ACKNOWLEDGEMENTS

We thank Professor Robert Tibshirani for his invaluable insights and discussions.

APPENDIX CONTENTS

A Societal impact

The use of large language models to automate systematic reviews offers clear potential to accelerate evidence synthesis in medicine and policy. However, when these systems produce incorrect or misleading results, clinicians and policymakers may base decisions on flawed findings, leading to inappropriate treatments or misguided recommendations.

Our study underscores the urgent need for continued research and cautious deployment. LLM-based systematic review systems need further rigorous validation, transparent uncertainty quantification, and mechanisms to detect and mitigate biases and errors. Only through careful development and oversight can these technologies be harnessed to benefit society without exacerbating existing risks or creating new harms.

B Dataset collection details

Below, we provide additional in-depth details regarding stages in dataset curation process.

B.1 Systematic review selection

MedEvidence is originally derived from 6,709 Cochrane publications extracted via Entrez from PubMed. We first discarded any papers where first References subsection was not both entitled "Studies included in this review" and non-empty, as our initial extraction filter included Cochrane SR protocols and SRs finding no valid studies, which were not of interest. We filter for SRs where all included references have a retrievable abstract and limit to SRs with 12 or less references to reduce annotator burden and improve odds of finding SRs where questions can be validated. On average, the end-to-end creation of a single question requires approximately 20 minutes. Appendix Figure 8 presents a cohort diagram for the materialization of the dataset.

B.2 Conclusion to question conversion

Appendix Figure 9 provides a direct example of a SR abstract parsed for manual question creation. We highlight the explicit statements ('conclusions') asserting differences between a treatment and control on an outcome, and the presence of standardized, author-provided assessment of evidence certainty for these individual conclusions. SR abstracts were consistently written in this form, allowing annotators to consistently interpret the conclusion into a question. To define the correct answer to the generated question, annotators obeyed the following criteria:

- Outcomes, or pairs of treatments and controls, where the authors stated that no studies provided sufficient (or any) evidence to perform analysis were labeled as `insufficient data` questions.
- Conclusions in which the authors stated that there was "no difference" or "no significant difference" between treatments and controls were labeled as `no difference` questions.
- Conclusions where the authors stated a difference between outcomes either definitively or with qualification (e.g. 'X increases Y' or 'X may reduce Y') were given the appropriate `higher` or `lower` label.
- Conclusions where the authors expressed that uncertainty was too great to evaluate a treatment outcome effect were placed in the `uncertain effect` label class. Conclusions where authors assessed a difference, but then stated that they were very uncertain of their findings were deemed ambiguous and discarded.

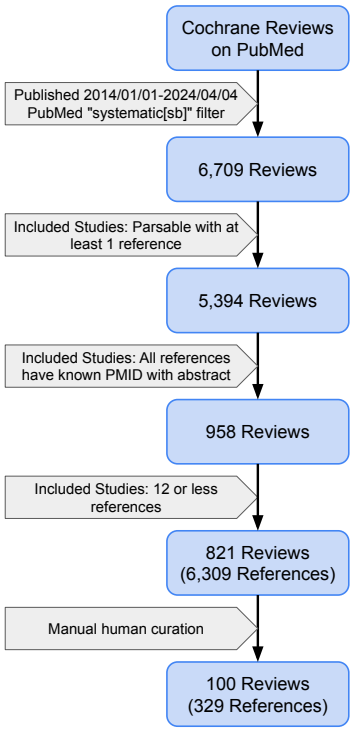

Figure 8: MedEvidence cohort diagram describing selection criteria for Cochrane SRs suitable for use in the MedEvidence dataset. Note that not all available papers in the second-to-last stage were manually reviewed for use in the final stage.

B.3 Relevant study selection and question validation

For author conclusions where more than one study was used, SRs provide meta-analyses over all relevant sources (an example meta-analysis is shown in Appendix Figure 10), allowing us to confirm whether the studies used in the original SR contain sufficient information to replicate the conclusions of human analysis.

Main results

Five RCTs, reported in 12 records, with 462,754 participants, met the inclusion criteria.

We identified trials on whole-cell plus recombinant vaccine (WC-rBS vaccine (Dukoral)) from Peru and trials on bivalent whole-cell vaccine (BivWC (Shanchol)) vaccine from India and Bangladesh. We did not identify any trials on other BivWC vaccines (Euvichol/ Euvichol-Plus), or Hillchol.

Two doses of Dukoral with or without a booster dose reduces cases of cholera at two-year follow-up in a general population of children and adults, and at five-month follow-up in an adult male population (overall VE 76%; RR 0.24, 95% confidence interval (CI) 0.08 to 0.65; 2 trials, 16,423 participants; high-certainty evidence).

Two doses of Shanchol reduces cases of cholera at one-year follow-up (overall VE 37%; RR 0.63, 95% CI 0.47 to 0.85; 2 trials, 241,631 participants; high-certainty evidence), at two-year follow-up (overall VE 64%; RR 0.36, 95% CI 0.16 to 0.81; 2 trials, 168,540 participants; moderate-certainty evidence), and at five-year follow-up (overall VE 80%; RR 0.20, 95% CI 0.15 to 0.26; 1 trial, 54,519 participants; high-certainty evidence).

A single dose of Shanchol reduces cases of cholera at six-month follow-up (overall VE 40%; RR 0.60, 95% CI 0.47 to 0.77; 1 trial, 204,700 participants; high-certainty evidence), and at two-year follow-up (overall VE 39%; RR 0.61, 95% CI 0.53 to 0.70; 1 trial, 204,700 participants; high-certainty evidence).

A single dose of Shanchol also reduces cases of severe dehydrating cholera at six-month follow-up (overall VE 63%; RR 0.37, 95% CI 0.28 to 0.50; 1 trial, 204,700 participants; high-certainty evidence), and at two-year follow-up (overall VE 50%; RR 0.50, 95% CI 0.42 to 0.60; 1 trial, 204,700 participants; high-certainty evidence).

We found no differences in the reporting of adverse events due to vaccination between the vaccine and control/placebo groups.

Figure 9: An example of a “Main Results" section from a Cochrane review used in MedEvidence (DOI: https://doi.org/10.1002/14651858.CD014573). Annotators were instructed to extract conclusions from this standardized sub-section of the SR abstract.

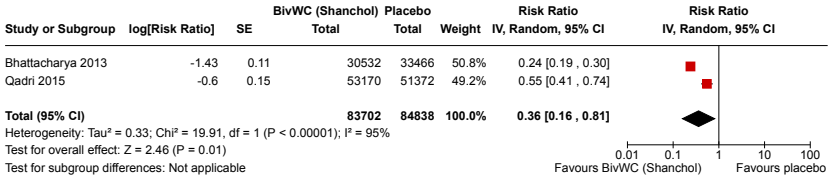

Figure 10: An example meta-analysis from a Cochrane review (figure from DOI: https://doi.org/10.1002/14651858.CD014573). Notably, the set of relevant studies and their individual weighted contributions to the overall result are available.

C ADDITIONAL DATASET DISTRIBUTIONS

We present additional statistical characteristics of the questions in our MedEvidence dataset in Appendix Figure 11. We highlight that the dataset is balanced with respect to evidence certainty levels, strengthening the reliability of our main observations on the relationship between evidence certainty and model performance. With regard to the joint distribution of correct treatment outcome effect and evidence certainty, we note that the highly concentrated distributions for the `insufficient data` and `uncertain effect` classes are inherent to the nature of SR. For example, in the case of the `insufficient data` class, authors cannot draw definitive conclusions from analyses they were unable to perform; thus, their findings are most uncertain when the quality of evidence is poor.

D EVALUATED MODELS AND PROMPTS

The full list of 25 models we evaluate on MedEvidence is provided in Appendix Table 3. The exact prompt used to elicit LLM responses for evaluation under the basic prompt regime is provided in Appendix Figure 12. Under the expert-guided prompt regime, models were first instructed to generate a formatted article summmary using the summarization step (using Appendix Figure 13a), then asked to provide answers based on the generated summaries for all relevant articles (via Appendix Figure 13b). In all cases, chunks of original article text or previously-generated summarization were provided with a header line containing the article's title, date of publication (if available), and PubMed ID, allowing the LLM to recognize and assign blocks of content to different sources and synthesize in-context.

Table 3: List of evaluated models with their model size and context length limit we set for our experiments. Precision is 16-bit floating point unless specified otherwise.

Model	Model Type	Parameter Sizes	Context Limit
DeepSeek R1 (DeepSeek-AI, 2025a)	Generalist Reasoning	671B	131K
DeepSeek V3 (DeepSeek-AI, 2025b)	Generalist Non-Reasoning	671B	131K
GPT-4.1 (OpenAI, 2024a)	Generalist Non-Reasoning	Unknown	1M
GPT-4.1 mini (OpenAI, 2024a)	Generalist Non-Reasoning	Unknown	131K
GPT-o1 (OpenAI, 2024b)	Generalist Non-Reasoning	Unknown	150K
GPT-oss-120B (OpenAI et al., 2025)	Generalist Reasoning	120B	128K
HuatuoGPT-o1 (Chen et al., 2024a)	Medical Reasoning	7B, 70B	32K, 16K
Llama 3.0 (AI@Meta, 2024)	Generalist Non-Reasoning	8B, 70B	8K
Llama 3.1 (AI@Meta, 2024)	Generalist Non-Reasoning	8B, 70B, 405B	131K
Llama 3.3 (AI@Meta, 2024)	Generalist Non-Reasoning	70B	131K
Llama 3.3 (R1-Distill) (DeepSeek-AI, 2025a)	Generalist Reasoning	70B	131K
Llama 4 Maverick (AI@Meta, 2025)	Generalist Non-Reasoning	400B (17B active)	500K
Llama 4 Scout (AI@Meta, 2025)	Generalist Non-Reasoning	109B (17B active)	1M
OpenBioLLM (Ankit Pal, 2024)	Medical Non-Reasoning	8B, 70B	8K
OpenThinker2 (Team, 2025a)	Generalist Reasoning	32B	131K
Qwen2.5 (Team, 2025b)	Generalist Non-Reasoning	7B, 32B, 72B	32K
Qwen3 (Team, 2025c)	Generalist Reasoning (hybrid)	235B (22B active, 8-bit)	32 K
QwQ (Team, 2025d)	Generalist Reasoning	32B	131K

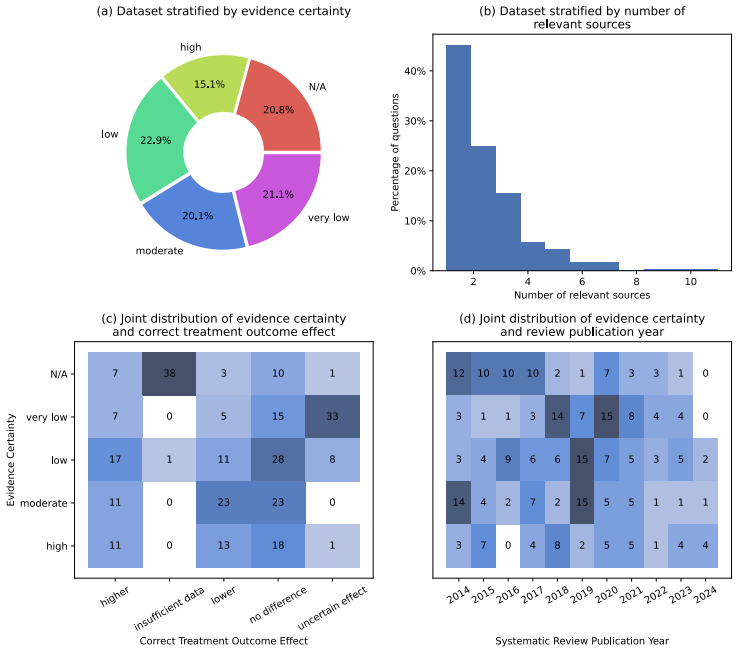

Figure 11: Additional statistical characteristics of MedEvidence. (a) shows the dataset distribution stratified by evidence certainty. (b) stratifies the questions by number of relevant sources. (c) is a joint distribution of evidence certainty and correct answer label. (d) shows the distribution of evidence certainties by systematic review publication year.

```
Given the ARTICLE SUMMARIES. Provide a concise and precise answer to the
    provided QUESTION.

After you think, return your answer with the following format:
- **Rationale**: Your rationale
- **Full Answer**: A precise answer, citing each fact with the Article
    ID in brackets (e.g. [2]).
- **Answer**: A final classification exactly matching one of the
    following options: Higher, Lower, No Difference, Insufficient Data,
    Uncertain Effect

Think step by step.
**QUESTION**: {question}
**ARTICLE SUMMARIES**: {context}
```

Figure 12: Prompt used to generate LLM responses to questions under the basic prompt setup.

```
You are the author of a Cochrane Collaboration systematic review, leveraging
    statistical analysis and assessing risks of bias in order to rigorously assess
    the effectiveness of medical interventions. As part of your review process,
    perform the following task:
As a subject expert, (1) summarize the evidence provided by a given ARTICLE as it
    pertains to a given QUESTION and (2) provide a possible answer.

Otherwise, if the provided article contains relevant information, you must return a
     list including the following items:

- **Study Design**: Type of study, level of evidence, and grade of recommendation
    according to the levels of evidence REC TABLE (provided Below).
- **Study Population**: Study size and patient population.
- **Summary**: A concise but comprehensive summary based on the previously
    specified information, with a focus on the main findings.
- **Possible Answer**: A concise feasible answer given the evidence.

**REC TABLE **: Levels of Evidence (from strongest [1a] to lowest [5]).

| Grade of Recommendation | Level of Evidence | Type of Study |
|-------------|-------------|-----------------------------------------|
| A | 1a | Systematic review and meta-analysis of (homogeneous) randomized
    controlled trials |
| A | 1b | Individual randomized controlled trials (with narrow confidence
    intervals) |
| B | 2a | Systematic review of (homogeneous) cohort studies of 'exposed' and '
    unexposed' subjects |
| B | 2b | Individual cohort study / low-quality randomized control studies |
| B | 3a | Systematic review of (homogeneous) case-control studies |
| B | 3b | Individual case-control studies |
| C | 4 | Case series, low-quality cohort or case-control studies, or case reports
    |
| D | 5 | Expert opinions based on non-systematic reviews of results or mechanistic
     studies |"

Think step by step.
**QUESTION**: {question}
**ARTICLE TITLE**: {title}
**ARTICLE CONTENT**:
{context}
```

(a) Prompt used for the summarization step.

```
You are the author of a Cochrane Collaboration systematic review, leveraging
    statistical analysis and assessing risks of bias in order to rigorously assess
    the effectiveness of medical interventions. As part of your review process,
    perform the following task:
Given the ARTICLE SUMMARIES. Provide a concise and precise answer to the provided
    QUESTION.

After you think, return your answer with the following format:
- **Rationale**: Your rationale
- **Full Answer**: A precise answer, citing each fact with the Article ID in
    brackets (e.g. [2]).
- **Answer**: A final classification exactly matching one of the following options:
     Higher, Lower, No Difference, Insufficient Data, Uncertain Effect

Think step by step.
**QUESTION**: {question}
**ARTICLE SUMMARIES**: {context}
```

(b) Prompt used for the final answer step.

