FROM SCORES TO CLAIMS: A VALIDITY ARGUMENT FRAMEWORK FOR LARGE LANGUAGE MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Benchmark scores are widely used as evidence of large language model capabilities, yet their meaning depends on the tasks, protocols, and scoring procedures that produce them. We introduce Score-to-Claim Auditing (S2C), a validity-argument framework grounded in language testing and educational measurement. Given an intended interpretation, S2C examines annotated item demands, changes in scores and rankings across protocols, and automated ratings through expert comparisons and controlled tests. Benchmark Validity Cards synthesize the evidence into supported interpretations and conditions of use. We apply S2C to multiple-choice tasks, mathematical word problems with numerical answers, and automated summary evaluation. In multiple-choice tasks, annotations reveal different demand mixtures beneath a shared response format. Small score changes across protocols can alter model rankings. Among 18 equal-accuracy comparisons, 12 contain offsetting item-level gains and losses. Construct-linked analysis also identifies opposing net changes across demand groups. In summarization, positive correlations with expert ratings coexist with numerical discrepancies. On the same summaries, the judges' point-estimate ordering by consistency correlation with experts reverses between summary-level and system-level aggregation. Controlled source reversals produce substantially larger consistency-rating changes than meaning-preserving expression changes or identical-input repeats, supporting a factual-consistency interpretation under the tested rubric and materials. These findings motivate validity profiles alongside benchmark scores, making the evidence and conditions behind capability and quality claims explicit. Analysis code and numerical data are available at `https://anonymous.4open.science/r/s2c-review-E603/`.

1 INTRODUCTION

Benchmark scores compare large language models (LLMs), but the claims attached to them often extend beyond the evaluated tasks. A science-question score may be read as evidence of scientific reasoning, and an automated rating as evidence of summary quality (Rein et al., 2024; Fabbri et al., 2021). The question is: *What evidence warrants a score interpretation, and under which conditions does it hold?*

We thus introduce **Score-to-Claim Auditing** (S2C), a validity-argument framework grounded in language testing and educational measurement (Bachman & Palmer, 1996; Samuel, 1995; Kane, 2013). Starting from an intended interpretation, S2C links item-demand annotations, protocol comparisons, and scoring analyses to the evidence needed to support it. **Benchmark Validity Cards** record the resulting interpretation, supporting evidence, scope and conditions, and evidence gaps. S2C complements benchmark profiling (Kim et al., 2025), prompt-sensitivity analysis (Sclar et al., 2024; Hua et al., 2025), and validity-centered evaluation and reporting (Salaudeen et al., 2025; Ghosh et al., 2026): its contribution is to connect these forms of evidence to the distinct interpretations they support. Figure 1 presents the audit workflow.

Our main audit evaluates six recorded model pipelines on 480 construct-annotated questions from MMLU, ARC-Challenge, HellaSwag and GPQA under five protocols (Hendrycks et al., 2021; Clark

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.

et al., 2018; Zellers et al., 2019; Rein et al., 2024). We vary the instructions, option order and wording across protocols, yielding 14,400 model–item–protocol records. The benchmarks share an answer format but differ in their annotated demands. Across six models and four benchmarks, comparing each of the four alternative protocols with the original presentation yields $6 \times 4 \times 4 = 96$ paired comparisons. Each comparison matches correctness outcomes for the same items across protocols, retaining 110–120 items with usable responses under both. Of these comparisons, 18 preserve accuracy, but twelve change which items are answered correctly. In one GPQA comparison, gains on mathematical-reasoning items offset losses on subject-knowledge items. Yet GPT-5.2 leads on GPQA under all five protocols. The evidence supports a sustained comparative advantage while qualifying a different interpretation: unchanged performance across annotated demands.

Extensions to GSM8K numeric answers and SummEval summary evaluation examine different scoring procedures (Cobbe et al., 2021; Fabbri et al., 2021). For summary ratings, positive correlations with the three-expert mean coexist with numerical discrepancies. On the same 1,200 summaries under the direct prompt, GPT-5.2 has the higher consistency correlation at summary level, but Claude Sonnet 4.6 does at system level. This point-estimate reversal makes the intended evaluation unit part of the comparison. Identical-input repeats assess rating recurrence separately from expert agreement.

A controlled study provides positive evidence about score meaning. Across 45 human-checked news-article groups, we reverse source support while keeping summaries unchanged. Within-source rating orderings reverse in both expression forms for all 45 GPT groups and 44 Sonnet groups. The remaining Sonnet group ties under the edited article. Factual-support contrasts substantially exceed the observed expression and identical-input repeat differences, supporting a factual-consistency interpretation under the tested rubric and materials.

Our contributions are a claim-specific validity-argument workflow, a construct-linked benchmark audit with a numeric-answer extension, and complementary observational and controlled evidence for interpreting automated summary ratings. We argue that validity profiles should accompany benchmark scores, making explicit what a comparison supports and what further evidence a broader interpretation requires.

2 Related Work

Validity-centered evaluation. Language testing and educational measurement ground validity in evidence for score interpretations and uses (Bachman & Palmer, 1996; Samuel, 1995; Kane, 2013). S2C adopts this argument-based perspective, complementing the statistical models of classical test theory and item response theory. *Measurement to Meaning* (Salaudeen et al., 2025) distinguishes claims about observable criteria from those about latent constructs and prioritizes evidence by the measurement–target relationship. S2C shares this claim-centered premise and links item-demand annotations, paired protocol comparisons, and scoring analyses to specific interpretations.

Benchmark diagnosis and protocol effects. Benchmark Profiling (Kim et al., 2025) combines gradient-based importance and targeted parameter ablation to estimate contributions of cognitively grounded abilities to benchmark performance. S2C provides an item-centered view through judged demands and recorded outputs. Sclar et al. (2024) demonstrate sensitivity to prompt formatting, and Hua et al. (2025) identify heuristic scoring, including rigid answer matching, as a source of apparent sensitivity. Our retrospective audit characterizes recorded pipeline variation through aggregate scores, paired correctness, model ordering and demand-group changes.

Reporting evaluation evidence. Model Cards (Mitchell et al., 2019) document intended uses, evaluation conditions and disaggregated performance. Evaluation Cards (Ghosh et al., 2026) unify benchmark, run, and model records with signals for reproducibility, completeness, provenance, risk, and comparability. S2C's Benchmark Validity Cards connect each interpretation to supporting analyses, scope and conditions, and evidence gaps. Their authored rationale explains why the findings support or qualify the interpretation, alongside the compiled numerical evidence.

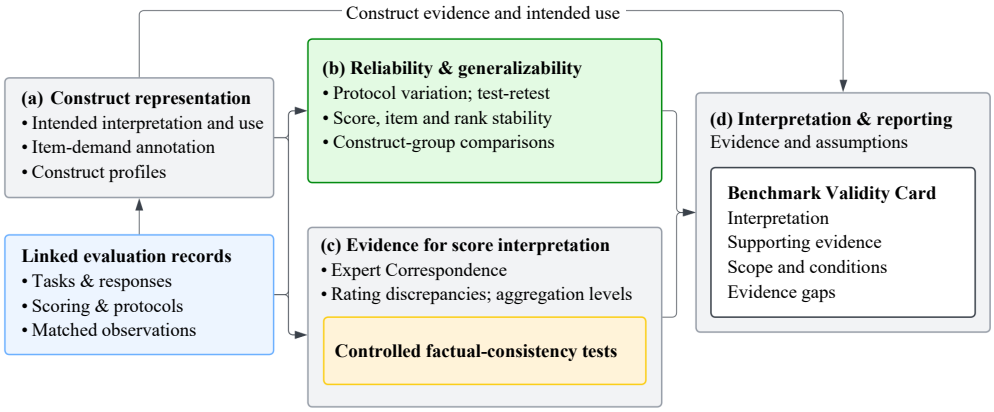


Figure 1: Overview of Score-to-Claim Auditing (S2C).

3 Methodology: Score-to-Claim Auditing

An S2C audit specifies an intended interpretation, compiles the relevant evidence, and records the resulting argument in a Benchmark Validity Card (Figure 1).

3.1 Construct representation

An audit specifies the capability or output quality being interpreted, the tasks and conditions over which the interpretation should hold, and the intended comparison. Following argument-based validation (Kane, 2013), S2C traces the inference from responses through extraction or rating, item scores, aggregation, and, where relevant, ranking to a capability or quality claim. Task demands inform the ability interpretation (Bachman & Palmer, 1996). Validity concerns how the combined evidence supports it (Samuel, 1995). Reliability concerns consistency across specified conditions. Establishing score meaning also requires evidence about the relevant task demands or scoring relations. The intended interpretation determines the evidence profile. Reference-quality, external-criterion, and transfer studies can extend it as the intended use requires.

An explicit codebook distinguishes demands such as subject knowledge, scientific reasoning, mathematical reasoning, and event continuation. On science questions, this separates recalling a domain fact from applying a principle or carrying out a quantitative derivation. Two annotators label items independently, and agreement is reported before adjudication produces the final labels. These describe judged item demands. Claims about internal model processes require separate evidence. Linking labels to responses shows how successful performance and protocol-related changes are distributed across demand groups. Primary labels define the main groups. Secondary and mismatch judgments remain exploratory. Assisted labeling is evaluated separately against held-out adjudicated labels.

3.2 Reliability & generalizability

Interpretations extending across presentations require comparisons under transformations designed to preserve the scoring target. We vary instruction style, option order, and wording. Score ranges describe aggregate variation. Paired outcomes distinguish net accuracy change from correctness turnover. Gains and losses are incorrect-to-correct and correct-to-incorrect transitions, whose rates give net change by subtraction and turnover by addition. Ranks describe model ordering within a specified panel and tie convention. Construct-linked comparisons locate changes across demands, while identical-input repeats examine recurrence within a scoring protocol. Each comparison addresses a distinct aspect of stability.

3.3 Evidence for score interpretation

For automated quality ratings, we specify the dimension, rubric, reference, and aggregation unit. Correlations with expert judgments assess correspondence in ordering. Signed and absolute differences describe numerical agreement. Comparing summary-level and system-mean associations on fixed membership separates evaluation level from sample composition.

Controlled contrasts test the rubric's intended relation more directly. Factual-consistency ratings should follow source support while remaining comparatively stable under meaning-preserving expression changes. We cross factual content with expression and reverse support for unchanged summaries. Human checks establish these material relations. Numerical expert targets would require separate ratings. Identical-input repeats provide a variability reference.

Linked records and uncertainty. Within each study, linked task, response, model, protocol, and scoring records preserve each comparison's eligible observations and denominator. Execution and parsing failures concern score production and remain distinct from construct coverage or scoring validity. Resampling retains paired observations within item or source-document blocks.

3.4 Interpretation & reporting

The analyst assembles relevant construct, protocol, and scoring results, preserving samples, denominators, and uncertainty statements, then explains what each finding supports or qualifies and which assumptions remain unexamined. The Card records four fields: *interpretation*, *supporting evidence*, *scope and conditions*, and *evidence gaps*. Scripts compile numerical evidence. The interpretation and rationale remain authored judgments whose supporting results are inspectable. Each finding retains its role in the interpretation, and the synthesis preserves the evidence as a profile. The earlier implementation's categorical rules are documented separately in Appendix G.4 and are not applied to the current Cards.

Suitable item-demand annotations can be reused. Changed protocols or new quality dimensions require corresponding evidence. The experiments provide worked interpretations, including Table 2. Appendix G.1 details the reporting workflow.

4 Experiments

4.1 Experiment Settings

We examine S2C across three task families. Multiple-choice evaluation asks what paired outcomes and item-demand annotations reveal beyond aggregate scores and rankings. Numeric-answer evaluation examines the same distinctions under exact numerical scoring. Summary evaluation examines expert correspondence, numerical agreement, aggregation, and recurrence, then tests whether consistency ratings follow controlled changes in source support.

The multiple-choice and numeric-answer studies characterize recorded model pipelines. The main audit includes heterogeneous historical repair and parsing procedures, so its comparisons concern those pipelines and do not isolate first-attempt prompt effects. Fixed-grid and paired multiple-choice analyses use separate sets of 10,000 item-block bootstrap draws per benchmark, retaining model and protocol observations together and preserving MMLU's four strata. Appendix C distinguishes pointwise paired intervals from approximate within-benchmark simultaneous contrast bands. The summary studies retain their own analysis populations and resampling units.

4.2 Multiple-Choice Task Evaluation

4.2.1 Evaluation Protocol

The main audit contains 120 items each from ARC-Challenge, HellaSwag, MMLU, and GPQA (Clark et al., 2018; Zellers et al., 2019; Hendrycks et al., 2021; Rein et al., 2024). ARC-Challenge and HellaSwag use validation samples. MMLU contributes 30 items from each of four selected subject groups, and GPQA draws from Diamond.

Six recorded pipelines, i.e., GPT-5.2, Claude Sonnet 4.6, Gemini 2.5 Pro, DeepSeek V4-Pro, Llama 4 Maverick, and Mistral Large 2512, are evaluated under original presentation, formal-exam framing, plain instruction, shuffled options, and semantic paraphrase. Shuffling remaps the displayed reference answer. Paraphrases are designed to preserve meaning, answer options, and the reference answer. This yields 14,400 model–item–protocol records. Appendix B documents sampling, configurations, protocol construction, and available preservation checks.

Of these records, 14,311 yield usable parsed answers. Panel accuracy counts usable correct responses out of all 120 scheduled items per cell. The 89 extraction or execution failures remain in that denominator and are logged separately from observed wrong answers. Paired analyses use items with usable responses under both the original and compared protocol, giving 96 comparisons of 110–120 items each. Ranks are computed within the six-model panel, with tied correct counts sharing a position.

Two annotators independently assign one primary demand from twelve categories before adjudication. The codebook also records optional secondary demands, load ratings, artifact judgments, and a rationale. Main analyses use human-adjudicated primary labels. Secondary labels and construct-mismatch judgments are exploratory. Appendix A gives the codebook and field-specific agreement.

4.2.2 Annotation and Demand Profiles

Pre-adjudication primary-label agreement is $\kappa = 0.748$ for the original 360 ARC-Challenge/HellaSwag/MMLU items and $\kappa = 0.568$ for the later 120 GPQA items, with exact agreement of 81.1% and 71.7%, respectively.

