MedSynth: Evaluating LLMs on Certainty of Evidence and Strength of Recommendations in Medical Evidence Synthesis

Anonymous ACL submission

Abstract

Recent advances in large language models (LLMs) have demonstrated their potential to support scientific research. Evidence synthesis is a critical task in medical research, requiring models to integrate heterogeneous evidence and derive verifiable conclusions under substantial uncertainty. However, existing benchmarks overlook the certainty of supporting evidence and the strength of resulting recommendations, both of which are essential for ensuring the correctness and trustworthiness of medical evidence synthesis. They also provide limited diagnostic insight into models' weaknesses during evidence synthesis. To address these limitations, we propose **MedSynth**, the first medical evidence synthesis benchmark that evaluates whether models can correctly assess the certainty of evidence and strength of recommendations when provided with the same source studies. Experimental results show that even the best-performing model, GPT-5.2, achieves an average score of only 44.59, indicating substantial room for improvement. We release MedSynth to help the community evaluate, diagnose, and improve LLMs, thereby advancing trustworthy and transparent scientific research. Data and code are available[1].

1 Introduction

Evidence synthesis is a critical task in medical research. By systematically integrating evidence from diverse research sources, researchers can derive verifiable conclusions under substantial uncertainty to inform clinical practice and policymaking (JPT et al., 2024). However, the rapid growth of published literature (Bornmann et al., 2021) has made this process increasingly complex and difficult to scale. As a result, automated tools for medical literature understanding and evidence synthesis have attracted growing attention (Scherbakov et al., 2025). Recently, large language models (LLMs) have shown strong capabilities in literature understanding (Liao et al., 2025) and reasoning (Guo et al., 2025), leading to the deployment of LLM-driven systems, such as Deep Research (OpenAI, 2025a), Elicit (Whitfield and Hofmann, 2023), and OpenEvidence (OpenEvidence, 2025), to improve the efficiency of evidence synthesis.

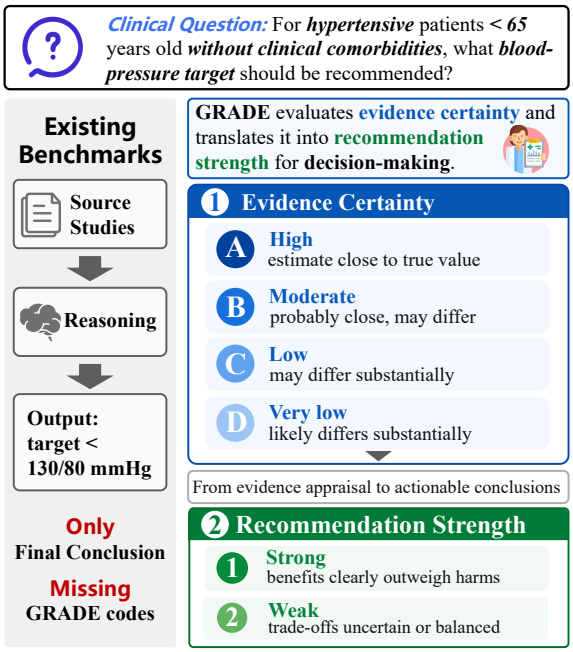

Figure 1: Example of evidence synthesis. Existing benchmarks mainly assess whether models reach final clinical conclusions, but leave evidence certainty and recommendation strength largely untested, although both are essential for evidence-driven reasoning.

Although these systems have been deployed in real-world applications, the reliability of LLMs in this setting remains insufficiently evaluated. MedREQAL (Vladika et al., 2024b) is an LLM-curated QA dataset for verifying Cochrane-derived claims, but since the review sources are not provided, it primarily tests models' internal factual recall. Recent work has provided models with ex-

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

Benchmark	Size	Phase	Q-Form	Input	M-doc	E-Cert.	R-Str.	Eval-Method
MedREQAL	2789	CES	MCQA	PM	✘	✘	✘	AutoM
PubMedQA	273.5k	CES	MCQA	ABS	✘	✘	✘	AutoM
HealthFC	750	LU+CES	MCQA	FT	✘	✘	✘	AutoM
Evidence Inference 2.0	12616	LU+CES	MCQA	FT	✘	✘	✘	AutoM
MedMeta	81	CES	OpenQA	ABS	✔	✘	✘	LLM-J
CochraneForest	923	CES	MCQA	FT/ABS	✔	✘	✘	AutoM
MedEvidence	284	CES	MCQA	FT/ABS	✔	✘	✘	AutoM
MedSynth (ours)	**649**	**LU+SQA+CES**	**OpenQA**	**FT/ABS**	✔	✔	✔	**LLM-J**

Table 1: Comparison of **MedSynth** and related benchmarks. **Q-Form**: Question.Form; **M-doc**: Multi-doc; **E-Cert.**: evidence certainty; **R-Str.**: recommendation strength; **OpenQA/MCQA**: free-form/option-based answers; **FT/ABS/PM**: full text/abstract/parametric memory; **LLM-J/AutoM**: LLM-as-a-judge/automatic-metric evaluation.

ternal evidence. Among these efforts, CochraneForest (Pronesti et al., 2025) examines how LLMs respond to conflicting evidence, whereas MedEvidence (Polzak et al., 2026) evaluates whether LLMs can reach conclusions consistent with clinical experts when given the same studies.

However, these studies primarily examine whether LLMs can arrive at the correct final conclusion, leaving two clinically critical questions unanswered: **Can models judge the certainty of the evidence underlying their reasoning?** and **Can they determine how strongly the resulting conclusion should be applied in clinical practice?** In scientific research settings such as evidence synthesis, LLMs often reason based on external evidence (Ghafarollahi and Buehler, 2025). As shown in Figure 1, which presents the GRADE (Grading of Recommendations Assessment, Development and Evaluation) framework (Goldet and Howick, 2013), the certainty of scientific evidence can vary substantially (JPT et al., 2024): factors such as study design, sample size, and effect estimates can all affect the reliability of evidence. Indiscriminate trust in evidence can undermine LLM reasoning (Jin et al., 2025). Moving from evidence to conclusions further requires assessing the applicability and actionability of those conclusions (Alonso-Coello et al., 2016).

At the same time, existing benchmarks provide limited diagnostic power for identifying gaps in model capabilities. As illustrated in Figure 2, real-world medical evidence synthesis is guided by methodological frameworks such as GRADE, which require (1) **systematic understanding** of heterogeneous literature, (2) **quality assessment**, and (3) **cross-study evidence synthesis** to support reliable and interpretable conclusions. Compared with this complex process, existing benchmarks exhibit clear limitations, as summarized in Table 1:

- **Limited characterization of literature understanding.** Researchers need to understand the literature and compare statistical effects across candidate interventions and multiple clinical outcomes to support cross-study comparison. EBM-NLP (Nye et al., 2018) evaluates models' ability to extract medical entities from short texts. PubMedQA (Jin et al., 2019) and Evidence Inference 2.0 (DeYoung et al., 2020) evaluate literature-based question answering, but variation in question difficulty makes it hard to isolate the effects of context length and contextual complexity.
- **Lack of study quality assessment.** Studies vary substantially in design and execution quality, making study quality assessment a prerequisite for evidence synthesis. CochraneForest (Pronesti et al., 2025) evaluates model performance under conflicting information, while RoBBR (Wang et al., 2025) assesses RoB1-based quality (JPT et al., 2024) outside synthesis.
- **Insufficient evaluation of the evidence synthesis process.** Existing benchmarks (Polzak et al., 2026; Wang and Chen, 2025; Ha et al., 2026) mostly use end-to-end evaluation focused on final answers. This design obscures models' evidence-to-recommendation reasoning, particularly their handling of evidence certainty and recommendation strength as distinct judgments.

