Analytical Reasoning - Models Mix (Rerun 2)

LLM Benchmark Evaluation Report

2026-02-25 | Duration: 1h 40m | Models Tested: 8 | Total Cost: $18.73

Evaluation Questions (1/3)

Role: You are a logical analyst who solves problems step by step, showing your reasoning clearly.

Question 0:
A company has 120 employees. 40% work remotely, 30% work in the office full-time, and the rest work hybrid. Of the remote workers, 25% plan to switch to hybrid next quarter. Of the office workers, 15% plan to switch to remote. After these changes, how many employees will be in each category? Show your reasoning step by step.

Role: You are a strategic reasoning assistant.

Question 1:
Three friends -- Alice, Bob, and Carol -- are deciding where to eat dinner. Alice ranks her preferences: Italian > Japanese > Mexican. Bob ranks: Japanese > Mexican > Italian. Carol ranks: Mexican > Italian > Japanese. They agree to eliminate options one at a time by majority vote between pairs. If they first vote Italian vs Japanese, then the winner vs Mexican, what is the final outcome? Now analyze: does the order of voting matter? Could a different pairing sequence produce a different winner? Explain the phenomenon at work.

Role: You solve scheduling and constraint satisfaction problems systematically.

Question 2:
A conference has 4 sessions (A, B, C, D) and 3 time slots. Constraints: A and B cannot be in the same slot (speaker conflict). C must be before D (prerequisite). A must be in the first or second slot (venue availability). No slot can have more than 2 sessions. Find all valid schedules and explain your reasoning for eliminating invalid ones.

Role: You analyze cause and effect in complex systems.

Question 3:
A city introduces free public transit. Within 6 months: traffic congestion decreases 15%, but parking revenue drops 40%. Downtown retail sales increase 8%, but suburban mall revenue drops 5%. Public transit ridership increases 60%, but bus maintenance costs triple. The city council asks: was this policy a net positive or net negative? Analyze the second-order effects, identify which metrics matter most for a final judgment, and state your conclusion with the key assumptions it depends on.

Role: You evaluate arguments with precision, identifying logical strengths and weaknesses.

Question 4:
Consider this argument: 'Countries with higher chocolate consumption per capita have more Nobel Prize winners per capita. Therefore, eating chocolate improves cognitive performance at a national level.' Identify every logical flaw in this argument. For each flaw, name the specific type of fallacy or error, explain why it's a flaw, and describe what evidence would be needed to make the argument valid.

Role: You reason about probability and decision-making under uncertainty.

Question 5:
A medical test for a rare disease has 99% sensitivity (true positive rate) and 95% specificity (true negative rate). The disease affects 1 in 1,000 people. A patient tests positive. What is the probability they actually have the disease? Now: the doctor orders a second independent test, which also comes back positive. What is the updated probability? Show all calculations and explain why the result is counterintuitive to most people.

Evaluation Questions (2/3)

Role: You reason about strategic interaction between rational agents who can anticipate each other's decisions.

Question 10:
Two competing firms must simultaneously set their prices -- High ($100) or Low ($70) -- for an identical product. If both price High, each earns $400K profit. If both price Low, each earns $200K profit. If one prices High and the other Low, the Low-priced firm earns $500K and the High-priced firm earns $80K. (a) Identify all Nash equilibria in this game and explain why each is an equilibrium. (b) What is the socially optimal outcome and why won't rational firms reach it without coordination? (c) If the firms play this game repeatedly for a known finite number of rounds (say, 10 rounds), what does game theory predict will happen in the final round, and what does backward induction imply for all 10 rounds? (d) How does the answer change if the game is repeated indefinitely?

Role: You analyze strategic bidding and auction theory.

Question 11:
Three bidders -- Anya, Bram, and Carla -- are bidding in a sealed-bid second-price auction (Vickrey auction) for a single item. Each bidder's private valuation is known only to them: Anya values the item at $900, Bram at $600, Carla at $750. (a) What is each bidder's dominant strategy in a Vickrey auction, and why? (b) Who wins the auction and what price do they pay? (c) Suppose instead the auction is a first-price sealed-bid auction (winner pays their own bid). Anya now faces a trade-off between winning and overpaying. If Anya believes Bram and Carla will both bid 80% of their true valuations, what should Anya bid to win while maximizing her surplus? (d) A policy advisor claims Vickrey auctions always generate higher total welfare than first-price auctions. Evaluate this claim -- is it necessarily true, sometimes true, or false?

Role: You identify and correct statistical reasoning errors with precision.