The shared answer format covers distinct demand mixtures (Table 1). HellaSwag is entirely labeled event continuation, while MMLU is dominated by subject-specific knowledge. ARC-Challenge emphasizes scientific and causal reasoning. GPQA combines subject knowledge with mathematical, scientific, and formal-logical reasoning. These profiles specify the annotated demands represented by each benchmark total.

Table 1: **Adjudicated primary-demand counts.** Seven of twelve categories occur as primary labels. The other five have zero counts. Each label denotes the dominant demand.

Primary item demand	ARC-C	HellaSwag	MMLU	GPQA
Subject-specific knowledge	2	0	83	56
Scientific reasoning	82	0	0	15
Mathematical reasoning	0	0	8	48
Formal-logical reasoning	0	0	0	1
Causal reasoning	36	0	0	0
Event continuation	0	120	0	0
Cultural/social knowledge	0	0	29	0
Total	120	120	120	120

A separate GPT-5.2 pre-labeling pilot uses 90 calibration examples and 270 held-out items, with 90 test items from each of the three original benchmarks. Against adjudicated labels, accuracy is 79.3%, macro-F1 is 0.561, and $\kappa = 0.731$. Benchmark accuracies are 100% on HellaSwag, 87.8% on MMLU, and 50.0% on ARC-Challenge, motivating benchmark-specific review of pre-labels. Appendix A details this separate evaluation.

4.2.3 Score, Rank, and Paired Outcomes

Score variation and rank movement. Score variation and rank movement differ within the panel (Figure C1, Appendix C). GPT-5.2's ARC-Challenge range is 1.67 percentage points, while its mean absolute rank shift from the original protocol is 1.50 places across the four alternatives. Mistral's GPQA range reaches 7.50 points, the audit's largest, while its rank remains sixth. DeepSeek's HellaSwag rank changes by up to three places. These comparisons show how the model panel shapes the interpretation of score differences.

Comparative leadership persists across protocols. GPT-5.2 ranks first on GPQA under all five protocols. Its baseline lead over Gemini is 23.33 percentage points, with an approximate 95% within-benchmark simultaneous contrast band of $[6.67, 40.00]$. The repeated ordering supports a comparative advantage on the audited tasks across the recorded protocols.

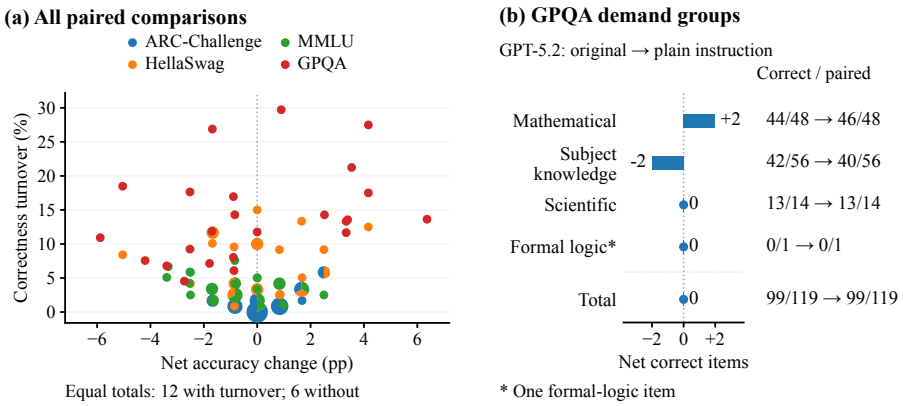


Figure 2: **Paired correctness and demand-group changes.** **(a)** Circle area encodes the number of coincident comparisons, and sectors identify benchmarks across 70 exact locations without jitter. **(b)** GPT-5.2's original/plain GPQA comparison by adjudicated primary demand. Comparisons share observations. Examples are retrospective. Pointwise intervals and complete results are in Appendix C.

Equal accuracy can conceal correctness turnover. Paired comparisons identify the item changes behind equal totals (Figure 2(a) and Figure C2). Of 96 comparisons, 18 preserve accuracy. Twelve contain gains and losses, and six preserve every retained item's correctness. For GPT-5.2 on HellaSwag, original and plain instructions each yield 103/120 correct through nine gains and nine losses. Turnover is 15.0% (pointwise 95% interval $[9.17, 21.67]$), and net change is zero (interval $[-6.67, 6.67]$ percentage points). Its shuffled-option ARC-Challenge comparison preserves 115/120 correct with zero transitions. Across GPQA's 24 comparisons, median absolute net change is 2.62 percentage points and median turnover is 13.45%. ARC-Challenge's corresponding values are 0.83 points and 0.84%.

4.2.4 Construct-Linked Changes and Interpretation

Construct-linked comparisons locate correctness changes within annotated demand groups. Among usable model–item pairs labeled as mathematical reasoning, correctness changes occur in 63/324 pairs (19.4%) under shuffled options and 18/325 (5.5%) under formal-exam instructions. Subject-knowledge rates are 104/839 (12.4%) and 56/840 (6.7%), respectively. These unadjusted pooled rates retain their benchmark and model composition. Denominators count usable model–item pairs, with shared item identities across models. Formal logic contributes one unique item and is interpreted at item level. Appendix C supplies all construct–protocol cells.

Table 2: **Benchmark Validity Card: GPT-5.2 on GPQA.** Authored synthesis. The band concerns only the baseline contrast. Expanded card: Table G4.

Interpretation GPT-5.2 shows a sustained comparative advantage on the audited expert-level science questions within the six-model panel.

Supporting evidence GPT-5.2 ranks first under all five protocols. Its baseline is 82.50% (99/120), with a 23.33-percentage-point lead over Gemini (approximate 95% within-benchmark simultaneous contrast band $[6.67, 40.00]$). Original and plain instructions each yield 99/119 correct, with opposing mathematical and subject-knowledge changes (Figure 2(b)).

Scope and conditions The audit uses six recorded answer-label pipelines on 120 GPQA Diamond items (demand counts in Table 1). Fixed-grid accuracy retains all 120 scheduled items. The original/plain comparison retains 119 items usable under both protocols.

Evidence gaps Transfer to other task mixtures or open-ended scientific problem solving remains untested. Item-demand annotations do not establish a unique internal solution process.

For GPT-5.2 on GPQA, original and plain instructions both yield 99/119 correct on the same paired items (Figure 2(b)). Mathematical-reasoning items rise from 44/48 to 46/48 correct, while subject-knowledge items fall from 42/56 to 40/56. The scientific group remains 13/14, and the formal-logical item remains incorrect. Both annotators' original labels retain the positive mathematical and negative subject-knowledge directions, with differences in membership and attribution. The unchanged total accompanies a redistribution of successful performance across demands.

Table 2 supports GPT-5.2's comparative advantage on the audited GPQA task mixture within the six-model panel across five presentations. Rank evidence establishes the lead's persistence, and construct annotations specify the knowledge and reasoning demands represented. In the original/plain comparison, opposing mathematical and subject-knowledge changes preserve the total (Section 4.2.4). The Card thus distinguishes comparative standing from stability within demand groups.

4.3 Numeric-Answer Task Evaluation

Evaluation Protocol. The GSM8K extension (Cobbe et al., 2021) evaluates 120 word problems under original, formal-exam, and paraphrased presentations, yielding 2,160 records across six pipelines. The DeepSeek pipeline is V3.2 here and V4-Pro in the main audit. A deterministic parser extracts and normalizes the final number for exact reference matching, and all records yield usable answers.

Results. Accuracies range from 94.2% to 99.2%, with a maximum within-model protocol range of 2.50 percentage points. Small changes reorder models near ceiling. Original and formal-exam instructions preserve DeepSeek V3.2's 116/120 correct through one gain and one loss, and GPT-5.2's 119/120 with zero transitions. Results are descriptive, with no bootstrap intervals. Appendix D documents the full panel and scoring procedure. The extension preserves the distinction among scores, item outcomes, and rankings under numeric-answer scoring.

4.4 Summary Evaluation

4.4.1 Expert-Reference Assessment

Evaluation Protocol. The author-released SummEval annotations (Fabbri et al., 2021) cover 100 CNN/DailyMail articles and summaries from sixteen systems. We reserve ten documents for calibration and ninety for evaluation. GPT-5.2 and Claude Sonnet 4.6 each return integer 1–5 ratings for coherence, consistency, fluency, and relevance in one response. Definitions follow SummEval, with expanded anchors specified for this study. The direct prompt (P0) and rephrasings (P1/P2) vary only the opening instruction. The article, summary, rubric, and output format stay fixed. Expert ratings, reference summaries, and system identities are withheld from judges. The mean of three released expert ratings per dimension serves as a fallible comparison reference. Configured ROUGE-L (Lin, 2004) and BERTScore (Zhang et al., 2020) each use the maximum F1 over eleven reference summaries.

Analysis populations and statistical procedures are detailed in Appendices E.1 and E.3.

Ratings correlate with expert judgments. Expert correlations are positive in all four dimensions (Figure 3(a)). Direct-prompt consistency correlations are 0.652 for GPT-5.2 and 0.543 for Sonnet, with a Sonnet-minus-GPT difference of -0.1091 (pointwise 95% interval $[-0.1455, -0.0709]$). Both exceed the configured ROUGE-L and BERTScore correlations of 0.139 and 0.098 on the same summaries. Across sixteen within-judge prompt contrasts, correlation changes range from -0.014 to 0.021, showing modest sensitivity to the tested opening-instruction rephrasings (Figure E1).

Correlation coexists with numerical discrepancies. Positive correlations coexist with numerical discrepancies (Figure E2). GPT's coherence correlation of 0.601 accompanies an average rating 0.919 points above the expert mean. Sonnet's signed difference of -0.016 accompanies a mean absolute difference of 0.705 points. Fluency ratings average 0.810 points below the expert mean for GPT and 1.451 points below it for Sonnet. Ordering and scale agreement thus provide distinct evidence. Appendix E reports pairwise prompt-score changes and within-document associations.

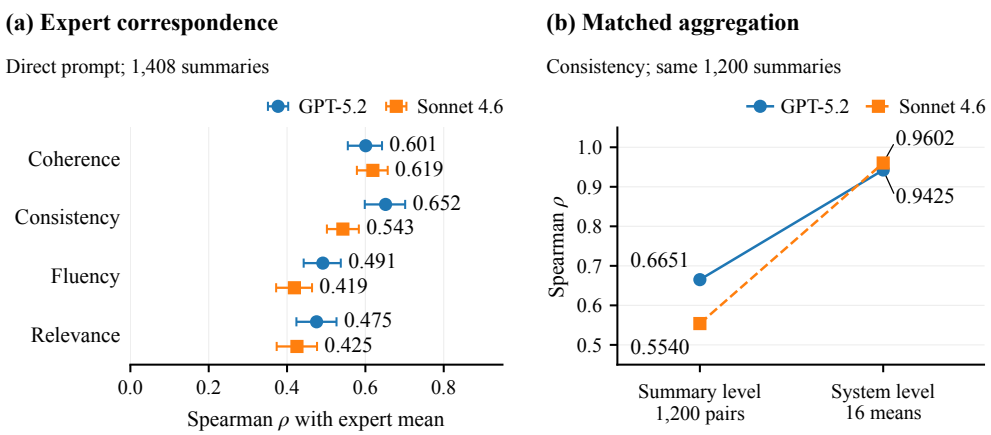


Figure 3: **Expert correspondence and matched aggregation.** **(a)** Direct-prompt Spearman correlations with the three-expert mean on 1,408 summaries from 89 documents, with pointwise 95% document-block bootstrap intervals. **(b)** Consistency at summary and system levels on a separate 1,200-summary subset. Lines connect each judge's point estimates only.

4.4.2 Aggregation and Repeatability

Evaluator ordering changes with aggregation. A separate matched set of 75 complete documents and sixteen systems fixes membership across aggregation levels, giving 1,200 summaries. Each system mean weights those same documents equally. Under the direct prompt, GPT's consistency correlation is 0.6651 across summaries and 0.9425 across system means. Sonnet's is 0.5540 and 0.9602. Their point-estimate ordering reverses (Figure 3(b)). We report point estimates only for this matched analysis. BERTScore's consistency correlation changes from 0.1321 to -0.2931, while ROUGE-L remains close to 0.158. These results tie evaluator comparison to its intended unit of use. Table E3 gives all 32 matched method–dimension comparisons.

Individual-rating recurrence. Identical-input repeats examine rating recurrence. Ten evaluation documents, selected independently of first-pass outcomes, provide 160 summary identities. Rescoring both judges under all three prompts yields 958 paired four-dimensional rating vectors, retaining condition-specific missingness. Across 24 judge–prompt–dimension cells, exact agreement is 81.2%–95.6% and mean absolute change is 0.044–0.188 points over start-to-start lags of 70.3–85.6 hours (Figure E3). For GPT's direct-prompt coherence ratings, ten increases and ten decreases leave the mean unchanged while 20/160 scores move. These descriptive results characterize recurrence at the individual-rating level.

4.4.3 Controlled Factual-Consistency Tests

Evaluation Protocol. We hold summaries fixed while reversing their source support. Each of 45 retained article groups from the same news pool crosses two article versions with four summaries. A/A' express one factual content in two forms, and B/B' express the alternative. Article D_1 supports A/A' and conflicts with B/B'. Edited article D_2 reverses these relations for the unchanged summaries. Two reviewers independently checked support, expression equivalence, and article coherence. Revisions and exclusions preceded scoring. The purposive set contains 31 quantity/time, ten relation/role, and four entity/participant contrasts. Human judgments establish these material relations. Appendix F documents the review process.

Scoring and analysis details are given in Appendix F.2.