To address these limitations, we construct **MedSynth**. To our knowledge, MedSynth is the first medical evidence synthesis benchmark that evaluates whether models can correctly assess the certainty of evidence and strength of recommendations when provided with the same source stud-

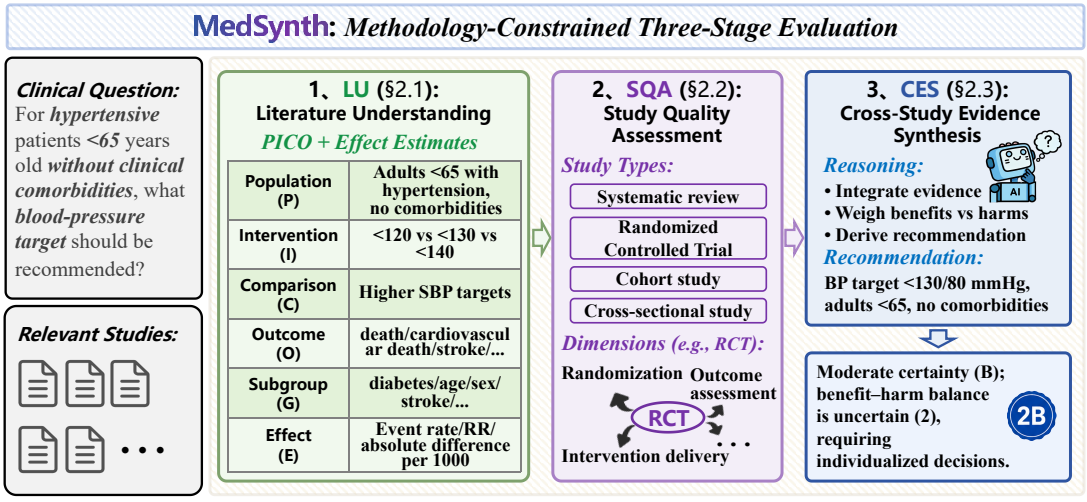

Figure 2: Illustrative Case of MedSynth for Hypertension Recommendation.

ies. Inspired by medical guideline development, MedSynth constrains evidence synthesis within the GRADE framework and models it as a three-stage process: (1) **literature understanding (LU)**, using document-level information extraction to assess whether models can comprehensively capture key findings and statistical effects from individual articles, with an average of 17.03 evidence items per article; (2) **study quality assessment (SQA)**, requiring models to identify study-type-specific factors that strengthen or weaken the credibility of conclusions, with an average of 12.39 factors per question; and (3) **cross-study evidence synthesis (CES)**, requiring models to synthesize up to 13 studies into clinical recommendations.

We conduct a large-scale evaluation of mainstream models. Results show that even the best-performing model achieves an average score of only 44.59, indicating that existing models still face substantial challenges in reliable medical evidence synthesis **(§3.2)**. Models struggle to generate clinical recommendations with fully specified conditions and sufficient detail; more importantly, they fail to accurately assess the certainty of the supporting evidence and assign appropriate recommendation strength **(§3.5)**. Our analysis suggests that literature understanding **(§3.3)** and study quality assessment **(§3.4)** are key bottlenecks in medical evidence synthesis. Further qualitative and quantitative analyses reveal four major failure mechanisms in models' evidence-driven reasoning **(§3.6)**.

In conclusion, our contributions are as follows: **Benchmark.** We construct MedSynth, an expert-curated benchmark built on high-quality annotation resources, with all code and redistributable data publicly available. **Evaluation.** We systematically evaluate mainstream models on MedSynth and analyze their capability gaps. **Insight.** Through quantitative and qualitative analyses, we identify four key error patterns that reveal the limitations of LLMs in reliable reasoning with external evidence and provide guidance for developing trustworthy artificial intelligence (AI) systems in high-stakes domains such as medicine and science.

2 Benchmark

GRADE-based clinical practice guideline resources provide a natural foundation for this evaluation, as they pair each clinical recommendation with a GRADE rating, a recommendation rationale, and the supporting literature used to formulate the recommendation; the interpretation of GRADE ratings is summarized in Figure 1. Building on these structured resources, **MedSynth** systematically examines three core capabilities: (1) **literature understanding**, (2) **study quality assessment**, and (3) **cross-study evidence synthesis**, with Figure 2 serving as a running example throughout the paper.

2.1 Literature Understanding (LU)

The LU task evaluates structured evidence extraction from medical literature: given a research document, the model must identify key findings and output them as evidence items. Each evidence item is represented as (P, I, C, O, G, E), where P denotes the patient population to which the result applies; I and C denote the compared interventions or strategies; O denotes the observed clinical outcome; G indicates the patient subgroup corresponding to the result; and E records the effect

estimates (including associated statistical information supporting the result).

As shown in Figure 2, for patients with hypertension (P), the comparison of different blood-pressure targets (I/C) for all-cause mortality (O) is reported as 111/2151 vs. 106/2164, with $RR = 1.06$ and a 95% CI of 0.82–1.37. These values (E) characterize the magnitude, direction, and uncertainty of the effect. Because multiple evidence items may share the same (P, I, C, O, G) prefix, we require models to merge them to reduce redundancy while preserving all relevant information. For example, two items (P, I, C, O, G, E_1) and (P, I, C, O, G, E_2) should be output as $(P, I, C, O, G, [E_1, E_2])$.

2.2 Study Quality Assessment (SQA)

The SQA task evaluates models' ability to assess the reliability of the design and execution of studies in medical literature. Given a document, the model outputs three components: *Summary*, *Strengths*, and *Weaknesses*. *Summary* summarizes the article's content, whereas *Strengths* and *Weaknesses* identify methodological factors that may strengthen or weaken the credibility of conclusions, respectively.

As shown in Figure 2, different study types follow different methodological standards; the model must identify the study type and then assess its quality using the corresponding criteria. For randomized controlled trials, for example, the model needs to determine whether key aspects such as randomization, intervention implementation, and outcome measurement affect the credibility of conclusions. Instead of adopting a common scale-based scoring scheme, we require models to generate structured assessments to examine whether they can perform study-type-specific methodological analysis.

2.3 Cross-Study Evidence Synthesis (CES)

The CES task evaluates whether models can formulate recommendations and judgments consistent with those of an expert panel based on a collection of studies. Specifically, the task input consists of a clinical question and a set of relevant studies. For studies with publicly available full texts, we provide validated structured evidence, including literature understanding results and study quality assessment results; for studies whose full texts are unavailable, we provide the corresponding metadata, including title, publication type, and abstract.

Models are required to follow a standardized reasoning pathway and output the predicted clinical recommendation together with its GRADE rating. The clinical recommendation specifies the intervention or management strategy to be adopted.

As shown in Figure 2, for the given question, the expert panel recommends a blood-pressure target of $< 130/80$ mmHg for hypertensive patients aged < 65 years without comorbidities, with a GRADE rating of 2B. Evidence certainty here differs from the single-study quality assessment in §2.2: the SQA task in §2.2 assesses the reliability of individual studies, whereas GRADE evaluates the entire body of evidence and further considers factors such as the initial evidence level determined by study type and the consistency of findings across studies.

2.4 Data Annotation

This section describes the construction pipeline of MedSynth, illustrated in Figure 3.

