Guidepoint. LABS



CLINICAL TRIAL DESIGN BENCHMARK SERIES

How frontier models perform in clinical trial design

2 expert-graded benchmarks covering early phase and late phase lung cancer trials

November 2026

Guidepoint. LABS

IN THIS REPORT

01 Summary

02 Key results

03 Model results

04 Results by criterion

05 Results by task

06 Results by function, phase and dificulty

07 Conclusions and recommendations

09 How the data was made

EXECUTIVE SUMMARY

How frontier models perform in clinical trial design

Guidepoint Labs evaluated 8 frontier and specialist models on 36 clinical trial design tasks developed by practicing specialists. Three credentialed Advisors independently and blindly graded each response against 7 criteria, with a written rationale for every score. Each criterion was rated on a 1 to 5 scale: 1 Poor, 2 Developing, 3 Acceptable, 4 Strong and 5 Excellent. A response's score is the average of its 21 ratings.

5% of responses reached 4.0 (Strong), 8 of 156 across both benchmarks

63% of responses scored below 3.0 (Acceptable), 98 of 156

2.6 vs 3.1 average on the 4 domain-critical criteria vs the 3 general ones; every model scored lower on domain-critical

61% of citation ratings across both benchmarks were 2 or below, the lowest criterion for every model

Guidepoint LABS

Two benchmarks with one measurement standard

The rubric, rating scale, review process and analysis settings were consistent across both benchmarks. Because the tasks, models and evaluators differed, model scores should be compared within, rather than across, benchmarks.

At a glance	STUDY 01 Early phase benchmark	STUDY 02 Late phase benchmark
Phases	Phase 1 through 2a, several indications	Phase 2b through 3b, lung cancer
Tasks	12 tasks (3 Easy, 2 Medium, 7 Difficult), each written for several expert readers	24 tasks (16 Medium, 8 Difficult) across 5 expert functions; 17 limited to 1 function
Models	Claude Sonnet 4.6, GPT-5.4 Pro, Gemini 3.1 Pro, Claude Opus 4.7, MedGemma 27B (specialist medical model)	GPT-5.6, Claude Opus 4.8, Qwen 3.5 (open weights), Gemini 3.1 Pro
Evidence	60 responses, 180 evaluations, 1,251 criterion scores	96 responses, 288 evaluations, 2,016 criterion scores
Evaluators	35 Advisors from a vetted pool of 41	52 Advisors and 14 QA reviewers; 17 also served on the first panel
Expert agreement	Krippendorff's alpha 0.66	Krippendorff's alpha 0.67
Results finalized	June 2026	August 2026

Key results

No model reached the Strong standard

Model averages ranged from 2.56 to 3.04 on the 1 to 5 scale, below the Strong threshold of 4.0. Additionally, 98 of 156 responses scored below 3.0 (Acceptable).

Citations scored lowest for every model

Across both benchmarks, 61% of citation ratings were 2 or below, and evaluators identified fabricated and misapplied references.

The capability profile repeated

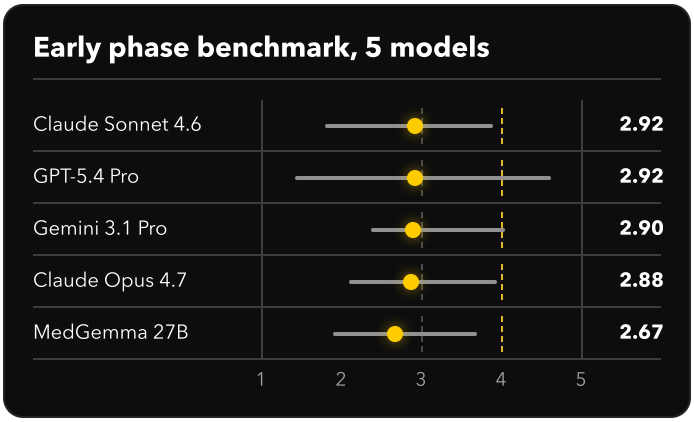
All 7 criteria ranked in the same order in both benchmarks. Although the phases, tasks, models, and evaluators differed, the ordering was unchanged. Within this study, that consistency suggests the pattern was not unique to a single task set.

