

MULTI-MODEL REASONING

Two models in a debate are *more accurate* than each model alone

A single model can fail because it cannot challenge its own assumptions. Chorus tests whether two independent lines of reasoning make answers more reliable at inference time (when the model answers, with no retraining).

Chorus asks two AI models the same question. Each model answers alone first. Then the two models check each other's work in turns. On 99 hard reasoning questions, Chorus is correct 76.8% of the time. The better model alone is correct 64.6% of the time. The method also has failure modes and a high cost. This note measures both.

	GPT-6-Sol OpenAI · high effort · speaks first	
	Opus 5.5 Anthropic · high effort	
	Chorus both models in turns · 18-turn limit	

By Krishna Kumar Soni

Designed, built and evaluated independently

Run scored-2026-10-04-09480b18

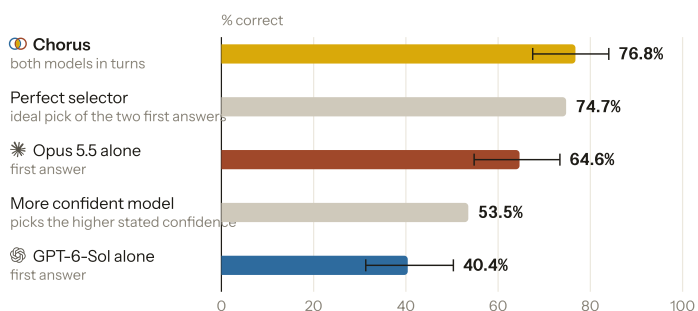
Closed-book (no tools, no web)

Graded by a script, not by an AI

FIGURE 1 · HEADLINE RESULT

Chorus is correct on 76.8% of the questions

Correct answers out of 99. Thin lines are 95% confidence intervals (the range that likely holds the true accuracy). Grey bars are rules that only pick one of the two first answers. The perfect selector (an ideal rule that always picks the correct answer when one of the two models has it) is the best that any such rule can do.



Source: Chorus run scored-2026-10-04-09480b18 on HLE-Diamond. n = 99.

+12.1 pts

More correct answers than Opus 5.5 alone. Chorus gains 18 questions and loses 6. $p = 0.023$ (exact McNemar test: the chance that a gap this large is luck).

76 vs 74

Chorus beats the perfect selector. The debate finds answers that neither model had at the start.

10 of 25

Questions that both models got wrong at first. Chorus got them right.

0.77 AUROC

How well the Chorus confidence separates right from wrong answers (0.5 = chance, 1.0 = perfect). Opus 5.5 alone: 0.57.

WHY I BUILT THIS

I built Chorus after I saw the same pattern many times. Two strong models often give different answers to the same hard question. There was no systematic way to measure whether that disagreement could become useful reasoning. Chorus is my attempt to measure it: one protocol, one fixed question set, every turn recorded and graded by a script.

SUMMARY

The test uses 99 questions from HLE-Diamond (a 1,000-question exam of expert-level problems). GPT-6-Sol alone is correct on 40.4%. Opus 5.5 alone is correct on 64.6%. Chorus is correct on 76.8%. The 95% interval for the gain over Opus 5.5 is +3.0 to +21.2 points (paired bootstrap: the 99 questions resampled 20,000 times). Chorus keeps all 30 answers that both models had right.

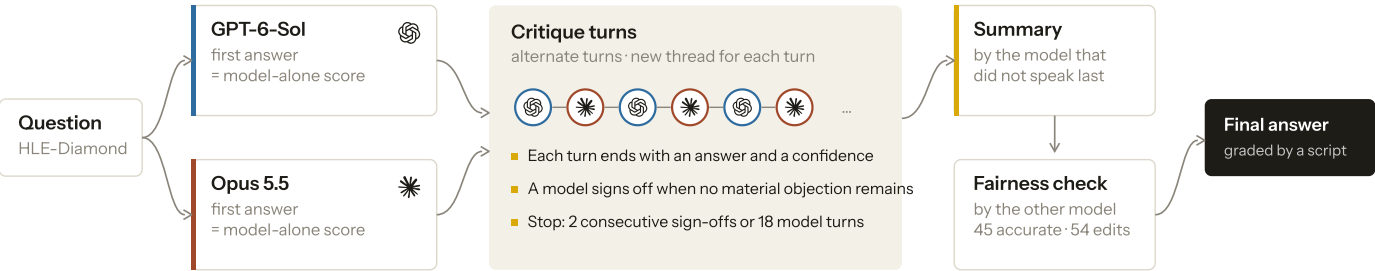
The second half shows how the method fails. A model can talk the other model out of a right answer. Both models can agree on a wrong answer.

