias) | performance bias / detection bias |
| blinding of participants and personnel (performance bias) | performance bias |
| selective reporting (reporting bias) | reporting bias |
| bias in selection of the reported result | reporting bias |
| blinding (performance bias and detection bias) all outcomes | performance bias / detection bias |

Table 12: Examples of Bias Categorization

how to classify the risk level as high, low or unclear/some concern, and is asked to determine the risk level of bias. Note, the definition is based on the specific bias name and risk level. Our categorization of different bias names into six main categories does not affect the definition of individual biases. See Table 12 for examples of bias categorization.

### C.2 Subtask: Support Sentence Retrieval

When judging the risk of bias for a specific study, the review authors provide support judgments which justify their risk of bias rating. These support judgments are grounded in specific aspects of the study, which the support judgment describes. This aims to test a model's ability to retrieve the correct source for the support judgment from a paper's text. Note the model is instructed to retrieve no more than the optimal number of sentences. The optimal number can be calculated from Aspect Sentence Mapping and is already included in our dataset. See Figure 5 for an illustration.

### C.3 Subtask: Support Judgment Selection

Figure 6 shows an example of a multiple-choice question with one correct answer and three synthetically generated answers. Both options C and D refer to the same information from the paper, in this case the operational challenges affecting the trial's methodology, but describe different reasoning given this information. This demonstrates that retrieving the correct information from the paper is not sufficient for solving this task.

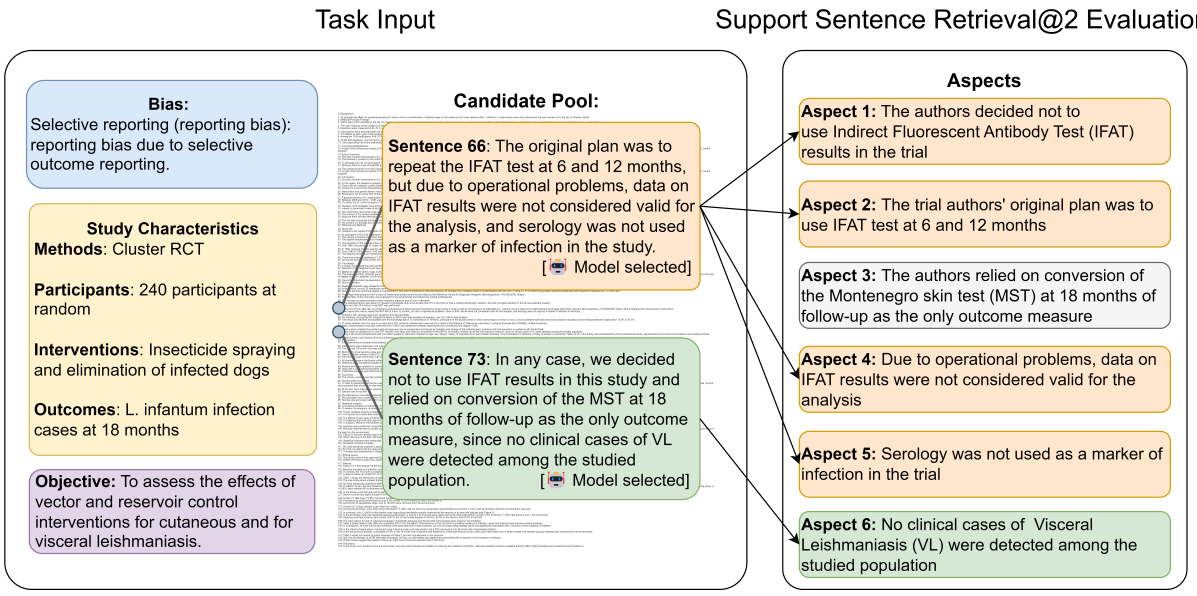


Figure 5: Subtask: Support Sentence Retrieval. In this task, the goal is to retrieve no more than the optimal number of sentences from the biomedical report which support a risk of bias judgment. Performance is evaluated against a support judgment, which has been split into aspects.