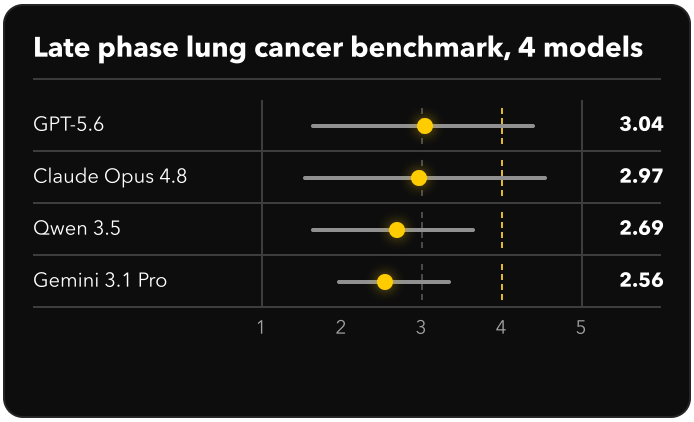
Models rarely tested designs for failure

Feasibility and Practicality and Risks and Mitigation each averaged below 3.0 (Acceptable) across all models in both benchmarks. No individual model averaged above 3.0 on either criterion in either benchmark.

Model results

The highest average score in either benchmark was 3.04 out of 5, nearly a full point below Strong (4.0). Averages also mask meaningful variation from task to task. In the early phase benchmark, GPT-5.4 Pro and Claude Sonnet 4.6 tied at 2.92, but GPT-5.4 Pro ranged from 1.43 to 4.62, while Sonnet ranged from 1.79 to 3.90. In the charts, the dot marks each model's average, and the line spans its lowest to highest response score. Each response score is the average of 21 ratings; 7 criteria assessed by each of 3 Advisors. Models are compared only within a benchmark because the tasks, models, and evaluators differed between them.

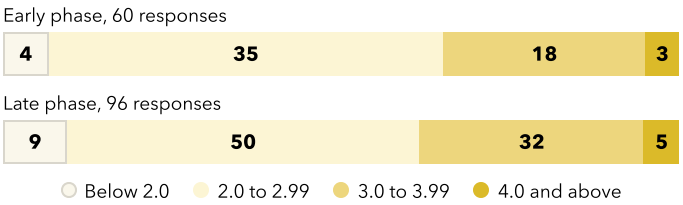




Scale: 1 Poor · 2 Developing · 3 Acceptable · 4 Strong · 5 Excellent

Model	Average	Elo	Response score		Number of responses		
			Lowest	Highest	Below 2.0	3.0 or higher	4.0 or higher
Early phase, 12 responses per model							
Claude Sonnet 4.6	**2.92**	1583	1.79	3.90	1	5	0
GPT-5.4 Pro	**2.92**	1512	1.43	4.62	2	5	2
Gemini 3.1 Pro	**2.90**	1494	2.38	4.05	0	4	1
Claude Opus 4.7	**2.88**	1537	2.10	3.95	0	4	0
MedGemma 27B	**2.67**	1374	1.90	3.71	1	3	0
Late phase, 24 responses per model							
GPT-5.6	**3.04**	1569	1.62	4.43	2	13	3
Claude Opus 4.8	**2.97**	1511	1.52	4.57	3	12	2
Qwen 3.5	**2.69**	1450	1.62	3.67	2	8	0
Gemini 3.1 Pro	**2.56**	1470	1.95	3.38	2	4	0

