Zero- and Few-Shot Generative Large Language Models Are Weak Assessors of Risk of Bias in Clinical Trials

SIMON ŠUSTER, School of Computing and Information Systems, The University of Melbourne, Australia
TIMOTHY BALDWIN, Department of Natural Language Processing, MBZUAI, United Arab Emirates
KARIN VERSPOOR, School of Computing Technologies, RMIT University, Australia

ABSTRACT

Existing systems for automating the assessment of risk-of-bias (RoB) in medical studies are supervised approaches, requiring substantial training data to work well. However, recent revisions to RoB guidelines have resulted in a scarcity of available training data. In this study, we investigate the effectiveness of generative large language models (LLMs) for assessing RoB. Their application requires little or no training data and, if successful, could serve as a valuable tool to assist human experts during the construction of systematic reviews. Following Cochrane's latest guidelines (RoB2) designed for human reviewers, we prepare instructions that are fed as input to LLMs, which then infer the risk associated with a trial publication. We distinguish between two modelling tasks: directly predicting RoB2 from text; and employing decomposition, in which an RoB2 decision is made after the LLM responds to a series of signalling questions. We curate new testing datasets and evaluate the performance of four general- and medical-domain LLMs. The results fall short of expectations, with LLMs seldom surpassing trivial baselines. On the direct RoB2 prediction test set (n=5,993), LLMs perform akin to the baselines (F1: 0.1–0.2). In the decomposition task setup (n=28,150), similar F1 scores are observed. Our additional comparative evaluation on RoB1 data also reveals results substantially below those of a supervised system. This testifies to the difficulty of solving this task based on (complex) instructions alone. Using LLMs as an assisting technology for assessing RoB2 thus currently seems beyond their reach.

HIGHLIGHTS

What is already known.

- Assessing the risk of bias (RoB) for studies included in systematic reviews is a key task in automated evidence synthesis. The effectiveness of Large Language Models (LLMs) such as ChatGPT in RoB automation has not been fully evaluated yet.

What is new.

- None of LLMs tested can make accurate predictions by following the available RoB guidance during inference. The results are little affected by the bias domain, the pretraining data provenance (general & biomedical), prompting strategy (zero- or few-shot prompting), and LLM type (proprietary & open-source).
- LLMs cannot yet compete with supervised systems, but additional task-specific adaptation may change this conclusion in the future.

Authors' addresses: Simon Šuster, simon.suster@unimelb.edu.au, School of Computing and Information Systems, The University of Melbourne, Melbourne, VIC, Australia; Timothy Baldwin, timothy.baldwin@mbzuai.ac.ae, Department of Natural Language Processing, MBZUAI, Abu Dhabi, United Arab Emirates; Karin Verspoor, karin.verspoor@rmit.edu.au, School of Computing Technologies, RMIT University , Melbourne, VIC, Australia.

1 INTRODUCTION

A key task in the preparation of systematic reviews is determining whether the results of included studies may be affected by biases, e.g., poor randomisation or blinding; a thorough understanding of the biases present in the studies can improve our confidence in the review's conclusions. Assessing such risk has been formalised into the risk-of-bias (RoB) guidance [1] in the context of Cochrane reviews [2]. A number of studies have proposed new datasets and techniques for automated RoB1 assessment [3, 4] as well as researched its practicality, expert acceptability, and reliability [5–11]. These approaches hold promise to expedite the reviewing process by assisting human reviewers, and lead to more current and comprehensive systematic reviews.

However, with the publication of version 2 of the RoB guidance (RoB2) [12], which differs from RoB1 in several important aspects (see Section 2.1), the existing RoB1 approaches are growing increasingly out of date. Since these systems are trained in a supervised manner, a major hindrance to simply retraining them on the data from the new task is the lack of a larger annotated dataset. Although the community of human reviewers has been transitioning to the updated version, the adherence has been incomplete as of yet, and the number of assessments has been modest compared to RoB1. Motivated by recent successes of prompt-based learning for generative large language models (LLMs) [13, 14], including on information extraction in medicine [15] and on summarisation for evidence synthesis [16–18], we turn to exploring the capacity of LLMs to follow the rich guidance available for RoB2. The goal of our work is to determine whether the risk of bias can be accurately predicted given prompts based on the RoB2 guidance but *without* being explicitly trained on large amounts of task-specific annotated data.

A key aspect of the current RoB2 guidance is that it lays out question-specific instructions as well as provides bias-specific decision rules when reaching a judgement. In this way, judgements are deemed more tractable and justifiable. Aligning with this framework, our LLM approach decomposes the problem into a series of lower-level question-answering tasks where the answers to these questions inform the decisions for the risk at each bias domain. This approach diverges from prior work in automated RoB assessment, which formulates the task simply as a mapping between some input text and a risk label.

Most existing attempts at automating RoB assessment concern RoB1 [3, 4, 11, 19]. These are supervised learning models (based on, e.g., SVMs, CNNs, and fine-tuned BERT [20]), in which a trial abstract or full article text is mapped to a binary bias decision according to the Cochrane RoB framework. Typically, these models achieve around 0.7 F1, with some variation depending on the bias domain. The only RoB2 work known to us, developed contemporaneously to ours, is Pitre et al. [21], who analyse ChatGPT's agreement with RoB2 decisions of authors in Cochrane reviews. Their work is a descriptive, statistical study, whereas we follow a predictive approach. Other notable differences are: a) their analysis is done only at the level of bias domains, while we also take signalling questions as the basic answering unit;[1] b) their sample includes only first outcomes from each review and is smaller (157 trials/instances per bias domain vs. around 1,000 instances per domain in our dataset); c) we test LLMs programmatically rather than manually via a GUI, allowing for more flexibility in exploring various prompting strategies and task variations; d) we include a few-shot prompting setup with justification exemplars; and e) compare LLM performance on RoB2 to that of RoB1, and contrast the findings to RobotReviewer [22], which also targets RoB1.

With manual RoB assessment being an arduous process, taking on average 6 hours per study for RoB2 [23], automated approaches offer promise towards a more rapid evaluation process while maintaining accuracy and consistency. The significance of these findings may have implications for the hypothetical feasibility of on demand literature review

[1] Although Pitre et al. [21] recognise this option as a promising research direction, it would be difficult to realise in their approach due to reliance on manual analysis.

generation by LLMs [24, 25]. It may also curb the enthusiasm about the role of LLMs in evaluating the quality of evidence in systematic reviews [26] and even in medical AI more broadly [27].

