Can LLMs assess risk of bias in medical research?

Carl-Etienne Juneau, PhD[1] orcid.org/0000-0003-1394-7888

[1]*Dr. Muscle AI, Sheridan, Wyoming, United States*

Correspondence: carl-etienne.juneau@umontreal.ca

May 7, 2026

Keywords: *risk of bias, large language models, systematic review, evidence synthesis, prompt engineering*

Abstract

Background: Risk-of-bias assessment is central to reviews of medical research, but time-intensive. Large language models (LLMs) could help reduce this burden, but optimal prompting strategies remain unclear.

Methods: We tested a weaker and a stronger model on 14 observational COVID-19 studies under four prompts, each adding one layer of guidance: baseline, criteria definitions, training material, and a worked example. We hypothesized that model performance on risk-of-bias criterion-level agreement would improve with increasing guidance.

Results: Overall risk-of-bias agreement improved with added guidance for both models (weaker: 28.6% to 78.6%; stronger: 57.1% to 92.9%). At the primary criterion-level endpoint, however, the models showed divergent patterns. The weaker model performed best with criteria definitions only (78.6%, 95% CI 71.4 to 85.7), whereas the stronger model performed best with full guidance (69.6%, 59.2 to 80.1). For the weaker model, agreement declined with added guidance (full decline 15.2 percentage points, 95% CI 5.3 to 25.1). Neither model exceeded the majority-class baseline (91.1%, 85.8 to 96.3).

Conclusion: In this sample of 14 studies, performance was modest and protocol-dependent. The apparent ranking of models changed with prompt guidance, and overall agreement alone overstated performance relative to criterion-level evaluation. Taken together, these findings do not support replacing human reviewers with LLMs for risk-of-bias assessment.

Registration: Open Science Framework, https://osf.io/8grbe (registered April 7, 2026).

INTRODUCTION

In medical research, we assess risk of bias (RoB) to judge how much confidence to place in a study's findings. We do this because flaws in study design, conduct, analysis, and reporting can systematically overestimate or underestimate effects (Higgins et al., 2024). To make these judgments, reviewers use structured tools that cover domains such as participant selection, outcome measurement, missing data, and confounding.

This work is important, but it is also time-intensive. Large language models (LLMs) could help reduce this burden. Recent reviews suggest that LLMs are being tested across many parts of evidence synthesis, especially search, screening, and data extraction, but validated

applications remain limited and fully autonomous use is not yet supported (Lieberum et al., 2025). Risk-of-bias assessment is a particularly open question. Early studies have reported mixed results, with performance varying by tool, domain, and study type. Related appraisal work suggests that more explicit instructions may improve agreement (Hasan et al., 2024).

In this study, we ask how well two LLMs (a weaker model and a stronger model) reproduce expert risk-of-bias labels from our prior systematic review (Juneau et al., 2023), and whether agreement improves as prompts cumulatively add criteria definitions, training material, and a worked example. We also examine how much performance depends on the evaluation protocol itself. We hypothesized that the strong model would outperform the weak model, and that both models would show improved agreement with increasing prompting guidance.

METHODS

This study uses expert risk-of-bias assessments from our systematic review of COVID-19 contact tracing as a benchmark (Juneau et al., 2023). Two LLMs of different capability levels assessed each of the 14 observational studies under four cumulative prompt conditions, using the 8-criterion risk-of-bias rubric reported in that review. Agreement with expert labels was compared across conditions and models. Reporting follows TRIPOD-LLM guidelines (Collins et al., 2025). The study was pre-registered on the Open Science Framework prior to data collection (OSF registration: https://osf.io/8grbe; registered April 7, 2026).

Sample

The test set comprised 14 observational studies (single-arm intervention design) assessed in Juneau et al. (2023). In that review, two independent investigators assessed each study, with discrepancies resolved by consensus. Both investigators held PhDs, one in public health and the other in economics. These consensus labels served as the reference standard. The studies and their labels are shown in Table 1.

Table 1. Risk of bias assessment of observational studies (n=14).

First author	Year	Selection bias (representative group)	Attrition bias (complete follow-up)	Detection bias (blinded assessor)	Confounding (adjustment)	Reporting bias (defined group)	Reporting bias (defined follow-up)	Reporting bias (defined outcome)	Analyses (well defined)	Overall RoB
Bernard Stoecklin	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious
Bi	2020	Yes	No	No	Yes	Yes	Yes	Yes	Yes	Serious
Burke	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious

First author	Year	Selection bias (representative group)	Attrition bias (complete follow-up)	Detection bias (blinded assessor)	Confounding (adjustment)	Reporting bias (defined group)	Reporting bias (defined follow-up)	Reporting bias (defined outcome)	Analyses (well defined)	Overall RoB
Chen	2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Serious
Choi H	2020	Yes	Yes	No	No	Yes	Unclear	Yes	Unclear	Serious
Choi JY	2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Serious
Cowling	2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Serious
Davalgi	2020	Yes	Yes	No	No	Yes	Unclear	Yes	Yes	Serious
Dinh	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious
Lam	2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Serious
Nachega	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious
Ng	2020	Yes	Yes	No	Yes	Yes	Yes	Yes	Yes	Serious
Wilasang	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious
Wong	2020	Yes	Yes	No	No	Yes	Yes	Yes	Yes	Serious

Rubric

The 8-criterion rubric was adapted from Mulder et al. (2019), a Cochrane review in which the criteria were originally developed for childhood cancer. In Juneau et al. (2023), some criteria were generalized to apply to observational studies more broadly. The criteria are listed in Table 2. A full mapping of each criterion to its Mulder antecedent is provided in Appendix A.

Table 2. Risk of bias criteria for single-arm observational studies of interventions.