Ratings follow factual support. Under D_1, both judges rate supported content higher in all 90 form-specific comparisons, two per group. Mean separations are 2.011 points for GPT and 2.356 for Sonnet (Figure 4). Switching to D_2 lowers A/A' ratings in all 90 comparisons per judge. It raises B/B' ratings in 90/90 GPT comparisons and 88/90 Sonnet comparisons, with two ties in one group. Within-source ordering reverses in both expression forms for all 45 GPT groups and 44

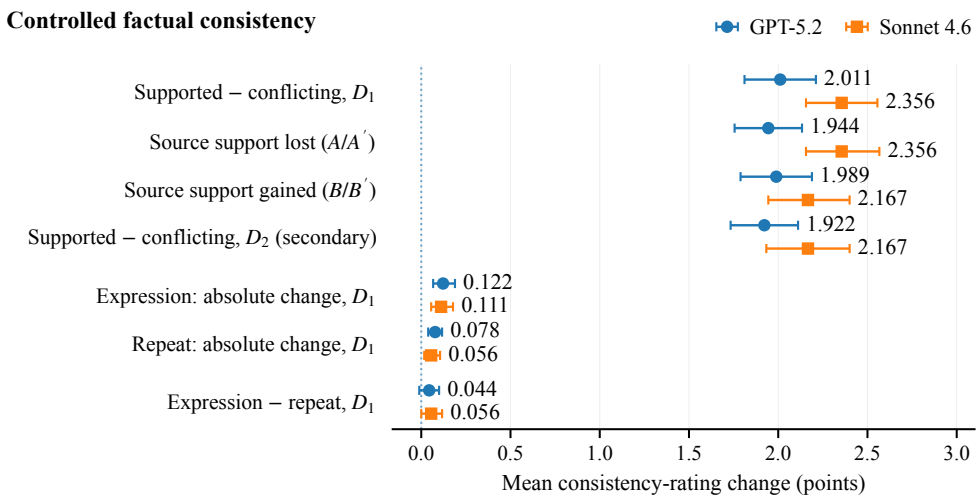


Figure 4: **Controlled factual-consistency contrasts.** Means over 45 equally weighted source groups with pointwise 95% source-group bootstrap intervals. Factual and expression contrasts use the first pass. Positive source-loss values denote decreases after support removal, and source-gain values denote increases after support is added. Expression and repeat rows show mean absolute differences on matched D_1 inputs. The final row is their paired difference.

Sonnet groups. The retained Sonnet exception, F15, ties under D_2 (Appendix F). The second pass preserves the directional pattern and exception.

Expression and repeat variation. Under D_1, meaning-preserving expression changes produce mean absolute differences of 0.122 points for GPT and 0.111 for Sonnet. Matched repeat differences are 0.078 and 0.056. Their paired excesses are 0.044 (pointwise 95% interval $[-0.011, 0.100]$) and 0.056 ($[0.000, 0.117]$), respectively. Across all eight input types per group, repeated consistency scores agree exactly in 318/360 GPT pairs (88.3%) and 341/360 Sonnet pairs (94.7%). Every changed score moves one category. These within-session repeats have a median start-to-start lag of 31.7 minutes, distinct from the natural-summary repeats over 70.3–85.6 hours.

Interpretation. Expert correspondence and bidirectional source contrasts support a factual-consistency interpretation under the tested rubric and materials (Table G6). Numerical discrepancies constrain interchangeability with the expert scale, while aggregation-dependent changes in judges' point-estimate ordering make the evaluation unit essential to comparison. Both studies use a shared news pool.

5 Conclusion

We have introduced Score-to-Claim Auditing (S2C), a validity-argument framework grounded in language testing and educational measurement. S2C links construct, protocol, and scoring evidence to intended interpretations and uses. Stable model leads can accompany score variation, and unchanged totals can conceal item and demand-group changes. Expert correspondence coexists with numerical discrepancies, while judges' point-estimate ordering changes with aggregation. Controlled source reversals support factual-consistency ratings under the tested rubric and materials, with effects exceeding expression and repeat variation. Benchmark Validity Cards document supporting evidence, scope, and gaps, motivating validity profiles alongside benchmark scores. Future longitudinal studies could test whether gains across recurring task demands explain capability development better than aggregate score increases.

AI USE STATEMENT

We used generative AI tools for writing assistance to improve manuscript clarity and organization, and for literature retrieval during related work preparation. Experimental uses of language models are documented in the methods and appendices. The authors directed the research, reviewed materials and code, verified the numerical analyses, and made the final scientific judgments. We reviewed all AI-assisted content and take responsibility for the final manuscript.

ETHICS STATEMENT

The authors performed the construct annotation and the earlier profile-classification annotation. Six participants provided pilot feedback solely to identify problems and refine the experimental materials; their responses were not used to validate S2C. Before participating, they were informed of the research purpose and potential use of their responses in a paper and agreed to participate via WeChat. Participation was voluntary, none was taught or supervised by an author, and each received RMB 50 for participating. Any reporting of pilot feedback in the paper or accompanying materials is limited to aggregate findings. Two professors of linguistics performed paid checks of the controlled materials to establish and refine the intended contrasts; their judgments were not numerical targets for the model ratings. No institutional ethics review was sought for these activities.

REPRODUCIBILITY STATEMENT

Appendices A–D document the construct annotations, benchmark samples, recorded pipelines, scoring, and uncertainty analyses. Appendices E and F specify the natural-summary and controlled-study materials, scoring configurations, analysis populations, and statistical procedures; Appendix G documents the evidence synthesis. Saved model observations, configurations, and analysis scripts support offline recomputation of the main benchmark audit without additional model calls. Replay checks compared the saved row-level records, score and contrast summaries, and uncertainty outputs, retaining the recorded bootstrap multiplicities. The natural-summary and controlled analyses retain their own samples, configurations, and saved outputs. Empirical figures are generated from saved numerical results without additional model calls or resampling. This verification concerns analysis of recorded observations, not regeneration of identical responses through model APIs.

REFERENCES

Lyle F. Bachman and Adrian S. Palmer. *Language Testing in Practice*. Oxford University Press, 1996.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge, 2018. URL https://arxiv.org/abs/1803.05457.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409, 2021. doi: 10.1162/tacl_a_00373. URL https://aclanthology.org/2021.tacl-1.24/.

Avijit Ghosh, Anka Reuel, Jenny Chim, Wm. Matthew Kennedy, Srishti Yadav, Jennifer Mickel, Yanan Long, Andrew Tran, Anastassia Kornilova, Damian Stachura, Kevin Klyman, Felix Friedrich, Jeba Sania, Jan Batzner, Anoop Mishra, Eliya Habba, Yixiong Hao, Nathan Heath, Shalaleh Rismani, Usman Gohar, Andrea Loehr, David Manheim, Ruchira Dhar, Sree Harsha Nelaturu, Aarush Sinha, Leshem Choshen, Drishti Sharma, Ishan Khire, Amit Saha, Subramanyam Sahoo, Michael Hardy, Michael Alexander Riegler, Kabir Manghnani, Michelle Lin, Yanan Jiang, Yilin Huang, Asaf Yehudai, Jessica Ji, Aris Hofmann, Mubashara Akhtar, Max Lamparth, Nuno Moniz, Yacine Jernite, Stella Biderman, Zeerak Talat, Sanmi Koyejo, Mykel

Kochenderfer, and Irene Solaiman. Evaluation cards: An interpretive layer for ai evaluation reporting, 2026. URL https://arxiv.org/abs/2606.09809.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Andong Hua, Kenan Tang, Chenhe Gu, Jindong Gu, Eric Wong, and Yao Qin. Flaw or artifact? rethinking prompt sensitivity in evaluating LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 19889–19899, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1006. URL https://aclanthology.org/2025.emnlp-main.1006/.

Michael T. Kane. Validating the interpretations and uses of test scores. *Journal of Educational Measurement*, 50(1):1–73, 03 2013. ISSN 0022-0655. doi: 10.1111/jedm.12000. URL https://cir.nii.ac.jp/crid/1363670318909230336.

Dongjun Kim, Gyuho Shim, Yongchan Chun, Minhyuk Kim, Chanjun Park, and Heuiseok Lim. Benchmark profiling: Mechanistic diagnosis of LLM benchmarks. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 15635–15650, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.789. URL https://aclanthology.org/2025.emnlp-main.789/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450361255. doi: 10.1145/3287560.3287596. URL https://doi.org/10.1145/3287560.3287596.

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. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=Ti67584b98.

Olawale Elijah Salaudeen, Anka Reuel, Ahmed M Ahmed, Suhana Bedi, Zachary Robertson, Sudharsan Sundar, Benjamin W. Domingue, Angelina Wang, and Sanmi Koyejo. Measurement to meaning: A validity-centered framework for AI evaluation. In *NeurIPS 2025 Workshop on Evaluating the Evolving LLM Lifecycle: Benchmarks, Emergent Abilities, and Scaling*, 2025. URL https://openreview.net/forum?id=2Bw6uC49QF.

Messick Samuel. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9):741–749, 1995. ISSN 0003-066X. doi: 10.1037//0003-066x.50.9.741. URL https://cir.nii.ac.jp/crid/1360862102112411904.

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. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In Anna Korhonen, David Traum, and Lluís Màrquez (eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1472. URL https://aclanthology.org/P19-1472/.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkeHuCVFDr.

A Construct Annotation and Assisted Labeling

A.1 Operational construct codebook

The codebook describes the demands an item is judged to place on a model, rather than the model's internal solution process. Each item receives exactly one primary construct from Table A1. A secondary label is added only for another demand necessary to answer, not merely a topic mentioned in the stem; at most two secondary labels are allowed, without repeating the primary. Annotators read the stem and options, identify the necessary demands, assign labels, rate loads and artifacts, and record a rationale.

Table A1: The twelve-label primary-demand codebook. Operational guidance describes annotated item demands, not a direct measurement of latent model processes.

Primary label	Operational guidance
Factual recall	Mainly knowing or remembering a fact.
Scientific reasoning	Applying a scientific principle, mechanism, or scientific causal relation.
Discourse coherence	Selecting what fits discourse context, narrative flow, or textual coherence.
Event continuation	Predicting a plausible next event or continuation.
Causal reasoning	Identifying cause, effect, consequence, mechanism, or intervention.
Pragmatic inference	Inferring intention, implication, or unstated communicative context.
Lexical-semantic inference	Word meaning or local semantic relations are central.
Formal-logical reasoning	Formal logic, symbolic reasoning, rule application, or abstract structure.
Cultural/social knowledge	Social, cultural, institutional, ethical, or normative knowledge.
Subject-specific knowledge	Domain knowledge not primarily scientific, mathematical, or formal-logical reasoning.
Mathematical reasoning	Calculation, numerical reasoning, algebra, or mathematical proof.
Other	None of the categories fits; a rationale is required.

Linguistic complexity, world-knowledge load, reasoning load, and cultural-context load use ordinal levels 0–3 (minimal/absent through high). Instructions favor the lower level when a higher one is not clearly required. Test-taking artifacts use none, weak, moderate, or strong for option elimination, implausible distractors, answer position, or lexical matching. Construct mismatch uses none, minor, or major relative to the benchmark's stated capability interpretation. The GPQA addendum warns against labeling every science item scientific reasoning: specialized knowledge, application of a principle, and quantitative reasoning remain distinct choices.

A.2 Annotation agreement

Two author annotators independently labeled the items before adjudication. The original 360-item ARC-Challenge/HellaSwag/MMLU set and the later 120-item GPQA set retain separate annotator A/B sheets and pre-adjudication agreement summaries. Finalized labels cover all 480 items, but the reliability estimates are not pooled. Table A2 reports agreement between annotators before adjudication, rather than agreement with the finalized labels or evidence that the taxonomy is uniquely correct.

Agreement varies across fields and strata. The main construct-group analyses use adjudicated primary labels; secondary labels and construct-mismatch judgments remain exploratory. Primary-label agreement does not imply uniform reliability across the auxiliary fields.

Coarse-label sensitivity. A supplementary GPQA analysis applies the recorded four-group mapping to the same two pre-adjudication label sets. Subject-specific knowledge and factual recall form *expert-science knowledge*; scientific and causal reasoning form *scientific/causal reasoning*; mathematical and formal-logical reasoning form *formal/quantitative reasoning*; the remaining codebook labels form the *other-or-misaligned* category. The resulting agreement is 88/120 items (73.3%), with Cohen's $\kappa = 0.586$. This describes sensitivity to a coarser grouping, not a replacement for the fine-label estimate. The main construct profiles and construct-linked comparisons continue to use adjudicated fine labels.

Table A2: Pre-adjudication agreement by annotation stratum. Weighted coefficients use quadratic weights. GPQA cultural-context weighted agreement is undefined despite exact agreement of 1.0; the dash is not zero.

Field / statistic	Original 360	GPQA 120
Primary construct, Cohen's κ	0.748	0.568
Primary construct, exact agreement	0.811	0.717
Construct mismatch, Cohen's κ	0.002	0.244
Test-taking artifact, Cohen's κ	0.921	0.200
Linguistic complexity, weighted κ	0.723	0.580
World-knowledge load, weighted κ	0.746	0.547
Reasoning load, weighted κ	0.504	0.593
Cultural-context load, weighted κ	0.741	—
Secondary constructs, mean Jaccard	0.315	0.458

A.3 Assisted-labeling evaluation

The GPT-5.2 labeling pilot uses 90 calibration examples and 270 held-out items, with 90 test items from each of ARC-Challenge, HellaSwag, and MMLU. The test set is disjoint from calibration by benchmark/item identity; GPQA is not part of this pilot. Adjudicated human labels are the comparison reference. Overall exact accuracy is 0.7926 (214/270), macro-F1 is 0.5611, and Cohen's κ is 0.7314. Table A3 retains the benchmark-specific results.

Table A3: Assisted primary-label classification against adjudicated labels. HellaSwag's κ is undefined because both the reference labels and predictions are constant; the dash is not zero.

Benchmark	Items	Accuracy	Macro-F1	κ
ARC-Challenge	90	0.500	0.257	0.176
HellaSwag	90	1.000	1.000	—
MMLU	90	0.878	0.609	0.683

The benchmark-specific results motivate review of pre-labels rather than reliance on pooled accuracy alone. HellaSwag's perfect accuracy concerns a single reference category, event continuation; it does not assess discrimination among construct labels. The pilot predictions do not supply or replace the human labels used in the main audit.

Table A4 reports the error distribution for the same 270 held-out predictions. In ARC-Challenge, 21 scientific-reasoning items were predicted as factual recall, and 14 causal-reasoning items as scientific reasoning. These disagreements identify specific label boundaries for benchmark-specific review without establishing that every adjudicated reference judgment is uniquely correct.