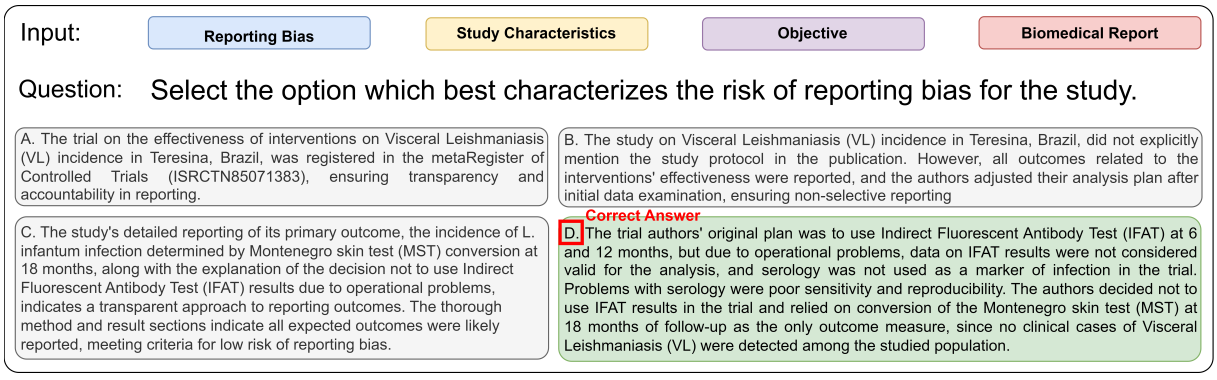


Figure 6: Subtask: Support Judgment Selection. In this task, the model is shown a paper, and asked which of 7 support judgments provides the best explanation of the paper's risk of bias. Only four options are shown here for illustration purposes. Both options C and D refer to the same information from the paper, in this case the operational challenges affecting the trial's methodology, but describe different reasoning given this information. This demonstrates that retrieving the correct information from the paper is not sufficient for solving this task.

Given a statement about a paper, Your job is to decompose the statement into pieces of information. If there is only one piece of information, you should not do decomposition.

Here are some important rules for decomposition:
1. Each piece of extracted information should reveal only one specific aspect about the statement.
2. Do not delete any information from the statement.
3. There must not be overlap between different pieces of information.
4. Only output the extracted information.

Statement:
**{aspect}**

Figure 7: Prompt for decomposing support judgment into aspects.

## D Support Sentence Retrieval Prompt, Optimization and Example

### D.1 Aspect Decomposition Prompt

See Figure 7 for the aspect decomposition prompt.

### D.2 Aspect Filtering Prompt

See Figure 8 for prompt to filter non-specific commentaries.

### D.3 Aspect-to-Sentence Mapping Prompt and Optimization

See Figure 9 for the prompt optimized on the development set.

To enhance the agreement between human annotators and the GPT-4-0125-preview model, we developed and tested various prompt engineering strategies on a development set consisting of 30 unique (aspect, paper) pairs. It is important to note that no paper or aspect from the development set overlapped with the 50 tasks in the hypothesis testing set.

GPT-4-0125-preview was chosen due to its robust instruction-following capabilities and our familiarity with its performance across different settings. Our experimentation revealed that using in-context learning examples did not improve agreement rates and tended to cause overfitting. Instead, we found that embedding the same instruction at both the beginning and the end of the prompt effectively helped the model maintain focus on the task of identifying text elements that significantly cover the details specified in an aspect.

We also implemented chain-of-thought reasoning, prompting GPT-4 to articulate its thought process following a specific keyword. This approach not only enhanced the quality of the model's reasoning but also stabilized its performance and reduced common-sense errors.

To address the issue of entity forgetting in long-context tasks (a typical paper might contain around 5000 tokens), we employed a sliding window technique. Each window, containing 10 text elements with a 5-element overlap, allowed GPT-4 to process and evaluate each text element within a manageable context size. The hyperparameters, window length and overlap size, were optimized using the development set. This overlapping approach ensures that each text element is evaluated twice, significantly reducing the likelihood of false negatives.

Figure 9 illustrates the final prompt template used to match text elements (such as sentences, tables, figure captions, etc.) with the aspects. Each template inputs one aspect, the 10 text elements within a sliding window, and the contextual background.

This methodology not only ensures high fidelity in aspect-text alignment but also leverages the model's capabilities to provide consistent and accurate annotations across extensive text bodies.

### D.4 A Motivating Example

Here we provide an example of how we build the support sentence retrieval task.

**Bias**
Selective reporting (reporting bias)

**Support Judgment For the Bias**
The trial authors' original plan was to use Indirect Fluorescent Antibody Test (IFAT) at 6 and 12 months, but due to operational problems, data on IFAT results were not considered valid for the analysis, and serology was not used as a marker of infection in the trial. Problems with serology were poor sensitivity and reproducibility. The authors decided not to use IFAT results in the trial and relied on conversion of the Montenegro skin test (MST) at 18 months of follow-up as the only outcome measure, since no clinical cases of Visceral Leishmaniasis (VL) were detected among the studied population.

**Decomposition of Support Judgment into Aspects**

- Aspect 1: The trial authors' original plan was to use IFAT test at 6 and 12 months
- Aspect 2: Due to operational problems, data on IFAT results were not considered valid for the analysis
- Aspect 3: Serology was not used as a marker of infection in the trial
- Aspect 4: Problems with serology were poor sensitivity and reproducibility.
- Aspect 5: The authors decided not to use Indirect Fluorescent Antibody Test (IFAT) results in the trial
- Aspect 6: The authors relied on conversion of the Montenegro skin test (MST) at 18 months of follow-up as the only outcome measure
- Aspect 7: No clinical cases of VL were detected among the studied population

**Aspect Filtering**
Using prompt 8, Aspect 4 is filtered since it is a commentary of the reviewer.

**Mapping Aspects to Sentences in Report**
Utilizing the procedure described in section D.3, we map the remaining 6 aspects to all sentences in the report (Werneck et al., 2014).