Question 12:
A basketball coach notices that players who score above 30 points in one game almost always score fewer points in their next game. She concludes: "Playing 30+ points is exhausting and causes fatigue that hurts performance in the next game." A sports statistician challenges this interpretation. (a) Explain the statistical phenomenon the statistician likely has in mind and why the coach's causal interpretation is almost certainly wrong. (b) Give a concrete numerical example: suppose a player's per-game scoring follows a normal distribution with mean 18 points and standard deviation 6 points. If a player scored 32 points in Game 1 (more than 2 standard deviations above their mean), what is the expected score in Game 2, assuming scores are independent across games? (c) The coach's proposed intervention is to reduce the fatigue-causing players' minutes in the game after a high-scoring game. Explain why this intervention will appear to "work" even if fatigue has absolutely nothing to do with the scoring pattern. (d) What study design would allow the coach to determine whether fatigue is genuinely causing lower performance, as distinct from the statistical artifact?

Role: You identify and explain statistical paradoxes in real-world data.

Question 13:
A hospital system reports the following surgical outcomes for two surgeons -- Dr. Novak and Dr. Osei -- over one year. Dr. Novak performed 200 low-risk surgeries with a 2% mortality rate and 50 high-risk surgeries with a 20% mortality rate. Dr. Osei performed 50 low-risk surgeries with a 4% mortality rate and 200 high-risk surgeries with a 30% mortality rate. (a) Calculate the overall (combined) mortality rate for each surgeon. (b) The hospital administrator looks only at overall rates and concludes one surgeon is clearly safer. Which surgeon does the administrator prefer, and is this conclusion justified? Identify the statistical phenomenon that explains the discrepancy. (c) Explain precisely what must be true about the data structure for this phenomenon to occur. (d) A new hospital policy says surgeons will be ranked and publicly reported by overall mortality rate. What perverse incentive does this create, and how might surgeons rationally respond?

Evaluation Questions (3/3)

Role: You solve estimation problems by building a structured model from first principles, making and stating your assumptions explicitly.

Question 20:
Estimate the total annual cost to the global economy of email spam, including all direct and indirect costs. You do not have access to statistics -- build your estimate from first principles using only numbers you can plausibly justify from general knowledge. Structure your answer as: (1) a model of the problem with defined components, (2) an estimate for each component with explicit assumptions, (3) a final aggregated estimate with an uncertainty range, and (4) a sanity check against any related figure you can derive independently.

Role: You assess risks and make decisions under uncertainty, distinguishing between quantifiable risk and genuine ambiguity.

Question 21:
A pharmaceutical company must decide whether to proceed to Phase 3 clinical trials for a new drug. Phase 3 costs $150M and takes 3 years. Historical data for this drug class shows: 60% of drugs entering Phase 3 receive regulatory approval. If approved, the drug is estimated to generate $80M/year in net profit for 10 years before patent expiry (assume no time-value-of-money discounting for simplicity). If the Phase 3 trial fails, the sunk cost is $150M with no recovery. (a) Calculate the expected monetary value (EMV) of proceeding to Phase 3. Should the company proceed based solely on EMV? (b) The company's chief medical officer argues that the 60% approval rate is based on industry-wide data, but this drug has shown unusually strong Phase 2 results. She estimates the true approval probability for this drug is 75%. Recalculate the EMV under this revised estimate. What is the break-even approval probability at which EMV = 0? (c) The CFO raises a different concern: the $80M/year profit estimate has high uncertainty -- it could be anywhere from $20M to $200M depending on market adoption. He suggests the company should be risk-averse and apply a 30% discount to uncertain revenue projections. Re-evaluate the decision under the CFO's framework. What type of reasoning is the CFO applying, and what are the limitations of this approach? (d) A rival company is also developing a drug in the same class and is expected to reach market 18 months before this drug if both succeed. If the rival reaches market first, this drug's expected revenue drops from $80M/year to $30M/year. How does this competitive risk change the analysis? What is the revised EMV incorporating the competitive scenario, assuming a 50% chance the rival succeeds first?

Role: You identify and reason about how sampling procedures affect the validity of statistical conclusions.

Question 22:
A polling firm surveys 1,000 voters by calling landline phone numbers selected randomly from the public directory. The poll shows Candidate A leading Candidate B by 8 percentage points among those polled. The polling firm reports a margin of error of ±3.1% (95% confidence interval) and claims Candidate A is "virtually certain" to win. (a) Identify at least three distinct sources of sampling bias in this methodology -- be specific about the direction of each bias (i.e., does it overstate or understate Candidate A's support?) and why. (b) The margin of error formula used is ±1.96 × ?(p(1-p)/n) ? ±3.1% for n=1,000 and p=0.5. Explain what this margin of error does and does not capture. Does it account for the biases you identified in (a)? (c) Suppose the true population split is 50/50 (a toss-up). Under what conditions could a well-designed random sample of 1,000 people still show an 8-point lead for one candidate? Calculate the probability of this occurring (hint: use the standard error of the poll under the null hypothesis that the true split is 50/50). (d) The polling firm defends its methodology by noting that the same approach correctly predicted the last three elections. Evaluate this defense -- what are the conditions under which past predictive success of a biased method would and would not give confidence in its future accuracy?