Table A4: Held-out assisted-labeling confusion counts. Rows are adjudicated primary labels; columns are GPT-5.2 predictions. Only rows with a nonzero reference count and columns with a nonzero prediction count are shown. Column abbreviations denote causal reasoning, cultural/social knowledge, event continuation, factual recall, mathematical reasoning, pragmatic inference, scientific reasoning, and subject-specific knowledge, respectively. All 270 held-out items are included.

Reference label	Cau.	Cul.	Event	Fact	Math	Prag.	Sci.	Subj.
Causal reasoning	12	0	0	2	0	1	14	0
Cultural/social knowledge	0	11	0	0	0	1	0	8
Event continuation	0	0	90	0	0	0	0	0
Mathematical reasoning	0	0	0	0	4	0	0	2
Scientific reasoning	3	0	0	21	1	0	32	2
Subject-specific knowledge	1	0	0	0	0	0	0	65

B MAIN AUDIT: SAMPLES, MODELS, AND PROTOCOLS

B.1 MODEL PANEL, SAMPLING, AND PROTOCOL

Table B1 records the six model pipelines used in the multiple-choice audit. The historical alias `deepseek32` denotes V4-Pro here and V3.2 in the GSM8K extension; the alias is therefore interpreted within its study rather than treated as a common model identity. Returned-model and execution metadata are incomplete in some historical records.

Table B1: Model names and recorded identifiers for the six multiple-choice evaluation pipelines.

Model name	Recorded audit identifier
GPT-5.2	`openai/gpt-5.2`
Claude Sonnet 4.6	`anthropic/claude-sonnet-4.6`
Gemini 2.5 Pro	`google/gemini-2.5-pro`
DeepSeek V4-Pro	`deepseek-v4-pro`
Llama 4 Maverick	`meta-llama/llama-4-maverick`
Mistral Large 2512	`mistralai/mistral-large-2512`

A historical metadata inconsistency affects 3,600 replacement rows in the original three-benchmark portion of the audit: the `model_version_or_id` field retains the earlier GPT identifier for 1,800 Claude and 1,800 DeepSeek rows. Their `response_model` fields and study-specific alias mapping identify the corresponding pipelines. Analyses group these records by the documented study aliases; the original records remain unchanged.

Sampling. ARC-Challenge uses 120 saved validation items from `allenai/ai2_arc` (ARC-Challenge configuration); HellaSwag uses 120 from `Rowan/hellaswag`. The main sampling script specifies seed 2026 and sampling without replacement. MMLU pools validation items within the four selected groups in Table B2, then selects 30 items per group. There is no per-subject quota, and the equal-group sample is not an all-subject MMLU sample. Matching metadata and a documented sampling script do not establish a particular historical invocation or dataset revision.

Table B2: Four subjects pooled within each MMLU group. Thirty items are selected per group, not equally per subject.

MMLU group	Selected subjects	Items
Humanities	Philosophy; high-school European history; high-school US history; world religions	30
Social sciences	Sociology; high-school psychology; high-school government and politics; econometrics	30
STEM	Abstract algebra; college physics; high-school chemistry; computer security	30
Professional	Professional law; professional medicine; business ethics; clinical knowledge	30

GPQA's saved metadata identify `Idavidrein/gpqa`, `gpqa_diamond`, and `train`. The sampling-source account specifies completeness filtering and a default seed of 42, but the actual invocation and upstream source-row identities are not recovered from reset local indices. Inference remains conditional on the preserved samples.

Protocol variants. Table B3 summarizes the five multiple-choice conditions. The intended changes concern instruction register, instruction style, option order, and question wording. For shuffled options, the displayed reference label is remapped to the new order; correctness is evaluated against that displayed label without applying a second remapping. Semantic paraphrases are intended to preserve the answer options and reference answer; preservation flags do not independently establish semantic equivalence.

Table B3: Five multiple-choice conditions and their operational variations.

Condition	Operational variation
Original baseline	Original benchmark item in answer-label form.
Formal exam	Examination-style instruction, varying register.
Plain instruction	Simplified direct instruction, varying instruction style.
Shuffled options	Deterministic option reordering with the displayed gold label remapped.
Semantic paraphrase	Reworded question/stem intended to preserve answer options and gold answer.

Scoring and retained observations. Corrected scoring accepts the preserved exact, explicit, or fallback parser outcomes only when there is no API error and the parsed label belongs to the item's choice alphabet. ARC-Challenge includes three-, four-, and five-option items; the other three benchmarks use four options. The resulting 14,311 usable and 89 unusable records retain the complete 14,400-row grid. Fixed-grid accuracy counts usable correct responses out of 120 scheduled items per cell: unusable records remain in the denominator and contribute no correct response, while remaining distinct from observed wrong answers. Paired comparisons require usable responses to the same item under both the original and the alternative protocol. Heterogeneous historical repair and parsing histories limit interpretations as standardized first-attempt or isolated prompt effects.

C Protocol Results and Paired Comparisons

C.1 Scoring, ties, and uncertainty

Scores and ranks. Panel-level accuracy counts usable correct responses out of 120 scheduled items per model–benchmark–protocol cell. Usable-only accuracy and coverage remain separate quantities in the numerical outputs. A model's protocol score range is its maximum minus minimum fixed-grid accuracy across the five conditions. Within each benchmark and condition, rank is one plus the number of models with a strictly higher correct count, so ties share a position. Writing $R_{m,q}$ for the rank of model m under condition q and q_0 for the original condition, the model's mean absolute rank shift is $\frac{1}{4}\sum_{q\neq q_0}|R_{m,q}-R_{m,q_0}|$. Table C1 distinguishes the maximum of these model-level means from the largest single-condition absolute rank shift. Neither is a condition-level mean over models.

Fixed-grid uncertainty. The fixed-grid analysis uses 10,000 shared item-block bootstrap draws within each benchmark, retaining all six models and five conditions for each selected item. MMLU draws thirty items within each of its four strata. The approximate simultaneous-band family contains 135 contrasts per benchmark: 75 between-model comparisons within a protocol (15 model pairs across five protocols) and 60 between-protocol comparisons within a model (ten protocol pairs across six models). The bands use the 95th percentile of the maximum absolute centered bootstrap deviation within that family; coverage is not asserted jointly across all four benchmarks. For the main-text GPQA example, the saved Gemini-minus-GPT contrast is reversed to express GPT's lead, with both the signs and order of the interval endpoints reversed accordingly. Percentile spreads for maximum score ranges and rank summaries are retained as resampling summaries, not treated as ordinary confidence intervals for those nonsmooth quantities.

Paired outcomes and uncertainty. Each paired comparison retains items with usable responses under both the original and the alternative protocol. Net accuracy change, Δ, is the gain rate minus the loss rate; correctness turnover is their sum. The paired analysis uses a separately saved family of 10,000 item-block draws, shared across the 24 comparisons within each benchmark and retaining the same item/stratum structure. Each resample recomputes the comparison's both-usable denominator. Pointwise intervals use the 2.5th and 97.5th percentiles with linear interpolation, rather than the simultaneous-band procedure above. These intervals are conditional on the sampled items and recorded pipelines. The comparisons are retrospective; historical repair and execution records do not support interpreting every contrast as a standardized first-attempt prompt effect.

Table C1: Observed benchmark summaries. The three maximum statistics use the fixed 120-item grid: the largest within-model score range, the largest model-level mean absolute rank shift, and the largest single-condition absolute rank shift. The two medians summarize 24 baseline–variant comparisons per benchmark on their respective both-usable item sets. Δ is paired net accuracy change; pp denotes percentage points.

Benchmark	Max range (pp)	Max mean rank shift	Max single rank shift	Median $\lvert\Delta\rvert$ (pp)	Median turnover (%)
ARC-C	2.50	1.50	2	0.83	0.84
HellaSwag	6.67	1.00	3	0.87	7.12
MMLU	3.33	1.50	2	1.25	3.39
GPQA	7.50	1.00	1	2.62	13.45

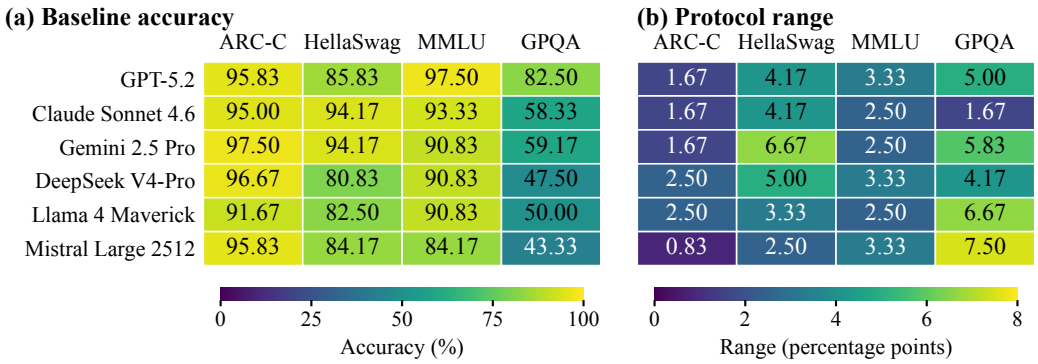


Figure C1: **Baseline accuracy and observed protocol ranges.** **(a)** Usable correct responses as a percentage of 120 scheduled items per cell; unusable responses remain in the denominator. **(b)** Maximum minus minimum accuracy across five protocols within each model–benchmark pair, in percentage points. Ranges are observed spreads, not confidence intervals. Aligned rows share the model labels; color scales are separate.

Figure C1 reports the baseline accuracies and observed protocol ranges for all 24 model–benchmark profiles.

C.2 Construct-linked correctness changes

Table C2 pools adjudicated-label transition counts within each construct–protocol cell. The numerator counts correctness changes relative to the original protocol; the denominator counts usable model–item pairs, not independent items. Across the 480 annotated items, the unique item counts are 141 for subject-specific knowledge, 97 for scientific reasoning, 56 for mathematical reasoning, one for formal logic, 36 for causal reasoning, 120 for event continuation, and 29 for cultural/social knowledge. Pooled cells reflect differing benchmark composition and model outcomes; they are descriptive rather than adjusted construct effects. The formal-logical row therefore does not support a construct-wide conclusion despite its multiple model observations. The saved numerical records retain benchmark, model, and both-usable denominator information for each row, under all three label views: adjudicated, annotator A, and annotator B.

Table C2: Correctness changes / usable model–item pairs (%). Counts pool gains and losses within each adjudicated primary demand and alternative protocol. Formal-logical reasoning has only one unique item despite its multiple model observations.

Primary demand	Formal	Plain	Shuffle	Paraphrase
Subject-specific knowledge	56/840 (6.7)	59/840 (7.0)	104/839 (12.4)	73/837 (8.7)
Scientific reasoning	8/578 (1.4)	4/579 (0.7)	33/579 (5.7)	17/578 (2.9)
Mathematical reasoning	18/325 (5.5)	26/321 (8.1)	63/324 (19.4)	41/322 (12.7)
Formal-logical reasoning	0/5 (0.0)	1/5 (20.0)	1/5 (20.0)	4/5 (80.0)
Causal reasoning	2/215 (0.9)	5/215 (2.3)	6/215 (2.8)	5/215 (2.3)
Event continuation	33/716 (4.6)	44/716 (6.1)	70/714 (9.8)	57/715 (8.0)
Cultural/social knowledge	2/172 (1.2)	3/172 (1.7)	7/172 (4.1)	6/171 (3.5)

C.3 Per-comparison scores and uncertainty

The saved numerical outputs retain 120 cell results, 24 model–benchmark profiles, 540 within-benchmark contrasts, all 96 paired correctness comparisons with pointwise intervals, and 840 construct-group records across the three label views. Complete gain/loss and missingness counts accompany the paired outcomes. The main-text benchmark profiles and the selected GPQA and HellaSwag illustrations use these saved outputs. Their intervals are taken from the recorded analyses rather than recomputed for manuscript presentation.

Figure C2 illustrates the two GPT-5.2 equal-accuracy cases discussed in Section 4.2.3: compensating gains and losses versus preserved correctness for every retained item.

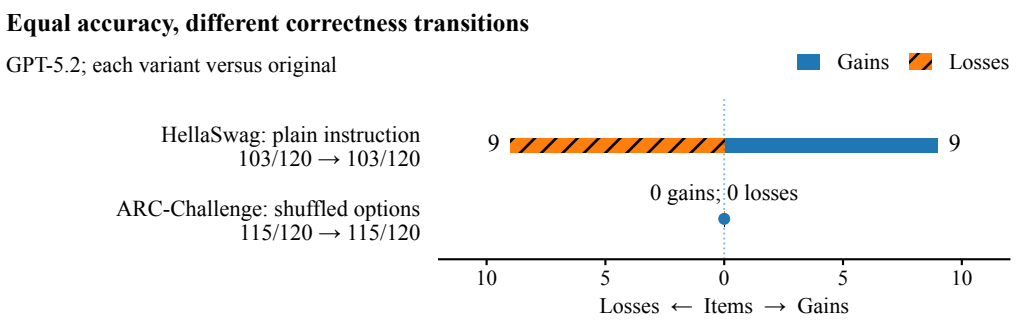


Figure C2: **Two equal-accuracy GPT-5.2 comparisons.** Original versus plain instructions on HellaSwag retains 103/120 correct through nine gains and nine losses; original versus shuffled options on ARC-Challenge retains 115/120 with no correctness transitions. Bars are numbers of items, not rating differences or confidence intervals. Both comparisons use 120 items with usable responses under both protocols. The examples are retrospective.

D Numeric-Answer Extension

The GSM8K extension evaluates 120 word problems with six model pipelines under three conditions: original rendering, formal-exam framing, and semantic paraphrase. Table D1 lists the model panel and its scores. The historical DeepSeek alias refers to V3.2 in this extension, not the V4-Pro pipeline in the multiple-choice audit. Paraphrases were constructed to preserve quantities, mathematical relations, and the reference numeric answer.