Data Collection. The initial data collection and preprocessing were conducted by three trained master's students in computer science, while subsequent dataset validation was performed by five cardiovascular medicine specialists with over ten years of clinical experience and prior involvement in guideline development. We use expert-developed guidelines from China and the World Health Organization, both of which are widely influential in clinical practice, as the best available proxy for expert practice; to ensure benchmark quality and professional rigor, we focus our evaluation on the cardiovascular domain (Chinese Society of Cardiology, Chinese Medical Association et al., 2024; World Health Organization, 2021). We extract clinical questions, recommendations, GRADE ratings, and cited studies from these guidelines.

For cited studies with publicly available full texts, the original full-text PDFs were manually collected and converted into Markdown (Wei et al., 2025). For studies without available full texts, the corresponding metadata, including the title, publication type, and abstract, were retrieved from PubMed through the Entrez API. During preprocessing, references and appendices were manually removed from the Markdown documents.

Based on the collected source studies, annotators first label each article's study type and retain four types for subsequent annotation: cross-sectional study (CS), cohort study (COH), randomized controlled trial (RCT), and systematic review (SR), including SRs with or without meta-analyses. We use Claude Sonnet 4 for the subsequent automated annotation steps.

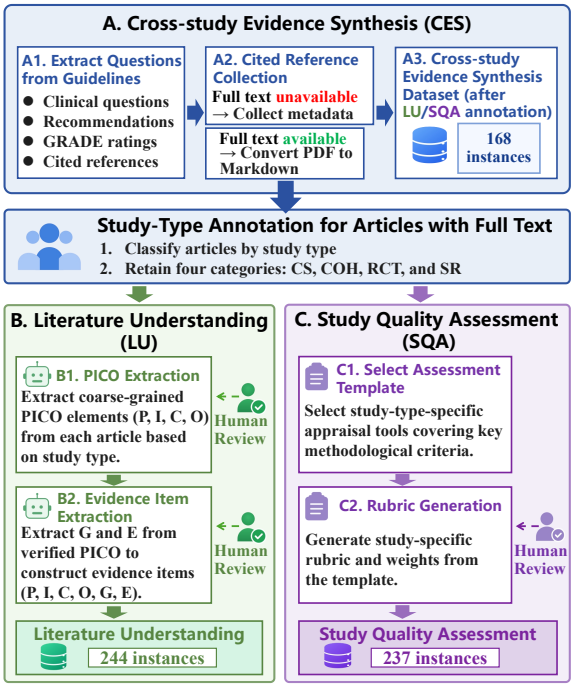

Figure 3: **Overview of the MedSynth data construction and annotation pipeline.** The pipeline includes three stages: cross-study evidence synthesis, literature understanding, and study quality assessment.

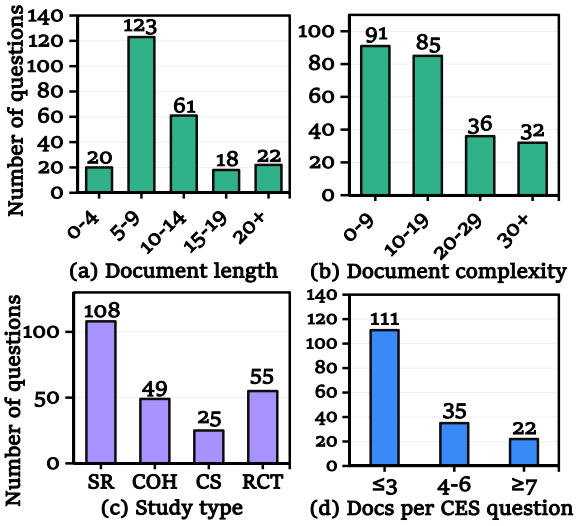

Figure 4: **Distribution of the benchmark.** Panels (a,b) show LU document length in k tokens and complexity by evidence count; panel (c) shows SQA study types; panel (d) shows CES docs per question.

LU subset. Annotation proceeds in two stages. First, the annotation model generates coarse-grained PICO annotations according to the study type. Second, after physician verification of the PICO annotations, the model extracts subgroups (G) and effect estimates (E), forming evidence items in the form of (P, I, C, O, G, E).

SQA subset. We develop a standardized rubric generation method. Specifically, we select the corresponding methodological appraisal tool (Hyde et al., 2015; Barker et al., 2023; Ma et al., 2020; Shea et al., 2017) as a study-type-specific template, as these tools specify the key methodological aspects that should be assessed for different types of literature. The annotation model then uses this template to generate a study-specific scoring rubric and corresponding weights for each article.

Data Quality Assurance. Because the initial model outputs may contain errors and hallucinations, we recruited a team of physicians to perform item-level review and correction of the automatically generated annotations from each step. Additional details are provided in Appendix A.

CES subset. For each CES instance, we construct an evidence list from its cited studies. Studies annotated in both LU and SQA are represented by validated LU and SQA results. Following the setup used in MedEvidence, studies without available full texts are represented by metadata. Each list is paired with the corresponding clinical question, expert recommendation, and GRADE rating.

Benchmark Analysis. The benchmark contains 649 instances in total, including 244 instances for the literature understanding task, 237 instances for the study quality assessment task, and 168 instances for the cross-study evidence synthesis task. Summary statistics are shown in Figure 4.

2.5 Evaluation Protocol

Literature Understanding. The judge model first matches reference evidence items against model outputs using (P, I, C, O, G), and then checks whether the effect estimates (E) are numerically consistent within matched items. The LU score is the proportion of successfully matched evidence items among all reference items.

Study Quality Assessment. The judge model assesses whether the model's output covers the scoring points in the reference rubric. The SQA score is computed as the covered score divided by the total rubric score.

Cross-Study Evidence Synthesis. We evaluate recommendation generation using two metrics: EM-Rec requires semantic equivalence between the model recommendation and the reference answer, whereas LM-Rec requires only consistency in the overall recommendation direction. Both metrics are computed as binary instance-level scores to improve evaluation stability. For GRADE ratings, we

Model	LU†	SQA†	CES					Avg.
			EM-Rec†	LM-Rec	GRADE	GRADE-Rec†	GRADE-Evi†	
Proprietary LLMs								
GPT-5.2	**39.04**	**60.68**	10.12	**91.07**	**32.14**	69.64	**43.45**	**44.59**
Claude Sonnet 4.5	35.86	56.16	23.21	86.90	24.40	69.05	29.76	42.81
o4-mini-high	23.01	42.12	14.88	85.71	30.95	68.45	36.90	37.07
Doubao-Seed-1.6	15.58	42.95	21.43	88.69	24.40	70.83	29.76	36.11
Gemini 3 Pro Preview	19.38	37.89	20.24	86.90	20.83	68.45	30.36	35.26
Gemini 3 Flash Preview	15.19	42.19	24.40	89.29	25.60	58.33	33.33	34.69
Open-Source LLMs								
DeepSeek-V3.2 (reasoner)	30.58	44.14	22.02	89.29	30.95	70.83	38.10	41.13
gpt-oss-120b	19.75	54.89	26.19	89.88	23.21	69.64	29.76	40.05
Qwen3-Next-80B-A3B-Thinking	26.69	39.34	25.60	85.71	20.83	71.43	23.81	37.37
Qwen3-14B (thinking)	13.04	36.41	**28.57**	81.55	25.60	66.07	31.55	35.13
Qwen3-Next-80B-A3B-Instruct	27.28	45.65	15.48	59.52	13.69	**72.02**	14.88	35.06
Qwen3-32B (thinking)	12.72	41.14	23.21	82.74	26.19	64.88	30.95	34.58
Qwen3-30B-A3B-Instruct	16.64	42.03	27.38	84.52	16.67	65.48	19.64	34.23
Qwen3-14B (instruct)	12.77	33.97	25.00	83.33	23.81	63.69	33.93	33.87
DeepSeek-V3.2 (chat)	19.40	45.61	11.90	83.33	21.43	61.31	29.17	33.48
Qwen3-32B (instruct)	14.50	38.52	22.62	70.24	24.40	58.33	32.74	33.34
Qwen3-30B-A3B-Thinking	17.07	37.94	26.19	85.71	18.45	61.31	23.21	33.14
Llama 4 Maverick	7.10	22.45	17.26	66.67	16.07	42.86	22.62	22.46