We only show sentences from the report that are matched with at least one aspect.

- Sentence 4: The main outcome is the incidence of infection assessed by the conversion of the Montenegro skin test (MST) after 18 months of follow-up in residents aged $\geq$1 year with no previous history of visceral leishmaniasis (VL). (**Mapped with aspect 6**)
- Sentence 66: The original plan was to repeat the IFAT test at 6 and 12 months, but due to operational problems, data on IFAT results were not considered valid for the analysis, and serology was not used as a marker of infection in the study. (**Mapped with aspect 2, 3, 5, 7**)
- Sentence 73: In any case, we decided not to use IFAT results in this study and relied on conversion of the MST at 18 months of follow-up as the only outcome measure, since no clinical cases of VL were detected among the studied population. (**Mapped with aspect 1, 3, 5, 6**)

**Aspect Recall Ratio @ Optimal**
One of our evaluation metric, Aspect Recall Ratio @ Optimal, measures the the optimal number of sentences to cover all aspects. In this example, we only need two sentences, sentence 66 and sentence 73, to cover all aspects.

Given the context of a review statement about a biomedical paper, determine if the statement is:

1. A general summary/results of the paper

2. Only contains specific information from the paper

3. Contains the reviewer's comment about the paper

The statement:
**{aspect}**

The context of the statement:
**{support_judgement}**

Give your reasoning, then output your decision of the statement type 1, 2 or 3 after the keyword "DECISION:"

Figure 8: Prompt for filtering non-specific commentaries.

A reviewer evaluates a scientific paper for this bias: **{bias_name}**. You are given a one-sentence statement from the reviewer's assessment of the paper. You are also shown the entire assessment as context. You are given a list of indexed text elements from a paper (including rows of table data, table captions, figure captions, section titles, sentences, others). Text element is followed by its index, in this format, index: text element [End of text element index].

The statement contains information specific to the paper. Focus solely on the one-sentence statement alone, and Go through each text element in the list independently, determine if the text element itself alone contains ALL of the information present in the statement. Note, information in the statement and in the paper can be presented in different formats, subject to rounding, derivation from simple calculation, spelling, acronyms, different wording, and summarization.

Statement:
**{aspect}**

The entire assessment as context:
**{support}**

Paper's list of text elements:
**{cand_pool}**

Make sure you go through the entire list of text elements, no matter how long. If you did not find any text element, explicitly explain why you cannot find all of the information of the statement from any of the text elements. For each text element you choose because you think it contains all of the information present in the statement, write to explain why. If you fail to provide such a satisfactory explanation, you should not include that text element. Output all your explanation and reasoning after the keyword "REASON:"

Once you finalize your selection, you should include the numerical index of each text element you choose. Finally, output all chosen numerical index or indices in a python list [index1, index2, ...] after the keyword "DECISION:"

Figure 9: Prompt for generating synthetic options.

## E Task-specific Fine-tuning

We fine-tune Llama-3-8B on the three train sets of RoBBR's three tasks using LoRA (Hu et al., 2021), on 1 node of 8 NVIDIA H-100. Our fine-tuning hyperparameters are standard: LoRA rank $= 8$, LoRA alpha $= 16$, a batch size $= 16$, with the AdamW optimizer (weight decay $= 0.01$), and a cosine learning rate scheduler with 10% warmup. Llama-3-8B is fine-tuned for 1 epoch using cross entropy loss. The model is asked to predict the label (i.e., risk-level for main task, sentence indices for SSR, correct support judgment for SJS). We compared different fine-tuning techniques, and found when we train model to predict both the label and a paragraph of reasoning text, the optimal training loss is achieved on the train sets, since it increases the amount of training tokens directly go through cross entropy loss (Hsieh et al., 2023). We use support judgment as proxy for reasoning text.

## F Analysis: Two Types of Long-Context Problems

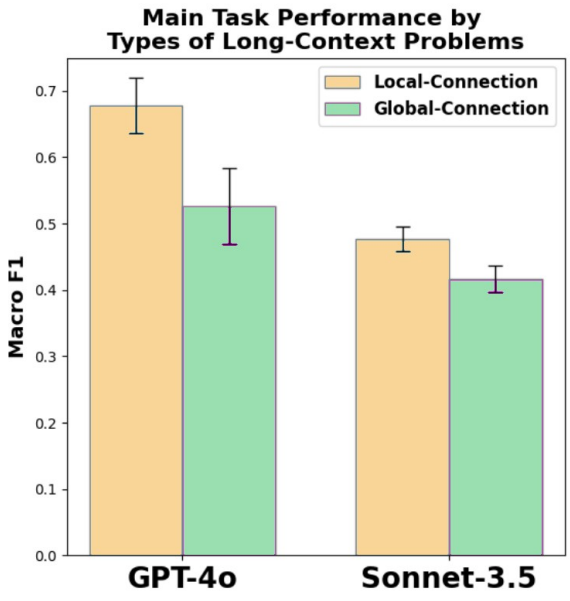


Figure 10: Local-connection problems are easier for LLMs.

