How Stable Are Answers to Unresolved Scientific Questions? Science80: Repetition and Evidence Changes

Anonymous authors
Paper under double-blind review

Abstract

When a language model gives different explanations for the same unresolved scientific question, does the difference reflect a change in the input or variation across repeated generations? Without an accepted complete explanation, disagreement alone cannot distinguish these possibilities. We introduce Science80, a controlled study of 80 questions across five scientific domains, and compare evidence reordering, deletion, and addition against responses to identical prompts. GPT-6 Astra xhigh provides the main experiment, with same-question comparisons from Gemini and Claude. Under identical inputs, the question-weighted rate of substantive revision is already 58.3%, including changes in secondary mechanisms, scope, or scientific preference. Evidence reordering does not show a significant average increment beyond repetition, whereas adding a statement increases the substantive-revision rate by approximately 20 percentage points. This increment remains after adjustment for response length and other text features. Further analyses show that a change in the preferred explanation can accompany little change in reported confidence, while limited directional evidence constrains conclusions about position and conflict. Our results support evaluating input responsiveness relative to repeated-generation variability, while assessing explanatory stability, evidence responsiveness, and scientific correctness separately.

1 Introduction

Scientific assistance requires more than recalling established facts. For an unresolved question, a useful response can propose a provisional mechanism, retain a competing explanation, and suggest an experiment that distinguishes them. Such responses can influence which mechanism a researcher investigates and which observations receive priority. Their usefulness therefore depends both on how explanations vary under the same evidence and on how they change when the evidence changes. **A comparison of two answers must account for the model's own repeated-generation variability before attributing the difference to an input intervention.**

Unresolved questions make this distinction difficult. In tasks with verified answers, changes can be evaluated against accuracy. For an open scientific phenomenon, however, "provisionally prefer A," "the evidence does not distinguish A and B," and "B dominates only under condition X" can all be informative. They modify preference, commitment, or scope, rather than simply being alternative strings for a known answer. Science80 question P05 illustrates the problem. Under the same prompt, Astra's Base response provisionally selects electroweak baryogenesis, while R1 selects high-scale leptogenesis; both acknowledge insufficient evidence (Appendix D). A preferred mechanism can therefore change without an input intervention. Observing one such change after an intervention is insufficient to identify the intervention's contribution.

Prior work establishes sensitivity to prompt formatting, demonstration order, premise order, and information position (Sclar et al., 2024; Lu et al., 2021; Chen et al., 2024; Liu et al., 2023). Semantic uncertainty research emphasizes that distinct strings can express the same meaning (Kuhn et al., 2023; Farquhar et al., 2024). These findings motivate a more specific measurement question: *does a change to the input produce scientific-content revisions beyond those already observed when the input is unchanged?* A distance from one original answer cannot answer this question by itself. A

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Sixteen independent response slots per question. Repeated instances of an ordering do not give that ordering extra aggregate weight.

Condition	Slots	Comparison purpose
Base, R1–R3	4	Identical prompts; estimate generation variability
Shuffle	10	Six reordered inputs; two orderings repeated three times
Less	1	Remove the designated evidence statement
More	1	Append one statement after the original input

single same/different label also merges refinements of conditions, shifts in provisional preference, and replacements of the leading mechanism.

Science80 holds the question and response structure fixed while varying evidence order and quantity. The main analysis scores entire Astra responses on a five-level scale, then compares each intervention with repetition within the same question. Same-question Gemini and Claude responses support an additional analysis of the preferred explanation. All experimental prompts and responses are in Chinese; the English presentation here does not constitute an English-language replication. The study covers 3,802 complete responses across the three configurations, with missing outputs retained as missing.

The results distinguish two observations that would otherwise be conflated. Substantive revisions are common under repetition, yet reordering does not show a significant average increment after multiplicity correction. Adding a statement produces an increment of about 20 percentage points. Substantive revision includes grade 2 changes to secondary mechanisms or scope, not only replacement of the leading explanation. A null average reordering result does not establish invariance for every question, and an addition effect does not establish that updating was scientifically correct.

Seven exploratory analyses examine evidence position, reference choice, initial non-preference, evidence conflict, question structure, confidence changes, and textual complexity. They reveal which conclusions survive alternative measurements and where the corpus cannot resolve an explanatory hypothesis. Adjusting text features leaves the addition effect largely unchanged, but position and conflict analyses depend on much smaller subsets than the nominal 80-question benchmark.

Our contributions are threefold:

1. We evaluate input-related changes relative to a repeated-prompt baseline on the same scientific-content scale, separating observed disagreement from an intervention's incremental association.
2. We characterize multiple levels of explanatory change and their measurement limits through whole-answer grades, preferred-explanation states, and seven targeted diagnostics.
3. We provide a traceable five-domain corpus, same-question configuration comparisons, and offline analysis materials that distinguish stability and consensus from scientific correctness.

2 Questions, experimental design, and evaluation

Corpus and response protocol. Science80 contains 16 questions in each of physics, chemistry, biology, Earth and planetary science, and atmospheric science. Each question has four short evidence statements, a designated deletion, an additional statement, and a mechanism-comparison focus. This is a curated convenience sample, not a random sample of scientific problems. The supplied evidence need not distinguish the candidate mechanisms. Responses contain a primary hypothesis, an alternative, a causal chain, confidence from 0 to 100, and a discriminating experiment. Instructions request a provisional preference while permitting explicit non-preference when the evidence is insufficient. They prohibit browsing, tools, and presenting speculation as established fact. This instruction itself can influence how often models select a working hypothesis.

Runs and completeness. Each slot uses an independent context and receives no responses from other conditions (Table 1). Base is a predetermined slot, not a selected representative answer. Astra

and Gemini each provide 1,280 complete responses; Claude provides 1,242, leaving 38 of its planned slots missing. Astra uses xhigh throughout, Gemini uses high, and Claude uses xhigh on P01 and medium elsewhere. Provider receipts also indicate mixed Claude model names for some runs. Client settings, recovery rules, and receipts are documented in Appendix A. These are comparisons of recorded configurations, not a controlled ranking of immutable model snapshots. More changes content, input length, and final position together. Less and More each have only one realization, so their condition-specific generation distributions are not estimated.

Two complementary endpoints. Three separate Astra xhigh judging contexts compare the four textual fields of Astra responses, with confidence, model identity, and experimental condition hidden. Answer identifiers and pair direction are randomized. A majority grade requires at least two equal ratings; otherwise the comparison is unresolved, denoted U. The rubric uses the highest substantive level of change: 0, equivalent scientific content; 1, peripheral description or experimental implementation; 2, secondary mechanism, causal detail, scope, or non-primary candidate priority; 3, provisional decision or preference within a mechanism family; and 4, preferred mechanism family, dominant role, or main causal direction. A switch between non-preference and provisional preference is grade 3. Cautious wording alone is not such a switch. The rubric and boundary examples appear in Appendix B.

The second endpoint codes only the primary-hypothesis field into a shared, question-specific explanatory state. A single anonymous judging context creates each codebook and assigns labels across configurations. Synonymous descriptions are merged; explicit preference, non-preference, and conditional partitions remain distinct. Unknown state "?" is not a substantive tie and is not assigned zero change. State disagreement is exploratory and is not interchangeable with a whole-answer grade. Independent judging contexts are not independent model families or domain experts.

Aggregation and inference. Let g_{ic} be the grade for condition c against Base on question i. For Shuffle, we first average valid repetitions within each ordering p and then average the valid orderings; Repeat averages the three original-prompt comparisons:

$$S_i = \frac{1}{|P_i|}\sum_{p\in P_i}\frac{1}{|R_{ip}|}\sum_{r\in R_{ip}} g_{ipr}, \qquad R_i = \frac{1}{|R_i^{\text{valid}}|}\sum_{r\in R_i^{\text{valid}}} g_{ir}, \qquad \Delta = \frac{1}{|I|}\sum_{i\in I}(S_i - R_i). \quad (1)$$

Here I contains questions valid at both endpoints. Less and More use the same paired construction. Replacing g with $\mathbb{1}[g \geq 2]$ yields the substantive-revision endpoint. Shared Base responses are not treated as independent questions. Main confidence intervals use 10,000 question-level bootstrap draws stratified by domain; primary tests use a centered studentized bootstrap, with Holm correction across three interventions (Efron, 1979; Holm, 1979). The original sign-flip analysis is retained as a sensitivity analysis, with its symmetry assumption stated. Extensions use 5,000 draws. Reported intervals are pointwise and unadjusted; Holm correction applies to the specified families of p-values. Rules for the extensions were fixed before their aggregation, but these remain post hoc analyses.

3 Do input changes exceed repeated-generation variability?

Of 1,200 whole-answer comparisons, 1,188 have a numerical majority and 12 remain U. All three judges agree exactly on 936 comparisons; pairwise quadratic-weighted agreement is 0.929, 0.906, and 0.882. This repeatability does not rule out shared judging biases. Figure 1 combines grade distributions, paired contrasts, field-level changes, and cross-configuration state comparisons while keeping their endpoints distinct.

Under Repeat, the mean grade is 1.773 and the question-weighted rate of grade 2 or higher is 58.3%. These are properties of repeated explanations, not scientific error rates. Shuffle has mean grade 1.726. Its paired difference from Repeat is -0.047 (95% CI $[-0.144, 0.049]$), with no significant average increment after correction. This result is not an equivalence test and does not establish that every question is insensitive to ordering.

More has a mean-grade increment of 0.411 (95% CI $[0.205, 0.622]$; Holm-adjusted $p = 0.0024$), while Less has an increment of 0.116. On the grade-≥ 2 endpoint, More exceeds Repeat by 19.83 percentage points on the 79 valid paired questions (95% CI $[12.24, 27.64]$). Different effective

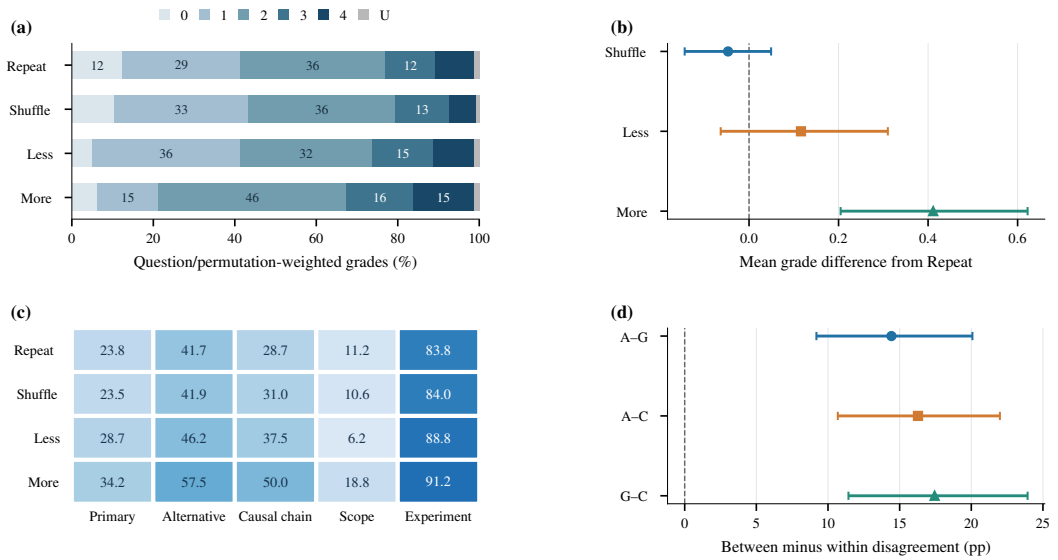

Figure 1: Whole-answer variation and primary-state disagreement. (a) Astra grade distributions, with U unresolved. (b) Paired mean-grade differences from Repeat; bars are pointwise 95% intervals. (c) Field-specific revision rates; each field excludes its own unresolved labels. (d) Between-configuration minus within-configuration primary-state disagreement. Panels use different endpoints and should not be interpreted as accuracy.

denominators mean that this contrast should not be reconstructed by subtracting two marginal percentages. Appendix B reports the complete test results and worst-case assignments for unresolved grades.