Role: You reason about causal relationships using directed graphs and identify when observed correlations do or do not imply causation.

Question 23:
Consider a study that finds a strong positive correlation between ice cream sales (I) and drowning deaths (D) across months of the year. A researcher proposes the causal graph: I ? D (ice cream consumption causes drowning). A statistician proposes instead: T ? I and T ? D, where T = summer temperature (a common cause). (a)

Judges & Evaluation Criteria

Judges

Gemini 3 Flash Preview
google

Claude Opus 4.5
anthropic

GPT-5.2-chat-latest
openai

Agreement Rate: 36%

Criteria

Reasoning Validity
Weight: 3.0
The logical chain from premises to conclusion is valid, with no gaps, non sequiturs, or unsupported leaps.

Solution Correctness
Weight: 2.5
The final answer or conclusion is factually and mathematically correct.

Reasoning Transparency
Weight: 2.0
The model shows its work clearly enough for a reader to verify each step independently.

Assumption Handling
Weight: 1.5
Unstated assumptions are identified, justified, or flagged rather than silently adopted.

Systematic Progression
Weight: 1.0
The response follows a structured analytical approach rather than jumping to conclusions.

Note: Responses evaluated using blind A/B/C comparison

Results Rankings

Rank	Model	Score	Tokens	Tok/s	Cost	Avg Latency
1	Claude Opus 4.6 [Reasoning (hi	9.10	92,155	53.8	$1.94	68558ms
2	GPT-OSS 120B	9.10	118,994	45.5	Free	104617ms
3	Kimi K2.5 [Reasoning]	8.86	130,919	55.4	$0.33	94465ms
4	GPT-5.2 [Reasoning (high)]	8.67	73,350	54.5	$0.85	53834ms
5	Gemini 3.1 Pro Preview [Thinki	8.34	110,142	36.9	$0.91	119436ms
6	Grok 4-1-fast-reasoning [Reaso	7.97	76,265	107.5	$0.03	28373ms
7	Qwen3 Next 80B A3B Thinking	7.65	214,293	59.1	Free	145120ms
8	GPT-OSS 20B	6.55	44,010	77.8	Free	22633ms

Scores by Criterion

Model	Reasoning (x3.0)	Solution C (x2.5)	Reasoning (x2.0)	Assumption (x1.5)	Systematic (x1.0)	Weighted
Claude Opus 4.6 [Reasonin	9.19	9.26	9.19	8.47	9.23	**9.10**
GPT-OSS 120B	9.03	9.17	9.32	8.73	9.25	**9.10**
Kimi K2.5 [Reasoning]	8.93	9.19	8.74	8.29	8.88	**8.86**
GPT-5.2 [Reasoning (high)	8.72	9.08	8.48	8.25	8.55	**8.67**
Gemini 3.1 Pro Preview [T	8.32	8.63	8.24	8.03	8.37	**8.34**
Grok 4-1-fast-reasoning [	8.03	8.48	7.79	7.27	7.93	**7.97**
Qwen3 Next 80B A3B Thinki	7.43	8.20	7.85	6.77	7.85	**7.65**
GPT-OSS 20B	6.20	6.59	7.12	6.07	7.08	**6.55**

Claude Opus 4.6 [Reasoning (high)]

#1

anthropic

Weighted Score	Win Count	Total Tokens	Est. Cost
9.10	**18**	**92,155**	**$1.94**

Tokens/sec	Avg Latency	P50	P95
53.8	**68558ms**	**52071ms**	**165609ms**

* Most Expensive

Per-Criterion Scores

Claude Opus 4.6 [Reasoning (high)] - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	7.95	1159	19727	$0.02
1	Three friends -- Alice, Bob, and Carol -- ...	9.15	1317	22927	$0.03
2	A conference has 4 sessions (A, B, C, D)...	7.85	2406	31890	$0.05
3	A city introduces free public transit. W...	8.72	2387	57504	$0.05
4	Consider this argument: 'Countries with ...	8.72	2387	51768	$0.05
5	A medical test for a rare disease has 99...	9.22	3055	44894	$0.06
6	A startup has $500K remaining runway and...	9.05	3711	92181	$0.08
7	Explain why the following scenario creat...	9.57	3135	85758	$0.07
8	A company policy states: 'Employees who ...	7.82	2773	51796	$0.06
9	A hospital emergency department must des...	9.37	8703	165609	$0.18
10	Two competing firms must simultaneously ...	9.52	2444	47435	$0.05
11	Three bidders -- Anya, Bram, and Carla -- ...	9.50	2399	43961	$0.05
	Average	**8.87**	**2990**	**59621**	**$0.06**

Claude Opus 4.6 [Reasoning (high)] - Judge Feedback

Gemini 3 Flash Preview

The premier performer providing the most nuanced, professional, and pedagogically superior responses across all domains.