How response scores were distributed



Head to head, late phase

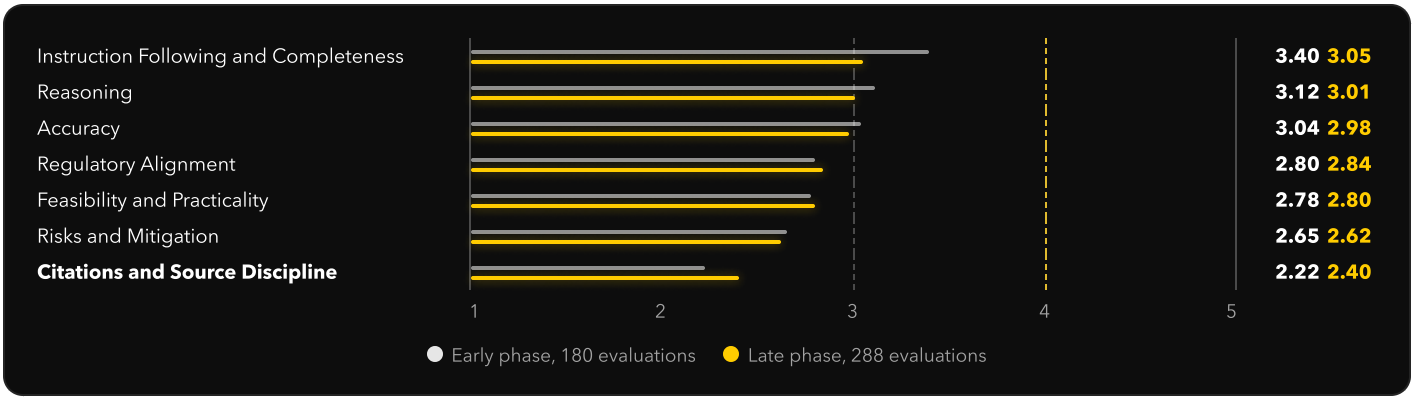
Share of tasks where the row model outscored the column model, ties counted as half

	GPT-5.6	Opus 4.8	Qwen 3.5	Gemini 3.1
GPT-5.6	n/a	60.4%	68.8%	79.2%
Opus 4.8	39.6%	n/a	66.7%	70.8%
Qwen 3.5	31.2%	33.3%	n/a	54.2%
Gemini 3.1	20.8%	29.2%	45.8%	n/a

The Elo scores in the table above rate models based on task-by-task results, starting from 1500. They align with the averages on each benchmark's leader, but the order below shifts, so it works best as a check on the averages.

Results by criterion

Every response was scored on 7 criteria, ranging from instruction following to the extent to which citations supported the response's claims. The criteria ranked in the same order in both benchmarks, despite changes in phases, tasks, models, and evaluators. Models scored highest on Instruction Following and Completeness, Reasoning, and Accuracy, and lowest on Feasibility and Practicality, Risks and Mitigation, and Citations and Source Discipline. These lower-scoring criteria that test whether a design would work and whether the evidence supports it. The chart compares both benchmarks criterion by criterion, and the tables show each model's score on each criterion. Citations was the weakest criterion for every model in both benchmarks, averaging 2.22 in the early-phase benchmark and 2.40 in the late-phase benchmark. A majority of the individual expert ratings on this criterion were 2 or below: 66% in the early-phase benchmark and 58% in the late-phase benchmark.



Bold marks the top model for each criterion.

1	2	3	4	5
Poor	Developing	Acceptable	Strong	Excellent

Early phase, average score by model and criterion

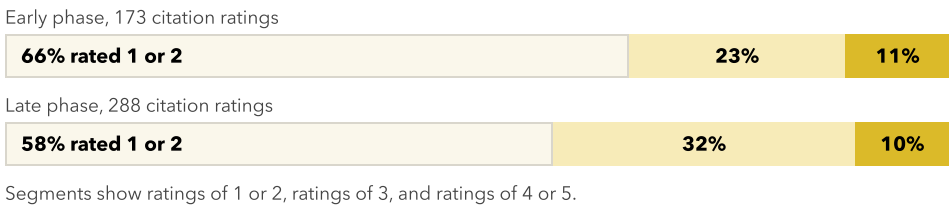
Criterion	Claude Sonnet 4.6	GPT-5.4 Pro	Gemini 3.1 Pro	Claude Opus 4.7	MedGemma 27B	All models
Instruction Following and Completeness	3.44	**3.64**	3.53	3.44	2.94	**3.40**
Reasoning	3.03	**3.42**	3.19	3.06	2.92	**3.12**
Accuracy	2.94	3.08	**3.28**	3.08	2.83	**3.04**
Regulatory Alignment	**3.04**	2.81	2.69	2.81	2.67	**2.80**
Feasibility and Practicality	**2.89**	2.62	2.86	2.83	2.67	**2.78**
Risks and Mitigation	**2.81**	2.58	2.47	2.69	2.69	**2.65**
Citations and Source Discipline	**2.32**	2.31	2.26	2.22	1.97	**2.22**