Figure 13: Prompts used to generate LLM responses to questions under the expert-guided prompt setup, designed to attempt to explicitly enforce model awareness of evidence quality and strength.

E LLM INSTRUCTION-FOLLOWING RATES

The rate at which LLMs provided valid answer output of any kind is presented as part of Figure 4. Precisely, we measured the per-model instruction-following rate, i.e. the percentage of questions for which the full "Answer" field in the model's final output exactly matched one of the defined answer classes (case-insensitive). We note that a substantial portion of models exhibit a high rate of instruction-following failures: OpenBioLLM 8B and 70B; HuatuoGPT-o1 7B and 70B; Llama 4 Maverick and Scout; Llama 3.0 8B; and Llama 3.1 8B all fail to achieve a 60% instruction-following rate, and only Llama 3.3 70B (Instruct and R1-Distill) achieves perfect instruction-following. We highlight that OpenBioLLM 8B has a 0% instruction-following rate. Lastly, we observe that even when significant portion of the outputs are valid, models still have high error rates, with only an average of $58.1(\pm 5.0)\%$ of valid model outputs being correct. These results demonstrate that, while a high instruction-following rate may diminish performance in small models, poor performance cannot be attributed to instruction-following errors alone.

F LLM PERFORMANCE AS A FUNCTION OF NUMBER OF RELEVANT SOURCES

As shown in Appendix Figure 14, we find no clear general trend between the number of relevant sources and model performance. Notably, this includes performance with a single source (no model achieves even 60% accuracy), highlighting challenges in LLMs' ability to perform systematic review beyond resolving evidence conflicts. The only exceptions to this are the models with the overall poorest performance (colored in red and orange hues, such as HuatuoGPT-o1 7B and Llama 3.0 8B).

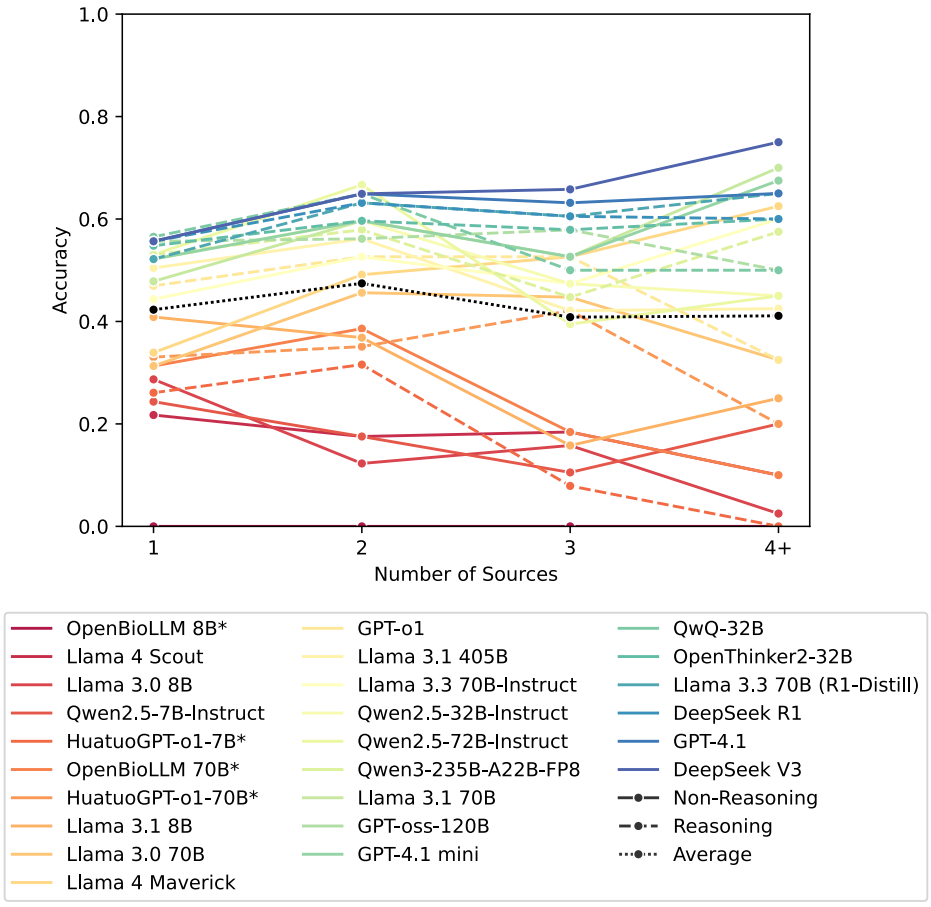

Figure 14: Model accuracy as a function of number of relevant sources.

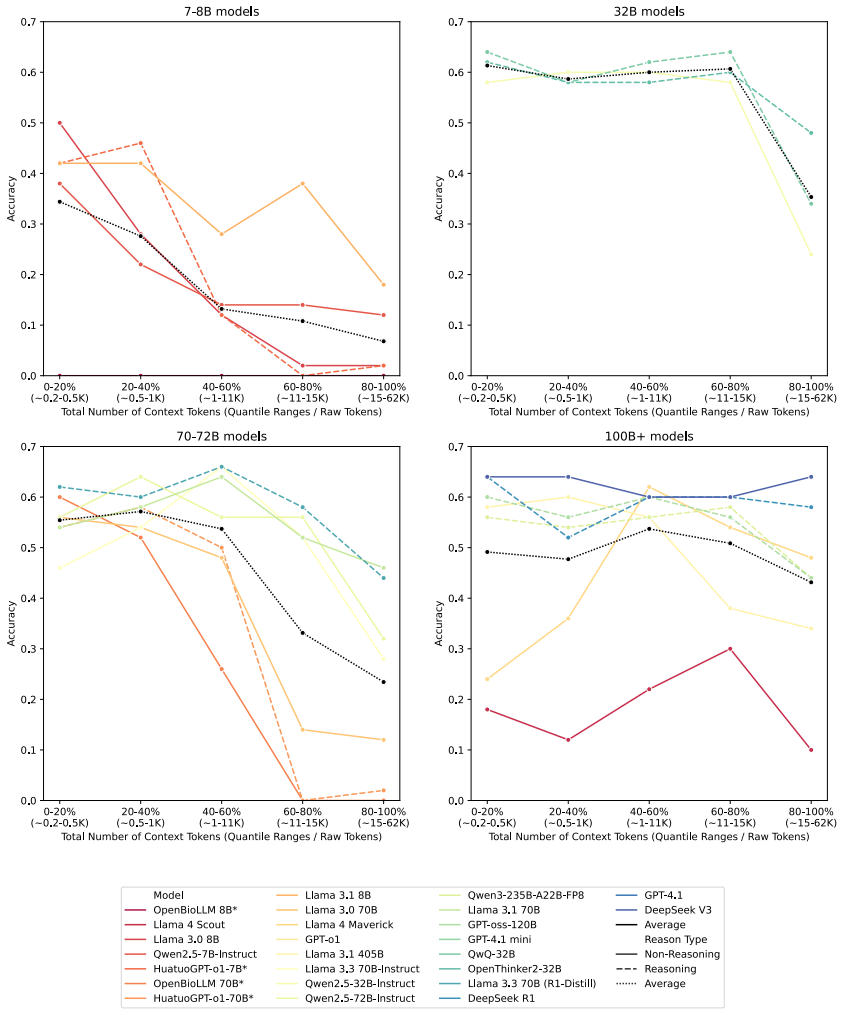

Figure 15: Model performance as a function of the number of tokens in the relevant studies, separated by model size range. Horizontal axis measures the accuracy by 5-quantiles.

G LLM PERFORMANCE AS A FUNCTION OF TOKEN LENGTH OF RELEVANT SOURCES

Given the lack of dependency on the number of sources on average accuracy, we directly investigate the dependency of model performance on the combined token length of all relevant sources; we present these results in Appendix Figure 15. As noted in the main analysis, performance consistently declines at high token counts, except for models with over 100B parameters. Notably, 32B models maintain over 50% average accuracy up to the 80–100% quantile (15K tokens and above). By contrast, 70–72B models fall below 50% accuracy around the 60–80% quantile (11–15K tokens). This decline in the 70–72B range is primarily driven by the underperformance of medically finetuned models (HuatuoGPT-o1 and OpenBioLLM).

H AVERAGE CONFUSION MATRICES FOR TREATMENT OUTCOME EFFECTS

We assess which treatment outcome effect classes are most frequently misclassify by visualizing the confusion matrix averaged across all models. As shown in Figure 16, we observe that models with lower than 40% accuracy significantly skew the confusion matrix toward invalid outputs. However, when considering exclusively models with above 40% performance, we observe two significant

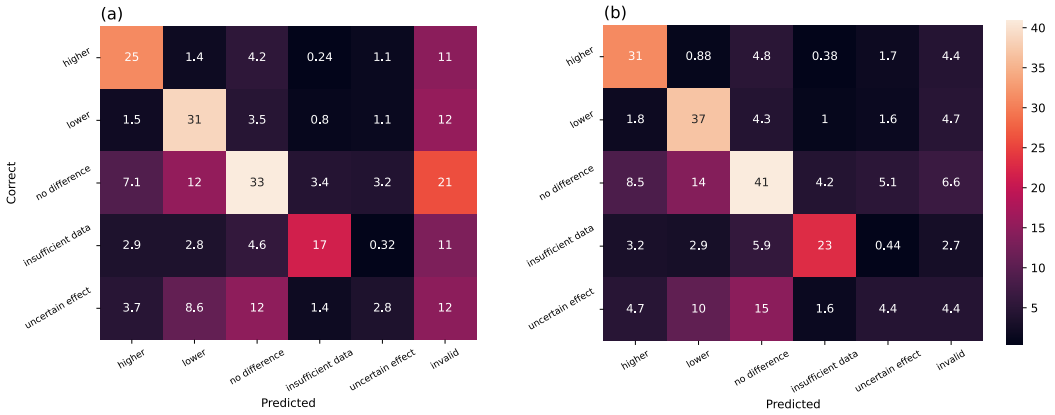

Figure 16: Average confusion matrices using basic prompts. (a) Average confusion matrix aggregated across all models. (b) Average confusion matrix aggregated across models achieving at least 40% overall accuracy.

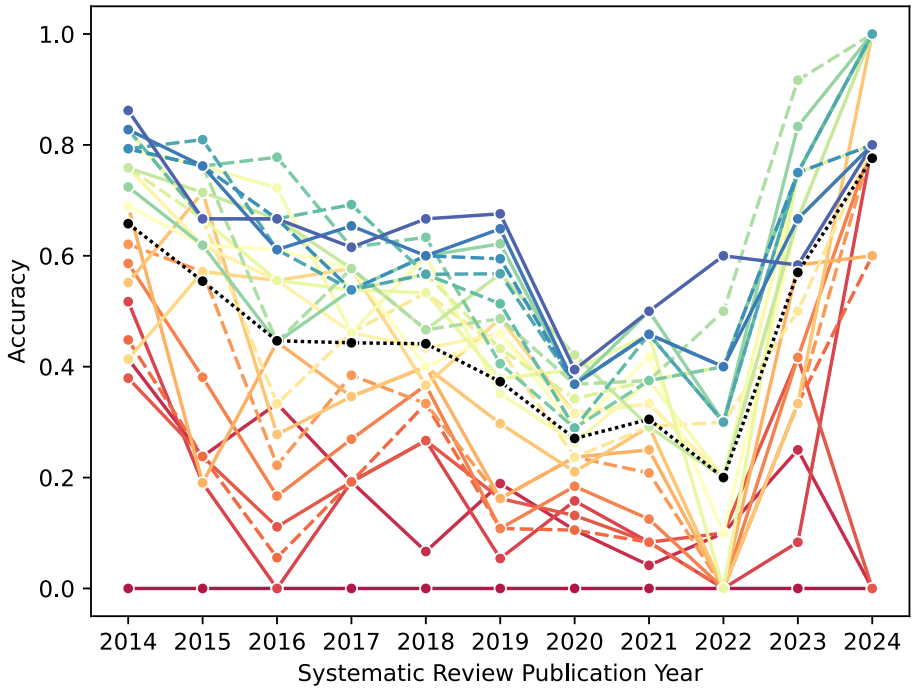

Figure 17: Accuracy by publication year

trends. First, models are consistently unwilling to predict `uncertain effect`. Second, models consistently confuse the `uncertain effect` and `no difference` classes.

For completeness, we provide all individual confusion matrices in Appendix Section U.