Domain	Validity	Bias type	Criterion (yes condition)
Study group	Internal	Selection bias (representative)	Yes if the described study group consisted of >90% of eligible individuals
Study group	External	Reporting bias (well defined)	Yes if the intervention and number of participants was defined
Follow-up	Internal	Attrition bias (adequate)	Yes if the outcome was assessed for at least 60% of the study group of interest
Follow-up	External	Reporting bias (well defined)	Yes if the length of follow-up was mentioned
Outcome	Internal	Detection bias (blind)	Yes if the outcome assessors were blinded to the investigated determinant
Outcome	External	Reporting bias (well defined)	Yes if the outcome definition was objective and precise
Risk estimation	Internal	Confounding (adjustment for other factors)	Yes if important prognostic factors (i.e. age, gender) or follow-up were taken adequately into account

Domain	Validity	Bias type	Criterion (yes condition)
Risk estimation	External	Analyses (well defined)	Yes if the method of analysis was described and the effect of the intervention was quantified

Each criterion is scored yes, no, or unclear. The overall risk-of-bias label is derived from criterion-level outputs as follows: low when all "yes", moderate if any "unclear", serious if any "no". This rule was provided to the model in Conditions B, C, and D; Python also applied it independently to criterion-level outputs in those conditions to produce a derived overall label for comparison.

Prompt conditions

Common elements across conditions

All prompt conditions use the same full-text of the original study (PDF), the same model settings, and the same scoring procedure. All conditions ask the model to report an overall risk-of-bias label (low, moderate, or serious). Conditions B, C, and D additionally return per-criterion judgments with supporting quotes; in those conditions, Python independently derives an overall label from the criterion-level outputs for comparison with the model-reported overall.

Condition A: baseline agreement

The model receives the study text and a minimal task instruction: "Assess risk of bias in the following study as low, moderate, or serious. Base your judgment only on the provided study text and not on other studies, knowledge, or assumptions. Return only valid JSON." This tests whether the model can reproduce expert judgments natively, without additional guidance.

Condition B: criteria definitions

In addition to the instructions provided in Condition A, the model receives the 8 criterion definitions, the allowed outputs per criterion (yes, no, unclear), the overall RoB derivation rule, and the JSON output schema. It does not receive training material or worked examples.

Condition C: training material

In addition to the instructions provided in Condition B, the model receives as training material the full text of Mulder et al. (2019), where the RoB tool was initially developed.

Condition D: worked example

In addition to the instructions provided in Condition C, the model receives one external worked example showing how the rubric was applied and how the structured output should be produced. The worked example includes both the input study text and the

expected structured output. The example was drawn from Green et al. (2019), an open-access study assessed in Mulder et al. (2019).

Although criterion-level labels for Green et al. (2019) already appear in Mulder et al. (2019) (given in C), Condition D provides additional guidance by pairing the full study text with the expected JSON output. This gives the model an explicit worked example of how the rubric is operationalized.

Models

The weak model is Google Gemini 3 Flash (`gemini-3-flash-preview`). The strong model is Google Gemini 3.1 Pro (`gemini-3.1-pro-preview`, `thinking_level: high`). Both models have a training data cutoff of January 2025. All runs were conducted on April 7–8, 2026. Temperature was fixed at 0 and max tokens at 32768 (increased from 1024 in the pre-registered protocol, where that limit caused truncated JSON responses for conditions B, C, and D).

Run policy

One API call was made per study, model, and prompt condition. If the output was invalid JSON or missing any criterion field, the prompt was retried once identically. If the second call also failed, a parse failure was recorded; outputs were not imputed or manually corrected. Transient API errors (e.g. 429 rate-limit, 503 overload) were retried. All prompts were frozen before any real-study run.

Statistical analysis

The primary endpoint was mean per-study criterion-level agreement in Conditions B–D. For each study, we calculated the proportion of the 8 criterion judgments matching the expert label. We then averaged this proportion across the 14 studies. This treats the study, rather than the individual criterion judgment, as the unit of analysis.

To compare prompt conditions within each model, we computed the per-study difference in criterion-level agreement for each prespecified contrast (C vs B and D vs C) and reported the mean paired difference with a 95% confidence interval. Between-model comparisons within each condition were secondary and exploratory, using the same framework. Unweighted Cohen's kappa was reported as a secondary descriptive measure at the criterion level. Per-criterion confusion matrices and a majority-class baseline (which always predicts the most frequent expert label for each criterion across the 14 studies) were also reported.

Overall-label analyses were secondary and descriptive. For each condition, we reported percent agreement between the model-reported overall label and the expert label. For Conditions B–D, we also compared the Python-derived overall label with the expert label. Because all 14 expert overall labels were serious, weighted kappa was not informative and is not reported for overall labels. Condition A returns only an overall label and is treated as a descriptive baseline, not included in the primary criterion-level comparison.

As a post hoc sensitivity analysis, we applied Holm's step-down procedure to control the family-wise error rate across all nine contrasts reported in the study: the four prespecified within-model contrasts (C vs B and D vs C for each model), two within-model B vs D contrasts, and three between-model same-condition comparisons (Flash vs Pro at B, C, and D). For each contrast, the two-sided p-value was obtained from a one-sample t-test of the per-study paired differences (H_0: mean difference = 0). Holm-adjusted p-values were computed across the full family of nine; comparisons with adjusted $p < 0.05$ are considered to survive correction.

As a second post hoc sensitivity analysis, we reran Condition B after removing a missingness rule originally included in the protocol. Agreement statistics were recomputed and compared to assess the rule's influence on measured performance. Because agreement improved markedly for both models, we adopted the rule-removed version as the operative Condition B and built Conditions C and D on top of it. Results from both versions of Condition B are reported to allow comparison.

RESULTS

Total API cost was under $10. Outputs were usable in all 14 studies for both models. Model-reported and Python-derived overall risk of bias were identical, indicating that the models' direct overall judgments were internally consistent with their criterion-level outputs.

Table 3 reports overall risk of bias agreement across all conditions. Because all 14 gold overall labels are `serious`, weighted kappa is 0.000 for all cells and is not reported; overall percent agreement and counts are reported instead. These overall-label analyses are descriptive and should be interpreted cautiously.