The deterministic numeric-answer parser first searches for an explicit final-answer expression; if none is found, it uses the last numeric expression. Correctness is determined by exact comparison with the reference after numeric normalization. All 2,160 recorded outputs are usable under this procedure; historical retry information remains in the saved CSV. Each of the eighteen model–condition cells therefore contains 120 usable outputs, so fixed-grid and usable-only accuracy coincide.

Paired analyses compare the original condition with each of the two alternatives for every model, giving twelve baseline–variant comparisons, each retaining all 120 items. Results for this extension are descriptive; no bootstrap intervals were computed.

Table D1: GSM8K numeric-answer accuracy (%). Each model–condition cell contains 120 items, with all recorded outputs usable under the numeric parser. Fixed-grid and usable-only accuracy therefore coincide.

Model	Original	Formal exam	Paraphrase
GPT-5.2	99.17	99.17	97.50
Claude Sonnet 4.6	97.50	99.17	96.67
Gemini 2.5 Pro	97.50	96.67	98.33
DeepSeek V3.2	96.67	96.67	95.83
Llama 4 Maverick	96.67	94.17	95.00
Mistral Large 2512	95.83	96.67	95.83

E Natural-Summary Evaluation

E.1 Data, scoring protocol, and analysis populations

The study uses the author-released SummEval annotation file (Fabbri et al., 2021). Articles are joined by their released CNN/DailyMail filepath; nonempty story lines before the first `@highlight` marker are joined with spaces. A deterministic hash order of document keys assigns ten documents to calibration and ninety to evaluation. The planned evaluation contains sixteen

system summaries per evaluation document, giving 1,440 summary identities. This is a split within an existing public collection, not a newly sampled test domain.

Each request contains the article and one candidate summary. Each judge returns all four dimension ratings in a single response; the judges receive the same dimension order, definitions, five-point anchors, and output schema. The direct instruction (P0) is "Evaluate the candidate summary against the source article using the rubric below." P1 is "Use the rubric below to assess the candidate summary in relation to the source article" and P2 is "Apply the rubric below when rating the candidate summary with respect to the source article." Everything after this opening is unchanged. These narrow instruction contrasts do not change the rubric or supply exemplars. Table E1 records the scoring and metric configurations.

Table E1: Scoring and metric configuration. These are the study's configured measurements, not a reconstruction of official SummEval metric values.

Component	Configuration
Judges	`openai/gpt-5.2` and `anthropic/claude-sonnet-4.6`, routed through OpenRouter to OpenAI and Anthropic respectively, without provider fallback.
Generation	Nonstreaming; maximum 512 output tokens. Temperature, top-p, and seed are omitted. GPT uses reasoning effort `none`; Claude does not receive a reasoning option.
Visible information	Rubric, article, and candidate; no reference summaries, expert scores, system identities, split identifiers, previous outputs, or expected judgments.
Output	Joint integer ratings for coherence, consistency, fluency, and relevance in that order. All four scores must be present, each from 1 to 5; unusable joint outputs remain unavailable.
Expert reference	Arithmetic mean of exactly three released expert scores for each dimension. The reference is not assumed error-free or numerically identical to the new rubric.
ROUGE-L	`rouge-score` 0.1.2; `rougeL` F1 (not `rougeLsum`); default tokenizer, no stemming or added normalization.
BERTScore	`bert-score` 0.3.13; RoBERTa-base, layer 10; slow tokenizer; CPU float32; no IDF or baseline rescaling; batch size one. Overflow is rejected rather than truncated.
References	Maximum of eleven reference-wise F1 scores for each metric; precision and recall are not maximized separately. All reference positions are treated symmetrically.

The first pass yields 8,537 usable joint judge ratings out of 8,640 planned requests. The remaining records comprise 86 unattempted requests following the quarantine of one document, sixteen terminally unavailable requests, and one unresolved request. All planned summaries have the three-expert criterion and both metric values. Requiring the six judge–prompt conditions and the same-dimension criterion gives 1,408 summaries from 89 contributing documents in each dimension. The metric-comparison mask additionally requires both metric values and retains the same identities. A missing response is not assigned a low quality rating.

The matched-aggregation analysis uses a separate fixed population: 75 documents with all sixteen systems and all eight scoring methods available (six judge–prompt conditions and two metrics), giving 1,200 summary pairs. Each system mean weights those same 75 documents equally. Pairwise prompt-score comparisons instead retain 1,417–1,420 usable summary pairs. These analysis-specific supports are not substituted for the primary 1,408-summary population.

E.2 Rubric and output instructions

The dimension definitions follow SummEval. The anchors below are those specified for this study, not the original instructions given to SummEval's expert annotators. Before the rubric, the prompt directs judges to treat both texts as data, ignore embedded instructions, use the supplied article rather than outside information, and assess the four dimensions separately.

Coherence. organization and connections among the summary's sentences.

1: The summary is fragmented or its organization makes the overall account very difficult to follow.
2: Frequent abrupt shifts, unclear connections or ordering problems substantially disrupt the account.
3: The overall account can be followed, but some ordering or connection problems interrupt it.
4: The account is well organized and connected, with only a minor transition or ordering weakness.
5: The account is consistently organized, connected and easy to follow as a whole.

Consistency. factual support of the summary's claims by the source article.

1: Most material claims are unsupported by or contradict the article.
2: A central claim, or several material claims, are unsupported or contradicted.
3: The main account is supported, but at least one noncentral material claim is unsupported or contradicted.
4: Material claims are supported, with only a minor peripheral imprecision that does not change the main account.
5: All material claims are supported by the article, without factual distortion.
Assess support from the supplied article, not outside knowledge. Do not equate plausible wording with source support.

Fluency. linguistic quality and readability of individual sentences.

1: Pervasive language problems make most sentences difficult to understand.
2: Frequent grammatical, wording, punctuation, capitalization or formatting problems seriously impair readability.
3: Sentences are generally understandable, but noticeable language problems interrupt reading.
4: Sentences read smoothly, with only minor language or formatting problems.
5: Sentences are clear and linguistically well formed, without noticeable readability problems.
Assess linguistic form here; do not substitute factual accuracy or content coverage for fluency.

Relevance. selection of salient article content without unnecessary or repeated content.

1: The summary largely misses the article's central content or is dominated by unnecessary or repeated material.
2: Major central content is omitted, or substantial unnecessary or repeated material crowds out salient content.
3: Some central content is captured, but meaningful omissions or unnecessary repetition remain.
4: The main salient content is selected, with only a minor omission or unnecessary detail.
5: The summary selects the article's central content concisely, without material omissions or unnecessary repetition.
Assess salient selection for a summary, not reproduction of every article detail.

The closing instruction identifies the next message as a JSON object containing the source article followed by the candidate summary. It requests only a JSON object with the keys `coherence`, `consistency`, `fluency`, and `relevance`; each value must be an integer from 1 to 5, with no additional explanation.

E.3 Statistical definitions and uncertainty

Spearman correlation is the Pearson correlation of average-tie ranks, with one weight per retained document–system pair. Expert ties preserve the integer numerator of the three-score mean. Signed discrepancy is judge score minus expert mean; absolute discrepancy takes its magnitude. These

numerical summaries adopt equal spacing of the recorded rating categories. Within-document correlations compare the available system summaries of one article; averages over documents with a defined correlation are not pooled-summary correlations.

The saved uncertainty analysis uses 10,000 shared resamples of ninety document blocks, including the document with no retained judge scores. Each sampled document carries all systems, conditions, and dimensions with its multiplicity. Eligibility masks remain fixed, while ranks are recomputed on the resampled multiset. The oriented correlation contrasts are calculated within each shared draw, rather than by subtracting separate interval endpoints. Intervals use the 2.5th and 97.5th percentiles with linear interpolation. They are pointwise document-resampling intervals, not simultaneous bounds or uncertainty over fresh model generations. All 32 association intervals and twenty contrast intervals are defined. No system-mean or matched-aggregation interval is inferred from these intervals, and no intervals are assigned to the descriptive signed or absolute rating differences. Full-precision values remain in the saved numerical outputs.

E.4 Expert correspondence and scale discrepancies

Table E2 expands the direct-prompt results in Figure 3(a) and Figure E2 to all six judge–prompt conditions and both configured reference-based metrics. Figure E1 retains all sixteen within-judge instruction contrasts and four between-judge direct-prompt contrasts. The former use P1 or P2 minus P0 within a judge; the latter use Sonnet minus GPT under P0. Each estimate retains this original orientation.

Table E2: All 32 expert associations on 1,408 matched summaries, with the 24 judge-scale discrepancy summaries. Intervals are pointwise document-block intervals. A dash denotes a non-comparable numerical scale, not missing correlation data. Signed and absolute judge–expert differences use rating points.

Dimension	Method	ρ	95% interval	$\overline{d}$	$\overline{\lvert d \rvert}$
Coherence	GPT-5.2 / P0	0.601	[0.555, 0.642]	+0.919	1.031
Coherence	GPT-5.2 / P1	0.609	[0.563, 0.651]	+0.916	1.031
Coherence	GPT-5.2 / P2	0.599	[0.557, 0.639]	+0.955	1.066
Coherence	Sonnet 4.6 / P0	0.619	[0.578, 0.657]	-0.016	0.705
Coherence	Sonnet 4.6 / P1	0.607	[0.564, 0.648]	-0.064	0.711
Coherence	Sonnet 4.6 / P2	0.605	[0.562, 0.645]	-0.067	0.708
Coherence	ROUGE-L	0.195	[0.126, 0.262]	–	–
Coherence	BERTScore	0.344	[0.283, 0.399]	–	–
Consistency	GPT-5.2 / P0	0.652	[0.598, 0.702]	-0.116	0.292
Consistency	GPT-5.2 / P1	0.646	[0.595, 0.695]	-0.119	0.289
Consistency	GPT-5.2 / P2	0.665	[0.610, 0.716]	-0.094	0.272
Consistency	Sonnet 4.6 / P0	0.543	[0.502, 0.583]	-0.589	0.660
Consistency	Sonnet 4.6 / P1	0.542	[0.501, 0.583]	-0.598	0.670
Consistency	Sonnet 4.6 / P2	0.546	[0.505, 0.586]	-0.612	0.678
Consistency	ROUGE-L	0.139	[0.078, 0.199]	–	–
Consistency	BERTScore	0.098	[0.044, 0.155]	–	–
Fluency	GPT-5.2 / P0	0.491	[0.442, 0.537]	-0.810	0.903
Fluency	GPT-5.2 / P1	0.513	[0.464, 0.559]	-0.821	0.912
Fluency	GPT-5.2 / P2	0.498	[0.453, 0.542]	-0.798	0.902
Fluency	Sonnet 4.6 / P0	0.419	[0.372, 0.464]	-1.451	1.482
Fluency	Sonnet 4.6 / P1	0.426	[0.380, 0.471]	-1.458	1.489
Fluency	Sonnet 4.6 / P2	0.422	[0.378, 0.465]	-1.429	1.459
Fluency	ROUGE-L	0.088	[0.021, 0.155]	–	–
Fluency	BERTScore	0.127	[0.054, 0.200]	–	–

Continued on next page

Table E2 (continued)

Dimension	Method	ρ	95% interval	$\overline{d}$	$\overline{\lvert d \rvert}$
Relevance	GPT-5.2 / P0	0.475	[0.424, 0.526]	-0.183	0.594
Relevance	GPT-5.2 / P1	0.480	[0.428, 0.531]	-0.170	0.583
Relevance	GPT-5.2 / P2	0.484	[0.432, 0.534]	-0.155	0.569
Relevance	Sonnet 4.6 / P0	0.425	[0.374, 0.477]	-0.796	0.939
Relevance	Sonnet 4.6 / P1	0.434	[0.383, 0.484]	-0.848	0.975
Relevance	Sonnet 4.6 / P2	0.443	[0.391, 0.493]	-0.829	0.957
Relevance	ROUGE-L	0.227	[0.142, 0.310]	–	–
Relevance	BERTScore	0.331	[0.262, 0.396]	–	–

Figure E2 displays the direct-prompt signed and absolute judge–expert differences reported in Table E2.

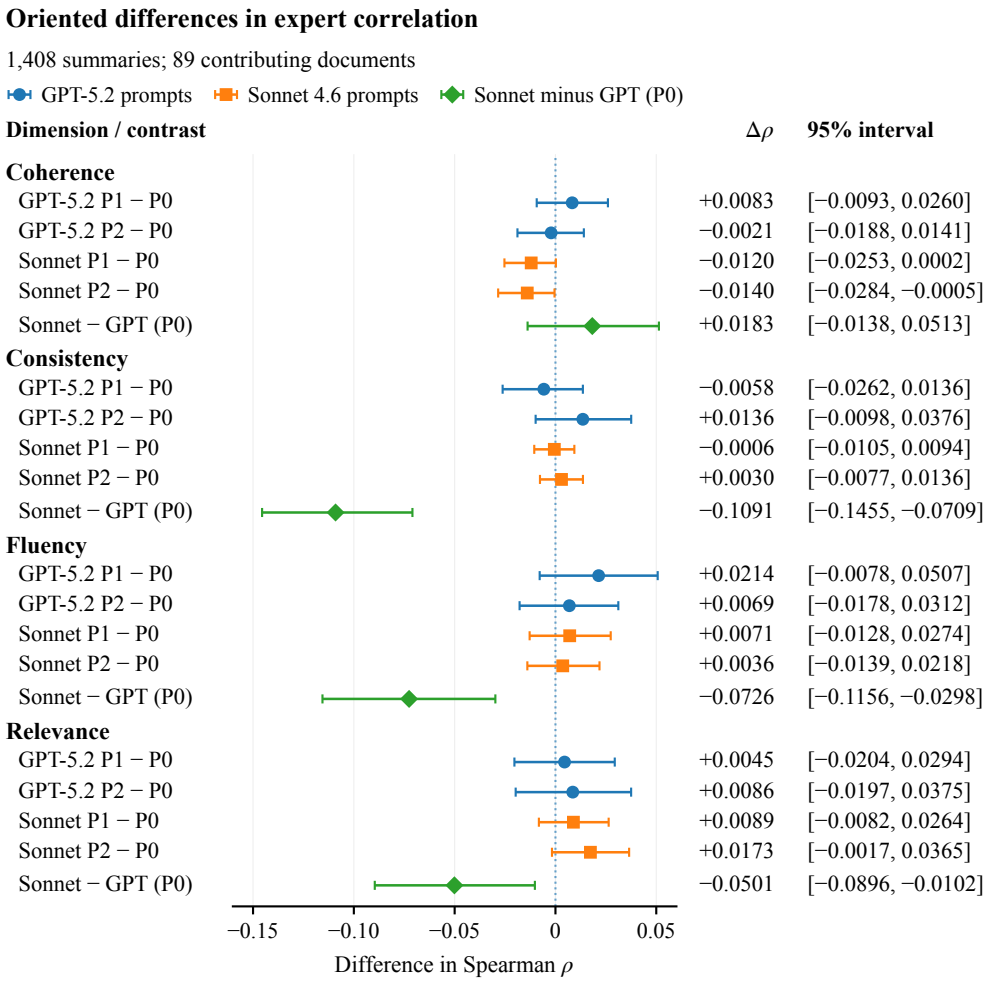


Figure E1: **Oriented differences in expert correlation.** All twenty contrasts on the common 1,408-summary set: sixteen within-judge instruction contrasts (P1 or P2 minus P0) and four Sonnet-minus-GPT contrasts under P0. Positive values denote higher correlation with the three-expert mean for the first-named condition. Bars are pointwise 95% intervals from 10,000 shared document-block bootstrap draws; each difference is calculated within a draw. Numerical columns retain the estimates and interval endpoints. Values are correlation differences, not rating-point differences; the intervals are not simultaneous bounds.