Field-level measurements locate the changes. Under Repeat, the primary-hypothesis field changes in 23.8% of weighted comparisons, while the proposed experiment changes in 83.8%. Under More, the corresponding values are 34.2% and approximately 91%. Each field excludes its own unresolved labels. Changing an experiment's implementation or conditional prediction is not automatically a reversal of the preferred theory. Across the four original-prompt responses, between-configuration primary-state disagreement exceeds within-configuration disagreement on average for all three pairs. This describes different explanatory coverage, not which explanation is true.

4 What do the additional diagnostics explain?

4.1 Evidence direction and position

An average Shuffle effect cannot identify how supporting evidence at a particular position affects a preference. We therefore annotate evidence independently of the responses. The annotator sees the question, comparison focus, and four evidence statements in randomized order. It selects candidates A and B, marks suitability for the binary comparison, and labels each statement as supporting A, supporting B, or neutral, with strength 0–3. It never sees response text, confidence, condition, or change grades. A separate mapping step assigns existing anonymous response states to A, B, tie, other, or unknown. These are context-isolated model annotations, not expert-verified evidence labels.

Of 59 questions judged suitable for the selected binary contrast, 11 contain directional evidence, including only one with support on both sides. Across all 320 statements, 306 are marked neutral relative to their specified candidate pair. Neutrality here does not mean that the scientific background is uninformative in every respect. Figure 2 makes this restriction visible before reporting positional associations.

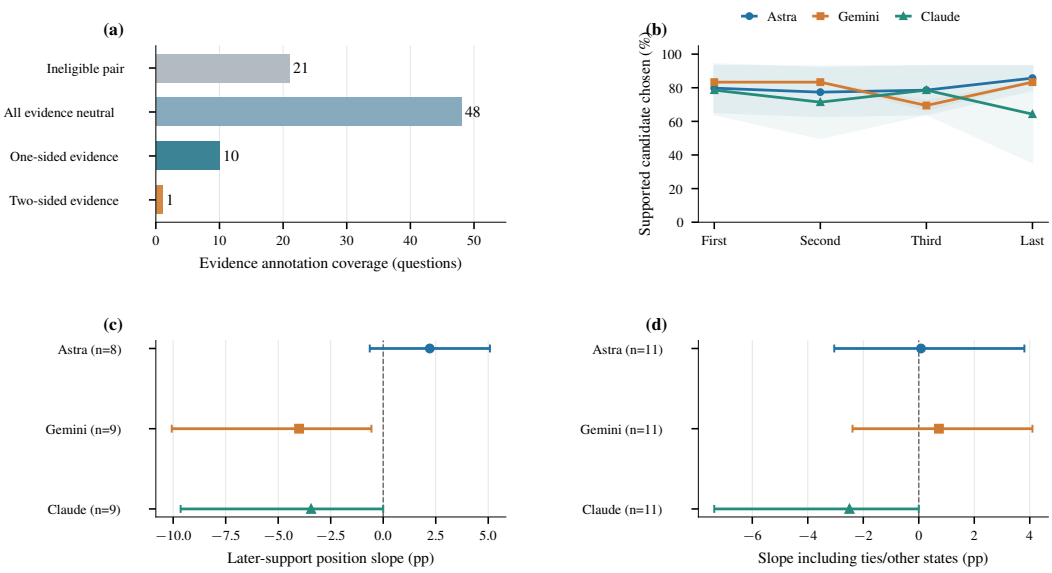

Figure 2: Coverage and positional diagnostics. (a) Questions eligible for the selected candidate contrast and directionality of their evidence. (b) Selection of the supported candidate by evidence position in the valid subset. (c) Position-index slopes conditional on choosing A or B. (d) Slopes including ties and other states. Error bars are pointwise 95% intervals; none of the nine planned tests survives Holm correction.

Assign positive signed strength w_{ij} to evidence supporting A and negative strength to evidence supporting B. For ordering p, define

$$z_{ip} = \frac{\sum_j w_{ij}(\text{pos}_{ipj} - 2.5)}{\sum_j |w_{ij}|}. \tag{2}$$

Among responses choosing A or B, a question-fixed-effect model estimates the probability of choosing A as z increases. Repetitions are averaged within ordering. Astra, Gemini, and Claude have only 8, 9, and 9 valid questions; estimated slopes are 2.22, -4.01, and -3.43 percentage points per unit of z. None of the nine planned position/conflict/decision tests is significant after Holm correction. A pointwise interval can exclude zero while the corrected test does not reject.

Including ties and other states and using "chooses A" as the endpoint changes the slopes to 0.08, 0.73, and -2.50 percentage points. Gemini's sign changes under this inclusion rule. The data therefore do not establish a general primacy or recency effect. Because several evidence items move simultaneously, a slope on z is not a causal effect of moving one isolated item.

4.2 Reference choice, non-preference, and reported confidence

State-based estimates can depend on which original-prompt response is the reference. We rotate Base, R1, R2, and R3, requiring all four original states to be valid. This analysis includes 80 Astra, 80 Gemini, and 70 Claude questions; More additionally requires a valid More state. It reuses the primary-state codebook rather than regrading entire answers. Figure 3 reports the resulting sensitivity alongside confidence and decision-state diagnostics.

Astra's Shuffle-minus-Repeat estimate ranges from -2.85 to 2.71 percentage points across references, and More-minus-Repeat ranges from 7.08 to 14.58. The corresponding ranges are $[-3.96, -0.56]$ and $[6.67, 15.83]$ for Gemini, and $[1.04, 4.33]$ and $[-0.95, 11.43]$ for Claude. A reference-dependent sign or magnitude should not be mistaken for a reference-free property of a question. Detailed intervals and common denominators appear in Appendix E.

The confidence diagnostic crosses primary-state equality with small/large absolute confidence differences. The main threshold is 10 points, with checks at 5 and 15. Repeat and Shuffle receive

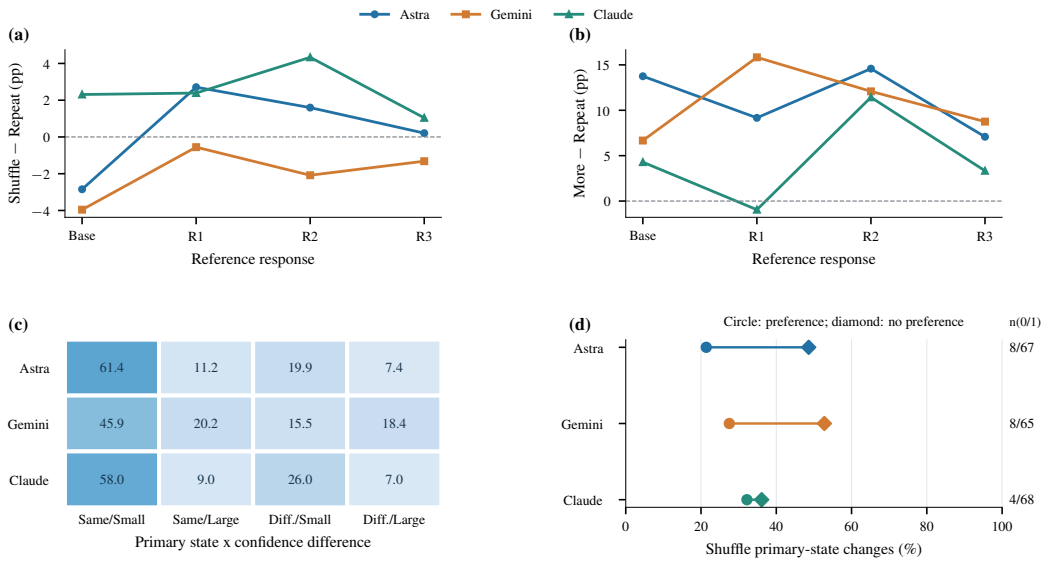

Figure 3: Reference and confidence diagnostics. (a) Primary-state Shuffle-minus-Repeat estimates under four original-prompt reference responses. (b) More-minus-Repeat under the same reference rotation. (c) State equality versus absolute confidence change below/at least 10. (d) Shuffle state-change rates grouped by Base preference; diamonds denote the small no-preference groups. Intervals are pointwise.

equal family weight, with question and ordering weights applied within families. State changes accompanied by a confidence difference below 10 account for 19.9%, 15.5%, and 26.0% for Astra, Gemini, and Claude. Even exactly unchanged confidence can accompany a different explanatory state: among raw state-changing pairs with readable confidence, the counts are 48/336, 85/405, and 168/368. These raw denominators differ from the weighted quadrant denominator and are not independent error rates. Confidence in a newly preferred explanation also need not refer to the same proposition as confidence before the switch.

We further compare questions whose Base state has no preference with those having a provisional preference. The no-preference groups contain only 8, 8, and 4 questions. For whole-answer judging, the corresponding Astra group contains four questions; moreover, grade 3 already includes a decision-state change. These selected, small groups do not establish that initial indecision causes instability.

4.3 Evidence conflict and question structure

The hypothesis that balanced competing evidence predicts instability requires questions with evidence on both sides. Let a_i and b_i be the total strengths supporting A and B, and define

$$C_i = \frac{2\min(a_i, b_i)}{a_i + b_i}, \qquad a_i + b_i > 0. \tag{3}$$

Questions with no directional evidence have undefined conflict, not maximal balance. Of the 11 eligible directional questions, ten have $C_i = 0$ and only C03 has nonzero conflict ($2/3$). A correlation from this coverage cannot test a general relationship between balanced evidence and instability. Appendix E preserves the annotations rather than replacing an unidentifiable test with a positive finding.