Late phase, average score by model and criterion

Criterion	GPT-5.6	Claude Opus 4.8	Qwen 3.5	Gemini 3.1 Pro	All models
Instruction Following and Completeness	**3.36**	3.18	2.90	2.75	**3.05**
Reasoning	**3.19**	**3.19**	2.86	2.78	**3.01**
Accuracy	**3.24**	3.12	2.75	2.81	**2.98**
Regulatory Alignment	**3.18**	2.97	2.74	2.47	**2.84**
Feasibility and Practicality	2.99	**3.00**	2.75	2.47	**2.80**
Risks and Mitigation	**2.88**	2.74	2.46	2.42	**2.62**
Citations and Source Discipline	2.46	**2.60**	2.35	2.19	**2.40**

How citation ratings were distributed

In both benchmarks, most citation ratings were 2 or below, while approximately 1 in 10 reached 4 or higher.



Segments show ratings of 1 or 2, ratings of 3, and ratings of 4 or 5.

Results by task

Scores varied more by task than by model. The highest scores were associated with structured documents, such as consent summaries and committee charters. The lowest-scoring tasks involved vulnerable populations such as pregnant and pediatric patients, study-specific dose-escalation designs, and late-phase translational medicine.

1	2	3	4	5
Poor	Developing	Acceptable	Strong	Excellent

Early phase, 12 tasks

Task	Phase	Level	Sonnet 4.6	GPT-5.4 Pro	Gemini 3.1	Opus 4.7	MedGemma	All
D1-01 MAD and PK/PD plan, pregnant bipolar II population	1b	Difficult	1.79	2.14	**2.67**	2.29	2.05	**2.19**
D1-02 Sample size, time to event endpoint, bronchiectasis	2	Difficult	2.79	**3.43**	3.02	2.95	2.38	**2.91**
D1-03 Probability of reaching Phase 3 (BPD)	1b	Difficult	**2.86**	2.81	2.71	2.71	2.81	**2.78**
D1-04 Sample size range per arm, Phase 2a (BPD)	2a	Difficult	**3.38**	2.83	3.10	2.29	2.38	**2.80**
D1-05 Pediatric eligibility criteria, ages 0 to 6	1	Difficult	2.52	2.38	**2.62**	2.38	2.33	**2.45**
D1-06 Consent form and summary, GLP-1 in type 2 diabetes	1	Easy	3.10	**4.14**	2.76	3.76	3.43	**3.44**
D1-07 Sample size with dropout, chronic pain	2	Easy	2.52	2.29	2.86	2.10	**3.19**	**2.59**
D1-08 Single ascending dose chart, pediatric asthma	1a	Medium	2.48	1.90	2.38	**2.52**	1.90	**2.24**
D1-09 Phase 1a data package and AE rates to progress (BPD)	1a	Difficult	3.48	3.48	**4.05**	3.95	3.71	**3.73**
D1-10 Multi-arm work fl ow chart, psoriasis and RA	2	Easy	**2.81**	1.43	**2.81**	2.48	2.00	**2.30**
D1-11 Adjudication committee charter, Hill criteria form	2	Medium	3.48	**4.62**	2.81	3.52	2.95	**3.48**
D1-12 AE list from FDA and EU safety databases, pregnancy	1b	Difficult	**3.90**	3.62	3.00	3.57	2.90	**3.40**