2 METHODS

2.1 Task description

Assessment of RoB is regarded as an essential component of a systematic review on the effects of an intervention. The most commonly used tool for assessing RoB in randomised trials is the Cochrane risk-of-bias framework, which was introduced in 2008 (RoB1) [1]. RoB1 considers biases arising at different stages of a trial (known as bias domains), chosen on the basis of both empirical evidence and theoretical considerations. The assessments are supported by quotes from sources describing the trial or by justifications written by the assessor.

Following the publication of the revised framework (RoB2) [12], the community of systematic reviewers has been gradually shifting over to the updated guidelines. The newer framework has been developed with a view to improve RoB1 usability and to reflect current understanding of how the causes of bias can influence study results [28]. It is structured into five domains through which bias might be introduced into the result:

- randomisation process,
- deviations from the intended interventions,
- missing outcome data,
- measurement of the outcome,
- selection of the reported result.

The possible judgements are: Low risk, Some concerns, and High risk. The authors are also asked to provide a short justification, as in RoB1. RoB2 introduces the concept of *signalling questions*, which aim to elicit information relevant to an assessment of RoB. Responses to these questions feed into *decision rules* developed to guide users to judgements about the risk of bias. The response options for these questions are: Yes, Probably yes, Probably no, No, and No information.[2] Although we assume that we do not have access to ground-truth answers to signalling questions, we use the questions to break down the complex assessment problem into finer-grained tasks that are answered individually by an LLM. Such decomposition via prompting is conceptually similar to Khot et al. [29]'s modular approach to addressing complex reasoning tasks.

Assessing risk of bias includes a subjective component. The question of annotation reliability has therefore been studied extensively for both RoB1 [30, 31] and RoB2 [32–35], with the latter showing better inter-rater agreement for those authors who are experts in the topic of the review.

2.2 Problem decomposition

One line of work on the applications of LLMs to complex reasoning tasks has explored strategies for basing a decision on intermediate, easier answers first (*chain-of-thought prompting*) [36, 37]. However, when tasks become more complex, a small number of demonstrations of the complex task is not sufficient to learn to perform all necessary reasoning steps. Hence, recent work has proposed *decomposing prompts* into simpler sub-tasks, often without predefined decomposition structure at inference time, which can be generated iteratively by a dedicated decomposition model [29, 38, 39]. Such decomposers can be implemented using some volume of manually-constructed examples or synthetic data. In our case,

[2]The modal expression *probably* does not alter the path taken in the decision rule, i.e. affects the final bias decision.

Number of RoB2 annotations[4]	5,993
Number of studies	218
Number of signalling questions	28,150
Number of Low risk/High risk/Some concerns	4,247/486/1,260
% of Low risk/High risk/Some concerns	71/8/21

Table 1. Data statistics.

the RoB2 signalling-question decision rules can be seen as decomposition models (Section 2.7), where each subquestion to be answered is already given as a node in the decision rule.

2.3 Data construction

We built our RoB2 dataset using a semi-automated approach based on Cochrane reviews spanning 2020–August 2023. We were granted by Cochrane all review manager (RevMan) files with author assessments, from which we selected only those conforming to RoB2 domain names. Together with these annotations, we extracted the study DOIs and imported them into Paperpile [40]. We then automatically downloaded PDFs using institutional access when necessary. Whenever this failed, we tried to obtain the PDFs manually. Finally, we converted the obtained PDFs to a full-text structure in XML using GROBID [41].

Given these full-text XMLs, we extract the Methods sections based on section headings identified by GROBID,[3] which we consider the most relevant section type for RoB assessment. This was necessary to accomodate context-size restrictions in the LLMs (2,000 tokens for FlanT5XL, and 4,000 all other models listed in Section 2.6) we have tested.

To explore how input-size constraints affect the results, we additionally run an experiment where we include entire articles rather than just the Methods sections. In this case, we use a more costly proprietary model that has a longer input context window (16,000), further described in Section 3.4.

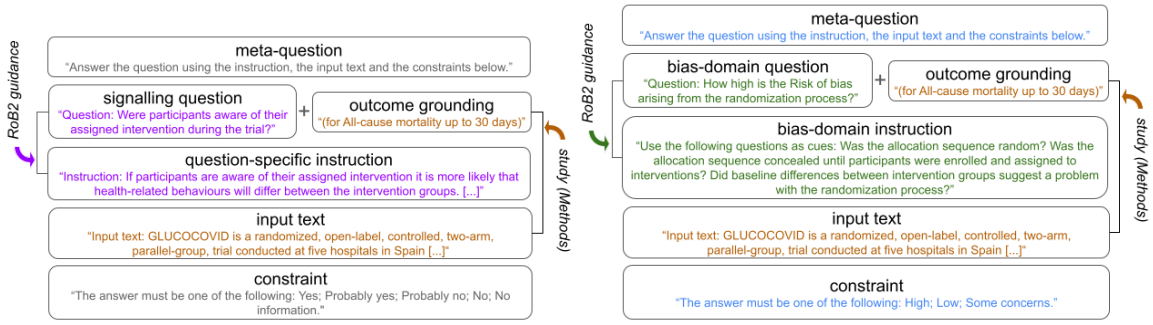
Some statistics of the resulting dataset are shown in Table 1. We release an abridged dataset to the community, formatted as `<study DOI, review DOI, RoB domain, outcome identifier, RoB decision>` [42]. This dataset does not include lengthy text excerpts from either primary studies or Cochrane reviews, but it enables other researchers and practitioners to work on RoB assessment using the primary study publications they have access to.

2.4 Zero-shot prompt creation

We distinguish between predicting RoB with signalling questions (SigQ in the tables of results) and without them, i.e., directly predicting the risk (Direct), which we include for comparison and more closely resembles the setup in previous work on RoB1. In both scenarios, no labelled inputs are provided, constituting a zero-shot setup. During construction, we truncate the prompt text if it exceeds the maximum context length for the intended LLM. We only let truncation affect the input text, keeping all other prompt parts intact. In a development round, we tuned the prompt creation strategy by re-ordering the different parts of the prompts. For this, we created a smaller dataset from three systematic reviews (which were excluded from the larger dataset), amounting to 29 primary studies and 899 data instances.