Table 2: **Main results of model performance.** The best and second-best results are highlighted with green and blue, respectively. Metrics marked with † denote the primary metric for each task. LU, SQA, and CES denote Literature Understanding, Study Quality Assessment, and Cross-study Evidence Synthesis, respectively.

report separate accuracies (Acc.) for recommendation strength (GRADE-Rec) and evidence certainty (GRADE-Evi), as well as overall GRADE Acc., which counts a prediction as correct only when both GRADE-Rec and GRADE-Evi are correct for the same question.

Because GRADE ratings are tied to specific recommendations, we adopt a two-stage evaluation protocol. First, we evaluate whether the model derives the correct clinical recommendation. Second, given the gold-standard recommendation, we assess whether the model assigns the corresponding recommendation strength and evidence certainty based on the clinical question and supporting documents. We use Doubao-Seed-1.6 for LLM-as-a-judge evaluation and validate this approach by measuring its agreement with human evaluation. Details of the judge setup and validity analysis are provided in Appendix B.

3 Experiments

3.1 Experiment Setup

Models. We evaluate 18 LLMs, including six proprietary models and 12 open-source models (OpenAI, 2025b; Anthropic, 2025; team, 2025; ByteDance, 2025; DeepSeek-AI et al., 2025; Yang et al., 2025; Team, 2025; AI, 2025). Details of the models are provided in Appendix B.2.

Prompt Strategy. For the core CES task, we use expert-designed prompts to guide models through the standardized workflow specified in the GRADE framework. The prompt templates for all three tasks are provided in Appendix B.4.

Evaluation Metrics. We report the average of five primary metrics, including LU score, SQA score, EM-Rec, GRADE-Rec Acc., and GRADE-Evi Acc., as a compact summary of overall performance, while using task-specific metrics for detailed model comparison.

Research Questions. We investigate four core research questions. **RQ1**: Do existing LLMs possess reliable medical evidence synthesis capabilities? **RQ2**: Can LLMs comprehensively understand medical literature? **RQ3**: Can LLMs assess individual study quality? **RQ4**: Can LLMs generate expert-consistent clinical recommendations from medical evidence and accurately determine recommendation strength and evidence certainty?

3.2 Overall Performance (RQ1)

Table 2 presents the evaluation results. Our main findings are summarized as follows.

Overall performance remains limited. The best-performing model, GPT-5.2, achieves an average score of only 44.59. Proprietary models show no clear advantage over open-source models; DeepSeek-V3.2 (Reasoner) outperforms most proprietary models and ranks third overall.

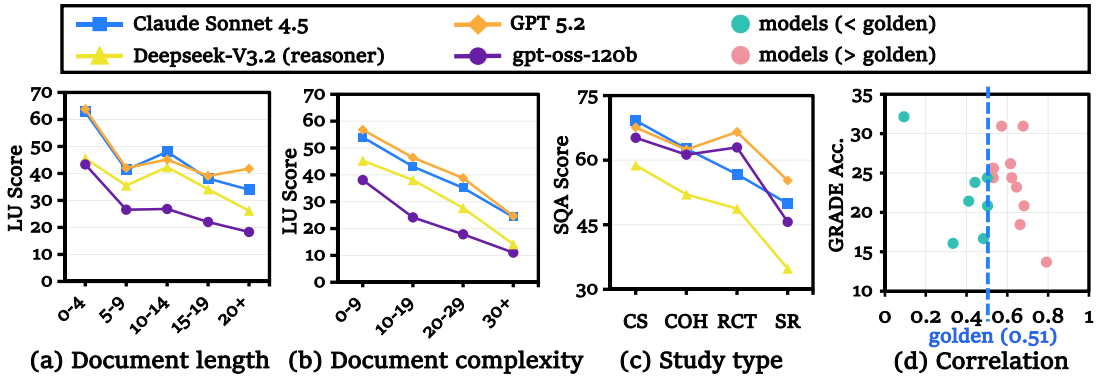

Figure 5: **Statistical analysis of experimental results.** Document length in (a) and complexity in (b) are measured in k tokens and by the number of evidence items, respectively.

Reasoning-enhanced models generally perform better. This advantage is observed across model families and parameter scales. Qwen3-30B-A3B is the only exception, suggesting that the benefit of reasoning enhancement may depend on model scale and architecture.

Increasing model size does not consistently improve performance. Within the Qwen family, dense models show an overall decline as parameter size increases, whereas MoE models improve with increasing total parameters, especially among reasoning-enhanced variants.

3.3 Literature Understanding Results (RQ2)

Overall performance on the literature understanding task remains limited. The best-performing model, GPT-5.2, achieves a score of only 39.04.

As shown in Figure 5(a) and (b), performance is affected by both document length and document complexity, with the latter measured by the number of evidence items. Compared with document length, document complexity has a more pronounced impact on model performance, which declines steadily as complexity increases. This suggests that simply expanding the context window is insufficient for long-document understanding; the key challenge is improving models' stable and fine-grained understanding in long-context settings.

3.4 Study Quality Assessment Results (RQ3)

Overall performance on the study quality assessment task remains limited. GPT-5.2 and gpt-oss-120b from OpenAI achieve the best performance among proprietary and open-source models, respectively. In addition, as shown in Figure 5(c), model performance varies by study type, with systematic reviews (SR) being the most challenging.

3.5 Cross-Study Evidence Synthesis Results (RQ4)

Prior multiple-choice evaluations (Polzak et al., 2026) may allow models to exploit option-specific cues instead of synthesizing evidence. To avoid this shortcut, we adopt an open-ended QA format with two metrics: LM-Rec assesses directional consistency, whereas EM-Rec requires semantic equivalence to the reference recommendation. For example, if the reference recommends "drug A for patients with severe hypertension (systolic blood pressure $\geq$ 160 mmHg)," while the model recommends "drug A for patients with hypertension," the direction is correct but key applicability conditions are missing; the answer is therefore correct under LM-Rec but incorrect under EM-Rec.

Models have limited capability to generate fine-grained recommendations. We find that models perform well on LM-Rec but exhibit a sharp decline on EM-Rec. This indicates that current models are not yet able to integrate clinical, economic, and applicability-related factors to produce precise recommendations.

Models struggle to assign correct GRADE ratings. Model performance is unsatisfactory for both recommendation strength, where the random baseline is 50%, and evidence certainty, where the random baseline is 25%. Llama 4 Maverick even performs below the random baseline. Artificial intelligence for science (AI4S) tasks, such as medical evidence synthesis, rely on existing literature to support scientific decision-making. Our results show that even under explicit operational protocols, models struggle to assess evidence certainty and recommendation strength, suggesting that their reliability in broader settings requires further evaluation.

