SciGuide: Evaluating Literature Understanding, Inductive Reasoning, and Knowledge Utilization in Scientific Research

Anonymous ACL submission

Abstract

We introduce **SciGuide**, a large language models (LLMs) benchmark for scientific research scenarios, designed to evaluate model performance in evidence-based clinical guideline development. Compared with existing benchmarks, SciGuide provides three key advances: (1) **Scan-oriented scientific literature understanding.** We introduce two novel tasks without explicit retrieval targets, requiring comprehensive document scanning. PICO extraction and quality appraisal tasks that require models to capture detailed PICO elements (17.03 PICOs and 451.75 factors per study on average) and methodological features (12.39); (2) **Inductive reasoning under uncertainty.** Grounded in the GRADE framework, models are required to synthesize multiple studies (3.04 on average, up to 13) under varying or conflicting evidence quality; (3) **Priors–driven knowledge utilization.** Models are required to rely on prior knowledge to complete expert-level scientific research tasks (7 task settings). We further conduct quantitative experiments to analyze the impact of prior knowledge and reasoning ability. We evaluate 18 LLMs. The best-performing model achieves only 37.64. We expect SciGuide to facilitate the application and improvement of LLMs in real-world scientific research.Data and code are available [1].

1 Introduction

Recently, large language models (LLMs) have demonstrated advanced logical reasoning and end-to-end problem-solving capabilities (Jaech et al., 2024; Liu et al., 2024). These advancements have inspired the application of Artificial Intelligence (AI) in scientific innovation. Several benchmarks have been proposed to evaluate LLMs in scientific research. LitQA2 (Skarlinski et al., 2024), QASPER (Dasigi et al., 2021), and SciCUEval (Yu et al., 2025) focus on extracting key information from literature. SciReviewGen (Kasanishi et al., 2023) and SurveySum (Fernandes et al., 2024) concentrate on cross-literature content summarization, while others (Lu et al., 2024) attempt to cover multiple stages of research innovation.

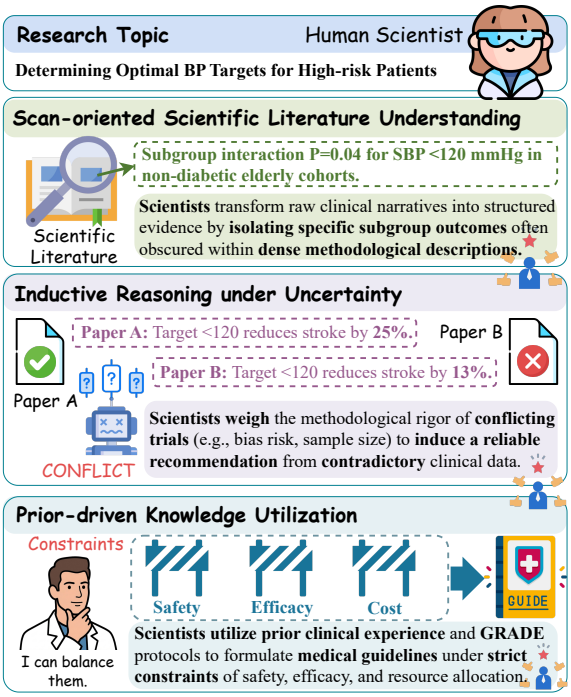

Figure 1: Shows how scientists extract data through scan-oriented understanding, weigh research conflicts using inductive reasoning, and develop guidelines for optimal blood pressure targets with prior knowledge.

Despite this progress, existing benchmarks fail to capture the nuanced demands of specific scientific scenarios. As illustrated in Figure 1, domain experts conduct systematic literature investigations guided by specific research needs and professional experience, synthesizing and assessing evidence to underpin subsequent research decisions. Compared to this complex process, existing benchmarks exhibit clear limitations:

[1] https://anonymous.4open.science/r/SciGuide-628/

Benchmark	Task & Setting (TS)				S.U.			I.R.		K.U.	
	Dm.	Task	Format	Size	Und.	Prt.	Apr.	Mlt.	Cnf.	Exp.	Modeling
LitQA2	Gen.	Retrieval	Single-doc MCQ	248	Search	✘	✘	✘	✘	✔	Sub-task
QASPER	Gen.	Extraction	Single-doc QA	5k	Search	✘	✘	✘	✘	✘	Heuristic
MMCR	Gen.	Reasoning	Multi-doc QA	276	Search	✘	✘	✔	✘	✘	Sub-task
SurveySum	Gen.	Summarize	Multi-doc Sum.	79	Search	✘	✘	✔	✘	✘	Sub-task
PubMedQA	Med.	Reasoning	Abstract MCQ	1k	Search	✘	✘	✘	✘	✘	Heuristic
CGBench	Clin.	Reasoning	Single-doc QA	2k	Search	✔	✔	✘	✘	✔	Sub-task
Quicker	Clin.	Guideline	Multi-doc Rec.	85	Search	✔	✔	✔	✔	✔	Unified
SciGuide (ours)	**Guid.**	**Guideline**	**Multi-doc Rec.**	**1331**	**Scan**	✔	✔	✔	✔	✔	**Unified**

Table 1: **Comparison of SciGuide and other benchmarks. TS**: **Dm.**: Domain (Gen.: General, Med.: Medical, Clin.: Clinical, Guid.: Guideline); **Format**: Sum.: Summarize, Rec.: Recommendation; **S.U.**(Scientific Understanding): **Und.**: Depth (Search vs. Scan); **Prt.**: Protocol; **Apr.**: Quality Appraisal. **I.R.**(Inductive Reasoning): **Mlt.**: Multi-evidence; **Cnf.**: Conflict handling. **K.U.**(Knowledge Utilization): **Exp.**: Expert-level.

- **Search-oriented literature understanding.** Scientific research is grounded in a deep understanding of existing literature. Existing works (Adams et al., 2025; Wadden et al., 2020; DeYoung et al., 2020) simplify the task into information retrieval around preset questions, where locating a small number of answer-bearing spans is often sufficient for success.
- **Single-answer evidence aggregation.** Scientific decision-making rarely has a single optimal answer, requiring synthesis of incomplete and uncertain evidence across studies to reach a defensible judgment. Current works (Bao et al., 2025; Lu et al., 2020) are formulated under assumptions of evidence sufficiency and consistency.
- **Closed knowledge QA.** LLMs are expected to effectively leverage their capabilities in scientific research practice. Existing benchmarks (Singhal et al., 2025; Rein et al., 2024; Wang et al., 2024; Jin et al., 2019) mainly assess knowledge mastery via multiple-choice questions, where constrained answer spaces reduce reasoning requirements and allow directional guessing.

However, real-world scientific research demands scan-oriented literature understanding, inductive reasoning under uncertainty, and prior-driven knowledge utilization. For instance, the development of evidence-based clinical guidelines requires researchers to perform evidence extraction and quality appraisal from the literature, synthesizing recommendations from evidence of varying quality or even conflicting findings. To fill this gap, we propose **SciGuide** to systematically identify bottlenecks in LLMs for scientific research. We offer three key advantages:

- **Scan-oriented Scientific Literature Understanding.** Bench introduces two novel tasks tailored for medical research. These two tasks operate without predefined retrieval targets and use recall-dominant evaluation to enforce protocol-specified evidence and methodological coverage. The PICO Extraction task requires models to understand factors such as research design and effect sizes (17.03 PICOs and 451.75 factors per study on average). The Quality Appraisal task (12.39 features) requires models to assess study quality as reviewers do.
- **Inductive Reasoning under Uncertainty.** Models must synthesize evidence from multiple studies (3.04 on average, up to 13) to formulate recommendations and assess confidence in both the evidence and recommendations.
- **Prior-driven Knowledge Utilization.** Models are required to complete expert-level research tasks (seven settings) using prior knowledge, and we quantitatively analyze the effects of prior knowledge and reasoning ability on problem-solving performance.

We evaluated 18 proprietary and open-source LLMs, with the primary findings as follows:

- **LLMs lack expert-level scientific capability.** No model achieved a total score exceeding 40; the SoTA model, Claude 4.5 sonnet, reached only 37.64. Proprietary models did not demonstrate a consistent advantage, whereas reasoning-enhanced models showed greater potential.

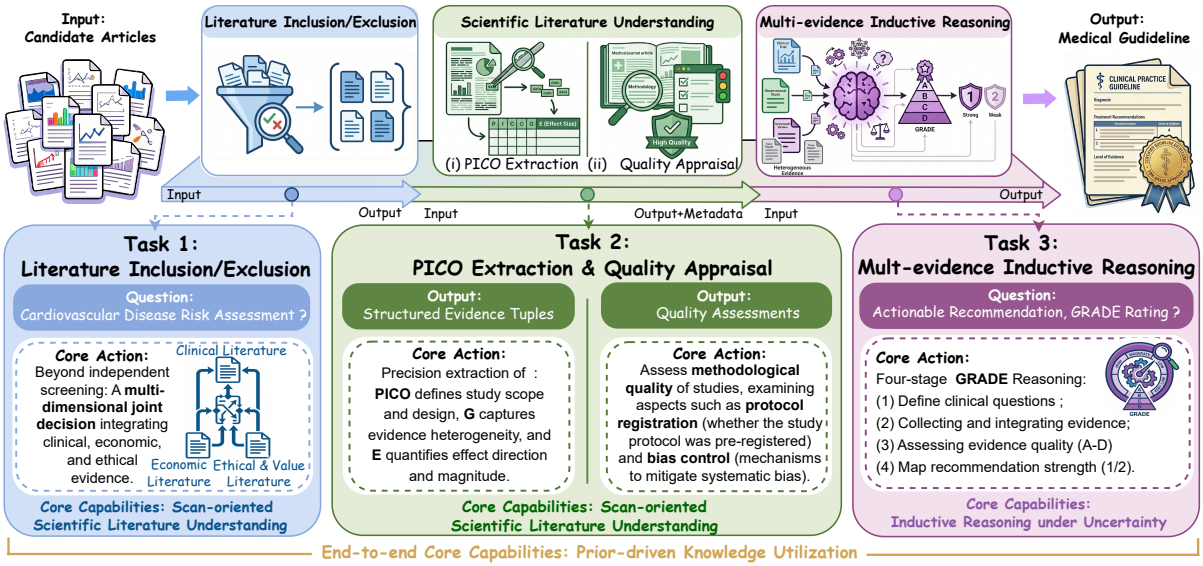

Figure 2: Integrating multi-dimensional literature screening, PICO-based evidence extraction with quality assessment, and GRADE-based inductive reasoning to transform raw articles into actionable clinical recommendations.

- **Models struggle to provide fine-grained recommendations.** Unlike professional guideline developers, LLMs cannot effectively utilize document information for inductive reasoning to provide granular recommendations with critical thresholds and applicable conditions.
- **Prior knowledge and reasoning capabilities jointly constrain performance.** Similar to human professional experience, a rich prior knowledge base is essential for models to complete scientific research, while inductive reasoning capability sets the performance ceiling.

2 Preliminaries

In this section, we introduce several key concepts from the field of evidence-based medicine, including PICO, Effect size, Quality appraisal, Evidence synthesis, and GRADE.

PICO is used to structure clinical questions and biomedical literature, helping define study scope and interpret study design. It represents a question or a study with four elements: Population, Intervention, Comparison, and Outcome.

Effect size quantifies the strength of the association between an intervention and an outcome, characterizing both the direction and magnitude of the effect.

Quality appraisal assesses the reliability of scientific studies by identifying the strengths and weaknesses in their methodologies, such as bias control, protocol adherence, and sample size.

Evidence synthesis refers to the systematic integration of evidence from multiple scientific studies. It emphasizes the joint consideration of study findings, evidence quality, and cross-study consistency.

GRADE is a standardized framework for assessing the certainty of synthesized evidence and supporting recommendation development. It maps evidence appraisal outcomes to clinical recommendation grades, reporting the final rating as a combination of an evidence level (A/B/C/D) and a recommendation strength (1/2).

3 Benchmark

We propose SciGuide, a benchmark for medical guideline development, to evaluate LLMs in realistic scientific research settings. The task design systematically examines three core capabilities: (1) **Scan-oriented Scientific Literature Understanding**, (2) **Inductive Reasoning under uncertainty**, and (3) **Prior-driven Knowledge Utilization**.

3.1 Benchmark Tasks

Literature Inclusion/Exclusion (LIE). This task requires the model to screen a set of candidate articles. Prior work (Li et al., 2025; Wang et al., 2025) implicitly assumes independence across articles and formulates screening as independent decisions over individual articles. However, real-world guideline development requires jointly considering relationships among articles and their roles in recommendation formation (e.g., economic affordability and clinical feasibility). We formulate this

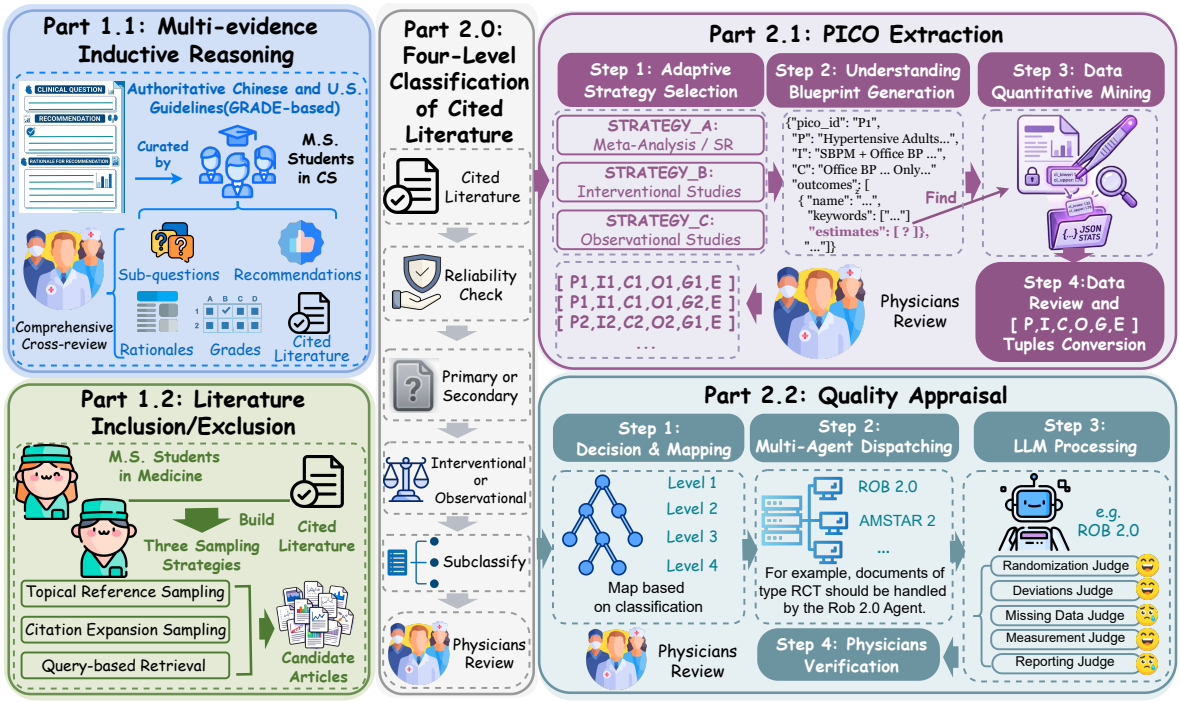

Figure 3: **Construction Pipeline of SciGuide.** (1) Guideline structuring and candidate article construction from Cited Literature;(2) literature understanding via classification, PICO extraction, and quality appraisal.

task as a set-level decision problem. Let the candidate set be $D = \{d_1, \ldots, d_n\}$ and the question be q. The task is defined as

$$\mathrm{LIE}(D, q) = \hat{\mathbf{y}}, \qquad \hat{\mathbf{y}} \in \{0, 1\}^{|D|}. \tag{1}$$

$\hat{\mathbf{y}}$ denotes the decision vector over the entire candidate set. Performance is evaluated using the F1 score. This subset includes 211 questions, with an average of 6.99 candidate documents per question (range: 3–15).

Scientific Literature Understanding (SLU). (i) PICO Extraction (PE). A single medical article often reports multiple findings, whose defining elements are distributed across different document components such as narrative text, tables, and forest plots. Let the article be denoted as d, and the PICO extraction task is defined as

$$\mathrm{PE}(d) = \mathcal{Q}_{\mathrm{pico}}. \tag{2}$$

Each extracted finding is represented as a tuple (P, I, C, O, G, E), where G denotes the corresponding subgroup and E represents the effect size. Performance is evaluated using tuple-level recall. This subset includes 244 questions. Each question requires the extraction of 17.03 PICO elements and 451.75 factors on average. **(ii) Quality Appraisal (QA).** Prior work (Abaho et al., 2019; Nye et al., 2020) commonly assumes that evidence is equally credible and does not explicitly model its reliability. The QA task requires models to identify key methodological characteristics of the study design and their potential limitations, as assessed by reviewers. This task is defined as

$$\mathrm{QA}(d) = \mathcal{Q}_{\mathrm{quality}}. \tag{3}$$

Performance is evaluated using weighted recall, assigning higher weights to the identification of study limitations and critical items. This subset includes 237 questions. Each question requires the identification of 12.39 features on average.

Multi-evidence Inductive Reasoning (MEI). This task requires the model to perform inductive reasoning over a collection of scientific articles and generate recommendations. The model is provided with an SLU representation for each article, defined by the corresponding $\mathcal{Q}^{(i)}_{\mathrm{pico}}$ and $\mathcal{Q}^{(i)}_{\mathrm{quality}}$:

$$\mathcal{Q}^{(i)}_{\mathrm{SLU}} = \left(\mathcal{Q}^{(i)}_{\mathrm{pico}}, \mathcal{Q}^{(i)}_{\mathrm{quality}}, m^{(i)}\right). \tag{4}$$

$m^{(i)}$ denotes the metadata(e.g., title, abstract, and study type) associated with the i-th article.Based on this representation, the model performs inductive reasoning. The MEI task is defined as:

$$\mathrm{MEI}\left(\{\mathcal{Q}^{(i)}_{\mathrm{SLU}}\}_{i=1}^{|D|}, \mathcal{P}_{\mathrm{GRADE}}\right) = (\hat{r}, \hat{g}), \tag{5}$$

Property	Value
# Total Instances	1,331
# Task Types	4 (LIE/PE/QA/MEI)
Tasks	
# LIE candidates ($\leq$5 / 6–10 / $\geq$11)	68/129/14
# PE PICOs ($\leq$10 / 11–20 / $\geq$21)	91/85/68
# QA study type (SR/C/ACS/RCT)	108/49/25/55
# MEI docs ($\leq$3 / 4–6 / $\geq$7)	111/35/22

Table 2: Dataset statistics of the benchmark.

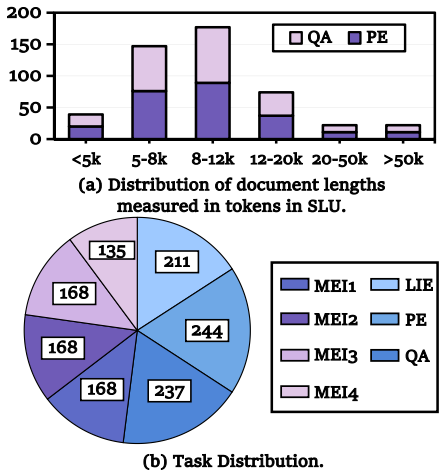

Figure 4: Distribution of the benchmark.

where $\mathcal{P}_{\text{GRADE}}$ denotes a precise GRADE protocol, including rule specifications and stage-wise constraints (covering four procedural stages such as clinical question specification and evidence synthesis). $\hat{r}$ denotes the recommendation formulated for medical guideline development, and $\hat{g}$ denotes the GRADE rating, whose output space is restricted to $\{1, 2\} \times \{A, B, C, D\}$. We further design four additional settings to analyze the model's reasoning ability under different information conditions. We evaluate $\hat{r}$ using an LLM-as-judge binary assessment and evaluate $\hat{g}$ using accuracy. This subset includes 639 questions.