How Chorus works: five design decisions

FIGURE 2 · SYSTEM

Two first answers, then turns of critique, then one final answer

The two first answers are also the “model alone” scores. Thus every comparison uses the same model calls. A critique turn (one model’s reply to the other model’s latest argument) starts a new conversation each time. It contains a short shared state (a summary of both positions that the models update), the model’s own last turn and the other model’s latest argument.



FIVE DESIGN DECISIONS

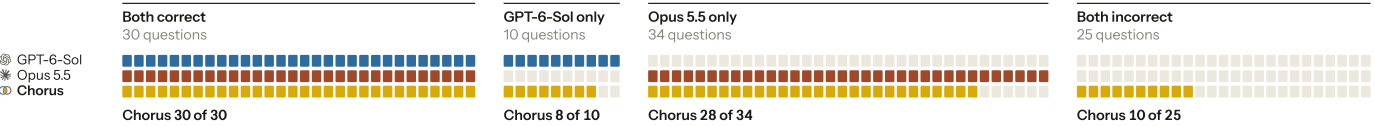
- Closed-book evaluation.** No tools and no web. Every gain must come from reasoning, not from a lookup.
- Independent first answers.** Both models answer in parallel and cannot see each other. These answers are also the solo scores, so every comparison is paired.
- New thread for each critique turn.** A turn sees only the shared state, its own last turn and the other model’s latest argument. This stops one long transcript from anchoring both models.
- Confidence is not evidence.** Each model checks claims itself and attacks the biggest weakness first. It changes its answer only when it can name the step that failed.
- Final answer after a fairness check.** One model writes the summary. The other model checks it for misrepresentation. It accepted 45 summaries and corrected 54.

EXPERIMENT SETUP	
Benchmark	HLE-Diamond by lastexam.ai, released 22 September 2026. 500 reasoning and 500 knowledge questions.
Sample	99 reasoning questions (19.8% of the reasoning set): all 75 text-only multiple-choice questions and 24 whole-number questions. A seeded hash (a fixed random order that anyone can reproduce) sets the order.
Models	GPT-6-Sol (OpenAI) and Opus 5.5 (Anthropic), both at high effort.
Conditions	Closed-book. Tools and web access are off. 0 tool calls in 1,471 model calls.
Turn limit	18 turns: 2 first answers and at most 16 critique turns.
Grading	A script reads the last “Answer:” line. No AI grades the answers. 0 answers are unreadable.
Statistics	Wilson 95% intervals, exact McNemar test, paired bootstrap with 20,000 resamples.

FIGURE 3 · ALL QUESTIONS

Every question, one column each

A filled cell is a correct answer. The four groups show which models were right at the start. Move the pointer over a column for details. The answer keys are not shown, to protect the benchmark.

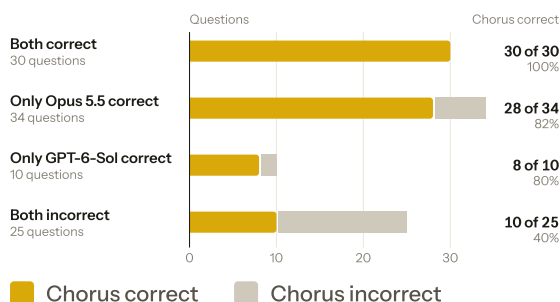


Where the extra correct answers come from

FIGURE 4 · START STATE

Chorus keeps right answers and fixes many wrong ones

Questions grouped by which models were right at the start. The coloured part of each bar is the number Chorus got right.



- Chorus creates new answers.** The perfect selector gets 74 questions right. Chorus gets 76. Both models were wrong on 25 questions. Chorus found the right answer for 10 of them.
- The weaker model adds most of the gain.** Chorus gains 18 questions over Opus 5.5. GPT-6-Sol is the source of 14. In 8, only GPT-6-Sol was right. In 6, GPT-6-Sol found the right answer first. The two models make different mistakes (error correlation $\phi = 0.18$, where 0 is unrelated and 1 is identical).
- Arguments decide, not confidence.** When only one model was right, Chorus ended right on 36 of 44 questions: 28 of 34 for Opus 5.5 and 8 of 10 for GPT-6-Sol. A rule that picks the more confident model gets only 53 of 99.
- Chorus fixes shared mistakes.** On 18 questions, both models gave the same wrong answer (a shared error). Chorus fixed 8. No selector can fix a shared error. On 9 multiple-choice questions, Chorus moved to an option that neither model chose first. It was right on 7. Random choice gives 2.0 ($p < 0.001$).
- Agreement is safe.** When both models started right, Chorus kept the right answer on 30 of 30 questions.