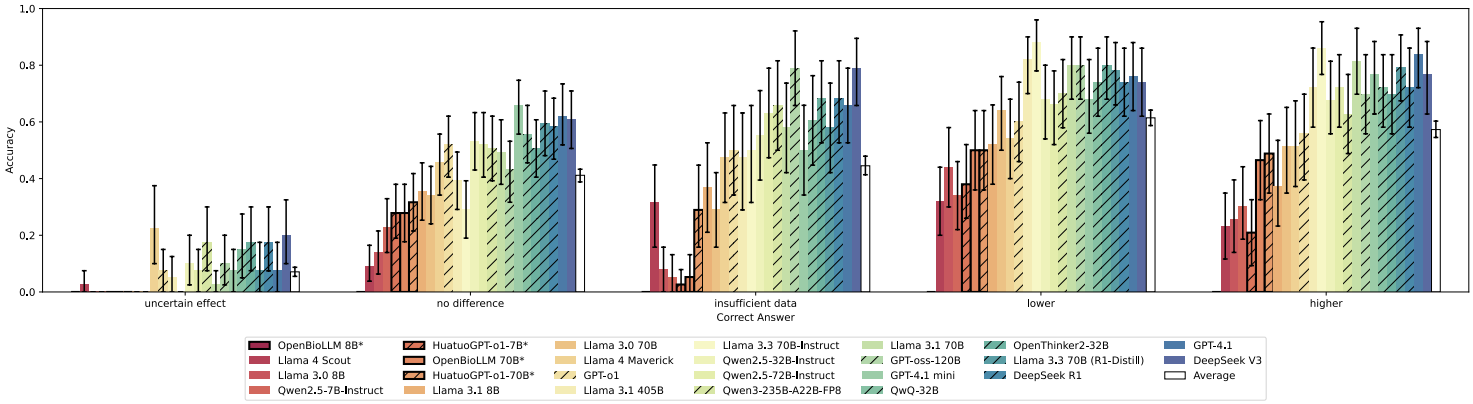

Figure 18: Per-class recall for each individual model. 95% confidence intervals are calculated via bootstrapping with N=1000.

I Performance by review publication year

As shown in Appendix Figure 17, performance steadily declines for more recent publication years, except for 2023 and 2024. These improvements may partially be explained by the fact that the majority of questions from 2024 involve high- or moderate-certainty evidence (as shown in Appendix Figure 11(d)); as a result, these questions are likely easier for models to answer.

J Per-class recall for individual models

We present individual model per-class recall in Appendix Figure 18. Notably, all models, without exception, perform poorly on the `uncertain effect` class. We highlight that Llama 3.3 70B-Instruct outperforms all other models on the `higher` and `lower` classes, but its overall accuracy is held back significantly by its poor performance on the `no difference` and `insufficient data` classes.

K Model performance under the expert-guided prompt setup

To evaluate the dependency of model performance on prompting quality, we leverage an expert-guided prompt setup as described in the main paper and Appendix Section D. Critically, as shown in Appendix Figure 19 and discussed in the main paper, we find that even with a prompt explicitly designed to encourage models to assess the quality of studies, the dependency of model performance on evidence certainty remains. More broadly, as shown in Appendix Figure 20, we find that our more intentionally-designed prompt does not consistently improve model performance; while performance improves for the five models that performed worst under the basic prompt (namely OpenBioLLM 8B, Llama 4 Scout, Llama 3.0 8B, Qwen2.5-7B-Instruct, and HuatuoGPT-o1 7B), we observe that performance actually decreases for several of the models that performed best with the basic prompt, including a nearly 20% drop in performance for DeepSeek V3 (the highest-performing model when using the basic prompt).

L Model performance under additional prompt variations

To further assess whether our results are simply the result of prompting, we evaluate a range of models without chain-of-thought (by default, all our prompts use chain-of-thought) and with different amounts of few-shot examples. As shown in Appendix Table 4, omitting CoT does not provide any significant difference in performance. We do observe in Appendix Table 5 that few-shot evaluation can improve model performance. However, the overall findings remain consistent: across all models, we observe high error rates; zero-shot DeepSeek V3 remains one of the best models; and a performance gap between LLMs and the original human experts persists.

Table 4: Model accuracy with and without chain-of-thought (CoT).

Model	With CoT	No CoT	Delta
Llama 3.0 70B	0.368	0.360	0.008
Qwen2.5-72B-Instruct	0.528	0.512	0.016
GPT-4.1 mini	0.564	0.564	0.000
Llama 3.3 70B (R1-Distill)	0.580	0.576	0.004
DeepSeek V3	0.624	0.604	0.020

Table 5: Model accuracy with several levels of few-shot prompting.

Model	0-shot	1-shot	2-shot	3-shot
Llama 4 Maverick	0.448	0.583 (+0.135)	0.583 (+0.135)	0.632 (+0.184)
Llama 3.3 70B-Instruct	0.492	0.555 (+0.063)	0.571 (+0.079)	0.579 (+0.087)
GPT-4.1 mini	0.564	0.619 (+0.055)	0.595 (+0.031)	0.619 (+0.055)
Llama 3.3 70B (R1-Distill)	0.580	0.599 (+0.019)	0.591 (+0.011)	0.591 (+0.011)
DeepSeek V3	0.624	0.567 (−0.057)	0.555 (−0.069)	0.575 (−0.049)

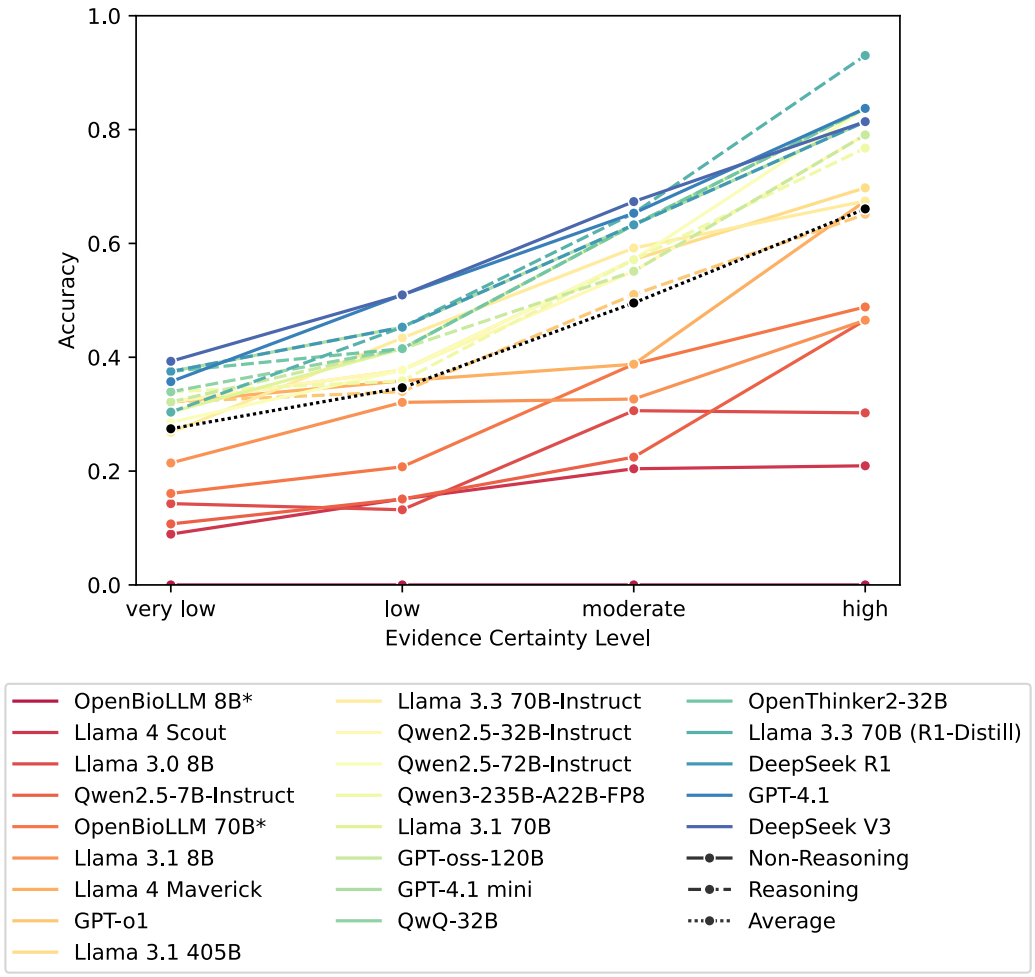

Figure 19: Model accuracy at different evidence qualities when using the expert-guided prompt setup. HuatuoGPT-o1 70B and Llama 3.0 70B are omitted as they were not tested on the expert-guided setup.

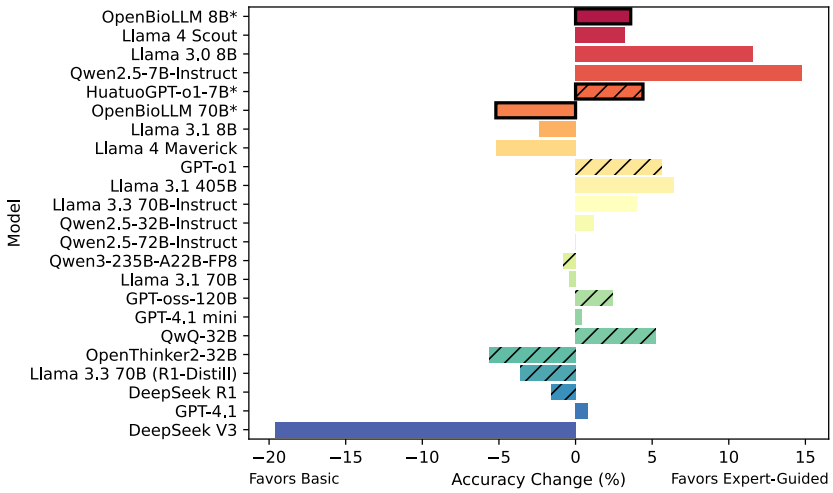

Figure 20: Changes in model performance when using the basic prompt setup versus the expert-guided prompt setup. HuatuoGPT-o1 70B and Llama 3.0 70B are omitted as they were not tested on the expert-guided setup.

M Question correctness across models

As shown in Appendix Figure 21, 53 questions are answered incorrectly by all models, and only 2 are answered correctly by all models (omitting OpenBioLLM 8B, which gets every question wrong). Otherwise, we observe that performance varies significantly across models . A qualitative analysis of these various question types is presented in Appendix Section T.

N Performance by medical specialty

Appendix Figure 22 shows average model accuracy stratified by medical specialty. Models perform significantly worse on questions relating to Psychology & Neurology and Surgery relative to other medical specialties, with accuracies of 27.60% (24.58, 30.52) and 34.09% (31.15, 37.03) respectively. The highest average model performance is observed in the Oncology & Hematology specialty, where models achieve an average accuracy of 63.28% (95% CI: 58.33–68.23).

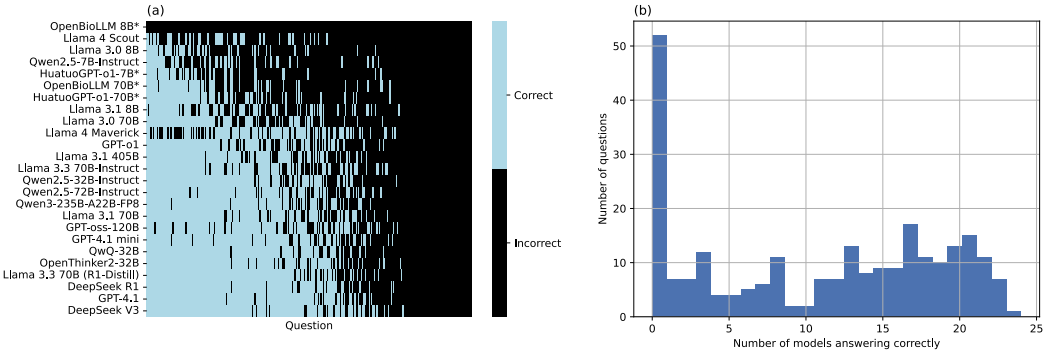

Figure 21: Analyses of model behavior across questions. (a) Questions (columns) that were deemed correct (light blue) or incorrect (black) for each model (rows), sorted by percentage of models with correct responses for that question (x-axis) and by the percentage of questions a model got correct (y-axis). (b) Distribution of questions by the number of models that answered that question correctly.

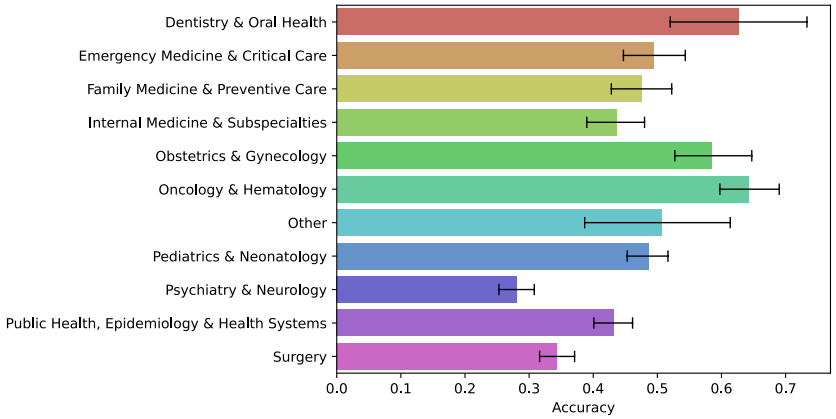

Figure 22: Average model accuracy across all models (and 95% confidence interval) stratified by medical specialty.

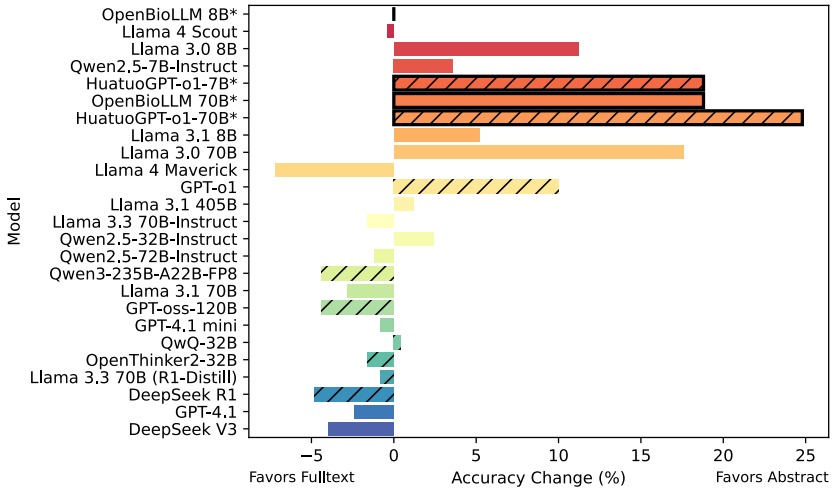

Figure 23: Changes in model performance when providing full-text when available versus always providing only the abstract (basic prompt setup).