SigQ prompts are constructed as shown in Figure 1a. All signalling questions with instructions are taken from the guidance document and are specified in the Appendix. The concrete form of a data instance fed to the LLM for inference

[3]More precisely, we take all sections starting with Methods until Results, which may include various Methods sub-sections, as GROBID does not differentiate between section levels.
[4]Initially, the number of RoB2 assessments in the source data was 23,658. We discarded a large part of them as we were unable to either find a corresponding PDF or to convert it to text.



(a) Answering signalling questions. (b) Predicting RoB2 directly.

Fig. 1. Two types of prompt composition. The texts in grey and blue are invariable and occur as such in all generated instances. Other parts of the prompt are sourced from either RoB2 guidance or the study under assessment.

is simply a concatenation of the parts (concrete examples are included in the Appendix). The structure of DIRECT prompts is broadly similar (see Figure 1b), but with the following differences: a) instead of a signalling question, we directly ask about the risk corresponding to the bias domain, and b) we augment that with a sequence of signalling questions belonging to that bias domain. In this way, we give some further instruction to the LLM without diving into specifics of signalling question, which is the goal of the first prompt type.

In addition to the prompt types described above, we include simplified variants of each (SIGQ-SIMPLE and DIRECT-SIMPLE). In these simplified versions, the instruction parts ("question-specific instruction" and "bias-domain instruction", respectively) are omitted, requiring the model to rely solely on the questions. While these prompts lack detailed instructions, the questions themselves nevertheless effectively capture the essence of each bias domain or signalling question aspect. This simplification reduces the length of the prompts and minimises the risk of providing extraneous information to the model is reduced. For example, certain signalling-question instructions may reference other signalling questions (handled separately) or previous works, which could potentially confuse the model.

2.5 Few-shot prompt creation

Due to the limited context size in current LLMs, rather than providing multiple input texts as demonstrations, we follow a different approach. In our data, an RoB judgement can be accompanied by a *justification*, which is written by review authors and often points to a part of the original paper supporting that decision. We provide the justifications as demonstrations of what (or, what piece of text) is important in reaching a decision, which should encourage the LLM to pay attention to those aspects of the study that matter for bias assessment. Therefore, in each prompt, we provide one example justification (together with the true label), for each of the three possible RoB output classes. Such input–output exemplars thus closely resemble the conventional way of preparing few-shot prompts, with the difference that the input text is a subpart (or a paraphrase) of the full-text input. Since in-context learning is known to be sensitive to the order of the demonstrations [43], we try to control for this by randomly sampling and ordering the demonstration examples, following previous work [44].

We only perform few-shot prompting when predicting RoB2 directly but not in the SigQ setup, due to the ease of obtaining labeled examples.

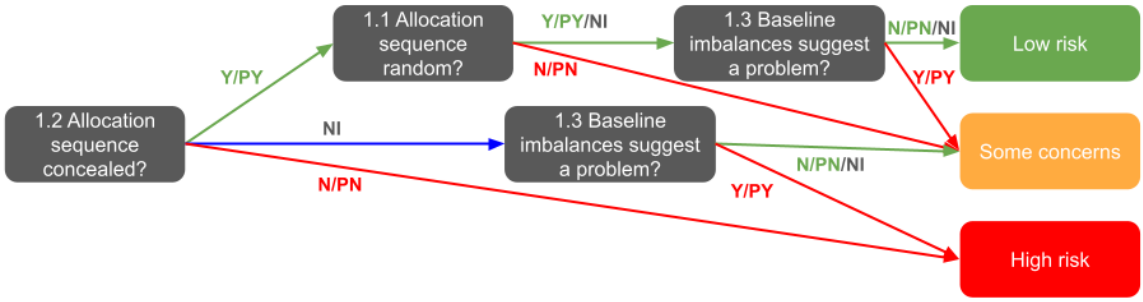


Fig. 2. Decision rule for suggested judgement of risk of bias arising from the randomisation process.

2.6 Models and inference

We conduct the experiments using two general-domain and two medical-domain LLMs. While some of these were fine-tuned for better instruction following, none of the evaluated LLMs have been fine-tuned specifically on RoB tasks.

(1) FlanT5XL [45], an improved T5 model [46] pretrained on a diverse collection of tasks with instruction tuning, which has achieved state-of-the-art zero- and few-shot performance on many QA benchmarks. We prompt this LLM, which has 3B parameters, on an 80GB Nvidia A100 GPU.
(2) ChatGPT (gpt-3.5-turbo)[5], a closed-source proprietary system, which has shown strong performance on a variety of tasks, including text classification [47–49]. We prompt ChatGPT, which has 175B parameters, via the OpenAI API.
(3) Meditron-70B [50], an open-source LLM developed through continued pretraining on LLaMA-2-70B [51] using a medical corpus containing selected PubMed articles, abstracts, a dataset of medical guidelines, and other general domain data. We use four 80GB Nvidia A100 GPUs and the vLLM library for distributed inference [52].
(4) Med42 [53], an open-access LLM with 70B parameters, developed with instruction tuning on data including medical flashcards, exam questions, and open-domain dialogues. Inference is done in the same way as for Meditron-70B.

In addition, we include two trivial baselines: the most frequent label (Freq) and the weighted random (WeighRand) baselines. In the latter, we generate a non-uniform random sample using label probabilities estimated on the dataset.

2.7 Evaluation

Signalling questions are answered by the LLM independently and then combined into a single bias decision using the domain-specific decision rules recommended by the RoB2 authors. As an example, we show in Figure 2 the decision rule for the bias arising from the randomisation process. All machine-readable decision rules are available in our project repository: https://bitbucket.org/aimedtech/fewshot_rob.

We report macro-averaged F1 throughout our results, and add accuracy when comparing to prior work on RoB1.