Late phase lung cancer, 24 tasks by expert function

Task	Phase	Level	GPT-5.6	Opus 4.8	Qwen 3.5	Gemini 3.1
Biostatistics						
BIO-P1	2b	Med	2.52	**3.29**	3.19	2.24
BIO-P1-FB	2b	Med	2.29	**3.38**	2.95	1.95
BIO-P2	3a	Med	2.38	**3.00**	2.52	2.57
BIO-P2-FB	3a	Med	2.48	**2.86**	1.62	2.33
BIO-P3-FB	3a	Diff	**2.71**	1.52	2.29	2.10
BIO-P4-FB	3a	Med	1.62	3.19	**3.29**	3.10
Clinical Development						
CLD-P1	2b	Diff	**3.24**	2.95	2.71	2.38
CLD-P1-FB	2b	Diff	**3.38**	3.05	2.14	2.95
CLD-P2-FB	3a	Diff	**4.24**	3.52	1.90	1.95
CLD-P3	3a	Diff	**4.24**	3.00	3.05	2.33
CLD-P3-FB	3a	Diff	**4.43**	3.19	2.62	2.81
CLD-P4-FB	3b	Diff	**3.48**	2.48	2.57	2.00

Task	Phase	Level	GPT-5.6	Opus 4.8	Qwen 3.5	Gemini 3.1
Medical Affairs						
MAF-P1	3a	Med	**3.33**	2.57	3.14	2.43
MAF-P1-FB	3a	Med	**3.10**	2.38	2.95	2.29
MAF-P2-FB	3b	Med	**3.33**	2.81	2.43	3.00
MAF-P4-FB	2b	Med	2.57	2.52	2.29	**2.62**
Regulatory Affairs						
REG-P1	2b	Med	**3.05**	2.76	2.71	2.86
REG-P1-FB	2b	Med	3.00	2.67	2.14	**3.38**
REG-P2-FB	Mixed	Med	**3.71**	**3.71**	3.38	3.24
REG-P3-FB	3a	Med	3.67	**4.43**	3.67	2.76
Translational Medicine						
TRM-P1	2b	Med	2.76	**3.81**	3.19	2.67
TRM-P1-FB	2b	Med	2.95	**4.57**	2.29	2.38
TRM-P2-FB	3b	Med	2.86	1.71	**3.43**	2.62
TRM-P4-FB	3b	Diff	1.67	1.95	2.00	**2.38**

Each cell shows a response score: the average of 21 ratings from 3 Advisors. FB identifies a task limited to one expert function. Bold identifies the highest-scoring model for each task. All 3 of GPT-5.6's Strong responses were Clinical Development tasks. TRM-P4-FB was the most challenging task in the series, with no model scoring above 2.4.

Guidepoint. LABS

Results by function, phase and difficulty

Each model's strengths varied by expert function. In the late phase benchmark, GPT-5.6 led Clinical Development by a substantial margin, while Claude Opus 4.8 scored higher in Biostatistics, Regulatory Affairs and Translational Medicine. A single overall score does not capture those differences.

Bold marks the top model in each row.



Early phase, by evaluator function

	Sonnet 4.6	GPT-5.4 Pro	Gemini 3.1	Opus 4.7	Med Gemma	All
Biostatistics	2.62	2.76	**2.95**	2.51	2.76	**2.72**
Clinical Development	2.74	2.35	**2.92**	2.53	2.36	**2.58**
Clinical Operations	2.92	**3.07**	2.75	3.02	2.60	**2.87**
Medical Affairs	2.96	**3.43**	2.71	2.96	2.93	**3.00**
Regulatory Affairs	3.69	3.79	3.31	**3.83**	3.26	**3.58**
All functions	**2.92**	**2.92**	2.90	2.88	2.67	**2.86**

Late phase, by task function

	GPT-5.6	Opus 4.8	Qwen 3.5	Gemini 3.1	All
Biostatistics (6)	2.33	**2.87**	2.64	2.38	**2.56**
Clinical Development (6)	**3.83**	3.03	2.50	2.40	**2.94**
Medical Affairs (4)	**3.08**	2.57	2.70	2.58	**2.73**
Regulatory Affairs (4)	3.36	**3.39**	2.98	3.06	**3.20**
Translational Medicine (4)	2.56	**3.01**	2.73	2.51	**2.70**
All functions	**3.04**	2.97	2.69	2.56	**2.82**

Because early phase tasks were written for several expert readers, that benchmark is segmented by the evaluator function. Its All functions row also includes 25 evaluations by Advisors in other functions. Each late phase task was written for one function.

