

How Much Should the Agent See? A Disclosure Axis for LLM Data-Analyst Agents

Athira M V¹ Aditya Challa¹ Jaya Tripathi² Saroj Mohanty²

Santosh Kumar Choudhary² Sravan Danda¹ Ruchika Mehra Jain² Snehanshu Saha^{1,3}

¹APPCAIR and CS&IS, BITS Pilani K K Birla Goa Campus, Goa, India

²MAMTA Health Institute, New Delhi, India

³School of AI and CSE, Mahindra University, Hyderabad, India

{p20250027, adityac, dandas, snehanshus}@goa.bits-pilani.ac.in

{jaya, sarojm, santoshchoudhary, ruchika.mehra}@mamtahimc.in

Abstract

Existing LLM data-analyst benchmarks fix data access at full disclosure and evaluate only on answer accuracy. But real deployments rarely grant full access. No prior work, to our knowledge, asks “what’s the least amount of information needed for the agent to work?”. In this article, we treat dataset disclosure as a controlled treatment along a *graduated information-disclosure* (GID) axis: C_0 (column names) and C_∞ (full table) bracket a nested chain $C_1 \subseteq \dots \subseteq C_5$ that adds per-column metadata, dataset purpose, summary statistics, sampled cells, and sampled rows. We maintain agent architecture, tool registry, and decoding fixed, and score the final reports on a 14-attribute rubric across a $6 \times 4 \times 12$ panel of frontier models, planted-signal datasets, and disclosure levels. At a high level, from C_0 to C_∞ , seven attributes rise and five fall. However, the rise/fall pattern has nuance, the optimal intermediate level depends on several factors including models, datasets and attributes. Level C_1 - one sentence of types and units beyond column names - matches (statistically) C_∞ on signal recovery and decoy rejection at $\sim 49\times$ lower prompt-token cost. The key takeaway from this work is - **current LLM data-science agents can act on partial information alone**, so row-level records need never leave the institution. Hence, the current framework also offers a way to enforce data boundary constraints (required by GDPR) without compromising on quality.

1 Introduction

LLM data-analyst agents are evaluated at one operating point - full data access - and on one axis - answer accuracy. Existing benchmarks (Majumder et al., 2025; Sahu et al., 2025; Lai et al., 2026; Guo et al., 2024; Gu et al., 2024) fix access at full disclosure, and tool-augmented agents (Shen et al., 2025) add retrieval rather than ask what happens with less.

Real deployments rarely see the full table: institutional, contractual, or regulatory limits restrict what data leaves the local environment. We ask: **how does an LLM analyst’s hypothesis quality change when its view of the data is varied along a controlled disclosure axis?** To our knowledge, no prior study combines a graduated access axis, a multi-attribute rubric, and a controlled substrate that isolates disclosure from real-world data-quality confounds (Appendix A).

We close this gap with a graduated information-disclosure (GID) axis (Figure 1). The endpoints are C_0 (column names only) and C_∞ (full table), connected by a nested chain $C_1 \subseteq C_2 \subseteq \dots \subseteq C_5$ that adds, in order, per-column metadata, dataset purpose, summary statistics, sampled cells, and sampled rows (Tables 1 and 2). Agent architecture, tool registry, turn budget, and decoding are held fixed; only the dataset card varies. Each (model, dataset, level) cell is scored by a 14-attribute rubric (Section 2). The study spans six models, four datasets, twelve disclosure levels ($C_0, C_1-C_5, C_\infty, A_1-A_5$), and turn/seed sweeps - ~ 460 runs in total.

Our main finding is: Adding column types and units to the prompt (disclosure level C_1) matches full-table access (C_∞) on signal recovery and decoy rejection, at about $49\times$ lower prompt-token cost (~ 428 vs. $\sim 21k$ tokens; Section 3.2). At a high level, from C_0 to C_∞ , seven attributes rise - signal recovery and decoy rejection rise most - and five fall, with column coverage and grounded analysis falling most (Figure 2). Section 3 reveals further nuance within this pattern: (i) smaller models stop gaining past C_1 , while larger models keep gaining through C_3 (Sections 3.1 and 3.3). (ii) Datasets follow the same pattern - signals visible from column-level descriptors stop benefiting past C_1 ; signals hidden in conditional or subgroup structure keep benefiting through C_3 . (iii) The key takeaway is - **current models can act on partial information alone**, so an operator can keep row-level records

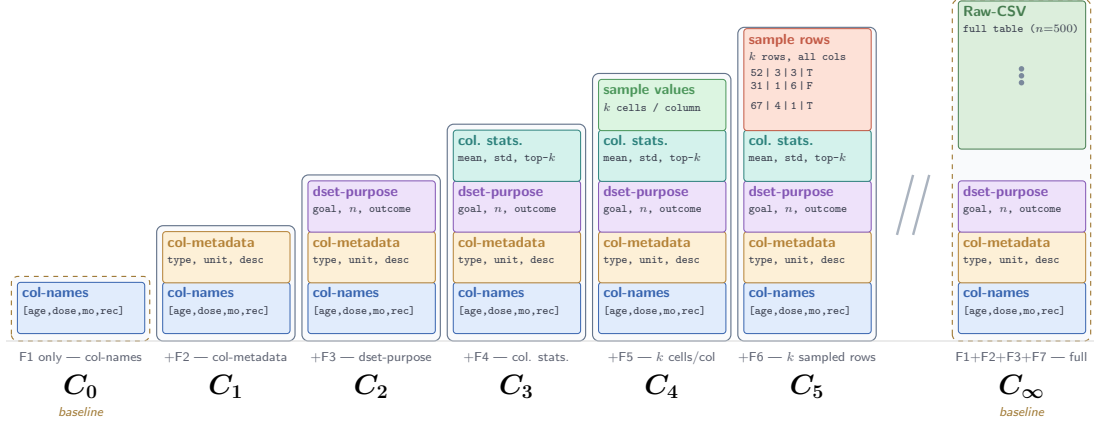


Figure 1: The graduated information-disclosure (GID) axis: C_0 (column names) and C_∞ (full disclosure) anchor a monotone-nested chain $C_1 \subseteq \dots \subseteq C_5$ that progressively adds metadata, purpose, statistics, cells, and rows. Architecture and decoding are held fixed; only the dataset card varies.

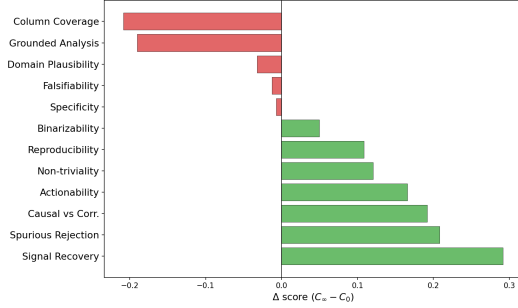


Figure 2: **Disclosure is not uniform across attributes.** Mean per-attribute change from C_0 to C_∞ across the model×dataset panel. Seven attributes rise (signal recovery and decoy rejection lead); five fall (column coverage and grounded analysis fall most). (See Section 3).

inside the institution and still obtain quality close to the full-table ceiling. The GID-conditioner is where the enforcement happens - it filters the context to exactly what a disclosure policy permits. This is data minimisation by construction, and it addresses the data-boundary part of regimes such as GDPR.