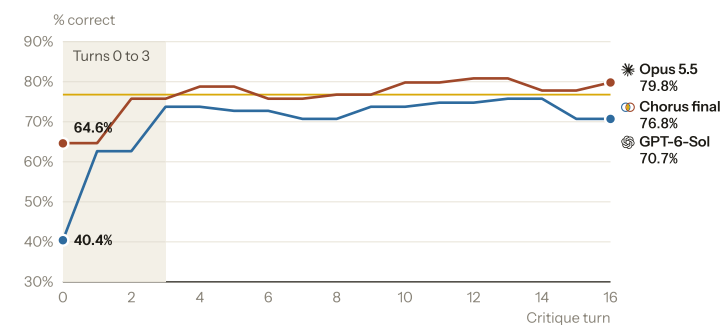
SECTION 3 · TURN BY TURN

Both models get better, and fast

FIGURE 5 · TURN BY TURN

Most of the gain happens in the first three turns

Share of questions where each model's current answer is right. Turn 0 is the first answer. A run that stopped early keeps its last answers.



Source: 1,074 of 1,075 critique turns state an answer. One turn has no answer and keeps the previous one.

- Both models improve.** GPT-6-Sol goes from 40 to 70 right answers. Opus 5.5 goes from 64 to 79.
- The right answer shows up early.** The right answer appears in 90 questions. In 86 of them, it appears by turn 3.
- The models play different roles.** GPT-6-Sol changes its answer 84 times. Opus 5.5 changes 50 times. Opus 5.5 signs off (marks that it has no objection left) in 63% of its turns. GPT-6-Sol signs off in 8.5%.
- Chorus confidence is trustworthy.** On its 54 most confident questions, Chorus is right on 49 (91%). Opus 5.5 is right on 37 of its top 54 (69%). When Chorus says 90% or more, it is right 17 of 17 times.

Sometimes one model talks the other out of a right answer

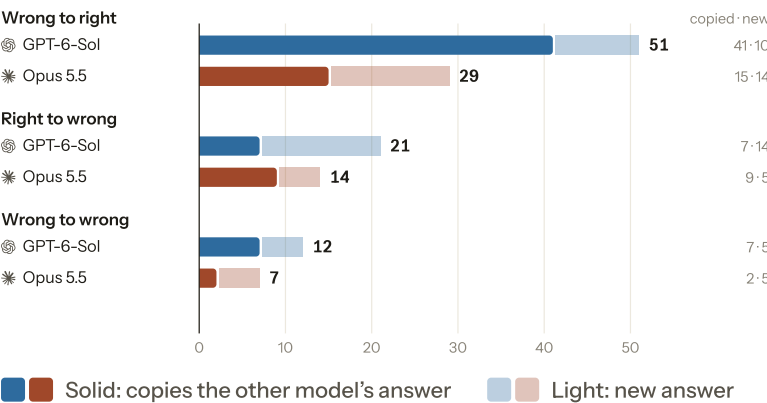
At least one model was right at the start on 74 questions. Chorus ended wrong on 8 of them (10.8%). In total, Chorus was wrong on 23 questions. The table names six ways the method fails. Some questions show more than one.

FAILURE MODE	WHAT IT MEANS	EVIDENCE
Persuasion persuasion-induced corruption	A model drops a right answer and adopts the other model's wrong answer.	16 of 35 right-to-wrong changes adopt the other model's answer. For Opus 5.5, it is 9 of 14.
Drift self-induced drift	A model drops a right answer for a new wrong answer that nobody held.	19 of 35 right-to-wrong changes. GPT-6-Sol makes 14. In 3 questions, both models left the right answer near the turn limit.
False consensus	Both models agree on a wrong answer at the end.	15 of the 23 wrong final answers. 8 of 18 shared errors survive to the end.
Ordering effect	The fixed speaking order decides who writes the final summary.	GPT-6-Sol writes all 16 summaries where the models still disagree. Its own answer is right 1 of 7 times. Opus 5.5's answer is right 7 of 9 times.
Wasted turns	Turns continue after both answers are final.	539 of 1,075 critique turns (50%). 36 of 52 runs that hit the turn limit had already agreed.
Long-tail cost	A few questions use a large share of the compute.	The 10 most expensive questions use 21% of all tokens (units of text a model reads or writes) and 32% of all time.