O FULL-TEXT VS ABSTRACT SOURCES

We evaluate how model performance differs when using full-text articles versus abstracts alone, using the basic prompt setup in both cases. The results of this experiment are presented in Appendix Figure 23. We find that most models with the poorest overall performance actually experience a boost in accuracy (as high as 24.8% in the case of HuatuoGPT-o1-70B) when given only abstracts, even though abstracts contain less information. This suggests that some models struggle on our dataset because of an inability to handle long context, as full-text articles are much longer than abstracts alone. By contrast, the best-performing models usually perform better when given access to full-text (DeepSeek R1, for instance, gains 4.8% accuracy with full-text access). We note that, due to licensing constraints in scientific publishing, many existing deployments and evaluations of LLM to answer questions using scientific literature rely solely on abstracts (Lozano et al., 2023; OpenEvidence, 2025). Our analysis demonstrates that access to full article text benefits frontier models, underscoring the urgent need to expand such access. We highlight initiatives promoting this shift—for instance, beginning in 2025, all U.S. federally-funded research must be made freely available, which could significantly enhance the performance of already-deployed models.

Table 6: Model accuracy on questions where the sources are all abstracts, all fulltext, or have a mix of fulltext and abstracts, respectively.

Model	Abstracts only	Fulltext only	Mixed sources
DeepSeek V3	0.638	0.460	0.772
GPT-4.1	0.615	0.524	0.667
DeepSeek R1	0.600	0.524	0.632
Llama 3.3 70B (R1-Distill)	0.638	0.429	0.614
OpenThinker2-32B	0.608	0.460	0.614
QwQ-32B	0.631	0.460	0.526
GPT-4.1 mini	0.608	0.444	0.596
GPT-oss-120B	0.592	0.429	0.596
Llama 3.1 70B	0.600	0.381	0.614
Qwen3-235B-A22B-FP8	0.562	0.460	0.561
Qwen2.5-72B-Instruct	0.592	0.444	0.474
Qwen2.5-32B-Instruct	0.608	0.413	0.439
Llama 3.3 70B-Instruct	0.554	0.365	0.491
Llama 3.1 405B	0.592	0.349	0.421
GPT-o1	0.631	0.302	0.281
Llama 4 Maverick	0.377	0.460	0.596
Llama 3.0 70B	0.577	0.000	0.298
Llama 3.1 8B	0.377	0.333	0.246
HuatuoGPT-o1-70B	0.600	0.032	0.035
OpenBioLLM 70B	0.531	0.000	0.000
HuatuoGPT-o1-7B	0.385	0.016	0.000
Qwen2.5-7B-Instruct	0.269	0.063	0.193
Llama 3.0 8B	0.338	0.000	0.053
Llama 4 Scout	0.169	0.238	0.158
OpenBioLLM 8B	0.000	0.000	0.000
Average	0.504	0.298	0.387

To even further investigate this phenomenon, we isolate model performance on questions with different types of sources. As shown in Appendix Table 6, on average, models perform better on questions that use abstracts only as compared to full-text only, demonstrating that current models generally perform worse when provided with the information that would be necessary to replicate full meta-analysis, emphasizing the need for more work on long-context input.

P Performance with internal memory

Table 7: Baseline model accuracy using the provided sources (from Figure 4) versus model accuracy using internal memory only.

Model	With Evidence	Memory Only	Delta
Llama 3.0 70B	0.368	0.116	-0.252
Qwen2.5-72B-Instruct	0.528	0.388	-0.140
GPT-4.1 mini	0.564	0.428	-0.136
Llama 3.3 70B (R1-Distill)	0.580	0.396	-0.184
DeepSeek V3	0.624	0.324	-0.300

To evaluate the effects of potential data leakage, we perform a small-scale experiment to quantify if models are able to answer questions in our benchmark (which were newly formulated for the purpose of this work) by prompting them to answer the questions without providing any articles as context. As shown in Appendix Table 7, models perform far worse when asked to answer the questions from memory; this trend is consistent across the range of baseline performances and in models from varying providers, with Llama 3.0 70B even performing worse than random.

Q PERFORMANCE WITH VARYING SOURCE ORDER

Table 8: Accuracy with original order of sources versus randomly shuffled sources.

Model	Original Order	Shuffled	Delta
Llama 3.0 70B	0.368	0.356	+0.012
Qwen2.5-72B-Instruct	0.528	0.500	+0.028
GPT-4.1 mini	0.564	0.548	+0.016
Llama 3.3 70B (R1-Distill)	0.580	0.564	+0.016
DeepSeek V3	0.624	0.572	+0.052

To ensure that our results did not depend on limitations of multi-step refinement (especially in questions with full-text articles as evidence), we compare baseline model performance to performance with the sources randomly shuffled to a new order. As shown in Appendix Table 8, randomizing the source order does not result in performance changes beyond the ranges expected by our confidence intervals; thus, our results suggest that source order has no statistically-significant effect on model performance.

R CONSTRAINED HUMAN BASELINE EVALUATION ON MEDEVIDENCE

To better benchmark model performance, we evaluate the performance of humans under similar conditions to the models (i.e. short time constraints, with access to only the abstracts when models similarly do not have full-text access). We recruited three external clinicians (with no connection to this work) with diverse specialties (e.g., pediatrics, oncology) and an average of four years of experience. Each clinician answered 20 randomly selected questions using the same inputs provided to the language models (i.e. abstract-only when models would only have abstract access). The average response time was approximately 10 minutes per question and the mean accuracy was 71.67% (min=65%, max=75%). Notably, every question was answered correctly by at least one clinician. Inter-rater agreement was moderate with a Fleiss' kappa of 0.43.

S HUMAN EVALUATION FOR SOURCE CONCORDANCE

To verify that the correlation between model performance and source-level concordance is not itself model induced (since DeepSeek V3 is used to generate the source-level concordance), we evaluated whether human experts and and DeepSeek V3 determine source concordance in the same way. To do this, we re-recruited two of the external clinicians used for the human baseline evaluation (Appendix R) and asked them to each answer 50 randomly questions from the dataset, provided only one randomly-selected relevant source for that question. As before, the clinicians only had access to the same inputs as the models (i.e. only the abstract when models would only have abstract access). We then evaluated the agreement between humans and DeepSeek V3 on concordance, where concordance takes on 2 values: `TRUE` if the individual source's evaluation of the treatment outcome effect was aligned with the overall conclusion of the SR, and `FALSE` if the individual source would lead to a conclusion different from the overall SR. On average, the clinicians agreed with the LLM-produced source concordance 74% of the time.

T QUALITATIVE ANALYSIS

To provide a better qualitative understanding of success and failure modes on our dataset, we explore the outputs (under the basic prompt setup) of our models directly. First, we closely inspect our best-performing model, DeepSeek V3, and provide some commentary on the quality of the model's output. Specifically, we consider four types of questions: (1) Questions where all models get the answer wrong. (2) Questions where most models are correct, but DeepSeek V3 is incorrect. (3) Questions where most models are correct, including DeepSeek V3. (4) Questions where most models are wrong, but DeepSeek V3 is correct. Second, we investigate additional failure modes demonstrated by the two other best-performing models and the three worst-performing models.

T.1 DEEP-DIVE ON DEEPSEEK V3

T.1.1 QUESTIONS WHERE ALL MODELS ARE WRONG

Question ID 1 - 0/24 models correct

Review PMID: 28660687
Question: Is the short-term rate of overall lymphocyst formation higher, lower, or the same when comparing retroperitoneal drainage to no drainage?
Correct Answer: `no difference`
PMIDs of Relevant Sources: 9190979, 12214830

DeepSeek V3 Response:

```
- **Rationale**:
  - Article ID 9190979 reports that the complication rate, including
      lymphocyst formation, was significantly higher in the drained
      group (Group A) compared to the not-drained group (Group B) (
     P = 0.01). However, sonographic monitoring showed free
     abdominal fluid (a proxy for lymphocyst formation) in 18% of
     drained patients vs. 36% of not-drained patients (P = 0.03),
     suggesting a lower rate in the drained group. This creates a
     contradiction in the findings.
  - Article ID 12214830 found no significant difference in
     lymphocyst formation between the drained and not-drained
     groups (P = 0.2), with asymptomatic lymphocysts detected only
     in the not-drained group (though statistically insignificant).

- **Full Answer**:
  - Article ID 9190979 reports a higher complication rate (including
      lymphocyst-related morbidity) in the drained group [9190979],
      but sonographic data suggests a lower rate of free abdominal
     fluid (a proxy for lymphocyst formation) in the drained group
     (18% vs. 36%, P = 0.03) [9190979]. Article ID 12214830 found
     no significant difference in lymphocyst formation between
     groups (P = 0.2) [12214830]. The evidence is conflicting, with
      one study suggesting a lower rate and another showing no
     difference.

- **Answer**: Uncertain Effect
```

Remarks: In the SR, authors perform meta-analysis and find no overall significant difference between retroperitoneal drainage and no drainage. DeepSeek V3 does not appear to know how to perform any kind of reconcilliation. We note that Article 12214830 does not specify the raw number of events under both treatments, but does state that no lymphocysts (i.e. 0) were detected in the drainage group.

Question ID 198 - 0/24 models correct

Review PMID: 35514111
Question: Is the rate of any PCR-positive SARS-COV-2 infection higher, lower, or the same when comparing test-based attendance to standard 10-day self-isolation?
Correct Answer: `uncertain effect`
PMIDs of Relevant Sources: 34534517

DeepSeek V3 Response:

```
- **Rationale**: The study compared daily testing (test-based
    attendance) with standard 10-day self-isolation for school-based
     contacts of SARS-CoV-2 cases. The primary outcomes included
```

```
    rates of symptomatic PCR-confirmed infections and COVID-19-
    related absences. The results showed no significant difference
    in the rate of symptomatic PCR-confirmed infections between the
    two groups, with the intervention group (daily testing) being
    non-inferior to the control group (self-isolation). The adjusted
     incidence rate ratio (aIRR) for symptomatic PCR-confirmed
    infections was 0.96 (95% CI -0.751.22), indicating no
    statistically significant difference.

- **Full Answer**: The rate of PCR-positive SARS-CoV-2 infection
    was not significantly different when comparing test-based
    attendance (daily testing) to standard 10-day self-isolation (
    aIRR 0.96, 95% CI -0.751.22) [34534517].

- **Answer**: No Difference
```

Remarks: The authors of the SR state that the results are inconclusive for the following reasons: (a) "risk of bias: considerable missing data for PCR-tests for staff (data were available for 76% control schools and 83% intervention schools)" and (b) "imprecision: the confidence intervals are wide and the 95% CI crosses the line of no effect." In this case, the full-text article is open-source, so LLMs theoretically had access to essentially all the same information as the SR authors, but evidently, none of these concerns are addressed in DeepSeek V3's response. Furthermore, the DeepSeek V3 response focuses only on symptomatic PCR-confirmed infections, demonstrating limitations in its knowledge of medical contexts, as the question calls for an assessment of any PCR positive infections, which includes asymptomatic cases.

Question ID 32 - 0/24 models correct

Review PMID: 26447539
Question: Is subjective fatigue higher, lower, or the same when comparing methylphenidate to placebo?
Correct Answer: `no difference`
PMIDs of Relevant Sources: 17674415

DeepSeek V3 Response:

```
- **Rationale**: The study (Article ID 17674415) directly compares
    methylphenidate to placebo in treating fatigue in Parkinson's
    disease patients. The results show that methylphenidate
    significantly reduced fatigue scores (FSS and MFI) compared to
    placebo, where reductions were nonsignificant. This indicates
    that subjective fatigue is lower with methylphenidate than with
    placebo.

- **Full Answer**: Subjective fatigue is lower with methylphenidate
     compared to placebo, as evidenced by significant reductions in
    Fatigue Severity Scale (FSS) and Multidimensional Fatigue
    Inventory (MFI) scores in the methylphenidate group (P < 0.04),
    while placebo group reductions were nonsignificant [17674415].

- **Answer**: Lower
```

Remarks: This is a relatively challenging question—in the original SR, the authors compare the improvements in FSS in the placebo group directly to the improvements in the treatment group, rather than comparing the two treatment groups separately to their own baselines (which is what the authors of the cited source do). The source's abstract does state that smaller but non-significant improvements were present in the placebo group, albeit without

providing numbers. Regardless, the LLM's response reads as a regurgitation of facts and lacks the critical thinking needed to critically assess cited studies.

T.1.2 QUESTIONS WHERE MOST MODELS ARE CORRECT, BUT DEEPSEEK V3 IS WRONG

Question ID 6 - 16/24 models correct

Review PMID: 26346232
Question: Is stroke prevention higher, lower, or the same when comparing Transcatheter Device Closure (TDC) to medical therapy?
Correct Answer: `no difference`
PMIDs of Relevant Sources: 22417252, 23514285, 23514286

DeepSeek V3 Response:

```
- **Rationale**: The three articles provide mixed evidence on the
    effectiveness of Transcatheter Device Closure (TDC) compared to
    medical therapy for stroke prevention in patients with patent
    foramen ovale (PFO). Article ID 22417252 found no significant
    difference in stroke prevention between TDC and medical therapy
    (P=0.37 for the primary endpoint). Article ID 23514285 also
    reported no significant reduction in recurrent embolic events or
     death with TDC (P=0.34 for the primary endpoint). However,
    Article ID 23514286 showed no significant benefit in the
    intention-to-treat analysis (P=0.08) but did find a significant
    reduction in recurrent stroke in the per-protocol and as-treated
     analyses (P=0.03 and P=0.007, respectively). Given the
    inconsistency in results across studies and analyses, the
    overall evidence is not conclusive.

- **Full Answer**: The evidence is mixed. Two studies (Article ID
    22417252 and Article ID 23514285) found no significant
    difference in stroke prevention between TDC and medical therapy
    [22417252][23514285]. A third study (Article ID 23514286)
    reported no significant benefit in the intention-to-treat
    analysis but found TDC superior in per-protocol and as-treated
    analyses [23514286].

- **Answer**: Uncertain Effect
```

Remarks: In this case, unlike most models (16/24 to be exact), DeepSeek V3 seems to 'overthink' the question and hone in on a single case where a sub-analysis of in Article 23514286 finds a significant benefit. Notably, SRs typically use intention-to-treat analysis (https://training.cochrane.org/handbook/current/chapter-06), which DeepSeek V3 fails to consider. Furthermore, as in prevous cases, no clear efforts to perform meta-analysis or aggregate studies are shown in DeepSeek V3's response.