C03 concerns selectivity in heterogeneous aqueous reactions. Its annotation contrasts local enrichment, mass transfer, and effective-collision differences (A) with interfacial-solvation changes in transition-state barriers (B). The directional classification concerns that chosen pair, not every mechanistic explanation in the literature. Likewise, question-structure groups are highly unbalanced: dominance (64), source/identity (8), causal direction (5), pathway details (2), and coex-

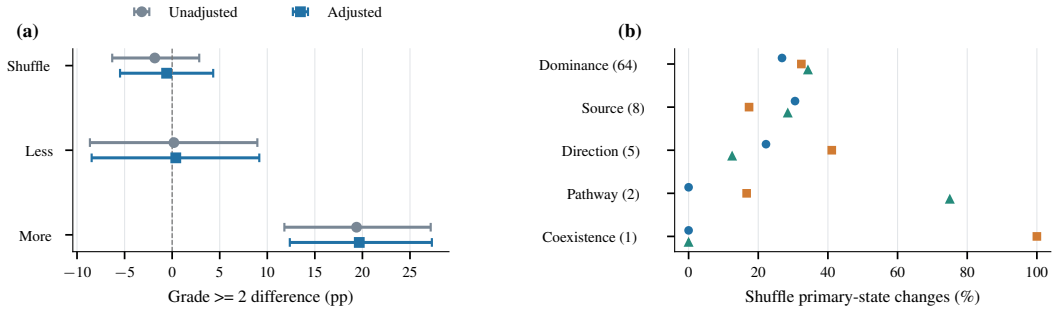

Figure 4: Sensitivity analyses. (a) Question-fixed-effect grade-at-least-2 coefficients before and after adjustment for absolute text-feature differences; error bars are pointwise 95% bootstrap intervals. (b) Descriptive primary-state changes under Shuffle across unbalanced question types; labels show valid question counts. Neither analysis establishes a causal mechanism.

istence/coupling (1). Figure 4(b) is descriptive; it does not rank these types or provide reliable group-level generalization.

4.4 Does textual complexity explain the observed revisions?

A longer answer may contain more mechanisms, conditions, and experimental details for a judge to compare. We fit question-fixed-effect sensitivity models for the grade-≥ 2 indicator. Each question has normalized total weight; repetitions and orderings retain their within-family averaging. Covariates are the absolute differences from Base in log-transformed character length, counts of a fixed hedge vocabulary, and counts of the two annotated candidate families. The candidate count is a lexical proxy taking values 0–2, not a complete inventory of mechanisms. All three regressors are absolute, not signed, differences.

In this regression, the More coefficient changes from 19.38 to 19.67 percentage points after adjustment (95% CI $[12.37, 27.30]$); Shuffle changes from -1.83 to -0.58, with its interval still crossing zero. The unadjusted regression estimate differs from the paired 19.83-point contrast because its weights and missing-data handling differ. These text proxies do not account for the addition-associated revisions. However, output length and hedging may themselves be consequences of an input change, so adjustment is a sensitivity analysis, not a causal mediation or ablation claim. Appendix E gives the exact regressors and grade-≥ 3 sensitivity results.

Scientific interpretation still requires returning to the answers. In C11, Astra prefers CO–CO coupling in both Base and More, but acknowledges CO–CHO dominance under a particular crystal-facet synergy in More. All three judges assign grade 2. A global preferred mechanism plus a conditional exception is distinct from replacing that global preference. In P05, repeated outputs within a configuration can stabilize around mechanisms that disagree across configurations. Neither consistency nor consensus independently validates the scientific explanation; the proposed discriminating experiments remain suggestions to be tested.

5 Related work and scope of the evidence

Prompt sensitivity and input dependence. Prompt formatting, few-shot demonstration order, premise order, and long-context position can affect model performance (Sclar et al., 2024; Lu et al., 2021; Chen et al., 2024; Liu et al., 2023). Calibration and irrelevant-context studies similarly show that output depends on presentation and surrounding material (Zhao et al., 2021; Shi et al., 2023). Science80 does not dispute these findings or infer universal order invariance. It asks whether revisions in an open scientific explanation exceed repeated-generation variation under the same response protocol. The baseline, semantic resolution, and question-level weighting determine which comparison is being made. Positional associations are further restricted to questions whose evidence supports an interpretable candidate contrast.

Semantic uncertainty and confidence. Self-consistency and semantic aggregation use multiple generations to study reasoning and uncertainty (Wang et al., 2022; Kuhn et al., 2023; Farquhar et al., 2024). Research on self-correction shows that another generation need not supply a correct correction (Huang et al., 2023). Confidence elicitation and calibration distinguish what models know from what they report (Kadavath et al., 2022; Lin et al., 2022; Xiong et al., 2023; Yona et al., 2024). Our quadrant analysis relates an expressed score to a question-specific explanatory state, rather than treating a stable number as a stable proposition. It does not estimate calibration against a scientific ground truth. Nor does the five-level rubric establish superiority to semantic entropy or entailment-based metrics; their incremental value requires an independent comparative annotation study.

Scientific assistance and model-based evaluation. Knowledge benchmarks evaluate established answers (Hendrycks et al., 2020; Rein et al., 2023), while scientific-assistance systems seek hypotheses that can be evaluated through further research (Gottweis et al., 2026). Our object is the stability of conditional explanations under a specified evidence packet, not the discovery or verification of natural laws. Broad evaluation frameworks (Liang et al., 2022) and model judging (Zheng et al., 2023; Liu et al., 2023) support structured assessment, but judging can be sensitive to presentation and semantic granularity (Wang et al., 2023). An output causal chain is not a faithful record of internal reasoning (Turpin et al., 2023). Repeated disagreement is also related to work on time to inconsistency under perturbations (Li et al., 2025); here, the original-prompt baseline and the scientific-content scale define a different empirical question.

Annotation independence and uncertainty. The extension uses 80 evidence-only annotation calls and 16 mapping calls, all Astra xhigh without tools. Mapping sees frozen candidates and anonymous codebooks but not configuration, condition, confidence, or frequencies. This isolates information between tasks; it is not cross-family or expert validation. New annotations do not replace the original 1,200 grades or response texts. The nine-test extension family retains unestimable conflict tests as non-rejections for multiplicity accounting. Bootstrap intervals quantify question resampling under the recorded labels, not uncertainty in candidate construction or annotation error. No resampling scheme makes this convenience corpus representative of all science.

6 Discussion and Conclusion

Science80 shows why "the answer changed" is an incomplete conclusion for unresolved scientific questions. Under an unchanged prompt, the grade-≥ 2 revision rate is already 58.3%. Against that baseline, evidence reordering has no detected average increment after correction, whereas adding a statement produces additional revisions. The contrast survives adjustment for limited textual proxies, but it does not determine whether the revisions improve the science. The practical lesson is to measure ordinary repeated-generation variation before interpreting an observed response to new input.

The diagnostics clarify what stability measures leave open. Estimates depend on reference choice and semantic resolution; small confidence differences can coexist with different preferred explanations. Apparent richness of a benchmark is also different from evidence for a particular hypothesis: 80 scientific questions yield only 11 eligible directional-evidence questions, and only one with support on both sides. Consequently, the current data cannot establish a general relationship between evidence conflict and instability, or a universal primacy/recency mechanism.

Three limitations determine the interpretation. First, content grades and states lack independent domain-expert validation, and the same model family generates and judges the main responses. Second, the protocol encourages provisional selection while allowing non-preference. Variation may partly reflect reasonable exploration among underdetermined candidates rather than unreliable belief revision. Third, More bundles content, length, and position, and Less/More lack within-condition repetition. Configuration heterogeneity, Claude missingness, mixed receipts, and unequal length compliance further limit cross-model generalization. All observations concern Chinese prompts and answers.

The most direct next tests are independent validation of the rubric, a contrast between provisional-selection and explicit non-preference prompts, and paired additions of discriminating evidence, neu-

tral material, and repeated information at balanced positions. These tests would address whether the measured changes matter scientifically and which input property drives them. Until then, the supported conclusion is narrower: scientific explanations exhibit substantial repeated-generation variability, and some input changes add revisions beyond that baseline. Explanatory stability, responsiveness to evidence, cross-model agreement, and scientific correctness require distinct validation.

AI USE STATEMENT

Generative AI tools assisted the formulation and refinement of the study, experiment and rubric design, data preparation and recovery, implementation of analysis and visualization code, interpretation, drafting, translation, and editing. The scientific responses are model-generated research data. GPT-6 Astra xhigh was also used for the recorded grading, state coding, evidence annotation, and mapping described in the paper. Thus, agreement among judges is not human-expert validation. The release includes the retained prompts, responses, labels, and analysis code, and its numerical summaries were checked by offline recomputation. These checks verify consistency with the archived records; they do not certify the scientific correctness of the generated hypotheses. The authors remain responsible for the submission and its claims.

REPRODUCIBILITY STATEMENT

Appendices A–E specify the protocol, recovery rules, evaluation, question-level results, examples, and extensions. The supplementary ZIP contains the original Chinese question packets and response text, anonymous judging inputs and outputs, provider provenance with identifying operational metadata removed, offline analysis and plotting code, an English README, and checksums. It permits recomputing the reported statistics without another model call. Hosted-model generation itself need not be repeatable because service snapshots and stochastic behavior are not controlled. English examples in the appendix are reading translations, not additional model responses.

ETHICS STATEMENT

The study evaluates model-generated explanations and does not involve human-subject experiments. Open scientific questions and hypothetical discriminating experiments are presented as research material, not established findings or professional advice. Public release preserves uncertainty, missingness, and annotation limitations. Operational user identifiers, private filesystem locations, and conversation identifiers are excluded from the anonymous supplement. No correctness or model-capability ranking is inferred solely from stability.

REFERENCES

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. Premise Order Matters in Reasoning with Large Language Models. *arXiv preprint*, 2024. https://arxiv.org/abs/2402.08939.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics, 7(1):1–26*, 1979. https://doi.org/10.1214/aos/1176344552.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature, 630:625–630*, 2024. https://doi.org/10.1038/s41586-024-07421-0.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, et al.. Accelerating scientific discovery with Co-Scientist. *Nature*, 2026. https://doi.org/10.1038/s41586-026-10644-y.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *arXiv preprint*, 2020. https://arxiv.org/abs/2009.03300.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics, 6(2):65–70*, 1979. https://www.jstor.org/stable/4615733.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large Language Models Cannot Self-Correct Reasoning Yet. *arXiv preprint*, 2023. https://arxiv.org/abs/2310.01798.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, et al.. Language Models (Mostly) Know What They Know. *arXiv preprint*, 2022. https://arxiv.org/abs/2207.05221.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. *arXiv preprint*, 2023. https://arxiv.org/abs/2302.09664.