+ Exceptional use of tables, LaTeX, and professional terminology to enhance clarity and authority.
+ Deeply insightful analysis that identifies non-obvious second-order effects and business nuances.
+ Logically airtight proofs and exhaustive verification steps across mathematical and causal tasks.
- Occasional minor rounding in early steps before final calculations.
- Slightly less 'back-of-the-envelope' transparency compared to models that prioritize raw estimation.

GPT-5.2-chat-latest

Exceptionally rigorous and comprehensive, with minor logical missteps and occasional verbosity preventing a near-perfect performance.

+ Highly systematic structure with explicit tables, formal setups, and verification steps across domains.
+ Strong quantitative accuracy in most mathematical, probabilistic, and optimization tasks.
+ Sophisticated causal reasoning, counterfactual analysis, and assumption transparency.
+ Clear separation of conceptual distinctions (e.g., sufficient vs. necessary, welfare vs. fiscal lenses).
- Several logical classification errors involving sufficiency vs. necessity distinctions.
- Occasional rounding ambiguity and minor numerical framing inconsistencies.
- Tendency toward verbosity and redundancy that reduces concision.
- Some asserted structural claims lack full formal justification.

Claude Opus 4.5

Exceptional performer with outstanding depth, sophisticated conceptual frameworks, and comprehensive ethical analysis, demonstrating the strongest overall reasoning quality despite occasional verbosity.

+ Exceptional depth in causal reasoning with thorough counterfactual analysis, identification of overdetermination issues, and clear articulation of fundamental epistemological problems
+ Outstanding conceptual frameworks including 'temporal incentive mismatch', 'ontologically inaccessible' counterfactuals, and sophisticated game-theoretic analysis
+ Comprehensive ethical analysis covering multiple frameworks (utilitarian, Rawlsian, rule of rescue) with practical implementation guidance
+ Excellent use of visual representations including decision trees, condensation DAGs, and comparison tables that enhance understanding

GPT-OSS 120B

#2

lmstudio

Weighted Score	Win Count	Total Tokens	Est. Cost
9.10	**14**	**118,994**	**Free**

Tokens/sec	Avg Latency	P50	P95
45.5	**104617ms**	**103408ms**	**162325ms**

* Free

Per-Criterion Scores

GPT-OSS 120B - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	9.38	2764	115843	Free
1	Three friends -- Alice, Bob, and Carol -- ...	9.90	2406	50430	Free
2	A conference has 4 sessions (A, B, C, D)...	8.48	5549	122867	Free
3	A city introduces free public transit. W...	9.05	4445	97250	Free
4	Consider this argument: 'Countries with ...	9.07	4088	89422	Free
5	A medical test for a rare disease has 99...	9.45	3505	75361	Free
6	A startup has $500K remaining runway and...	7.97	4784	104083	Free
7	Explain why the following scenario creat...	9.23	3304	70950	Free
8	A company policy states: 'Employees who ...	8.18	4254	91847	Free
9	A hospital emergency department must des...	9.32	7129	158717	Free
10	Two competing firms must simultaneously ...	9.80	3967	84773	Free
11	Three bidders -- Anya, Bram, and Carla -- ...	9.23	2927	59936	Free
	Average	**9.09**	**4094**	**93457**	**Free**

GPT-OSS 120B - Judge Feedback

Gemini 3 Flash Preview

A highly rigorous and transparent model that excels in professional presentation and formal mathematical depth.

+ Excellent use of Markdown tables and LaTeX for professional, readable presentation.
+ High level of reasoning transparency with every step and symbolic notation clearly documented.
+ Strong academic integration, accurately connecting scenarios to complex theorems like Arrow's and Condorcet.
- Explanations can be slightly repetitive and the length of responses is sometimes excessive.
- Occasional generic evaluations in specialized areas like legal claims.

GPT-5.2-chat-latest

Most consistently rigorous and technically precise model, combining formal clarity with strong quantitative validation, though sometimes overly verbose or speculative.

+ Clear symbolic formalization and careful handling of logical structure (especially necessary vs. sufficient conditions).
+ Consistently correct numerical work with transparent intermediate steps and sensitivity analysis.
+ Systematic tabular organization that enhances verifiability.
+ Strong causal reasoning with explicit assumption handling and robustness checks.
- Occasional overextension beyond prompt requirements (e.g., theoretical expansions).
- Some illustrative quantitative assumptions risk false precision.
- Verbosity and repetition dilute core insights.
- Minor framing inconsistencies in a few applied evaluations.

Claude Opus 4.5

Top-tier performer with exceptional formal rigor, comprehensive coverage, and outstanding use of tables and verification methods, though occasionally verbose and with minor presentation issues.