3.2 Benchmark Construction

The distribution of the benchmark is shown in Table 2 and Figure 4.

Part 1: Guideline-based construction. We construct the benchmark from authoritative Chinese and U.S. guidelines (Organization et al., 2021; Chinese Society of Cardiology et al., 2024) developed under the GRADE methodology. The MEI portion is curated by three M.S. students in computer science based on the original guideline content. The LIE portion is curated by seven M.S. students in medicine, who build candidate pools using three sampling strategies: Topical Reference Sampling, Citation Expansion Sampling, and Query-based Retrieval. The medical M.S. team then conducts cross-review of the Part 1 annotations. Articles are collected from publicly accessible sources like PubMed, excluding those with copyright restrictions that prevent key evidence coverage.

Part 2: Human–AI collaborative annotation. Both PE and QA first undergo decision-tree–based literature classification (Part 2.0) to assign type-specific processing strategies. PE proceeds with coarse-grained PICO construction, followed by fine-grained extraction of subgroups and effect sizes, while QA applies methodology-aligned prompts to perform itemized yes/no annotations. Claude Sonnet 4 serves as the backbone model for generating initial automatic annotations.

Quality control. To ensure high quality, we implement a strict quality control protocol. Given the domain-specific nature of the benchmark, we focus on hypertension treatment and management; all participating physicians (with over 10 years of clinical experience and prior involvement in guideline writing) and medical M.S. annotators are from cardiovascular-related research areas. In Part 2, two physicians independently review each item; disagreements are adjudicated by a third physician for final decision.

4 Experiments

4.1 Experiment Setting

Models. We evaluate 18 large language models (greedy decoding), including 6 proprietary (OpenAI, 2025; Gemini Team, Google, 2025; Anthropic, 2025) and 12 open-source models (Agarwal et al., 2025; Yang et al., 2025; Liu et al., 2025).

Metrics. For LIE, we use Macro-F1, jointly reflecting inclusion (I-F1) and exclusion (E-F1) performance. For SLU, both tasks are evaluated using recall. For MEI, EM-Rec (exact match) and LM-Rec (directional match) evaluate recommendation correctness. GRADE measures grading accuracy conditional on a correct recommendation, with scores set to zero otherwise. For MEI1, GRADE is evaluated conditioned on LM-Rec.

MEI. We evaluate three MEI configurations. MEI1 inputs multiple structured documents and outputs R (recommendations) and G (GRADE).

Model	LIE			SLU		MEI						Avg.
						MEI1			MEI2	MEI3		
	M-F1†	I-F1	E-F1	PE†	RA†	EM-Rec†	LM-Rec	GRADE	GRADE†	EM-Rec†	LM-Rec	
Proprietary LLMs												
Claude 4.5 Sonnet	64.17	72.73	55.61	35.86	56.16	23.21	86.90	27.98	24.40	22.02	86.90	**37.64**
GPT-5.2	62.03	69.50	54.56	39.04	60.68	10.12	91.07	29.17	32.14	8.33	89.29	**35.39**
Gemini 3 Flash Preview	65.53	74.06	56.99	15.19	42.19	24.40	89.29	24.40	25.60	14.88	75.60	**31.30**
Gemini 3 Pro Preview	64.93	72.31	57.56	19.38	37.89	20.24	86.90	17.26	20.83	10.12	87.50	**28.90**
Doubao Seed 1.6	49.32	49.22	49.42	15.58	42.95	21.43	88.69	18.45	24.40	19.64	84.52	**28.89**
OpenAI o4-mini-high	49.59	49.21	49.97	23.01	42.12	14.88	85.71	23.21	30.95	12.50	88.69	**28.84**
Open-source LLMs												
DeepSeek-V3.2(reasoner)	60.19	65.59	54.79	30.58	44.14	22.02	89.29	28.57	30.95	13.69	86.90	**33.60**
gpt-oss-120b	53.35	55.63	51.07	19.75	54.89	26.19	89.88	22.62	23.21	18.45	87.50	**32.64**
Qwen3-Next-80B-A3B-Thinking	59.35	66.24	52.45	26.69	39.34	25.60	85.71	17.26	20.83	16.07	85.71	**31.31**
Qwen3-32B-Thinking	58.36	64.12	52.60	12.72	41.14	23.21	82.74	17.86	26.19	23.21	83.93	**30.81**
Qwen3-30B-A3B-Instruct	55.35	59.65	51.05	16.64	42.03	27.38	84.52	16.67	16.67	20.83	80.95	**29.82**
Qwen3-14B(Thinking)	54.34	57.43	51.24	13.04	36.41	28.57	81.55	17.26	25.60	20.83	87.50	**29.80**
Qwen3-VL-32B-Instruct	57.65	62.79	52.51	14.50	38.52	22.62	70.24	19.05	24.40	20.83	83.30	**29.75**
Qwen3-Next-80B-A3B-Instruct	56.23	60.86	51.59	27.28	45.65	15.48	59.52	14.29	13.69	18.64	82.74	**29.50**
DeepSeek-V3.2(chat)	60.40	66.23	54.57	19.40	45.61	11.90	83.33	23.81	21.43	12.50	88.10	**28.54**
Qwen3-14B(Instruct)	55.60	60.00	51.20	12.77	33.97	25.00	83.33	24.40	23.81	18.45	82.14	**28.27**
Qwen3-30B-A3B-Thinking	43.23	38.91	47.54	17.07	37.94	26.19	85.71	20.24	18.45	17.26	83.93	**26.69**
Llama 4 Maverick	61.19	69.14	53.24	7.10	22.45	17.26	66.67	16.07	16.07	17.86	81.55	**23.70**

Table 3: **Main results of model performance.** M-F1 denotes macro-F1. The best and second-best results are highlighted with green and blue, respectively. Metrics marked with † denote the primary metric for each task.

MEI2 further conditions on R and predicts G. MEI3 generates R directly from the question without external evidence.

Prompt Strategies. We use unmodified original prompts to fairly evaluate the native capabilities of reasoning-enhanced and instruction-tuned models.

4.2 Main Result

presents the evaluation results across models. Our main findings are summarized as follows:

Overall performance remains unsatisfactory. The SoTA model Claude Sonnet 4.5 fails to exceed an average score of 40. Proprietary models show no clear advantage over open-source alternatives, with Deepseek-v3.2 (reasoner) ranking third.

Reasoning-enhanced models demonstrate advantages. This trend holds across Qwen and DeepSeek, except for Qwen3-30B-A3B, where the non-reasoning variant performs better. Further analysis is provided in the next subsection.

Set-level LIE remains non-trivial. Exclusion decisions are harder than inclusion decisions across models, since exclusion often hinges on nuanced methodological and scope-related criteria that are less explicitly stated than inclusion signals.

Scan-oriented SLU Poses Substantial Challenges Performance varies markedly across systems; GPT-5.2 significantly outperforms Llama-4-Maverick on both PE (39.04 vs. 7.10) and RA (60.68 vs. 22.45). Despite the distinct focuses of PE and RA, model performance is aligned (Spearman $\rho = 0.76$), suggesting that both tasks assess shared capabilities in scientific literature understanding and knowledge utilization.

Models exhibit limited capacity for fine-grained recommendations. Models perform well on LM-Rec but degrade sharply on EM-Rec (82.84 vs. 21.43, on average), demonstrating that current models are not yet capable of synthesizing clinical, economic, and contextual factors through inductive reasoning to produce precise recommendations.

4.3 Fine-grained Analysis

More details are shown in Appendix C. The key findings are as follows:

Statistical analysis. As shown in Figure 5(a), performance on the LIE task gradually decreases as the size of the candidate set increases. Figure 5(b) shows that performance on the PE task exhibits a similar decreasing trend as the number of PICO tuples to be extracted increases. Figure 5(c) indicates that the RA task shows substantial performance differences across study types.

MoE, reasoning and scale. As model parameter scale and architecture vary, instruction-tuned models exhibit relatively stable overall performance, whereas thinking-oriented models show more pronounced performance variation. Under dense architectures, thinking-oriented models overall outperform instruction-tuned models; however, when switching to MoE architectures at comparable parameter scales, their average performance

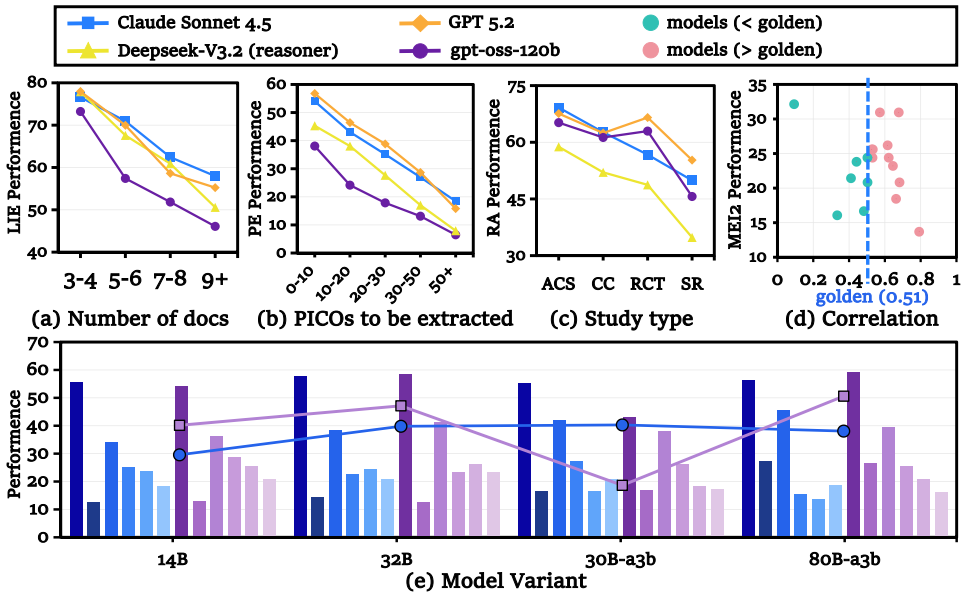

Figure 5: (c) In RA, The x-axis denotes different study types. and (e) **task-wise and averaged performance** across different **Qwen model variants**, where **blue/purple bars** denote **instruct/thinking models**, bars follow the **main-table task order**, and **lines** show averaged performance on a **separate y-axis**.

drops noticeably (-4.12) and falls below that of instruction-tuned models. Across both architectures, reasoning-enhanced models obtain performance gains as parameter scale increases, while instruction-tuned models exhibit more irregular performance changes.

How priors and inductive reasoning facilitate scientific discovery. To quantify the role of priors and inductive reasoning in MEI, we contrast MEI1 with MEI3, which removes external knowledge. Based on EM-Rec (MEI1), we group models into three tiers (using EM-Rec thresholds of 22 and 25, resulting in 6 high-tier, 5 mid-tier, and 7 low-tier models). Comparing MEI1 with MEI3 reveals that, with comparable MEI3 priors (18.65 vs. 18.93 EM-Rec, averaged within each tier), high-tier models achieve larger gains by better leveraging MSD for inductive reasoning (+7.84 vs. +4.17). Low-tier models show weaker MEI3 performance (14.23 EM-Rec) and derive the smallest benefit from MSD (+1.67). Similar to experience-driven human scientific practice, models likewise require sufficient prior knowledge to support hypothesis formation and critical judgments, clarify task execution pathways, and effectively advance the inductive reasoning process.

Insufficient decision-making for recommendation strength. As shown in Figure 6 In case-level analysis, we observe an evidence–recommendation confusion phenomenon, where models implicitly assume that higher evidence quality directly translates into stronger recommendations. We analyze the correlation between predicted evidence quality and recommendation strength in GRADE (MEI2). As shown in Figure 5(d), most models exhibit higher correlation coefficients than that of the golden, indicating an overreliance on evidence quality with insufficient consideration of other decision factors. Once the correlation surpasses the golden value, there is a noticeable decline in performance.

4.4 Error Analysis

We analyze failure cases (MEI1) from GPT-5.2, as it performed well on other tasks but showed low performance in MEI1 (EM-Rec). We categorize the failures into four types: (1) **Ignoring critical details (12%):** The recommendation lacks essential thresholds or operational details, reducing its practical executability. (2) **Conditioning errors (77%):** The recommendation applies incorrect populations or triggering conditions, leading to a mismatch with the intended clinical context. (3) **Decision path compression (1%):** A multi-step or conditional decision process is oversimplified into a single conclusion, omitting intermediate reasoning stages. (4) **Evidence–recommendation confusion (10%):** Evidence quality is simplistically mapped to recommendation strength, neglecting clinical feasibility, risk–benefit trade-offs. More

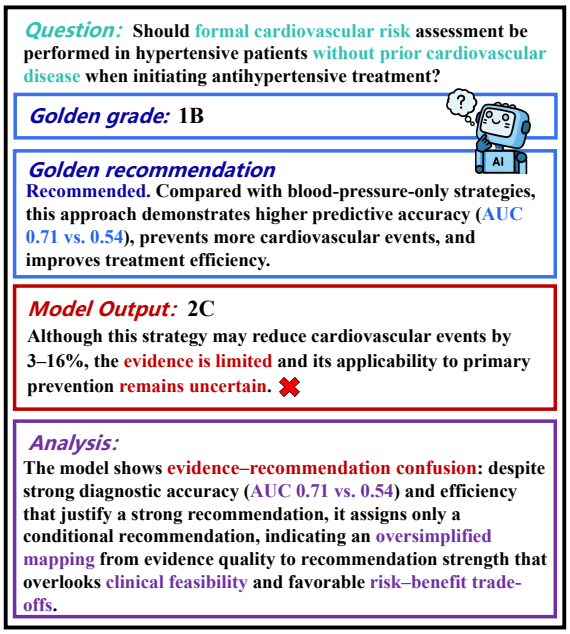

Figure 6: Case study of evidence–recommendation confusion.

error case analyses are provided in Appendix D.

4.5 Evidence Input Analysis

We replace the structured documents in MEI1 with full-text inputs to form MEI4, and evaluate the best-performing models from the main experiments (Claude Sonnet 4.5, GPT-5.2, DeepSeek-V3.2 (reasoner), and gpt-oss-120b), with MEI4 inputs capped at 130k tokens for fair comparison. As shown in Table 4. Replacing structured evidence (MSD) with full-document (MD) in MEI1 leads to consistent performance drops across all models, with larger relative degradation on EM-Rec than on LM-Rec, averaging 19.3% versus 3.7%. Under the MD setting, the EM-Rec of Claude Sonnet 4.5 and gpt-oss-120b even falls below their MEI3 prior performance.

Model	MEI1:R+G (MSD)		MEI4:R+G (MD)		MEI3:R	
	EM	LM	EM	LM	EM	LM
Claude Sonnet 4.5	22.96	85.93	18.52	84.44	22.02	86.90
GPT-5.2	8.15	88.89	7.41	85.93	8.33	89.29
DeepSeek-V3.2 (r)	22.96	87.41	20.74	83.70	13.69	86.90
gpt-oss-120b	28.15	88.15	25.93	81.48	18.45	87.50

Table 4: Comparison of performance under different MEI settings.

5 Related Work

Real scientific research requires models to perform scan-oriented literature understanding, multi-evidence inductive reasoning, and prior-driven knowledge utilization.

However, existing benchmarks fail to faithfully capture this process. Early benchmarks such as MedQA-USMLE (Jin et al., 2021) and MedMCQA (Pal et al., 2022) evaluate models in multiple-choice settings, where the answer space is discrete and directional judgments are often sufficient. Moving beyond multiple-choice formats, ScienceQA (Saikh et al., 2022), MMMU (Yue et al., 2024), MMLU (Hendrycks et al., 2020), and GPQA (Rein et al., 2024) adopt unified question–answering setups to assess general scientific knowledge, but remain largely detached from fine-grained research workflows. Further extending the task scope, PaperQA (Lála et al., 2023), LitQA (Lála et al., 2023), and PubMedQA (Jin et al., 2019) introduce documents as external knowledge sources. However, these benchmarks typically rely on predefined queries, making successful information retrieval from documents sufficient to answer the question. CGBENCH (Queen et al., 2025) leverages expert-annotated clinical genetics literature to evaluate models' ability to interpret and judge evidence. Evidence Inference (DeYoung et al., 2020), QASPER (Dasigi et al., 2021), and MMCR (Tian et al., 2025) introduce cross-document reasoning, but do not explicitly account for differences in evidence quality or conflicts among conclusions.

LLMs are being integrated into scientific research workflows. LLM4SD (Zheng et al., 2025) integrates literature and structured data for molecular property prediction. Med-PaLM (Tu et al., 2024) assists researchers in literature analysis and knowledge organization. TrialMind (Wang et al., 2025) and Quicker(Li et al., 2025) employ LLM-based workflows to synthesize literature evidence starting from clinical questions.

6 Conclusion

We introduce SciGuide to reveal the limitations of LLMs in scientific research, with a focus on evidence-based clinical guideline development. Despite partial improvements brought by prior knowledge utilization and reasoning enhancements, a clear gap remains between existing models and expert-level scientific practice. We hope this work establishes foundations for advancing reliable, domain-grounded scientific reasoning.

7 Limitations

The current evaluation is conducted under a fixed candidate document set and does not incorporate large-scale literature retrieval in open environments. This design helps focus on core capabilities such as literature understanding and multi-evidence reasoning, while retrieval-related abilities remain complementary to retrieval-oriented benchmarks. In addition, the evaluation adopts question-level tasks that are largely independent, without explicitly modeling multi-round iteration or cross-question decision-making, leaving room for further extension toward long-term and continuous scientific reasoning processes.

References

Micheal Abaho, Danushka Bollegala, Paula Williamson, and Susanna Dodd. 2019. Correcting crowdsourced annotations to improve detection of outcome types in evidence based medicine. In *CEUR workshop proceedings*, volume 2429, pages 1–5. CEUR-WS. org.

Lisa Adams, Felix Busch, Tianyu Han, Jean-Baptiste Excoffier, Matthieu Ortala, Alexander Löser, Hugo JWL Aerts, Jakob Nikolas Kather, Daniel Truhn, and Keno Bressem. 2025. Longhealth: A question answering benchmark with long clinical documents. *Journal of Healthcare Informatics Research*, pages 1–17.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Anthropic. 2025. Introducing claude sonnet 4.5. https://www.anthropic.com/news/claude-sonnet-4-5. Accessed: 2025-09-30.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. Surveygen: Quality-aware scientific survey generation with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736.

Timothy Hugh Barker, Jennifer C Stone, Kim Sears, Miloslav Klugar, Jo Leonardi-Bee, Catalin Tufanaru, Edoardo Aromataris, and Zachary Munn. 2023. Revising the jbi quantitative critical appraisal tools to improve their applicability: an overview of methods and the development process. *JBI Evidence Synthesis*, 21(3):478–493.