Contributions. (i) We introduce the GID axis - C_0 floor, C_1 – C_5 nested chain, C_∞ ceiling - paired with four planted-signal synthetic datasets with ground-truth signal/decoy attribution and a 14-attribute deterministic rubric (Section 2). (ii) We show that attributes do not move together with disclosure: seven rise and five fall (Section 3.1, Figure 2). (iii) We find a cheap level of disclosure that matches the full table - C_1 matches C_∞ on signal recovery and decoy rejection at $\sim 49\times$ lower prompt-token cost, with the $C_0 \rightarrow C_1$ step - adding types and units - giving the largest gain in decoy

rejection; C_3 is closest to full disclosure on average (Section 3.2). (iv) We identify which models benefit from extra disclosure and which statistical signals it helps to recover (Section 3.3).

2 Methodology

We isolate disclosure as the only treatment variable. Three components are held constant: (i) the agent; (ii) datasets with planted signals and decoys exposed through a graduated conditioning interface; (iii) a 14-attribute evaluation rubric with reproducible scoring. The setup originated in a project with a healthcare-sector partner whose data-usage agreement prevents direct reporting (Appendix G); a sanity check on two real-world datasets (Appendix H) verifies that the synthetic datasets cover the attribute profile of real-datasets.

Agent Design: The agent is a fixed data-analyst loop with three integer parameters: turns T , hypotheses per turn H , and tool calls per hypothesis K . Tool registry, structured-output schema, and system prompt are held constant across disclosure conditions. A *GID-conditioner* builds the context prompt once per run from the raw dataset and a disclosure level C (Section 2); disclosure is a property of the run, not of any turn or tool call.

Each turn has three phases. (a) The agent emits 5–10 candidate hypotheses through one structured-output call - the only free LLM generation in the loop. (b) Each hypothesis is dispatched concurrently to up to K tools from a 25-tool statistical registry. (c) Outputs are assembled into a Markdown blob that becomes the next turn’s user prompt. After turn T , a synthesis call (Phase d) reads only

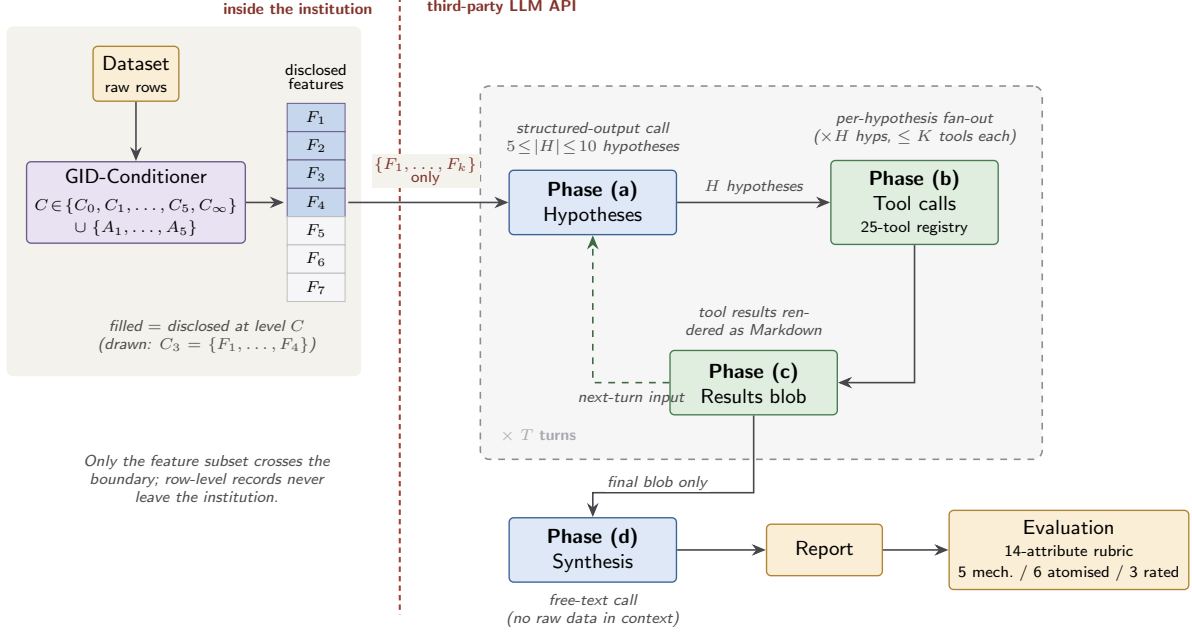


Figure 3: **Data-analysis agent flow.** The GID-conditioner builds the prompt once from the raw table and a disclosure level C , and only the feature subset that C permits (Table 1) is passed onto API. Row-level records need never leave the institution. Phases (a)-(c) repeat for T turns; (d) reads the final results

Feature	Description
F1	Column names.
F2	Per-column metadata: dtype, unit, and short description.
F3	Free-text dataset purpose, methodology, and domain story.
F4	Per-column summary statistics (mean / std / min / max / median / n_{unique} / n_{missing} for numeric; top- k counts for categorical; min / max / span for datetime).
F5	k randomly sampled cells per column, drawn from independently sampled row indices ($k=5$).
F6	k randomly sampled whole rows ($k=5$).
F7	The full table embedded in the prompt as CSV, together with tool-mediated access to a Python sandbox (see Table 4).

Table 1: The seven elementary disclosure features F1–F7. Each of the twelve disclosure conditions presented to the agent (see Table 2) is a subset of $\{F_1, \dots, F_7\}$.

the final blob and emits the report (Figure 3). Tools are implemented via statsmodels (Seabold and Perktold, 2010); see Table 3 and Table 4 for the registry and argument schemas.

Datasets: We construct synthetic datasets with planted signals and documented decoys, because no public tabular corpus jointly offers enumerable ground truth, documented decoys, and constant data quality across disclosure conditions. We extend the Majumder et al. (2025) (*context, variables, relationship*) ground-truth format and the Sahu et al. (2025) planted-signal paradigm with two pieces: documented planted decoys, and the seven-feature disclosure factoring below.

Disclosure conditions			Ablations (A1–A5)	
Cond.	Role	Feature subset	Cond.	Feature subset
C_0	baseline	$\{F_1\}$	A1	$\{F_1, F_2, F_4\}$
C_1	chain	$\{F_1, F_2\}$	A2	$\{F_1, F_3, F_4\}$
C_2	chain	$\{F_1, F_2, F_3\}$	A3	$\{F_1, F_2, F_3, F_5\}$
C_3	chain	$\{F_1, F_2, F_3, F_4\}$	A4	$\{F_1, F_2, F_3, F_6\}$
C_4	chain	$\{F_1, F_2, F_3, F_4, F_5\}$	A5	$\{F_1, F_7\}$
C_5	chain	$\{F_1, F_2, F_3, F_4, F_6\}$		
C_∞	baseline	$\{F_1, F_2, F_3, F_7\}$		

Table 2: The twelve disclosure conditions, each defined as a subset of $\{F_1, \dots, F_7\}$ (see Table 1). C_0 and C_∞ are baseline anchors (names-only and full-table respectively); the studied chain is C_1 through C_5 , designed to be monotone-nested: $C_1 \subseteq C_2 \subseteq \dots \subseteq C_5$. The step $C_5 \rightarrow C_\infty$ is the only non-monotone step: it replaces $F_4 \cup F_6$ (static per-column aggregates and sampled rows) with F_7 (the full table plus tool access), so that the effect of tools is not confounded with the effect of pre-computed statistics. C_∞ is included as a ceiling reference rather than as a rung of the chain.