Yubo Li, Ramayya Krishnan, and Rema Padman. Time-To-Inconsistency: A Survival Analysis of Large Language Model Robustness to Adversarial Attacks. *arXiv preprint*, 2025. https://arxiv.org/abs/2510.02712.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, et al.. Holistic Evaluation of Language Models. *arXiv preprint*, 2022. https://arxiv.org/abs/2211.09110.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching Models to Express Their Uncertainty in Words. *arXiv preprint*, 2022. https://arxiv.org/abs/2205.14334.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *arXiv preprint*, 2023. https://arxiv.org/abs/2303.16634.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the Middle: How Language Models Use Long Contexts. *arXiv preprint*, 2023. https://arxiv.org/abs/2307.03172.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. *arXiv preprint*, 2021. https://arxiv.org/abs/2104.08786.

Belinda Phipson, and Gordon K. Smyth. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. *Statistical Applications in Genetics and Molecular Biology, 9(1), Article 39*, 2010. https://gksmyth.github.io/pubs/PermPValuesPreprint.pdf.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A Graduate-Level Google-Proof Q&A Benchmark. *arXiv preprint*, 2023. https://arxiv.org/abs/2311.12022.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting. *International Conference on Learning Representations*, 2024. https://arxiv.org/abs/2310.11324.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed Chi, Nathanael Schärli, and Denny Zhou. Large Language Models Can Be Easily Distracted by Irrelevant Context. *arXiv preprint*, 2023. https://arxiv.org/abs/2302.00093.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *arXiv preprint*, 2023. https://arxiv.org/abs/2305.04388.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-Consistency Improves Chain of Thought Reasoning in Language Models. *arXiv preprint*, 2022. https://arxiv.org/abs/2203.11171.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, et al.. Large Language Models are not Fair Evaluators. *arXiv preprint*, 2023. https://arxiv.org/abs/2305.17926.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs. *arXiv preprint*, 2023. https://arxiv.org/abs/2306.13063.

Gal Yona, Roee Aharoni, and Mor Geva. Can Large Language Models Faithfully Express Their Intrinsic Uncertainty in Words?. *arXiv preprint*, 2024. https://arxiv.org/abs/2405.16908.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate Before Use: Improving Few-Shot Performance of Language Models. *arXiv preprint*, 2021. https://arxiv.org/abs/2102.09690.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, et al.. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. *arXiv preprint*, 2023. https://arxiv.org/abs/2306.05685.

A PROTOCOL, PROVENANCE, AND RECOVERY

Original input and response structure. The exact experimental language is Chinese. Each input contains a common instruction, an evidence list, and the question; condition identifiers are not presented as scientific evidence. The instruction asks the model to reason conditionally from supplied information and existing scientific knowledge, use no search or tools, and avoid portraying speculation as fact. It asks for a provisional favorite while explicitly allowing "multiple mechanisms cannot yet be distinguished." The five JSON fields are `main` (at most 80 Chinese characters), `alternative` (80), `chain` (160), integer `confidence` (0–100, explicitly not a calibrated probability), and `experiment` (160). A proposed experiment should distinguish the two explanations and specify contrasting expected outcomes. The explanatory chain is visible answer content, not access to hidden reasoning.

Condition construction. Base and R1–R3 share an identical visible prompt. Six non-original permutations are stored separately for each question, with two repeated three times and the others once. The condition string identifies the permutation and repetition; it must not be interpreted as a universally fixed ordering. Less removes the specified zero-based evidence index; More appends its designated statement. The question, instruction, and output schema remain fixed. Each slot starts an independent context, so this is not a multi-turn updating dialogue. The full prompts are provided with SHA-256 hashes in the supplement. English question titles and case summaries are reading aids; they were not supplied to the models.

Configuration attribution. The main 80-question Astra dataset comprises 16 P01 pilot slots and 1,264 suite slots, with response-side model attribution `gpt-6-astra` and xhigh effort. Gemini is recorded as `gemini-3.1-pro-preview`, high. Claude uses xhigh for P01 and medium otherwise; effective service receipts are heterogeneous. The archive contains 1,085 Claude receipts naming Fable only, 189 naming both Fable and Opus (172 complete), and six without model-usage entries. These are client/receipt descriptions, not verified immutable server weights. All 1,280 archived Astra trajectory request-target fields retain a reused Gemini template target, whereas the associated response-side records identify Astra. We therefore preserve the exact common prompt but identify configurations using response-side records, and disclose the mismatch in the provenance audit. The legacy 30-question Astra low dataset is kept separately for historical reproducibility; none of its 480 answers replaces a main xhigh answer.

Recovery and missingness. Astra has 1,275 directly parsed answers, four trailing-comma repairs, and one conversion of the token "sixty" to the numeric confidence 60 (C04/S5r1). No scientific text is completed or rewritten. Gemini has 1,280 complete answers, two exceeding requested field lengths. Claude has 1,135 directly parsed and 107 recovered complete answers, with 38 slots still missing; 509 complete answers exceed at least one requested field length. Recovery accepts complete original fields after removing wrappers/extra keys or repairing delimiters. It does not finish truncated answers. Complete overlength responses remain available and are not automatically graded as scientific failures. Missing answers and unknown labels are excluded at their relevant endpoint, never filled as agreement.

Audit boundary. Original archive checks verified 2,560 Astra slot files and 5,120 comparison-configuration slot files against manifests. Hash consistency establishes agreement with those archives; it cannot authenticate the hosted model's internal weights or certify scientific correctness. The anonymous release removes account names, filesystem paths, command lines, conversation identifiers, and raw transport logs. It retains scientific text, model/configuration receipts, timings, token records, prompt hashes, judging prompts, and labels. An English README distinguishes original data, reading translations, and derived statistics.

B GRADING AND INFERENTIAL DETAILS

Operational scale. Judges compare primary and competing explanations, causal chain, applicable scope, and proposed experiment. The highest substantive revision determines the ordinal level. Grade 0 means equivalent scientific content despite wording; 1 covers peripheral elaboration or experimental implementation; 2 covers secondary mechanisms, causal detail, scope, or non-primary priority; 3 covers non-preference versus provisional preference or within-family preference; 4 covers the leading family, dominance among potentially coexisting mechanisms, or principal causal direction. Uncertainty wording alone neither cancels an explicit preference nor creates a preference change. Each field also has a separate change flag. Three independent contexts of the same model family judge anonymous pairs; exact numerical majority is used, not a median or a forced quota of each grade. Disagreement without two matching numbers remains U.

Table 2: Raw majority-grade counts and question/permutation-weighted summaries. Counts and weighted percentages have different denominators.

Condition	0	1	2	3	4	U	Mean	Grade ≥ 2 (%)
Repeat	30	69	86	29	23	3	1.773	58.3
Shuffle	85	266	283	103	56	7	1.726	56.6
Less	4	29	26	12	8	1	1.886	58.2
More	5	12	37	13	12	1	2.190	78.5

Table 3: Paired Astra contrasts. Intervals are pointwise; Holm correction is applied separately across the three interventions for each endpoint. Sign-flip tests are sensitivity analyses.

Vs. Repeat	Endpoint	n	Difference	95% CI	$p_{\text{boot,H}}$	$p_{\text{flip,H}}$
Shuffle	Grade	80	-0.047	[-0.144, 0.049]	0.5019	0.4919
Shuffle	≥ 2 (pp)	80	-1.70	[-6.32, 2.95]	0.9909	1.0000
Less	Grade	79	0.116	[-0.063, 0.310]	0.5019	0.4919
Less	≥ 2 (pp)	79	0.42	[-8.23, 9.28]	0.9909	1.0000
More	Grade	79	0.411	[0.205, 0.622]	0.0024	0.0019
More	≥ 2 (pp)	79	19.83	[12.24, 27.64]	0.0006	0.0001

Resampling and uncertainty. Questions, not output pairs, are the resampling units. Main intervals use 10,000 domain-stratified question-bootstrap draws. The centered studentized bootstrap tests a mean-zero contrast. A documented amendment introduced this test before full aggregation, while retaining the initially specified 50,000-draw sign-flip results. Sign flips require a symmetric null distribution and are not treatment-randomization tests here. Monte Carlo probabilities use the plus-one construction $(b+1)/(B+1)$ rather than zero (Phipson and Smyth, 2010). Holm correction controls the stated family of tests (Holm, 1979); displayed percentile intervals are not simultaneous intervals. Neither procedure accounts for unmeasured annotation bias or makes the convenience sample representative.

Unresolved grades. Assigning each U the endpoint of its allowable 0–4 range that minimizes or maximizes the full-80-question mean contrast gives Shuffle $[-0.088, -0.008]$, Less $[0.063, 0.163]$, and More $[0.362, 0.463]$. These are missing-label bounds for this fixed corpus, not confidence intervals for a population mean. They therefore need not match the signs allowed by resampling intervals. Exact unanimity occurs in 936/1,200 comparisons, and numerical majority in 1,188. Quadratic-weighted pairwise kappas are 0.929, 0.906, and 0.882, conditional on readable labels. Agreement is annotation repeatability, not expert validity.

C QUESTION INDEX AND PER-QUESTION RESULTS

The following index reports Astra mean whole-answer grades against Base: R is Repeat, S is Shuffle, L is Less, and M is More. S first averages within ordering, then across valid orderings; it is not the raw ten-slot average. A dash is unresolved, not zero. Titles are English reading translations; original questions, four evidence statements, deletion indices, additions, permutations, and every response are included in the supplementary JSON files. The machine-readable CSV additionally supplies configuration-specific primary-state revision rates and completeness counts. These per-question summaries are descriptive, without individual significance claims.