Chinese Medical Association Chinese Society of Cardiology, Hypertension Committee of Cross-Straits Medicine Exchange Association, Cardiovascular Disease Prevention, and Chinese Association of Rehabilitation Medicine Rehabilitation Committee. 2024. Clinical practice guideline for the management of hypertension in china. *Chinese Medical Journal*, 137(24):2907–2952.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. *arXiv preprint arXiv:2105.03011*.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 123–132, Online. Association for Computational Linguistics.

Leandro Carísio Fernandes, Gustavo Bartz Guedes, Thiago Soares Laitz, Thales Sales Almeida, Rodrigo Nogueira, Roberto Lotufo, and Jayr Pereira. 2024. Surveysum: A dataset for summarizing multiple scientific articles into a survey section. In *Brazilian Conference on Intelligent Systems*, pages 431–444. Springer.

Gemini Team, Google. 2025. Gemini 3 pro model card. Model card, Google DeepMind. Accessed: 2026-01-02. Version 7 DT update published Dec 4, 2025.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

M Hyde, P Higgs, RD Wiggins, and D Blane. 2015. A decade of research using the casp scale: key findings and future directions.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, and 1 others. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Di Jin, Eileen Pan, Nassim Oufattole, Wei-Hung Weng, Hanyi Fang, and Peter Szolovits. 2021. What disease does this patient have? a large-scale open domain question answering dataset from medical exams. *Applied Sciences*, 11(14):6421.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2567–2577.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. Scireviewgen: a large-scale dataset for automatic literature review generation. *arXiv preprint arXiv:2305.15186*.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G Rodriques, and Andrew D White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*.

Dubai Li, Nan Jiang, Kangping Huang, Ruiqi Tu, Shuyu Ouyang, Huayu Yu, Lin Qiao, Chen Yu, Tianshu Zhou, Danyang Tong, and 1 others. 2025. From questions to clinical recommendations: Large language models driving evidence-based clinical decision making. *arXiv preprint arXiv:2505.10282*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, and 1 others. 2025. Deepseek-v3. 2: Pushing the frontier of open large language models. *arXiv preprint arXiv:2512.02556*.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-xscience: A large-scale dataset for extreme multi-document summarization of scientific articles. *arXiv preprint arXiv:2010.14235*.

Lin-Lu Ma, Yun-Yun Wang, Zhi-Hua Yang, Di Huang, Hong Weng, and Xian-Tao Zeng. 2020. Methodological quality (risk of bias) assessment tools for primary and secondary medical studies: what are they and which is better? *Military Medical Research*, 7(1):7.

Benjamin Nye, Ani Nenkova, Iain Marshall, and Byron C Wallace. 2020. Trialstreamer: mapping and browsing medical evidence in real-time. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 63–69.

OpenAI. 2025. Introducing gpt-5.2. https://openai.com/zh-Hans-CN/index/introducing-gpt-5-2/. Accessed: 2025-12-11.

World Health Organization and 1 others. 2021. *Guideline for the pharmacological treatment of hypertension in adults*. World Health Organization.

Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. 2022. Medmcqa: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pages 248–260. PMLR.

Owen Queen, Harrison G Zhang, and James Zou. 2025. Cgbench: Benchmarking language model scientific reasoning for clinical genetics research. *arXiv preprint arXiv:2510.11985*.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2024. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: A novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23(3):289–301.

Beverley J Shea, Barnaby C Reeves, George Wells, Micere Thuku, Candyce Hamel, Julian Moran, David Moher, Peter Tugwell, Vivian Welch, Elizabeth Kristjansson, and 1 others. 2017. Amstar 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both. *bmj*, 358.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, Le Hou, Kevin Clark, Stephen R Pfohl, Heather Cole-Lewis, and 1 others. 2025. Toward expert-level medical question answering with large language models. *Nature Medicine*, 31(3):943–950.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Yang Tian, Zheng Lu, Mingqi Gao, Zheng Liu, and Bo Zhao. 2025. Mmcr: Benchmarking cross-source reasoning in scientific papers. *arXiv preprint arXiv:2503.16856*.

Tao Tu, Shekoofeh Azizi, Danny Driess, Mike Schaekermann, Mohamed Amin, Pi-Chuan Chang, Andrew Carroll, Charles Lau, Ryutaro Tanno, Ira Ktena, and 1 others. 2024. Towards generalist biomedical ai. *Nejm Ai*, 1(3):AIoa2300138.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. *arXiv preprint arXiv:2004.14974*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, and 1 others. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509.

Haoran Wei, Yaofeng Sun, and Yukun Li. 2025. Deepseek-ocr: Contexts optical compression. *arXiv preprint arXiv:2510.18234*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Jing Yu, Yuqi Tang, Kehua Feng, Mingyang Rao, Lei Liang, Zhiqiang Zhang, Mengshu Sun, Wen Zhang, Qiang Zhang, Keyan Ding, and 1 others. 2025. Scicueval: A comprehensive dataset for evaluating scientific context understanding in large language models. *arXiv preprint arXiv:2505.15094*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2025. Large language models for scientific discovery in molecular property prediction. *Nature Machine Intelligence*, pages 1–11.

This Appendix is organized as follows:

- Appendix A provides detailed statistics of the benchmark and describes the construction and quality control procedures.
- Appendix B presents the experiment details.
- Appendix C reports additional experimental results.
- Appendix D contains extended case studies and error analyses.

A Dataset Annotation and Construction

A.1 Dataset Statistics

A.1.1 Overall Statistics

Dataset statistics. As shown in Figure 7(d), the benchmark contains a total of 1,331 instances across four major components. Specifically, the Literature Inclusion/Exclusion (LIE) task includes 211 instances, the PICO Extraction (PE) task consists of 244 instances, and the Quality Appraisal (QA) task comprises 237 instances. The Multi-evidence Inductive Reasoning (MEI) component includes 639 instances in total, which are further divided into four configurations: MEI-1, MEI-2, and MEI-3 each contain 168 instances, while MEI-4 contains 135 instances.

A.1.2 Per-task Statistics

LIE statistics. As shown in Figure 7(b). LIE contains an average of 6.99 candidate documents per question, with the total number of documents per question ranging from 3 to 15. Among these, an average of 5.02 documents per question are considered correct, with the number of correct documents per question ranging from 1 to 13. In contrast, the average number of incorrect documents per question is 1.97, with a range of 1 to 2. This dataset simulates the fine-grained inclusion/exclusion process following large-scale initial screening, where the number of irrelevant or distracting documents is much lower compared to the broader search phase. As a result, the identification of these incorrect documents becomes more challenging.

PE statistics. As shown in Figure 7(c) and Figure 7(e). The quality appraisal tasks require models to capture detailed PICO elements, with an average of 17.03 extracted tuples and 451.75 structured factors per study. Such distributions indicate moderate overall structural complexity, while the presence of denser instances reflects substantial variability in methodological detail. This combination mirrors real-world scientific literature and poses consistent challenges for structured extraction and fine-grained reasoning.

QA statistics. As shown in Figure 7(f), the Quality Appraisal (QA) component comprises a total of 237 instances, including 108 instances assessed using AMSTAR 2, 49 instances evaluated with CASP, 25 instances assessed using JBI, and 55 instances evaluated with ROB 2.0. In this benchmark, JBI denotes analytical cross-sectional studies, CASP denotes case–control and cohort studies, ROB 2.0 denotes randomized controlled trials, and AMSTAR 2 denotes systematic reviews and meta-analyses. The Quality Appraisal task (12.39 features per study on average) requires models to assess study quality as reviewers do.

MEI statistics. In MEI, models are required to synthesize evidence from multiple studies, with each task involving an average of 3.04 studies, and up to 13 studies, in order to formulate recommendations and assess confidence levels.

A.2 Literature Inclusion/Exclusion (LIE) Construction

A.2.1 Candidate Literature Pool Construction

The LIE task aims to construct a dataset that simulates the literature inclusion and exclusion decisions made by medical experts, enabling models to distinguish between supporting evidence and non-applicable, distracting documents given a specific clinical question. To this end, we take sub-questions derived from clinical guidelines as the basic units and construct a corresponding pool of candidate documents for each sub-question, resulting in multiple independent inclusion/exclusion decision instances. Each instance consists of the question text, a set of candidate document identifiers, and gold-standard labels indicating whether each document should be included or excluded.

During candidate document construction, all documents labeled as included are directly sourced from the references cited in the original guideline text and are manually verified to ensure their relevance to the question, particularly with respect to core PICO elements such as the target population, interventions, and outcomes. In contrast, constructing excluded documents constitutes the primary challenge of the LIE task. These documents are often topically similar to the question at a surface level but exhibit a clear mismatch in at least one PICO element, and therefore should not be consid-

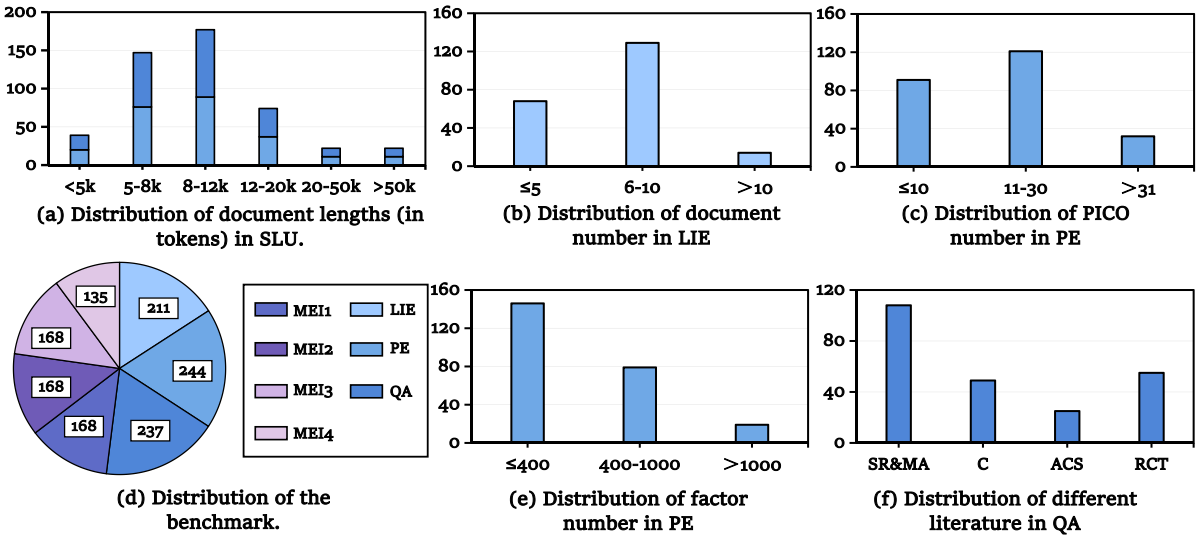

Figure 7: **Dataset statistics across different benchmark components.**

ered valid evidence.

To systematically introduce representative distractors, we construct excluded documents using three complementary strategies: (1) retrieving topically similar literature through keyword-based searches; (2) reusing references from other related but non-identical questions as cross-question distractors; and (3) selecting potential distractors from the reference lists of included documents. Through this process, each sub-question can be associated with multiple inclusion/exclusion decision instances, forming a candidate literature pool that covers diverse types of realistic distractors.

A.2.2 Quality Control

To ensure methodological soundness and consistent task difficulty in the LIE dataset, we applied a systematic quality control and filtering process after the initial construction phase. In the first stage, a total of 504 inclusion/exclusion instances were generated. Each instance was subsequently reviewed in detail, and only 211 high-quality instances were retained in the final dataset, while the remaining instances were discarded for failing to meet quality criteria.

Instance removal was primarily driven by several factors. First, some candidate instances contained distractor documents with insufficient ambiguity, where the lack of relevance was obvious and could be resolved without fine-grained PICO-level reasoning. Such instances do not faithfully reflect real-world literature screening decisions and were therefore excluded. Second, during manual verification, certain documents initially labeled as excluded were found to potentially provide meaningful evidence for the target question at the methodological or outcome level. To avoid introducing controversial or questionable negative samples, these instances were removed from the dataset. In addition, a small number of instances were discarded due to unclear exclusion rationales, insufficient justification of PICO mismatches, or ambiguous evidence boundaries.

We further analyzed the effectiveness of different distractor construction strategies. Whole-question keyword searches in PubMed were found to be suboptimal in many cases, particularly for common or broadly phrased clinical questions. Such searches often returned results that were either weakly related or failed to produce documents with clear PICO-level conflicts. In practice, we observed that more effective retrieval could be achieved by selectively omitting or weakening certain PICO-related keywords rather than querying the full question verbatim.

In contrast, the cross-question citation strategy proved to be consistently effective. This approach preserves strong domain relevance by drawing from related questions, while simultaneously ensuring salient differences in PICO elements such as population, intervention, or outcomes. As a result, it reliably yields realistic and challenging distractors and constitutes a primary source of excluded documents in the final dataset.

Finally, selecting distractors from the reference lists of included documents was also feasible in principle and often produced semantically close

but non-applicable studies. However, this strategy required substantial manual effort, as each candidate needed to be carefully examined for subtle methodological differences. Due to its high annotation cost and low efficiency, this approach was applied only selectively and under strict human verification.

Through these multi-stage filtering and validation procedures, the final LIE dataset achieves a balanced trade-off between relevance, ambiguity, and methodological consistency, closely reflecting the decision complexity faced by medical experts during literature inclusion and exclusion.

A.3 PICO Extraction (PE) Construction

A.3.1 Extraction Targets and Scope

We define the PICO Extraction (PE) task as a high-fidelity structural transformation of heterogeneous clinical literature d into a formalized evidence knowledge tuple $\mathcal{Q}_{\text{pico}}$. To maintain methodological integrity, the proposed framework initiates with a deterministic routing logic that categorizes input literature into three distinct methodological paradigms based on study design and analytic intent. *Strategy A (Evidence Synthesis)* is assigned to secondary research that aggregates existing data, specifically encompassing systematic reviews and meta-analyses. *Strategy B (Interventional)* is reserved for primary research where investigators actively assign treatments, primarily including randomized controlled trials (RCTs) and non-randomized interventional designs. *Strategy C (Observational)* targets analytical research observing natural exposures, covering cohort studies, case-control designs, and analytical cross-sectional studies.

The terminal output of this process is a structured sextuple (P, I, C, O, G, E), where each element is extracted under strategy-specific semantic constraints. P (Population) delineates the demographic and clinical boundaries of the subjects; I (Intervention/Exposure) and C (Comparator/Reference) define the study arms; O (Outcome) specifies the clinical endpoints; G (Subgroup) identifies stratified populations; and E (Effect Size) represents granular statistical estimates including point estimates (e.g., HR, OR, MD), 95% confidence intervals, and significance levels. This framework ensures that observational "Exposures" are never conflated with interventional "Treatments," and that pooled estimates in meta-analyses are clearly distinguished from individual study results. The resulting dataset covers 244 documents, necessitating the extraction of 17.03 PICO elements and 451.75 granular factors per article on average, reflecting the high-density information environment of clinical evidence.

A.3.2 Human–AI Collaborative Annotation

The PE benchmark is constructed through a collaborative, knowledge-augmented pipeline utilizing Claude 3.5 Sonnet as the core reasoning engine, implemented via a decoupled, multi-stage "Blueprint-to-Mining" protocol. This protocol is designed to minimize subjective freedom and enforce strict dependence on reported text:

Methodological Classification and Routing. The system first classifies the literature based on study design and analytic intent. This acts as a routing signal to activate the corresponding strategy (A, B, or C), ensuring that subsequent agents adhere to the correct methodological paradigm and evidence-based medicine (EBM) logic.

Ontological Blueprint Generation. A specialized "Senior Methodologist" agent constructs the structural skeleton of the study without extracting numeric values. It identifies the PICO/PECO elements and enumerates all measured outcomes, generating "search keywords" based on exact aliases used in the text. To prevent the over-reporting of incidental findings, this stage enforces a "Primary-Stingy" rule, labeling an outcome as "Primary" only when it is explicitly identified as the endpoint for sample size calculations.

High-Fidelity Statistical Mining. Guided by the blueprint, a "Precision Extraction" agent performs an exhaustive scan of narrative text, multi-dimensional tables, and forest plots to locate quantitative estimates. The agent operates under strict methodological constraints: for Strategy C, it is programmed to prioritize the most fully adjusted models to account for confounding; for Strategy B, it performs multi-section cross-referencing across baseline characteristics and methodology descriptions to recover denominator populations (N) that may be missing from primary results tables.

Tuple Transformation. The final stage involves the deterministic mapping of extracted data into the formalized (P, I, C, O, G, E) format. Each extracted value is linked to a mandatory

quote_source field, ensuring that every annotation is traceable to explicit textual evidence and preserving the auditability of the synthesis.

A.3.3 Quality Control

To ensure the "Golden Standard" integrity of the dataset, we implement a rigorous quality control (QC) protocol centered on human verification of automated outputs. The process begins with an automated *Validator Agent* that conducts logical consistency checks, identifying missing components or statistical contradictions, such as effect values inconsistent with reported 95% confidence intervals. For documents identified with distributed evidence gaps, we utilize a "Targeted Patching" mechanism, which generates a manifest of missing fields and rescans the full text with expanded context to capture information scattered across footnotes or appendices.

The final validation is conducted by a clinical review board including seven medical M.S. students and three senior cardiovascular physicians with extensive experience in clinical guideline development. Reviewers focus on **Information Distribution Correction**, manually adjusting cases where the model assigned "Not Reported" (NR) due to the fragmented nature of clinical reporting. For instance, when sample sizes or event counts are dispersed across methodology sections and supplementary results, human experts integrate these findings to update the final annotations. Disagreements in PICO classification or effect size mapping are adjudicated by a third physician, ensuring that overall conclusions are mechanically derived from explicit evidence. This process preserves full auditability, retaining both the original model outputs and the human-revised annotations for subsequent consistency analysis and error pattern investigation.

A.4 Quality Appraisal (QA) Construction

A.4.1 Study Types and Appraisal Frameworks

We construct a hierarchical, method-driven decision tree to automatically classify input literature prior to dataset synthesis. The process is grounded in methodological assessability and begins by excluding materials that do not constitute reliable research, including comments, editorials, letters, news articles, preprints, and audio or video content. These items are categorized as unreliable research and removed from subsequent processing. For the remaining documents, we first distinguish research intent. Studies that synthesize existing evidence, evaluate study quality, or aggregate effect sizes, or that explicitly identify as systematic reviews, meta-analyses, clinical guidelines, expert consensuses, or health technology assessments, are classified as secondary research. All other eligible documents are treated as primary research and proceed to further methodological differentiation.

Primary studies are then divided according to whether investigators actively assign interventions or controlled exposures. Studies involving deliberate allocation of treatments, behavioral interventions, or controlled exposures are classified as interventional studies, while those observing exposures without investigator control are classified as observational studies. Interventional studies are further examined for explicit descriptions of randomization, allowing randomized controlled trials to be distinguished from non-randomized interventional designs. Randomized trials are subsequently categorized by specific trial structures, whereas non-randomized interventional studies are differentiated based on the presence and nature of control groups. Observational studies are classified based on analytic intent and temporal direction. Studies that explicitly evaluate exposure–outcome associations using comparative groups are identified as analytical observational studies and further categorized as cohort, case–control, or analytical cross-sectional designs according to temporal ordering. Studies that do not meet these criteria are treated as descriptive observational research.

Following methodological classification, quality assessment frameworks are deterministically assigned based on study design. Cohort and case–control studies are evaluated using the CASP framework (Hyde et al., 2015), analytical cross-sectional studies are assessed using JBI (Barker et al., 2023) criteria, randomized controlled trials are evaluated under the ROB 2.0 (Ma et al., 2020) framework, and systematic reviews or meta-analyses are assessed using AMSTAR 2 (Shea et al., 2017). This one-to-one mapping ensures that each study is evaluated exclusively with criteria appropriate to its methodological paradigm, preventing cross-framework contamination.