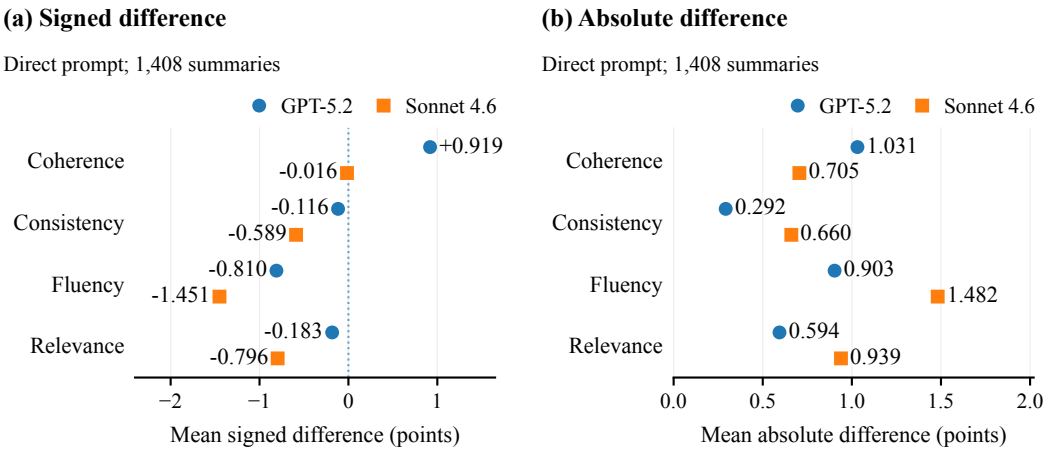


Figure E2: **Signed and absolute differences from expert ratings.** Direct-prompt results on the common 1,408 summaries from 89 documents. **(a)** Mean judge-minus-expert-mean difference. **(b)** Mean absolute difference. Both use points on the recorded 1–5 scale, treating its categories as equally spaced. These are descriptive estimates with no assigned uncertainty intervals; Table E2 retains all conditions.

E.5 Aggregation and additional observational views

Table E3 compares pooled-summary and system-mean associations on identical document membership. Each method uses the same 1,200 summary pairs; the system-level coefficient compares sixteen means, each calculated over the same 75 documents. Table E4 reports numerical score changes under the instruction rephrasings on pairwise-usable supports. Table E5 reports conditional within-document associations and retains the defined and undefined document counts. These supplementary views use different comparison units and eligibility rules, not additional independent datasets.

Table E3: Matched aggregation analysis: all cells use 75 documents and 16 systems. Summary-level coefficients use 1,200 pairs; system-level coefficients use sixteen means over those same documents. No uncertainty intervals are assigned to these matched-aggregation coefficients.

Dimension	Method	Summary level	System level
Coherence	GPT-5.2 / P0	0.5939	0.8319
Coherence	GPT-5.2 / P1	0.6028	0.8492
Coherence	GPT-5.2 / P2	0.5947	0.8138
Coherence	Sonnet 4.6 / P0	0.6163	0.8588
Coherence	Sonnet 4.6 / P1	0.6021	0.8698
Coherence	Sonnet 4.6 / P2	0.6043	0.8471
Coherence	ROUGE-L	0.1946	0.4412
Coherence	BERTScore	0.3499	0.3294
Consistency	GPT-5.2 / P0	0.6651	0.9425
Consistency	GPT-5.2 / P1	0.6634	0.9601
Consistency	GPT-5.2 / P2	0.6800	0.9527
Consistency	Sonnet 4.6 / P0	0.5540	0.9602
Consistency	Sonnet 4.6 / P1	0.5577	0.9705
Consistency	Sonnet 4.6 / P2	0.5561	0.9617
Consistency	ROUGE-L	0.1583	0.1576
Consistency	BERTScore	0.1321	-0.2931

Continued on next page

Table E3 (continued)

Dimension	Method	Summary level	System level
Fluency	GPT-5.2 / P0	0.4916	0.7749
Fluency	GPT-5.2 / P1	0.5170	0.6933
Fluency	GPT-5.2 / P2	0.4935	0.6148
Fluency	Sonnet 4.6 / P0	0.4180	0.5701
Fluency	Sonnet 4.6 / P1	0.4269	0.5210
Fluency	Sonnet 4.6 / P2	0.4193	0.5920
Fluency	ROUGE-L	0.0987	0.4772
Fluency	BERTScore	0.1428	0.1414
Relevance	GPT-5.2 / P0	0.4629	0.8489
Relevance	GPT-5.2 / P1	0.4634	0.8668
Relevance	GPT-5.2 / P2	0.4669	0.8036
Relevance	Sonnet 4.6 / P0	0.4120	0.7629
Relevance	Sonnet 4.6 / P1	0.4245	0.7290
Relevance	Sonnet 4.6 / P2	0.4319	0.7741
Relevance	ROUGE-L	0.2381	0.5765
Relevance	BERTScore	0.3322	0.1471

Table E4: Prompt-score differences (P1 or P2 minus P0) on pairwise-usable supports, which range from 1,417 to 1,420 summaries. These descriptive rating-point changes do not replace the common 1,408-summary analyses. No uncertainty intervals are assigned.

Dimension	Judge / prompt	Mean signed	Mean absolute	Pairs
Coherence	GPT-5.2 / P1	-0.003	0.117	1418
Coherence	GPT-5.2 / P2	+0.037	0.128	1417
Coherence	Sonnet 4.6 / P1	-0.048	0.099	1420
Coherence	Sonnet 4.6 / P2	-0.051	0.105	1419
Consistency	GPT-5.2 / P1	-0.004	0.075	1418
Consistency	GPT-5.2 / P2	+0.023	0.078	1417
Consistency	Sonnet 4.6 / P1	-0.010	0.076	1420
Consistency	Sonnet 4.6 / P2	-0.024	0.080	1419
Fluency	GPT-5.2 / P1	-0.010	0.089	1418
Fluency	GPT-5.2 / P2	+0.012	0.077	1417
Fluency	Sonnet 4.6 / P1	-0.006	0.098	1420
Fluency	Sonnet 4.6 / P2	+0.023	0.110	1419
Relevance	GPT-5.2 / P1	+0.012	0.153	1418
Relevance	GPT-5.2 / P2	+0.027	0.159	1417
Relevance	Sonnet 4.6 / P1	-0.053	0.092	1420
Relevance	Sonnet 4.6 / P2	-0.032	0.085	1419

Table E5: Within-document judge–expert correlations, computed across the available system summaries of each document. Means condition on a defined correlation; the last column counts documents with defined / undefined values out of 90. These are not pooled correlations.

Dimension	Method	Conditional mean	Defined / undefined
Coherence	GPT-5.2 / P0	0.598	89 / 1
Coherence	GPT-5.2 / P1	0.603	89 / 1
Coherence	GPT-5.2 / P2	0.591	89 / 1
Coherence	Sonnet 4.6 / P0	0.614	89 / 1
Coherence	Sonnet 4.6 / P1	0.602	89 / 1
Coherence	Sonnet 4.6 / P2	0.598	89 / 1
Consistency	GPT-5.2 / P0	0.690	84 / 6
Consistency	GPT-5.2 / P1	0.682	84 / 6
Consistency	GPT-5.2 / P2	0.698	84 / 6
Consistency	Sonnet 4.6 / P0	0.590	84 / 6
Consistency	Sonnet 4.6 / P1	0.590	84 / 6
Consistency	Sonnet 4.6 / P2	0.586	84 / 6
Fluency	GPT-5.2 / P0	0.505	85 / 5
Fluency	GPT-5.2 / P1	0.521	85 / 5
Fluency	GPT-5.2 / P2	0.513	85 / 5
Fluency	Sonnet 4.6 / P0	0.417	87 / 3
Fluency	Sonnet 4.6 / P1	0.423	87 / 3
Fluency	Sonnet 4.6 / P2	0.421	87 / 3
Relevance	GPT-5.2 / P0	0.433	88 / 2
Relevance	GPT-5.2 / P1	0.447	89 / 1
Relevance	GPT-5.2 / P2	0.444	88 / 2
Relevance	Sonnet 4.6 / P0	0.415	89 / 1
Relevance	Sonnet 4.6 / P1	0.423	89 / 1
Relevance	Sonnet 4.6 / P2	0.418	89 / 1

E.6 NATURAL-SUMMARY REPEAT STUDY

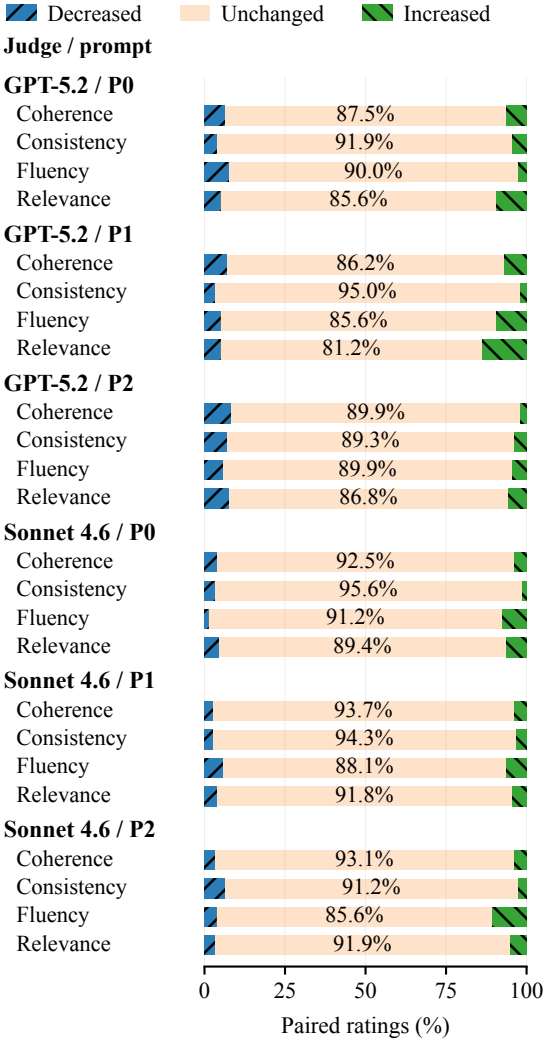
The repeat subset comprises the first ten documents in the stored evaluation order, selected independently of first-pass scores. Their sixteen system summaries give 160 summary identities. Both judges and all three prompts are repeated, giving 960 requests. All repeat ratings are usable; two first-pass vectors remain unavailable, leaving 958 paired four-dimensional vectors or 3,832 dimension-level pairs. GPT/P2 and Sonnet/P1 each retain 159 pairs per dimension; the remaining conditions retain 160. Every summary contributes at least one pair, and 158 summaries contribute all six condition pairs. The repeat analysis does not impose a new 158-summary complete-case mask.

Start-to-start lags use the selected usable requests for the same document, system, judge, and prompt. The minimum, median, and maximum are 70.305, 73.560, and 85.641 hours. Repeat requests preserve the original payload bytes and recorded model/provider identities. This measures recurrence of the scoring procedure, not a verified unchanged hidden serving snapshot.

Figure E3 reports all 24 judge–prompt–dimension cells, including decreases, unchanged scores, increases, and signed and absolute mean changes. Signed change is repeat minus first-pass rating. Direction counts and numerical magnitudes remain separate: the natural-summary changes are not assumed to be one category. The saved numerical outputs retain all twenty-five transition counts and both score margins for each cell. No additional resampling or significance tests were performed for this repeat study. These 70–86-hour repeats remain distinct from the within-session controlled repeats in Appendix F.

Natural-summary rating recurrence

Identical-input repeats; ten documents and sixteen systems



Judge / prompt		*n*	Down / same / up	Mean *d*	Mean |*d*|
GPT-5.2 / P0					
Coherence	87.5%	160	10 / 140 / 10	+0.000	0.125
Consistency	91.9%	160	6 / 147 / 7	+0.000	0.087
Fluency	90.0%	160	12 / 144 / 4	−0.050	0.100
Relevance	85.6%	160	8 / 137 / 15	+0.044	0.144
GPT-5.2 / P1					
Coherence	86.2%	160	11 / 138 / 11	+0.000	0.138
Consistency	95.0%	160	5 / 152 / 3	−0.006	0.056
Fluency	85.6%	160	8 / 137 / 15	+0.044	0.144
Relevance	81.2%	160	8 / 130 / 22	+0.087	0.188
GPT-5.2 / P2					
Coherence	89.9%	159	13 / 143 / 3	−0.063	0.101
Consistency	89.3%	159	11 / 142 / 6	−0.038	0.113
Fluency	89.9%	159	9 / 143 / 7	−0.013	0.101
Relevance	86.8%	159	12 / 138 / 9	−0.019	0.132
Sonnet 4.6 / P0					
Coherence	92.5%	160	6 / 148 / 6	+0.000	0.075
Consistency	95.6%	160	5 / 153 / 2	−0.019	0.044
Fluency	91.2%	160	2 / 146 / 12	+0.062	0.087
Relevance	89.4%	160	7 / 143 / 10	+0.019	0.106
Sonnet 4.6 / P1					
Coherence	93.7%	159	4 / 149 / 6	+0.013	0.063
Consistency	94.3%	159	4 / 150 / 5	+0.006	0.057
Fluency	88.1%	159	9 / 140 / 10	+0.006	0.119
Relevance	91.8%	159	6 / 146 / 7	+0.006	0.082
Sonnet 4.6 / P2					
Coherence	93.1%	160	5 / 149 / 6	+0.006	0.069
Consistency	91.2%	160	10 / 146 / 4	−0.037	0.087
Fluency	85.6%	160	6 / 137 / 17	+0.069	0.144
Relevance	91.9%	160	5 / 147 / 8	+0.019	0.081

P0: direct prompt; P1/P2: opening-instruction rephrasings.