The battery has four grounding datasets, each with 500 rows and 15 columns: **G1** a negative dose-recovery trend with month confounding (clinical trials); **G2** Simpson’s paradox in admission rates (Simpson, 1951); **G3** an ad-spend $\rightarrow$ page-views $\rightarrow$ conversion mediation chain; **G4** a U-shaped temperature–failure relationship with near-zero Pearson correlation. Together, G1–G4 cover correlational, confounded, mediation, and nonlinear relationships.

Each dataset card factors into seven elementary features $F_1 \dots F_7$ (Table 1); each disclosure condition is a subset of these features (Table 2). C_0 is column names only, C_∞ is the full table, and $C_1 \subseteq C_2 \subseteq \dots \subseteq C_5$ adds features in order. The step $C_5 \rightarrow C_\infty$ replaces the static aggregates and sampled rows ($F_4 \cup F_6$) with the full table plus a Python sandbox (F_7), so C_∞ is a ceiling reference rather than a chain rung. Five ablations $A_1 \dots A_5$ probe individual feature contributions (Table 2).

The unit of analysis is the (model, dataset, disclosure-level) cell. The headline panel is balanced: $6 \times 4 \times 7 = 168$ cells, with a 24-cell (model, dataset) sample per chain rung $C_1 \dots C_5$ for within-block contrasts.

Evaluation The rubric is chosen to be reproducible. Of the 14 attributes (Table 5), 5 are computed deterministically from the structured report, 6 decompose into four binary atoms aggregated by a fixed rule, and 3 are LLM-rated only when the construct cannot be reduced to binary checks. Each attribute is tagged by scope: hypothesis-level (H), or hypothesis and aggregate level (HA). Mechanical attributes summarize tool use, reproduce the DiscoveryBench HMS score (Majumder et al., 2025), and run a per-decoy state machine over tool calls and the final report; rated attributes capture non-triviality, domain plausibility, and actionability.

The six atomized attributes (Falsifiability-H through Grounded-analysis-HA; Table 5) each decompose into four binary atoms (Table 6). A *mechanical* atom is a set-membership check on the schema-validated JSON (e.g., are tool-argument columns within the disclosed schema?). A *text* atom asks whether the hypothesis names a numeric threshold or method, answered by a small structured-output rater. A *hybrid* atom is the logical OR of the two. Atoms are aggregated as

$$\text{score} = \text{clamp} \left(1 + \sum_{i=1}^4 a_i, 1, 5 \right), \quad (1)$$

so the score equals 1 when no atom fires and 5 when all four do.

For the three rated attributes, $K = 3$ raters from different providers (Anthropic, 2025; Team et al., 2025; Singh et al., 2026) score each hypothesis; the attribute score is the mean 1–5 rating. We restrict raters to small models so reasoning-heavy LLMs cannot impute information missing from the report (Li et al., 2024). Only these three attributes

involve a holistic judgement; the remaining eleven are mechanical or atomised and are reproducible by construction. A human-agreement study on 60 sampled hypotheses by 18 annotators (Appendix I) finds that the mean human rating tracks the LLM-ensemble rating at Spearman $\rho = 0.56\text{--}0.62$ on all three, while a randomly drawn pair of *human annotators* agree no better than chance. A larger and more comprehensive human study remains future work. Further, rubric bias does not affect the reported findings. Every disclosure condition - C_0 , $C_1 \dots C_5$, and C_∞ - is scored by the same judge under the same rubric. So whatever bias the rubric carries applies uniformly. Moreover, signal-recovery results in section 3.2 rely on mechanically computed attributes.

3 Experiments

Each run combines the three components of Section 2 - agent, grounding dataset at a fixed disclosure level, evaluator (Figure 3). The panel spans six models (opus-4-7, haiku-4-5, gpt-5-5, gpt-5-4-mini, gemini-3-1-pro, gemini-3-flash), four datasets (Section 2), all disclosure levels (Table 2), and $T \in \{3, 5\}$ turns.

3.1 Attribute vs Disclosure Analysis

Setup. For each (model, dataset, attribute) block we compute the Spearman rank correlation ρ between ordinal disclosure level $C_0 \dots C_\infty$ and mean attribute score across the three turns; the reported statistic is the mean ρ over the 24 (model, dataset) blocks. Confidence intervals are 90% global block-bootstrap with noise pooled across attributes ($B=5000$). All analyses condition on a fixed (model, dataset) cell.

Findings. Of the nine hypothesis-level attributes, five rise consistently with disclosure - binarizability, actionability, non-triviality, causal-vs-correlational reasoning, and reproducibility - two fall consistently - grounded analysis (the strongest coupling, $\rho = -0.36$) and domain plausibility - and two are direction-inconsistent across models - specificity and falsifiability, whose per-model ρ magnitudes match the consistently-signed attributes but whose sign flips with model size (Figures 4 and 19). Adding the three run-level attributes - signal recovery and decoy rejection (rising) and column coverage (falling) - and counting the two direction-inconsistent attributes by their panel-mean sign

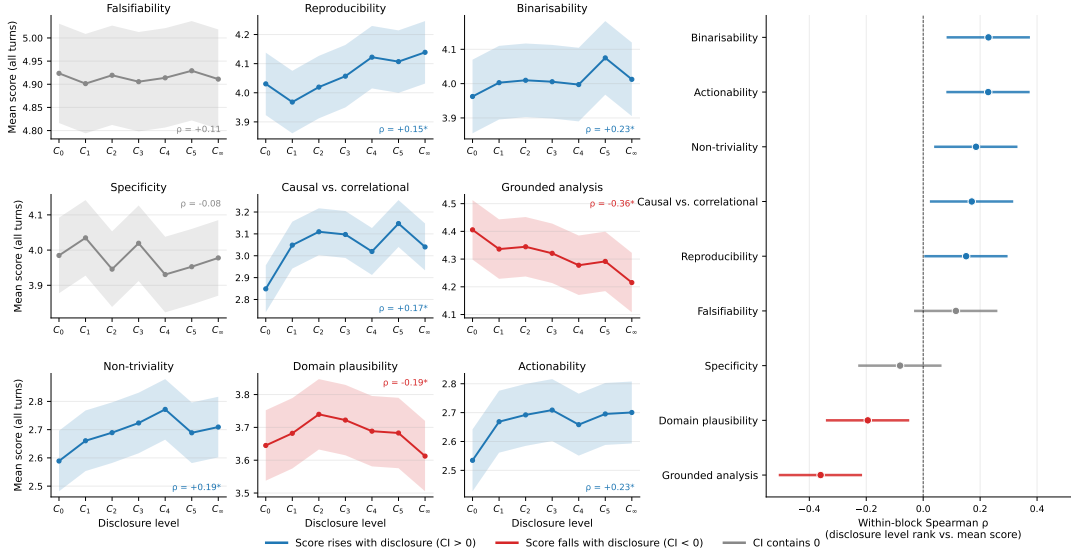


Figure 4: **Disclosure does not affect all attributes equally.** Mean attribute score against disclosure level C_0 – C_∞ for the nine hypothesis-level attributes, averaged over three turns and 24 (model, dataset) cells. *Left*: small multiples with 90% pooled-bootstrap CIs ($B=5000$). *Right*: forest plot of within-block Spearman ρ between disclosure rank and mean score. Color: blue = CI strictly above zero (rises), red = below (falls), grey = contains zero (direction-inconsistent).