Building on this structure, we implement a multi-agent data synthesis pipeline in which study classifications act as routing signals for automated assessment. Systematic reviews and meta-analyses are processed directly by an AMSTAR 2 agent. Observational studies are routed to CASP or JBI

agents according to their specific designs. Randomized controlled trials are decomposed into multiple bias domains, each independently evaluated by specialized ROB 2.0 sub-agents focusing on randomization, deviations from intended interventions, missing outcome data, outcome measurement, and selective reporting. Outputs from these sub-agents are subsequently aggregated to produce structured, item-level annotations. By centering the pipeline on study design and enforcing low-freedom, discrete decision tasks with human verification, this approach enables scalable, auditable, and methodologically consistent construction of the QA dataset.

A.4.2 Annotation Protocol

Under the above data synthesis framework, we define a unified annotation protocol that governs how each evaluation agent operates and how annotations are produced. The protocol is designed to minimize subjective freedom, enforce strict dependence on reported text, and ensure consistency across heterogeneous study designs, while remaining interpretable to non-medical researchers.

For secondary research, systematic reviews and meta-analyses are annotated using a dedicated AMSTAR 2 agent. This agent performs a single-pass, item-level methodological appraisal strictly following the 16 AMSTAR 2 criteria and produces an overall confidence rating. All judgments are text-grounded: the agent is explicitly prohibited from inferring unreported methods or supplementing missing information with external knowledge. Each item is labeled using a constrained three-way decision (yes, partial, no), accompanied by brief justifications and direct textual evidence. A fixed subset of critical items is highlighted and deterministically mapped to the final confidence level, ensuring that overall ratings are traceable to explicit methodological deficiencies rather than holistic impressions.

For randomized controlled trials, risk-of-bias annotation follows a decomposed RoB 2.0 protocol implemented through multiple specialized agents. Each agent is responsible for exactly one bias domain—randomization process, deviations from intended interventions, missing outcome data, outcome measurement, or selective reporting—and is restricted to domain-specific evidence and rules. Domain agents output low, some concerns, or high risk labels based solely on reported trial information, without access to other domains. A separate aggregation agent then combines these domain-level outputs using the official RoB 2.0 decision rules to produce an overall risk-of-bias judgment, ensuring that global conclusions are mechanically derived rather than implicitly reasoned.

For analytical cross-sectional studies, a JBI agent is used to conduct methodological appraisal based on the eight-item JBI checklist. Each item is annotated with a closed-set label (yes, no, unclear, or not applicable), supported by short explanations and direct quotations from the study. In addition to item-level judgments, the agent produces an overall appraisal decision indicating whether the study should be included, excluded, or flagged for further information, summarizing the main methodological strengths and limitations in a structured manner.

For observational cohort and case–control studies, separate CASP-based agents are employed to reflect design-specific appraisal logic. Both agents evaluate studies against predefined CASP criteria using strictly bounded response options (yes, no, can't tell), require explicit textual evidence for each judgment, and produce an overall quality rating derived from deterministic rules over key methodological domains. Particular emphasis is placed on recruitment, exposure and outcome measurement, confounding control, and follow-up adequacy, with insufficient reporting explicitly propagated as uncertainty rather than silently resolved.

Across all agents, annotations are generated as structured JSON objects with fixed schemas, itemized judgments, and explicit evidence quotes. This protocol enforces low-degree-of-freedom decisions, isolates domain-specific reasoning, and makes uncertainty explicit, enabling scalable annotation while preserving methodological fidelity and auditability.

A.4.3 Quality Control

After automatic annotation, we introduce a human quality control stage to identify and correct systematic issues arising from information sparsity, textual ambiguity, or conservative model behavior. This process focuses on ensuring that annotations accurately reflect what is explicitly reported in the source texts and that methodological judgments are applied consistently across studies.

The most frequently modified items are those related to missing or insufficiently reported information. Across the AMSTAR 2 and CASP/JBI frameworks, the model tends to assign "no" or "can't tell" labels when relevant information is not immediately identifiable in a single section. Manual

review reveals that, in some cases, required details are present but distributed across appendices, supplementary materials, or results sections. These items are therefore revised to "partial" or "yes" when sufficient evidence can be located, preventing the misclassification of dispersed reporting as complete non-reporting.

Items that require multi-step or cross-section evidence integration are also more prone to revision. For example, AMSTAR 2 items concerning search strategy adequacy, excluded-study lists, and publication bias assessment often depend on information scattered across the main text and supplementary files. When the model assigns conservative ratings due to localized evidence gaps, human reviewers integrate evidence across sections and adjust item-level judgments accordingly, typically from "no" to "partial," while avoiding upgrades in the absence of explicit methodological support.

Judgments that affect overall study-level conclusions, such as the overall risk-of-bias rating in RoB 2.0 or the overall confidence level in AMSTAR 2, are reviewed with particular care but are modified less frequently. In these cases, human review primarily verifies whether an overall downgrade was triggered by an incorrect assessment of a critical item. When a critical item is corrected, the corresponding overall rating is updated deterministically; otherwise, global conclusions are left unchanged.

In addition, items related to study design boundaries, such as confounding control in observational studies or follow-up adequacy in cohort studies, are selectively reviewed. Human intervention is limited to cases where the model either overlooks explicitly stated methodological limitations or unduly penalizes studies that appropriately acknowledge and discuss their constraints, thereby avoiding conflation of domain-specific design characteristics with methodological flaws.

Overall, revisions introduced during quality control are concentrated at the item level rather than at the level of aggregate conclusions. Most corrections address conservative interpretations of information distribution and insufficient cross-section evidence aggregation rather than systematic directional errors. All modified items retain both the original model outputs and the human-revised annotations, enabling subsequent consistency analysis and error pattern investigation while preserving full auditability of the dataset.

A.5 Multi-evidence Inductive Reasoning (MEI) Construction

A.5.1 Annotation Protocol

The construction of the MEI (Multi-evidence Inductive Reasoning) task starts directly from clinical practice guidelines, leveraging their inherently structured formulation of clinical questions and recommendations. For each clinical question defined in the guideline, we extract the original question text together with its corresponding recommendations, recommendation rationales, and the explicitly cited references. These elements jointly form the basic semantic units for MEI data construction, ensuring that questions, conclusions, and supporting evidence are consistently grounded in the same authoritative source.

In clinical guidelines, a single clinical question often corresponds to multiple parallel or conditional recommendations, reflecting different patient subgroups, comorbidities, or clinical contexts. Treating such cases as a single question–answer pair would obscure the underlying decision structure. We therefore decompose these complex questions at the level of individual recommendations and construct separate MEI instances for each recommendation. For example, a guideline question on antihypertensive therapy in patients with coronary artery disease may recommend different first-line agents for patients with angina versus those with a history of myocardial infarction. These recommendations are converted into distinct problem instances rather than being merged into a single response.

The question construction process consists of two main steps. First, we perform question decomposition based on guideline recommendations. For each recommendation associated with a clinical question, a corresponding sub-question is generated under structured prompting, such that the medical semantics of the original question are preserved while the decision target is restricted to a single recommendation. At this stage, the original clinical question, the derived sub-question, and the associated recommendation text are recorded.

Second, we conduct recommendation rationale alignment and decomposition. Recommendation rationales in guidelines are typically written as integrated justifications that collectively support multiple recommendations, without explicitly delineating evidence boundaries. Given the predefined sub-questions, we therefore require a language model

to restructure the guideline's rationale text by identifying which explanatory components correspond to each sub-recommendation and which referenced studies contribute directly to its support. This step yields a recommendation-specific rationale and an associated set of supporting references for each sub-question.

Through this process, each MEI instance is centered on a single recommendation and is paired with a corresponding justification and multiple supporting studies. This design preserves the original evidence-based reasoning logic of clinical guidelines while transforming it into a standardized input format suitable for evaluating models' ability to perform inductive reasoning over multiple pieces of evidence.

A.5.2 Quality Control

During MEI dataset construction, we introduce explicit revision and exclusion strategies at the human quality control stage to address inconsistencies arising from automatic question decomposition and evidence alignment. These strategies operate at both the question and evidence levels and are applied to ensure internal coherence within each MEI instance.

First, we apply a question rewriting strategy when the automatically decomposed sub-questions fail to align tightly with their corresponding recommendations. In such cases, sub-questions may remain overly close to the original high-level question or introduce constraints irrelevant to the target recommendation, resulting in ambiguous decision targets. Human reviewers rewrite the question formulation based on the guideline text so that each question clearly corresponds to a single recommendation, without altering the recommendation itself or its evidentiary basis.

Second, we adopt a reference re-identification and realignment strategy when supporting studies are incorrectly assigned. During automatic decomposition of recommendation rationales, the model may attribute references that actually support other sub-recommendations, or omit studies that are explicitly cited by the guideline as critical evidence. In these cases, human review reexamines the guideline citations and reconstructs the reference set to ensure that each MEI instance includes only studies that directly support the associated recommendation.

Third, we apply a discarding strategy for instances in which key supporting studies cannot be accessed due to copyright or availability constraints. When guideline-cited references are essential for understanding or evaluating the recommendation but their full texts cannot be legally obtained, the resulting evidence gaps prevent reliable multi-evidence reasoning. Such instances are therefore removed from the dataset rather than being retained with incomplete evidence.

Finally, when both question ambiguity and evidence unavailability occur simultaneously, the instance is excluded rather than being retained through speculative rewriting or weakened evidence requirements. These revision and exclusion strategies collectively reduce noise introduced during automatic construction and preserve clear alignment between questions, recommendations, and supporting evidence within the MEI dataset.

B Experiment Details

B.1 Experiment Environment

All experiments reported in this benchmark were conducted on the system configuration summarized below. To ensure stable and reproducible evaluation, model inference was performed exclusively via external API interfaces. The local system was used solely for experiment orchestration, data preprocessing, and result aggregation. The system configuration is summarized below.

- **CPU:** Dual-socket Intel Xeon Gold 6148 (2.40 GHz), 40 cores per socket, 80 threads total
- **GPU:** 8× NVIDIA A40, each with 48 GB VRAM
- **GPU driver:** 525.125.06, **CUDA:** 11.8
- **cuDNN:** 8.x (compiled with CUDA 11.8)
- **Operating System:** Ubuntu 20.04.6 LTS

B.2 Experiment Models

All model variants evaluated in the experiments, including their specific model configurations, parameter scales, and deployment platforms, are summarized in Table 5.

B.3 Experiment Metrics

B.3.1 Literature Inclusion/Exclusion (LIE)

Literature Inclusion/Exclusion (LIE). Let inclusion be treated as the positive class. Inclusion performance is evaluated using the F1 score (I-F1), computed from precision and recall defined over true positives, false positives, and false negatives.

To symmetrically assess exclusion performance, exclusion is treated as the positive class via label

Platform	Model	Release	Version	Size
		Proprietary models		
Anthropic	Claude 4.5 Sonnet	2025-12	claude-4.5-sonnet	-
OpenAI	GPT-5.2	2025-12	gpt-5.2	-
	o4-mini-high	2025-04	o4-mini-high	-
Google	Gemini 3 Pro Preview	2025-11	gemini-3-pro-preview	-
	Gemini 3 Flash Preview	2025-11	gemini-3-flash-preview	-
ByteDance	Doubao Seed 1.6	2025-06	doubao-seed-1.6-thinking	-
		Open source		
DeepSeek	DeepSeek-V3.2 (reasoner)	2025-12	deepseek-v3.2-reasoner	-
	DeepSeek-V3.2 (chat)	2025-12	deepseek-v3.2-chat	-
OpenAI	gpt-oss-120b	2025-08	gpt-oss-120b	120B
Alibaba	Qwen3-Next-80B-A3B-Thinking	2025-12	qwen3-next-80b-a3b-thinking	80B
	Qwen3-Next-80B-A3B-Instruct	2025-12	qwen3-next-80b-a3b-instruct	80B
	Qwen3-32B-Thinking	2025-11	qwen3-32b-thinking	32B
	Qwen3-30B-A3B-Thinking	2025-11	qwen3-30b-a3b-thinking	30B
	Qwen3-30B-A3B-Instruct	2025-11	qwen3-30b-a3b-instruct	30B
	Qwen3-32B-Instruct	2025-10	qwen3-32b-instruct	32B
	Qwen3-14B (Thinking)	2025-10	qwen3-14b-thinking	14B
	Qwen3-14B (Instruct)	2025-10	qwen3-14b-instruct	14B
Meta	Llama 4 Maverick	2025-04	llama-4-maverick	400A17B

Table 5: Details of the foundation models evaluated in our study, categorized by proprietary and open-source status.

inversion, yielding a corresponding F1 score (E-F1) computed over true negatives, false negatives, and false positives.

Finally, macro-F1 is reported as the arithmetic mean of I-F1 and E-F1 to reflect balanced performance across inclusion and exclusion decisions.

B.3.2 PICO Extraction (PE)

PICO Extraction (PE) The evaluation of PICO extraction is primarily measured through Recall, which assesses the comprehensiveness of the model in capturing the complete set of evidence units defined in the gold standard. This evaluation is executed via a two-stage automated auditing framework. In the first stage, tuple-level recall is determined using a five-coordinate clinical semantic alignment strategy. Each gold-standard tuple is mapped to the predicted output based on the semantic alignment of population, intervention, comparator, outcome, and estimate context. A gold-standard tuple is considered successfully recalled if a corresponding match is identified in the model output, indicating that the core evidence unit has been correctly located.

In the second stage, field-level recall is evaluated for these matched tuples by auditing the accuracy of specific payload fields, including statistical estimates, group characteristics, and evidence sources. A field is counted as correctly recalled only if it meets the predefined semantic or numerical matching criteria, such as the 0.02 tolerance threshold for floating-point values. The final PE performance is reported as the overall Recall score, representing the ratio of successfully extracted and verified gold-standard elements to the total number of required elements across all instances. This focused metric highlights the model's effectiveness in preventing the omission of critical clinical evidence, which is paramount in the context of evidence-based medicine.

B.3.3 Quality Appraisal (QA)

Quality Appraisal (QA). For the QA task, we evaluate model outputs using a single metric, *Evidence Coverage*, which measures how comprehensively the model output captures the strengths and limitations specified in the gold-standard risk-of-bias assessment.

Each gold-standard evidence item is assigned a type-dependent weight, with regular strengths weighted as 1, critical strengths as 2, regular limitations as 2, and critical limitations as 3. The maximum attainable evidence score for each instance is defined as the sum of the weights of all gold-standard evidence items.

Model-covered evidence is quantified by sum-

ming the weights of gold-standard strengths and limitations that are correctly captured by the model output, based on semantic alignment with the original article and the gold standard. The evidence coverage rate is then computed as the ratio between the model-captured evidence score and the maximum attainable evidence score.

The final EvidenceCoverage score is obtained by linearly scaling the coverage rate to a 0–10 range and rounding to the nearest integer. Overall QA performance is reported as the average EvidenceCoverage score across all successfully evaluated instances. For presentation convenience, EvidenceCoverage scores in the main table are multiplied by 100.

B.3.4 Multi-evidence Inductive Reasoning (MEI)

Multi-evidence Inductive Reasoning (MEI). For MEI, we evaluate both (i) whether the model reaches the correct recommendation conclusion and (ii) whether it assigns the correct evidence quality grade under the GRADE framework.

Recommendation correctness. We use an LLM-based judge to compare the predicted recommendation with the gold recommendation at two granularities: *strict correctness*, which requires full semantic equivalence (including population, thresholds, direction, and applicable conditions), and *direction correctness*, which only checks whether the overall recommendation direction is consistent with the gold reference while ignoring finer conditions. We report the average strict and direction correctness rates over all scored instances.

GRADE correctness with gating. GRADE correctness is evaluated by exact matching between the predicted and gold grades, reported at three levels: *full* grade (e.g., 2B), the *numeric level* (1 vs. 2), and the *letter strength* (A/B/C/D). To ensure that grade evaluation is meaningful only when the recommendation is aligned, we apply a gating protocol: under the strict setting, grade match is counted only for instances with strictly correct recommendations; under the lenient setting, grade match is counted only for instances with direction-correct recommendations. Overall MEI performance is reported as the averaged scores under these strict and lenient gating settings.

B.4 Evaluation Protocol

The detailed prompt words for each method are shown in the following pages.

B.4.1 Inference

LIE. In the LIE task, the prompt applies role-based expert conditioning to anchor the model in evidence-based medicine practice, PICO-guided structured reasoning to constrain relevance judgments, and output schema enforcement to ensure deterministic, machine-readable inclusion/exclusion decisions.

PE. In the PE task, the prompt implements a two-phase Structured Clinical Evidence Extraction (SCEE) framework. First, it employs triage and strategy selection to anchor the model to study-specific extraction rules, such as prioritizing pooled summary estimates for meta-analyses or adjusted models for observational studies. This is followed by atomic evidence extraction, which guides the model through PICO reconstruction and outcome profiling to generate discrete, self-contained evidence tuples. To ensure data fidelity, the prompt enforces a rigid JSON schema covering estimate context, group-level statistics, and significance values, effectively transforming unstructured medical text into high-granularity, machine-readable clinical evidence data.

QA. In the QA task, the prompt employs methodological expert role conditioning to guide the model to assess study quality from a reviewer-like perspective, and adopts a task definition without predefined retrieval targets, specifying only the evaluation objective and requiring the model to rely on its own capabilities for holistic, full-text methodological understanding and judgment. Building on this, the prompt incorporates study-design–adaptive quality appraisal, where the model first identifies the study type from the text and applies corresponding methodological evaluation criteria, enforces evidence-bounded reasoning strictly limited to explicitly reported information, and finally uses strict JSON schema constraints to produce standardized, machine-readable methodological quality assessments (overall evaluation, strengths, limitations, and conclusion).

MEI. In the MEI task, the prompt casts the model as a clinical guideline developer and evidence-based medicine methodologist, requiring it to generate clinical guideline recommendations under the GRADE framework based solely on the provided structured evidence list (evi_list). The task explicitly requires explicit and traceable evidence-based reasoning, rather than surface-level summarization.

The prompt instructs the model to assess evidence quality following standard GRADE principles, with possible downgrading based on risk of bias, inconsistency, indirectness, imprecision, and publication bias, and upgrading in the presence of large effects, dose–response relationships, or bias-reducing considerations. Evidence quality is categorized into four levels (A–D).

In parallel, the model is required to determine recommendation strength (strong vs. conditional), based on the balance of benefits and harms for critical outcomes and the credibility of the supporting evidence, with particular emphasis on the lowest-quality evidence among critical outcomes.

Finally, the prompt enforces strict output constraints, mandating the combined numeric–letter GRADE format (e.g., 1A, 2B) and prohibiting alternative expressions. The model must produce a fixed-field JSON output containing the recommendation statement, the assigned GRADE level, a clinician-oriented explanation, and a detailed reasoning trace covering critical outcomes, effect magnitude, evidence consistency, and benefit–harm balance. All judgments are strictly limited to the information provided in evi_list, with no external knowledge or subjective assumptions permitted.