Question ID 106 - 15/24 models correct

Review PMID: 38197546
Question: Is the risk of cholera at two-year follow-up higher, lower, or the same when comparing two doses of Dukoral with or without a booster dose to placebo?
Correct Answer: `lower`
PMIDs of Relevant Sources: 7967990, 10823767

DeepSeek V3 Response:

```
- **Rationale**:
```

```
  - Article ID 7967990 evaluates the short-term efficacy of two
      doses of WC/rBS cholera vaccine without a booster, showing 86%
       protective efficacy over a mean follow-up of 18 weeks.
      However, it does not provide data on long-term (two-year)
      efficacy or the effect of a booster dose.
  - Article ID 10823767 evaluates the two-year efficacy of the same
      vaccine, comparing two doses (without booster) to three doses
      (with booster). After two doses, the vaccine showed no
      efficacy (VE = -4%), but after the booster dose, efficacy
      increased to 61%. This suggests that the booster dose
      significantly improves long-term protection compared to two
      doses alone.

- **Full Answer**:
  - Without a booster dose (two doses only), the vaccine showed no
      protective efficacy (VE = -4%) at two-year follow-up
      [10823767].
  - With a booster dose (three doses), the vaccine demonstrated 61%
      protective efficacy at two-year follow-up [10823767].
  - Therefore, the risk of cholera at two-year follow-up is **higher
      ** with two doses alone compared to two doses plus a booster.

- **Answer**: Higher
```

Remarks: This is case where DeepSeek V3 seems to fail due to a misinterpretation of the question, although we believe the semantics of this question are unambiguous. DeepSeek V3 does not seem to pick up on the mention of placebo in the question at all.

Question ID 163 - 16/24 models correct

Review PMID: 25739381
Question: Is neonatal mortality higher, lower, or the same when comparing community-led chlorhexidine cord cleansing to community-led dry cord care?
Correct Answer: `lower`
PMIDs of Relevant Sources: 22322124, 16546539, 22322126

DeepSeek V3 Response:

```
- **Rationale**: The question asks whether neonatal mortality is
    higher, lower, or the same when comparing community-led
    chlorhexidine cord cleansing to community-led dry cord care. The
     three articles provide evidence from randomized trials in
    different settings (Bangladesh, Nepal, and Pakistan) comparing
    chlorhexidine cord cleansing with dry cord care. All three
    studies report reductions in neonatal mortality with
    chlorhexidine use, though the magnitude and statistical
    significance vary.

- **Full Answer**:
  - In Bangladesh, neonatal mortality was lower with single
      chlorhexidine cleansing (22.5 per 1000 livebirths) compared to
       dry cord care (28.3 per 1000 livebirths; RR 0.80 [0.65-0.98])
       [22322124].
  - In Nepal, neonatal mortality was 24% lower with chlorhexidine (
      RR 0.76 [0.55-1.04]) compared to dry cord care, with a
      significant 34% reduction in infants enrolled within 24 hours
      (RR 0.66 [0.46-0.95]) [16546539].
  - In Pakistan, chlorhexidine cleansing reduced neonatal mortality
      (RR 0.62 [0.45-0.85]) compared to dry cord care [22322126].

- **Answer**: **Lower**
```

Remarks: An example of an instruction-following failure for DeepSeekV3 (even though the overall analysis leads to a correct solution).

T.1.3 Questions where most models are correct, including DeepSeek V3

Question ID 160 - 21/24 models correct

Review PMID: 28829911
Question: Is the risk of treatment-related mortality higher, lower, or the same when comparing chemoradiotherapy plus esophagectomy to chemoradiotherapy alone?
Correct Answer: `higher`
PMIDs of Relevant Sources: 17401004, 15800321

DeepSeek V3 Response:

```
- **Rationale**: Both articles compare chemoradiotherapy plus
    esophagectomy (surgery) to chemoradiotherapy alone in patients
    with locally advanced esophageal cancer. Article ID 17401004
    reports a 3-month mortality rate of 9.3% in the surgery group (
    arm A) compared to 0.8% in the chemoradiation-alone group (arm B
    ), with a statistically significant difference (P = .002).
    Article ID 15800321 similarly reports a significantly higher
    treatment-related mortality in the surgery group (12.8%)
    compared to the chemoradiotherapy-alone group (3.5%) (P = .03).
    Both studies consistently demonstrate that adding surgery to
    chemoradiotherapy increases treatment-related mortality.

- **Full Answer**: The risk of treatment-related mortality is
    higher when comparing chemoradiotherapy plus esophagectomy to
    chemoradiotherapy alone, as evidenced by significantly higher
    mortality rates in the surgery groups (9.3% vs. 0.8% [17401004];
     12.8% vs. 3.5% [15800321]).

- **Answer**: Higher
```

Remarks: While this question involves multiple sources, both sources individually agree with the overall correct conclusion, making it fairly straightforward to answer; consequently, almost all models get this question correct.

Question ID 214 - 23/24 models correct

Review PMID: 37594020
Question: Is pain relief at 48 hours higher, lower, or the same when comparing NSAIDs to placebo?
Correct Answer: `higher`
PMIDs of Relevant Sources: 8871138

DeepSeek V3 Response:

```
- **Rationale**: The study compared ibuprofen (an NSAID) to
    acetaminophen and placebo for pain relief in children with acute
     otitis media over 48 hours. The results showed that ibuprofen
    was statistically superior to placebo (P < 0.01) in relieving
    pain, with 7% of children still suffering pain in the ibuprofen
    group compared to 25% in the placebo group. There was no
    significant difference between acetaminophen and placebo for
    pain relief.
```

```
- **Full Answer**: Pain relief at 48 hours was higher with NSAIDs (
    ibuprofen) compared to placebo, as evidenced by a statistically
    significant reduction in pain (7% vs. 25% of children still
    suffering pain, P < 0.01) [8871138].

- **Answer**: Higher
```

Remarks: A straightforward question that all models (except OpenBioLLM 8B) get correct: one source accurately directly reports on relevant findings without conclusions being revised by the SR authors' critical analysis.

T.1.4 QUESTIONS WHERE DEEPSEEK V3 IS CORRECT, DESPITE MOST MODELS BEING WRONG

Question ID 235 - 3/24 models correct

Review PMID: 32104914
Question: Is quality of life higher, lower, or the same when comparing computerised cognitive training (CCT) of at least 12 weeks to no intervention?
Correct Answer: `insufficient data`
PMIDs of Relevant Sources: 23531885, 25511081, 20418350

DeepSeek V3 Response:

```
- **Rationale**: The question asks whether quality of life (QoL) is
     higher, lower, or the same when comparing computerized
    cognitive training (CCT) of at least 12 weeks to no intervention
    . The provided articles focus on cognitive and physical
    performance improvements but do not explicitly measure or report
     QoL outcomes. Article ID 23531885 discusses cognitive
    improvements from CCT but does not mention QoL. Article ID
    25511081 examines dual-task gait and divided attention but also
    does not address QoL. Article ID 20418350 evaluates cognitive
    performance after mental and physical activity interventions but
     similarly lacks QoL data.

- **Full Answer**: None of the articles provide data on quality of
    life (QoL) outcomes when comparing CCT of at least 12 weeks to
    no intervention [23531885, 25511081, 20418350].

- **Answer**: Insufficient Data
```

Remarks: DeepSeek V3 correctly recognizes that none of the relevant studies provide direct, quantitative analysis on QoL outcomes; given the very low model correctness rate (3/24), other models likely attempted to apply colloquial definitions of "quality of life" rather than the narrowly-defined medical definition requiring explicit quantitative measurement.

Question ID 89 - 2/24 models correct

Review PMID: 31557310
Question: Is the incidence of surgical site infection (SSI) for patients undergoing breast cancer surgery without reconstruction higher, lower, or the same when comparing prophylactic antibiotics administered preoperatively to placebo?
Correct Answer: `lower`
PMIDs of Relevant Sources: 7480237, 9776150, 23052809, 10873356, 19673597, 2403655, 2403697

DeepSeek V3 Response:

```
- **Rationale**: The question asks about the incidence of surgical
    site infection (SSI) in breast cancer surgery without
    reconstruction when comparing prophylactic antibiotics to
    placebo. The relevant articles provide mixed results: some show
    a significant reduction in SSI with antibiotics (e.g.,
    [7480237], [2403655]), while others show no significant
    difference (e.g., [23052809], [10873356], [2403697]). Notably,
    [7480237] specifically highlights a significant reduction in
    infections for breast surgery, while [2403655] also supports a
    reduction in infections for breast surgery. However, other
    studies, such as [23052809] and [10873356], found no significant
     difference. The variability in results suggests that the effect
     may depend on the type of breast surgery (e.g., mastectomy vs.
    axillary lymph node dissection) or other factors.

- **Full Answer**: The incidence of SSI is **lower** with
    prophylactic antibiotics compared to placebo in some breast
    cancer surgeries, as shown by significant reductions in
    infections in studies involving plastic surgery ([7480237]) and
    certain breast procedures ([2403655]). However, other studies
    found no significant difference for modified radical mastectomy
    ([23052809]), clean elective breast surgery ([10873356]), or
    mastectomy ([2403697]). The overall evidence suggests a context-
    dependent effect, with some procedures benefiting from
    prophylaxis.

- **Answer**: Lower
```

Remarks: While DeepSeek V3 does get this question right, and shows flashes of attempting deeper reasoning (e.g. the remark that the effect "may depend on the type of breast surgery"), but this analysis is not directly supported by similar statements from the original SR authors. Additionally, there is still no formal aggregation or meta-analysis attempted, and the level of detail in the reasoning provided by DeepSeek V3 makes it hard to tell if the model's predicted answer was made via some form of intuition or if the model just got a bit lucky.

Question ID 173 - 1/24 models correct

Review PMID: 31425604
Question: Is the risk of necrotizing enterocolitis higher, lower, or the same when comparing monitoring of gastric residual volume and quality to only monitoring of gastric residual quality?
Correct Answer: `uncertain effect`
PMIDs of Relevant Sources: 29866595

DeepSeek V3 Response:

```
- **Rationale**: The study compared two groups: one where gastric
    residual volume (GRV) was measured only in the presence of
    specific clinical signs (study group) and another where GRV was
    routinely measured (control group). The primary outcome was time
     to reach full feeding volumes, but secondary outcomes included
    necrotizing enterocolitis (NEC). The study found no significant
    differences in NEC incidence between the two groups, with only 2
     cases reported in the control group and none in the study group
    . However, the small sample size (87 infants) and low event rate
     limit the ability to draw definitive conclusions about NEC risk
    .

- **Full Answer**: The study found no significant difference in the
     incidence of necrotizing enterocolitis between the group where
```

```
    gastric residual volume was measured only in the presence of
    clinical signs (0 cases) and the group where it was routinely
    measured (2 cases) [29866595]. The small sample size and low
    event rate suggest uncertainty in the effect.

- **Answer**: Uncertain Effect
```

Remarks: Even without explicit prompting, DeepSeek V3 recognizes the weakness of the limited sample size/total number of events—the fact that only DeepSeek V3 gets this question correct shows both the current limitations of models' ability to assess uncertainty, as well as the promise that they may be able to do so consistently in the future.

T.2 Additional examples of failure modes

T.2.1 GPT-4.1

Question ID 82 - 1/24 models correct

Review PMID: 31829446
Question: Is the rate of major complications higher, lower, or the same when comparing restrictive fluid therapy (RFT) to goal-directed fluid therapy (GDFT)?
Correct Answer: `uncertain effect`
PMIDs of Relevant Sources: 26471495, 22710266, 25595308, 25342408, 23132508

GPT-4.1 Response:

```
- **Rationale**:
To answer whether the rate of major complications is higher, lower,
     or the same when comparing restrictive fluid therapy (RFT) to
    goal-directed fluid therapy (GDFT), we need direct comparisons
    from randomized controlled trials or similar studies. Four
    relevant articles are summarized:

1. **Article 26471495**: Compared GDFT, restrictive (PRESSURE), and
     standard (CONTROL) fluid therapy in hip/knee replacement. The
    rate of major complications was 5% (GDFT) vs. 15% (restrictive),
     but this difference was not statistically significant (P=0.27).
2. **Article 22710266**: Compared GDFT (Doppler) to zero-balance (
    restrictive) in colorectal surgery. No significant difference in
     major complications (P=0.62).
3. **Article 25595308**: Compared GDFT to standard (not
    specifically restrictive) in CRS/HIPEC. GDFT had significantly
    fewer major complications, but the comparator was standard, not
    restrictive.
4. **Article 25342408**: Compared GDFT to restrictive in elective
    colorectal surgery. Major complications: 2% (GDFT) vs. 8% (
    restrictive), not statistically significant (P=0.36).
5. **Article 23132508**: Compared GDFT to fluid restriction in
    colectomy. Number of patients with complications was nearly
    identical (26 vs. 27; P=1.000).

Across these studies, the rate of major complications is
    numerically lower with GDFT in some cases, but none of the
    differences between GDFT and restrictive fluid therapy reach
    statistical significance. Therefore, the evidence consistently
    shows no significant difference in major complication rates
    between RFT and GDFT.

- **Full Answer**:
```

```
Multiple randomized controlled trials directly comparing
    restrictive fluid therapy (RFT) and goal-directed fluid therapy
    (GDFT) in various surgical settings (colorectal, colectomy, hip/
    knee replacement) found no statistically significant difference
    in the rate of major complications between the two approaches.
    Article 26471495 reported major complication rates of 5% (GDFT)
    vs. 15% (restrictive), P=0.27; Article 22710266 found no
    significant difference (P=0.62); Article 25342408 found 2% (GDFT
    ) vs. 8% (restrictive), P=0.36; and Article 23132508 found
    nearly identical rates (26 vs. 27 patients, P=1.000). Thus, the
    current evidence indicates no difference in major complication
    rates between RFT and GDFT
    [26471495][22710266][25342408][23132508].

- **Answer**: No Difference
```

Remarks: In the SR, the authors consider the evidence poor because of study limitations, imprecision of results, and indirectness of evidence. Much like DeepSeek V3, GPT-4.1 does not take into consideration any of these potential limitations and assumes the provided results are entirely trustworthy.