+ Exceptional formal rigor with clear notation tables, explicit constraint formalization, and comprehensive proofs using multiple methods
+ Outstanding use of tables for organization, verification, and summary that significantly enhances clarity and auditability
+ Excellent assumption handling with explicit flagging of ambiguities, sensitivity analysis, and consideration of alternative interpretations

Kimi K2.5 [Reasoning]

#3

kimi

Weighted Score	Win Count	Total Tokens	Est. Cost
8.86	**6**	**130,919**	**$0.33**

Tokens/sec	Avg Latency	P50	P95
55.4	**94465ms**	**108072ms**	**153362ms**

Per-Criterion Scores

Kimi K2.5 [Reasoning] - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	8.98	2737	50744	$0.0069
1	Three friends -- Alice, Bob, and Carol -- ...	8.77	1462	24963	$0.0037
2	A conference has 4 sessions (A, B, C, D)...	8.08	7932	142259	$0.02
3	A city introduces free public transit. W...	8.32	2102	39183	$0.0053
4	Consider this argument: 'Countries with ...	9.28	2538	49038	$0.0064
5	A medical test for a rare disease has 99...	9.28	4042	72964	$0.01
6	A startup has $500K remaining runway and...	8.70	3628	69220	$0.0091
7	Explain why the following scenario creat...	8.65	1889	37042	$0.0048
8	A company policy states: 'Employees who ...	7.40	5064	93629	$0.01
9	A hospital emergency department must des...	9.08	7178	134758	$0.02
10	Two competing firms must simultaneously ...	9.33	2657	47531	$0.0067
11	Three bidders -- Anya, Bram, and Carla -- ...	8.78	3205	60903	$0.0081
	Average	**8.72**	**3703**	**68520**	**$0.0093**

Kimi K2.5 [Reasoning] - Judge Feedback

Gemini 3 Flash Preview

A highly professional and business-savvy model that excels in realistic scenario modeling and technical precision.

+ Superb technical precision in economic theory, causality, and statistical frameworks.
+ Highly professional formatting and the inclusion of practical 'Emergency Protocols' for business scenarios.
+ Excellent multi-faceted sanity checks that validate results against macro-level benchmarks.
- Occasional early rounding of values that slightly obscures the mathematical path.
- Slightly less detailed academic framing compared to the highest-ranked models.

GPT-5.2-chat-latest

Analytically strong and conceptually sophisticated with broad correctness, though occasionally overconfident in assumptions and affected by EMV framing errors.

+ Consistently correct formal reasoning in logic, game theory, and combinatorics.
+ Strong causal and counterfactual analysis with nuanced discussion of overdetermination.
+ Clear optimization reasoning tied to binding constraints.
+ Well-structured explanations with explicit assumption articulation.
- Incorrect baseline EMV and related probability framing.
- Occasional overstatement of necessary conditions in policy logic tasks.
- Some speculative quantitative assumptions inflate estimates.
- Minor verbosity and occasional redundancy.

Claude Opus 4.5

Exceptionally strong performer with excellent verification practices, sophisticated analytical frameworks, and comprehensive coverage across mathematical, logical, and strategic reasoning tasks.

+ Outstanding verification and transparency with explicit step-by-step calculations, totals checks, and clear presentation of both exact and rounded answers
+ Sophisticated conceptual framing using terms like 'congestion-rebound paradox', 'modal lock-in threshold', and 'Tragedy of the Engagement Commons' that capture complex dynamics
+ Excellent sensitivity analysis and conditional reasoning, providing specific thresholds and actionable decision criteria

GPT-5.2 [Reasoning (high)]

#4

openai

Weighted Score	Win Count	Total Tokens	Est. Cost
8.67	**6**	**73,350**	**$0.85**

Tokens/sec	Avg Latency	P50	P95
54.5	**53834ms**	**49219ms**	**128782ms**

Per-Criterion Scores

GPT-5.2 [Reasoning (high)] - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	8.38	1010	16178	$0.01
1	Three friends -- Alice, Bob, and Carol -- ...	8.42	776	12665	$0.0090
2	A conference has 4 sessions (A, B, C, D)...	8.72	2579	38246	$0.03
3	A city introduces free public transit. W...	8.38	2147	49219	$0.02
4	Consider this argument: 'Countries with ...	8.93	2415	52275	$0.03
5	A medical test for a rare disease has 99...	9.48	1593	21384	$0.02
6	A startup has $500K remaining runway and...	8.22	4884	102961	$0.06
7	Explain why the following scenario creat...	8.20	1532	32690	$0.02
8	A company policy states: 'Employees who ...	7.90	1287	18798	$0.01
9	A hospital emergency department must des...	9.22	4902	95782	$0.06
10	Two competing firms must simultaneously ...	9.18	1501	24348	$0.02
11	Three bidders -- Anya, Bram, and Carla -- ...	8.53	1436	25662	$0.02
	Average	**8.63**	**2172**	**40851**	**$0.03**

GPT-5.2 [Reasoning (high)] - Judge Feedback

Gemini 3 Flash Preview

A highly accurate and practical model that prioritizes logical systematicity and actionable advice over visual flair.