An active research direction in long-context modeling is to categorize different types of long-context problems (Li et al., 2024). Using our annotation pipeline's Aspect-to-Sentence mapping, we can naturally categorize RoBBR's datapoints into two types by the optimal number of support sentences used to cover all aspects.

**Local-Connection Problem:**
If one support sentence can cover all aspects that form the support judgment, it means this support sentence is all a model needs to reach the same support judgment. While the main task for this datapoint is still a long-context problem because RoBBR's average document length exceeds 8k tokens, it can be solved by locating one specific sentence from the report. There are 213 local-connection datapoints in RoBBR.

**Global-Connection Problem:**
If the support judgment cannot be traced back to one single sentence, some degree of global understanding of the report is required. There are 693 global-connection datapoints in RoBBR.

In Figure 10, we calculate the average LLM accuracy per type, and found evidence that suggests local-connection problems are easier for both GPT-4o and Sonnet-3.5. The intuition is that a problem becomes harder when a more global understanding of the entire long context is required.

## G Error Analysis

To better illustrate that **GPT-4o** and **Sonnet-3.5** possess different reasoning strengths, we compare their reasoning traces when both are given the same ground-truth retrieval results.

> **Bias:** Allocation concealment (selection bias)
>
> **Bias Definition and Selection Criteria for Risk Level:**
>
> Selection bias (biased allocation to interventions) due to inadequate concealment of allocations prior to assignment.
>
> low risk: Participants and investigators enrolling participants could not foresee assignment because one of the following, or an equivalent method, was used to conceal allocation: Central allocation (including telephone, web-based and pharmacy-controlled randomization); Sequentially numbered drug containers of identical appearance; Sequentially numbered, opaque, sealed envelopes.
>
> high risk: Participants or investigators enrolling participants could possibly foresee assignments and thus introduce selection bias, such as allocation based on: Using an open random allocation schedule (e.g. a list of random numbers); Assignment envelopes were used without appropriate safeguards (e.g. if envelopes were unsealed or non-opaque or not sequentially numbered); Alternation or rotation; Date of birth; Case record number; Any other explicitly unconcealed procedure.
>
> unclear risk: Insufficient information to permit judgement of 'Low risk' or 'High risk'. This is usually the case if the method of concealment is not described or not described in sufficient detail to allow a definite judgement – for example if the use of assignment envelopes is described,

but it remains unclear whether envelopes were sequentially numbered, opaque and sealed.

**Retrieved Sentence:**

Patients were randomly assigned to the control or experimental group following a simple randomization procedure (computerized random numbers) achieved using opaque envelopes.

Physical therapy was started under consultation with the rehabilitation medicine department after wound healing, and the allocation information to each group was not provided to reduce bias.

**Sonnet-3.5:**

REASON: The study describes using a computerized random number generator for randomization, which is a good start for reducing selection bias. They also mention using opaque envelopes, which aligns with one of the low-risk criteria (sequentially numbered, opaque, sealed envelopes). However, the description doesn't explicitly state that the envelopes were sequentially numbered or sealed. Additionally, while they mention that allocation information was not provided to reduce bias, it's not entirely clear if this refers to blinding after assignment or concealment before assignment.

The use of opaque envelopes suggests an attempt at concealment, but without the full details of how these envelopes were used (sequentially numbered, sealed), we can't be certain that the concealment was adequate to prevent foresight of assignments. The computerized randomization is a positive factor, but the implementation of the envelope system is crucial for proper concealment.

Given these factors, there is some evidence of attempts to conceal allocation, but not enough information to confidently classify it as low risk. At the same time, there's no clear indication of high risk practices.

DECISION: unclear

**GPT-4o:**

REASON: The study describes the use of opaque envelopes for allocation concealment, which is a method that can prevent participants and investigators from foreseeing assignments, provided that the envelopes are sequentially numbered, opaque, and sealed. The highlighted sentences indicate that opaque envelopes were used, but do not specify whether they were sequentially numbered and sealed. However, the mention of using opaque envelopes suggests an attempt to conceal allocation, and the additional information that allocation information was not provided to reduce bias further supports the intention to prevent foreseeing assignments. Despite the lack of explicit detail about sequential numbering and sealing, the description leans towards a method that is generally considered adequate for allocation concealment.