3 RESULTS

The results in Table 2 show that none of the LLMs outperform the simple baselines, and that all LLMs perform similarly poorly, without a consistent effect regarding the task specification (i.e., SigQ vs. Direct), or whether justification

[5]We use the 4,096-token context variant for all except the full-paper experiments (see Section 3.4), where we use the variant with a 16,385 token window.

model	RP	DII	MOD	MO	SRR
Freq	0.15	0.17	0.16	0.18	0.16
WeighRand	0.18	0.20	0.19	0.18	0.19
SigQ-FlanT5XL	0.02	0.15	0.15	0.13	0.19
Direct-FlanT5XL	0.15	0.17	0.17	0.15	0.14
FewShot-Direct-FlanT5XL	0.17	0.18	0.17	0.18	0.15
SigQ-ChatGPT	0.10	0.11	0.13	0.19	0.11
Direct-ChatGPT	0.14	0.07	0.05	0.05	0.11
FewShot-Direct-ChatGPT	0.11	0.07	0.05	0.05	0.10
Direct-ChatGPT-Long	0.03	0.14	0.12	0.18	0.10
SigQ-Meditron70B	0.20	0.15	0.15	0.12	0.05
Direct-Meditron70B	0.14	0.17	0.18	0.18	0.15
FewShot-Direct-Meditron70B	0.17	0.20	0.17	0.19	0.18
SigQ-Med42	0.06	0.04	0.00	0.05	0.05
Direct-Med42	0.15	0.17	0.16	0.18	0.16
FewShot-Direct-Med42	0.09	0.04	0.03	0.04	0.07

Table 2. Main results for the five RoB2 domains: RP = risk due to randomisation process; DII = risk due to deviations from the intended interventions; MOD = risk due to missing outcome data; MO = risk from measurement of the outcome; and SRR = risk due to selection of the reported result.

exemplars are provided (FewShot). Also, no bias domain stands out as markedly easier compared to the others. Finally, our findings indicate that the LLMs do not fall short due to a lack of domain awareness, as evidenced by the similarly low results observed in medical LLMs. All of this extends the emerging evidence of ChatGPT's unreliable predictions on RoB2 assessment, as noted by [21], who observed only slight to fair agreement between ChatGPT and human reviewers.

3.1 Conflating labels

In this setup, instead of 3-way classification, we limit the outcomes to two classes by merging the High risk and Some concerns labels, similarly to some previous work on RoB1 [4, 54]. We adapt the constraint part of the prompt when testing all Direct models; when evaluating the SigQ models, we just take the predicted answers (i.e., without repeating inference) and map the output derived from a decision rule to two classes. Table 3 shows that the F1 results are higher due to lower label ambiguity as a result of conflation, but as before, we do not observe any strong improvement over the baselines. Prompt simplification (cf. Section 2.4) does not alter the overall results, raising doubts about the models' ability to effectively interpret the provided instructions.[6]

We observed in the predictions of Direct-ChatGPT that it sometimes refrains from making a decision, and instead mentions the lack of information as a reason for being unable to determine the RoB. The number of these "undeterminables" varies per domain, and is the lowest for RoB due to deviations from the intended interventions (4%) and highest for RoB due to missing outcome data (24%). After removing these instances from the evaluation, we observe slight increases in F1 across all bias domains (RP: 0.01, DII: 0.01, MOD: 0.04, MO: 0.01, SRR: 0.02).

[6]For the results from simplified prompts on individual signalling questions, see next section and Table 4.

model	RP	DII	MOD	MO	SRR
Freq	0.37	0.43	0.4	0.45	0.39
WeighRand	0.49	0.50	0.50	0.49	0.50
SigQ-FlanT5XL	0.30	0.40	0.49	0.39	0.49
Direct-FlanT5XL	0.43	0.49	0.43	0.45	0.41
Direct-Simple-FlanT5XL	0.39	0.41	0.26	0.35	0.36
SigQ-ChatGPT	0.32	0.30	0.34	0.48	0.28
Direct-ChatGPT	0.45	0.48	0.46	0.40	0.48
Direct-Simple-ChatGPT	0.43	0.45	0.48	0.46	0.41
SigQ-Meditron70B	0.53	0.38	0.39	0.36	0.33
Direct-Meditron70B	0.49	0.42	0.43	0.47	0.40
SigQ-Med42	0.17	0.12	0.01	0.12	0.12
Direct-Med42	0.12	0.05	0.05	0.03	0.08

Table 3. Results for five RoB2 domains *with label conflation*: RP = risk due to randomisation process; DII = risk due to deviations from the intended interventions; MOD = risk due to missing outcome data; MO = risk from measurement of the outcome; and SRR = risk due to selection of the reported result.

	F1			
model	overall	NI	NPN	YPY
WeighRand	0.30	0.21	0.32	0.33
SigQ-FlanT5XL	0.20	0.00	0.39	0.19
SigQ-Simple-FlanT5XL	0.33	0.00	0.47	0.53
SigQ-ChatGPT	0.36	0.36	0.48	0.24
SigQ-Simple-ChatGPT	0.35	0.22	0.52	0.32
SigQ-Meditron70B	0.31	0.10	0.36	0.50
SigQ-Med42	0.18	0.36	0.10	0.10

Table 4. Answering performance on signalling questions. Label key: No information (NI), Yes and Probably yes (YPY), No and Probably no (NPN).

3.2 Evaluation at the level of signalling-questions

Since the gold labels in our dataset only pertain to the risk for each bias domain, we cannot directly evaluate the quality of answers to questions leading to bias-level decisions. To get around this, we searched for systematic reviews containing SQ decisions as supplementary material. Having identified two systematic reviews, one on viral infection prevention measures [55] and another on medication adherence in hypertensive patients [56], we extracted the answers to signalling questions for a total of 9 studies. This serves as a small SQ test set, to which we apply our LLMs and evaluate them.

The results presented in Table 4 indicate that, for some output labels, the predictions of esp. ChatGPT and Meditron70B outperform random guessing. Still, none of the LLMs exhibit strong performance across *all* labels. Interestingly, both ChatGPT and Med42 predict equally well when to abstain (by answering "No information"), surpassing other models in this regard.