B.4.2 LLM-as-a-Judge

PE. In the PE evaluation stage, we implement a two-stage automated auditing framework. The first stage focuses on tuple matching based on a 5-coordinate clinical semantic alignment strategy, mapping golden evidence units to model predictions across population, intervention, comparator, outcome, and estimate context dimensions through one-to-one semantic pairing. The second stage involves a detailed field-level verification (payload audit), where the judge assesses the precision of specific data points, including sample sizes, effect values, and statistical significance. Specifically, a strict numerical tolerance of ± 0.02 is applied to floating-point values, and semantic matching is used for categorical variables to ensure clinical accuracy. Similar to other tasks, this stage enforces zero-temperature decoding and strict JSON schema constraints to guarantee scoring reproducibility. This dual-layered approach allows for a precise decomposition of model errors into alignment failures (missed tuples) and extraction inaccuracies (wrong field values).The reliability of this automated system is further validated by a high Spearman correlation with human expert ratings ($\rho = 0.851$).

QA. In the QA evaluation stage, we adopt an LLM-as-a-Judge scheme and use doubao-seed-1.6, a judge model with reasoning-enhanced (intrinsic CoT) capabilities, to perform automatic scoring. The evaluation prompts separate system-level and user-level instructions to clearly specify the judging role and scoring criteria, and incorporate methodological framework–driven domain knowledge injection by dynamically loading AMSTAR-2, CASP, ROB-2, and JBI guidelines according to the study type. During evaluation, the judge conducts gold–prediction alignment based on a full-text review of the source document and performs item-level evidence coverage assessment. In addition, zero-temperature decoding and strict JSON output constraints are applied to ensure scoring stability and reproducibility. For QA quality assessment, the Spearman correlation reaches 0.712, indicating consistent relative ordering of answer quality, while the mean absolute error (MAE) is 0.951 on a 0–10 scale, suggesting that absolute scoring deviations are small on average.

MEI. In the MEI evaluation stage, we adopt an LLM-as-a-Judge framework with doubao-seed-1.6 to assess both recommendation correctness and GRADE assignment. The evaluation pipeline integrates automatic output repair via LLM-based structured extraction to normalize model outputs into a canonical JSON format, followed by rule-based grading for GRADE labels (separating strength and evidence level) and LLM-based semantic judgment for recommendation consistency. The judge applies dual-criteria evaluation (strict semantic equivalence and directional consistency), enforces consistency constraints between recommendation and GRADE scoring, and uses zero-temperature decoding with structured output constraints. For MEI evaluation, we use Cohen's κ to measure human–machine agreement, as the task involves binary judgments and κ explicitly accounts for chance agreement under class imbalance. Directional consistency (LM) achieves near-perfect agreement ($\kappa = 0.913$), indicating strong alignment in recommendation direction. In contrast, strict semantic equivalence (EM) yields substantial agreement ($\kappa = 0.651$), reflecting the greater difficulty of exact guideline-level matching that requires precise alignment on thresholds, target populations, and applicability conditions.

B.5 OCR Setting

OCR Implementation and Configuration. All open-access PDF documents were converted into machine-readable text using the open-source DeepSeek-OCR toolkit (Wei et al., 2025). OCR preprocessing was executed in a dedicated conda environment (deepseek-ocr) with vLLM version 0.8.5+cu118 as the inference backend. PDF pages were rendered at a resolution of 144 DPI to preserve fine-grained layout details before being processed by the OCR model.

For multimodal encoding, images were tokenized using the DeepSeek OCR processor with image-level tokenization and cropping enabled, following the highest-quality configuration provided by the official implementation. Model inference was performed with vLLM using a maximum context length of 8192 tokens, a block size of 256, and a GPU memory utilization ratio of 0.9.

To ensure deterministic and stable decoding, greedy generation with zero temperature was adopted together with a custom no-repeat n-gram constraint, configured with an n-gram size of 20 and a window size of 50, to suppress repetitive structural tokens.

The OCR pipeline was parallelized using configurable concurrency and worker settings for efficient batch processing.

C More Experimental Results and Discussions

C.1 How priors and inductive reasoning facilitate scientific discovery

To analyze how priors and inductive reasoning contribute to performance in MEI, we stratify models based on their EM-Rec scores under the MEI1 (R+G) setting, which reflects inductive reasoning performance when multiple structured documents (MSD) are provided. Using two fixed thresholds (22 and 25), models are divided into three performance tiers, yielding six high-tier, five mid-tier, and seven low-tier models. This tiering is applied uniformly across all models and is fully determined by MEI1 performance, with model assignments strictly consistent with the main results table.

The high-tier group (Tier 1) consists of qwen3-14b (thinking), qwen3-30b-a3b-instruct-2507, gpt-oss-120b, qwen3-30b-a3b-thinking-2507, qwen3-next-80b-a3b-thinking, and qwen3-14b (instruct). All these models achieve MEI1 EM-Rec scores of at least 25, with an average R+G EM-Rec of 26.49. Under the MEI3 (R) setting, which removes external documents and thus reflects prior-only capability, this group attains an average EM-Rec of 18.65, indicating a relatively strong and consistent level of prior reasoning ability.

The mid-tier group (Tier 2) includes Gemini 3 Flash Preview, claude-sonnet-4.5, qwen3-32b (thinking), qwen3-32b (instruct), and DeepSeek-V3.2 (reasoner). These models fall within the MEI1 EM-Rec range of 22 to 25, with an average R+G EM-Rec of 23.09. Notably, their average MEI3 EM-Rec reaches 18.93, which is nearly identical to that of the high-tier group, suggesting that the two tiers exhibit comparable levels of prior capability when external document inputs are removed.

The low-tier group (Tier 3) comprises doubao-seed-1.6, Gemini-3-pro, Llama-4-Maverick, qwen3-next-80b-a3b-instruct, o4-mini-high, DeepSeek-V3.2 (chat), and GPT-5.2. These models all score below 22 in MEI1, with an average R+G EM-Rec of 15.90. Their prior-only performance under MEI3 is also substantially lower, with an average EM-Rec of 14.23, indicating weaker baseline reasoning ability.

Comparing MEI1 with MEI3 across tiers reveals a clear pattern. Despite exhibiting nearly identical prior performance, high-tier models obtain substantially larger gains from MSD than mid-tier models, with average improvements of +7.84 versus +4.17 EM-Rec, respectively. In contrast, low-tier models not only show the weakest prior performance but also derive the smallest benefit from MSD, with an average gain of only +1.67. These results indicate that when prior capability is held constant, differences in inductive reasoning over multiple structured documents become the primary factor distinguishing high- and mid-tier models. More broadly, the findings mirror experience-driven human scientific practice: sufficient prior knowledge is a prerequisite for effective hypothesis formation, critical judgment, and the successful execution of inductive reasoning processes.

C.2 Analysis of Evidence–recommendation con- fusion

To quantify whether models conflate evidence quality with recommendation strength in GRADE-based decision making, we conduct a correlation analysis between the two components encoded in GRADE labels. Specifically, each GRADE grade is decomposed into a numeric component representing recommendation strength (1 = strong, 2

= weak) and a letter component representing evidence quality (A–D mapped to an ordinal scale). For each model, we compute the Spearman rank correlation between these two dimensions using the model-predicted grades, and compare it against the same correlation computed from the gold annotations.

Results show that the gold annotations exhibit a moderate positive correlation ($\rho \approx 0.51$), reflecting that while evidence quality influences recommendation strength, the two are not deterministically coupled in real guideline decision making. In contrast, most models display equal or substantially higher correlations, with several large or reasoning-oriented models exceeding $\rho = 0.65$ and, in extreme cases, approaching $\rho \approx 0.8$. This systematic inflation indicates that models tend to implicitly assume a monotonic mapping from higher-quality evidence to stronger recommendations, failing to adequately incorporate other guideline decision factors such as benefit–harm trade-offs, population applicability, values and preferences, or resource considerations.

Notably, once the predicted correlation surpasses the gold reference, overall recommendation performance begins to degrade, suggesting that excessive coupling is not a sign of better reasoning but rather a symptom of over-simplified decision heuristics. Conversely, a small number of models exhibit correlations below the gold level, reflecting under-utilization of evidence signals. Together, these findings support the claim that current LLMs lack calibrated, multi-factor decision-making in recommendation strength assignment, instead relying on an overly evidence-centric shortcut that deviates from expert guideline reasoning.

D Case Study

D.1 Literature Inclusion/Exclusion(LIE)

In this section, we present representative case studies for the LIE task, shown in Figure 8–Figure 11.

D.2 PICO Extraction(PE)

In this section, we present representative case studies for the PE task, shown in Figure 12–Figure 16.

D.3 Quality Appraisal(QA)

In this section, we present representative case studies for the QA task, shown in Figure 17–Figure 21.

D.4 Multi-evidence Inductive Reasoning(MEI)

D.4.1 Ignoring Critical Details

We presents cases where models fail to include critical thresholds or operational details required for actionable recommendations, shown in Figure 22–Figure 26.

D.4.2 Conditioning Errors

We presents cases in which models apply recommendations to incorrect conditions, shown in Figure 27–Figure 31.

D.4.3 Decision Path Compression

We presents case where multi-step or conditional decision processes is overly compressed into simplified conclusions, shown in Figure 32.

D.4.4 Evidence–recommendation Confusion

We presents cases where models conflate evidence quality with recommendation strength, overlooking other essential decision factors, shown in Figure 33–Figure 37.

Literature Inclusion/Exclusion (LIE) Inference Prompt

 # Role
You are an **expert Clinical Guideline Developer and Evidence-Based Medicine Methodologist**. You are familiar with the full workflow of systematic reviews and clinical guideline development, and you are capable of making rigorous, transparent, and methodologically sound judgments on research literature based on evidence-based principles.

 # Task Definition
Citation Screening refers to the process of making inclusion or exclusion decisions for a set of **previously retrieved candidate documents** with respect to a **clearly specified clinical question**.
Your task is to decide, for each candidate document, whether it should be included (include) or excluded (exclude) based on its relevance and eligibility for answering the clinical question.
All decisions must be made strictly according to the **PICO framework**:
* **P (Population)**: Does the study population match or closely correspond to the target population defined in the clinical question?
* **I (Intervention)**: Does the study investigate the intervention or exposure of interest specified in the clinical question?
* **C (Comparison)**: Where applicable, is the comparison group appropriate and relevant to the clinical question?
* **O (Outcome)**: Does the study report outcomes that are relevant to those specified in the clinical question?

 # Instructions
1. Carefully analyze the provided **Clinical Question**.
2. Review the provided **Candidate Documents**, based on their metadata.
3. For each candidate document, determine whether it meets the inclusion criteria for the clinical question:
* If it meets the criteria, assign **include**.
* If it does not meet the criteria or is clearly irrelevant, assign **exclude**.

 # Output Format (Must Be Strictly Followed)
You may include explanatory text outside the structured output.
However, your response **must contain exactly one JSON object** representing the screening results, and this JSON **must be wrapped with fixed markers** to enable automatic parsing.
The required markers and format are as follows:

```
<JSON>
{
  "Doc_ID": {
    "decision": "include" or "exclude",
    "reason": "Brief explanation based on PICO."
  }
}
</JSON>
```

Requirements
<JSON> and </JSON> must each appear exactly once.
The content between the markers must be one single, complete, and valid JSON object.

PICO Extraction (PE) Inference Prompt

You are an expert Clinical Research Assistant specializing in Evidence-Based Medicine (EBM). Your task is to perform **Structured Clinical Evidence Extraction (SCEE)**.

Phase 1: Triage & Strategy Selection (CRITICAL STEP)

Before extracting any data, you must classify the input paper into one of the following 3 strategies. **This determines WHAT you should extract.**

* **STRATEGY_A_SYNTHESIS (Meta-Analysis / Systematic Review)**
- Target: Secondary research aggregating multiple studies.
- **DO:** Extract only the **Pooled / Summary Estimates** (the overall diamond/conclusion).
- **DON'T:** Do **NOT** extract individual study rows (e.g., individual lines in a forest plot). This is a strict resource-saving constraint.

* **STRATEGY_B_RCT (Interventional Studies)**
- Target: Randomized trials (Parallel, Cross-over, Cluster).
- **DO:** Prioritize **Intention-to-Treat (ITT)** results for efficacy and **Safety Set** for adverse events.
- **DON'T:** Do not focus on Per-Protocol (PP) analysis unless ITT is missing.

* **STRATEGY_C_OBS (Observational Studies)**
- Target: Cohorts, Case-Control, Cross-Sectional.
- **DO:** Map "Intervention" to **"Exposure"**. Prioritize the **Most Fully Adjusted Model** (multivariate) to control for confounding factors (e.g., age, sex, comorbidities).
- **DON'T:** Do not extract "Crude/Unadjusted" estimates if adjusted ones exist.

Phase 2: Atomic Evidence Extraction

Based on the selected strategy, extract the **"Evidence Profile"** into a structured list.
Each entry must be a self-contained **Atomic Evidence Tuple** representing a specific statistical analysis.

Core Reasoning Process

1. **PICO Reconstruction:** Accurately identify the Population (P), Intervention (I), and Comparator (C) definitions.
2. **Outcome Profiling:** Distinguish between Primary and Secondary outcomes.
3. **Atomic Identification:** Identify every distinct statistical estimate reported. Identify **all** study groups involved in that specific analysis.

Output Format: The Evidence Schema

Output a valid JSON object following this strict skeleton.
Fill in the values based on the text. Use "NR" for missing data.

```
{
  "meta": {
    "file_id": "...",
    "doc_type": "<e.g., Meta-analysis>",
   "assigned_strategy": "<STRATEGY_A_SYNTHESIS | STRATEGY_B_RCT | STRATEGY_C_OBS>"
  },
  "evidence_tuples": [
    {
      "trace_id": "String (P{x}-O{y}-E{z})",

      "pico_context": {
        "pico_id": "String",
        "study_design_specifics": "String",
        "P": "String",
```

```
        "I": "String (Note: Put 'Exposure' here for Strategy C)",
        "C": "String"
      },

      "outcome_info": {
        "outcome_id": "String",
        "outcome_name": "String",
        "outcome_type": "String (Primary/Secondary/Safety)",
        "heterogeneity": { "i2_score": "Number/NR",
                           "p_value_heterogeneity": "Number/NR" }
      },

      "estimate_info": {
        "estimate_id": "String",
        "is_primary_estimate": true/false,
        "description": "String",
        "subgroup": "String",
        "model_type": "String (e.g., ITT / Adjusted)",
        "time_point": "String"
      },

      "group_data": [
       { "group_name": "String", "role": "String", "sample_size": "Number/NR",
        "event_rate": "String" },
       { "group_name": "String", "role": "String", "sample_size": "Number/NR",
        "event_rate": "String" }
      ],

      "statistical_result": {
        "comparison_direction": "String",
        "effect_measure": "String",
        "effect_value": "Number/NR",
        "ci_lower": "Number/NR",
        "ci_upper": "Number/NR",
        "p_value": "String",
        "statistical_significance": "String"
      },

      "quote_source": "String"
    }
  ]
}
```

Execution Instructions:
Analysis: Please perform your reasoning and strategy selection analysis first in plain text.
Output: After the analysis, you must output the final JSON object strictly wrapped inside <json> and </json> tags.

Quality Appraisal (QA) Inference Prompts

1. QA-inf (System Prompt)

You are an **expert in evidence-based medicine and clinical research methodology**.
Your task is to evaluate the **methodological quality** of a study and **produce a structured methodological quality assessment in JSON format**.

Requirements

- Writing style must be **academic, objective, concise, and non-conversational**.
- The evaluation must focus **exclusively on methodological quality**.
- Do **not** introduce external knowledge or subjective speculation.

Output Constraints (IMPORTANT)

- You may include necessary analytical or explanatory text in your response.
- **However, your final response must include a strictly valid JSON object**.
- The JSON object must **exactly and strictly follow the structure below**:
- It must contain **only** the fields: "overall_evaluation", "strengths", "limitations", "conclusion".
- **No fields may be omitted**.
- **No additional fields are allowed**.
- The JSON itself must be **syntactically valid and directly machine-parsable**.

Required JSON Format (must be followed exactly)

```
{
  "overall_evaluation": "(One paragraph summarizing the study's
                         methodological rigor and overall reliability.)",
  "strengths": [
    "...",
    "...",
    "..."
  ],
  "limitations": [
    "...",
    "...",
    "..."
  ],
  "conclusion": "(One paragraph summarizing the methodological
                 credibility of the study.)"
}
```

---

2. QA-inf (User Prompt)

Your task is to conduct a systematic appraisal of the methodological quality of the provided medical research study.
At the beginning of the evaluation, you should first determine the type of study based on information explicitly reported in the document (for example, a randomized controlled trial, an observational study, a systematic review, or another analytical study design). This determination does not need to be reported separately as part of the structured evaluation output.
Based on this understanding, you should assess the methodological rigor of the study in light of the general characteristics of the identified study type. In doing so, you may consider methodological aspects relevant to that design, such as whether the study design is appropriate, whether the selection and definition of the study population are clear, and whether the measurement of exposures or interventions and outcomes is appropriate.
Throughout the evaluation, you should not infer, assume, or supplement any aspects of the study design, conduct, or analysis that are not clearly described in the document.

Finally, you should synthesize your assessment into a structured JSON output that summarizes the study's overall methodological reliability, key methodological strengths, major limitations, and an overall conclusion.
— Start of Document —
{literature_content}
— End of Document —

Multi-evidence Inductive Reasoning (MEI) Inference Prompts

You are an expert clinical guideline developer and evidence-based medicine methodologist.
Your task is to generate a clinical guideline recommendation under the GRADE framework based on the provided structured evidence (evi_list).
You must perform explicit, traceable evidence-based reasoning and assign a corresponding GRADE recommendation.
==========
GRADE Methodological Background
==========
GRADE is a framework for evidence evaluation and recommendation grading used in clinical guideline development. Its core process includes:
1. Defining the clinical question
The clinical question must clearly specify the Population (P), Intervention (I), Comparator (C), and Outcomes (O). Outcomes are categorized by their importance for clinical decision-making into critical outcomes, important but non-critical outcomes, and unimportant outcomes. Critical outcomes play a decisive role in evidence quality assessment and recommendation strength.
2. Collecting and integrating evidence
GRADE emphasizes assessment of the body of evidence as a whole, rather than individual studies. Whenever possible, priority should be given to systematic reviews and well-designed randomized controlled trials. When high-quality systematic reviews are unavailable, all relevant studies should be jointly considered to judge evidence reliability.
3. Assessing evidence quality
Evidence quality is classified into four levels: A (High), B (Moderate), C (Low), and D (Very Low). Randomized controlled trials generally start at a higher level, while observational studies start at a lower level. Evidence quality may be downgraded due to risk of bias, inconsistency, indirectness, imprecision, or publication bias; it may be upgraded in the presence of large effects, a clear dose–response relationship, or when all plausible biases would reduce the observed effect.
4. Determining recommendation strength
Recommendation strength is classified as strong (1) or weak/conditional (2). It depends on the balance of benefits and harms for critical outcomes and the credibility of the evidence, and is typically determined by the lowest-quality evidence among the critical outcomes.
The final GRADE recommendation must be expressed as a combination of recommendation strength and evidence quality.
==========
GRADE Level Definitions (MUST be strictly followed)
==========
I. Evidence Quality (LETTER):
- A (High): Very confident that the observed effect is close to the true effect.
- B (Moderate): Moderately confident; the observed effect is likely close to the true effect, but a substantial difference is possible.
- C (Low): Limited confidence; the observed effect may be substantially different from the true effect.