Table 3. Overall risk of bias agreement by model and condition (model-reported versus gold; n=14 studies).

Condition	Flash	Pro
A — baseline (no guidance)	4/14 (28.6%)	8/14 (57.1%)
B — criteria definitions	8/14 (57.1%)	12/14 (85.7%)
C — + training material	9/14 (64.3%)	12/14 (85.7%)
D — + worked example	11/14 (78.6%)	13/14 (92.9%)

Both models improved overall risk of bias from A to D. Pro showed higher overall agreement at every condition.

Primary analysis: criterion-level agreement (conditions B–D)

As shown in Table 4, the two models displayed divergent patterns: Flash agreement declined from B to D, while Pro agreement increased.

Table 4. Mean per-study criterion-level agreement (%) by model and condition (n=14 studies).

	Flash	Pro
B — criteria definitions	78.6% (95% CI 71.4–85.7)	57.1% (95% CI 40.2–74.1)
C — + training material	70.5% (95% CI 61.3–79.8)	66.1% (95% CI 51.8–80.3)
D — + worked example	63.4% (95% CI 50.6–76.2)	69.6% (95% CI 59.2–80.1)

The best-performing condition for Flash was B (criteria definitions only). The best-performing condition for Pro was D (full guidance). Table 5 shows each model's best-performing condition versus the majority-class baseline. This baseline predicts labels without using the study text by always predicting the most frequent expert label for each criterion across the 14 studies. It achieved 91.1% agreement (95% CI 85.8 to 96.3).

Table 5. Best-performing condition versus majority-class baseline (n=14 studies).

	Flash	Pro	Majority-class baseline
Best-performing condition	78.6% (95% CI 71.4–85.7) (B)	69.6% (95% CI 59.2–80.1) (D)	91.1% (95% CI 85.8–96.3)

Both models fell well below the majority-class baseline.

Prespecified prompt contrasts

Table 6 reports the prespecified within-model condition contrasts (C vs B and D vs C) per the protocol.

Table 6. Prespecified paired differences in criterion-level agreement (percentage points; positive = first condition better; n=14 studies).

Contrast	Mean diff	95% CI
Flash: B vs C	+8.0	[+0.2, +15.8]
Flash: C vs D	+7.1	[−1.7, +16.0]
Pro: B vs C	−8.9	[−19.3, +1.4]
Pro: C vs D	−3.6	[−15.4, +8.2]

One prespecified contrast had a 95% CI excluding zero: Flash B vs C (+8.0pp, 95% CI [+0.2, +15.8]). Adding the Mulder et al. (2019) training material reduced Flash's criterion-level agreement, contrary to the hypothesis that more guidance would improve performance. The corresponding contrast for Pro was in the opposite direction (−8.9pp) but its CI included zero. No D vs C contrast had a CI excluding zero for either model.

Criterion-level kappa

Table 7. Criterion-level Cohen's kappa by model and condition (n=14 studies).

Condition	Flash	Pro
B	0.464	0.232

Condition	Flash	Pro
C	0.310	0.320
D	0.140	0.336

Flash kappa declined from B to D; Pro kappa increased. The pattern mirrors the criterion-level agreement results.

Exploratory between-model comparison

In exploratory between-model comparisons, Flash exceeded Pro at Condition B by 21.4 percentage points (95% CI [+6.1, +36.8]), whereas the between-model differences at Conditions C and D had confidence intervals including zero. However, this Condition B difference did not survive Holm adjustment (p = 0.08).

Criterion-specific error patterns

Table 8 shows the per-criterion error patterns at Condition B. `outcome_assessors_blinded` was the most systematically miscoded criterion: expert labels were `no` in all 14 studies, yet Flash predicted `unclear` in 12/14 at B; Pro predicted `no` in 10/14. `intervention_and_participants_defined` and `outcome_definition_objective_precise` were correctly coded by Flash in 14/14 studies at B, but frequently miscoded by Pro with gold=`yes`, model=`no` errors. Full 3×3 confusion matrices for all conditions are provided in the Supplement.

Table 8. Per-criterion error summary, Flash and Pro, Condition B (14 studies).

Criterion	Gold majority	Flash: correct in majority class	Flash: main error	Pro: correct in majority class	Pro: main error
`study_group_representative`	yes (14)	11/14	yes → no (2)	5/14	yes → no (7)
`intervention_and_participants_defined`	yes (14)	14/14	—	8/14	yes → no (6)
`outcome_assessed_for_60pct`	yes (13)	12/13	yes → unclear (1)	7/13	yes → no (6)
`follow_up_length_reported`	yes (12)	12/12	unclear → yes (2)	9/12	yes → no (3)
`outcome_assessors_blinded`	no (14)	2/14	no → unclear (12)	10/14	no → unclear (4)
`outcome_definition_objective_precise`	yes (14)	14/14	—	9/14	yes → no (5)
`important_prognostic_factors_accounted_for`	mixed	13/14	—	10/14	yes → no (3)
`analysis_described_and_effect_quantified`	yes (13)	10/13	yes → no (3)	6/13	yes → no (7)

Post hoc sensitivity analysis: removing the missingness rule

In the original protocol, Condition B required explicit text evidence for every judgment, including "no", and defaulted to "unclear" when evidence was absent. The actual RoB tool is less strict: a "no" does not require the paper to explicitly state an absence (e.g., "we did not adjust for confounders"). After inspecting the initial results, we conducted a post hoc sensitivity analysis removing this rule from Condition B. The results are shown in Table 9.

Table 9. Condition B results with and without the missingness rule.

Model	Criterion with rule	Criterion without rule	Change	Overall RoB with rule	Overall RoB without rule	Change
Flash	68.8%	78.6%	+9.8pp	3/14 (21.4%)	8/14 (57.1%)	+35.7pp
Pro	49.1%	57.1%	+8.0pp	10/14 (71.4%)	12/14 (85.7%)	+14.3pp

Removing the rule improved criterion-level agreement for both models by approximately 8–10 percentage points, and substantially improved overall RoB agreement, particularly for Flash. We retained results without the missingness rule and report them in this text.