Further, evaluation at the level of SQs opens up the possibility to analyse the difficulty of different questions. We show in Figure 3 (only for FlanT5XL, but all LLMs show a comparable trend) that the accuracy for each question is

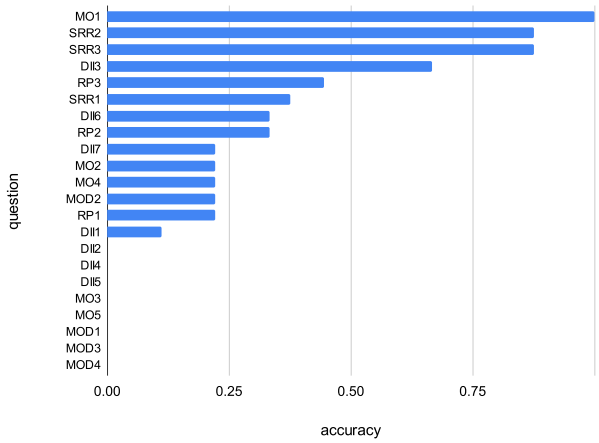


Fig. 3. Answering accuracy on signalling questions for the FlanT5XL model. Refer to the Appendix for the expanded question names.

highly variable. Upon contrasting the performance to the instructions through manual analysis, we can provisionally explain some of that variation. Namely, for the question MO1 (*Was the method of measuring the outcome inappropriate?*), the model appears to correctly comply with the instructions: *In most circumstances, for pre-specified outcomes, the answer to this question will be 'No' or 'Probably no'*, leading to perfectly accurate predictions. For RP2 (*Was the allocation sequence concealed until participants were enrolled and assigned to interventions?*), the instructions are again revealing (*This level of detail is rarely provided in reports, and a judgement may be required*), giving us a clue to model's confusion (low accuracy) on this question. The difficulty of extracting the right information from text is well illustrated by the model's performance on the Deviations from assigned interventions (DII) domain. For the questions DII1 and DII2, the model cannot pick up whether participants or carers were aware of the assigned intervention based on the methods description from the report. Similarly, it tends to err for DII3–5 on deviations from intended interventions, showing overconfidence where there is really no information in the report on which to base a judgement. For example, for deviations arising because of the trial context (DII3), the instructions are clear that the answer should be NI whenever there is no evidence of deviations from trial protocol. As no information about the trial protocol was included in the text, the answer should clearly be NI. In light of this, it is not surprising that the overall performance for Deviations from assigned interventions (DII) does not exceed the trivial baselines (0.15 F1, Table 2).

Also more generally, we observe that the model never predicts 'No information' although it should be the expected answer to 35 questions (out of 155 in the test set). Apart from the DII domain mentioned above, we find that the model is excessively confident also for bias due to randomisation process (RP) — for example, for RP2 the texts lack background information about baseline differences that would allow yes/no answers, yet the model invariably answers with a negative answer rather than with 'No information'.

	Accuracy (F1)		
	RobotReviewer	FlanT5XL	FlanT5XL-Simple
RSG	0.72 (0.67)	0.49 (0.32)	0.43 (0.41)
AC	0.67 (0.67)	0.45 (0.25)	0.49 (0.35)
BPP	0.75 (0.74)	0.44 (0.27)	0.64 (0.57)
BOA	0.64 (0.64)	0.52 (0.32)	0.62 (0.58)

Table 5. Performance on the RoB1 dataset for four RoB1 domains: Random sequence generation (RSG), Allocation concealment (AC), Blinding of participants and personnel (BPP), and Blinding of outcome assessment (BOA).

3.3 Performance on RoB1

To better understand the low performance observed in RoB2 assessment, we explored whether switching to the older RoB1 task and its corresponding guidance [57] would lead to different findings.

In preparing the RoB1 data, we follow the procedure described in Šuster et al. [11] to obtain about 3,000 instances per bias domain from open-access publications. The included bias domains are random sequence generation (with a proportion of low-risk evidence of 70% when considering binary outcomes), blinding of outcome assessment (51% low-risk), allocation concealment (57% low-risk), and blinding of participants and personnel (41% low-risk). This setting allows a direct comparison to the test results obtained on this dataset with RobotReviewer [19], a supervised system, as reported by Šuster et al. [11]. To construct the prompts for this task, we adopt the layout of Figure 1b but change the bias-domain questions and instructions to reflect RoB1, while removing the outcome grounding part. See Appendix for more examples.

We test only the FlanT5XL model in the zero-shot setup, using label conflation consistent with prior work on RoB1. The results in Table 5 reveal a stark contrast between RobotReviewer and FlanT5XL, with the latter trailing far behind. We note that the F1 scores of FlanT5XL are in a similar range as observed in the RoB2 experiments.

3.4 Expanding the input text

While we have limited the input text used in prompts to the Methods sections in primary studies, which we deem the most relevant section type for RoB assessment, a valid concern is that RoB decisions sometimes may not be derivable from the text if the information is missing. However, the information could still be present in other parts of the paper, or even outside of it.

A bias domain where it should be clearly beneficial to incorporate other sources of information is RoB in selection of the reported result, where the content under Methods can be contrasted to the results actually reported in sections such as Results, Discussion, and Conclusion. We therefore adapt our prompting strategy by incorporating all available article sections instead of the Methods only. This increases the prompt lengths, so we run the test using the gpt-3.5-turbo-16k variant, which can accommodate context sizes up to 16k tokens. In this way, as many as 99.6% instances fit without requiring truncation.[7] Although the results (Direct-ChatGPT-Long in Table 2) slightly improve for certain bias domains, they are still in the same range as the trivial baselines.

[7]The mean prompt length is 6,600 tokens.

4 DISCUSSION

We now analyse potential factors contributing to the observed poor performance of LLMs on the RoB2 assessment, addressing concerns such as unclear guidance, distractability, and missing external evidence. We additionally explore possible areas for improvement, including the acquisition of ground-truth information, extended few-shot prompting, and the consideration of LLM fine-tuning for improved predictive capabilities.

Unclear guidance. One possibility for the low observed performance of LLMs on RoB2 assessment is that the guidance is insufficient, unclear, or too abstract, which is a concern already raised in the reviewing community [23].[8] However, the inter-rater reliability studies [34, 35] appear to alleviate these concerns by showing that high agreement is attainable provided that the authors possess content and methodological expertise. In addition, our results from testing on RoB1 with different guidelines provide evidence that the RoB2 guidance is unlikely to be the (only) culprit.

Distractability. Another possibility is that LLMs get distracted by parts of input that are irrelevant for answering, which is not unknown in the recent LLM literature [58, 59]. For example, Shi et al. [59] show that LLMs are sensitive to irrelevant information in arithmetic reasoning problems. Although a simple instruction to ignore irrelevant context worked well in their case, we find no improvements from augmenting the prompts in this way.