ID	Question	R	S	L	M
P01	Microscopic identity of dark matter	2.67	2.33	3.00	2.00
P02	The dominant source of the Hubble tension	0.00	0.44	4.00	0.00
P03	Information release from evaporating black holes	1.67	1.94	2.00	2.00
P04	Constant versus dynamical dark energy	1.33	0.67	0.00	–
P05	Origin of the matter–antimatter asymmetry	1.67	0.89	4.00	0.00
P06	Why neutrino masses are so small	1.33	1.25	1.00	2.00
P07	Pairing in cuprate high-temperature superconductors	2.67	2.17	3.00	3.00
P08	Causality between the pseudogap and charge order	2.00	1.89	2.00	2.00
P09	Linear resistivity in strange metals	2.00	3.11	2.00	4.00
P10	Pairing in twisted bilayer graphene	4.00	3.28	4.00	4.00
P11	Deconfined quarks in neutron-star cores	2.00	1.83	1.00	2.00
P12	Sources of ultrahigh-energy cosmic rays	4.00	3.28	4.00	2.00
P13	Location of coherent fast-radio-burst emission	2.00	1.72	1.00	2.00
P14	Dominant heating channels of the solar corona	3.00	2.78	3.00	2.00
P15	Perpendicular ion heating in the collisionless solar wind	3.00	2.50	2.00	2.00
P16	Seeds of large-scale cosmic magnetic fields	1.67	1.33	1.00	2.00
C01	Origin of molecular homochirality	2.50	2.72	3.00	3.00
C02	Microscopic mechanism of chirality-induced spin selectivity	2.00	2.00	2.00	2.00
C03	Anomalous selectivity in heterogeneous aqueous reactions	1.00	1.50	1.00	4.00
C04	Structural origin of the glass transition	1.33	1.00	1.00	2.00
C05	A liquid–liquid critical point in supercooled water	2.00	1.06	1.00	4.00
C06	How water vapor forms ice on particles	1.33	1.72	2.00	2.00
C07	Acid/base ion distributions at water surfaces	3.00	3.00	3.00	3.00
C08	Reaction acceleration in microdroplets	1.00	1.06	2.00	1.00
C09	Vibrational strong coupling and ground-state reactions	1.00	1.00	1.00	1.00
C10	Thermal/nonthermal plasmonic photocatalysis	1.00	1.00	1.00	1.00
C11	Carbon–carbon bonding in copper CO2 reduction	2.00	2.00	2.00	2.00
C12	Lattice oxygen participation in oxygen evolution	2.00	2.00	2.00	2.00
C13	Nitrogen binding and reduction by nitrogenase	0.33	0.28	1.00	0.00
C14	Stability of lithium-metal interphases	2.00	1.89	1.00	3.00
C15	When metal-containing solid-state-battery interfaces help	3.00	2.94	3.00	3.00
C16	Microscopic control of enzymatic rate enhancement	1.33	1.06	2.00	1.00
B01	Barriers to adult mammalian regeneration	4.00	3.50	4.00	2.00
B02	Gut microbiota as an aging driver or feedback amplifier	0.00	0.67	1.00	2.00
B03	Reactivation of dormant cancer cells	1.33	1.50	1.00	1.00
B04	Causal role of transposable-element reactivation in aging	1.33	1.39	1.00	2.00
B05	Protective versus damaging microglia in Alzheimer's disease	2.00	1.39	1.00	2.00
B06	Body-first versus brain-first Parkinson's disease	1.50	1.39	1.00	1.00
B07	Mechanisms of persistent long-COVID symptoms	3.00	3.00	3.00	3.00
B08	Sleep and waste clearance in brain parenchyma	1.67	1.44	1.00	1.00
B09	Astrocyte ensembles and memory traces	2.67	1.56	2.00	1.00
B10	Neural implementation of conscious content	1.00	1.00	1.00	1.00
B11	How enhancers find the correct promoters	1.50	1.22	1.00	2.00
B12	Transcriptional condensates: cause or consequence	2.00	2.22	2.00	2.00
B13	Reading signal duration and frequency	2.00	2.00	–	2.00
B14	Selective synaptic vulnerability in dementia	2.33	2.39	3.00	3.00
B15	Persistence of tumor T-cell dysfunction	1.33	1.78	2.00	2.00
B16	Timing of zygotic genome activation	1.67	2.78	4.00	4.00
E01	Onset of sustained plate tectonics	1.33	1.22	2.00	3.00
E02	Original sources of Earth's water	1.33	1.28	1.00	4.00
E03	Material identity of low-shear-velocity mantle provinces	0.00	0.89	1.00	2.00
E04	Origin of lunar nearside/farside mantle differences	1.33	1.67	2.00	0.00

ID	Question	R	S	L	M
E05	Geodynamo power before inner-core formation	0.67	1.06	1.00	1.00
E06	Origin of inner-core seismic anisotropy	2.67	2.33	3.00	4.00
E07	Rapid large-scale rupture in deep earthquakes	0.67	1.50	1.00	0.00
E08	Why slow earthquakes stop at particular locations	1.67	1.33	1.00	4.00
E09	Active versus passive supercontinent breakup	1.67	2.00	1.00	2.00
E10	Critical threshold of the Great Oxidation Event	2.00	2.06	2.00	2.00
E11	Termination of Snowball Earth	0.67	1.00	2.00	2.00
E12	Inconsistent reports of Martian methane	3.00	3.33	4.00	4.00
E13	Venus's unidentified ultraviolet absorber	3.00	2.06	2.00	4.00
E14	Different heat loss from Uranus and Neptune	0.00	0.78	0.00	1.00
E15	Formation of Jupiter's dilute core	4.00	3.44	4.00	4.00
E16	Overcoming radial drift in planet formation	2.00	1.89	2.00	2.00
A01	Recent declines in ocean low-cloud reflectivity	2.00	1.67	2.00	3.00
A02	Dominant contributions to tropical high-cloud feedback	1.33	1.28	1.00	2.00
A03	Controls of mean tropical rain-belt position	3.00	1.94	3.00	3.00
A04	Aerosol increases/decreases in low-cloud water	0.67	1.00	3.00	3.00
A05	Observed low-cloud trends and long-term feedback	0.67	1.00	1.00	1.00
A06	Spontaneous aggregation of tropical convection	1.67	1.39	2.00	3.00
A07	Triggers of tropical-cyclone rapid intensification	1.00	1.44	1.00	2.00
A08	Ice-nucleation pathways controlling cirrus crystal numbers	1.67	1.44	2.00	2.00
A09	Excess ice in mixed-phase clouds	2.00	1.83	3.00	2.00
A10	Regional variation in MJO propagation speed	2.00	2.00	2.00	2.00
A11	Arctic sea-ice loss and midlatitude weather	1.67	1.67	2.00	2.00
A12	Simulating North Pacific storm-track shifts	2.00	1.83	2.00	2.00
A13	Predicting sudden stratospheric warming	0.67	1.17	0.00	2.00
A14	Initiation of lightning breakdown	1.67	1.22	2.00	2.00
A15	Secondary organic aerosols in humid polluted air	1.33	1.00	1.00	4.00
A16	Super-Clausius–Clapeyron extreme-rainfall scaling	1.33	1.50	0.00	3.00

D TEN ANSWER-LEVEL COMPARISONS

The following cases are illustrative selections covering different domains and change types, not a random sample for estimating prevalence. Lists provide condensed English translations of the original Chinese content. They preserve the stated preference, uncertainty, and main experimental qualification but are not verbatim quotations or new generations. Original full text is indexed by configuration, question ID, and condition in the supplement. Each case compares Astra Base with the named variant. The additional Gemini/Claude summaries always compare their Base and More slots; for repetition cases these are a separate comparison. Grades apply only to the specified Astra pair. Proposed mechanisms and tests remain model claims, not validated scientific findings.

D.1 P05: THE ORIGIN OF THE MATTER–ANTIMATTER ASYMMETRY

Question. Which early-Universe mechanism most plausibly generated the present matter excess, and what distinguishing relics would it leave?

Evidence packet.

E1. Electroweak baryogenesis links transport and CP violation during a phase transition.
E2. Its success depends on the transition, interactions, and transport approximations; a mechanism's name alone is insufficient.
E3. The observable Universe has a clear excess of matter over antimatter.
E4. Baryogenesis must connect microscopic asymmetry to nonequilibrium cosmic evolution.

More statement. Recent electroweak-baryogenesis calculations give different outcomes for different Standard Model extensions, requiring specific new-physics conditions.

The two Astra prompts are identical.

Astra Base (confidence 30).

Primary
Provisionally select new-physics-driven electroweak baryogenesis, without sufficient evidence that it is more likely than high-scale leptogenesis.
Alternative
High-scale leptogenesis: out-of-equilibrium, CP-asymmetric heavy-particle decays produce B-L asymmetry, partly converted to baryons by sphalerons.
Causal chain
A strong first-order transition creates expanding bubble walls; CP violation and transport establish chiral asymmetry ahead of the wall; sphalerons generate baryons, whose excess survives if washout is suppressed inside. Gravitational waves, new particles, and electric dipole moments are possible relics.
Discriminating experiment
Search for a millihertz stochastic gravitational-wave background compatible with an electroweak transition. Detectable-transition parameter regions predict a spectrum; minimal high-scale leptogenesis with a smooth electroweak crossover does not. Detection alone would not prove baryogenesis; non-detection excludes only part of parameter space.

Astra R1 (confidence 35).

Primary
Provisionally favor high-scale leptogenesis: first create lepton asymmetry and then convert it to baryon excess. Its priority is not established.
Alternative
Electroweak baryogenesis in a strong first-order transition, with CP-violating transport and sphalerons near bubble walls and suppressed washout.
Causal chain
If heavy Majorana neutrinos exist, their nonequilibrium CP-violating decays can create B-L asymmetry. Sphalerons convert part into baryon number; sufficiently weak washout permits survival. Majorana neutrinos would provide non-unique supporting evidence.
Discriminating experiment
The same millihertz search distinguishes detectable first-order-transition scenarios from high-scale-leptogenesis scenarios with a smooth crossover, subject to the same non-uniqueness and limited-exclusion caveats.

Same-question configuration comparison.

- *Gemini Base*: Electroweak baryogenesis through new interactions, a strong first-order transition, and additional CP violation.
- *Gemini More*: Electroweak baryogenesis in a Standard Model extension, with gravitational-wave and electric-dipole relics.
- *Claude Base*: Mechanisms remain unresolved; if forced to choose, favor high-scale leptogenesis because the Standard Model crossover and CP violation are insufficient.
- *Claude More*: Mechanisms remain unresolved; favor electroweak baryogenesis requiring new scalars and CP phases.

Interpretation. All three grades are 4. The preferred mechanism changes under identical input despite uncertainty language in both answers. The cross-configuration choices cannot establish which cosmic mechanism is correct.

D.2 P10: PAIRING IN TWISTED BILAYER GRAPHENE

Question. Near the magic angle, is pairing dominated by electronic interactions, phonons, or their cooperation?

Evidence packet.

E1. Superconductivity and other correlated states depend on carrier density and environment.
E2. Electronic and phonon-mediated models can both produce pairing under some conditions.
E3. Changing the dielectric environment changes screening and can also change competition among interactions.
E4. Twisting produces narrow bands and enhances interaction effects.

More statement. A tunable-screening study finds that changing screening can weaken the superconducting region, suggesting that interactions can help or suppress pairing.

Astra Base (confidence 35).

Primary
Mechanisms cannot yet be distinguished; use electronic–phonon cooperation as a working hypothesis without establishing cooperation or a dominant mechanism.
Alternative
Electronic fluctuations provide the main attraction and phonons are secondary; phonon-dominated pairing with electronic modulation of repulsion and competing states also remains possible.
Causal chain
Narrow bands reduce kinetic scales and enhance correlations, potentially also effective phonon pairing. Under cooperation, both determine pairing, while density and screening alter competing states. Environmental sensitivity alone does not identify the mechanism.
Discriminating experiment
Cross carbon-isotope substitution with dielectric screening while matching angle, density, and strain; measure phonon spectra and the gap. Cooperation predicts a systematic phonon-frequency response modulated by screening, whereas a purely electronic mechanism predicts weak isotope response. Indirect effects and compensation weaken discrimination; an isotope effect alone does not prove cooperation.