Figure E3: **Natural-summary rating recurrence.** All 24 judge–prompt–dimension cells on condition-specific first-pass/repeat pairs. Bars show decreased, unchanged, and increased ratings; labels give exact category agreement. GPT/P2 and Sonnet/P1 retain 159 pairs per dimension; other conditions retain 160. Numerical columns give denominators, direction counts, and mean signed and absolute changes, where d is repeat minus first-pass rating in points on the recorded 1–5 scale. Direction counts do not encode change magnitude. Results are descriptive, with no uncertainty intervals; start-to-start repeat lags span 70.3–85.6 hours.

F CONTROLLED FACTUAL-CONSISTENCY EVALUATION

F.1 MATERIALS AND INDEPENDENT REVIEW

The controlled set draws on the previously analyzed news pool, excluding the ten calibration/development documents and the quarantined document. Forty-eight candidate groups were prepared for review; forty-five entered model scoring. Each retained group contains an unchanged source article D_1, an edited counterpart D_2, and four summaries. A and A' express one factual content in two forms, while B and B' express the alternative content in corresponding forms. D_1 supports the target content in A/A' and conflicts with that in B/B'; D_2 reverses these relations. All four full summaries are reused unchanged under both articles.

Draft edits were prepared with AI assistance and checked by two independent reviewers. Each reviewer examined the full two-article/four-summary set, identified supporting or conflicting passages, compared the two expression pairs, and recorded additional meaning changes, internal inconsistencies, and the importance of the changed fact. Responses could indicate insufficient information or uncertainty. Reviewers were instructed to use the supplied article rather than outside knowledge of the event. One reviewer used assisted transcription and formatting; that reviewer formed and confirmed the judgments.

The initial review covers all 48 candidates, not only the eventual retained sample. Reviewers selected the same support-relation category in 384/384 corresponding judgments, the same expression category in 92/96 pairs, and the same importance category in 42/48 groups. These descriptive agreement counts precede all formal revisions and exclusions. Support-label agreement did not imply that every candidate was suitable: comments identified changes in degree, temporal scope, attribution, and discourse relations. Seven retained groups were revised and their affected content rechecked before scoring; unchanged content retains its earlier review rather than being counted as newly judged.

Three candidates were excluded before model scoring: one changed the precision of a quotation along with the target quantity, one inherited unresolved scope ambiguity about fixed cameras, and one introduced a narrative linkage change when the current-event participant was replaced. These exclusions are editorial decisions informed by review, not three unanimous exclusion votes. Revisions to retained groups addressed tense symmetry, degree of increase, reporting force, transaction stage, the identity of a video maker, and the temporal scope of a negated event. No candidate was removed on the basis of its model rating.

The final set contains 31 quantity/time, ten relation/role, and four entity/participant contrasts. It is purposive and sports-heavy. Some target changes require dependent edits, such as a date with its duration or a component score with its total. These are documented together rather than described as single independent word substitutions. Material review establishes local support and expression relations, not numerical expert targets for the new summaries or representativeness of all factual errors.

F.2 SCORING AND PAIRED ESTIMANDS

Both judges use the same direct prompt and four-dimensional rubric as the natural-summary study (Appendix E). The two articles crossed with four summaries give eight inputs per group and 360 distinct article–summary inputs overall. Each judge scores every input twice, yielding 720 usable joint four-dimensional ratings per judge and 1,440 across both judges. Consistency is the primary dimension. The first pass supplies the primary comparisons; the second pass is reported separately and supplies identical-input variation. Ratings are not averaged across passes or judges. The sample and analysis choices were finalized by the authors before model scoring; no preregistration claim is made.

Factual and source-change contrasts. For a fixed judge and dimension, let $y_r(D, S)$ denote the rating of summary S against article D in pass $r \in \{0, 1\}$. Here, $r = 0$ is the first scoring pass and $r = 1$ the second; the figure's display labels are pass 1 and pass 2, respectively. Within a source

group, the primary factual and source-change estimands are

$$F_1 = \tfrac{1}{2}\{y_0(D_1, A) - y_0(D_1, B) + y_0(D_1, A') - y_0(D_1, B')\}, \tag{F1}$$

$$L = \tfrac{1}{2}\{y_0(D_1, A) - y_0(D_2, A) + y_0(D_1, A') - y_0(D_2, A')\}, \tag{F2}$$

$$G = \tfrac{1}{2}\{y_0(D_2, B) - y_0(D_1, B) + y_0(D_2, B') - y_0(D_1, B')\}. \tag{F3}$$

Second-pass counterparts use y_1 in place of y_0. Positive F_1 favors supported content within D_1; positive L records a decrease when the A content loses support, and positive G an increase when the B content gains support. The secondary F_2 uses the same supported-minus-conflicting comparison under D_2, reversing the A/B order. The subscripts in F_1 and F_2 identify articles, not passes. The source–content interaction satisfies $F_1 + F_2 = L + G$; it does not replace reporting both source directions. The indicator that both source-direction means are positive requires $L > 0$ and $G > 0$. The indicator that all four source changes are positive instead requires each of the two form-specific losses and two form-specific gains to be positive. These indicators remain distinct from the check of within-source A/B ordering under both articles.

Expression and repeat variation. Within each pass, expression change under D_1, denoted E_1 in Figure F1, is the mean of the two absolute differences between primed and unprimed summaries. The primary expression comparison uses the first pass. Its matched repeat baseline is the mean absolute difference between passes for the four D_1 inputs. Expression minus repeat is calculated within each source group. Signed expression endpoints average $y_r(D, A') - y_r(D, A)$ and $y_r(D, B') - y_r(D, B)$ within each article and pass; signed repeat change is $y_1(D, S) - y_0(D, S)$. Repeatability over all eight inputs also reports signed change, absolute change, and exact category agreement. The first-pass expression measure shares scores with the repeat baseline, so they are not independent measurements. Tied and contrary form-specific changes are retained alongside group means.

Aggregation and uncertainty. A group endpoint requires every constituent pair to be available; aggregate results weight complete source groups equally. All 45 groups are complete in this study. The 10,000 bootstrap draws resample 45 source groups with replacement, carrying both judges, all dimensions, all article–summary combinations, and both passes together. Pointwise 95% intervals use the 2.5th and 97.5th percentiles of these shared draws; paired differences are computed within draws. These conditional group-resampling intervals do not imply representative sampling. The saved numerical outputs retain full-precision group values, direction counts, all condition transitions, and the recorded interval estimates.

F.3 Complete consistency results and second pass

Table F1 includes the first-pass factual and expression comparisons, their signed and absolute counterparts, both source-direction indicators, and the cross-pass repeat summaries. Table F2 applies the same consistency analyses to the second pass on the same materials. The second-pass results are not averaged into the first-pass estimates. The relative-ordering check compares supported and conflicting content under both D_1 and D_2, retaining the same exception described below.

Table F1: Complete first-pass consistency and cross-pass repeat endpoints. Each column averages the same 45 source groups. Brackets are pointwise 95% source-group bootstrap intervals. The two positive-direction indicators and exact repeat agreement are fractions; all other entries are rating points.

Endpoint	GPT-5.2	Sonnet 4.6
D_1: supported − conflicting	2.011 [1.811, 2.211]	2.356 [2.156, 2.556]
A/A' loses support	1.944 [1.756, 2.133]	2.356 [2.156, 2.567]
B/B' gains support	1.989 [1.789, 2.189]	2.167 [1.944, 2.400]
D_2: supported − conflicting	1.922 [1.733, 2.111]	2.167 [1.933, 2.400]
D_1 expression: signed	0.056 [-0.011, 0.122]	-0.044 [-0.111, 0.022]
D_1 expression: absolute	0.122 [0.067, 0.189]	0.111 [0.056, 0.178]
D_2 expression: signed	0.033 [-0.044, 0.111]	-0.033 [-0.111, 0.044]
D_2 expression: absolute	0.144 [0.078, 0.211]	0.122 [0.056, 0.200]
Source–content interaction	3.933 [3.567, 4.289]	4.522 [4.122, 4.922]
Both source-direction means positive	1.000 [1.000, 1.000]	0.978 [0.933, 1.000]
All four source changes positive	1.000 [1.000, 1.000]	0.978 [0.933, 1.000]
Repeat: signed (all inputs)	-0.022 [-0.053, 0.011]	-0.008 [-0.028, 0.011]
Repeat: absolute (all inputs)	0.117 [0.083, 0.150]	0.053 [0.025, 0.086]
Repeat: exact agreement	0.883 [0.850, 0.917]	0.947 [0.914, 0.975]
Repeat: absolute (D_1)	0.078 [0.039, 0.117]	0.056 [0.017, 0.106]
Repeat: absolute (D_2)	0.156 [0.111, 0.200]	0.050 [0.017, 0.089]
Expression − repeat (D_1)	0.044 [-0.011, 0.100]	0.056 [0.000, 0.117]

Table F2: Second-pass consistency results on the same 45 source groups, reported separately rather than averaged into the first pass. Brackets are pointwise 95% source-group bootstrap intervals. Positive-direction indicators are fractions; other entries are rating points.

Endpoint	GPT-5.2	Sonnet 4.6
D_1: supported − conflicting	2.011 [1.811, 2.211]	2.378 [2.178, 2.578]
A/A' loses support	2.033 [1.844, 2.222]	2.356 [2.144, 2.567]
B/B' gains support	1.989 [1.800, 2.178]	2.222 [2.000, 2.444]
D_2: supported − conflicting	2.011 [1.822, 2.200]	2.200 [1.967, 2.433]
D_1 expression: signed	0.011 [-0.044, 0.067]	-0.044 [-0.111, 0.022]
D_1 expression: absolute	0.078 [0.033, 0.133]	0.133 [0.067, 0.211]
D_2 expression: signed	0.011 [-0.089, 0.111]	-0.022 [-0.111, 0.067]
D_2 expression: absolute	0.233 [0.156, 0.311]	0.156 [0.089, 0.233]
Source–content interaction	4.022 [3.667, 4.378]	4.578 [4.178, 4.978]
Both source-direction means positive	1.000 [1.000, 1.000]	0.978 [0.933, 1.000]
All four source changes positive	1.000 [1.000, 1.000]	0.978 [0.933, 1.000]

F.4 Additional dimensions and the source-reversal exception

Figure F1 presents all four jointly collected dimensions, with the two scoring passes shown separately. For the first-pass D_1 factual contrast, GPT's coherence and fluency differences are 0.044 and 0.000 points, compared with 0.600 and 0.311 for Sonnet. The corresponding consistency differences are 2.011 and 2.356. This suggests different degrees of coupling among dimensions in the two rating procedures; it is a secondary observation, not a diagnosis of a particular bias mechanism. The saved numerical outputs also include signed expression changes, all-dimensional repeat endpoints, and their available intervals.

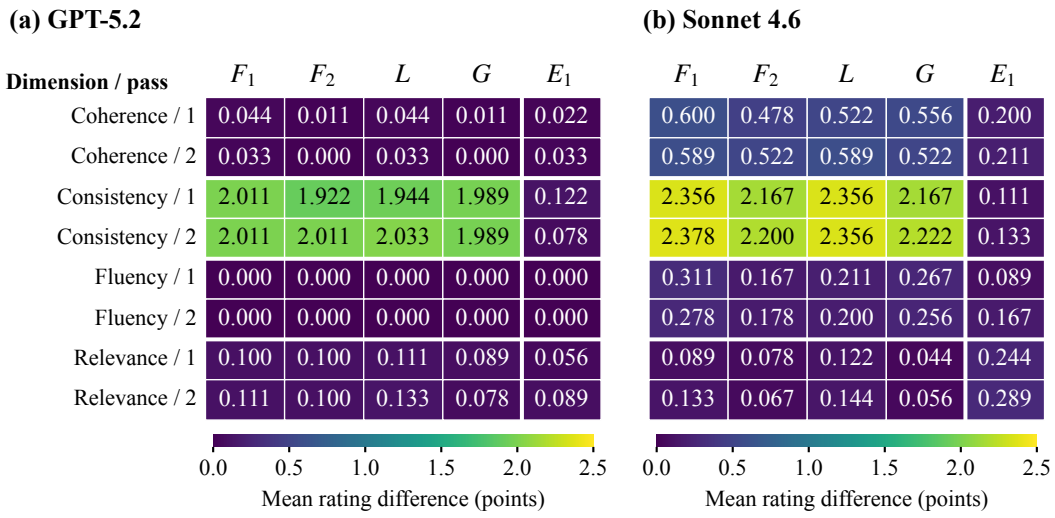


Figure F1: **Controlled contrasts across rating dimensions. (a)** GPT-5.2; **(b)** Sonnet 4.6. Each cell is an equally weighted mean over 45 source groups, in rating points. Aligned rows share the dimension/pass labels; 1 and 2 denote separate scoring passes. F_1 and F_2 are supported-minus-conflicting contrasts under D_1 and D_2; positive L denotes a rating decrease after support removal, and positive G an increase after support is gained. E_1 is absolute expression change under D_1, not repeat change. Both panels use the same linear color scale and show means only; intervals remain in the saved numerical outputs. Consistency is primary; other dimensions are secondary.