Late phase, by difficulty and phase

	GPT-5.6	Opus 4.8	Qwen 3.5	Gemini 3.1
By difficulty				
Medium (16)	2.85	**3.10**	2.82	2.65
Difficult (8)	**3.42**	2.71	2.41	2.36
By phase				
Phase 2b (9)	2.86	**3.22**	2.62	2.60
Phase 3a (10)	**3.22**	2.97	2.70	2.47
Phase 3b (4)	**2.83**	2.24	2.61	2.50

Early phase, by difficulty and phase

	Sonnet 4.6	GPT-5.4 Pro	Gemini 3.1	Opus 4.7	Med Gemma
By difficulty label					
Easy (3)	2.81	2.62	2.81	2.78	**2.87**
Medium (2)	2.98	**3.26**	2.59	3.02	2.42
Difficult (7)	2.96	2.96	**3.02**	2.88	2.65
By phase					
Phase 1 to 1b (7)	2.87	2.93	2.88	**3.03**	2.73
Phase 2 to 2a (5)	**3.00**	2.92	2.92	2.67	2.58

GPT-5.6 performed better on Difficult tasks, while the other 3 models declined. In the early phase benchmark, tasks labelled Easy averaged 2.78, below Medium (2.86) and Difficult (2.89), indicating that the labels did not predict how difficult experts found the tasks. The groups are small and combine functions and phases, so these patterns should be treated as directional. The late phase task spanning several phases is excluded from the phase rows.

Failure patterns evaluators documented across all 5 early phase models

Convincing to a non-specialist

The most common assessment was that a response might appear reasonable to a non-specialist but would require specialist revision before committee review.

Little reasoning about failure

Unless the task expressly requested them, responses rarely addressed site burden, enrollment feasibility, stopping rules or futility analyses.

Fabricated and misapplied citations

Evaluators identified references that did not exist, guidance cited for positions it did not support, and standards attributed to bodies that had not issued them.

Clarifying questions helped

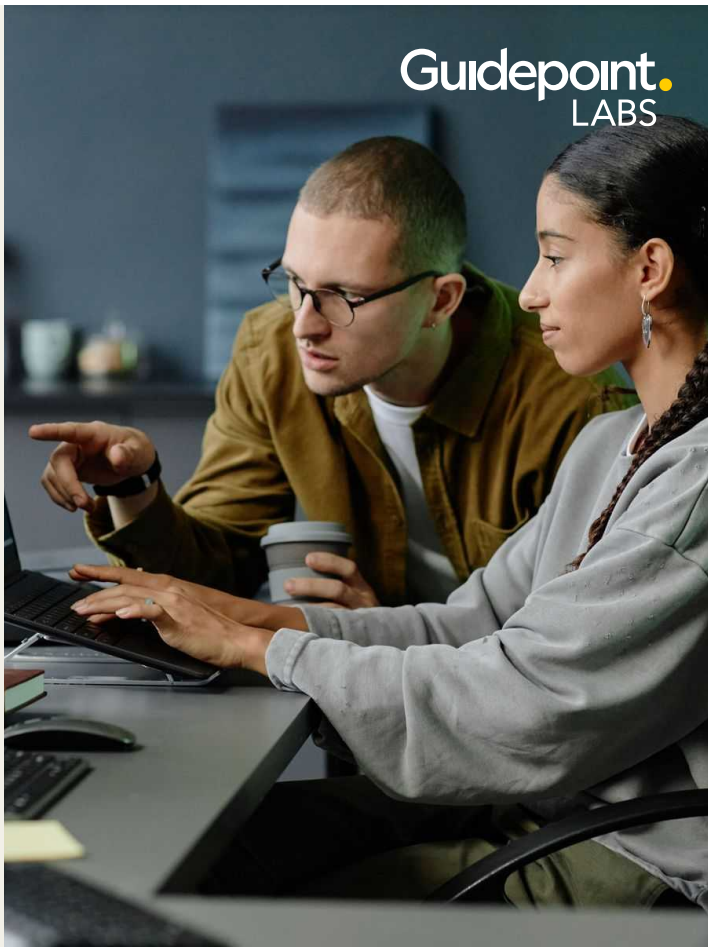
Models that identified ambiguity or asked a clarifying question scored higher, but no model did so consistently.