Astra More (confidence 60).

Primary
Electronic interactions dominate, with phonons cooperating or modulating; current evidence still cannot fully distinguish mechanisms.
Alternative
Phonons dominate pairing; electronic interactions mainly adjust the superconducting window through screening and competing order.
Causal chain
Narrow bands increase the density of states and correlations; dielectric changes alter Coulomb interactions and competing states, potentially enhancing or suppressing pairing. Phonon attraction may cooperate in some parameter regions.
Discriminating experiment
Substitute carbon-12/13 at matched angle, density, and screening; measure transition temperature and gap. The response proposes that a substantial isotope effect supports phonon dominance, whereas weak isotope response with strong screening dependence supports electronic dominance.

Same-question configuration comparison.

- *Gemini Base*: Favor electronic interactions, such as spin or valley fluctuations, while acknowledging underdetermination.
- *Gemini More*: Favor electronic–phonon cooperation; screening sensitivity reflects a balance of repulsion and attraction.
- *Claude Base*: Favor cooperation: correlations determine the parent state and pairing symmetry, while phonons supply or enhance attraction.
- *Claude More*: Favor cooperation: phonons supply pairing attraction and correlations regulate bands and competing phases.

Interpretation. Grades 4/4/4 reflect a change in stated dominance. The More experiment's simplified isotope criterion is reproduced as model output, not endorsed as a uniquely identifying scientific test.

D.3 C15: When metal-containing battery interfaces help

Question. By what mechanism can a metal-containing reaction interface improve, rather than worsen, lithium instability in a solid-state battery?

Evidence packet.

E1. Contact with lithium can form metal- or alloy-containing interface products.
E2. Interfacial electron transport may promote further electrolyte decomposition.
E3. Some metal-containing interfaces improve contact and lithium deposition.
E4. Longer symmetric-cell cycling does not automatically mean higher lithium Coulombic efficiency.

More statement. Recent perspectives argue that metal connectivity, ionic pathways, and mechanical contact must be considered together; metal presence alone does not determine benefit.

The two Astra prompts are identical.

Astra Base (confidence 64).

Primary
Metal is enclosed by an ion-conducting, electron-blocking layer, making the interface self-limiting; wetting and current homogenization outweigh decomposition. Multiple mechanisms remain unresolved.
Alternative
Mechanical gap filling, stress buffering, and contact pressure drive improvement; the response claims that limited lithium inventory masks side reactions and that metal need not chemically stabilize the interface.
Causal chain
Lithium reduces the electrolyte to metal/alloy. If metal is isolated and electron leakage is limited while ions pass, self-passivation reduces contact resistance and hotspots. A percolating electronic network instead sustains decomposition.
Discriminating experiment
At matched pressure, roughness, and electrolyte, vary metal connectivity. Use limited-lithium asymmetric cells to measure electronic/ionic impedance, decomposition current, true Coulombic efficiency, and deposition images. The primary account predicts an optimum nonpercolating window; the mechanical account predicts effects tracking contact or monotonically worsening with connectivity.

Astra R1 (confidence 68).

Primary
Prefer nonpercolating metal/alloy nanoparticles separated by an ion-conducting, electron-blocking interphase; lithiophilicity and compliance homogenize nucleation and stress, making decomposition self-limiting.
Alternative
Geometry, wetting, pressure, or current distribution improve cycling while metal still promotes continuous reduction; symmetric-cell longevity masks low Coulombic efficiency.
Causal chain
Reduction creates metal/alloy; interrupted electronic paths restrict leakage and further decomposition. Metal lowers nucleation barriers, smooths current, and maintains compliant contact. Percolating networks may reverse the benefit.
Discriminating experiment
Control metal continuity and loading while matching interface and pressure; measure leakage, plating/stripping efficiency, and interface thickness. Nonpercolating samples should have low leakage, a decomposition plateau, and improved efficiency; geometric effects should disappear after pressure matching.

Same-question configuration comparison.

- *Gemini Base*: Lithiophilic alloy or mixed conduction improves contact and flux uniformity, but electron leakage still lowers efficiency.
- *Gemini More*: Spatially isolated metal maintains contact and smooths fields while interrupting macroscopic electron paths.
- *Claude Base*: Passivation or interrupted electron percolation lets lithiophilic metal mainly improve wetting and lithium flux.
- *Claude More*: Isolated metal in an ion-conducting, electron-insulating matrix provides local nucleation sites without through-going electronic paths.

Interpretation. Grades 3/3/1 give majority 3, showing a disputed boundary between elaboration and preference. The Base claim that limited inventory masks side reactions deserves domain review; it is preserved in the original data, not silently repaired.

D.4 C11: THE FIRST CARBON–CARBON BOND IN COPPER CO2 REDUCTION

Question. Which intermediate combination most plausibly forms the first key carbon–carbon bond when copper catalysts produce multicarbon products?

Evidence packet.

E1. Copper electrocatalysts can reduce carbon dioxide to multicarbon products.
E2. Potential, local acidity, facets, and intermediate coverage jointly affect selectivity.
E3. Adsorbed-CO-related intermediates are precursors in several proposed mechanisms.
E4. Final products alone cannot distinguish CO–CO coupling from coupling involving more hydrogenated intermediates.

More statement. Recent facet-synergy work proposes that different intermediate stabilization at adjacent sites can favor CO–CHO-related bond formation.

Astra Base (confidence 55).

Primary
Favor coupling of adjacent adsorbed CO to OCCO; CO with CHO or CH2O can run in parallel under high proton supply.
Alternative
CO is first protonated to CHO or CH2O and then couples with neighboring CO to form a C2 oxygenate intermediate.
Causal chain
Potential and local pH regulate CO coverage and protonation. High coverage favors adjacent CO coupling, followed by protonation and deoxygenation. Faster protonation instead favors the hydrogenated-intermediate route.
Discriminating experiment
At matched CO coverage, replace water with heavy water and vary local pH while measuring C2 rates and selectivity. The response predicts weaker initial kinetic isotope effects for CO–CO and stronger isotope/pH effects for CO–CHO or CO–CH2O.

Astra More (confidence 55).

Primary
Still provisionally favor CO–CO coupling; evidence does not establish the dominant channel, and CO–CHO may dominate under facet synergy.
Alternative
CO is hydrogenated to CHO before coupling with adjacent CO; distinct facets may stabilize the two intermediates.
Causal chain
Carbon dioxide reduction forms CO, adjacent CO forms OCCO, and proton/electron transfer yields multicarbon products. Potential, pH, facets, and coverage change barriers and may favor hydrogenation before coupling.
Discriminating experiment
Use isotope-assisted, potential-step, time-resolved operando infrared measurements of the earliest bond-forming species. CO–CO predicts OCCO before hydrogenated C2 species; CO–CHO predicts CHO formation followed by consumption with OCCHO formation. Failure to detect a short-lived intermediate alone cannot reject a pathway.

Same-question configuration comparison.

- *Gemini Base*: Favor direct CO–CO coupling, allowing parallel pathways under different conditions.
- *Gemini More*: Favor asymmetric CO–CHO coupling through neighboring-site synergy, while acknowledging uncertainty.
- *Claude Base*: Favor CO dimerization on Cu(100) at low overpotential/high CO coverage, with remaining uncertainty.
- *Claude More*: Multiple condition-dependent pathways: CO–CO at low overpotential/high coverage, with greater CO–CHO contribution under other conditions.

Interpretation. Grades 2/2/2: Astra retains its global preference but adds a conditional exception and changes its proposed experiment. This is not a global mechanism reversal. Gemini's and Claude's different More statements are separately retained, without transferring their labels to Astra.

D.5 B01: BARRIERS TO ADULT MAMMALIAN REGENERATION

Question. Is the key restriction on complete tissue regeneration an intracellular state lock or the tissue microenvironment?

Evidence packet.

E1. Repair requires coordinated proliferation, differentiation, and spatial organization.
E2. Immune, matrix, metabolic, and gene-regulatory states can all affect regeneration.
E3. Restoring proliferation in a cell need not restore organ structure and function.
E4. Adult regeneration differs greatly across species and organs.

More statement. Recent reviews emphasize temporal/spatial signal coordination and stopping mechanisms as obstacles to regeneration at the correct scale.

The two Astra prompts are identical.

Astra Base (confidence 68).

Primary
Favor the tissue microenvironment and cross-cell spatial coordination as the main bottleneck, interacting with intracellular locking; multiple mechanisms remain unresolved.
Alternative
Adult epigenetic and differentiated states are the primary restriction on re-entering a complete regenerative program.
Causal chain
Immune, matrix, metabolic, and spatial signals jointly control proliferation, differentiation, migration, and arrangement after damage. An adult environment may inhibit their coordination; intracellular locking additionally restricts responsiveness.
Discriminating experiment
In a matched injury model, factorially combine epigenetic resetting and reconstruction of embryonic-like immune, matrix, and metabolic conditions. Compare full structural and functional recovery: environment dominance predicts a larger environmental-only effect; state-lock dominance predicts a larger reset-only effect.

Astra R2 (confidence 60).

Primary
Prefer mutually reinforcing intracellular locking and microenvironmental restriction, with insufficient evidence for universal priority and potentially organ-specific bottlenecks.
Alternative
Intracellular differentiation/epigenetic locking is primary; once released, the adult environment can support complete regeneration.
Causal chain
Stable cell states inhibit regenerative programs, while injury-induced immune, matrix, and metabolic changes reinforce locking and disturb spatial cues. Proliferation alone therefore need not restore organized differentiation and structure.
Discriminating experiment
Cross cell-state unlocking and a pro-regenerative environment in four groups, verify intervention efficacy, and measure long-term structure and function. Coupling predicts insufficient single interventions and strong joint recovery; the competing account predicts unlocking alone approaches the joint effect.

Same-question configuration comparison.

- *Gemini Base*: Favor environmental inhibition and missing spatial guidance, while retaining uncertainty.
- *Gemini More*: Favor the microenvironment, where rapid fibrosis and absent developmental gradients impede organization.
- *Claude Base*: Favor the microenvironment; fibrosis and missing positional information dominate, with intracellular locking a secondary reversible restriction.
- *Claude More*: Favor microenvironmental spatial/temporal signals and stopping mechanisms; intracellular locking is secondary and plastic.

Interpretation. Grades 4/4/4 identify a shift from environmental priority to a coupled account without universal priority. Interacting mechanisms need not be mutually exclusive; the change concerns their stated dominant role, not proof of two incompatible biological worlds.