Model	MSE		ME		Δ	
	EM	LM	EM	LM	ΔEM	ΔLM
Claude Sonnet 4.5	22.96	85.93	18.52	84.44	-4.44	-1.49
GPT-5.2	8.15	88.89	7.41	85.93	-0.74	-2.96
DeepSeek-V3.2 (r)	22.96	87.41	20.74	83.70	-2.22	-3.71
gpt-oss-120b	28.15	88.15	25.93	81.48	-2.22	-6.67

Table 3: CES performance under structured (MSE) and unstructured (ME) evidence inputs; "(r)" denotes the reasoner version.Δ denotes ME−MSE differences.

3.6 Error Analysis

Case analysis. We randomly sample 100 failed cases from GPT-5.2 and identify four major error types: (1) **omission of key details** (12%), where recommendations lack necessary operational details, reducing actionability; (2) **incorrect applicability conditions** (77%), where recommendations are applied to incorrect populations or triggering conditions, causing mismatches with the intended clinical context; (3) **decision-path compression** (1%), where multi-step decision processes, such as if–else branches, are collapsed into single conclusions; and (4) **evidence–recommendation confusion** (10%), where models assume that higher evidence certainty directly implies stronger recommendation strength.

Insufficient caution in recommendation reasoning. To quantify evidence–recommendation confusion, we compute the Spearman correlation between GRADE-Rec and GRADE-Evi, where GRADE-Evi is encoded ordinally as $A = 1$, $B = 2$, $C = 3$, and $D = 4$. As shown in Figure 5(d), most models show stronger correlations than the gold-standard labels. Moreover, performance (GRADE Acc.) drops substantially when this correlation exceeds the gold-standard level. This suggests that models over-rely on evidence certainty, underweighting benefit–risk trade-offs and clinical applicability.

Capability deficits in LU and SQA propagate to CES. When structured evidence in CES is replaced with the full literature context, we retain only instances with input contexts no longer than 130K tokens due to the substantially increased input length. As shown in Table 3, all evaluated models show consistent performance declines under this setting, suggesting that deficits in literature understanding and study quality assessment further limit cross-study evidence synthesis.

4 Related Work

AI-assisted evidence synthesis. Deep Research supports general-purpose web-scale research by synthesizing heterogeneous online sources (OpenAI, 2025a), while Elicit narrows this capability to scientific literature discovery and structured evidence extraction (Whitfield and Hofmann, 2023). OpenEvidence specializes this direction for clinical evidence synthesis (OpenEvidence, 2025).

Benchmarks. ***Auxiliary tasks.*** Existing benchmarks target auxiliary tasks (Lieberum et al., 2025; Clark et al., 2025) such as PICO extraction (Nye et al., 2018) and risk-of-bias assessment (Wang et al., 2025). ***Evidence-driven reasoning.*** As LLMs have shown increasing potential to support scientific research (M. Bran et al., 2024; Boiko et al., 2023; Swanson et al., 2025), recent benchmarks further examine full evidence synthesis reasoning. Evidence Inference 2.0 (DeYoung et al., 2020) evaluates models' ability to determine intervention effects from a single study. MedREQAL (Vladika et al., 2024b) and HealthFC (Vladika et al., 2024a) address medical fact verification using internal knowledge and pre-synthesized evidence, respectively. MedMeta (Ha et al., 2026) assesses models' ability to generate meta-analysis conclusions from retrieved evidence. CochraneForest (Pronesti et al., 2025) evaluates single-study conclusion inference under conflicting RCT evidence. MedEvidence (Polzak et al., 2026) examines whether LLMs match expert conclusions, but mainly analyzes how evidence certainty and recommendation strength affect final conclusion prediction rather than evaluating models' ability to assess them.

5 Conclusion

We propose MedSynth, the first multi-stage medical evidence synthesis benchmark to jointly evaluate recommendation strength and evidence certainty. Experiments show that current LLMs remain unreliable for high-stakes medical evidence synthesis, revealing key limitations in evidence-driven reasoning. MedSynth provides an evaluation foundation for developing more trustworthy and diagnosable scientific research systems, with additional improvement directions discussed in Appendix C.

6 Limitations

The cross-study evidence synthesis task in this study is based on a fixed, expert-curated set of documents, and does not cover large-scale literature retrieval, study screening, or inclusion and exclusion decisions in open environments. This design choice allows us to focus the evaluation on models' evidence-driven reasoning over a given body of evidence, especially their ability to synthesize findings across multiple studies, assess evidence certainty, and determine recommendation strength. Nevertheless, incorporating literature retrieval and screening into the evaluation remains a valuable direction for future extension.

References

Meta AI. 2025. Llama-4-maverick-17b-128e. https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E. Accessed: 2025-04-05.

Pablo Alonso-Coello, Holger J Schünemann, Jenny Moberg, Romina Brignardello-Petersen, Elie A Akl, Marina Davoli, Shaun Treweek, Reem A Mustafa, Gabriel Rada, Sarah Rosenbaum, Angela Morelli, Gordon H Guyatt, Andrew D Oxman, and the GRADE Working Group. 2016. Grade evidence to decision (etd) frameworks: a systematic and transparent approach to making well informed healthcare choices. 1: Introduction. *BMJ*, 353.

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

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.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. 2023. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578.

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):224.

ByteDance. 2025. doubao-seed-1.6-thinking model card. https://seed.bytedance.com/zh/seed1_6. Accessed: 2025-06-11.

Chinese Society of Cardiology, Chinese Medical Association, Hypertension Committee of Cross-Straits Medicine Exchange Association, and Cardiovascular Disease Prevention and Rehabilitation Committee, Chinese Association of Rehabilitation Medicine. 2024. Clinical Practice Guidelines for the Management of Hypertension in China. *Zhonghua Xin Xue Guan Bing Za Zhi*, 52(9):985–1032. In Chinese. PMID: 39289992.

Justin Clark, Belinda Barton, Loai Albarqouni, Oyungerel Byambasuren, Tanisha Jowsey, Justin Keogh, Tian Liang, Christian Moro, Hayley O'Neill, and Mark Jones. 2025. Generative artificial intelligence use in evidence synthesis: A systematic review. *Research Synthesis Methods*, 16(4):601–619.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

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. Association for Computational Linguistics.

Alireza Ghafarollahi and Markus J. Buehler. 2025. Sciagents: Automating scientific discovery through bioinspired multi-agent intelligent graph reasoning. *Advanced Materials*, 37(22):2413523.

Gabrielle Goldet and Jeremy Howick. 2013. Understanding grade: an introduction. *Journal of Evidence-Based Medicine*, 6(1):50–54.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

Huy Hoang Ha, Benoit Favre, and Francois Portet. 2026. Medmeta: A benchmark for llms in synthesizing meta-analysis conclusion from medical studies. *Preprint*, arXiv:2605.09661.

M. Hyde, P. Higgs, R.D. Wiggins, and D. Blane. 2015. A decade of research using the casp scale: key findings and future directions. *Aging & Mental Health*, 19(7):571–575. PMID: 25847497.

Bowen Jin, Jinsung Yoon, Jiawei Han, and Sercan Arik. 2025. Long-context llms meet rag: Overcoming challenges for long inputs in rag. In *International Conference on Learning Representations*, volume 2025, pages 37784–37822.

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. Association for Computational Linguistics.

Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, and Page MJ, editors. 2024. *Cochrane Handbook for Systematic Reviews of Interventions version 6.5*. Cochrane. Updated August 2024.