Conclusions and recommendations

Across 36 tasks, 8 models, and 468 blind expert evaluations, a consistent pattern emerged. The models produced clinical trial design content that often read well but did not consistently withstand specialist review. They generally followed instructions, reasoned clearly, and most addressed many relevant facts.

In clinical trial design, however, a single incorrect dose limit, misapplied guideline or invented reference may have significant consequences for patient safety or a submission. Models were weakest in areas central to whether a trial can be approved, conducted, and managed safely: regulatory fit, feasibility, risk mitigation and evidence.

The 7 criteria ranked in the same order in both benchmarks, even though the phases, tasks, models and evaluators all changed. We therefore read this as a property of current models, not a quirk of a single task set or panel.



Where targeted human data has the most effect

01 Train citation discipline against expert-verified sources

Evaluators found 2 kinds of citation failure. Some references were invented outright. Others were real but misapplied: guidance cited for content it does not contain, or standards attributed to bodies that never issued them. Training data should include claim-level grounding annotation in which specialists check that each reference exists, supports the specific claim and comes from the body named. It should also include contrastive pairs of fabricated and correct citations. Responses that say plainly when no governing source exists should be rewarded, not penalized.

---

02 Teach models to stress-test their own designs unprompted

Feasibility and Practicality and Risks and Mitigation both averaged below 3.0 in both benchmarks. Responses rarely covered site burden, enrollment feasibility, stopping rules or futility analyses unless the task asked for them. The most useful data here is critique-and-revision pairs. In each pair, a specialist takes a model's draft, explains how it would fail in practice and rewrites it. Prompts should deliberately leave out any request for risk content, and rubric-based rewards should give credit when the model raises those risks on its own.

---

03 Train regulatory reasoning on which guidance applies, not only what it says

Regulatory Affairs was the highest-scoring function in the late phase benchmark (3.20), yet the Regulatory Alignment criterion averaged only 2.80 and 2.84 across all tasks. This suggests models can handle regulatory content when asked for it directly but do not apply it when it is not the stated topic. The data should map scenarios to the applicable guidance and the specific requirement within it, by jurisdiction, phase and population. Regulatory Affairs specialists should write this data, and it should be embedded in tasks whose main topic is something else.

04 Prioritize vulnerable populations

The 4 lowest-scoring early phase tasks included 3 built around vulnerable populations: pregnant patients with bipolar II disorder, children aged 0 to 6, and children with asthma. These tasks also carry the highest consequences when a model is wrong. This calls for a small but high-value set of gold-standard demonstrations written by pediatric and maternal-fetal specialists. Each demonstration should state the benefit and risk case these populations require. A dedicated evaluation set should follow the training data to confirm it improved performance on these tasks.

---

05 Supervise the reasoning behind numerical design, not just the answer

Bespoke numerical design was among the lowest-scoring work in the early phase benchmark. In the late phase benchmark, Biostatistics was the lowest-scoring function (2.56). Biostatisticians should annotate each step of a calculation: the assumptions, the source of the effect size, the dropout adjustment and the sensitivity analysis. Final numbers alone teach a model where to land but not how to get there.