D.6 B14: SELECTIVE SYNAPTIC VULNERABILITY IN DEMENTIA

Question. Within the same diseased brain, are synaptic loss and preservation chiefly determined by intrinsic synapse type or the network/glial environment?

Evidence packet.

E1. Synaptic loss is associated with cognitive decline in several dementias.
E2. Not all synapses degenerate equally at the same stage.
E3. Synapse type, activity load, connectivity, and surroundings differ simultaneously.
E4. Preserved synapses may have less damage exposure or active protective mechanisms.

More statement. Recent reviews distinguish vulnerability from resilience rather than treating them as simple opposites, motivating separate searches for harmful and protective factors.

The Astra variant uses evidence order E3, E2, E1, E4, with the original question and instruction.

Astra Base (confidence 60).

Primary
Favor network activity load and the local glial environment, with intrinsic type setting susceptibility thresholds; mechanisms are not fully separable.
Alternative
Intrinsic molecular type and active protective programs dominate degeneration; network and glial differences mainly change overall damage.
Causal chain
Disease changes activity, connectivity, inflammation, and trophic support. High load or inflammation destabilizes synapses; receptors, adhesion, and repair modify thresholds. Similar synapses can therefore differ across environments, and preservation may combine low exposure with protection.
Discriminating experiment
Label same-subtype synapses, track activity longitudinally, and locally manipulate inflammation or load. Environmental dominance predicts that altered surroundings can reverse vulnerability rankings; intrinsic dominance predicts stable rankings and smaller effects.

Astra S2r2 (confidence 58).

Primary
Again favor network load and glial/local environmental differences, with intrinsic type setting thresholds; the two interact and remain incompletely distinguished.
Alternative
Intrinsic receptor, proteostasis, and structural differences determine degeneration; network/glial changes mainly follow it or amplify it.
Causal chain
Abnormal activity and connectivity alter metabolism, inflammation, and glial signaling. Highly loaded or damaged networks destabilize first, with intrinsic vulnerability/protection modifying thresholds and selective loss contributing to cognitive decline.
Discriminating experiment
Place matching synaptic types under different activity and glial conditions, image longitudinally, and manipulate local inflammation/complement. Environmental dominance predicts reordered vulnerability with environmental changes; intrinsic dominance predicts continued preferential loss of susceptible types.

Same-question configuration comparison.

- *Gemini Base*: Intrinsic susceptibility and the microenvironment cooperate; receptors/metabolism set thresholds and glial engulfment/network load determine survival.
- *Gemini More*: Intrinsic receptors establish baseline vulnerability; glia and activity may supply active resilience.
- *Claude Base*: Favor local network/glial conditions, with intrinsic type setting the threshold and multiple mechanisms unresolved.
- *Claude More*: Intrinsic type sets vulnerability, and activity plus glial/complement conditions determine threshold crossing; no single dominant cause is selected.

Interpretation. Grades 2/2/2: the primary preference is similar, but the competing causal account and detail change. S2r2 is an existing reordered-input repetition, not an added-evidence response. The comparison must not inherit an unresolved label from an earlier, different low-effort dataset.

D.7 E12: INCONSISTENT REPORTS OF MARTIAN METHANE

Question. Do differences between near-surface methane reports and orbital non-detections most plausibly arise from real local processes or measurement problems?

Evidence packet.

E1. Measurements differ in sampling time, altitude, and instrument-background handling.
E2. Curiosity has reported near-surface methane variation around Gale crater.
E3. Orbital observations tightly constrain methane over broader atmospheric regions.
E4. A real local-release explanation must also account for mixing, transport, and removal.

More statement. Recent reliability analyses question instrument/background treatment in Curiosity detections, but have not established that every methane report is an artifact.

The two Astra prompts are identical.

Astra Base (confidence 40).

Primary
Mechanisms remain unresolved; provisionally choose local release and nocturnal accumulation, without sufficient evidence that they are more likely than measurement bias.
Alternative
Instrument background or processing residuals drive much of the reported variation; mismatched sampling further complicates orbital comparisons.
Causal chain
Local release could accumulate in a shallow nocturnal boundary layer and dilute by day. Dilution is not removal: total release, transport, and lifetime must jointly satisfy orbital limits. An unverified fast-removal requirement weakens the account.
Discriminating experiment
Use independent instruments with different principles and comparable sensitivity at matching near-surface and elevated locations, covering day/night with blanks. Real processes predict cross-instrument replication and physically coherent variation; bias predicts non-replication and dependence on background or processing.

Astra R1 (confidence 68).

Primary
Favor mismatched sampling, near-surface gradients, and background treatment as the main sources of disagreement, without excluding brief local release.
Alternative
Short, intermittent, highly localized release in Gale is rapidly diluted, transported, or removed.
Causal chain
A surface instrument samples a local peak; boundary-layer mixing, wind, and removal keep column abundance below orbital sensitivity, while background corrections amplify disagreement.
Discriminating experiment
Coordinate repeated orbital overpasses with rover, vertical-profile, and meteorological measurements. Real release predicts coherent surface peaks and a predictable column signal; measurement problems predict no stable correspondence under matched sampling.

Same-question configuration comparison.

- *Gemini Base*: Favor local release/nocturnal accumulation followed by dilution and an unknown rapid-destruction mechanism.
- *Gemini More*: Lean toward a background-processing artifact, while retaining local release plus unknown removal as an alternative.
- *Claude Base*: Real weak release and sampling/instrument differences jointly contribute; dominance cannot be determined.
- *Claude More*: Lean toward low-ppb TLS-SAM background/processing biases, without excluding a small genuine signal.

Interpretation. Grades 4/4/4 indicate a shift in the stated preferred account under identical input. R1's causal chain still invokes physical gradients, so the preference label does not exhaust its content. Neither answer proves methane exists or that all detections are artifacts.

D.8 E13: VENUS'S UNIDENTIFIED ULTRAVIOLET ABSORBER

Question. Which chemical species mainly account for the unidentified ultraviolet absorption in Venus's clouds?

Evidence packet.

E1. Iron-containing and sulfur-containing species are both studied candidates.
E2. Common background constituents do not fully explain the cloud UV absorption.
E3. A candidate must maintain sufficient abundance in acidic and irradiated clouds.
E4. A similar laboratory absorption band does not establish sufficient atmospheric abundance.

More statement. Recent experiments study iron chloride absorption/stability under Venus-like conditions; three-dimensional chemistry simulations constrain OSSO abundance and spectral contribution.

Astra Base (confidence 40).

Primary
Provisionally lean toward sulfur photochemical products, while acknowledging insufficient evidence to distinguish sulfur dominance, iron dominance, or joint contributions.
Alternative
Trivalent-iron complexes in acidic droplets supply most additional absorption, with sulfur contributing less.
Causal chain
Available sulfur feeds photochemistry, possibly producing UV absorbers. Sufficient production and transport must offset destruction for them to accumulate. Sulfur availability supports supply plausibility, not identity.
Discriminating experiment
Measure excess UV absorption and sulfur/iron speciation and abundance together in one cloud layer, then use cross-sections at actual acidity to calculate contributions. Sulfur dominance predicts most absorption from measured sulfur products; iron dominance predicts the reverse.

Astra More (confidence 45).

Primary
Weakly favor iron-containing species such as trivalent-iron complexes in droplets; sufficient abundance is unproven, and mixtures remain possible.
Alternative
Sulfur species or mixtures dominate; constraints on OSSO cannot exclude all sulfur candidates.
Causal chain
If sufficiently abundant, iron-complex charge-transfer absorption could fill the UV deficit. OSSO modeling weakens that particular candidate; conducting iron-chloride experiments alone does not establish stability or abundance, so the preference is weak.
Discriminating experiment
Across UV-bright/dark cloud regions, measure iron/sulfur speciation, abundance, and spectra in situ; correct scattering and reconstruct contributions. Iron dominance predicts iron explains the main deficit and variation; sulfur dominance predicts sulfur does so.

Same-question configuration comparison.

- *Gemini Base*: Lean toward sulfur photoproducts such as OSSO or polymers, with dominance unresolved without in-situ concentrations.
- *Gemini More*: Iron and sulfur may jointly contribute; no single dominant absorber is established.
- *Claude Base*: Favor sulfur oxides and dissolved polysulfur, with remaining uncertainty.
- *Claude More*: Lean toward iron species in sulfuric-acid droplets, possibly supplemented by sulfur.

Interpretation. Grades 4/4/4 capture a weak but explicit family-level preference change. Evidence constraining one sulfur candidate does not rule out its entire family; a laboratory study of iron chloride does not certify atmospheric abundance. The answer itself retains both caveats.

D.9 A16: SUPER-CLAUSIUS–CLAPEYRON EXTREME-RAINFALL SCALING

Question. Is super-Clausius–Clapeyron growth mainly due to stronger moisture convergence or changing precipitation-type composition?

Evidence packet.

E1. Conditional statistical relationships cannot automatically be interpreted as causal responses to future climate change.
E2. Warming increases atmospheric moisture capacity, but dynamics also affect rainfall intensity.
E3. Observed intense rainfall sometimes grows faster with temperature/dew point than the thermodynamic scale.
E4. Stratiform and convective rain fractions change with environmental state.

More statement. Recent studies propose composition shifts and mesoscale moisture-convergence amplification; their statistical definitions and scenarios must first be aligned.

The two Astra prompts are identical.

Astra Base (confidence 50).

Primary
For observed conditional slopes, provisionally favor composition changes; evidence does not settle the mechanisms' relative importance.
Alternative
Within a precipitation type, stronger ascent and moisture convergence yield growth beyond the thermodynamic scale.
Causal chain
A larger convective fraction with a heavier intensity tail can steepen a mixed-sample extreme quantile even when neither type grows super-thermodynamically internally. Dynamics can also generate within-type excess growth. Both can coexist, and the conditional slope is not automatically a future-warming response.
Discriminating experiment
Combine radar classification and moisture budgets, fix type fractions across temperature/dew-point bins, and reconstruct distributions. Composition dominance predicts a weaker excess slope; within-type convergence dominance predicts persistence with stronger ascent/convergence.

Astra R3 (confidence 58).

Primary
Provisionally favor dynamics: stronger convergence or precipitation efficiency under warming produces excess growth, assisted by composition changes, whose relative contributions remain unresolved.
Alternative
Increasing convective relative to stratiform rainfall and associated microphysics creates the conditional slope without requiring stronger convergence.
Causal chain
Warming raises available moisture; stronger convergence, ascent, or efficiency can amplify condensation beyond the thermodynamic scale. Type fractions can also enlarge or generate the slope, with selection effects in conditional statistics.
Discriminating experiment
Use paired convection-resolving simulations that separately hold type fractions or low-level moisture convergence fixed. Persistence under fixed fractions but disappearance under fixed convergence supports convergence dominance; the opposite supports composition dominance.