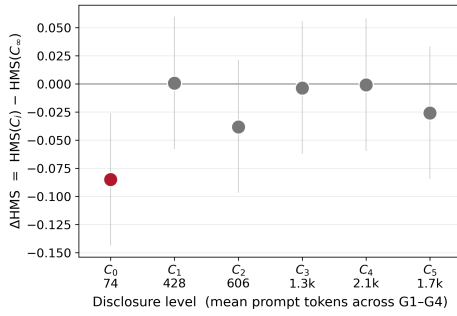


Figure 5: **Signal recovery is flat across C_1 – C_5 ; only the C_0 floor falls behind C_∞ .** Paired $\Delta\text{HMS} = \text{HMS}(C_i) - \text{HMS}(C_\infty)$ over the 6×4 cells at $T=3$; 90% paired-bootstrap CIs ($B=10,000$). Red: CI strictly below zero; grey: contains zero. x -axis labels each level with mean prompt-token cost; $C_\infty \approx 21k$.

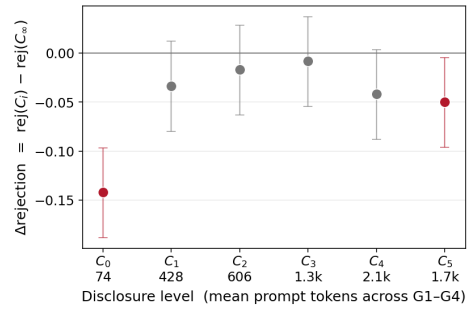


Figure 6: **Decoy rejection rises from C_0 to a descriptive max at C_3 , then dips at C_5 .** Paired $\Delta\text{Rejection} = \text{rej}(C_k) - \text{rej}(C_\infty)$ at panel level; 90% CIs with pooled half-width. C_0 and C_5 : CI strictly below zero. C_1 – C_4 : indistinguishable from C_∞ .

(both negative), this 5/2/2 partition is what produces the seven-rise / five-fall split reported in the abstract and Figure 2. Where richer disclosure does help, the attribute peaks at C_3 or C_4 rather than at C_∞ . Section 3.4 discusses a rule-of-thumb for practical deployment.

3.2 Signal Recovery and Decoy Rejection

Setup. We evaluate signal recovery and decoy rejection on the same $6 \times 4 \times 7$ panel at $T=3$, with all contrasts against the C_∞ ceiling. For signal recovery, 2–4 planted signals per dataset (12 total) give $n=72$ paired observations per contrast; per-signal

differences $\Delta\text{HMS} = \text{HMS}(C_i) - \text{HMS}(C_\infty)$ use pair key (model, dataset, signal-id) with 90% pooled paired-bootstrap CIs ($B=10,000$). For decoy rejection, five Bernoulli verdicts per cell give 840 observations across 20 planted decoys; we report two paired contrasts - panel-level $\Delta\text{Rejection} = \text{rej}(C_k) - \text{rej}(C_\infty)$ ($n=120$; Figure 6) and per-model $\Delta_k = \text{rej}(C_\infty) - \text{rej}(C_k)$ ($n=20$, half-width pooled across 36 contrasts; Figure 7). Per-model and per-dataset disaggregations are in Section 3.3.

Findings. On HMS, C_1 – C_5 all sit inside the C_∞ noise band; only the C_0 floor falls measurably be-

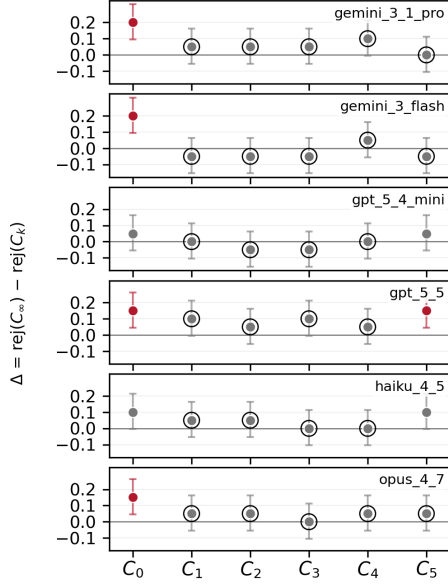


Figure 7: **All six models share minimum-sufficient level C_1 ; per-model maxima cluster at C_3 .** Per-model $\Delta_k = \text{rej}(C_\infty) - \text{rej}(C_k)$; 90% half-width pooled across all 36 contrasts. Open circles mark each model’s minimum-sufficient level.

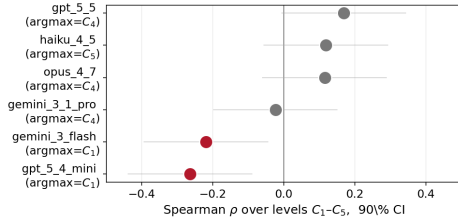


Figure 8: Spearman correlation between disclosure rank and HMS, stratified by model. The two smallest models have CI strictly below zero and peak at C_1 ; the larger four peak at C_4 or C_5 .

hind (Figure 5). On decoy rejection (Figure 6), the largest gain occurs at $C_0 \rightarrow C_1$ - adding only data types and units - after which rejection stays in the C_∞ noise band through C_4 , then dips at C_5 ($\Delta = -0.050$, CI strictly below zero). All six models share minimum-sufficient level C_1 (Figure 7): per-model C_0 floors are heterogeneous, C_1 ceilings cluster tightly, no model’s Δ_k exits the pooled noise band above C_1 , and per-model descriptive maxima cluster at C_3 . Combining both metrics, C_1 is the minimum-sufficient operating point at $\sim 49\times$ lower prompt-token cost than C_∞ (428 vs. $\sim 21k$ tokens. See Section 3.4).

The key takeaways from this are: (i) Levels C_3 , C_4 offer a good rule-of-thumb for the analysis. (ii) One should be careful about disclosing additional information as it might hurt ($C_1 \rightarrow C_2$ in signal

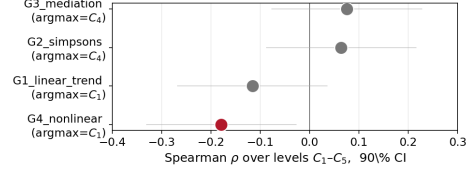


Figure 9: Spearman correlation between disclosure rank and HMS, stratified by model. Marginal-structure datasets (G1, G4) peak at C_1 ; conditional-structure datasets (G2, G3) peak at C_4 .

recovery, $C_3 \rightarrow C_4$ in decoy rejection).