Post hoc sensitivity analysis: Holm multiplicity correction

In a post hoc sensitivity analysis, we applied Holm's step-down procedure across all nine reported contrasts (Table 10). Under this omnibus family definition, only one contrast remained just below the conventional 0.05 threshold after adjustment: the within-model Flash B versus D comparison, for which agreement was 15.2 percentage points higher in B than D (Holm-adjusted $p \approx 0.050$). The prespecified Flash B versus C contrast (Holm-adjusted $p = 0.312$) and the exploratory Flash versus Pro contrast at Condition B (Holm-adjusted $p = 0.08$) did not survive adjustment. This analysis therefore weakens confidence in individual contrast-level claims and supports a cautious interpretation centered on the broader pattern of protocol-dependent performance rather than on any single model-comparison result.

Table 10. All contrasts with raw and Holm-adjusted p-values (k = 9).

Contrast	Mean diff (pp)	95% CI	p	Holm-adjusted p
Flash: B vs C	+8.0	[+0.2, +15.8]	0.045	0.312
Flash: C vs D	+7.1	[−1.7, +16.0]	0.104	0.429
Flash: B vs D	+15.2	[+5.3, +25.1]	0.006	0.050†
Pro: B vs C	−8.9	[−19.3, +1.4]	0.086	0.429
Pro: C vs D	−3.6	[−15.4, +8.2]	0.525	0.836
Pro: B vs D	−12.5	[−25.5, +0.5]	0.058	0.346
Flash vs Pro at B	+21.4	[+6.1, +36.8]	0.010	0.080
Flash vs Pro at C	+4.5	[−7.1, +16.0]	0.418	0.836
Flash vs Pro at D	−6.2	[−17.1, +4.6]	0.236	0.709

Positive values indicate the first-named condition or model is higher. †Holm-adjusted p-value based on unrounded value; displayed value is rounded to three decimal places.

DISCUSSION

In this study, overall risk-of-bias agreement improved with prompt guidance for both models, rising from 28.6% to 78.6% for Flash and from 57.1% to 92.9% for Pro (Table 3). These are encouraging headline numbers. However, because all 14 expert overall labels in this dataset were serious, overall agreement largely reflects whether the model assigned a serious label — a low bar given that label's prevalence. Overall-label agreement should therefore be interpreted cautiously and not treated as the primary measure of performance.

At the criterion level, which is the primary endpoint, the picture was more nuanced (Table 4). Flash performed best under Condition B (criteria definitions only), reaching 78.6% agreement (kappa 0.46, moderate) (Landis & Koch, 1977). Flash degraded with additional guidance. Pro showed the opposite pattern, improving from 57.1% under Condition B to 69.6% under full guidance (kappa 0.34, fair). These divergent trajectories suggest that the weaker model may have been more sensitive to prompt complexity, while the stronger model was better able to leverage additional context.

To put this in context, we compared each model against a majority-class baseline. It achieved 91.1% agreement (95% CI 85.8 to 96.3), indicating substantial class imbalance in the dataset and showing that exact agreement alone can overstate model performance. Indeed, both models fell well below the majority-class baseline at every condition. This is in line with the results of Šuster et al. (2024) who criticized LLMs for "seldom beating trivial baselines" in RoB assessment.

These results only partially supported our hypotheses. The stronger model did show higher overall agreement at baseline, but at the primary criterion-level endpoint the apparent ranking of models depended on prompt condition: Flash performed best under the leaner criteria-only prompt, whereas Pro performed best under full guidance. After a post hoc Holm adjustment across all nine comparisons, the only contrast remaining below the 0.05 threshold was the Flash B versus D summary comparison, supporting the conclusion that performance in this benchmark was strongly protocol-dependent.

How might we explain these findings? One possibility is that Flash, as a smaller and less capable model, is more compliant with explicit rubric definitions and less likely to draw on prior knowledge that conflicts with the expert labeling standard. Under Condition B, this compliance may have worked in its favor. With the addition of the Mulder et al. training material (Condition C) — which was developed for childhood cancer rather than COVID-19 contact tracing — Flash may have been pulled toward a different application domain, reducing alignment with the gold labels. Pro, by contrast, may be better at integrating and reconciling a longer, more complex prompt. This would explain why Pro improved progressively while Flash degraded.

The per-criterion error summary (Table 8) provides evidence that is consistent with these explanations. The most systematic single error in the table is on `outcome_assessors_blinded`: gold labels were `no` in all 14 studies, yet Flash predicted `unclear` in 12 of 14 cases, while Pro was correct in 10 of 14. Flash's near-total collapse on this criterion to `unclear` is a consistent pattern that points to a rubric-prompt mismatch. The blinding criterion is the only criterion in the rubric where the gold-majority label is `no`, and Flash appears to have treated absence of evidence of blinding as insufficient to assign `no`, defaulting instead to `unclear`. Pro's broader tendency at Condition B runs in the opposite direction: across several yes-majority criteria, including `study_group_representative`, `intervention_and_participants_defined`, `outcome_assessed_for_60pct`, and `analysis_described_and_effect_quantified`, Pro's main error was yes → no, suggesting it was systematically stricter or more pessimistic than the expert labeling standard. This directional pattern explains a seeming paradox: Pro assigned more `serious` overall labels correctly (Table 3) partly because it was more willing to assign `no` at the criterion level, and under the derivation rule any `no` yields `serious`. Flash, by contrast, was near-perfect on several yes-majority criteria at B — scoring 14/14 on both `intervention_and_participants_defined` and `outcome_definition_objective_precise`, but its systematic blinding error dragged down its criterion-level agreement. In short, the two models were not simply better or worse overall; they had diverging error profiles that are invisible in the headline agreement numbers. It would be interesting to examine whether a debate or reconciliation protocol — in which the two models are asked to review each other's judgments and resolve disagreements — could combine their complementary strengths and yield higher agreement than either model alone.