+ Highly accurate initial calculations and systematic decision-tree logic.
+ Provides practical, actionable recommendations and strong operational policy sections.
- Lacks practical application in some answers, such as using decimals for human counts.
- Formatting is often plain and less visually accessible for complex logical chains.

GPT-5.2-chat-latest

Highly reliable and logically disciplined with strong quantitative accuracy, though slightly lighter in depth and occasionally affected by structural EMV framing errors.

+ Consistently correct mathematical derivations and optimization setups.
+ Clear handling of logical form, dominance reasoning, and equilibrium concepts.
+ Strong causal reasoning with explicit DAG and counterfactual awareness.
+ Concise yet coherent structure that maintains clarity.
- Baseline EMV and break-even calculations use an incorrect framing.
- Less exhaustive structural elaboration compared to the strongest model.
- Occasionally compressed reasoning limits transparency of intermediate steps.
- Some assumptions acknowledged but not deeply stress-tested.

Claude Opus 4.5

Highly reliable performer with exceptional precision in logical distinctions, thorough assumption handling, and strong practical guidance, though occasionally sacrificing conciseness for comprehensiveness.

+ Exceptional logical precision distinguishing mechanical vs substantive claims, sufficient vs necessary conditions, and risk vs ambiguity with appropriate nuance
+ Outstanding assumption handling with explicit enumeration, clear flagging of ambiguities, and consideration of multiple interpretations
+ Strong practical orientation with checkpoint-based frameworks, specific triggers for reallocation, and actionable implementation guidance
+ Excellent use of multiple verification methods including natural frequencies, sanity checks, and alternative calculation approaches

Gemini 3.1 Pro Preview [Thinking (high)]

#5

google

Weighted Score	Win Count	Total Tokens	Est. Cost
8.34	**3**	**110,142**	**$0.91**

Tokens/sec	Avg Latency	P50	P95
36.9	**119436ms**	**76638ms**	**358150ms**

* Slowest

Per-Criterion Scores

Gemini 3.1 Pro Preview [Thinking (high)] - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	9.05	2645	159577	$0.02
1	Three friends -- Alice, Bob, and Carol -- ...	9.23	2199	412914	$0.02
2	A conference has 4 sessions (A, B, C, D)...	9.62	6876	114545	$0.06
3	A city introduces free public transit. W...	7.93	2388	47398	$0.02
4	Consider this argument: 'Countries with ...	7.93	2361	101842	$0.02
5	A medical test for a rare disease has 99...	9.15	3936	56546	$0.03
6	A startup has $500K remaining runway and...	7.92	3304	111880	$0.03
7	Explain why the following scenario creat...	7.83	2530	358150	$0.02
8	A company policy states: 'Employees who ...	8.65	3706	72965	$0.03
9	A hospital emergency department must des...	8.88	4429	71176	$0.04
10	Two competing firms must simultaneously ...	8.52	2953	93458	$0.02
11	Three bidders -- Anya, Bram, and Carla -- ...	9.07	2718	52569	$0.02
	Average	**8.65**	**3337**	**137752**	**$0.03**

Gemini 3.1 Pro Preview [Thinking (high)] - Judge Feedback

Gemini 3 Flash Preview

A strong analytical model that excels in narrative depth and intuitive analogies but occasionally lacks structural rigor.

+ Effective use of persona and intuitive analogies to explain complex mathematical fallacies.
+ Strong strategic and narrative framing that adds value for executive-level decision making.
+ Deep ethical insights and elegant symbolic proofs for philosophical contradictions.
- Narrative-heavy structure can be harder to verify than table-based responses.
- Occasional failure to build models from first principles or provide requested granular calculations.

GPT-5.2-chat-latest

Clear, accurate, and well-structured overall, but less formally rigorous and occasionally internally inconsistent compared to top-tier responses.

+ Clean step-by-step reasoning with explicit rounding notes and numerical verification.
+ Strong logical clarity in formal implication and game-theoretic problems.
+ Balanced and accessible explanations of statistical and causal concepts.
+ Consistently correct core quantitative results across tasks.
- Occasional internal tension or inconsistency in applied sections.
- Less formal depth and weaker symbolic structure than leading models.
- Some speculative or weakly justified quantitative assumptions.
- Limited exploration of edge cases and structural counterarguments.

Claude Opus 4.5

Strong performer with excellent pedagogical explanations, sophisticated strategic insights, and good theoretical depth, though occasionally incomplete and with some study design and interpretation errors.