Question ID 183 - 0/24 models correct

Review PMID: 30320433
Question: Is the number of fallers higher, lower, or the same when comparing exercise to usual care?
Correct Answer: `insufficient data`
PMIDs of Relevant Sources: 19335674

GPT-4.1 Response:

```
- **Rationale**: The article reports that women who exercised had
    fewer falls compared to those who did not exercise. However, it
    also states that the difference in the number of falls between
    the two groups was not statistically significant. This means
    that, while there was a numerical reduction in fallers in the
    exercise group, the evidence does not support a definitive
    conclusion that exercise leads to a higher or lower number of
    fallers compared to usual care.
- **Full Answer**: According to Article ID 19335674, women in the
    exercise group had fewer falls than those in the usual care
    group, but the difference in the number of falls between the two
     groups was not statistically significant [19335674].
- **Answer**: No Difference
```

Remarks: The original article discusses only the number of fall events, not the number of patients who fell. GPT-4.1 fails to recognize this distinction, and thus misinterprets the provided data.

T.2.2 DeepSeek-R1

Question ID 2 - 13/24 models correct

Review PMID: 25734590
Question: Is the rate of clinician impression of cognitive change higher, lower, or the same when comparing cholinase inhibitors to placebo in patients with MS?
Correct Answer: `higher`
PMIDs of Relevant Sources: 15534239, 21519001

DeepSeek-R1 Response:

```
- **Rationale**: Two studies (15534239 and 21519001) investigated
    donepezil (a cholinesterase inhibitor) versus placebo in MS
    patients. The first study (15534239) reported a statistically
    significant higher clinician-reported cognitive improvement with
     donepezil (p = 0.036). The second, larger multicenter study
    (21519001) found no significant difference in clinician
    impression (p = 0.097), though a non-significant trend favored
    donepezil. These conflicting results create uncertainty about
    the overall effect.
- **Full Answer**: The rate of clinician impression of cognitive
    change was higher with donepezil in one trial [15534239] but
    showed no significant difference in another larger trial
    [21519001], leading to conflicting evidence.
- **Answer**: Uncertain Effect
```

Remarks: In the original SR, the formal meta-analysis reveals an overall statistically significant effect on the clinician impression of cognitive change. DeepSeek R1 only tries to directly compare the two individual papers, and this simple strategy fails to correctly resolve the seemingly conflicting evidence.

Question ID 27 - 0/24 models correct

Review PMID: 26447539
Question: Is fatigue severity higher, lower, or the same when comparing doxepin to placebo?
Correct Answer: `insufficient data`
PMIDs of Relevant Sources: 23561946

DeepSeek-R1 Response:

```
- **Rationale**: The study (Article ID 23561946) compared doxepin
    to placebo in Parkinson's patients with insomnia and measured
    fatigue severity as a secondary outcome. Results showed that
    doxepin significantly reduced the fatigue severity scale
    compared to placebo (p = 0.02), indicating lower fatigue
    severity in the doxepin group.
- **Full Answer**: Fatigue severity was lower with doxepin compared
     to placebo, as evidenced by a statistically significant
    reduction in the fatigue severity scale [23561946].
- **Answer**: Lower
```

Remarks: The SR authors find 12 patients to be an insufficient sample size to make a conclusive evaluation of fatigue severity, while DeepSeek R1 blindly trusts the authors of the paper. This type of failure highlights that models may be too easily swayed by the style in which the scientific papers are written.

T.2.3 LLAMA 3.0 8B

Question ID 56 - 3/24 models correct

Review PMID: 27782297
Question: Is the risk of treatment-related sore throat/dry mouth higher, lower, or the same when comparing botulinum toxin type A (BtA) to botulinum toxin type B (BtB)?
Correct Answer: `lower`
PMIDs of Relevant Sources: 16275831, 18098274

Llama 3.0 8B Response:

```
**Rationale**: The question asks whether the risk of treatment-
    related sore throat/dry mouth is higher, lower, or the same when
     comparing botulinum toxin type A (BtA) to botulinum toxin type
    B (BtB). To answer this question, we need to analyze the adverse
     events reported in the two articles.

**Full Answer**: According to Article ID 16275831, dry mouth was
    more frequent with BoNTB (80%) compared to BoNTA (41%) [1]. In
    Article ID 18098274, mild dry mouth was more frequent with BoNTB
    , but there were no differences for moderate/severe dry mouth
    [2]. There is no mention of sore throat in either article.

**Answer**: Based on the available data, the risk of treatment-
    related dry mouth is higher with BtB compared to BtA. However,
    there is no data on sore throat to make a comparison. Therefore,
     the answer is: Higher (for dry mouth) and Insufficient Data (
    for sore throat).
```

Remarks: Llama 3.0 often fails to follow the instruction format, and in particular, does so while failing to conclusively pick a side for its final evaluation.

Question ID 94 - 17/24 models correct

Review PMID: 30746689
Question: Is the likelihood of disease-free survival higher, lower, or the same when comparing nerve-sparing radical hysterectomy to standard radical hysterectomy?
Correct Answer: `no difference`
PMIDs of Relevant Sources: 25872890

Llama 3.0 8B Response:

```
Based on the new context, I refined the original summary as follows:


**Rationale**: To answer this question, I will analyze the article
    summaries to identify the outcomes related to disease-free
    survival in patients undergoing nerve-sparing radical
    hysterectomy (NSRH) versus standard radical hysterectomy (SRH).

**Full Answer**: A systematic review and meta-analysis published in
     [1] found that NSRH was associated with a significantly higher
    rate of disease-free survival at 3 years (RR 1.23, 95% CI
    1.04-1.46) compared to SRH. Another study published in [3]
    reported that NSRH resulted in a higher rate of disease-free
    survival at 5 years (HR 1.34, 95% CI 1.06-1.69) compared to SRH.
     However, a randomized controlled trial published in [2] found
    no significant difference in disease-free survival between NSRH
    and SRH at 2 years (HR 0.93, 95% CI 0.65-1.34).

**Answer**: **Higher**. The majority of the studies suggest that
    NSRH is associated with a higher likelihood of disease-free
    survival compared to SRH, although the evidence is not uniform
    and the results may vary depending on the specific population
    and follow-up period.

**Refined Summary**: The available evidence suggests that nerve-
    sparing radical hysterectomy (NSRH) may be associated with a
    higher rate of disease-free survival compared to standard
    radical hysterectomy (SRH), particularly at longer follow-up
```

```
    periods. However, the evidence is not uniform, and the results
    may vary depending on the specific population and follow-up
    period. A recent randomized controlled trial published in [2]
    found no significant difference in disease-free survival between
     NSRH and SRH at 2 years, but other studies have reported higher
     rates of disease-free survival with NSRH at 3 and 5 years.

**Additional Insights**: The study published in [1] found that NSRH
     was associated with a higher rate of disease-free survival at 3
     years, which was consistent with the findings of another study
    published in [3]. However, the study published in [2] found no
    significant difference in disease-free survival between NSRH and
     SRH at 2 years. The results of these studies suggest that NSRH
    may be associated with a higher likelihood of disease-free
    survival compared to SRH, but the evidence is not uniform and
    the results may vary depending on the specific population and
    follow-up period.

**Conclusion**: The available evidence suggests that NSRH may be
    associated with a higher rate of disease-free survival compared
    to SRH, particularly at longer follow-up periods. However, the
    evidence is not uniform, and the results may vary depending on
    the specific population and follow-up period. Further research
    is needed to confirm these findings and to determine the optimal
     surgical approach for patients with cervical cancer.

**Additional Context**: The study published in [2] found that NSRH
    was associated with improved urinary function and reduced
    symptoms of urinary incontinence compared to SRH. This suggests
    that NSRH may be a more effective surgical approach for patients
     with cervical cancer, particularly those who are at risk of
    developing urinary incontinence.
```

Remarks: The output from Llama 3.0 shows that it is relying on or biased by internal memory in answering the questions on this dataset; we note that this question involves only one source, so there should not be multiple sources being cited.

T.2.4 LLAMA 4 SCOUT

Question ID 8 - 2/24 models correct

Review PMID: 28898386
Question: Is the length of hospital stay higher, lower, or the same when comparing Pancreatojejunostomy (PJ) to Pancreatogastrostomy (PG)?
Correct Answer: `no difference`
PMIDs of Relevant Sources: 16327486, 19092337, 7574936

Llama 4 Scout Response:

```
To answer the question of whether the length of hospital stay is
    higher, lower, or the same when comparing Pancreatojejunostomy (
    PJ) to Pancreatogastrostomy (PG), let's analyze the provided
    article summaries.

- **Article ID 16327486**: This study found that the mean
    postoperative hospitalization period stay was comparable in both
     groups, indicating no significant difference in hospital stay
    length between PJ and PG.
```

```
- **Article ID 19092337**: This study reported a mean hospital stay
    of 12 +/- 2 days after PPPD-GP (PG) and 16 +/- 3 days after
   PPPD-PJ (PJ), suggesting that the hospital stay was lower for PG
    compared to PJ.

- **Article ID 7574936**: This study noted that pancreatic fistula
   was associated with a significant prolongation of postoperative
   hospital stay but did not find a significant difference in the
   incidence of pancreatic fistula between the PG and PJ groups.
   However, it did not provide a direct comparison of mean hospital
    stays between the two groups.

Given these findings:

- **Rationale**: The rationale is based on the comparison of
   hospital stay lengths between PJ and PG across the provided
   studies. While one study found no difference, another study
   clearly indicated a shorter hospital stay for PG.

- **Full Answer**:
  - [16327486] found no difference in hospital stay.
  - [19092337] found a lower hospital stay for PG (12 +/- 2 days)
     compared to PJ (16 +/- 3 days).
  - [7574936] did not provide a direct comparison but discussed
     complications.

- **Answer**: Lower.

The conclusion that the length of hospital stay is lower for PG
   compared to PJ is primarily supported by [19092337], which
   directly compared hospital stays and found a significant
   difference. The other studies either found no difference or did
   not directly compare hospital stay lengths as primary outcomes.
   Therefore, based on the information provided:

The final classification is: **Lower**.
```

Remarks: As with previous models, unlike the SR authors that perform a full meta-analysis and explicitly pool together the data from sources, Llama 4 scout fails to reconcile conflicting evidence, in this case seeming to just pick one source it trusts the most.

Question ID 219 - 11/24 models correct

Review PMID: 27056645
Question: Is the risk of adverse events higher, lower, or the same when comparing chromoscopy to conventional endoscopy?
Correct Answer: `insufficient data`
PMIDs of Relevant Sources: 12196768, 14960519, 20179689, 16527699, 21159889, 16767577, 19139000

Llama 4 Scout Response:

```
To answer the question of whether the risk of adverse events is
   higher, lower, or the same when comparing chromoscopy to
   conventional endoscopy, let's analyze the provided article
   summaries.

## Step 1: Identify Relevant Information
The articles provided compare chromoscopy (or high-definition
   chromocolonoscopy) with conventional endoscopy in terms of
```

```
    detection rates of adenomas and other lesions. However, to
    assess the risk of adverse events, we need to focus on
    complications, side effects, or any negative outcomes reported.

## 2: Analyze Each Article for Adverse Events
- Article ID 12196768: "There were no complications."
- Article ID 14960519: No specific mention of adverse events.
- Article ID 20179689: No significant differences in adverse events;
     the study focuses on detection rates.
- Article ID 16527699: No specific mention of adverse events.
- Article ID 21159889: Mean extubation times were slightly but
    significantly longer in the PCC group, but no specific adverse
    events were mentioned.
- Article ID 16767577: The study mentions a significant difference
    in examination duration but does not specify adverse events.
- Article ID 19139000: Chromoendoscopy was associated with more
    normal tissue biopsies and longer procedure times but does not
    specify significant adverse events.

## 3: Determine the Risk of Adverse Events
From the provided summaries, only Article ID 12196768 explicitly
    states that "There were no complications," directly addressing
    adverse events. Other articles may imply potential for increased
     risk due to longer procedure times or more biopsies but do not
    explicitly report a higher incidence of adverse events.

## 4: Conclusion
Given the information, most studies do not report a significant
    increase in adverse events with chromoscopy compared to
    conventional endoscopy. The explicit mention of "no
    complications" in one study and the lack of detailed adverse
    event reporting in others suggest that there is not a clearly
    documented higher risk of adverse events with chromoscopy.

- **Rationale**: The provided article summaries do not detail a
    significant increase in adverse events when comparing
    chromoscopy to conventional endoscopy. Only one study explicitly
     mentions no complications, and others focus on detection rates
    without highlighting adverse events.
- **Full Answer**: The risk of adverse events is not significantly
    higher with chromoscopy compared to conventional endoscopy based
     on the provided summaries [12196768]. Some studies note longer
    procedure times or more biopsies, which could potentially
    increase the risk of adverse events, but significant differences
     are not commonly reported.
- **Answer**: No Difference.
```

Remarks: The original SR authors recognize that no studies report on adverse events. Reading the logs closely, Llama 4 seems to also recognize this fact, but then concludes that this means there should be no significant difference in risk, highlighting a difference in how models (fail to) assess potential risks relative to human clinicans.

T.2.5 OPENBIOLLM-8B

Question ID 13 - 22/24 models correct

Review PMID: 29489032
Question: Is food intake higher, lower, or the same when comparing grehlin to placebo?
Correct Answer: `higher`
PMIDs of Relevant Sources: 15181065

OpenBioLLM-8B Response:

```
**Rationale**: The study found that ghrelin significantly increased
    energy intake in cancer patients with anorexia, as evidenced by
    a 31% increase in energy intake during ghrelin infusion
   compared to saline control, with all patients consuming more
   food. Additionally, the meal appreciation score was higher with
   ghrelin treatment. These results suggest that ghrelin could be
   an effective treatment for cancer anorexia and potentially other
    conditions with appetite loss. **Full Answer**: The food intake
    is higher in ghrelin than in placebo, as demonstrated by a 31%
   increase in energy intake during ghrelin infusion compared to
   saline control, with all patients eating more and a higher meal
   appreciation score. [15181065]
```

Remarks: Some of OpenBioLLM's answers are correct when you read through them manually, but it clearly struggles to follow the required format, leading to answers marked incorrect.