Our criterion-level kappa values (0.46 and 0.34) are broadly consistent with prior work. In a rapid review, Adam et al. (2024) found weighted kappa values ranging from 0.11 to 0.48 across multiple automation tools for risk-of-bias assessment. Pre-LLM systems such as RobotReviewer showed similar variation: Gates et al. (2018) found moderate reliability for some domains but only fair agreement for overall risk of bias, and Tian et al. (2024) confirmed substantial domain-to-domain variation with kappa values ranging from 0.25 to 0.59 across 1,955 trials. Our results fall within that range, though they were obtained with a general-purpose LLM rather than a purpose-built system.

Results in the LLM era have been similarly mixed. Šuster et al. (2024) found that four LLMs seldom beat trivial baselines without additional guidance, which matches our Condition A results. In contrast, Lai et al. (2024) reported mean correct assessment rates of 84.5% and 89.5% for two LLMs using a structured prompt and a three-expert criterion standard (higher than our best-condition results), though their criterion standard and study type differed. Hasan et al. (2024) found 61% raw agreement for overall risk of bias using GPT-4 with ROBINS-I, and Huang et al. (2025) reported 65–70% overall accuracy with structured prompting. More recently, Kuitunen et al. (2025) found overall kappa of 0.43 and concluded ChatGPT-4o was not ready for routine use; Rose et al. (2025) reported 50.7% human-ChatGPT overall agreement; and Taneri (2025) found weighted kappa of 0.51 but only 53% sensitivity for high-risk studies. Across these studies and our own, the

consistent pattern is that performance remains below what would be needed for autonomous use, and that it depends heavily on the assessment protocol.

Another interesting finding of this study is that a single instruction materially changed model behavior. In the original Condition B prompt, the missingness rule imposed a strict text-evidence standard for distinguishing `no` from `unclear`. Under that rule, the models tended to assign `unclear` when the study text did not explicitly state that a criterion was not met, reducing agreement with expert labels. In the rerun without the rule, models supplied more of their own judgment about `no` versus `unclear`, increasing agreement.

This suggests that measured agreement depends on the fit between the prompt's standard and the standard implicit in the RoB tool. In this study, prompt components were part of the intervention under evaluation, not neutral scaffolding. Researchers designing LLM-based assessment protocols should therefore report instructions explicitly and assess their compatibility with the reference standard.

These findings may also reflect ambiguity in the rubric itself. The tool we used specifies three possible outputs per criterion (yes, no, unclear), but some criteria are more explicit about when they are met than about how to distinguish "not met" from "insufficiently reported." Human reviewers can often bridge that gap through tacit methodological judgment. LLMs cannot do so reliably unless the prompt or tool makes that judgment rule explicit. Our results therefore suggest that under-specified appraisal tools may be a bottleneck not only for LLM-based assessment, but for reproducible assessment more generally. Methodologists developing future appraisal tools should consider defining all possible outputs with equal specificity.

Taken together, these findings suggest that more prompting may not be inherently better to assess RoB. Its effect depended on the model. In that sense, this study was less a simple comparison of two models than an illustration of protocol dependence in LLM-based risk-of-bias assessment. Small prompt decisions and under-specified rubric boundaries could materially change measured performance, including the apparent ranking of models.

Finally, it suggests researchers evaluating LLMs for this task should assess performance at the criterion level rather than relying on overall-label agreement alone, which can overestimate true performance.

Limitations

This study has several limitations. First, each study × model × prompt-condition combination was sampled only once, at temperature 0. This reflects a plausible real-world deployment setting where users seek stable, low-randomness outputs, but it does not estimate expected performance under repeated sampling. As Miller (2024) notes, a fuller evaluation would resample outputs or use token-probability-based scoring where feasible. Our confidence intervals therefore reflect variation across studies in this benchmark, but not residual sampling variability within the same prompt and model configuration.

Second, all 14 expert overall labels in this study are serious, which limits the interpretability of overall-label agreement measures and makes between-condition comparisons at the overall level uninformative. This is why we focused on criterion-level agreement as the primary endpoint. This limitation may be inherent to this specific tool or even RoB assessment itself, as Mulder et al. (2019) also found substantial class imbalances across criteria in their 33 included studies.

Third, the sample of 14 studies constrains the precision of the estimates and the generalizability of the findings to study designs other than single-arm observational studies. Confidence intervals are wide throughout, and several exploratory contrasts that could not be confirmed here may warrant investigation in larger benchmarks.

Fourth, we used a rubric originally developed for childhood cancer (Mulder et al., 2019) and adapted for COVID-19 contact tracing (Juneau et al., 2023). The Condition C training material (Mulder et al., 2019) comes from the childhood-cancer domain, which may have introduced topical contamination for Flash in particular. Moreover, the criterion-level label distributions in Mulder et al. (2019) differ substantially from those in our dataset. For example, blinding is low-risk in 29 of 33 studies in Mulder et al. (2019) but high-risk in 14 of 14 in our sample. This may have compounded the topical mismatch by exposing the model to conflicting priors. The degree to which these findings generalize to other rubrics or study types is unknown.

Finally, the two models were both from the same provider (Google Gemini) and represent a snapshot of model capabilities at the time of running (April 2026). Results may differ for models from other providers or for future model versions.

CONCLUSION

In this study, overall risk-of-bias agreement was higher than criterion-level agreement, which reflects more accurately whether the models reproduced the underlying criterion judgments. At the criterion level, the two models reached at best 78.6% (kappa 0.46) and 69.6% (kappa 0.34) agreement, while both remained below a majority-class baseline of 91.1% at every condition. Performance depended on prompting: under the leaner criteria-only prompt, Flash matched the gold standard more closely; under fuller guidance, Pro improved and the models converged. These findings suggest that prompt design is central to performance, and should be specified explicitly. Overall, they do not support replacing human reviewers with LLMs for risk-of-bias assessment. Future work should examine whether carefully specified prompts can support human reviewers in high-volume screening contexts, rather than replace them.