+ Excellent pedagogical structure with clear analytical notes, strategic insights about agenda control, and sophisticated framing of concepts
+ Strong theoretical depth including proper application of Revenue Equivalence, Arrow's theorem, and formal decision theory frameworks
+ Good verification practices (with explicit constraint checking, logical deduction to reduce search space, and multiple calculation methods
+ Insightful observations about mathematical structure, market dynamics, and the distinction between symmetric and asymmetric environments

Grok 4-1-fast-reasoning [Reasoning]

#6

grok

Weighted Score	Win Count	Total Tokens	Est. Cost
7.97	**0**	**76,265**	**$0.03**

Tokens/sec	Avg Latency	P50	P95
107.5	**28373ms**	**23911ms**	**66270ms**

* Fastest

Per-Criterion Scores

Grok 4-1-fast-reasoning [Reasoning] - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	8.25	2610	30438	$0.0011
1	Three friends -- Alice, Bob, and Carol -- ...	8.68	1404	10303	$0.0006
2	A conference has 4 sessions (A, B, C, D)...	9.20	3995	28386	$0.0018
3	A city introduces free public transit. W...	7.07	2017	23911	$0.0009
4	Consider this argument: 'Countries with ...	7.82	2185	18229	$0.0010
5	A medical test for a rare disease has 99...	9.28	2302	19158	$0.0010
6	A startup has $500K remaining runway and...	7.43	2430	28231	$0.0011
7	Explain why the following scenario creat...	8.02	1960	21528	$0.0009
8	A company policy states: 'Employees who ...	7.70	3914	32051	$0.0017
9	A hospital emergency department must des...	8.72	4544	66270	$0.0020
10	Two competing firms must simultaneously ...	7.87	1719	11817	$0.0008
11	Three bidders -- Anya, Bram, and Carla -- ...	7.52	1534	9644	$0.0007
	Average	**8.13**	**2551**	**24997**	**$0.0011**

Grok 4-1-fast-reasoning [Reasoning] - Judge Feedback

Gemini 3 Flash Preview

A solid logical performer that provides good academic context but suffers from occasional severe mathematical lapses.

+ Clear, professional, and concise writing style that avoids unnecessary filler.
+ Effective use of systems-thinking terminology and real-world evidence to support economic analysis.
- Prone to significant mathematical errors in complex calculations and logic proofs.
- Presentation is often dry and lacks the granular detail or formatting found in top-tier models.

GPT-5.2-chat-latest

Clear and generally accurate on core technical tasks, but weakened by several derivation mistakes and speculative quantitative assumptions.

+ Strong clarity in step-based reasoning and constraint enumeration.
+ Correct handling of most combinatorial, auction, and regression problems.
+ Good intuitive explanations of voting cycles and regression to the mean.
+ Explicit assumption statements in several analytical sections.
- Incorrect derivation of key thresholds (e.g., discount factor condition).
- Baseline EMV and break-even probability miscalculations.
- Speculative quantitative claims without grounding.
- Less rigorous welfare and causal analysis depth.

Claude Opus 4.5

Competent performer with good theoretical grounding and correct core reasoning, but hampered by brevity that sacrifices depth, occasional calculation errors, and implausible parameter estimates.

+ Good theoretical grounding with appropriate references to Arrow's theorem, Revenue Equivalence, Folk Theorem, and Bertrand's paradox
+ Clear constraint ordering and assumption statements that aid systematic problem-solving
+ Correct core reasoning on most problems with proper identification of dominant strategies, regression to mean, and d-separation
- Insufficient depth in explanations with many sections too brief to fully justify conclusions or explore implications
- Notable errors including incorrect discount factor threshold (0.5 vs 1/3) and implausible estimates ($570B spam cost)

Qwen3 Next 80B A3B Thinking

#7

lmstudio

Weighted Score	Win Count	Total Tokens	Est. Cost
7.65	**1**	**214,293**	**Free**

Tokens/sec	Avg Latency	P50	P95
59.1	**145120ms**	**118289ms**	**321903ms**

* Free * Most Verbose

Per-Criterion Scores

Qwen3 Next 80B A3B Thinking - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	8.43	7630	141854	Free
1	Three friends -- Alice, Bob, and Carol -- ...	8.92	2924	45226	Free
2	A conference has 4 sessions (A, B, C, D)...	7.48	15538	276837	Free
3	A city introduces free public transit. W...	7.00	3051	47249	Free
4	Consider this argument: 'Countries with ...	7.68	2147	32804	Free
5	A medical test for a rare disease has 99...	6.73	5032	79964	Free
6	A startup has $500K remaining runway and...	7.23	13603	237222	Free
7	Explain why the following scenario creat...	7.47	2678	41248	Free
8	A company policy states: 'Employees who ...	7.27	5208	82522	Free
9	A hospital emergency department must des...	8.45	17680	321903	Free
10	Two competing firms must simultaneously ...	9.45	4482	69333	Free
11	Three bidders -- Anya, Bram, and Carla -- ...	7.93	5883	93333	Free
	Average	**7.84**	**7155**	**122458**	**Free**

Qwen3 Next 80B A3B Thinking - Judge Feedback

Gemini 3 Flash Preview

A capable model with good self-monitoring that is unfortunately undermined by major quantitative hallucinations.