For the sole Sonnet relative-ordering exception (material F15), the first pass assigns consistency scores of 5 and 4 to A and A' under D_1, and 2 to each of B and B'. Under D_2, all four scores are 2. The second pass reproduces these consistency scores. Thus, the A content loses points after losing support, while the B content does not gain points after gaining support; both within-D_2 comparisons tie. The group is retained in the analyses. Its stable ratings illustrate why recurrence and responsiveness are different properties, without reversing the positive directional findings for the remaining groups.

F.5 Timing and computational record

All 720 first-pass requests complete before the 720 second-pass requests. Each pass follows a fixed request-ID hash order, without a fixed delay between requests or a day-long repeat requirement. Collection runs from 14:21:16 to 15:23:29 UTC on 19 September 2026. Across 720 same-input, same-judge repeat pairs, start-to-start lags range from 0.840 to 59.181 minutes, with a median of 31.693 minutes. Judge-specific medians are 32.094 minutes for GPT and 31.168 minutes for Sonnet. No retries or missing scores occur; the reported cost is USD 4.9213. These are within-session repeats, distinct from the 70–86-hour natural-summary repeats in Appendix E.

G Benchmark Validity Cards and Evidence Synthesis

G.1 From analysis outputs to a Card

The six Cards below synthesize the results in Sections 4.2–4.4; they are authored interpretations of this audit, not quotations sampled from leaderboard reports. Four concern the multiple-choice benchmarks, one concerns numeric-answer evaluation, and one concerns the factual-consistency interpretation of summary ratings. Each retains the same four fields as the main-text example: *interpretation*, *supporting evidence*, *scope and conditions*, and *evidence gaps*.

Specify the comparison. Record the capability or quality being interpreted, the models or scoring procedures being compared, and the intended unit of use. A claim about a six-model ranking

differs from a claim about performance within a construct group. For summary ratings, ordering individual summaries, ordering system means, and agreement on the numerical scale remain distinct interpretations.

Compile the evidence. Link the interpretation to the relevant construct distributions, protocol scores, paired outcomes, and scoring analyses. Each result retains its sample, denominator, and uncertainty statement. The main audit's fixed-grid accuracy uses 120 scheduled items per cell; paired transitions use items with usable responses under both compared conditions. The natural-summary analyses retain their separate 1,408-summary and 1,200-summary populations. Controlled contrasts retain source groups and both expression forms. These observations are not pooled into a single effective sample size.

Write the warrant. Explain why each finding bears on the proposed interpretation. A lead that persists across presentations supports comparative task performance; a source-support reversal tests whether a consistency rating follows the article–summary relation. A construct profile identifies the annotated demands represented in a score. The synthesis preserves both supporting findings and relevant divergences, such as stable ranking alongside changing absolute accuracy, without averaging them into a validity score.

State the scope and next evidence. Name the conditions over which each result was observed and the assumptions required for the interpretation. Evidence gaps identify extensions that need further investigation, rather than automatically reversing a supported conclusion. Saved analysis outputs support numerical compilation; the interpretation and rationale remain authored judgments whose supporting results are inspectable. The earlier rule-based categories in Section G.4 are not applied to the current Cards.

G.2 Cards for the four-benchmark audit

These Cards use the corrected fixed-grid scores and paired analyses in Appendix C. Unless a model is named, a maximum score range is the largest within-model range over five protocols for that benchmark. Rank summaries retain the six-model panel and the competition-tie convention: rank is one plus the number of models with a strictly higher correct count. Demand counts are adjudicated primary labels, not observations of the models' internal reasoning.

Table G1: ARC-Challenge card, based on Sections 4.2.2–4.2.3 and Appendix C.

ARC-Challenge: science-question performance	
Interpretation	Scores provide evidence of performance on scientific- and causal-reasoning questions in the audited answer-choice format. Modest aggregate variation does not ensure stable ordering of nearby models.
Supporting evidence	Of 120 items, 82 have scientific reasoning, 36 causal reasoning, and two subject knowledge as their primary label. Every model's five-protocol score range is at most 2.50 percentage points; median paired turnover is 0.84%. GPT-5.2 has a mean absolute rank shift of 1.50 places despite a score range of 1.67 percentage points. Its shuffled-option comparison preserves 115/120 correct with no correctness transitions.
Scope and conditions	The saved 120-item validation sample, six recorded model pipelines, supplied answer references, and five presentations. Panel scores use the scheduled denominator; turnover uses each comparison's both-usable items.
Evidence gaps	Aggregate-score stability alone does not establish broad scientific expertise. Transfer to experimental or open-ended scientific tasks and the contribution of particular solution strategies require additional task- and response-level evidence.

Table G2: HellaSwag card. The equal-score example retains its own 120-item paired support and does not describe every model or protocol (Figure C2).

HellaSwag: event-continuation performance	
Interpretation	The audited scores summarize success in choosing plausible event continuations under specified presentations. The common primary demand gives the total a focused content interpretation, while paired outcomes qualify an interpretation of unchanged item correctness.
Supporting evidence	All 120 primary labels are event continuation. The largest five-protocol score range is 6.67 percentage points, and median paired turnover is 7.12%. GPT-5.2 obtains 103/120 under both original and plain instructions, but nine gains and nine losses produce 15.0% turnover (pointwise 95% interval $[9.17, 21.67]$). DeepSeek V4-Pro moves by up to three rank places.
Scope and conditions	The 120-item validation sample, six recorded pipelines, supplied answer-choice references, and tested instruction, order, and wording variants. A primary label describes the dominant annotated demand, not a uniquely observed model strategy.
Evidence gaps	A concentrated label distribution does not make the item set exhaustive of commonsense competence. Other discourse genres, practical reasoning tasks, and evidence about how solutions are reached would extend that interpretation.

Table G3: MMLU card. Sample composition is part of the interpretation, not a benchmark-wide population claim.

MMLU: a sampled academic-task mixture	
Interpretation	The score summarizes performance on the sampled mixture of academic demands. Its interpretation should retain that mixture and distinguish a numerical advantage from a protocol-stable ranking.
Supporting evidence	Primary labels comprise 83 subject-knowledge, 29 cultural/social-knowledge, and eight mathematical-reasoning items. The largest within-model score range is 3.33 percentage points. Median paired absolute net change is 1.25 percentage points and median turnover is 3.39%; the largest model-level mean absolute rank shift is 1.50 places.
Scope and conditions	Thirty items from each of four selected subject groups, not a sample covering every MMLU subject; six recorded pipelines and five presentations. The equal group allocation and supplied references define the reported aggregate.
Evidence gaps	General academic expertise requires coverage and transfer evidence beyond these subject groups and the answer-choice format. Construct heterogeneity is not itself invalidity: its adequacy depends on whether the intended academic interpretation calls for this mixture.

Table G4: Expanded GPQA card corresponding to main-text Table 2. The two accuracy denominators remain distinct; the simultaneous band concerns the stated baseline contrast.

GPQA: scientific task performance and a sustained model lead	
Interpretation	GPT-5.2 has a sustained comparative advantage in the audited expert-level science task mixture within the six-model panel. This lead across tested presentations remains distinct from stable success within each demand group.
Supporting evidence	GPT-5.2 ranks first in every protocol; its baseline is 99/120 and its own score range is 5.00 percentage points. Its baseline lead over Gemini is 23.33 percentage points, with an approximate 95% within-benchmark simultaneous contrast band of $[6.67, 40.00]$. In the 119-item original/plain paired subset, both totals are 99 correct: mathematical reasoning moves from 44/48 to 46/48 and subject knowledge from 42/56 to 40/56.
Scope and conditions	The 120-item Diamond sample comprises 56 subject-knowledge, 48 mathematical, 15 scientific, and one formal-logical primary label; six recorded pipelines and five presentations. Fixed-grid scores use 120 scheduled items, whereas the paired example uses 119 both-usable items. Annotation agreement and alternative label views are documented in Appendices A and C.
Evidence gaps	Transfer beyond the tested science-task mixture or to open-ended scientific problem solving remains untested. Item-demand annotations do not establish a unique internal solution process.

G.3 Cards for numeric answers and summary ratings

The extensions change the score-generating procedure and therefore the evidence required for interpretation. Exact numeric matching supplies an operational scoring reference for GSM8K. Summary-quality ratings require dimension-specific comparisons with expert judgments and, for factual consistency, controlled evidence about source support. The following Cards preserve these different scoring relations and their separate analysis populations.

Table G5: Numeric-answer card, based on Section 4.3 and Appendix D.

GSM8K: numeric-answer performance	
Interpretation	Scores support high numeric-answer accuracy on the audited mathematical word problems, with modest absolute changes across three tested presentations. Comparative rankings and item-level preservation remain separate properties.
Supporting evidence	All 2,160 recorded outputs are usable. Across 18 model–condition cells, accuracy ranges from 94.2% to 99.2%, and the largest within-model score range is 2.50 percentage points. DeepSeek V3.2 retains 116/120 under original and formal-exam instructions through one gain and one loss; GPT-5.2 retains 119/120 with no correctness transitions in that comparison.
Scope and conditions	The 120-item sample; original, formal-exam, and paraphrased presentations; six recorded pipelines, including DeepSeek V3.2 rather than V4-Pro. Correctness is normalized exact numeric agreement with the reference. No bootstrap intervals accompany this extension.
Evidence gaps	Correct final numbers do not alone characterize solution quality or broader mathematical reasoning. Those interpretations call for evidence on solutions and additional problem types. The extension does not substitute for validation of semantic quality metrics.

Table G6: Factual-consistency card, based on Section 4.4 and Appendices E–F. The natural-summary repeat study covers 24 judge–prompt–dimension cells (81.2%–95.6% exact agreement over 70.3–85.6 hours); controlled repeats are within-session comparisons. They are not pooled into a common repeatability estimate.

SummEval: factual-consistency ratings	
Interpretation	The two direct-prompt judges provide source-sensitive factual-consistency ratings for the evaluated news summaries. The evidence supports this dimension-specific interpretation; the numerical scale and evaluation level remain part of the intended use.
Supporting evidence	On 1,408 natural summaries, GPT-5.2 and Sonnet correlate with the expert consistency reference at 0.652 and 0.543; mean absolute discrepancies are 0.292 and 0.660 rating points. On the matched 1,200 summaries, GPT's correlation changes from 0.6651 at summary level to 0.9425 at system level, and Sonnet's from 0.5540 to 0.9602. These are point-estimate comparisons. Across 45 controlled groups, each judge prefers supported content under the original article in all 90 form-specific comparisons. Both-form ordering reversals hold in 45 GPT groups and 44 Sonnet groups; the remaining Sonnet group ties under the edited article. Factual contrasts exceed the observed expression and repeat variation.
Scope and conditions	Named judges, the shared direct rubric, and a three-expert-mean reference. Observational, matched-aggregation, repeat, and controlled populations remain separate. The purposive edited materials emphasize quantities and news, particularly sports. Natural and controlled studies draw on the same source pool.
Evidence gaps	The findings do not establish interchangeable human and model rating categories or a universal best judge. Other genres, relation types, rubrics, and quality dimensions need corresponding evidence. Human material checks establish contrasts, not numerical score targets.

G.4 Earlier implementation and a classification-repeatability pilot

Implemented evidence compilation. The earlier three-benchmark implementation assembled numerical profiles from score summaries, ranking changes, and measurement records. Its profile script produced 18 model–benchmark profiles: four received the stable-protocol label and fourteen the qualified-benchmark label. A separate report generator combined benchmark-level profiles, construct distributions, and evidence summaries into three Markdown Cards for ARC-Challenge, HellaSwag, and MMLU. General capability descriptions and reporting phrases were supplied as templates. This provided an executable route from result files to structured reports, with natural-language interpretations dependent on those rules and templates. Table G7 records the historical category boundaries; they are not applied to the current Cards.

Table G7: Heuristic cutoffs in the earlier profile implementation. They document how its labels were produced, not established adequacy thresholds for the present cards.

Historical quantity	Implemented category boundaries
Within-model score range	Low: ≤ 0.03; moderate: $(0.03, 0.08]$; high: > 0.08. Values are accuracy fractions.
Mean absolute rank shift	Low: ≤ 0.25; moderate: $(0.25, 1.0]$; high: > 1.0. Values are rank places.
Maximum problem rate	Very low: ≤ 0.01; acceptable: $(0.01, 0.05]$; concerning: > 0.05. Values are fractions of records.

At the model–benchmark level, the stable-protocol label required low score range, low rank shift, and either very-low or acceptable problem rate. The qualified-benchmark label applied to the remaining cases with low or moderate range and one of those two problem-rate categories; other cases required strong qualification. This model-level branch did not use each item's construct label; benchmark-level reporting added construct information. The present synthesis uses the full

construct, paired, and scoring evidence rather than transferring those categorical rules to the four-benchmark and summarization studies.

Classification repeatability. A separate pilot during the earlier revision selected twelve evidence profiles from ARC-Challenge, HellaSwag, and MMLU. Each packet specified benchmark and model identifiers, baseline accuracy, score range, mean absolute rank shift, and measurement-problem rate. Two annotators selected a category and wrote a short rationale. The packet-construction script removed the original qualification and recommended-claim columns. Instructions asked for the strongest supported interpretation and advised choosing the more qualified interpretation when uncertain.

Table G8: Earlier profile-classification pilot. Counts describe twelve selected profiles, not the 18-profile implementation output or the 45-group controlled-material review.

Pilot property	Recorded result
Evidence profiles / annotators	12 / 2
Stable-protocol category	4
Qualified-benchmark category	8
Overextended category	0
Between-annotator agreement	12/12
Cohen's κ	1.000
Agreement with the original rule categories	12/12

The completed annotations agreed on all twelve categories and matched the original rule categories (Table G8). This supports repeatability of a constrained profile-classification task. The packets did not contain a case-specific capability proposition or an item-level construct profile, and the sample contained no GPQA or overextended-category cases. The result therefore does not test the correctness of the present Cards' full interpretations or whether Cards improve readers' decisions relative to another reporting format. The current Cards synthesize the reported evidence; they are not outputs validated by that pilot. The two-reviewer material checks in Appendix F address a different question and are not combined with these annotations.