Acknowledgements

Thanks to Noah Y. Siegel (Google DeepMind) for helpful discussions and feedback on study design and interpretation.

Ethical approval

This study is a secondary analysis of published, publicly available data. No new data were collected from human participants. No patients or members of the public were involved in the design, conduct, or reporting of this study. No ethical approval was required.

Code availability

The prompting scripts and scoring pipeline are available at https://github.com/carljuneau/2026-llm-rob.

Conflict of interest

Carl-Etienne Juneau is Founder of Dr. Muscle AI, a company developing AI-enabled fitness software (https://dr-muscle.com). No company product was evaluated in this study. The author declares no other competing interests.

Funding

This study was supported by Dr. Muscle AI. The company had no role in the design, analysis, or interpretation of this study.

AI disclosure

Google Gemini, Claude Sonnet, and OpenAI ChatGPT were used to assist in drafting and editing this manuscript, and to help refine the research questions, prompt conditions, and statistical analysis. All final decisions were made by the author.

REFERENCES

Adam GP, Davies M, George J, et al. Machine Learning Tools To (Semi-)Automate Evidence Synthesis: A Rapid Review and Evidence Map [Internet]. Rockville, MD: Agency for Healthcare Research and Quality (US); 2024. PMID: 40424432.

Arno A, Thomas J, Wallace B, Marshall IJ, McKenzie JE, Elliott JH. Accuracy and Efficiency of Machine Learning-Assisted Risk-of-Bias Assessments in "Real-World" Systematic Reviews: A Noninferiority Randomized Controlled Trial. Ann Intern Med. 2022;175(7):1001–1009. doi:10.7326/M22-0092

Bernard Stoecklin S, Rolland P, Silue Y, et al. First cases of coronavirus disease 2019 (COVID-19) in France: surveillance, investigations and control measures, January 2020. Euro Surveill. 2020;25(6):2000094. doi:10.2807/1560-7917.ES.2020.25.6.2000094

Bi Q, Wu Y, Mei S, et al. Epidemiology and transmission of COVID-19 in 391 cases and 1286 of their close contacts in Shenzhen, China: a retrospective cohort study. Lancet Infect Dis. 2020;S1473-3099(20)30287-5. doi:10.1016/S1473-3099(20)30287-5

Burke RM, Midgley CM, Dratch A, et al. Active Monitoring of Persons Exposed to Patients with Confirmed COVID-19 — United States, January–February 2020. MMWR Morb Mortal Wkly Rep. 2020;69:245–246. doi:10.15585/mmwr.mm6909e1

Chen CM, Jyan HW, Chien SC, et al. Containing COVID-19 Among 627,386 Persons in Contact With the Diamond Princess Cruise Ship Passengers Who Disembarked in Taiwan. J Med

Internet Res. 2020;22(5):e19540. doi:10.2196/19540

Choi H, Cho W, Kim MH, Hur JY. Public Health Emergency and Crisis Management: Case Study of SARS-CoV-2 Outbreak. Int J Environ Res Public Health. 2020;17:3984. doi:10.3390/ijerph17113984

Choi JY. COVID-19 in South Korea. Postgrad Med J. 2020;96:399–402. doi:10.1136/postgradmedj-2020-137738

Collins GS, Dhiman P, Andaur Navarro CL, et al. The TRIPOD-LLM reporting guideline for studies using large language models. Nat Med. 2025;31(1):49–60. doi:10.1038/s41591-024-03425-5

Cowling BJ, Ali ST, Ng TWY, et al. Impact assessment of non-pharmaceutical interventions against coronavirus disease 2019 and influenza in Hong Kong: an observational study. Lancet Public Health. 2020;5(5):e279–e288. doi:10.1016/S2468-2667(20)30090-6

Davalgi S, Malatesh U, Rachana A, et al. Comparison of measures adopted to combat COVID-19 pandemic by different countries in WHO regions. Indian J Comm Health. 2020;32(2). doi:10.47203/IJCH.2020.v32i02SUPP.023

Dinh L, Dinh P, Nguyen PDM, Nguyen DHN, Hoang T. Vietnam's response to COVID-19: prompt and proactive actions. J Travel Med. 2020;27(3):taaa047. doi:10.1093/jtm/taaa047

Flemyng E, Noel-Storr A, Macura B, et al. Position Statement on Artificial Intelligence (AI) Use in Evidence Synthesis Across Cochrane, the Campbell Collaboration, JBI, and the Collaboration for Environmental Evidence 2025. Campbell Syst Rev. 2025;21(4):e70074. doi:10.1002/cl2.70074

Gates A, Vandermeer B, Hartling L. Technology-assisted risk of bias assessment in systematic reviews: a prospective cross-sectional evaluation of the RobotReviewer machine learning tool. J Clin Epidemiol. 2018;96:54–62. doi:10.1016/j.jclinepi.2017.12.015

Green DM, Bhatt NS, Bhakta N, et al. Serum Alanine Aminotransferase Elevations in Survivors of Childhood Cancer: A Report From the St. Jude Lifetime Cohort Study. Hepatology. 2019;69(1):94–106. doi:10.1002/hep.30176

Hasan B, Saadi S, Rajjoub NS, et al. Integrating large language models in systematic reviews: a framework and case study using ROBINS-I for risk of bias assessment. BMJ Evid Based Med. 2024;29(6):394–398. doi:10.1136/bmjebm-2023-112597

Higgins JPT, Thomas J, Chandler J, et al. (eds). Cochrane Handbook for Systematic Reviews of Interventions version 6.5. Cochrane, 2024. www.cochrane.org/authors/handbooks-and-manuals/handbook/current