3.3 Model-Capacity and Dataset-Type Interactions

Setup. We reuse the 6×4 panel at $T=3$. For each (model, dataset) block we compute Spearman ρ between ordinal disclosure level and HMS over C_1-C_5 , then average within each stratification. CIs are 90% from a within-block permutation null pooled globally.

Findings. Figures 8 and 9 show that disclosure levels should also be selected based on the dataset/model combination. Per model (Figure 8), the two smallest models in our panel - gpt-5-4-mini and gemini-3-flash - have 90% CI strictly below zero and peak at C_1 ; the three flagship models (gemini-3-1-pro, opus-4-7, gpt-5-5) peak at C_4 , and haiku-4-5 peaks at C_5 , so larger models benefit more from intermediate disclosure. Per dataset (Figure 9), G4-nonlinear has CI strictly below zero and G1-linear-trend also peaks at C_1 , while G2-simpsons and G3-mediation both peak at C_4 : signals visible from column-level descriptors stop improving past C_1 , while signals hidden in conditional or subgroup structure need the C_4 aggregate-statistics layer.

3.4 Choosing an Operating Point

As a rule of thumb - default to C_3 . This level scores highest on average, and most attributes peak here. Increase the level to C_4 when hypotheses misses subtle aspects such as - subgroup or conditional structure, as in G2-simpsons and G3-mediation. Reduce the level to C_1 if token costs are a priority. This level matches C_∞ on signal recovery and decoy rejection at $\sim 49\times$ less cost. Also, use C_1 too when dealing with sensitive information. **Remark:** Note that these are rules deduced from the observations in this article. As with most systems,

evaluating on the specific domain is recommended before deployment.

4 Conclusion

We measure how dataset disclosure shapes hypothesis quality for an LLM data-analyst agent, holding the agent, tools, and decoding fixed (Section 2). Four key takeaways are – (a) **Less data is often enough.** Adding column types and units to column names (C_1) matches full-table access (C_∞) on signal recovery and decoy rejection, at $\sim 49\times$ lower prompt-token cost (~ 428 vs. $\sim 21k$ tokens; Section 3.2). A production system that pays per input token can run at C_1 on these two metrics with no measurable loss. (b) **More disclosure does not uniformly help.** Seven of the twelve attributes rise from C_0 to C_∞ and five fall; column coverage and grounded analysis fall most (Section 3.1). A deployment that cares about grounded, conservative claims should not assume that giving the agent the full table improves the outcomes. (c) **Model size and signal type set the operating point.** The two smallest models in our panel peak at C_1 ; the four larger ones peak at C_4 or C_5 (Section 3.3). Datasets with column-readable signals (G1, G4) also peak at C_1 ; datasets with conditional subgroup structure (G2, G3) need the C_4 aggregate-statistics layer. As a rule of thumb, C_3 gives the best mean score (See Section 3.4). (d) **Partial-information operation keeps data inside the boundary by construction.** Because the agent recovers most of the headline metrics from C_1 – C_4 , an operator can keep row-level records inside the institution and still obtain quality close to the C_∞ ceiling, with the GID-conditioner filtering the context to exactly what the disclosure policy permits. While this does not completely solve GDPR compliance, it provides a way to enforce an important aspect - preserving the data boundary. We also note that, it is on this basis we deploy C_3 with a healthcare partner (Appendix G).

Two future directions of research follow. First, low-disclosure operation is cheap enough that future agents could infer the disclosure level a domain warrants rather than always asking for the full table. Second, the aggregates transmitted at C_3 – C_4 are the natural place to also apply techniques from differential privacy, and certifying that an agent consumes inputs below an approved disclosure level, which would make it an auditable property of the system for effective enforcement of policy.

5 Limitations

In our view, the two factors most likely to break the findings when applied in practice lie outside the controlled axes of this analysis. First, the results characterize a single agent architecture - tool-last-call, bounded-interaction loop (Section 2) - so transfer to agents with richer tool registries, alternative synthesis prompts, or unbounded interaction horizons is an empirical question we do not settle. Second, all datasets, prompts, and rubrics are English- and tabular-only; whether the disclosure axis behaves analogously for other languages or modalities is open. The remaining apparent caveats - the planted-signal synthetic substrate (Section 2), the atomized rubric (Section 2), and the $C_5 \rightarrow C_\infty$ modality jump that makes the chain non-monotone at the ceiling (Table 2) - are deliberate design choices required for controlled measurement and, in our judgment, do not affect the direction of the reported effects.

Acknowledgements

Aditya Challa, Sravan Danda and Snehanishu Saha acknowledge the grant from MAMTA Health Institute for Mother and Child. Aditya Challa and Snehanishu Saha further acknowledge the financial support of ANRF CRG-DST (CRG/2023/003210). Snehanishu Saha also acknowledges partial support from the DBT-Builder project (BT/INF/22/SP42543/2021), Govt. of India. Sravan Danda would like to thank ANRF for funding from SERB-SURE project number SUR/2022/002735.

References

- Anthropic. 2025. [Claude 4 system card](#). Technical report, Anthropic. Accessed: May 14, 2026.
- Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A Merrill, Jeffrey Heer, and Tim Althoff. 2024. [BLADE: Benchmarking language model agents for data-driven science](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971, Miami, Florida, USA. Association for Computational Linguistics.
- Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. 2024. DS-Agent: Automated data science by empowering large language models with case-based reasoning. In *Int. Conf. Mach. Learning*, volume 235 of *Proceedings of*