External sources. While we have explored the augmentation of input text from one section to the entire article, it would be possible to include sources external to the main study publication. One example is trial protocols, which may offer additional information or serve as a contrasting source (against the study text), cf. Pitre et al. [21]. These may be beneficial for determining the risk of various biases, including the biases due to the randomisation process, due to deviations from the intended interventions, and due to selection of the reported result. Similarly, trial registry entries may contain relevant information to clarify the study intentions, which could help determine the bias in selection of the reported result. Taking advantage of free text in trial registries has been shown to work well on other predictive tasks [60]. We acknowledge that such external sources could be potentially beneficial to LLM performance, but would require addressing the questions of accommodating even longer input sequences than those considered in our work. We therefore leave the implementation of additional input augmentation to future work.

Ground-truth for signalling questions. A crucial next step in the study of LLM-based RoB2 assessment would be acquiring answers to signalling questions from review authors. These would allow more precise evaluation of model decisions and a more targeted prompt development, focusing on the questions that are higher upstream in a domain decision rule and at the same time suffer from low accuracy. Such questions or prompts could be further decomposed, augmented with additional context or used in a chain-of-thought manner, not unlike Reppert et al. [61] who study a similar kind of task decompositions for extracting placebo information from randomised controlled trials (RCTs), analysing participant flow in RCTs, and question answering from NLP papers. Such ground-truth information would also enable a study of mixed-regime prediction where some questions are answered by humans and some by an LLM; or a study of how human reviewers can supervise the process of deriving domain-level bias decisions.

Extended few-shot prompting. Although justification-based exemplars did not lead to improvements in our case, adding hand-picked demonstrations for each signalling question separately may represent a way forward, much like few-shot demonstrations for individual reasoning steps [29]. Our way of addressing the RoB2 task using natural (i.e., as

[8] As a possible example, consider the instruction "*Consider available information on the proportion of study participants who continued with their assigned intervention throughout follow up, and answer 'Yes' or 'Probably yes' if the proportion who did not adhere is high enough to raise concerns.*" When exactly is the non-adherence proportion high enough to warrant a raised concern?

defined in the guidelines) decomposition into smaller answerable units should allow for a smooth integration of such an approach. Another possibility would be to present exemplars with full text, not only the annotated justifications. This, however, may require special long-input LLMs or the use of dedicated LLMs plugins to consider additional context.

LLMs as predictors. A body of literature maintains that—despite LLMs excelling at generation tasks—they can be less successful at predictive tasks [62, 63]. A concrete example is the work on named entity recognition, where very low results (F1<0.5) have been reported, yet the state-of-the-art supervised approaches reach F1 of 0.9 or more depending on the benchmark [64, 65]. Our empirical results can be seen as broadly reinforcing these findings. Nevertheless, some prediction-oriented tasks, e.g., extraction of structured medical evidence and general-domain relation extraction [66, 67] have seen much more compelling results when LLMs undergo task-specific fine-tuning.

Future work. We expect that LLM fine-tuning [68, 69] specifically on RoB assessment datasets would lead to better performance. While this requires annotated data, this is arguably much less than needed to train a traditional supervised predictor. For RoB2, this option will become increasingly more feasible as additional annotated data accumulates. As fine-tuning opens up several additional considerations, it is beyond the scope of the present work.

Finally, while both versions of RoB instruments are usually presented as distinct, they share a considerable level of similarity. It would therefore make sense to leverage the more plentiful RoB1 data to help improve the performance on RoB2 in a transfer-learning setup.

ACKNOWLEDGMENTS

The authors would like to thank the Cochrane Collaboration and John Wiley & Sons Limited for providing the source data of the Cochrane Database of Systematic Reviews used in our study. This research was funded by the Australian Research Council through an Industrial Transformation Training Center Grant (grant IC170100030). We also acknowledge the support of The University of Melbourne's Research Computing Services and the Petascale Campus Initiative.

REFERENCES

[1] Julian P T Higgins, Douglas G Altman, Peter C Gøtzsche, Peter Jüni, David Moher, Andrew D Oxman, Jelena Savović, Kenneth F Schulz, Laura Weeks, and Jonathan A C Sterne. The Cochrane Collaboration's tool for assessing risk of bias in randomised trials. *BMJ*, 343, 2011. ISSN 0959-8138.

[2] Cochrane. Cochrane database of systematic reviews. https://www.cochranelibrary.com/, 2023. Accessed: 6 December 2023.

[3] Iain J Marshall, Joël Kuiper, and Byron C Wallace. Automating risk of bias assessment for clinical trials. *IEEE Journal of Biomedical and Health Informatics*, 19(4):1406–1412, 2015.

[4] Iain J Marshall, Benjamin Nye, Joël Kuiper, Anna Noel-Storr, Rachel Marshall, Rory Maclean, Frank Soboczenski, Ani Nenkova, James Thomas, and Byron C Wallace. Trialstreamer: A living, automatically updated database of clinical trial reports. *Journal of the American Medical Informatics Association*, 27(12):1903–1912, 09 2020. ISSN 1527-974X. doi: 10.1093/jamia/ocaa163.

[5] Allison Gates, Ben Vandermeer, and Lisa Hartling. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the robotreviewer machine learning tool. *Journal of Clinical Epidemiology*, 96:54–62, 2018.

[6] Frank Soboczenski, Thomas A Trikalinos, Joël Kuiper, Randolph G Bias, Byron C Wallace, and Iain J Marshall. Machine learning to help researchers evaluate biases in clinical trials: a prospective, randomized user study. *BMC Medical Informatics and Decision Making*, 19(1):96, 2019.

[7] Susan Armijo-Olivo, Rodger Craig, and Sandy Campbell. Comparing machine and human reviewers to evaluate the risk of bias in randomized controlled trials. *Research Synthesis Methods*, 11(3):484–493, 2020.

[8] Christiaan H Vinkers, Herm J Lamberink, Joeri K Tijdink, Pauline Heus, Lex Bouter, Paul Glasziou, David Moher, Johanna A Damen, Lotty Hooft, and Willem M Otte. The methodological quality of 176,620 randomized controlled trials published between 1966 and 2018 reveals a positive trend but also an urgent need for improvement. *PLoS biology*, 19(4):e3001162, 2021.

[9] Anneliese Arno, James Thomas, Byron Wallace, Iain J. Marshall, Joanne E. McKenzie, and Julian H. Elliott. Accuracy and efficiency of machine learning–assisted risk-of-bias assessments in 'real-world' systematic reviews. *Annals of Internal Medicine*, 2022.