Hirt J, Meichlinger J, Schumacher P, et al. Agreement in Risk of Bias Assessment Between RobotReviewer and Human Reviewers: An Evaluation Study on Randomised Controlled Trials in Nursing-Related Cochrane Reviews. J Nurs Scholarsh. 2021;53(2):246–254. doi:10.1111/jnu.12628

Huang J, Pan B, Liu J, et al. Large Language Model–Assisted Risk-of-Bias Assessment in Randomized Controlled Trials Using the Revised Risk-of-Bias Tool: Evaluation Study. J Med Internet Res. 2025;27:e70450. doi:10.2196/70450

Juneau CE, Briand AS, Collazzo P, Siebert U, Pueyo T. Effective contact tracing for COVID-19: A systematic review. Glob Epidemiol. 2023;5:100103. doi:10.1016/j.gloepi.2023.100103

Kuitunen I, Nyrhi L, De Luca D. ChatGPT-4o in Risk-of-Bias Assessments in Neonatology: A Validity Analysis. Neonatology. 2025;1–6. doi:10.1159/000544857

Lai H, Ge L, Sun M, et al. Assessing the Risk of Bias in Randomized Clinical Trials With Large Language Models. JAMA Netw Open. 2024;7(5):e2412687. doi:10.1001/jamanetworkopen.2024.12687

Lam HY, Lam TS, Wong CH, et al. The epidemiology of COVID-19 cases and the successful containment strategy in Hong Kong — January to May 2020. Int J Infect Dis. 2020;98:51–58. doi:10.1016/j.ijid.2020.06.057

Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics. 1977;33(1):159–174. doi:10.2307/2529310

Leucuța D-C, Urda-Cîmpean AE, Istrate D, Drugan T. Risk of Bias Assessment of Diagnostic Accuracy Studies Using QUADAS 2 by Large Language Models. Diagnostics. 2025;15(12):1451. doi:10.3390/diagnostics15121451

Lieberum J-L, Toews M, Metzendorf M-I, et al. Large language models for conducting systematic reviews: on the rise, but not yet ready for use: a scoping review. J Clin Epidemiol. 2025;181:111746. doi:10.1016/j.jclinepi.2025.111746

Marshall IJ, Kuiper J, Wallace BC. RobotReviewer: evaluation of a system for automatically assessing bias in clinical trials. J Am Med Inform Assoc. 2016;23(1):193–201. doi:10.1093/jamia/ocv044

Miller E. Adding Error Bars to Evals: A Statistical Approach to Language Model Evaluations. arXiv:2411.00640. 2024.

Mulder RL, Bresters D, Van den Hof M, et al. Hepatic late adverse effects after antineoplastic treatment for childhood cancer. Cochrane Database Syst Rev. 2019;4:CD008205. doi:10.1002/14651858.CD008205.pub3

Nachega JB, Grimwood A, Mahomed H, et al. From Easing Lockdowns to Scaling-Up Community-Based COVID-19 Screening, Testing, and Contact Tracing in Africa. Clin Infect Dis. 2020;ciaa695. doi:10.1093/cid/ciaa695

Ng Y, Li Z, Chua YX, et al. Evaluation of the Effectiveness of Surveillance and Containment Measures for the First 100 Patients with COVID-19 in Singapore. MMWR Morb Mortal Wkly Rep. 2020;69(11):307–311. doi:10.15585/mmwr.mm6911e1

Rose CJ, Bidonde J, Ringsten M, et al. Using a Large Language Model (ChatGPT-4o) to Assess the Risk of Bias in Randomized Controlled Trials of Medical Interventions: Interrater Agreement With Human Reviewers. Cochrane Evid Synth Methods. 2025;3(5):e70048. doi:10.1002/cesm.70048

Šuster S, Baldwin T, Verspoor K. Zero- and few-shot prompting of generative large language models provides weak assessment of risk of bias in clinical trials. Res Synth Methods. 2024;15(6):988–1000. doi:10.1002/jrsm.1749

Taneri PE, Engel C, Engel H, et al. Human Versus Artificial Intelligence: Comparing Cochrane Authors' and ChatGPT's Risk of Bias Assessments. Cochrane Evid Synth Methods. 2025;3(7):e70044. doi:10.1002/cesm.70044

Tian Y, Yang X, Doi SAR, et al. Towards the automatic risk of bias assessment on randomized controlled trials: A comparison of RobotReviewer and humans. Res Synth Methods. 2024;15(6):1111–1119. doi:10.1002/jrsm.1761

Wilasang C, Sararat C, Jitsuk NC, et al. Reduction in effective reproduction number of COVID-19 is higher in countries employing active case detection with prompt isolation. J Travel Med. 2020;taaa095. doi:10.1093/jtm/taaa095

Wong SYS, Kwok KO, Chan FKL. What can countries learn from Hong Kong's response to the COVID-19 pandemic? CMAJ. 2020;192(19):E511–E515. doi:10.1503/cmaj.200563

APPENDIX A — CRITERION MAPPING: JUNEAU ET AL. (2023) VERSUS MULDER ET AL. (2019)

Table A1. Criterion mapping: Juneau et al. (2023) versus Mulder et al. (2019).