- D (Very Low): Very little confidence; the observed effect is likely to be substantially different from the true effect.
II. Recommendation Strength (NUMBER):
- 1 (Strong recommendation): Clear evidence shows that benefits outweigh harms, or harms outweigh benefits.
- 2 (Weak/Conditional recommendation): Benefits and harms are uncertain or closely balanced; even with relatively high-quality evidence, a strong recommendation cannot be made.
III. Final GRADE Format (MANDATORY):
The GRADE level MUST be reported in the combined format: {Recommendation Strength}{Evidence Quality}
Examples: 1A, 1B, 2C, 2D.
No other formats are allowed. Do NOT use expressions such as "high/moderate/low quality" or "strong/weak recommendation". The ONLY valid grade outputs are: 1A, 1B, 2C, 2D, etc.
==========
Output Requirements (STRICT)
==========
You may write free-form analysis, but your response MUST end with a JSON object wrapped in <json></json> tags.
The JSON MUST contain EXACTLY the following four fields (no more, no less):

```
{
  "recommendation": "...",
  "grade": "...",
  "recommendation_explanation": "...",
  "reasoning": "..."
}
```

Field definitions:
- recommendation: the final guideline recommendation statement.
- grade: the GRADE level, using the mandatory combined format (e.g., 1A, 2B).
- recommendation_explanation: a guideline-style explanation intended for clinicians.
- reasoning: explicit evidence-based reasoning, including assessment of evidence quality, critical outcomes, consistency, magnitude of effects, and benefit–harm balance.
All judgments MUST be based strictly on the provided evidence (evi_list). Do NOT introduce any external knowledge, guideline conclusions, or subjective speculation.

PICO Extraction (PE) Evaluation Prompts

STAGE 1: SYSTEM PROMPT
Role: Senior Clinical Evidence Auditor - Tuple Matching Specialist.
Task: Match Golden tuples with Predicted tuples based on **5-Coordinate Clinical Semantic Alignment**.
Step 0: Strategy Gate
Check meta.assigned_strategy.
* If Golden != Predicted: Output {"file_status": "STRATEGY_FAILED", "tuple_matches": []}. **STOP HERE.**
* If Match: Output {"file_status": "STRATEGY_PASS"}. **PROCEED.**
Step 1: Tuple Matching (5-Coordinate Key) - ONE-TO-ONE MATCHING ONLY
Iterate through every **Golden Tuple**.
(Note: Golden Tuples represent extracted estimates. If an outcome has no estimates or a tuple has no valid data, skip it).
CRITICAL: ONE-TO-ONE MATCHING RULE

- Each Golden tuple can match **ONLY ONE** Predicted tuple.
- Each Predicted tuple can be matched by **ONLY ONE** Golden tuple.
- If multiple Golden tuples could match the same Predicted tuple, choose the **MOST CLOSELY MATCHING** one (highest semantic similarity).
- If a Predicted tuple is already matched by another Golden tuple, it cannot be matched again.

Matching Strategy:

Find the **BEST matching** Predicted tuple for each Golden tuple. If multiple candidates exist, select the one with the **highest semantic similarity**.

For each Golden tuple, attempt to find the **BEST MATCHING** tuple in Predicted using the **5-Coordinate Key**. Use **semantic matching**, NOT exact string equivalence.

5-Coordinate Matching Rules:

1. **P (Population):** pico_context.P semantically matches estimate_info.subgroup or P.
* Wildcard Rule: If Golden P is generic (e.g., "Overall", "All Patients", "Total population"), it matches ANY population in Predicted.
2. **I (Intervention):** **PRIORITIZE pico_context.I.**
* Check if the core intervention concept described in Golden's pico_context.I semantically matches Predicted's pico_context.I.
* Fallback: Only if pico_context.I is unclear or missing, check group_data[].group_name where role indicates "Intervention" or "Exposure".
3. **C (Comparator):** **PRIORITIZE pico_context.C.**
* Check if the core comparator concept described in Golden's pico_context.C semantically matches Predicted's pico_context.C.
* Fallback: Only if pico_context.C is unclear or missing, check group_data[].group_name where role indicates "Control" or "Comparator".
4. **O (Outcome):** outcome_info.outcome_name semantically matches.
* Synonym Examples: "Death" == "Mortality", "CVD" == "Cardiovascular events", "CVD Incidence" matches "Population-attributable risk (PAR) of CVD incidence", "Stroke incidence" matches "Stroke", etc.
* Focus on the **core outcome concept**, ignore descriptive modifiers.
5. **E (Estimate Context):** estimate_info.time_point AND estimate_info.model_type semantically match.
* Examples: "Up to 20 years" matches "Up to 20 years follow-up", "Adjusted" matches "Adjusted (multivariate Cox proportional hazards model)", etc.
* Null Rule: If Golden field is "NR" or null, it matches **ANYTHING** in Predicted (wildcard).

Important Matching Principles:

- Use semantic/synonym matching, NOT exact string matching.
- If core concepts match despite different wording, consider it a match.
- Prioritize pico_context fields over group_data extraction.
- **When multiple candidates exist, choose the one with the highest overall semantic similarity across all 5 coordinates.**

Matching Process:

1. For each Golden tuple, evaluate ALL Predicted tuples and identify potential matches.
2. If multiple Predicted tuples match, rank them by matching quality (more matching coordinates = better).
3. Assign the Golden tuple to the **BEST MATCHING** Predicted tuple that is **NOT YET ASSIGNED**.
4. If no unassigned Predicted tuple matches well enough, mark as MISSING.
* If **NOT FOUND** or **NO GOOD MATCH**: Mark as {"golden_trace_id": "P1-O5-E1", "predicted_trace_id": "MISSING"}.
* If **FOUND**: Mark as {"golden_trace_id": "P1-O5-E1", "predicted_trace_id": "P1-O6-E1"}.

Output Format (JSON)

```
{
  "file_status": "STRATEGY_PASS",
  "tuple_matches": [
    {"golden_trace_id": "P1-O5-E1", "predicted_trace_id": "P1-O6-E1"},
    {"golden_trace_id": "P1-O1-E1", "predicted_trace_id": "P1-O1-E1"},
    {"golden_trace_id": "P1-O2-E1", "predicted_trace_id": "MISSING"}
  ]
}
```

STAGE 2: SYSTEM PROMPT

Role: Senior Clinical Evidence Auditor - Field Verification Specialist.

Task: Verify field-level correctness for matched tuple pairs.

Field Verification Protocol

You will receive a queue of matched tuple pairs. For each pair, compare the **Payload Fields** (excluding the 5 coordinates used for matching).

For each field below, output status:

* **IGNORE:** If Golden value is "NR" or null.
* **CORRECT:**
- Numbers (sample_size, event_rate): Exact match.
- Floats (effect_value, ci_lower, ci_upper, p_value, i2_score): Tolerance ± 0.02.
- Strings (outcome_type, significance, effect_measure): Semantic match.
* **WRONG:** Predicted value differs OR is missing/NR while Golden has value.

Fields to Check:

1. **Context:** study_design_specifics, outcome_type, is_primary_estimate, description.
2. **Outcome:** heterogeneity.i2_score, heterogeneity.p_value_heterogeneity.
3. **Groups:** Check sample_size, event_rate, role for BOTH Intervention and Control groups (iterate through group_data list).
4. **Stats:** comparison_direction, effect_measure, effect_value, ci_lower, ci_upper, p_value, statistical_significance.
5. **Source:** quote_source.

Output Format (JSON Log)

```
{
  "audit_logs": [
    {
      "golden_trace_id": "P1-O1-E1",
      "predicted_trace_id": "P1-O1-E1",
      "tuple_status": "MATCHED",
      "field_results": {
        "study_design_specifics": "CORRECT",
        "outcome_type": "CORRECT",
        "sample_size": "WRONG",
        "effect_value": "CORRECT",
        "ci_lower": "IGNORE",
        "heterogeneity_i2": "CORRECT"
      }
    }
  ]
}
```

Important Notes:
- For group_data, check each group object separately. Use field names like group_0_sample_size, group_1_sample_size if needed.
- Only output fields that exist in Golden (non-NR, non-null).
- Output ONLY valid JSON, no additional text or markdown.
- Process ALL tuple pairs in the queue.

Quality Appraisal (QA) Evaluation Prompts

1. QA-eval (System Prompt)
You are a Methodological Quality Judge.
Follow all evaluation rules, scoring criteria, and instructions provided entirely in the user message.
You must output strictly valid JSON only, with the exact required fields and no extra text.

---

2. QA-eval (User Prompt)
Below is all information needed to evaluate the **evidence coverage quality** of a model-generated appraisal.
FULL SCORING RULES — PLEASE READ CAREFULLY
Role
You are an **Evidence Coverage Judge** who strictly follows principles of evidence-based medicine and research methodology.
Your sole responsibility is to **objectively and precisely evaluate how well the model output covers the evidence specified in the Gold Standard**.
Objective
Your evaluation must compare **three sources**:
1. **Original Article** (the only source of factual truth)
2. **Gold Standard** (authoritative list of true strengths and limitations)
3. **Model Output** (to be evaluated; only the JSON object in the output should be considered)
You must assess **only one dimension**: **Evidence Coverage (0–10)**.
Final score range: **0–10 (integer)**.
Scope and Constraints
You MUST rely only on: Original article content, Gold Standard strengths and limitations, and Model output content.
You MUST NOT: Use external knowledge, guess or infer missing information, evaluate factual correctness beyond what is required for coverage matching, or introduce any additional scoring dimensions.
Domain Knowledge Context (For Semantic Grounding Only)
The following content provides domain-specific definitions and evaluation conventions. It is intended **only** to help interpret the semantic meaning of evidence items when determining whether the model output correctly covers Gold Standard entries.
You MUST NOT: Use this section to introduce new evidence, override the scoring rules below, or relax hit/miss validity requirements.
{{DOMAIN_KNOWLEDGE}}
EVIDENCE COVERAGE — SCORING RULES
1. Evidence Coverage (0–10)
This dimension evaluates how comprehensively the model output captures **both strengths and limitations** defined in the Gold Standard, using a **weighted coverage system**.
1.1 Weighting Rule
Each evidence item has a base weight. Items marked as **critical** in the Gold Standard carry higher weight.

- Regular Strength: Weight 1
- Critical Strength: Weight 2
- Regular Limitation: Weight 2
- Critical Limitation: Weight 3

Define:
- S_total_regular: number of regular strengths
- S_total_critical: number of critical strengths
- L_total_regular: number of regular limitations
- L_total_critical: number of critical limitations
- S_hit_regular: regular strengths correctly captured by the model
- S_hit_critical: critical strengths correctly captured by the model
- L_hit_regular: regular limitations correctly captured by the model
- L_hit_critical: critical limitations correctly captured by the model

1.2 Maximum Possible Evidence Points

max_points = S_total_regular * 1 + S_total_critical * 2 + L_total_regular * 2 + L_total_critical * 3

1.3 Model-Captured Evidence Points

hit_points = S_hit_regular * 1 + S_hit_critical * 2 + L_hit_regular * 2 + L_hit_critical * 3

1.4 Coverage Rate

coverage_rate = hit_points / max_points

1.5 Final Evidence Coverage Score (0–10)

evidence_coverage_score = round(coverage_rate * 10)

The final score is an **integer between 0 and 10**, directly reflecting weighted evidence coverage.

1.6 Validity Requirements (Critical)

An evidence item is counted as **"hit"** only if **all** conditions are satisfied:
- It is explicitly supported by the **Original Article**.
- It matches the **semantic meaning** of the corresponding item in the Gold Standard.

You MUST NOT: Count fabricated or hallucinated content, count items not supported by the article, reverse direction (e.g., treat a limitation as a strength), or merge multiple Gold Standard items into one vague statement.

Output Format (STRICT)

You MUST output JSON in **exactly** the following structure. No additional fields, no extra text.

```
{
  "evidence_coverage": {
    "score": int,
    "justification": "..."
  }
}
```

ORIGINAL DOCUMENT

{{MD_CONTENT}}

GOLD STANDARD

{{GOLDEN}}

MODEL OUTPUT TO BE EVALUATED

{{MODEL_OUTPUT}}

Your Task

Evaluate the model output **strictly according to the Evidence Coverage rules above**, and output the required JSON **without deviation**.

Multi-evidence Inductive Reasoning (MEI) Evaluation Prompts

You are an expert in clinical guideline development and evidence-based medicine methodology.
Your task is to determine whether a **model-generated recommendation statement** is semantically consistent with the **gold recommendation** at two distinct but related levels.
The input will include: question_id, sub_question_text (the clinical question), gold_recommendation (the reference recommendation), and pred_recommendation (the model-generated recommendation).
You must produce **two interrelated judgments with different decision criteria**:

---

1. strict_correct (Strict Semantic Consistency)
This judgment determines whether the model-generated recommendation is **fully semantically equivalent** to the gold recommendation.
Decision rules (must be followed strictly):
* Evaluate only whether the **recommendation conclusion itself** is consistent, including but not limited to: target population, thresholds, direction, strength, and applicability conditions.
* Focus on **coverage of the gold recommendation**. Additional broad or common-sense advice is not considered an error, provided the gold recommendation is correctly covered.
* If there is any inconsistency in **key numerical thresholds, ranges, recommendation direction (recommend vs. not recommend), target population, or applicability conditions**, the judgment must be 0.
* If the prediction differs only in wording but is **fully equivalent in meaning**, assign 1.
* For recommendation accuracy, **"not recommended" and "not recommended due to insufficient evidence" are considered equivalent**.
* Do **not** consider reasoning processes or explanations; judge only whether the recommendation content is semantically equivalent to the gold recommendation.

---

2. direction_correct (Directional Consistency)
This judgment determines whether the model-generated recommendation is consistent with the gold recommendation **in recommendation direction**.
Decision rules (must be followed strictly):
* If the recommendation direction of pred is opposite to that of gold, assign 0.
* direction_correct = 1 does **not** imply the recommendation is fully correct; it only indicates that the direction does not deviate from the gold recommendation.
* Evaluate only the **direction of recommendation**, such as whether to recommend, not recommend, or initiate a certain type of intervention.
* Ignore differences in population restrictions, subgroup conditions, or additional constraints.
* If the recommendation direction is consistent, assign 1 **even if pred omits or incorrectly specifies certain conditions**.

Consistency Constraint
* If strict_correct = 1, then direction_correct **must** be 1.
* The case where strict_correct = 1 and direction_correct = 0 is **not allowed**.
Output Requirements (Strict)
* Your response must end with <json></json> and contain **only one JSON object**.
* The JSON object must contain **exactly** the following three fields:

```
{
  "strict_correct": 0 or 1,
  "direction_correct": 0 or 1,
  "rationale": "one-sentence justification"
```

```
}
```

Literature Including/Exclusion One

Question
Under what circumstances is drug washout required and ARR rescreening performed?
document number: 282
Golden inclusion: 1
Reasons for Recommendation:
This study is highly relevant as it directly addresses a critical gap in clinical practice: the severe underutilization of Primary Aldosteronism (PA) screening in patients with Obstructive Sleep Apnea (OSA).
It provides compelling real-world evidence from a massive cohort (over 120,000 patients) demonstrating that despite clear Endocrine Society guidelines, less than 4% of eligible OSA patients are screened. The study serves as a vital alert for clinicians, highlighting that PA is significantly underdiagnosed in this high-risk group and that improved screening could lead to targeted treatments that benefit both hypertension and sleep apnea severity.

Model Response
Inclusion: 0 ✘
Recommendation: ✘
...
The model **rejected** this literature (Prediction: 0), citing that it "does not address PA diagnostic **medication washout or ARR rescreening**." Furthermore, the model erroneously claimed that the "**abstract content is unrelated to the title and provides no relevant PICO information.**" It failed to identify the study's direct relevance to PA screening practices, dismissing valid epidemiological evidence based on narrow, technical exclusion criteria that were likely inapplicable to the broader research question.
...

Error Analysis
The model committed errors of **"Hallucinated Irrelevance" and "Criterion Over-Specification."** It falsely claimed a disconnect between the title and abstract, while rigidly rejecting the paper for lacking niche technical details (e.g., "medication washout") irrelevant to the study's epidemiological focus. Furthermore, the assertion of "no relevant PICO information" is objectively incorrect, as the study clearly defines all core clinical elements.

Figure 8: Literature inclusion error case study by GPT-5.2. The model incorrectly rejects a relevant study due to **hallucinated irrelevance** and **criterion over-specification** (bolded). The specific error categories—falsely claiming "no relevant PICO information" while demanding irrelevant technical details—are highlighted in **red** within the error analysis.

Literature Including/Exclusion Two

Question
Which hypertensive patients are recommended to receive initial combination antihypertensive drug therapy?
document number: 195
Golden inclusion: 1
Reasons for Recommendation:
This review is essential for managing elderly hypertension, providing evidence-based guidance on initial combination therapy tailored to aging-related pharmacokinetic changes.
It explicitly advocates for low-dose fixed combinations (e.g., quarter-dose quadruple therapy) to achieve strict blood pressure targets safely while minimizing adverse events. This directly answers "which patients should start with initial combination therapy"—specifically, elderly patients requiring effective yet tolerable regimens.

Model Response
Inclusion: 0 ✘
Recommendation: ✘
...
The model **rejected** this literature (Prediction: 0), reasoning that the review "does not address which patients should start with initial combination therapy" and claiming it focuses solely on "low-dose initiation and titration." This assessment is factually incorrect. The model failed to recognize that the paper explicitly advocates for low-dose combination therapy(e.g., citing the quarter-dose quadruple therapy trial) as a specific strategy for the elderly population. By treating **"low-dose initiation" and "initial combination therapy"** as mutually exclusive concepts, the model missed the core recommendation that elderly patients are the specific group for whom this low-dose combination approach is indicated.
...

Error Analysis
The model committed errors of **"Hallucinated Irrelevance" and "Criterion Over-Specification."** It falsely claimed a disconnect between the title and abstract, while rigidly rejecting the paper for lacking niche technical details (e.g., "medication washout") irrelevant to the study's epidemiological focus. Furthermore, the assertion of "no relevant PICO information" is objectively incorrect, as the study clearly defines all core clinical elements.

Figure 9: Literature inclusion error case study by GPT-5.2. The model incorrectly rejects a relevant review due to a **concept separation error** (bolded), failing to recognize that the paper's recommendation for "low-dose initiation" in the elderly specifically includes **initial combination therapy**. The specific oversight—ignoring explicit citations of fixed-dose combination trials—is highlighted in **red** within the error analysis

Literature Including/Exclusion Three

Question
For patients with renal artery stenosis (RAS) after successful interventional therapy, what is the recommended use of RASI?
document number: 332
Golden inclusion: 1
Reasons for Recommendation:
This Cochrane review is essential as it aggregates high-quality RCT data defining the standard of care for Renal Artery Stenosis (RAS).
While primarily comparing intervention vs. medication, the detailed medical therapy protocols (often including ACEIs/ARBs) within these trials provide the foundational evidence for pharmacological management. For clinicians, these protocols implicitly establish the safety and efficacy baselines for RASI use, which remain relevant even after interventional therapy, making this a key reference for establishing "recommended use."

Model Response
Inclusion: 0 ✘
Recommendation: ✘
...
The model **rejected** this literature (Prediction: 0), asserting that it "does not address RAS inhibitor (ACEI/ARB) use after successful interventional therapy" and focuses solely on the comparison of intervention versus medical therapy. This assessment fundamentally misinterprets the utility of the review. The model failed to recognize that the pharmacological regimens detailed in the "medical therapy" arms of these trials constitute the very evidence base that informs post-procedural medication guidelines, treating the two study arms as mutually exclusive rather than complementary components of patient care.
...