Same-question configuration comparison.

- *Gemini Base*: Favor a shift from weaker stratiform to stronger convective rain as the source of the observed conditional slope.
- *Gemini More*: Lean toward mesoscale moisture-convergence amplification, intertwined with composition changes.
- *Claude Base*: Favor dynamical feedback for hourly growth while distinguishing composition effects on conditional statistics from climate-forced responses.
- *Claude More*: Emphasize composition for observed conditional slopes, with a smaller convergence contribution to true climate responses; uncertainty remains.

Interpretation. Grades 4/4/4 mix a preference change with a shift of explanatory target and scope. Base explicitly targets observed conditional slopes; R3 discusses warming dynamics. This is not a clean reversal about a single fixed statistical object, and the examples should be read with that qualification.

D.10 A07: TRIGGERS OF TROPICAL-CYCLONE RAPID INTENSIFICATION

Question. When ocean energy supply and environmental shear are favorable, what internal process most plausibly decides whether sudden rapid intensification occurs?

Evidence packet.

E1. Ocean thermal conditions and wind shear affect intensification likelihood.
E2. Storms in similar large-scale environments need not intensify rapidly at the same time.
E3. Inner-core convection, vortex structure, and cold wakes jointly affect energy and angular-momentum budgets.
E4. A strong convective cell preceding intensification does not prove it is the sole trigger.

More statement. Recent work links small convective disturbances to rapid intensification, while other studies emphasize weakened cold-wake feedback; these act at different stages.

The two Astra prompts are identical.

Astra Base (confidence 65).

Primary
Favor cooperative organization of convection and the vortex: persistent near-center heating, vertical alignment, and angular-momentum convergence create positive feedback that may cross an intensification threshold.
Alternative
Localized convective bursts directly trigger intensification through vorticity production/aggregation; global organization and alignment mainly follow.
Causal chain
Moistening reduces evaporative cooling and alignment improves heating–circulation coupling. Organized heating inside the maximum-wind radius strengthens inflow, inward angular-momentum transport, and contraction; winds and surface enthalpy flux reinforce each other. A cold wake can suppress the feedback.
Discriminating experiment
Use matched numerical ensembles while preserving total heating, radial distribution, and peak cell strength but disrupting annular organization/persistence. The primary account predicts less or later rapid intensification; a burst-driven account predicts smaller effects and continued linkage to strong cells.

Astra R1 (confidence 65).

Primary
Favor persistent inward concentration and axisymmetrization of convection, reinforced by boundary-layer angular-momentum input; pathways and thresholds remain uncertain.
Alternative
Small rotating vortices generated by intense local convection merge and transport vorticity inward, triggering intensification; organization follows.
Causal chain
Vertical alignment and moistening concentrate deep convection near/inside the maximum-wind radius. Heating drives secondary circulation and inward absolute-angular-momentum transport, strengthening the warm core and winds and further organizing convection. Cold wakes and residual shear may interrupt it.
Discriminating experiment
In matched high-resolution ensembles, preserve accumulated heating and radial distribution while separately weakening extreme cells or disrupting persistent annular organization. The primary account predicts stronger suppression by disrupting organization; the alternative predicts stronger suppression by weakening extreme cells.

Same-question configuration comparison.

- *Gemini Base*: Favor vertical alignment and axisymmetrization, with deep convection concentrated inside the maximum-wind radius.
- *Gemini More*: Small inner-core convective bursts drive rapid axisymmetrization and are the critical internal mechanism.
- *Claude Base*: Persistent alignment and axisymmetrization of deep convection inside the maximum-wind radius efficiently strengthen the vortex.
- *Claude More*: Favor organized, symmetric inner-core convection in a high-inertial-stability region, while retaining uncertainty.

Interpretation. Grades 1/1/1: the same leading mechanism is retained with implementation-level differences. Both Astra confidence scores are 65. Different emphases in other configurations do not by themselves establish contradiction; full response fields remain available in the supplement.

E EXTENDED ANALYSES AND COMPLETE ENDPOINTS

Annotation separation and eligible evidence. Eighty Astra xhigh calls annotate question/evidence packets without seeing model outputs. They select two candidate explanations, mark binary-comparison suitability, assign question type, and classify each evidence item as A-supporting, B-supporting, or neutral with strength 0–3. Evidence order is randomized for this task. Sixteen further calls map frozen anonymous state codebooks to A/B/tie/other/unknown and preference indicators, without condition labels, confidence, model identity, or state frequencies. All original annotations and mapping prompts are retained. The independence is between contexts and visible information, not between model families or experts.

Of 80 questions, 59 have an eligible candidate pair. Twelve questions have any directional item, but P03 is ineligible; thus only 11 eligible questions support the directional analyses. The eligible directional IDs are C03, B04, B06, B08, B09, E02, E08, E14, E16, A05, and A13. Across all 320 items, 306 are neutral relative to the annotated pair. Only C03 has support on both sides, with conflict $C = 2/3$; the other ten eligible directional questions have $C = 0$. All-neutral or ineligible cases have undefined conflict. A generic conflict–instability association is not identified by this coverage.

Table 5: Position-index slopes in percentage points per unit z. AB excludes ties/other; all includes them. The corrected p-values refer to the nine-test exploratory family.

Config.	n_{AB}	β_{AB}	95% CI	p_H	n_{all}	β_{all}	95% CI
Astra	8	2.22	[-0.64, 5.08]	1.000	11	0.08	[-3.05, 3.81]
Gemini	9	-4.01	[-10.07, -0.56]	0.336	11	0.73	[-2.39, 4.10]
Claude	9	-3.43	[-9.65, 0.00]	1.000	11	-2.50	[-7.37, 0.00]

Position estimand and multiplicity. Within each question, repetitions are averaged within permutation, then a question-fixed-effect slope relates the choice of A to the signed position index in Equation 2. The denominator normalizes total absolute evidence strength. Valid questions require usable outcomes and variation in the index; this reduces the conditional analysis to 8/9/9 questions. The unconditional version retains ties/other states as not choosing A. First-versus-last profiles are descriptive associations when several evidence items move together. The nine-test Holm family comprises three conditional position slopes, three conflict correlations, and three Shuffle differences between initial-preference groups. Non-estimable conflict tests remain non-rejections for multiplicity accounting. Intervals use 5,000 question-bootstrap draws and are pointwise, not multiplicity-adjusted.

Table 6: Primary-state contrasts under reference rotation (percentage points). Complete four-original-state sets contain 80/80/70 questions; each More contrast additionally requires a readable More state. Full per-reference intervals and question sets are in extended statistics.

Config.	Contrast	Base	R1	R2	R3
Astra	Shuffle	-2.85	2.71	1.60	0.21
Astra	More	13.75	9.17	14.58	7.08
Gemini	Shuffle	-3.96	-0.56	-2.08	-1.32
Gemini	More	6.67	15.83	12.08	8.75
Claude	Shuffle	2.31	2.39	4.33	1.04
Claude	More	4.29	-0.95	11.43	3.33

Reference and confidence endpoints. Reference rotation does not re-run the whole-answer grader. It changes the original-prompt state used as the reference for state disagreements. Confidence quadrants compare state equality with $|q_{variant} - q_{Base}| < \tau$ for $\tau = 5, 10, 15$. Question weights are equal; Repeat and Shuffle each receive half the weight, and Shuffle permutations receive equal weight. Missing confidence/state pairs are excluded locally. Exactly unchanged scores use a separate raw-pair denominator. A confidence number may describe a different proposition after the preferred explanation changes, so these diagnostics are not calibration estimates.

Table 7: Weighted state/confidence quadrants (percent). Small means strictly below the threshold. Values may not sum to 100 exactly after rounding.

Config.	τ	n	Same/small	Same/large	Changed/small	Changed/large
Astra	5	80	36.0	36.7	12.3	15.0
Astra	10	80	61.4	11.2	19.9	7.4
Astra	15	80	68.2	4.5	23.1	4.2
Gemini	5	80	25.3	40.8	6.5	27.3
Gemini	10	80	45.9	20.2	15.5	18.4
Gemini	15	80	57.0	9.1	22.0	11.8
Claude	5	76	42.1	25.0	15.5	17.5
Claude	10	76	58.0	9.0	26.0	7.0
Claude	15	76	66.6	0.4	32.5	0.5

Complexity specification. For a valid Astra whole-answer comparison r on question i, the adjusted weighted linear probability model is

$$\begin{aligned}\mathbb{1}[g_{ir} \geq k] = \alpha_i + \beta_S \mathbb{1}[S] + \beta_L \mathbb{1}[L] + \beta_M \mathbb{1}[M] \\ + \gamma_1 |\log(1 + L_{ir}) - \log(1 + L_{iB})| + \gamma_2 |H_{ir} - H_{iB}| + \gamma_3 |F_{ir} - F_{iB}| + \varepsilon_{ir},\end{aligned} \tag{4}$$

with Repeat as reference, $k = 2$ or 3, four-text-field character count L, fixed-vocabulary hedge count H, and lexical candidate-family count $F \in \{0, 1, 2\}$. All differences are absolute. Candidate aliases come from the evidence annotation; they do not enumerate every scientific mechanism. Each question has normalized total weight, retaining repetition/permutation weights. Question-bootstrap intervals use 5,000 draws. The ten original Chinese hedge strings and exact counting code are in the supplement. Output features can be downstream of the intervention, so adjustment cannot establish a direct causal effect.

Table 8: Fixed-effect sensitivity coefficients (percentage points). The paired mean contrast and this regression are different weighted estimands.

Threshold	Condition	Unadjusted	95% CI	Adjusted	95% CI
2	Shuffle	-1.83	[-6.30, 2.85]	-0.58	[-5.48, 4.31]
2	Less	0.18	[-8.64, 8.96]	0.38	[-8.46, 9.16]
2	More	19.38	[11.80, 27.18]	19.67	[12.37, 27.30]
3	Shuffle	-2.07	[-5.76, 1.52]	-1.11	[-5.08, 2.69]
3	Less	3.11	[-3.32, 9.58]	3.15	[-3.10, 9.30]
3	More	9.45	[0.22, 19.01]	9.78	[0.47, 19.55]

Question types and initial non-preference. Question-type counts are dominance 64, source/identity 8, causal direction 5, pathway details 2, and coexistence/coupling 1. Within-type summaries are descriptive. The initial non-preference state groups contain 8 Astra, 8 Gemini, and 4 Claude questions, versus 4 Astra questions under the separate whole-answer decision annotation. The classifications are not interchangeable; sparse groups and grade-3's definition prevent causal conclusions about indecision and instability. The supplement retains raw transition counts, effective denominators, complete-six-permutation sensitivity, exclusion of Claude P01, and the separate historical low/xhigh comparison. These are audit material, not additional independent replications.