DECISION: low

The provided example demonstrates that Sonnet-3.5 is better in reasoning compared to GPT-4o. Sonnet-3.5 correctly identifies critical information gaps in the allocation concealment process—while opaque envelopes were used, there's no confirmation they were sealed and sequentially numbered, which are important for allocation concealment according to the decision criteria. Sonnet-3.5 also recognizes that the mention of withheld allocation information could indicate post-assignment blinding rather than proper pre-assignment concealment. In contrast, GPT-4o fails to identify these crucial gaps and makes unwarranted assumptions about concealment adequacy without supporting evidence for key criteria. This example clearly illustrates Sonnet-3.5's better reasoning capabilities.

## H SJS Synthetic Option Generation Procedure and Prompt

We first generate three detailed synthetic options that are specifically crafted to imitate the support judgments from other papers concerning the same type of bias, ensuring that they are tailored to be relevant to the specific paper in question while maintaining the foundational reasoning. See Figure 11 for the prompt template. Following this, we condense these detailed options into shorter versions that preserve the original meaning, using prompt 12. To prevent heuristic algorithms from solving the task easily, we randomly select either the long or short version of each synthetic option to include in the multiple choice questions.

For selecting the three options derived from other papers' support judgments within the same bias category, we use prompt 13.

Finally, once the six incorrect options are constructed, we conduct a manual review with the help of prompt 14 on all data points. This review ensures that the options are incorrect and not false negatives.

A reviewer evaluates a biomedical paper for the bias: **{bias_name}**. You are a researcher designing a multiple-choice question to test if someone can select the correct support judgment for the bias. You are given the reviewer's support judgment for the bias as the correct answer.

You are also given a list of support judgments of the same bias but for other papers. Your job is to construct three incorrect choices using this list of support judgments following these criteria:

1. The incorrect choices must not entail correct support judgment. In other words, you need to make sure these choices are not false negatives.

2. You should choose the incorrect choices from the list so that their reasonings differ from the correct support judgment. In other words, you should not choose semantically similar choices.

3. You are given the biomedical paper and you should modify the incorrect choices to be paper-specific so that it is more deceptive. Make sure your modification does not turn the incorrect choice into a false negative.

The correct support judgment:
**{support_judgement}**

The list of other support judgments:
**{other_judgements}**

The biomedical paper:
**{full_paper}**

Obey the criteria for constructing the incorrect choices. First, you must find three incorrect support judgements from the list and make sure their reasonings differ from the correct support judgment. You must further check to make sure they do not entail the correct support judgment, i.e. from an incorrect support judgment you cannot infer the correct support judgment, and the incorrect support judgment does not contain all information in the correct support judgment.
Second, you must verify that each incorrect choice is infused with paper-specific information to make them more deceptive. Importantly, this infusion cannot and must not make the incorrect choice contain all information in the correct support judgment. Output your step by step verification and reasoning after the keyword "REASON:".
Finally, write and output your constructed incorrect choices in a python list format [incorrect choice1, incorrect choice2, incorrect choice3] after the keyword "DECISION:"

Figure 11: Prompt for generating synthetic options.

You are evaluating a paper's risk level for a certain bias: **{bias_name}**. You are given four different judgments. Your job is to make the last three judgments much shorter and concise, while retaining its central meaning and judgment, and imitate the style of the first judgment whenever possible.

Four support judgments:
{**four_choices}**

Explain how you rewrite the last three support judgments, and output your reasoning after the keyword "REASON:"

Finally, write and output the three rewritten judgments in a python list format [judgment2, judgment3, judgment4] after the keyword "DECISION:"

Figure 12: Prompt for shortening the synthetic option.

A reviewer evaluates a biomedical paper for the bias: **{bias_name}**. You are a researcher designing a multiple-choice question to test if someone can select the correct support judgment for the bias. You are given the reviewer's support judgment for the bias as the correct answer.

You are also given a list of support judgments of the same bias but for other papers. Your job is to choose up to three incorrect choices from this list of support judgments following these criteria:

1. The incorrect choice must not entail the correct support judgment, and it must not contain ANY information present in the correct support judgment.

2. If you cannot find any support judgment that is significantly different from the correct support judgment, output an empty list.

The correct support judgment:
**{support_judgement}**

The list of other support judgments:
**{other_judgements}**

If you cannot find any incorrect choices, explicitly explain why all choices entail or contain some information present in the correct judgment. For all incorrect choices you find (up to three choices), explain why these choices do not entail or contain ANY information present in the correct judgment. Output your explanation and reasoning after the keyword "REASON:"

Finally, write and output your chosen incorrect choices in a python list format [incorrect choice1, incorrect choice2, incorrect choice3] after the keyword "DECISION:"

Figure 13: Prompt for finding negative options from other papers' support judgments of the same bias.

A reviewer evaluates a biomedical paper for the bias: **{bias_name}**. You are given one correct support judgment for the bias and six other support judgments. Your task is to find any support judgment from the six other potential judgments that follows these criteria:

1. The selected judgment must entail the correct support judgment. In other words, the selected judgment must contain ALL information and ALL detail present in the correct support judgment.