+ Strong systematic progression and the ability to self-correct during the reasoning process.

+ Grounds theoretical analysis in relevant real-world examples and bold, well-reasoned conclusions.

- Severe mathematical errors, including 1000x scale discrepancies in financial calculations.

- Inconsistent assumptions and terminology slips that muddle causal and logical proofs.

GPT-5.2-chat-latest

Generally accurate and well-organized but uneven, with notable logical misclassifications and several structural quantitative errors.

+ Clear stepwise organization with explicit constraint checking and verification.

+ Strong performance on core math, Bayes, game theory, and combinatorics tasks.

+ Good intuitive explanations of statistical fallacies and regression effects.

+ Consistent effort to validate answers through multiple methods.

- Repeated confusion between necessary and sufficient conditions in policy logic tasks.

- Several significant EMV and probability framing errors.

- Occasional numerical mistakes and speculative assumptions.

- Less depth in causal counterfactual and overdetermination analysis.

Claude Opus 4.5

Solid performer with good structural organization and correct core calculations, but undermined by notable errors in reasoning, occasional misinterpretations, and inconsistent depth across problems.

+ Clear step-by-step structure with good use of tables, checkmarks, and organized sections that aid verification

+ Correct fundamental calculations in probability, game theory, and optimization problems with proper methodology

+ Good pedagogical explanations of fallacies, psychological traps, and statistical concepts with accessible language

- Critical errors in key problems including misidentifying Simpson's paradox, incorrect bias direction claims, and algebraic mistakes in break-even derivations

- Inconsistent depth with some sections showing erroneous intermediate calculations or convoluted reasoning paths before arriving at correct answers

GPT-OSS 20B

#8

lmstudio

Weighted Score	Win Count	Total Tokens	Est. Cost
6.55	**0**	**44,010**	**Free**

Tokens/sec	Avg Latency	P50	P95
77.8	**22633ms**	**21972ms**	**37049ms**

* Free * Most Concise

Per-Criterion Scores

GPT-OSS 20B - Per-Question Performance

Q#	Question	Score	Tokens	Latency (ms)	Cost
0	A company has 120 employees. 40% work re...	8.60	920	16781	Free
1	Three friends -- Alice, Bob, and Carol -- ...	8.63	855	10141	Free
2	A conference has 4 sessions (A, B, C, D)...	6.95	3102	44189	Free
3	A city introduces free public transit. W...	7.82	1772	23844	Free
4	Consider this argument: 'Countries with ...	7.05	1405	18649	Free
5	A medical test for a rare disease has 99...	8.67	1280	16568	Free
6	A startup has $500K remaining runway and...	6.57	1546	20169	Free
7	Explain why the following scenario creat...	7.33	1258	16309	Free
8	A company policy states: 'Employees who ...	4.55	752	8359	Free
9	A hospital emergency department must des...	5.55	2637	36382	Free
10	Two competing firms must simultaneously ...	3.25	1945	25286	Free
11	Three bidders -- Anya, Bram, and Carla -- ...	7.73	1739	21972	Free
	Average	**6.89**	**1601**	**21554**	**Free**

GPT-OSS 20B - Judge Feedback

Gemini 3 Flash Preview

A model with good structural organization that is severely compromised by fundamental logical and mathematical failures.

+ Well-organized structure with clear distinctions between categories and steps.
+ Effective use of summary tables and transparent weighting schemes for decision-making.
- Fundamental misunderstandings of core concepts like Nash Equilibria and first-order logic rules.
- Major mathematical inconsistencies and non-functional code snippets in complex prompts.

GPT-5.2-chat-latest

Structured and often clear, but undermined by multiple fundamental logical and numerical errors across advanced tasks.

+ Clear table-based organization and structured progression.
+ Correct handling of several basic combinatorics, auction, and probability computations.
+ Systematic constraint checking in simpler optimization problems.
+ Readable formatting aligned to prompt structure.
- Frequent logical misclassifications (e.g., necessary vs. sufficient, Nash equilibria errors).
- Major numerical and graph-structure mistakes in several sections.
- Incorrect EMV, probability, and cost calculations in key problems.
- Incomplete or structurally flawed responses in later tasks.

Claude Opus 4.5

Weakest performer with fundamental errors in game theory, graph analysis, and logical reasoning that undermine otherwise adequate structural organization and basic calculation abilities.

+ Good use of tables and clear formatting that organizes information systematically
+ Correct basic calculations in straightforward probability and arithmetic problems
+ Adequate coverage of standard concepts like dominant strategies and Bayes' theorem when applied correctly
- Major conceptual errors including incorrect Nash equilibrium identification, flawed backward induction analysis, and confused graph reachability reasoning
- Misapplication of fallacy types (post hoc, appeal to authority) and logical concepts (colliders vs forks, prove vs satisfy)