Question ID 3 - 11/24 models correct

Review PMID: 25734590
Question: Is the rate of patient self-reported impression of memory change higher, lower, or the same when comparing cholinase inhibitors to placebo in patients with MS?
Correct Answer: `uncertain effect`
PMIDs of Relevant Sources: 15534239, 21519001

OpenBioLLM-8B Response:

```
The rate of patient self-reported impression of memory change is
   higher when comparing cholinase inhibitors to placebo in
   patients with MS.
```

Remarks: Even ignoring formatting errors, though, OpenBioLLM-8B sometimes comes to incorrect conclusions with little provided rationale.

U INDIVIDUAL CONFUSION MATRICES FOR ALL MODELS

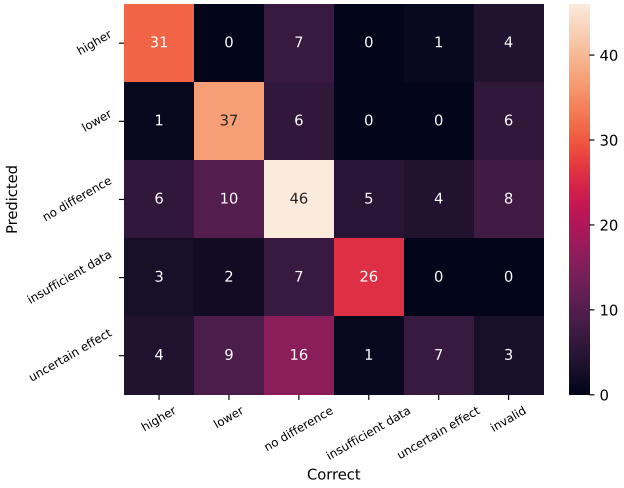

Figure 24: Confusion matrix for DeepSeek R1.

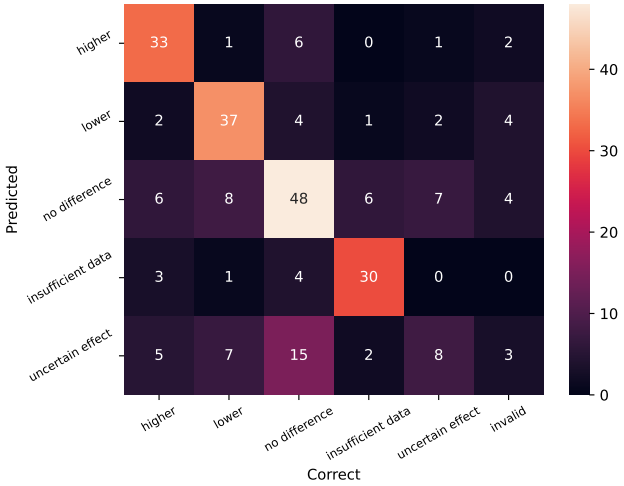

Figure 25: Confusion matrix for DeepSeek V3.

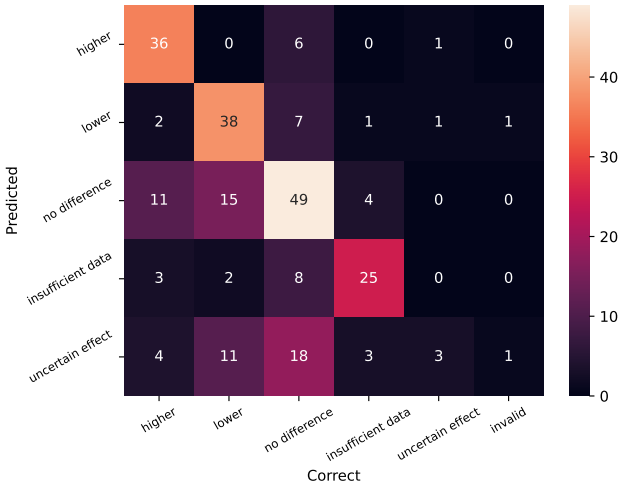

Figure 26: Confusion matrix for GPT-4.1.

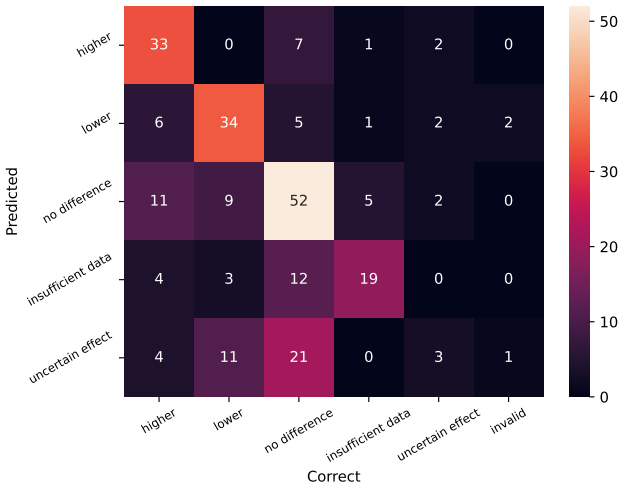

Figure 27: Confusion matrix for GPT-4.1 mini.

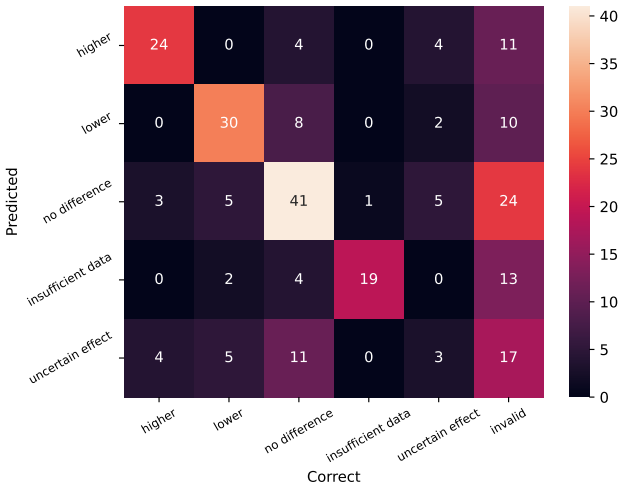

Figure 28: Confusion matrix for GPT-o1.

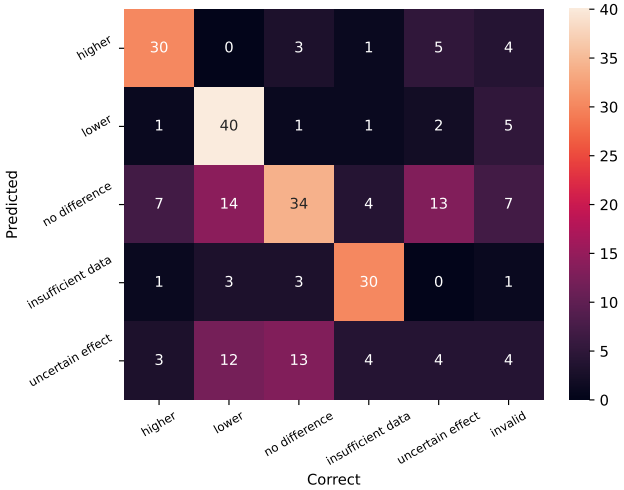

Figure 29: Confusion matrix for GPT-oss-120B.

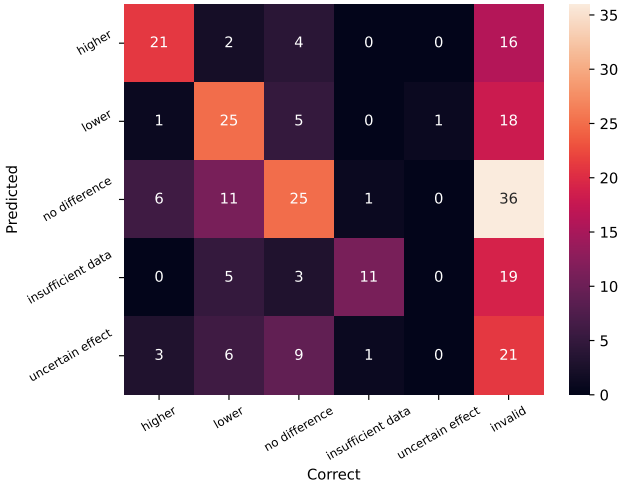

Figure 30: Confusion matrix for HuatuoGPT-o1-70B.

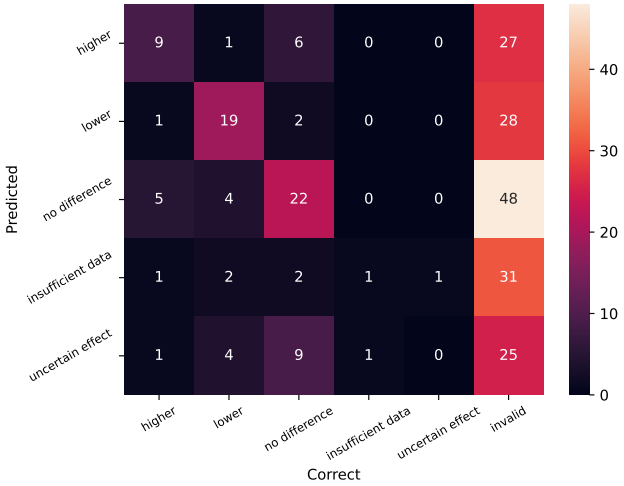

Figure 31: Confusion matrix for HuatuoGPT-o1-7B.

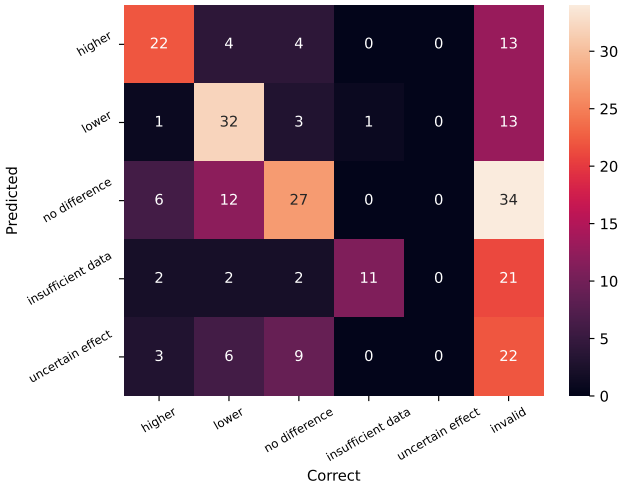

Figure 32: Confusion matrix for Llama 3.0 70B.

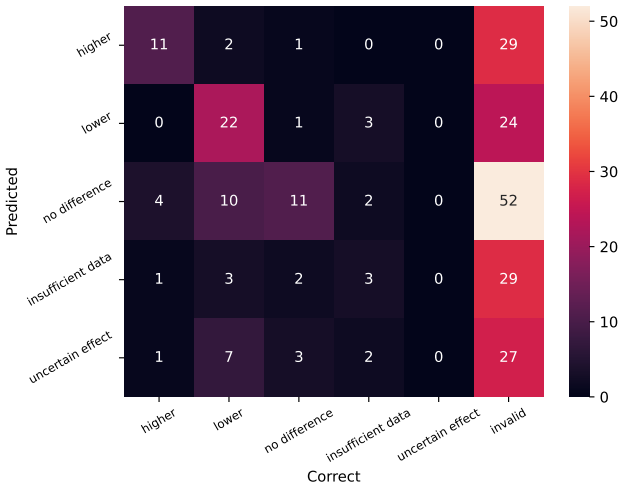

Figure 33: Confusion matrix for Llama 3.0 8B.

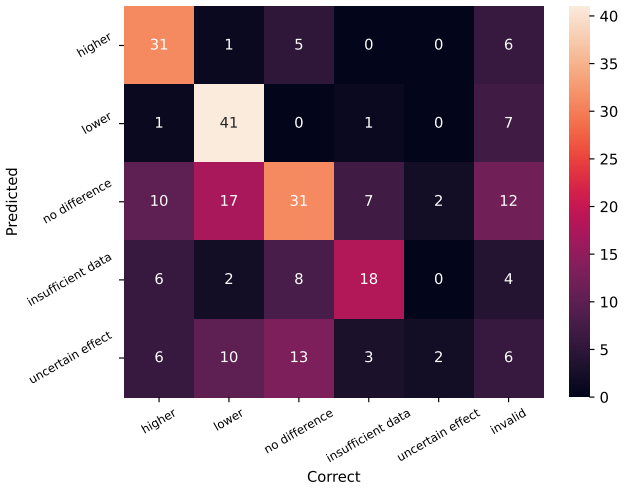

Figure 34: Confusion matrix for Llama 3.1 405B.

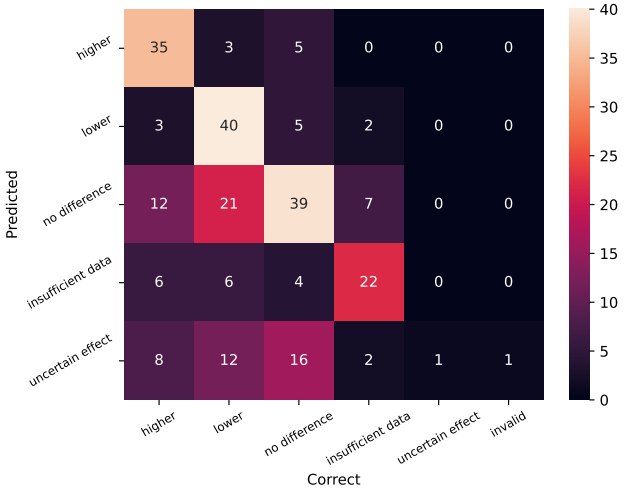

Figure 35: Confusion matrix for Llama 3.1 70B.