2. The selected judgment can contain additional information not present in the correct support judgment.

3. If you do not find any judgment that meets the criteria, output an empty list.

The correct support judgment is:
**{correct}**

The six potential candidates are:
**{six_other}**

Before making the decision, explain your evaluation of each of the six judgments, and provide your reasoning after the keyword "REASON:"

Finally, write and output the selected judgment(s) in a python list format after the keyword "DECISION:"

Figure 14: Prompt for filtering false negatives.

# I Experiment Details

## I.1 Prompt Optimization

All evaluation prompts are optimized using disjoint development set for each task. Each prompt includes specialized instructions designed to elicit chain-of-thought reasoning, thereby enhancing the models' reasoning abilities. These instructions are repeated twice: once at the beginning and once at the end of the prompt, ensuring that the models retain the instructions even after processing the entire paper. Empirical evidence shows that few-shot in-context learning does not improve model performance. Consequently, evaluation prompts do not incorporate in-context learning. It is important to note that prompt optimization is not tailored to any specific large language model; instead, the aggregated performance across all models is used to guide the optimization of evaluation prompts.

## I.2 Prompts for Evaluation

### I.2.1 Prompt to Evaluate Support Sentence Retrieval (SSR)

See Figure 15 for prompt to evaluate Support Sentence Retrieval (SSR). We use an additional multi-turn prompt 16 when the model outputs more than required.

### I.2.2 Prompt to Evaluate Support Judgment Selection (SJS)

See Figure 17 for prompt to evaluate Support Judgment Selection (SJS).

### I.2.3 Prompt to Evaluate Risk-of-Bias Determination (Main Task)

See Figure 18 for prompt to evaluate Risk-of-Bias Determination (Main Task).

## I.3 Prompt for Risk-of-Bias Determination (Main Task) Given Retrieval Results

See Figure 19 for prompt for Risk-of-Bias Determination (Main Task) given retrieval results.

## I.4 Prompt for OpenAI o4-mini to identify support sentences

See Figure 20 for prompt for OpenAI o4-mini to identify support sentences

## I.5 Results with Standard Error

We provide bootstrapped standard errors for the experimental results in Tables 13,14,15.

You will be given the following pieces of data:
1. The objective of a systematic review
2. The characteristics of a study, which is contained within a biomedical paper
3. The full text of the paper, in the format of an indexed list of text elements (including sentences, tables, figure captions and others).
4. The definition and selection criteria of risk levels for a type of bias: **{bias_name}**.

Your task is to look through the entire paper and locate text elements that are directly relevant to this bias: **{bias_name}**. Specifically, you must find information from the paper that is relevant to determining its level (low, unclear or high) of this particular bias: **{bias_name}**. In other words, you should find text elements that would support a judgment of high risk of bias or low risk of bias, or text elements that would explain why the risk level is unclear for this bias, or a combination of text elements that are relevant to different risk levels of the bias.

You must not have preconceived judgment on this study's level of risk for this bias. You must go through the entire paper in an objective manner, and not miss any text element that could be indicative of low, unclear or high risk of the bias.

Very importantly, your task is to find NO MORE than **{number_of_maximally_allowed_text_elements}** text elements. Since you are only allowed to find a limited amount of text elements, you must only select the most important text elements that together would cover the most amount of information relevant to determining this study's risk level for this bias.

Objective of Systematic Review:
**{objective}**

Bias Name:
**{bias_name}**

Bias Definition and Selection Criteria for Level of Risk:
**{bias_definition_selection_criteria}**

Characteristics of the Study:
**{PICO}**

Indexed list of text elements of the Paper:
**{indexed_paper}**

Provide an explanation of how the details you identified from the paper would jointly provide the most effective basis for judgment of the risk level for this bias. Output your reasoning first, and based on your reasoning, make your final selections of text elements. You should only keep the most important text elements. Make sure you do not find more than **{number_of_maximally_allowed_text_elements}** text elements. If you do, you have to choose the most important **{number_of_maximally_allowed_text_elements}** elements from it that jointly would provide the greatest amount of information that can support an informed determination of the risk level for this bias: **{bias_name}**.

Finally, output the list of indices of these text elements in a python list [index1, index2, ...] after the keyword "DECISION:"

Figure 15: Prompt for Support Sentence Retrieval (SSR) evaluation.