Error Analysis
The model committed errors of **"Indirect Evidence Neglect" and "Contextual Blindness."**
It rigidly excluded the review for lacking a section explicitly titled "post-intervention RASI use," failing to infer that the standardized medical regimens (including ACEIs) described in the trials define the pharmacological standard of care applicable post-procedure. The model ignored the implicit value of these protocols, treating the "medical therapy" comparison arm as irrelevant to the user's question about medication management.

Figure 10: Literature inclusion error case study by GPT-5.2. The model incorrectly rejects a Cochrane review due to **indirect evidence neglect** and
textbfcontextual blindness (bolded). It fails to recognize that the standardized medical protocols (including RASI) detailed in the trials provide the foundational evidence for post-interventional pharmacological management. The specific oversight is highlighted in
textcolorred
textbfred within the error analysis.

Literature Including/Exclusion Four

Question
What are the preferred antidiabetic drugs recommended to improve prognosis in patients with hypertension and type 2 diabetes?
document number: 226
Golden inclusion: 1
Reasons for Recommendation:
This guideline is highly relevant for managing patients with both hypertension and T2D. It explicitly recommends preferred drug classes (e.g., ACEIs/ARBs) that optimize outcomes for this comorbidity and emphasizes holistic risk reduction, including statin therapy (citing HOPE-3) to improve cardiovascular prognosis. This directly answers the user's query by establishing the "preferred" pharmacological strategy for improving prognosis in these high-risk patients.

Model Response
Inclusion: 0 ✘
Recommendation: ✘
...
The model **rejected** this literature (Prediction: 0), asserting that it is merely a "hypertension management guideline" that "does not address antidiabetic drugs or prognosis-related selection of glucose-lowering therapy." This assessment is reductionist. While the title focuses on hypertension, the content explicitly discusses the management of comorbidities, including the selection of antihypertensive agents (ACEIs/ARBs) preferred in metabolic conditions and the use of adjunctive therapies like statins (citing HOPE-3) to improve cardiovascular prognosis in high-risk groups, which fundamentally overlaps with the user's question about prognosis in T2D and hypertension.
...

Error Analysis
The model committed errors of **"Domain Rigidness" and "Holistic Neglect."** It rigidly classified the document as solely "hypertension-focused," failing to recognize that guidelines for comorbid conditions (HTN + T2D) inherently address the "preferred" therapies (both antihypertensive and adjunctive like statins) that improve overall prognosis. The model ignored the guideline's explicit recommendations on comorbidity management and risk reduction strategies (e.g., statins, ACEIs/ARBs) that directly answer the prognosis aspect of the prompt.

Figure 11: Literature inclusion error case study by GPT-5.2. The model incorrectly rejects a relevant guideline due to **domain rigidness** and **holistic neglect** (bolded). It fails to recognize that hypertension guidelines for comorbid conditions provide critical evidence on "preferred" therapies (e.g., ACEIs/ARBs, statins) that improve prognosis in patients with T2D. The specific oversight is highlighted in **red** within the error analysis.

Case Study: Strategy A-Meta-Analysis PICO Extraction

Ground Truth

For the evidence tuple P1-O2-E1 (Low HDL cholesterol and cardiovascular mortality in older women):
Sample Size: 50,266
Outcome Type: Secondary
P-value: 0.003
Study Design Specifics: Specifically defined as a Meta-analysis with pooled relative risk estimates and reported heterogeneity metrics (I^2).

Model Response ✘

In the matched predicted tuple P1-O10-E1:
Sample Size: Output as "NR" (Not Reported).
Outcome Type: Misclassified as "Primary".
P-value: Output as "NR" (Not Reported).
Study Design Specifics: Overly verbose output that incorrectly included follow-up medians and demographic percentages, failing to align semantically with the concise Golden definition.

Error Analysis

Omission of Metadata (Sample Size and P-value):
While the model accurately captured the core statistical estimates (RR=1.34, 95% CI: 1.03–1.74), it failed to extract the total sample size (50,266) and the specific P-value (0.003) clearly present in the text, marking them as "NR." This indicates that the model prioritizes the extraction of "effect size pairs" but lacks the sensitivity to capture surrounding descriptive metadata and exact probability values in complex meta-analytical reporting.
Interference of Prior Knowledge (Outcome Type):
In general clinical contexts, mortality is typically categorized as a primary outcome. The model's classification of this outcome as "Primary" is likely a result of Prior Bias (Internal Knowledge Hallucination), where the model relied on general clinical intuition rather than the specific secondary-outcome designation defined by the authors of the paper.
Information Redundancy and Alignment Failure (Study Design):
The model exhibited "information overload" in the study design field, conflating structural design details with population and context parameters (e.g., 9-year follow-up and age thresholds). This lack of structural constraint led to a semantic misalignment during the audit phase, resulting in a "WRONG" status despite the presence of some correct information.
Data Localization Challenges:
As evidenced by the quote_source, the model successfully located the sentence containing the effect size. However, it failed to perform a broader scan of the document (such as tables or the PRISMA flow diagram) to aggregate the total sample size. In meta-analyses, the total "N" often requires summation across subgroups or extraction from summary tables, which currently represents a significant boundary for LLM reasoning.

Figure 12: Strategy A-Analysis PICO extraction error case study by GPT-5.2.

Case Study: Strategy B-RCT PICO Extraction

Ground Truth
For the PARAMOUNT trial evidence tuples:
Primary Outcome (P1-O1-E1): NT-proBNP reduction at 12 weeks.
Effect Size: 0.77 (95% CI: 0.64–0.92) as a "Ratio of change."
Sample Size: 149 (Intervention) and 152 (Control).
Secondary Outcome (P1-O3-E3): NYHA Class Improvement at 36 weeks.
Effect: Qualitative "Improvement" with a P-value of 0.006.

Model Response ✗
In the model's generated evidence tuples:
Missing Evidence (Recall Failure): The model completely failed to extract the tuple for NYHA Class Improvement (P1-O3-E3), resulting in a MISSING status in the audit.
Metadata Omission: For the primary outcome, the model failed to extract the Effect Value (0.77), CI, and Sample Sizes, marking them all as "NR" (Not Reported).
Statistical Logic Error: The model failed to identify the Effect Measure (Ratio of change), leaving it as "NR."

Error Analysis
Contextual Parsing Difficulty (Source Material Format): The model noted the document type as a "presentation/slide deck." In Strategy B (RCT), evidence is often presented in dense tables or graphical plots within slides. The model's failure to capture numerical effect sizes (0.77) and specific sample sizes (149/152)—despite successfully finding the P-values—suggests a structural scanning limitation. It can locate the "result" but struggles to parse the specific "data coordinates" associated with that result when they are not in a standard sentence format.
Recall Gap in Qualitative Outcomes:
The omission of the NYHA Class Improvement tuple highlights a weakness in capturing Qualitative Evidence. While the model focused heavily on quantitative endpoints (NT-proBNP, Left Atrial Volume), it ignored the broader clinical conclusions regarding functional class improvement. This is a critical failure in clinical guideline development where functional improvements are often "Important" or "Critical" outcomes.
Inference vs. Extraction Conflict (ITT):
Following the Strategy B instruction to prioritize Intention-to-Treat (ITT) analysis, the model correctly assumed ITT in its model_type description. However, because ITT was not explicitly labeled on the slides, the model remained conservative in its data extraction, contributing to the "NR" status for several key fields.
Trace ID and Outcome Mapping:
The model created a duplicate structure for Outcome 1 (E1 at 12 weeks, E2 at 36 weeks) and Outcome 2 (E1 at 12 weeks, E2 at 36 weeks). While this reflects the longitudinal nature of the trial, it diverged from the Golden Standard's more aggregated mapping, leading to matching complications during the automated evaluation phase.

Figure 13: Strategy Strategy B-RCT PICO extraction error case study by GPT-5.2.

Case Study: Strategy B-Parallel RCT PICO Extraction

Ground Truth

For the DAWN trial (Parallel RCT) evidence tuples:

Primary Outcome 1 (P1-O1-E1): Utility-weighted mRS at 90 days. Mean 5.5 (Intervention) vs 3.4 (Control), Mean Difference (MD) of 2.0 (95% CI: 1.1–3.0).

Primary Outcome 2 (P1-O2-E1): Functional independence (mRS 0-2) at 90 days. 49% vs 13%, Risk Difference (RD) of 33 (95% CI: 21–44).

Safety Outcomes (O9, O10): Neurologic deterioration (14% vs 26%, P=0.04) and Symptomatic intracranial hemorrhage (6% vs 3%, P=0.50).

Technical Outcome (O7): Immediate successful recanalization (TICI 2b/3) in the thrombectomy group (84% per central lab, 82% per local).

Model Response ✗

Omission of Procedural Detail (O7 MISSING): The model failed to extract the specific TICI recanalization rates (TICI 2b/3), which are critical technical secondary endpoints for thrombectomy trials.

Structural Alignment Errors: In 'study_design_specifics' and 'description', the model produced overly verbose narrative text that failed the automated audit's semantic check.

Field Accuracy: For Safety Outcomes (O9 matched to P1-O5-E1), the model misclassified the 'is_primary_estimate' flag as "Primary" instead of a safety-related descriptive estimate.

Error Analysis

Strategy B Execution (RCT Specifics):

The model demonstrated strong adherence to Strategy B (RCT) by correctly identifying the Intention-to-Treat (ITT) population and focusing on Bayesian adjusted models as required by the trial design. It accurately captured the core efficacy statistics (RD of 33 and MD of 2.0), which are the most complex numerical components. However, its recall (0.68) was limited by its neglect of "Technical Procedural" data. In interventional trials like this, models often overlook TICI or TIMI flow scores, viewing them as process markers rather than clinical outcomes.

Semantic Conflation vs. Precision:

The audit logs show a "WRONG" status for 'description' fields despite the model capturing correct facts. This is due to Information Conflation; the model combined Bayesian model assumptions and infarct volume adjustment details into the description field, whereas the Golden Standard separated these into specific metadata slots. This highlights a persistent challenge where LLMs struggle to modularize information into the exact schema constraints of SCEE (Structured Clinical Evidence Extraction).

Safety Outcome Priority:

The model successfully captured neurologic deterioration and intracranial hemorrhage. However, it failed to differentiate the "primary safety endpoint" from "descriptive safety events" (e.g., mislabeling primary flags). This indicates that while the model has improved in finding P-values for safety (P=0.04 and P=0.50), it lacks the methodological rigor to distinguish the hierarchical importance assigned to safety metrics in clinical protocols.

Figure 14: Strategy Strategy B-Parallel RCT PICO extraction error case study by GPT-5.2.

Case Study: Strategy C-Observational Study PICO Extraction

Ground Truth

For the retrospective case-control study on Primary Hyperaldosteronism (PA):
Study Design:A retrospective case-control study involving 62 Chinese hypertensive patients (45 cases of PA, 17 controls with Essential Hypertension).
Primary Evidence (P1-O1-E1/E2/E3): Diagnostic performance of the Aldosterone-to-Renin Ratio (ARR) under "morning seated" conditions.
AUC: 0.97 (95% CI: 0.88–1.00).
Sensitivity: 96.8% at a cutoff of 23.6.
Specificity: 94.1% at a cutoff of 23.6.
Sample Size: Consistently 45 (Cases) and 17 (Controls).

Model Response ✘

Sample Size Discontinuity: While the model correctly identified in the first tuple, it incorrectly marked sample sizes as "NR" (Not Reported) in subsequent tuples (E3, E4, etc.).
Role Mapping Error: The model assigned roles as "Disease present/absent" instead of the required "Intervention (Exposure)/Control" logic, causing a mismatch in the audit.
Tuple Explosion: The model extracted 20 separate tuples (capturing every possible measurement time and analyte), whereas the Golden Standard focused on the 4 primary diagnostic performance metrics.
Inconsistency in "Primary" Flags: The model mislabeled secondary analytes (like plasma renin) as primary estimates.

Error Analysis

Strategy C Execution (Observational Complexity):
The model successfully followed **Strategy C (OBS)** by identifying the study as observational and mapping ARR as the "Exposure." However, the model struggled with the **Diagnostic Accuracy Study** subtype of Strategy C. In these designs, the "Exposure" is the test result and the "Outcome" is the disease status. The model confused the final diagnosis with the intervention/control roles, demonstrating a lack of specialized reasoning for diagnostic case-control logic.
Context Loss in Multi-Tuple Extraction:
A significant failure occurred in **contextual persistence**. After identifying the total sample size at the beginning of the document, the model failed to "carry over" this value into subsequent tuples derived from the same population. This resulted in multiple "NR" entries for sample sizes in later evidence units, drastically reducing its field-level recall (0.40).
Greedy Extraction vs. Schema Alignment:
The model adopted a "greedy" strategy, extracting 20 tuples to ensure it didn't miss any data. While this captured many numerical values correctly (AUC 0.97, 0.99, etc.), it led to **Precision/Recall imbalance**. By focusing on every analyte (renin, aldosterone, ARR) at every time point (0900h, 1300h, pre-saline, post-saline), the model diluted the core clinical evidence, making the structured output less useful for guideline synthesis.

Figure 15: Strategy C-Observational Study PICO extraction error case study by GPT-5.2.

Case Study: Strategy C-Cross-sectional Study PICO Extraction

Ground Truth

For the analytical cross-sectional study on drug-drug interactions (DDIs) in Croatia:
Study Design: Analytical cross-sectional study conducted in community pharmacies, focusing on 265 elderly outpatients (65 years) with hypertension taking 2 drugs.
Core Evidence (O1, O4): The prevalence of potential clinically significant DDIs (CS-DDIs) was 90.6% (240/265 patients).
Mechanism Categorization (O2, O3): Mechanisms: Pharmacodynamic (60.9%) and Pharmacokinetic (39.1%).
Lexi-Interact Categories (by patient): Category C (97.9%), Category D (20.4%), and Category X (0.8%).
Specific Combinations (O5): High prevalence of NSAID + antihypertensives (n=149) and ACEI + diuretics (n=146).

Model Response ✗

Numerical Success but Metadata Failure: The model captured nearly all correct numerical percentages (90.6%, 60.9%, 97.9%, etc.) but failed the automated audit due to **schema misalignment**.
Missing Specific Combinations (O5 Missing): The model failed to extract the specific drug-pair tuples (e.g., NSAID + Antihypertensives), leading to a significant drop in recall (0.25).
Categorization Errors: The model failed to identify the correct **Effect Measure** (labeled as "NR" or descriptive text instead of "Prevalence") and the correct **Role** (failed to use "Exposure" or "Characteristic").
Over-verbosity: Similar to previous cases, the 'study_design_specifics' and 'description' fields were too narrative, causing "WRONG" judgments in the semantic audit.

Error Analysis

Strategy C Execution (Descriptive vs. Analytical): While the model correctly identified **Strategy C (OBS)**, it struggled with the "Descriptive" nature of this cross-sectional study. Strategy C prompts often emphasize "Adjusted Models," but in a prevalence-focused study, the priority is on the population denominator (265) and event numerator (240). The model successfully found these numbers but failed to categorize them under the correct "Prevalence" effect measure, treating them as generic descriptive results.
Granularity and "Tuple Explosion": The model attempted to structure the categories (C, D, X) but struggled with the hierarchy. In the Golden standard, these are separate tuples. The model's failure to capture 6 out of 16 tuples (Recall: 0.25) shows that it tends to "summarize" a table of results into a few dense tuples rather than breaking them down into the **Atomic Evidence Tuples** requested by the prompt. This "grouping bias" causes high-quality information to be lost during the matching phase.
Terminology Mismatch in Audit: The field stats show 0% recall for 'effect_measure' and 'comparison_direction'. This reflects a gap between clinical understanding and structured extraction. The model understands the result is a "rate," but does not explicitly output the term "Prevalence" required by the evidence schema, illustrating a failure in **Schema Adherence** rather than a failure in clinical comprehension.

Figure 16: Strategy C-Cross-sectional Study PICO extraction error case study by GPT-5.2.

Case Study: AMSTER_2 Quality Appraisal

Golden Appraisal

The overall confidence in this review is assessed as **critically low**.
This judgment is based on the fact that the review contains two critical flaws: there is "no mention of a pre-registered protocol or publicly available plan" (Item 2), and "no list of excluded studies with corresponding reasons was provided" (Item 7).
Furthermore, regarding reporting quality, "search strategies were not reported in detail... and no report of funding sources for the included primary studies" was found. Despite valid statistical methods, these critical omissions fundamentally compromise the review's reliability.

Model Response ✘

...
Methodologically, the review is **competently designed and executed**... using **appropriate study identification**, selection processes, QUADAS2 appraisal, and hierarchical meta-analytic techniques." While acknowledging limitations, the output states that "the credibility of the quantitative conclusions is limited by the low... methodological quality of included studies and sparse... evidence bases." It concludes that "overall, the review provides a **structured synthesis with appropriate methods**, but the strength of inference is moderate-to-low and chiefly constrained by the **underlying primary study limitations**.
...

Comparative Error Analysis

The model fails to detect the review's intrinsic methodological violations, resulting in a fundamentally incorrect quality assessment. Specifically, it validates the review as " **competently designed**" while overlooking the **absence of a pre-registered protocol (Item 2) and the missing list of excluded studies (Item 7)**—critical domains in AMSTAR 2 that define a review's trustworthiness. Instead of critiquing the conduct of the systematic review itself, the model misdirects the evaluation toward the " **quality of included studies**," thereby endorsing a review process that should have been flagged as critically flawed due to transparency and reporting deficits.

Figure 17: AMSTAR 2 quality appraisal error case study by GPT-5.2.The model incorrectly validates the systematic review as "competently designed" (**bolded**) by failing to detect critical methodological deficits. The specific overlooked critical domains—absence of a protocol (Item 2) and excluded studies list (Item 7)—are highlighted in **red** within the comparative analysis.

Case Study: ROB_2 Quality Appraisal

Golden Appraisal

The overall risk of bias is rated as **some concerns** because two domains were flagged. Specifically, regarding randomization, "the study does not report the method used to generate the random sequence... or allocation concealment procedures." In the deviations domain, "blindness was broken (participants and researchers were aware of the assigned rest time)," although "adherence to rest periods was high" and analysis methods were appropriate. However, the study demonstrated low risk elsewhere: "missing data proportion is approximately 2.6%... meeting the low risk criterion," outcomes were obtained using a "consistent automated method," and the study "has a valid clinical trial registration."

Model Response ✗

...

Methodologically, the trial is generally well-designed... leveraging randomization... with standardized, automated BP assessment." However, "randomization procedures are **incompletely described** (e.g., method of sequence generation, allocation concealment)." The evaluation criticizes that "missing data handling is per-protocol... **no intention-to-treat analysis**... are reported." It concludes that "confidence... is reduced by incomplete reporting of key randomization safeguards, **lack of explicit period/carryover assessment**, and the non-randomized placement of the second 5-minute reference condition.

...

Comparative Error Analysis

The model's assessment diverges significantly from the standard RoB 2 framework. While it correctly identifies randomization reporting issues, it fails to detect the **broken blindness (Deviations domain)**, a critical source of potential bias in open-label trials. Furthermore, the model incorrectly flags the exclusion of three participants as a limitation ("no intention-to-treat analysis"), whereas the correct appraisal classifies **Missing Data as Low Risk** (<5% missingness). Finally, the model overlooks the study's confirmed strengths in **Reporting Bias (valid registration) and analysis methods**, instead offering unfounded critiques regarding statistical approaches and period effects.