---

06 Reward clarifying questions on underspecified tasks

In the early phase benchmark, models that flagged ambiguity or asked a clarifying question scored higher, but no model did this consistently. This behavior can be taught. The data should include deliberately underspecified prompts paired with expert-written responses that either ask the right question or state their assumptions before proceeding. Preference pairs should favor those responses over confident guesses.

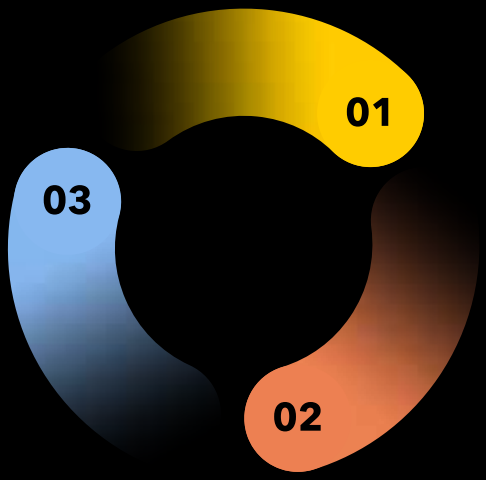
---

07 Train for the floor as well as the average

Every model produced at least 1 response well below its average, and in the late phase benchmark every model had at least 2 responses below 2.0. Running evaluations like this one identifies each model's weakest task types. Data collection should then target those areas specifically rather than adding more examples of work the model already does well.

From evaluation to improvement

This benchmark is the first step in a repeatable cycle. Guidepoint Labs can run every step, so the effect of new training data is measured on the same rubric that found the gap.



01 Evaluate

Blind expert review finds each model's weakest criteria and task types.

02 Build targeted data

Advisors in the relevant specialty produce data aimed at those weaknesses.

03 Re-evaluate

A new blind evaluation on held-out tasks tests whether performance improved.

Guidepoint. LABS

How the data was made

Every score in this report came out of the same 5 steps Guidepoint Labs uses to produce expert data for AI training and evaluation.

01 02 03 04 05

Design

Practicing specialists wrote the 36 tasks and the scoring framework, then tested them against early model output. Late phase tasks carried reference notes defining a strong answer.

Qualification

Advisors were certified blind against a scored rubric before any live work. 42 of 55 certification submissions passed.

Elicitation

Three Advisors scored each response independently, blind to the model and to each other, with a written rationale for every score and no AI tools.

Adjudication

Divergent ratings went to QA review, with 91 reviews by 14 reviewers in the late phase benchmark. Defensible professional disagreement was kept.

Validation

All 288 late phase evaluations were audited before the score set was locked, with the original and final score recorded for each.

The verified Guidepoint experts behind every score

70	credentialed Advisors across the series, 17 of them on both panels
535	years of combined drug development experience in the first panel
0.66 and 0.67	expert agreement (Krippendorff's alpha), highest on citations in both
1.40 to 0.56	late phase gap between experts' response scores, first pass to final

What the review changed

Evaluations adjusted	184 of 288 late phase evaluations had at least 1 criterion score changed at review
Criterion scores changed	477 of 2,016, with 212 raised and 265 lowered
Disagreement kept	528 of 672 response and criterion units (79%) still show experts at least 1 point apart after review

An error it took a specialist to catch

Asked for Phase 1 eligibility criteria for an approved drug in children aged 0 to 6, a model set a blood draw limit of 3 mL/kg and attributed it to "AAHRPP/IRB standards." The pediatric trials specialist who graded it noted that no such standard exists, that the response cited ICH M3(R2), a nonclinical guideline, as governing the study, and that it never made the benefit and risk case a pediatric trial requires.

> A fabricated limit, complete with units and a named authority, reads as settled fact. Here, it would decide how much blood is drawn from an infant.

2M+ vetted Advisors

300+ industries

150 countries

Discuss your AI data priorities

Talk to us

This study is illustrative of the benefit an expert perspective can bring to large language model evaluation and training. It is provided for informational purposes only, and the limited sample size should not be relied upon for broader implications.