- Machine Learning Research*, pages 16813–16848. PMLR.
- Eugenie Lai, Gerardo Vitagliano, Ziyu Zhang, Om Chabra, Sivaprasad Sudhir, Anna Zeng, Anton A. Zabreyko, Chenning Li, Ferdi Kossmann, Jialin Ding, Jun Chen, Markos Markakis, Matthew Russo, Weiyang Wang, Ziniu Wu, Mike Cafarella, Lei Cao, Samuel Madden, and Tim Kraska. 2026. [KramaBench: A benchmark for AI systems on data-to-insight pipelines over data lakes](#). In *Int. Conf. on Learning Representations*.
- Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. [Llms-as-judges: A comprehensive survey on llm-based evaluation methods](#). *Preprint*, arXiv:2412.05579.
- Ruizhe Li, Chiwei Zhu, Benfeng Xu, Xiaorui Wang, and Zhendong Mao. 2025. [Automated creativity evaluation for large language models: A reference-based approach](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 21475–21488, Suzhou, China. Association for Computational Linguistics.
- Yining Lu, Dixuan Wang, Tianjian Li, Dongwei Jiang, Sanjeev Khudanpur, Meng Jiang, and Daniel Khashabi. 2025. [Benchmarking language model creativity: A case study on code generation](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2776–2794, Albuquerque, New Mexico. Association for Computational Linguistics.
- Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhkar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. 2025. [DiscoveryBench: Towards data-driven discovery with large language models](#). In *Int. Conf. on Learning Representations*.
- Gaurav Sahu, Abhay Puri, Juan A. Rodríguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vázquez, Nicolas Chapados, Christopher Pal, Sai Rajeswar, and Issam H. Laradji. 2025. InsightBench: Evaluating business analytics agents through multi-step insight generation. In *Int. Conf. on Learning Representations*. OpenReview.net.
- Skipper Seabold and Josef Perktold. 2010. [Statsmodels: Econometric and statistical modeling with python](#). *SciPy 2010*.
- ChengAo Shen, Zhengzhang Chen, Dongsheng Luo, Dongkuan Xu, Haifeng Chen, and Jingchao Ni. 2025. [Exploring multi-modal data with tool-augmented LLM agents for precise causal discovery](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 636–660, Vienna, Austria. Association for Computational Linguistics.
- Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Int. Conf. Mach. Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.
- E. H. Simpson. 1951. [The interpretation of interaction in contingency tables](#). *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2):238–241.
- Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 466 others. 2026. [Openai gpt-5 system card](#). *Preprint*, arXiv:2601.03267.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1332 others. 2025. [Gemini: A family of highly capable multimodal models](#). *Preprint*, arXiv:2312.11805.
- Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024. [SciMON: Scientific inspiration machines optimized for novelty](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 279–299, Bangkok, Thailand. Association for Computational Linguistics.
- Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob N. Foerster, Jeff Clune, and David Ha. 2025. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv*, abs/2504.08066.
- Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. 2025. [Assessing and understanding creativity in large language models](#). *Machine Intelligence Research*, 22(3):417–436.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.
- Yangqiaoyu Zhou, Haokun Liu, Tejes Srivastava, Hongyuan Mei, and Chenhao Tan. 2024. [Hypothesis generation with large language models](#). In *Proceedings of the 1st Workshop on NLP for Science (NLP4Science)*, pages 117–139, Miami, FL, USA. Association for Computational Linguistics.

A Related Work

Data-analyst agents and benchmarks. A first thread evaluates LLM agents on multi-step data analysis under full dataset access. Majumder et al. (2025) formalize hypothesis search and verification across 264 real-world and 903 synthetic tasks in six domains, with the best evaluated system reaching 25% accuracy. Sahu et al. (2025) score end-to-end business analytics over 100 enterprise datasets, and Lai et al. (2026) extend the setting to noisy data lakes of 1,700 files across 24 sources. On the agent-design side, Guo et al. (2024) couple an LLM with case-based retrieval over expert Kaggle solutions, Yamada et al. (2025) combine progressive tree search with a vision-language reviewer, Gu et al. (2024) benchmark multi-step analytical workflows, and Shen et al. (2025) augment the agent with tool-mediated causal discovery. Across this literature, dataset access is fixed at the full disclosure.

Context ablation and hypothesis generation. A second thread varies what context a model receives. Shi et al. (2023) quantify the effect of *irrelevant* context on reasoning but do not study graduated levels of relevant data access. Zhou et al. (2024) formulate hypothesis refinement as a multi-armed-bandit exploration loop, and Wang et al. (2024) condition idea generation on retrieved prior literature against a novelty objective. These works manipulate the *type* of context (relevant/irrelevant, retrieved/none) for a task with otherwise full data access.

Multi-attribute LLM evaluation. A third related thread asks how to score open-ended LLM output. Zheng et al. (2023) ground LLM-as-judge evaluation through MT-Bench and Chatbot Arena, surveyed at scale by Li et al. (2024). In adjacent creativity-rubric work, Zhao et al. (2025) score fluency, flexibility, originality, and elaboration on Torrance-style prompts; Lu et al. (2025) pair denial prompting with convergent/divergent metrics; and Li et al. (2025) propose a reference-based Likert protocol with improved pairwise agreement. These rubrics aggregate to a single quality score. In this article, we instead retain per-attribute resolution so that the signed shifts induced by disclosure remain observable (Section 2).

Gaps addressed. To our knowledge, no prior work simultaneously (i) treats dataset-information

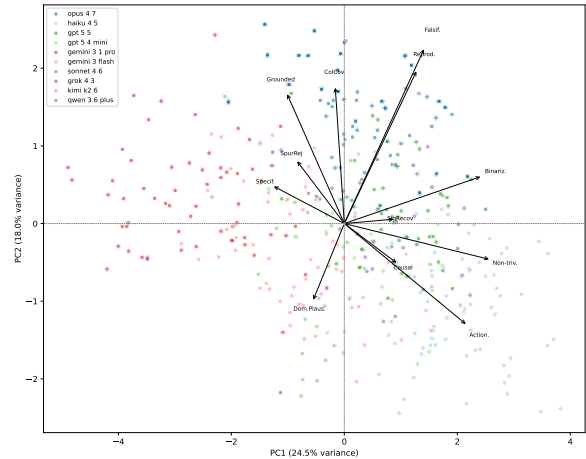


Figure 10: PCA biplot (PC1 = 24.5%, PC2 = 18.0%, total = 42.5%) of the 12 non-degenerate attributes, coloured by model. The moderate explained variance confirms that the rubric is not reducible to one or two latent dimensions, consistent with the multi-attribute framing in Section 2.

access as a controlled experimental axis, (ii) preserves multi-attribute resolution rather than collapsing hypothesis quality into a single utility scalar, and (iii) isolates disclosure from confounds in real-world data quality via a planted-signal substrate. The analysis in this paper closes the three gaps jointly for the data-analyst-agent setting.

B Statistical Methodology and Per-Contrast Robustness

All confidence intervals in the main text use a pooled bootstrap half-width: a single scalar half-width w is estimated from the empirical bootstrap distribution of all disclosure-level contrasts within a metric-baseline combination, and the same w is applied to every bar in the figure. Pooling is justified on power grounds: each individual contrast is estimated from a relatively small number of paired observations, so per-contrast widths would vary primarily as a function of sample-size imbalance rather than true effect variability. A shared half-width guards against artefactually wide or narrow individual bars and enables direct visual comparison of contrast magnitudes across disclosure levels.