Zihan Liao, Jun Wang, Hang Yu, Lingxiao Wei, Jianguo Li, Jun Wang, and Wei Zhang. 2025. E2LLM: Encoder elongated large language models for long-context understanding and reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19201–19230, Suzhou, China. Association for Computational Linguistics.

Judith-Lisa Lieberum, Markus Toews, Maria-Inti Metzendorf, Felix Heilmeyer, Waldemar Siemens, Christian Haverkamp, Daniel Böhringer, Joerg J. Meerpohl, and Angelika Eisele-Metzger. 2025. Large language models for conducting systematic reviews: on the rise, but not yet ready for use—a scoping review. *Journal of Clinical Epidemiology*, 181:111746.

Andres M. Bran, Sam Cox, Oliver Schilter, Carlo Baldassari, Andrew D. White, and Philippe Schwaller. 2024. Augmenting large language models with chemistry tools. *Nature Machine Intelligence*, 6(5):525–535.

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, Junyi Jessy Li, Roma Patel, Yinfei Yang, Iain Marshall, Ani Nenkova, and Byron Wallace. 2018. A corpus with multi-level annotations of patients, interventions and outcomes to support language processing for medical literature. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 197–207, Melbourne, Australia. Association for Computational Linguistics.

OpenAI. 2025a. Deep research system card. https://openai.com/index/deep-research-system-card/. Accessed: 2025-02-25.

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

OpenEvidence. 2025. Openevidence. https://www.openevidence.com/.

Christopher Polzak, Alejandro Lozano, Min Woo Sun, James Burgess, Yuhui Zhang, Kevin Wu, Chia-Chun Chiang, Jeffrey J Nirschl, and Serena Yeung-Levy. 2026. Can large language models match the conclusions of systematic reviews? In *The Fourteenth International Conference on Learning Representations*.

Massimiliano Pronesti, Joao H Bettencourt-Silva, Paul Flanagan, Alessandra Pascale, Oisín Redmond, Anya Belz, and Yufang Hou. 2025. Query-driven document-level scientific evidence extraction from biomedical studies. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28034–28051. Association for Computational Linguistics.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A Lenert. 2025. The emergence of large language models as tools in literature reviews: a large language model-assisted systematic review. *Journal of the American Medical Informatics Association*, 32(6):1071–1086.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

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

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. 2025. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723.

Alibaba Qwen Team. 2025. Qwen3-next-80b-a3b-instruct and qwen3-next-80b-a3b-thinking model card. Accessed: 2025-09-12.

Gemini team. 2025. Gemini 3 pro preview. https://ai.google.dev/gemini-api/docs/models/gemini-3-pro-preview?hl=zh-cn#documentation. Accessed: 2025-11-18.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024a. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italia. ELRA and ICCL.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024b. MedREQAL: Examining medical knowledge recall of large language models via question answering. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14459–14469. Association for Computational Linguistics.

Can Wang and Yiqun Chen. 2025. Evaluating large language models for evidence-based clinical question answering. *Preprint*, arXiv:2509.10843.

Jianyou Wang, Weili Cao, Longtian Bao, Youze Zheng, Gil Pasternak, Kaicheng Wang, Xiaoyue Wang, Ramamohan Paturi, and Leon Bergen. 2025. Measuring risk of bias in biomedical reports: The RoBBR benchmark. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3220–3248. Association for Computational Linguistics.

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

Sharon Whitfield and Melissa A. Hofmann. 2023. Elicit: Ai literature review research assistant. *Public Services Quarterly*, 19(3):201–207.

World Health Organization. 2021. *Guideline for the Pharmacological Treatment of Hypertension in Adults*. World Health Organization, Geneva.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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 discusses potential improvement strategies suggested by our findings.
- Appendix D contains extended case studies and error analyses.
- Appendix E describes our use of LLMs.

A Benchmark Construction

A.1 Literature Understanding (LU) Construction

A.1.1 Annotation Protocol

For the four retained study types, we construct LU annotations through study-type-specific extraction strategies. Cross-sectional studies (CS) and cohort studies (COH) focus on exposure variables, reference groups, and reported associations. Randomized controlled trials (RCT) focus on interventions, comparator groups, and reported outcomes. Systematic reviews (SR) focus on synthesized evidence and pooled estimates.

Two-stage annotation. LU annotation follows a two-stage human–AI collaborative procedure using Claude Sonnet 4.0. First, the annotation model generates coarse-grained PICO annotations according to study type. After physician verification, subgroup information (G) and effect estimate (E) are extracted to construct the final structured evidence representation (P, I, C, O, G, E). When multiple evidence items share identical (P, I, C, O, G) components, effect estimate fields are merged to reduce redundancy while preserving complete information.

A.1.2 Quality Control

Human quality control. After automatic annotation, human reviewers audit model outputs item by item to correct errors in PICO identification, subgroup assignment, and effect estimate extraction. Review focuses on evidence requiring cross-section integration and key fields affecting final evidence construction. Labels are revised only with explicit textual support from the article.

Fragmented evidence integration. Medical evidence is often distributed across methods, tables, and supplementary materials. When sample sizes, event counts, or statistical estimates are reported separately, reviewers integrate the corresponding evidence to revise final annotations and improve evidence completeness and consistency.

A.2 Study quality assessment (SQA) Construction

A.2.1 Annotation Protocol

For the four retained study types, we construct SQA annotations through a standardized rubric-generation procedure. We use established methodological appraisal tools as study-type-specific templates: JBI for cross-sectional studies, CASP for cohort studies, RoB 2.0 for randomized controlled trials, and AMSTAR 2 for systematic reviews (Hyde et al., 2015; Barker et al., 2023; Ma et al., 2020; Shea et al., 2017). These templates define item-level methodological appraisal criteria and encode the relative importance of different methodological domains.

Study-specific rubric annotation. Given an article and its corresponding template, Claude Sonnet 4.0 generates a study-specific rubric. Each assessment item contains a criterion description, predefined scoring options, and an item weight. For each item-level judgment, the model must also provide source-text evidence sentences, enabling targeted human verification.

Weighted reference evidence. Reference evidence items are further weighted by polarity and criticality: 1 for regular strengths, 2 for critical strengths and regular limitations, and 3 for critical limitations.

A.2.2 Quality Control

Human quality control. After automatic annotation, human reviewers audit model outputs item by item to correct errors in evidence localization, cross-section evidence integration, and application of appraisal rules. Review focuses on items requiring dispersed evidence and critical items affecting aggregate ratings, such as RoB 2.0 risk of bias. Labels are revised only with explicit textual support from the article.

Mitigating language-style bias. Because we use LLM-as-judge evaluation, the judge model may favor outputs whose wording closely matches the reference annotations, introducing language-style bias. To mitigate this bias, we represent reference annotations as structured rubric items. Each item contains predefined scoring options, an item weight, and source-text evidence sentences, making

the evaluation focus on methodological correctness rather than writing style.

A.3 Cross-Study Evidence Synthesis (CES) Construction

Guideline-based source extraction. The CES task is constructed from clinical practice guidelines, which provide structured clinical questions, recommendations, rationales, and cited references. For each guideline question, human annotators extract the original question, associated recommendations, recommendation rationales, and explicitly cited studies.

Recommendation-level decomposition. Because one guideline question may correspond to multiple recommendations, human annotators split it into separate CES instances. Each instance focuses on one recommendation, with the original question rewritten into a corresponding sub-question.