	Domain	Criterion used in Juneau et al. (2023)	Earlier wording in Mulder et al. (2019)	Adaptation made in Juneau et al. (2023)	Code key
1	Study group	Yes if the described study group consisted of >90% of eligible individuals	Yes if the described study group consisted of >90% of the original cohort of childhood cancer survivors, or was a random sample with respect to cancer treatment	Generalised from a childhood-cancer cohort frame to a broader eligible-population frame; the explicit random-sample clause was not carried over	`study_group_representative`
2	Study group	Yes if the intervention and number of participants was defined	Yes if the type of chemotherapy and/or location of radiotherapy was mentioned and the number of participants with chronic viral hepatitis was mentioned	Generalised from cancer-treatment exposure and hepatitis-status reporting to a broader intervention-and-sample-description criterion	`intervention_and_participants_defined`
3	Follow-up	Yes if the outcome was assessed for at least 60% of the study group of interest	Yes if the outcome was assessed for >90% of the study group of interest (++) or for 60% to 90% of the study group of interest (+)	Collapsed Mulder's two favourable follow-up thresholds into one broader yes condition at ≥60%	`outcome_assessed_for_60pct`
4	Follow-up	Yes if the length of follow-up was mentioned	Yes if the length of follow-up was mentioned	No substantive wording change	`follow_up_length_reported`
5	Outcome	Yes if the outcome assessors were blinded to the investigated determinant	Yes if the outcome assessors were blinded to the investigated determinant; Mulder added that if outcomes were biochemical measurements produced by a machine, blinding of the investigator was not relevant	Kept the general blinding criterion but removed the liver-biochemistry-specific caveat from the criterion wording	`outcome_assessors_blinded`
6	Outcome	Yes if the outcome definition was objective and precise	Yes if the outcome definition was objective and precise, i.e., if the upper limits of normal for liver function tests were described in the definition of hepatic late adverse effects	Generalised from a liver-test-specific definition to a broader outcome-definition criterion	`outcome_definition_objective_precise`
7	Risk estimation	Yes if important prognostic factors (i.e. age, gender) or follow-up were taken adequately into account	Yes if important prognostic factors (i.e. age, gender, co-treatment) or follow-up were taken adequately into account	Retained the same structure but removed the cancer-specific example of co-treatment	`important_prognostic_factors_accounted_for`
8	Risk estimation	Yes if the method of analysis was described and the effect of the intervention was quantified	Yes if a relative risk, odds ratio, attributable risk, linear or logistic regression model, mean difference, or Chi² was calculated	Generalised from a list of named statistical measures to a broader requirement that the analysis be described and the effect quantified	`analysis_described_and_effect_quantified`

SUPPLEMENT

Per-criterion confusion matrices (gold rows × model columns; n=14 studies)

Rows = gold label; columns = model label. Conditions B, C, and D only (Condition A returns no criterion-level output). Each cell counts the number of studies with that gold–model label combination.

Flash (`gemini-3-flash-preview`)

Condition B — criteria definitions

Representative study group

Gold \ Model	yes	no	unclear
yes	11	2	1
no	0	0	0
unclear	0	0	0

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	14	0	0
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	12	0	1
no	1	0	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	12	0	0
no	0	0	0
unclear	2	0	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0
no	0	2	12
unclear	0	0	0

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	14	0	0
no	0	0	0
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	6	0	0
no	1	7	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	10	3	0
no	0	0	0
unclear	0	1	0

Condition C — + training material

Representative study group

Gold \ Model	yes	no	unclear
yes	9	4	1

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	13	1	0

Gold \ Model	yes	no	unclear
no	0	0	0
unclear	0	0	0

Gold \ Model	yes	no	unclear
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	12	1	0
no	1	0	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	11	1	0
no	0	0	0
unclear	2	0	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0
no	3	2	9
unclear	0	0	0

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	13	1	0
no	0	0	0
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	3	1	2
no	1	7	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	9	4	0
no	0	0	0
unclear	0	1	0

Condition D — + worked example

Representative study group

Gold \ Model	yes	no	unclear
yes	7	4	3
no	0	0	0
unclear	0	0	0

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	12	2	0
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	9	2	2
no	1	0	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	11	1	0
no	0	0	0
unclear	2	0	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0
no	7	1	6

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	13	1	0
no	0	0	0

Gold \ Model	yes	no	unclear
unclear	0	0	0

Gold \ Model	yes	no	unclear
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	3	3	0
no	2	6	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	9	4	0
no	0	0	0
unclear	1	0	0

Pro (`gemini-3.1-pro-preview`)

Condition B — criteria definitions

Representative study group

Gold \ Model	yes	no	unclear
yes	5	7	2
no	0	0	0
unclear	0	0	0

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	8	6	0
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	7	6	0
no	1	0	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	9	3	0
no	0	0	0
unclear	0	2	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0
no	0	10	4
unclear	0	0	0

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	9	5	0
no	0	0	0
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	3	3	0
no	1	7	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	6	7	0
no	0	0	0
unclear	0	1	0

Condition C — + training material

Representative study group

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	8	5	1
no	0	0	0
unclear	0	0	0

Gold \ Model	yes	no	unclear
yes	10	4	0
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	10	2	1
no	1	0	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	10	2	0
no	0	0	0
unclear	1	1	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0
no	0	6	8
unclear	0	0	0

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	12	2	0
no	0	0	0
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	4	2	0
no	0	8	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	6	7	0
no	0	0	0
unclear	0	1	0

Condition D — + worked example

Representative study group

Gold \ Model	yes	no	unclear
yes	8	4	2
no	0	0	0
unclear	0	0	0

Intervention and participants defined

Gold \ Model	yes	no	unclear
yes	13	1	0
no	0	0	0
unclear	0	0	0

Outcome assessed for ≥60%

Gold \ Model	yes	no	unclear
yes	11	2	0
no	0	1	0
unclear	0	0	0

Follow-up length reported

Gold \ Model	yes	no	unclear
yes	10	2	0
no	0	0	0
unclear	2	0	0

Outcome assessors blinded

Gold \ Model	yes	no	unclear
yes	0	0	0

Outcome definition objective and precise

Gold \ Model	yes	no	unclear
yes	13	1	0

Gold \ Model	yes	no	unclear
no	2	5	7
unclear	0	0	0

Gold \ Model	yes	no	unclear
no	0	0	0
unclear	0	0	0

Important prognostic factors accounted for

Gold \ Model	yes	no	unclear
yes	4	2	0
no	1	7	0
unclear	0	0	0

Analysis described and effect quantified

Gold \ Model	yes	no	unclear
yes	6	7	0
no	0	0	0
unclear	0	1	0