[10] Patricia Sofia Jacobsen Jardim, Christopher James Rose, Heather Melanie Ames, Jose Francisco Meneses Echavez, Stijn Van de Velde, and Ashley Elizabeth Muller. Automating risk of bias assessment in systematic reviews: a real-time mixed methods comparison of human researchers to a machine learning system. *BMC Medical Research Methodology*, 22(1):1–12, 2022.

[11] Simon Šuster, Timothy Baldwin, and Karin Verspoor. Analysis of predictive performance and reliability of classifiers for quality assessment of medical evidence revealed important variation by medical area. *Journal of Clinical Epidemiology*, 159:58–69, 2023.

[12] Jonathan A C Sterne, Jelena Savović, Matthew J Page, Roy G Elbers, Natalie S Blencowe, Isabelle Boutron, Christopher J Cates, Hung-Yuan Cheng, Mark S Corbett, Sandra M Eldridge, Jonathan R Emberson, Miguel A Hernán, Sally Hopewell, Asbjørn Hróbjartsson, Daniela R Junqueira, Peter Jüni, Jamie J Kirkham, Toby Lasserson, Tianjing Li, Alexandra McAleenan, Barnaby C Reeves, Sasha Shepperd, Ian Shrier, Lesley A Stewart, Kate Tilling, Ian R White, Penny F Whiting, and Julian P T Higgins. Rob 2: a revised tool for assessing risk of bias in randomised trials. *BMJ*, 366, 2019. ISSN 0959-8138. doi: 10.1136/bmj.l4898. URL https://www.bmj.com/content/366/bmj.l4898.

[13] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9), jan 2023. ISSN 0360-0300. doi: 10.1145/3560815. URL https://doi.org/10.1145/3560815.

[14] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901, 2020.

[15] Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, 2022.

[16] Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, et al. Evaluating large language models on medical evidence summarization. *NPJ Digital Medicine*, 6(1):158, 2023.

[17] Byron C Wallace, Sayantan Saha, Frank Soboczenski, and Iain J Marshall. Generating (factual?) narrative summaries of RCTs: Experiments with neural multi-document summarization. *AMIA Summits on Translational Science Proceedings*, 2021:605, 2021.

[18] Sanjana Ramprasad, Iain J Marshall, Denis Jered McInerney, and Byron C Wallace. Automatically summarizing evidence from clinical trials: A prototype highlighting current challenges. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2023, page 236. NIH Public Access, 2023.

[19] Ye Zhang, Iain Marshall, and Byron C. Wallace. Rationale-augmented convolutional neural networks for text classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 795–804, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1076. URL https://aclanthology.org/D16-1076.

[20] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1423. URL https://www.aclweb.org/anthology/N19-1423.

[21] Tyler Pitre, Tanvir Jassal, Jhalok Ronjan Talukdar, Mahnoor Shahab, Michael Ling, and Dena Zeraatkar. ChatGPT for assessing risk of bias of randomized trials using the RoB 2.0 tool: A methods study. *medRxiv*, 2023. doi: 10.1101/2023.11.19.23298727. URL https://www.medrxiv.org/content/early/2023/11/22/2023.11.19.23298727.

[22] Iain J Marshall, Joël Kuiper, and Byron C Wallace. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. *Journal of the American Medical Informatics Association*, 23(1):193–201, 06 2015.

[23] Thomas Frederick Crocker, Natalie Lam, Magda Jordão, Caroline Brundle, Matthew Prescott, Anne Forster, Joie Ensor, John Gladman, and Andrew Clegg. Risk-of-bias assessment using Cochrane's revised tool for randomized trials (RoB 2) was useful but challenging and resource-intensive: observations from a systematic review. *Journal of Clinical Epidemiology*, 161:39–45, 2023.

[24] Hye Sun Yun, Iain J Marshall, Thomas Trikalinos, and Byron C Wallace. Appraising the potential uses and harms of LLMs for medical systematic reviews. *arXiv preprint arXiv:2305.11828*, 2023.

[25] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[26] Abdulqadir J Nashwan and Jaber H Jaradat. Streamlining systematic reviews: Harnessing large language models for quality assessment and risk-of-bias evaluation. *Cureus*, 15(8), 2023.

[27] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[28] Ella Flemyng, Theresa Helen Moore, Isabelle Boutron, Julian PT Higgins, Asbjørn Hróbjartsson, Camilla Hansen Nejstgaard, and Kerry Dwan. Using risk of bias 2 to assess results from randomised controlled trials: guidance from cochrane. *BMJ Evidence-Based Medicine*, 2023.

[29] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022.

[30] Susan Armijo-Olivo, Maria Ospina, Bruno R da Costa, Matthias Egger, Humam Saltaji, Jorge Fuentes, Christine Ha, and Greta G Cummings. Poor reliability between cochrane reviewers and blinded external reviewers when applying the cochrane risk of bias tool in physical therapy trials. *PloS one*, 9(5):e96920, 2014.

[31] Lisa Hartling, Michele P Hamm, Andrea Milne, Ben Vandermeer, P Lina Santaguida, Mohammed Ansari, Alexander Tsertsvadze, Susanne Hempel, Paul Shekelle, and Donna M Dryden. Testing the risk of bias tool showed low reliability between individual reviewers and across consensus

assessments of reviewer pairs. *Journal of Clinical Epidemiology*, 66(9):973–981, 2013.

[32] Huseyin Naci, Courtney Davis, Jelena Savović, Julian PT Higgins, Jonathan AC Sterne, Bishal Gyawali, Xochitl Romo-Sandoval, Nicola Handley, and Christopher M Booth. Design characteristics, risk of bias, and reporting of randomised controlled trials supporting approvals of cancer drugs by European Medicines Agency, 2014-16: cross sectional analysis. *BMJ*, 366, 2019.

[33] Silvia Minozzi, Michela Cinquini, Silvia Gianola, Marien Gonzalez-Lorenzo, and Rita Banzi. The revised Cochrane risk of bias tool for randomized trials (RoB 2) showed low interrater reliability and challenges in its application. *Journal of Clinical Epidemiology*, 126:37–44, 2020.