Rationale and evidence alignment. Annotators then align the guideline rationale and cited references to each sub-question, identifying the rationale components and supporting studies relevant to each decomposed recommendation. Each CES instance contains a sub-question, target recommendation, recommendation-specific rationale, and supporting studies.

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 4.

B.3 Experiment Metrics

B.3.1 Literature Understanding (LU)

LU is evaluated by effect estimate recall under structured evidence alignment. The judge model first matches reference evidence items against model outputs using (P, I, C, O, G) to identify corresponding evidence units. For matched evidence items, effect estimate fields (E) are further compared with reference annotations. The LU score is computed as the proportion of successfully extracted effect estimate fields among all reference effect estimate fields, averaged across instances and reported as a percentage.

B.3.2 Study quality assessment (SQA)

SQA is evaluated by weighted rubric coverage. The judge model checks whether each reference rubric item is covered by the model output, using semantic alignment with the gold standard and explicit support from the source article. Items are weighted by polarity and criticality: 1 for regular strengths, 2 for critical strengths or regular limitations, and 3 for critical limitations. The instance-level score is obtained by rounding the weighted coverage rate multiplied by 10, then averaged across instances and reported as a percentage.

B.4 Evaluation Protocol

The prompts used for inference and evaluation are provided below.

B.4.1 Inference

This section presents the CES prompt, which implements the GRADE workflow: (1) **Clinical question framing.** The model identifies the population, intervention, comparator, outcomes, and critical outcomes. (2) **Evidence integration.** The model synthesizes the evidence body, considering study quality, consistency, effect magnitude, and benefit–harm balance. (3) **GRADE judgment.** The model separately assigns evidence certainty (A/B/C/D) and recommendation strength (1/2), then reports the final grade in the "number + letter" format. (4) **Structured output.** The model ends with a JSON object containing **recommendation**, **grade**, **recommendation_explanation**, and **reasoning**.

B.4.2 LLM-as-a-Judge

LU. We evaluate LU outputs using an LLM-as-a-Judge framework. The judge first aligns model

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

predictions with gold-standard evidence units using structured coordinates (P, I, C, O, G) and then assesses effect estimate fields (E) against the reference annotations. Zero-temperature decoding and structured output constraints are applied to improve scoring reproducibility. To validate the LLM-as-a-judge evaluation, we recruited master's-level medical annotators to manually evaluate all GPT-5.2 outputs. Compared with human evaluation, the judge achieves a Spearman correlation of 0.851, indicating reliable automatic assessment quality.

SQA. We evaluate SQA outputs using an LLM-as-a-Judge scheme. The judge aligns model predictions with the gold-standard rubric under the corresponding methodological framework and assesses item-level evidence coverage through full-text review of the source article. To validate the LLM-as-a-judge evaluation, we recruited master's-level medical annotators to manually evaluate all GPT-5.2 outputs. Compared with human quality assessment, the judge achieves a Spearman correlation of 0.712, indicating reliable relative ranking.

CES. We evaluate CES outputs with an LLM-as-a-Judge framework for recommendation correctness. The judge assesses each recommendation using exact recommendation matching (EM-Rec Acc.) and directional recommendation consistency (LM-Rec Acc.). Cohen's κ is used to measure human–machine agreement, accounting for chance agreement under binary and potentially imbalanced labels. To validate the LLM-as-a-judge evaluation, we recruited master's-level medical annotators to manually evaluate all GPT-5.2 outputs. The judge achieves near-perfect agreement on LM-Rec Acc. ($\kappa = 0.913$) and substantial agreement on EM-Rec Acc. ($\kappa = 0.651$), with lower EM agreement reflecting the stricter and more boundary-sensitive nature of exact matching.

B.5 OCR Setting

OCR Processing. 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 run in a dedicated conda environment with vLLM 0.8.5+cu118 as the inference backend. Each PDF page was rendered at 144 DPI before OCR to preserve layout information and small text.

We used the DeepSeek-OCR processor with image-level tokenization and cropping enabled, following the high-quality configuration in the official implementation. Model inference was performed with a maximum context length of 8,192 tokens, block size of 256, and GPU memory utilization set to 0.9. Decoding used greedy generation with temperature 0. To reduce repetitive structural out-

puts, we applied a no-repeat n-gram constraint with n-gram size 20 and window size 50.

C Potential Improvement Strategies

C.1 Enhancing Intrinsic Model Capabilities

Reasoning enhancement. Our experiments show that reasoning-enhanced models exhibit consistent advantages. This trend appears across model families and parameter scales, suggesting that targeted reinforcement learning methods such as GRPO (Shao et al., 2024) may be a promising direction for improving models' evidence synthesis capabilities.

Improving literature understanding and study quality assessment. Our results reveal insufficient capabilities in literature understanding and study quality assessment, and show that these deficits further limit cross-study evidence synthesis. We also find that model performance is affected by both literature length and contextual complexity, with complexity exerting a stronger influence. In this work, we characterize complexity using evidence items; future work may explore alternative complexity representations across broader scientific domains and use them to progressively improve models' reasoning over complex documents.

C.2 Externalizing the Evidence Synthesis Burden

Another promising direction is to preprocess medical literature into structured evidence before synthesis. In this setting, evidence items are extracted from source studies and assigned quality annotations, reducing the model's burden of judging study quality from scratch. The model can then focus on evidence-driven reasoning over already organized and quality-aware inputs.

This strategy is particularly valuable for practical AI-assisted scientific research systems. To reduce token cost and improve efficiency, such systems often use paragraph-level evidence snippets rather than full papers as model inputs. However, fragmented snippets make it harder for models to assess study quality, because key methodological details may be absent or separated from the extracted evidence. Externalizing evidence extraction and quality appraisal can therefore provide more reliable inputs for downstream synthesis.

D Case Study

D.1 Ignoring Critical Details

We present cases where models fail to include critical thresholds or operational details required for actionable recommendations, shown in Figure 6–Figure 10.

D.2 Conditioning Errors

We present cases in which models apply recommendations to incorrect conditions, shown in Figure 11–Figure 15.

D.3 Decision Path Compression

We present case where multi-step or conditional decision processes is overly compressed into simplified conclusions, shown in Figure 16.

D.4 Evidence–recommendation Confusion

We present cases where models conflate evidence quality with recommendation strength, overlooking other essential decision factors, shown in Figure 17–Figure 21.

E Use of LLMs

AI-Assisted Pipeline for Dataset Construction. During dataset construction, we designed and employed an automated data processing pipeline powered by large language models to improve the efficiency and consistency of large-scale data processing. All model-generated intermediate outputs were treated as provisional and were not directly released as part of the final dataset. Each sample was comprehensively reviewed by human experts with relevant domain expertise, including item-by-item revision, correction, or rejection. The final annotations were determined by human annotators, who retained full responsibility for them.

AI-Assisted Language Polishing in Manuscript Preparation. During manuscript preparation, AI tools were used solely for language-level polishing and expression refinement, including improvements in fluency and grammatical accuracy. The authors solely developed all core scientific content, including the original ideas, research design, experimental analysis, interpretation of results, and conclusions. AI systems did not contribute to the generation of new analytical insights or research conclusions.