Figure 18: RoB 2 quality appraisal error case study by GPT-5.2. The model incorrectly critiques valid methodological aspects—specifically missing data handling and statistical analysis (**bolded**)—as limitations. Simultaneously, it overlooks the actual critical bias of broken blindness (Deviations domain), which is highlighted in **red** within the comparative analysis.

Case Study: JBI Quality Appraisal

Golden Appraisal

The study is deemed suitable for inclusion, demonstrating **strong methodological quality**. Key strengths include clearly defined "inclusion criteria... and exclusion criteria," detailed descriptions of the "study setting... and subjects' demographics," and valid "secondary HTN diagnosis uses objective standard criteria" (e.g., specific guidelines for PA, Cushing's, ADPKD). Confounding factors were explicitly identified and addressed using "binary logistic regression with stepwise backward elimination" and "missing data imputation." A minor limitation was noted regarding the "unclear validity/reliability of TFT measurement," but this did not significantly compromise overall rigor.

Model Response ✗

...

Methodological credibility is supported by the large EMR-based cohort, explicit diagnostic criteria... and transparent statistical procedures including handling of missing data." However, the evaluation heavily criticizes that "outcome ascertainment is susceptible to verification/ascertainment bias because confirmatory evaluation... was not uniformly applied." It incorrectly identifies "predictor modeling used stepwise backward elimination" as a methodological limitation yielding unstable selection. The model concludes that findings are "methodologically limited for definitive prevalence estimation" due to "incomplete and non-standardized diagnostic evaluation.

...

Comparative Error Analysis

The model's critique diverges from the standard JBI appraisal by misidentifying accepted analytical methods as flaws. Specifically, while the Golden Appraisal validates the use of **stepwise backward elimination** as an appropriate strategy to address confounding (Item 6), the model flags this same method as a **limitation** likely to yield "unstable variable selection." Furthermore, the model fails to acknowledge the study's strength in **detailed setting and subject descriptions (Item 2)**, a core JBI requirement. Instead, it overemphasizes verification bias inherent to retrospective designs, potentially undervaluing the study's utility within its stated context.

Figure 19: JBI critical appraisal error case study by GPT-5.2. The model incorrectly critiques the valid statistical strategy of stepwise backward elimination (**bolded**) as a limitation, contradicting standard appraisal which recognizes it as an appropriate adjustment method. Additionally, the model overlooks the study's verified strength in detailed setting/subject description (Item 2), highlighted in **red**within the comparative analysis.

Case Study: CASP_cohort Quality Appraisal

Golden Appraisal

The overall quality of the study is assessed as **low**, driven by uncertainty across six domains ("can't tell").

Specifically, cohort recruitment appropriateness could not be confirmed due to missing baseline data tables. Critical gaps included "no mention of blinding outcome assessors" (accurate outcome measurement) and "no info on loss to follow-up" for 3-month outcomes (follow-up adequacy).

Furthermore, the appraisal noted a lack of information regarding population and setting applicability beyond Japanese hospitals, and no conclusion on consistency with other evidence. Conversely, the study was strong in having a clear focus, accurate exposure measurement, and identified/controlled confounders.

Model Response ✗

...

Methodological credibility is moderate for identifying associations... The study benefits from explicit exposure and outcome definitions... and multivariable adjustment." The evaluation praises the "prespecified BP measurement schedule" and "explicit computation of multiple BP variability metrics." However, it identifies limitations such as the "retrospective observational design" precluding causal inference, "participant exclusions... may introduce selection bias," and "insufficiently detailed" BP measurement procedures (device type/conditions). It also flags "outcome event counts are low relative to the number of adjusted covariates" as a statistical concern.

...

Comparative Error Analysis

The model's assessment overlooks critical reporting deficiencies identified in the Golden Appraisal, leading to an overly optimistic "moderate" credibility rating instead of "low." Specifically, the model fails to detect the **lack of blinding for outcome assessors** and missing information on **loss to follow-up**, both pivotal for internal validity in cohort studies. While the model correctly identifies strengths in exposure measurement and confounding control, it erroneously focuses on "small event numbers" and "multiplicity" as primary limitations, rather than the fundamental **applicability and follow-up completeness issues** flagged by the CASP tool.

Figure 20: CASP cohort study appraisal error case study by GPT-5.2. The model incorrectly rates the study's credibility as "moderate" by focusing on statistical nuances like "small event numbers" (**bolded**). It fails to detect critical validity threats flagged by the CASP tool, specifically the lack of outcome assessor blinding and missing follow-up data, which are highlighted in **red** within the comparative analysis

Case Study: CASP_cc Quality Appraisal

Golden Appraisal

The study demonstrates **high overall reliability** with strong methodological validity across most domains.

The research question is clearly focused, and the case-control design is appropriate for studying genetic risk factors for stroke. Cases were rigorously defined using "MRI-confirmed cerebral stroke" with explicit inclusion/exclusion criteria, and controls were recruited from the "same geographic region and hospitals." Genetic exposures were measured with high accuracy, using "blinded investigators" and "100% concordance" in sequencing validation. Key confounders (sex, BMI, BP) were identified and controlled via multiple logistic regression, yielding precise and credible results.

Model Response ✗

...

Methodologically, the study provides a reasonably described candidate-gene case–control analysis... with strong outcome ascertainment." The model acknowledges clear definitions and rigorous genotyping (blinding/validation). However, it lists significant limitations: "limited adjustment for key stroke risk confounders... (e.g., smoking, diabetes, lipid levels... are not reported as measured or adjusted)." It also criticizes that "case and control selection is clinic-based... raising risk of selection bias," and flags potential issues with "incomplete control of confounding and population stratification.

...

Comparative Error Analysis

The model's assessment diverges from the standard appraisal by overemphasizing theoretical limitations while underappreciating the study's actual confounding control. Specifically, the Golden Appraisal validates that **key confounders (sex, BMI, BP, medication) were identified and controlled (Item 6)**. In contrast, the model incorrectly critiques the study for **"limited adjustment for key stroke risk confounders"** (e.g., smoking, lipids), treating the absence of these specific variables as a major flaw despite the study's adequate control of primary factors. Furthermore, the model raises concerns about **"clinic-based recruitment"** causing selection bias, whereas the standard appraisal accepts the hospital-based controls as appropriate for the source population.

Figure 21: CASP case-control study appraisal error case study by GPT-5.2. The model incorrectly critiques the study for limited adjustment for key stroke risk confounders (**bolded**) such as smoking and lipids, whereas the standard appraisal confirms that primary confounders were adequately controlled. Additionally, the model overstates the risk of selection bias from clinic-based recruitment, overlooking the verified strength of recruiting controls from the same source population (Item 4), which is highlighted in **red** within the comparative analysis.

Ignoring Critical Details

Question
What are the norms and recommendations regarding the specific methods of blood pressure measurement, such as pre-measurement preparation, body position, cuff selection, number of measurements, and special populations?
Golden Grade: 1C
Golden recommendation:
Measurement methods include:
...
(2) Choose an appropriate cuff; a wrist cuff is optional for individuals with an upper arm circumference **> 42 cm**;
...
(4) Measure at least twice each time and take the average; **if the difference between readings is > 10 mmHg, measure a third time and take the average of the last two readings.** For the first measurement, measure both arms and take the higher value;
...

Model Response
Grade: 1B ✘
Recommendation: ✘
...
Cuff selection: use appropriately sized cuff based on arm circumference (standard cuff 12-13 cm width for most adults; larger or smaller cuffs for arm circumference **>32 cm** or <27 cm); (4) Number of measurements: obtain 3 readings at 1-2 minute intervals and record the average of the last 2 readings; (5) Initial assessment: measure BP in both arms at first visit and use the higher value for reference;
...

Error Analysis
Strict consistency fails due to discrepancies in specific thresholds and conditions, including missing wrist-cuff arm-circumference limits, omission of the third-measurement requirement when inter-arm differences exceed $10\mathrm{mmHg}$, and mismatched atrial fibrillation measurement protocols.

Figure 22: Ignoring critical details Error Case by Claude 4.5 Sonnet. The omitted information is highlighted in **red**.

Ignoring Critical Details

Question
At what time in the evening is home blood pressure measurement recommended?
Golden Grade: 2D
Golden recommendation:
It is recommended to measure blood pressure in the evening before dinner; if this is not feasible, **measure it within 1 hour before bedtime**.

Model Response
Grade: 2D
Recommendation: ✘
For evening home blood pressure measurement in hypertensive patients, measurements may be taken either before dinner or at bedtime, depending on guideline preference and local practice...

Error Analysis
The model omitted the critical threshold of measuring blood pressure within 1 hour before bedtime as specified in the gold recommendation.

Figure 23: Ignoring critical details Error Case by Claude 4.5 Sonnet. The omitted information is highlighted in **red**.

Ignoring Critical Details

Question
What is the recommended blood pressure control target for patients with hypertension complicated by coronary heart disease?
Golden Grade: 2B
Golden recommendation: For patients with hypertension and coronary heart disease, the recommended blood pressure control target is <130/**80** mmHg.

Model Response
Grade: 2C ✘
Recommendation: ✘
...
These benefits support targeting systolic blood pressure to **less than 130 mm Hg** in this population ...

Error Analysis
The model's recommendation lacks the critical diastolic blood pressure target (<80 mmHg) specified in the gold standard.

Figure 24: Ignoring critical details Error Case by Claude 4.5 Sonnet. The omitted information is highlighted in **red**.

Ignoring Critical Details

Question
For non-dialysis patients with chronic kidney disease (CKD) and urinary protein > 300 mg/d, what is the recommended target value for blood pressure control?
Golden Grade: 2B
Golden recommendation: The recommended blood pressure control target for CKD non-dialysis patients with urine protein > 300 mg/day is <130/80 mmHg; **if tolerated, the systolic blood pressure can be further reduced to 120 mmHg**.

Model Response
Grade: 2C ✗
Recommendation: ✗
For non-dialysis patients with chronic kidney disease (CKD) and urinary protein excretion greater than 300 mg/d, a lower blood pressure target of less than 130/80 mmHg may be considered over the standard target of less than 140/90 mmHg to reduce the risk of kidney disease progression...

Error Analysis
The recommendation lacks the essential detail of further reducing systolic BP to 120 mmHg if tolerated as stated in the gold.

Figure 25: Ignoring critical details Error Case by Claude 4.5 Sonnet. The omitted information is highlighted in **red**.

Ignoring Critical Details

Question
For hypertensive patients with unilateral renal artery stenosis (RAS), what kind of monitoring is recommended as the basis for using RASI?
Golden Grade: 1C
Golden recommendation: In hypertensive patients with unilateral renal artery stenosis (RAS), the use of renin-angiotensin system inhibitors (RASI) is recommended **on the basis of close monitoring of urine output**, electrolytes, and serum creatinine.

Model Response
Grade: 1C
Recommendation: ✗
For hypertensive patients with unilateral renal artery stenosis receiving renin-angiotensin system inhibitors (ACE inhibitors or ARBs), regular monitoring of renal function (serum creatinine/eGFR) and serum potassium is strongly recommended...

Error Analysis
The model's recommendation lacks urine output monitoring (essential).

Figure 26: Ignoring critical details Error Case by Claude 4.5 Sonnet. The omitted information is highlighted in **red**.

Conditioning Errors

Question
In the management of hypertension, which blood pressure monitoring method is recommended as the first choice?
Golden Grade: 2C
Golden recommendation:
Measurement methods include:
In the management of hypertension, home blood pressure monitoring is recommended as the first choice; if conditions do not permit, management should be based on OBPM combined with 24-hour ambulatory blood pressure monitoring.

Model Response
Grade: 1B ✘
Recommendation: ✘
Ambulatory blood pressure monitoring (24-hour or nighttime measurements) is recommended as the preferred method for **cardiovascular risk assessment** in patients with hypertension...

Error Analysis
The model applies to cardiovascular risk assessment instead of hypertension management, leading to wrong preferred monitoring method.

Figure 27: Conditioning Error Case by Claude 4.5 Sonnet. The inconsistent conditional information is highlighted in **red**.

Conditioning Errors

Question
What is the recommended monitoring frequency for patients with well-controlled blood pressure?
Golden Grade: 2D
Golden recommendation:
Patients with well-controlled blood pressure are recommended to perform home blood pressure monitoring for 1 to 2 days per week.

Model Response
Grade: 2C ✘
Recommendation: ✘
For patients with well-controlled blood pressure on antihypertensive treatment, we suggest **performing 1-2 blood pressure measurements per week or per month for long-term monitoring**.

Error Analysis
The pred adds an unnecessary condition (on antihypertensive treatment), includes an incorrect frequency option (per month) and uses a different unit (measurements vs days), so strict consistency fails.

Figure 28: Conditioning Error Case by Claude 4.5 Sonnet. The inconsistent conditional information is highlighted in **red**.

Conditioning Errors

Question
Are beta-blockers recommended as first-line antihypertensive agents in adults with hypertension requiring pharmacological treatment, compared with placebo?
Golden Grade: 2C
Golden recommendation:
Beta-blockers are not recommended as first-line antihypertensive agents in adults with hypertension.

Model Response
Grade: 2C
Recommendation: ✘
In adults with hypertension **requiring pharmacological treatment**, beta-blockers are suggested NOT to be used as routine first-line antihypertensive agents when compared with placebo...

Error Analysis
The model applies an incorrect population (needing meds) vs gold's all adults with hypertension, causing context mismatch.

Figure 29: Conditioning Errors by Claude 4.5 Sonnet. The inconsistent conditional information is highlighted in **red**.

Conditioning Errors

Question
What is the recommended potassium intake for patients with hypertension?
Golden Grade: 2B
Golden recommendation: The recommended target for potassium intake is 3500–4700 mg/d.

Model Response
Grade: 2C ✘
Recommendation: ✘
For adults with hypertension and normal renal function, we suggest increasing dietary potassium intake to approximately 90 mmol/day (3500 mg/day) through potassium-rich foods...

Error Analysis
Model added unnecessary population restrictions and exclusions not present in the gold recommendation, causing context mismatch.

Figure 30: Conditioning Errors by Claude 4.5 Sonnet. The inconsistent conditional information is highlighted in **red**.

Conditioning Errors

Question
Is pharmacological management of hypertension by a nurse recommended compared to usual care in patients with hypertension in LMICs?
Golden Grade: 1B
Golden recommendation: Pharmacological management of hypertension by a nurse is strongly recommended over usual care in patients with hypertension in LMICs.

Model Response
Grade: 2C ✘
Recommendation: ✘
In adults with hypertension in low- and middle-income countries, pharmacological management **by trained nurses** may be considered as an alternative to usual physician-led care...

Error Analysis
Model added unnecessary population restrictions and exclusions not present in the gold recommendation, causing context mismatch.

Figure 31: Conditioning Errors by Claude 4.5 Sonnet. The inconsistent conditional information is highlighted in **red**.

Decision Path Compression

Question
What is the recommended target value for blood pressure control in non-dialysis CKD patients with urinary protein < 300 mg/d?
Golden Grade: 2B
Golden recommendation:
For non-dialysis patients with chronic kidney disease (CKD) and urine protein < 300 mg/d, the recommended blood pressure control target is < 140/90 mmHg. If tolerated, the systolic blood pressure can be further reduced to 130 mmHg.

Model Response
Grade: 2C ✘
Recommendation: ✘
In adults with non-dialysis CKD and urinary protein excretion <300 mg/day, we suggest using a standard blood pressure target of <140/90 mm Hg rather than routinely pursuing intensive targets (e.g., <130/80 mm Hg or systolic <120 mm Hg)

Error Analysis
The model omits the conditional step (if tolerated, reduce systolic to 130) and simplifies to not pursuing intensive targets.

Figure 32: Decision Path Compression Error Case by GPT-5.2.

Evidence–recommendation Confusion

Question
At what time point in the morning is home blood pressure measurement recommended?
Golden Grade: 1B
Golden recommendation:
It is recommended to measure blood pressure in the morning before taking medication, before breakfast, and after emptying the bladder.

Model Response
Grade: 2D ✘
Recommendation: ✘
No specific time point in the morning can be recommended for home blood pressure measurement based on the provided evidence.

Error Analysis
The model gives an opposite recommendation (no specific time vs. specific morning times) due to incorrect evidence-recommendation mapping.

Figure 33: Evidence–recommendation confusion Error Case by Claude 4.5 Sonnet.

Evidence–recommendation Confusion

Question
For patients with hypertension complicated by acute hemorrhagic stroke, what range is recommended to control the systolic blood pressure within during the acute phase?
Golden Grade: 2C
Golden recommendation: For patients with hypertension complicated with acute hemorrhagic stroke, it is recommended to administer antihypertensive treatment during the acute phase and control the systolic blood pressure within 130-140 mmHg.

Model Response
Grade: Insufficient evidence to assign GRADE level ✘
Recommendation: ✘
No recommendation can be made due to complete absence of evidence. Evidence from relevant clinical trials is required to determine appropriate systolic blood pressure targets for patients with hypertension complicated by acute hemorrhagic stroke.

Error Analysis
Model incorrectly claims no evidence to recommend, conflicting with the gold's evidence-based BP control recommendation.

Figure 34: Evidence–recommendation confusion Error Case by Claude 4.5 Sonnet.

Evidence–recommendation Confusion

Question
Are blood pressure lowering drugs recommended in patients with baseline systolic blood pressure 130–140 mmHg?
Golden Grade: 2A
Golden recommendation: Blood pressure lowering drugs are not recommended in patients with baseline systolic blood pressure 130–140 mmHg.

Model Response
Grade: 2B ✘
Recommendation: ✘
In patients with baseline systolic blood pressure 130-140 mmHg, we suggest the use of blood pressure lowering drugs to reduce cardiovascular events.

Error Analysis
Model's opposite recommendation for same population ignores risk-benefit trade-offs considered in gold's decision.

Figure 35: Evidence–recommendation confusion Error Case by Claude 4.5 Sonnet.

Evidence–recommendation Confusion

Question
Are loop diuretics recommended as first-line agents compared to placebo in individuals with hypertension?
Golden Grade: 2B
Golden recommendation: Loop diuretics are not recommended as first-line agents in individuals with hypertension (Insufficient evidence to support routine use).

Model Response
Grade: 2D ✘
Recommendation: ✘
Insufficient evidence exists to recommend for or against the use of loop diuretics as first-line agents compared to placebo in individuals with hypertension. Clinical decisions should be individualized and based on other available evidence sources not included in this assessment.

Error Analysis
The model incorrectly maps insufficient evidence to a neutral conclusion instead of the gold's clear 'not recommended'.

Figure 36: Evidence–recommendation confusion Error Case by Claude 4.5 Sonnet.

Evidence–recommendation Confusion

Question
Is a blood pressure target <140/90 mmHg recommended compared to a target of <150-160/95-105 mmHg in patients with hypertension who are >65 years old?
Golden Grade: 1B
Golden recommendation: A blood pressure target of <140/90 mmHg is recommended over <150-160/95-105 mmHg in hypertensive patients aged >65 years old, especially those with preserved organ function.

Model Response
Grade: 2C ✘
Recommendation: ✘
In older adults (65 years) with hypertension, either a blood pressure target of <140/90 mmHg or a target of <150-160/95-105 mmHg may be considered acceptable. If resources and patient tolerance permit, a lower target of <140/90 mmHg may be preferred given trends toward cardiovascular benefit, but this recommendation is conditional due to insufficient high-quality evidence demonstrating superiority of either target in this population.

Error Analysis
The model confuses insufficient high-quality evidence with recommendation strength, failing to clearly recommend <140/90 mmHg over higher targets.

Figure 37: Evidence–recommendation confusion Error Case by Claude 4.5 Sonnet.