FIGURE 6 · ANSWER CHANGES

Opus 5.5 gets persuaded. GPT-6-Sol drifts.

All 134 answer changes during critique turns, by direction. The solid part copies the other model's current answer. The light part is a new answer.



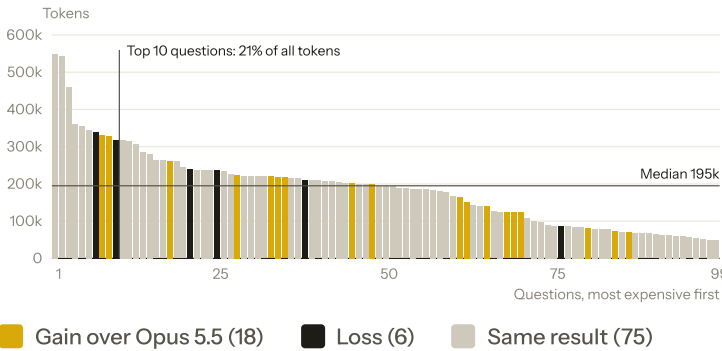
- 10 **Persuasion cuts both ways.** 56 of 80 wrong-to-right changes copy the other model's answer. The same pull causes 16 of 35 right-to-wrong changes.
- 11 **Confidence does not predict who gets persuaded.** When the right model was at least 10 points less confident, Chorus lost 3 of 11 questions. When it was at least 10 points more confident, Chorus also lost 3 of 14.
- 12 **False consensus has a warning sign.** When both models agree on a wrong answer, the average final confidence is 55.7%. When they agree on a right answer, it is 77.2%.
- 13 **The last speaker was more often right.** In the 16 questions that ended in a split (the two models on different answers), Opus 5.5 always spoke last. Its final answer was right 11 times. The summary was right only 8 times.

Hard questions cost much more

FIGURE 7 · TOKENS PER QUESTION

The questions Chorus gets wrong cost the most

Total tokens for each question, from most to least. Colour shows the result compared with Opus 5.5 alone.



RESULT VS OPUS 5.5 ALONE	QUESTIONS	TOKENS EACH	SHARE
Gain	18	180k	17.5%
Same, right	58	171k	53.4%
Same, wrong	17	233k	21.3%
Loss	6	239k	7.7%

14 The tail is long. The typical question uses 195k tokens and 11 minutes. The largest uses 550k tokens. The slowest takes 79 minutes. In total, Chorus uses 7.3× the tokens of the two first answers.

15 Most compute does not change the result. 75% of tokens go to questions where Chorus gets the same result as Opus 5.5 alone.

16 A simple stop rule halves the turns. Stop when both answers match for 3 turns in a row. On the recorded runs, this keeps 76 right answers with 507 critique turns, not 1,075.

Each failure has a fix in the next version

FAILURE MODE	FIX IN THE NEXT VERSION
Persuasion	Candidate map. Before any attack, both models list every possible answer with its assumptions, best evidence and strongest objection. No model picks a winner at this step.
Drift	Candidate record. Chorus keeps the record, not a model. An answer with no open objection stays available. At the end, both models try to disprove the leading answer.
Ordering effect, false consensus	Independent judge. A fresh copy of a randomly chosen model picks the answer from the record. It does not see the leading answer, the last speaker or how many models agree. A random model speaks first.
Wasted turns, long tail	Automatic stop. A turn is settled when both models hold the leading answer and nothing new appears. Two settled turns end the critique. The turn limit drops to 12.

NEXT EXPERIMENTS

Opus 5.5 with Opus 5.5	How much of the gain needs a second, different model?
Same-budget baseline	Does Opus 5.5 with a majority vote over 7× its tokens reach 76?
Next version, new questions	Do the fixes reduce each failure mode on 66 unused questions?

REPRODUCIBILITY

PROTOCOL SHA256	DATASET
342aebdbdbc077f3...79810dc	cais/hle-diamond @ 04eeb7e
SAMPLE SEED	GRADER
chorus-hle-diamond-2026-10	chorus-hle-grader/1.1
CHORUS COMMIT	RESOLVED MODELS
680f882 → 464091f	gpt-6-sol · claude-opus-5-5

Prompts, models and settings stayed fixed for the whole run. Each condition has one run per question. These scores are not an official HLE-Diamond score.