**{Previous conversation}**

{'role': 'user', 'content': You have output more than **{number_of_maximally_allowed_text_elements}** text elements, which violates the requirement to find no more than **{number_of_maximally_allowed_text_elements}** number of text elements. You must choose the best **{number_of_maximally_allowed_text_elements}** text elements in the same format as the previous output.}

Figure 16: Prompt for Support Sentence Retrieval (SSR) evaluation when the model output more than required.

| Model | Avg | Bias Type | | | | |
|---|---|---|---|---|---|---|
| | | Selection n = 157 | Attrition n = 46 | Performance n= 54 | Detection n = 61 | Reporting n = 14 |
| OpenAI-v3 | 22.72 | 40.13 ± 3.56 | 6.3 ± 2.72 | 26.67 ± 5.29 | 27.38 ± 4.81 | 13.1 ± 7.56 |
| GritLM-7B | 18.87 | 35.84 ± 3.59 | 6.99 ± 2.67 | 15.19 ± 3.76 | 26.83 ± 5.08 | 9.52 ± 7.09 |
| GPT-4o | **47.48** | 60.5 ± 3.54 | **42.36 ± 6.47** | 41.48 ± 6.06 | 50.05 ± 5.7 | **43.03 ± 11.55** |
| Sonnet-3.5 | 39.17 | 55.42 ± 3.62 | 30.65 ± 5.71 | 37.01 ± 5.69 | 37.21 ± 5.5 | 35.54 ± 11.55 |
| Llama-3.1-70B | 45.62 | 61.17 ± 3.48 | 34.21 ± 5.95 | 45.19 ± 5.96 | **50.11 ± 5.68** | 37.41 ± 11.49 |
| Llama-3-8B | 22.66 | 49.44 ± 3.68 | 14.49 ± 4.86 | 22.13 ± 4.95 | 20.77 ± 4.7 | 6.46 ± 4.4 |
| Llama-3-8B Fine-Tuned | 40.8 | **69.01 ± 3.41** | 24.78 ± 5.85 | **47.28 ± 6.26** | 48.63 ± 5.92 | 14.29 ± 9.35 |

Table 13: Results for Sentence Retrieval Standard Error Included.

You will be given the following pieces of data:
1. The objective of a systematic review
2. The characteristics of a study, which is contained within a biomedical paper
3. The full text of the paper, in the format of an indexed list of text elements (including sentences, tables, figure captions and others).
4. The definition and selection criteria of risk levels for a type of bias: **{bias_name}**.

Your task is to select the argument that would best inform a human reviewer about how to reach the correct determination for this study's level of risk for this bias: **{bias_name}**. Note, the correct judgment might not contain specific details about the study or the paper. You must base your selection only on the correctness and informativeness of the judgment.

Objective of Systematic Review:
**{objective}**

Bias Name:
**{bias_name}**

Bias Definition and Selection Criteria for Level of Risk:
**{bias_definition}**

Characteristics of the Study:
**{PICO}**

Full text of the Paper:
**{full_paper}**

List of arguments to choose from (Must only select ONE).
**{options}**

You must make sure your chosen argument is reasonable and correctly reflects the limitations, or the lack of limitations, within this study described in the full text of the paper. You must further ensure your chosen judgment follows the criteria for this bias: **{bias_name}**. Note, the correct argument might not contain specific details of the study or the paper. Output your reasoning first, after the keyword "REASON:"

Finally, based on your reasoning, only output one single capitalized Letter Choice (e.g. C) of your chosen judgment, after the keyword "DECISION:"

Figure 17: Prompt for Support Judgment Selection (SJS) evaluation.

You are presented with the objective of a systematic review. You are also presented with the study characteristics of a study contained in the full text of a paper. Your task is to determine the risk level of this study for a particular bias: **{bias_name}**. There are three levels of risk for this bias, "low", "unclear", "high". You should choose one of the risk levels following the definition and selection criteria of the bias and risk levels, if applicable.
Objective of Systematic Review:
**{objective}**

Bias Name:
**{bias_name}**

Bias Definition and Selection Criteria for Risk Level:
**{bias_definition_selection_criteria}**

The Characteristics of a Study:
**{PICO}**

The Full text of the paper that contains the aforementioned study.
**{full_paper}**

Follow the definition and selection criteria of the bias and risk level, locate relevant information of the study from the paper to help you determine this study's level of risk for **{bias_name}**, whether it is "low", "unclear" or "high". Output your reasoning after the keyword "REASON:"

Finally, based on your reasoning, make your decision and output whether this study has "low", "unclear" or "high" risk for **{bias_name}**. Output "low", "unclear" or "high" after the keyword "DECISION:"

Figure 18: Prompt for Risk-of-Bias Determination (Main Task) evaluation.

You are presented with highlighted sentences from a biomedical paper. Your task is to determine the risk level of this study for a particular bias: **{bias_name}**. There are three levels of risk for this bias, "low", "unclear", "high". You should choose one of the risk levels following the definition and selection criteria of the bias and risk levels, if applicable.

Bias Name:
**{bias_name}**

Bias Definition and Selection Criteria for Risk Level:
**{bias_definition}**

Highlighted Sentence(s):
**{retrieved_sentences}**

Now you must determine this study's level of risk for **{bias_name}**. Output your reasoning after the keyword "REASON:"

Finally, based on your reasoning, make your decision and output whether this study has "low", "unclear" or "high" risk for **{bias_name}**. Output "low", "unclear" or "high" after the keyword "DECISION:"

Figure 19: Prompt for Risk-of-Bias Determination (Main Task) Given Retrieval Results.