We report 90% bootstrap confidence intervals throughout. Within each contrast family the goal is to characterise a trend across the C_1 – C_5 chain rather than to test a single null hypothesis, and 90% intervals are standard practice in evaluation studies that condition on a fixed panel of models and datasets. Because each contrast is paired (the

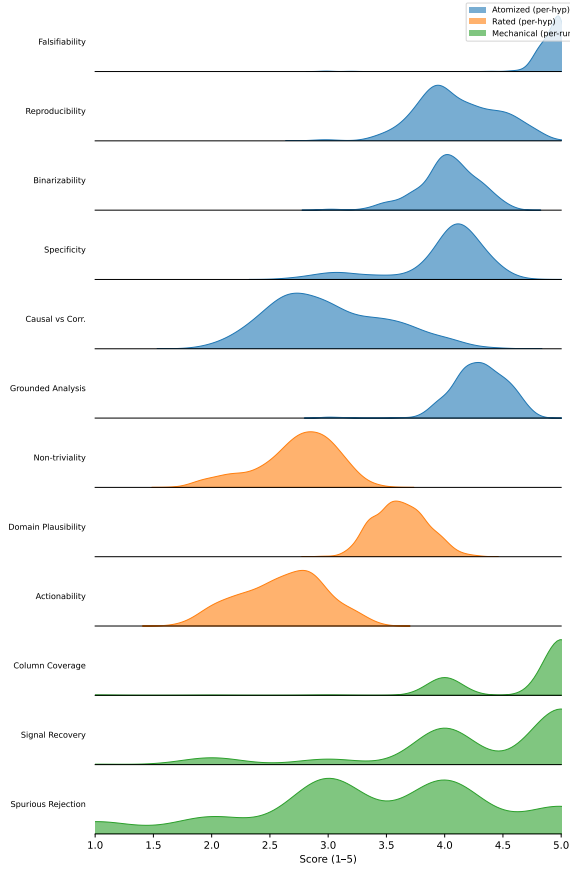


Figure 11: Score distributions (KDE ridges) for each attribute across all 460 runs. Ceiling concentration in falsifiability and spurious_rejection bimodality illustrate the range of discriminative power across attributes; attributes with broad unimodal distributions contribute the most to the disclosure-level signal in Section 3.

same model–dataset cell appears on both sides), the pooled intervals are already conservative relative to two-sample alternatives. Figure 14 reproduces the four headline contrast families under both pooled and per-contrast half-widths; the ordering of disclosure levels is preserved in every panel, confirming that the main-text conclusions hold under the alternative construction.

C Exploratory Data Analysis

The figures below characterise the 168-run core panel (six models $\times$ four datasets $\times$ seven disclosure levels $\times$ one seed $\times$ three turns) used throughout the analysis. All figures use the 12 non-degenerate attributes, excluding parsimony_batch (constant 5.0) and diversity_batch (near-ceiling with negligible variance).

Seed stability is exact: the intra-class correlation

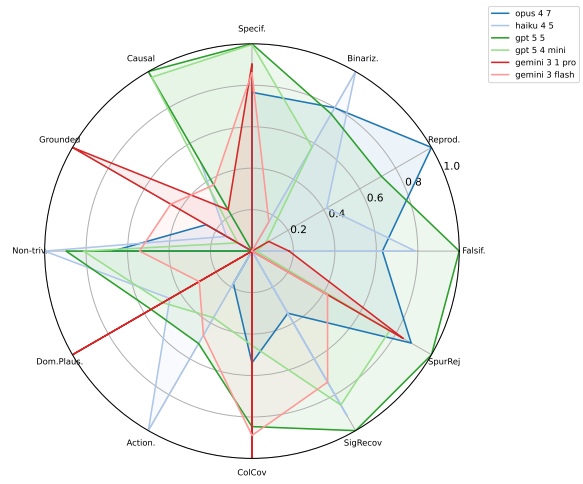


Figure 12: Model fingerprint radar for the six core-panel models, normalised per attribute. Each model’s characteristic strength profile is visible at a glance; precise rank values appear in Figure 16.

coefficient across seeds is 1.0 for all 12 attributes over the 52 multi-seed cells in the panel, confirming that attribute scores are fully reproducible for a given (model, dataset, level, turn-count) configuration.

D Tool Descriptions

Table 3 lists all 25 tools available to the agent in Phase (b), with a one-line description of each. Table 4 below gives the full argument schemas, grouped into six functional categories: *Descriptives* (summary statistics and frequency tables), *Statsmodels primitives* (regression, ANOVA, *t*-test, and chi-square wrappers), *Added primitives* (histogram and multinomial logit), *Diagnostics* (effect sizes, VIF, and regression residual checks), *Compound* (tools that combine stratification or conditioning with a statistical test), and *Misc.* (partial correlation, binning helpers, and a plot-hint tool). The schema is fixed across all disclosure levels; the agent’s tool registry does not change with C_k .

E Metrics and Atoms

Table 5 lists the 14 high-level evaluation attributes grouped by extraction method. Six attributes are *atomised*: each is decomposed into four binary yes/no atoms and aggregated via $\text{score} = \text{clamp}(1 + \sum_{i=1}^4 a_i, 1, 5)$, yielding deterministic, reproducible scores. Three attributes are *rated* by an ensemble of three small LLMs using rigorously templated prompts. The remaining five are *mechanical*: computed deterministically from the

Tool	Description	Tool	Description
describe	Per-column n , nulls, mean, std, min, quartiles, max.	mnlogit	Multinomial logistic regression; per-class coefficients and SEs.
value_counts	Top- k most frequent values and counts for one column.	cohens_d_or_eta_squared	Effect size: Cohen’s d (2 groups) or η^2 (>2 groups).
groupby_agg	Group by columns; aggregate target with mean, median, std, or count.	vif	Variance inflation factors for predictors in a formula.
crosstab	Frequency cross-tabulation of two categorical columns.	regression_diagnostics	OLS diagnostics: Shapiro, Breusch–Pagan, Durbin–Watson.
corr_matrix	Pairwise Pearson or Spearman correlation among numeric columns.	stratified_correlation	Pearson/Spearman $r(x, y)$ within each level of a stratifier.
quantiles	Requested quantile values of a numeric column.	binning_chi_square	Quantile-bin x and y , then chi-square on the joint table.
gls	Generalised Least Squares from a formula; coefs, SEs, p , R^2 .	ttest_per_group	Two-sample t -test, repeated within each stratum.
glm_logit	Logistic regression from a formula; coefs, SEs, p , log-lik.	simpsons_diagnostic	Compare marginal $r(x, y)$ vs. within-stratum r ; flags sign reversal.
glm_poisson	Poisson GLM from a formula; coefs, SEs, p , log-lik.	outlier_robust_regression	RLM-Huber or quantile-median regression from a formula.
anova	OLS-based ANOVA table (factor, SS, df, F , p) per term.	partial_correlation	Pearson $r(a, b)$ after residualising on one or more controls.
ttest	Two-sample t -test between the two levels of a grouping column.	binning_then_mean	Quantile-bin x ; report mean, SE, count of y per bin.
chi_square	Chi-square independence test for two categorical columns.	plot_hint	Plot-relevant statistics for a chosen visualisation kind.
histogram	Equal-width binning of a numeric column; edges and counts.		

Table 3: Tool registry available to the agent in Phase (b). Full argument schemas are listed in Table 4.

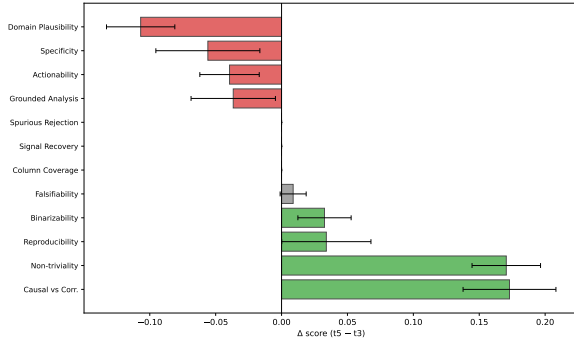


Figure 13: Turn-count sensitivity (t3 vs. t5) for all 12 attributes over 72 paired cells. Causal-vs-correlational (+0.17) and non-triviality (+0.17) improve with more turns; domain-plausibility (−0.11) and specificity (−0.06) decrease - a trade-off that Section 3.1 identifies as direction-inconsistent across models. Mechanical attributes (column coverage, signal recovery, spurious rejection) are invariant to turn count.

structured JSON report and (for signal-recovery and spurious-rejection) the ground-truth planted-signal record. Table 6 lists all 24 atoms, four per atomised attribute, together with the yes/no question each atom answers and its extraction method (*mech.* for schema-validated JSON checks, *LLM* for structured-output calls, *hybrid* for a logical OR of both).