[34] Silvia Minozzi, Marien Gonzalez-Lorenzo, Michela Cinquini, Daniela Berardinelli, Celeste Cagnazzo, Stefano Ciardullo, Paola De Nardi, Mariarosaria Gammone, Paolo Iovino, Alex Lando, Marco Rissone, Giovanni Simeone, Marta Stracuzzi, Giovanna Venezia, Lorenzo Moja, and Giorgio Costantino. Adherence of systematic reviews to Cochrane RoB2 guidance was frequently poor: a meta epidemiological study. *Journal of Clinical Epidemiology*, 152:47–55, 2022.

[35] B Richter and B Hemmingsen. Comparison of the Cochrane risk of bias tool 1 (RoB 1) with the updated cochrane risk of bias tool 2 (RoB 2). https://community.cochrane.org/sites/default/files/uploads/inline-files/RoB1_2_project_220529_BR%20KK%20formatted.pdf, 2023.

[36] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.

[37] Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V Le, Ed H Chi, Denny Zhou, et al. Challenging big-bench tasks and whether chain-of-thought can solve them. *arXiv preprint arXiv:2210.09261*, 2022.

[38] Dheeru Dua, Shivanshu Gupta, Sameer Singh, and Matt Gardner. Successive prompting for decomposing complex questions. *arXiv preprint arXiv:2212.04092*, 2022.

[39] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[40] Paperpile. https://paperpile.com, 2024. Accessed: 14 April 2023.

[41] Grobid. https://github.com/kermitt2/grobid, 2008–2023.

[42] Simon Suster. Risk-of-bias v.2 assessment with large language models [data set]. https://doi.org/10.5281/zenodo.11243025, May 2024.

[43] Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

[44] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*, pages 12697–12706. PMLR, 2021.

[45] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[46] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[47] Yizhong Wang, Hamish Ivison, Pradeep Dasigi, Jack Hessel, Tushar Khot, Khyathi Raghavi Chandu, David Wadden, Kelsey MacMillan, Noah A Smith, Iz Beltagy, et al. How far can camels go? exploring the state of instruction tuning on open resources. *arXiv preprint arXiv:2306.04751*, 2023.

[48] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120. URL https://www.pnas.org/doi/abs/10.1073/pnas.2305016120.

[49] Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. Can ChatGPT reproduce human-generated labels? A study of social computing tasks. *arXiv preprint arXiv:2304.10145*, 2023.

[50] Zeming Chen, Alejandro Hernández Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, et al. Meditron-70b: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*, 2023.

[51] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLaMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[52] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with paged attention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626, 2023.

[53] Clément Christophe, Avani Gupta, Nasir Hayat, Praveen Kanithi, Ahmed Al-Mahrooqi, Prateek Munjal, Marco Pimentel, Tathagata Raha, Ronnie Rajan, and Shadab Khan. Med42 — a clinical large language model. https://huggingface.co/m42-health/med42-70b, 2023.

[54] Simon Suster, Timothy Baldwin, and Karin Verspoor. Promoting fairness in classification of quality of medical evidence. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 413–426, 2023.

[55] Cochrane Infectious Diseases Group, Asger Sand Paludan-Müller, Kim Boesen, Irma Klerings, Karsten Juhl Jørgensen, and Klaus Munkholm. Hand cleaning with ash for reducing the spread of viral and bacterial infections: a rapid review. *Cochrane Database of Systematic Reviews*, 2020(7), 1996.

[56] Jialin Hong, Yuen Chak Tiu, Po Yat Bowie Leung, Man Fai Wong, Wing Yan Ng, Dawn Cheung, Hiu Yan Mok, Wai Yan Lam, Kwan Yu Li, and Carlos KH Wong. Interventions that improve adherence to antihypertensive medications in coronary heart disease patients: a systematic review. *Postgraduate Medical Journal*, 98(1157):219–227, 2022.

[57] Julian PT Higgins and Sally Green. *Cochrane Handbook for Systematic Reviews of Interventions, Version 5.1.0*. The Cochrane Collaboration, 2011.

[58] Lalchand Pandia and Allyson Ettinger. Sorting through the noise: Testing robustness of information processing in pre-trained language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1583–1596, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.119. URL https://aclanthology.org/2021.emnlp-main.119.

[59] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*, pages 31210–31227. PMLR, 2023.

[60] Siyang Wang, Simon Šuster, Timothy Baldwin, and Karin Verspoor. Predicting publication of clinical trials using structured and unstructured data: Model development and validation study. *Journal of Medical Internet Research*, 24(12):e38859, 2022.

[61] Justin Reppert, Ben Rachbach, Charlie George, Luke Stebbing Jungwon Byun, Maggie Appleton, and Andreas Stuhlmüller. Iterated decomposition: Improving science Q&A by supervising reasoning processes. *arXiv preprint arXiv:2301.01751*, 2023.

[62] Dhananjay Ashok and Zachary C Lipton. PromptNER: Prompting for named entity recognition. *arXiv preprint arXiv:2305.15444*, 2023.

[63] Junjie Ye, Xuanting Chen, Nuo Xu, Can Zu, Zekai Shao, Shichun Liu, Yuhan Cui, Zeyang Zhou, Chao Gong, Yang Shen, et al. A comprehensive capability analysis of GPT-3 and GPT-3.5 series models. *arXiv preprint arXiv:2303.10420*, 2023.

[64] Matthew Honnibal. How Many Labelled Examples Do You Need for a BERT-sized Model to Beat GPT4 on Predictive Tasks?, 2023. URL https://youtu.be/3iaxLTKJROc?list=LL&t=1275.

[65] Shuhe Wang, Xiaofei Sun, Xiaoya Li, Rongbin Ouyang, Fei Wu, Tianwei Zhang, Jiwei Li, and Guoyin Wang. GPT-NER: Named entity recognition via large language models. *arXiv preprint arXiv:2304.10428*, 2023.

[66] Somin Wadhwa, Jay DeYoung, Benjamin Nye, Silvio Amir, and Byron C Wallace. Jointly extracting interventions, outcomes, and findings from rct reports with llms. In *Machine Learning for Healthcare Conference*, pages 754–771. PMLR, 2023.

[67] Somin Wadhwa, Silvio Amir, and Byron Wallace. Revisiting relation extraction in the era of large language models. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.868. URL https://aclanthology.org/2023.acl-long.868.

[68] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[69] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022.