You will be given the following pieces of data:
1. The objective of a systematic review
2. The full text of the paper, in the format of an indexed list of text elements (including sentences, tables, figure captions and others).
3. The definition and selection criteria of risk levels for a type of bias: {**bias_name**}.
4. The risk level of the bias

Your task is to look through the entire paper and locate text elements that are directly relevant to this bias: {**bias_name**}. Specifically, you must find information from the paper that is relevant to determining its level {**risk_level**} of this particular bias: {**bias_name**}. In other words, you should find text elements that would support a judgment of {**risk_level**} risk of bias.

You must go through the entire paper in an objective manner, and not miss any text element that could be indicative of {**risk_level**} risk of the bias.

As a demonstration, you will first see three examples, each example will have a sample bias name and a sample list of text elements that are representative and together cover the most amount of information relevant to determining a study's risk level of bias.

Sample Bias A: high risk of bias: {**bias_name_1**}
Sample List of Text Elements A: {**text_ele_1**}

Sample Bias B: low risk of bias: {**bias_name_2**}
Sample List of Text Elements B: {**text_ele_2**}

Sample Bias C: unclear risk of bias: {**bias_name_3**}
Sample List of Text Elements C: {**text_ele_3**}

Objective of Systematic Review:
{**objective**}

Bias Name:
{**bias_name**}

Bias Definition and Selection Criteria for Level of Risk:
{**bias_definition**}

Risk Level of the Bias:
{**risk_level**}

Indexed list of text elements of the Paper:
{**full_paper**}

Provide an explanation of how the details you identified from the paper would jointly provide the most effective basis for judgment of the {**risk_level**} risk for this bias. Output your reasoning first, and based on your reasoning, make your final selections of text elements.

Finally, output the list of indices of these text elements in a python list [index1, index2, ...] after the keyword "DECISION:"

Figure 20: Prompt for OpenAI o4-mini to identify support sentences.

| Model | Bias Type | | | | | |
|---|---|---|---|---|---|---|
| | Full<br>n = 310 | Selection<br>n = 86 | Attrition<br>n = 60 | Performance<br>n = 52 | Detection<br>n = 56 | Reporting<br>n = 50 |
| GPT-4o | 47.17 | 58.46 ± 4.37 | 60.2 ± 5.04 | 48.28 ± 5.24 | 42.68 ± 5.29 | 26.25 ± 4.7 |
| Sonnet-3.5 | **59.92** | **73.08 ± 3.89** | **73.47 ± 4.28** | **50.57 ± 5.42** | **51.22 ± 5.44** | **51.25 ± 5.26** |
| Llama-3.1-70B | 53.16 | 66.15 ± 3.89 | 62.24 ± 4.98 | 44.83 ± 5.19 | 46.34 ± 5.64 | 46.25 ± 5.7 |
| Llama-3-8B | 26.54 | 26.92 ± 3.87 | 34.69 ± 4.77 | 24.14 ± 4.64 | 21.95 ± 4.69 | 25.0 ± 4.91 |
| Llama-3-8B Fine-Tuned | 29.6 | 40.77 ± 4.36 | 28.57 ± 4.37 | 24.14 ± 4.42 | 19.51 ± 4.52 | 35.0 ± 5.38 |

Table 14: Results for Support Judgment Selection Standard Error Included.

| Model | Avg | Bias Type | | | | | |
|---|---|---|---|---|---|---|---|
| | | Selection<br>n = 933 | Attrition<br>n = 629 | Performance<br>n = 309 | Detection<br>n = 645 | Reporting<br>n = 594 | Deviation<br>n= 331 |
| GPT-4o | **42.07** | 50.52 ± 1.87 | 36.48 ± 2.14 | **54.67 ± 2.8** | 43.66 ± 2.02 | **34.43 ± 1.87** | 32.65 ± 2.77 |
| Sonnet-3.5 | 41.93 | **52.83 ± 1.86** | **37.07 ± 2.04** | 50.56 ± 2.75 | 43.29 ± 2.09 | 33.58 ± 1.41 | **34.23 ± 1.96** |
| Llama-3.1-70B | 38.81 | 48.1 ± 1.85 | 31.7 ± 1.94 | 48.04 ± 2.67 | **44.36 ± 2.0** | 31.58 ± 1.33 | 29.04 ± 2.37 |
| Llama-3-8B | 30.05 | 36.37 ± 1.28 | 32.44 ± 1.62 | 39.13 ± 2.59 | 37.21 ± 1.92 | 19.76 ± 1.96 | 15.38 ± 1.64 |
| Llama-3-8B Fine-Tuned | 36.33 | 49.51 ± 1.91 | 33.34 ± 1.89 | 42.44 ± 2.69 | 40.1 ± 2.01 | 26.95 ± 0.44 | 25.67 ± 1.06 |

Table 15: Results for Risk Level Determination Standard Error Included.