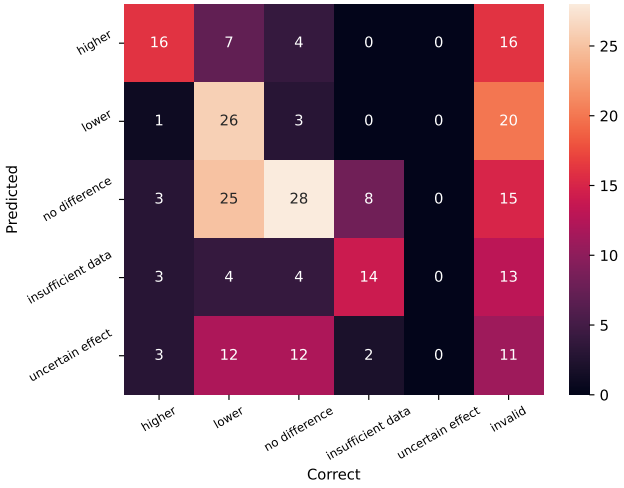

Figure 36: Confusion matrix for Llama 3.1 8B.

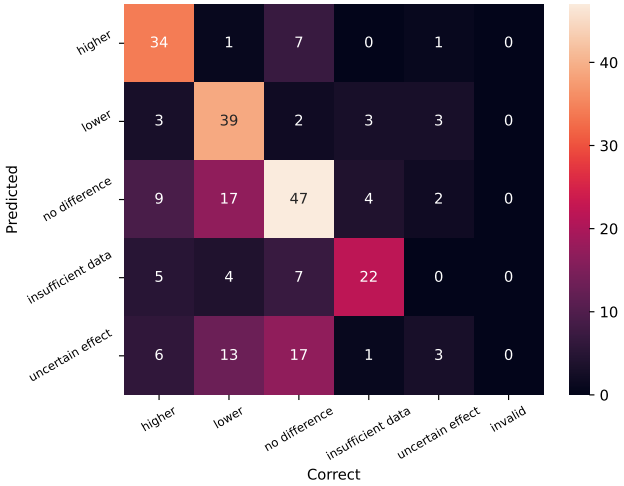

Figure 37: Confusion matrix for Llama 3.3 70B (R1-Distill).

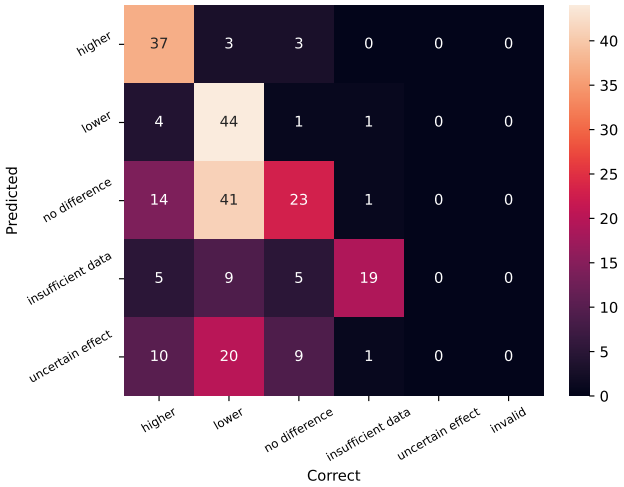

Figure 38: Confusion matrix for Llama 3.3 70B-Instruct.

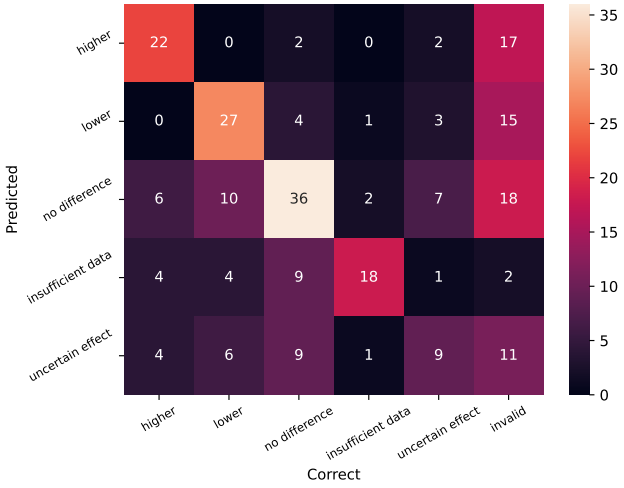

Figure 39: Confusion matrix for Llama 4 Maverick.

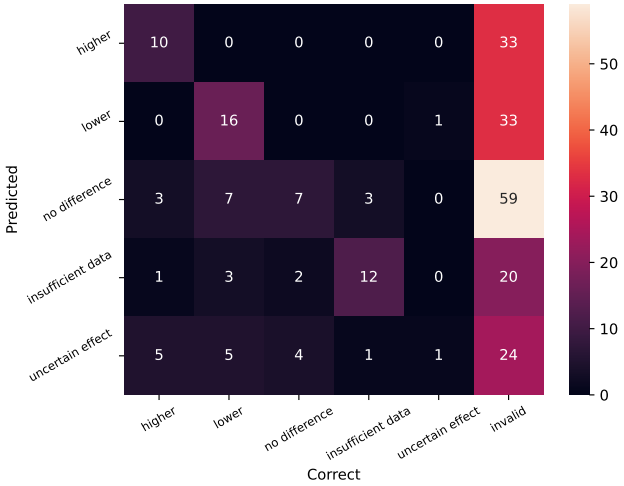

Figure 40: Confusion matrix for Llama 4 Scout.

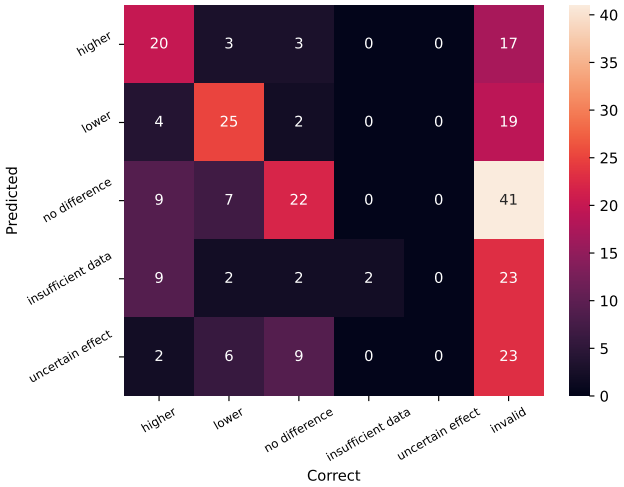

Figure 41: Confusion matrix for OpenBioLLM 70B.

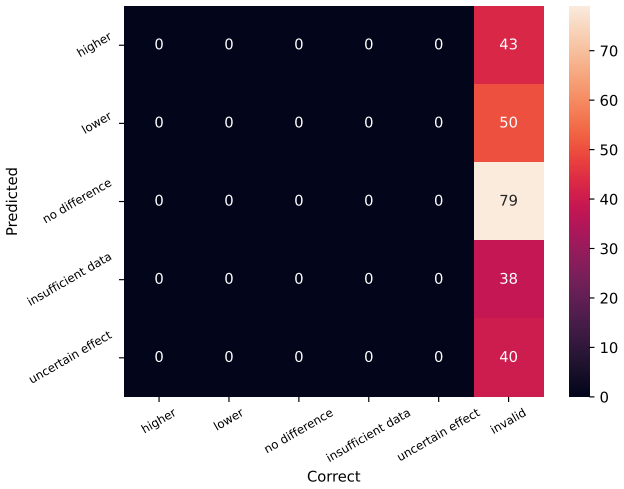

Figure 42: Confusion matrix for OpenBioLLM 8B.

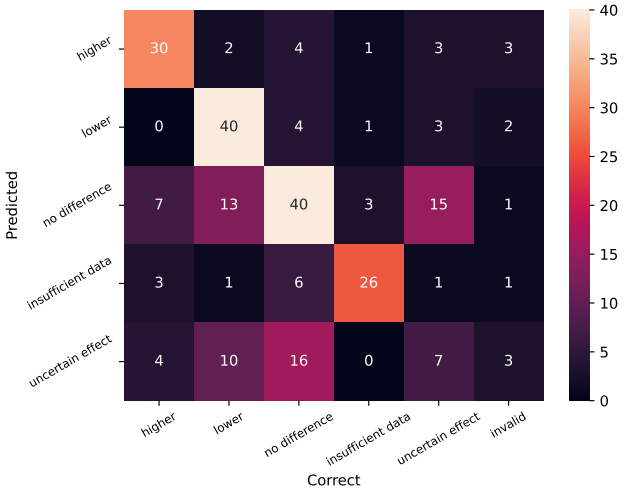

Figure 43: Confusion matrix for OpenThinker2-32B.

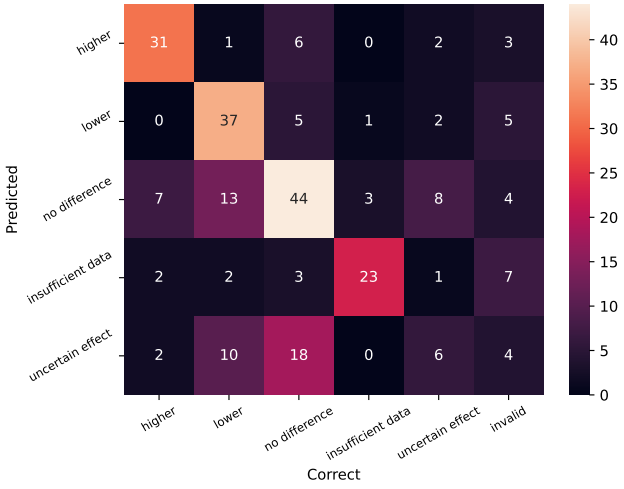

Figure 44: Confusion matrix for QwQ-32B.

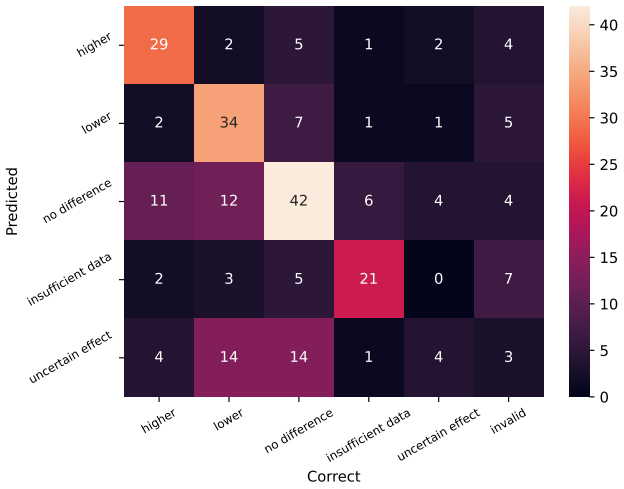

Figure 45: Confusion matrix for Qwen2.5-32B-Instruct.

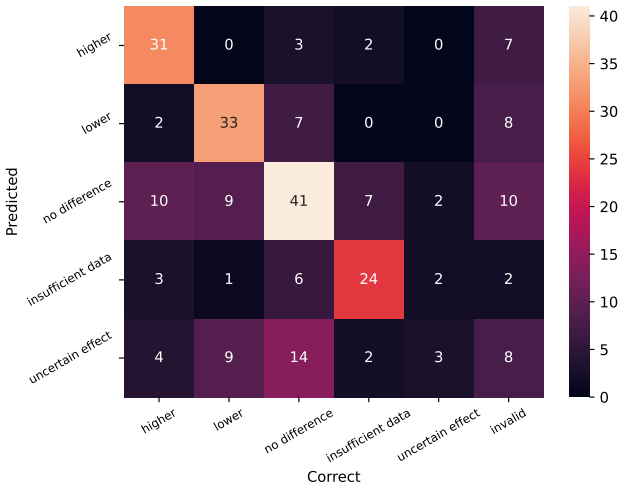

Figure 46: Confusion matrix for Qwen2.5-72B-Instruct.

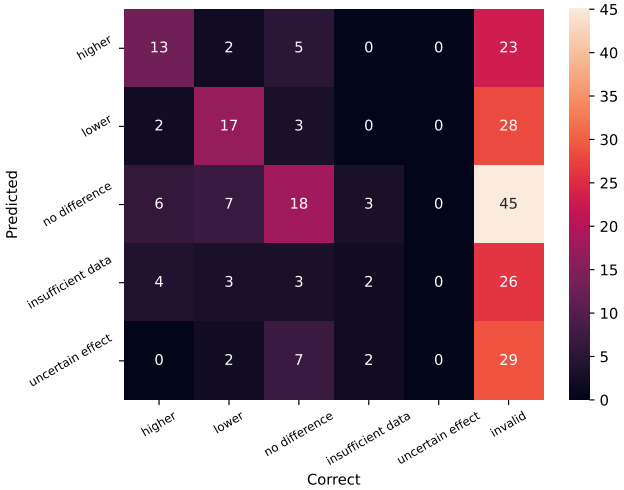

Figure 47: Confusion matrix for Qwen2.5-7B-Instruct.

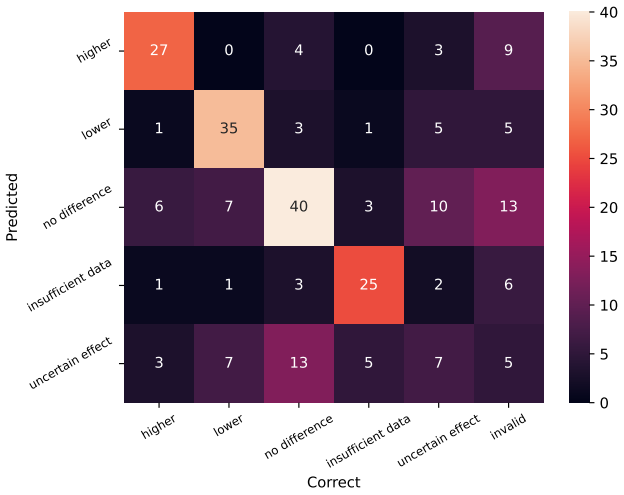

Figure 48: Confusion matrix for Qwen3-235B-A22B-FP8.