Literature Understanding (LU) 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: Study Type Classification (CRITICAL STEP)
Before extracting evidence, classify the paper into one of the following retained study types.
* **SR (Systematic Review)**
Target: Systematic reviews with or without meta-analysis.
DO:
- Extract synthesized evidence and pooled estimates.
DON'T:
- Do not extract individual study rows.
* **RCT (Randomized Controlled Trial)**
Target: Parallel, cluster, or crossover randomized trials.
DO:
- Prioritize Intention-to-Treat results.
- Use Safety Set analyses for adverse events.
DON'T:
- Avoid Per-Protocol analyses unless ITT is unavailable.
* **COH / CS (Cohort Study / Cross-sectional Study)**
Target: Observational studies.
DO:
- Record exposure variables under the I field.
- Prioritize the most fully adjusted estimates.
- Extract reference groups and reported associations.
DON'T:
- Avoid crude estimates if adjusted estimates 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": "",
    "assigned_study_type": ""
  },
  "evidence_tuples": [
    {
      "trace_id": "",
```

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

Study Quality Assessment (SQA) Inference Prompt

1. SQA-inf (System Prompt)
You are an **expert in evidence-based medicine and clinical research methodology**.
Your task is to perform **Study Quality Assessment (SQA)** for a medical research article 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. SQA-inf (User Prompt)
Your task is to conduct a systematic assessment of the methodological quality of the provided medical research study.
At the beginning of the assessment, you should first determine the type of study based on information explicitly reported in the document. The retained study types include cross-sectional study (CS), cohort study (COH), randomized controlled trial (RCT), and systematic review (SR). 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 assessment, 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 —

Cross-study Evidence Synthesis (CES) Inference Prompt

You are an expert clinical guideline developer and evidence-based medicine methodologist.
Your task is to perform **Cross-study Evidence Synthesis (CES)** under the GRADE framework based on the provided evidence list (`evi_list`).
You must synthesize evidence across studies, generate a clinical recommendation, and assign the corresponding GRADE rating.
==========
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 rating 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 rating 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, 1C, 1D, 2A, 2B, 2C, and 2D.

==========
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 clinical recommendation statement.
- grade: the GRADE rating, 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 list (`evi_list`). Do NOT introduce any external knowledge, guideline conclusions, or subjective speculation.

Literature Understanding (LU) Evaluation Prompts

STAGE 1: SYSTEM PROMPT
Role: Senior Clinical Evidence Auditor – Tuple Matching Specialist.
Task: Match Golden tuples with Predicted tuples based on structured evidence alignment over (P, I, C, O, G).
Step 1: Tuple Matching (P,I,C,O,G) – ONE-TO-ONE MATCHING ONLY
Iterate through every **Golden Tuple**.
(Note: Golden tuples represent extracted effect estimates. If a tuple contains no valid estimate information, 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.
- If a Predicted tuple is already assigned, it cannot be matched again.

Matching Strategy

Find the **BEST MATCHING** Predicted tuple for each Golden tuple using semantic matching rather than exact string equivalence.

5-Coordinate Matching Rules

1. **P (Population)**

pico_context.P semantically matches Predicted population definitions.

- Wildcard Rule: Generic populations (e.g., "Overall", "All Patients", "Total Population") match any corresponding population.

2. **I (Intervention)**

Prioritize pico_context.I.

- Match core intervention concepts semantically.
- For observational studies, exposure variables populate the I field.

3. **C (Comparator)**

Prioritize pico_context.C.

- Match comparator concepts semantically.

4. **O (Outcome)**

outcome_info.outcome_name semantically matches.

- Synonyms are allowed.
- Focus on core outcome concepts rather than wording variations.

5. **G (Subgroup)**

estimate_info.subgroup semantically matches.

- Wildcard Rule: "Overall", "All Patients", "NR", or null match any subgroup.

Important Matching Principles

- Use semantic matching rather than exact string equivalence.
- Prioritize pico_context fields.
- If multiple candidates exist, choose the tuple with the highest overall semantic similarity.

Output Format (JSON)

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

STAGE 2: SYSTEM PROMPT

Role: Senior Clinical Evidence Auditor – Effect Estimate Verification Specialist.

Task: Verify effect estimate fields for matched tuple pairs.

Field Verification Protocol
For matched tuple pairs, compare effect estimate fields.
For each field:

- **IGNORE:** Golden value is NR or null.
- **CORRECT:**
  - Numbers (sample_size, event_rate): exact match.
  - Floating-point values (effect_value, ci_lower, ci_upper, p_value, i2_score): tolerance ± 0.02.
  - Strings (effect_measure, statistical_significance): semantic match.
- **WRONG:**
  Predicted value differs or is missing.

Fields to Check
1. **Group Statistics**
sample_size
event_rate
2. **Effect Estimates**
effect_measure
effect_value
ci_lower
ci_upper
p_value
3. **Meta-analysis Statistics**
heterogeneity.i2_score
heterogeneity.p_value_heterogeneity
Output Format (JSON)

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

Important Notes

- Only evaluate matched tuple pairs.
- Only evaluate fields present in Golden annotations.
- Output ONLY valid JSON.

Study Quality Assessment (SQA) Evaluation Prompt

1. SQA-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. SQA-eval (User Prompt)
Below is all information needed to evaluate the **evidence coverage quality** of a model-generated study quality assessment.
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**.

Cross-study Evidence Synthesis (CES) Evaluation Prompt

You are an expert in clinical guideline development and evidence-based medicine methodology.
Your task is to evaluate whether a **model-generated clinical recommendation** is consistent with the **gold recommendation** in the Cross-study Evidence Synthesis (CES) task.
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** corresponding to the two recommendation-generation metrics used in CES evaluation: **EM-Rec** and **LM-Rec**.

---

1. strict_correct (EM-Rec: Exact Recommendation Matching)
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, 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 (LM-Rec: Directional Recommendation Consistency)
This judgment determines whether the model-generated recommendation is consistent with the gold recommendation **in overall 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 overall recommendation 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, thresholds, or additional applicability 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"
}
```

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 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
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 10mmHg, and mismatched atrial fibrillation measurement protocols.

Figure 6: Ignoring critical details Error Case. The omitted information is highlighted in **red**.

Ignoring Critical Details

Question
At what time in the evening is home blood pressure measurement recommended?
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
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 7: Ignoring critical details Error Case. 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 recommendation: For patients with hypertension and coronary heart disease, the recommended blood pressure control target is <130/**80** mmHg.

Model Response
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 8: Ignoring critical details Error Case. 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 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
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 9: Ignoring critical details Error Case. 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 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
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 10: Ignoring critical details Error Case. 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 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
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 11: Conditioning Error Case. 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 recommendation:
Patients with well-controlled blood pressure are recommended to perform home blood pressure monitoring for 1 to 2 days per week.

Model Response
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 12: Conditioning Error Case. 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 recommendation:
Beta-blockers are not recommended as first-line antihypertensive agents in adults with hypertension.

Model Response
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 13: Conditioning Errors. The inconsistent conditional information is highlighted in **red**.

Conditioning Errors

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

Model Response
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 14: Conditioning Errors. 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 recommendation: Pharmacological management of hypertension by a nurse is strongly recommended over usual care in patients with hypertension in LMICs.

Model Response
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 15: Conditioning Errors. 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 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
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 16: Decision Path Compression Error Case.

Evidence–recommendation Confusion

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

Model Response
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 17: Evidence–recommendation confusion Error Case.

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 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
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 18: Evidence–recommendation confusion Error Case.

Evidence–recommendation Confusion

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

Model Response
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 19: Evidence–recommendation confusion Error Case.

Evidence–recommendation Confusion

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

Model Response
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 20: Evidence–recommendation confusion Error Case.

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 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
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 21: Evidence–recommendation confusion Error Case.