F Ablation: Subtractive vs. Additive Disclosure Conditions

The A-level conditions (A1–A5) instantiate a subtractive manipulation: each condition starts from the full-disclosure context (C_∞) and removes specific information components (column types, summary statistics, sampled rows, and data-quality flags) in sequence. This contrasts with the additive chain C_1 – C_5 reported in the main text, in which components are introduced progressively from a minimal context. Figure 18 plots per-attribute score differences $\Delta(\text{A-level} - \text{matched C-level})$. The direction of effects is broadly consistent with the additive analysis: attributes that rise with disclosure under additive conditions tend to improve under the equivalent subtractive condition as well, suggesting that the attribute–information relationship is approximately symmetric across manipulation direction and is not an artifact of the additive framing.

G Applied Context and Design Provenance

This work originated in a collaboration with a healthcare organization that holds electronic health records (EHR) subject to strict data-governance re-

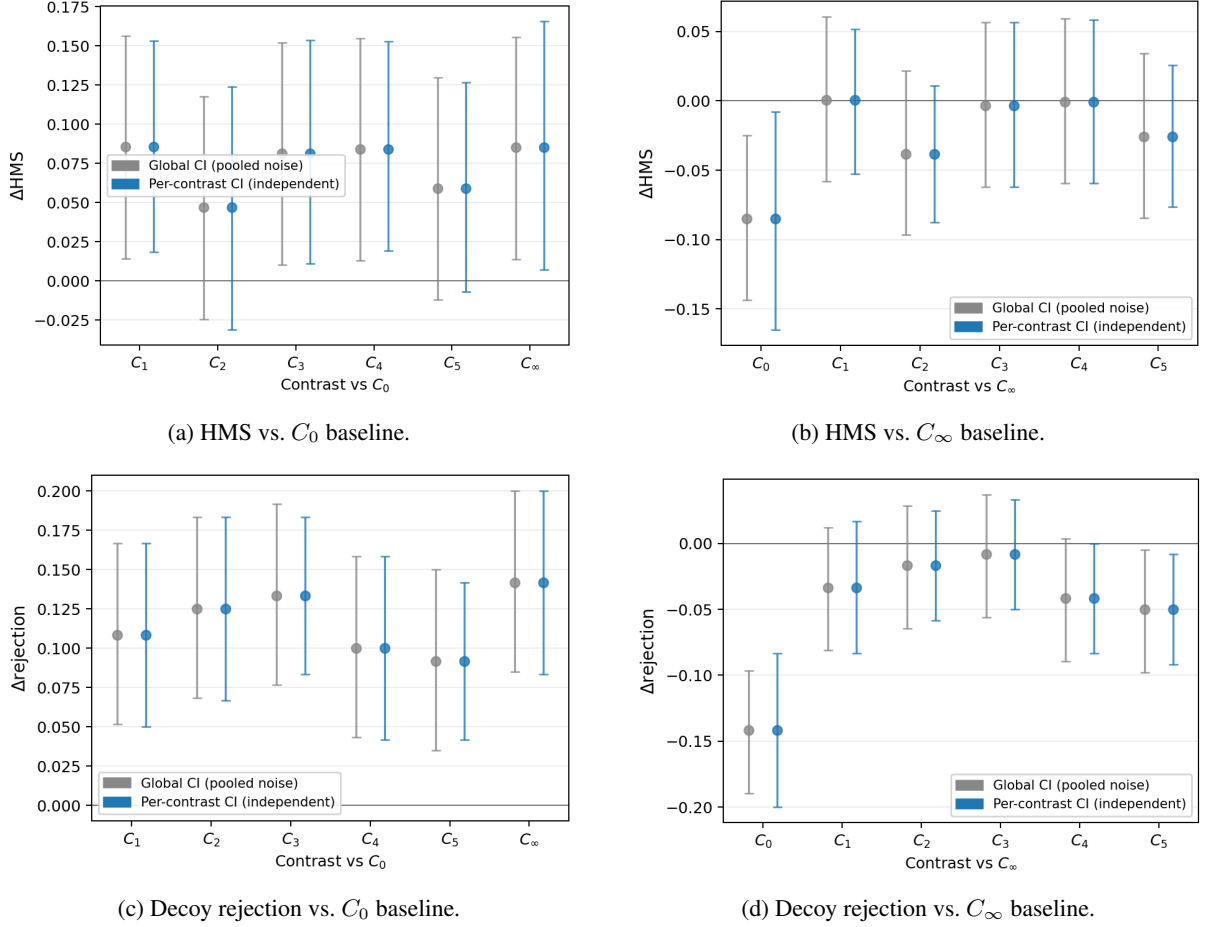


Figure 14: Pooled vs. per-contrast bootstrap CI construction for the four headline contrast families (HMS and decoy-rejection rate, each compared against the C_0 floor and C_∞ ceiling baselines). In every panel the ordering of disclosure levels under pooled half-widths (main text) is preserved when per-contrast half-widths are substituted, confirming robustness of the conclusions to the choice of half-width estimator.

quirements. The organization needed statistical insight from its records but could not expose patient-level data to a third-party LLM API. The practical question - *what is the minimum information an LLM analyst needs to produce useful hypotheses?* - motivated the graduated disclosure framework.

Based on domain expertise and iterative experimentation on the partner’s data, we identified C_3 -level disclosure ($\{F_1, \dots, F_4\}$: column names, data types, free-text metadata, and per-column summary statistics; Table 2) as the operating point deployed in production - independently reaching the default that Section 3.4 recommends on the synthetic panel. The choice was driven by two observations: (i) summary statistics were sufficient for the agent to generate clinically plausible hypotheses that domain experts judged actionable, and (ii) no patient-identifiable information leaves the institution at this disclosure level, since all transmitted features are population-level aggregates.

The synthetic benchmark presented in this paper was designed to *rigorously test and generalize* the intuitions developed during this engagement. We do not claim the synthetic results independently validate the proprietary findings; the causal direction is the reverse. What the benchmark provides is a controlled, reproducible environment in which disclosure is the sole treatment variable—a property the proprietary setting, with its fixed deployment configuration and uncontrolled domain complexity, cannot offer.

Why not report proprietary results? Three reasons: (i) the data-use agreement prohibits disclosure of any patient-derived statistics, including aggregate performance metrics; (ii) the proprietary experiments used a single disclosure level (C_3) in a deployment context, not the full twelve-level experimental design, so they do not constitute a controlled comparison; and (iii) reporting them as

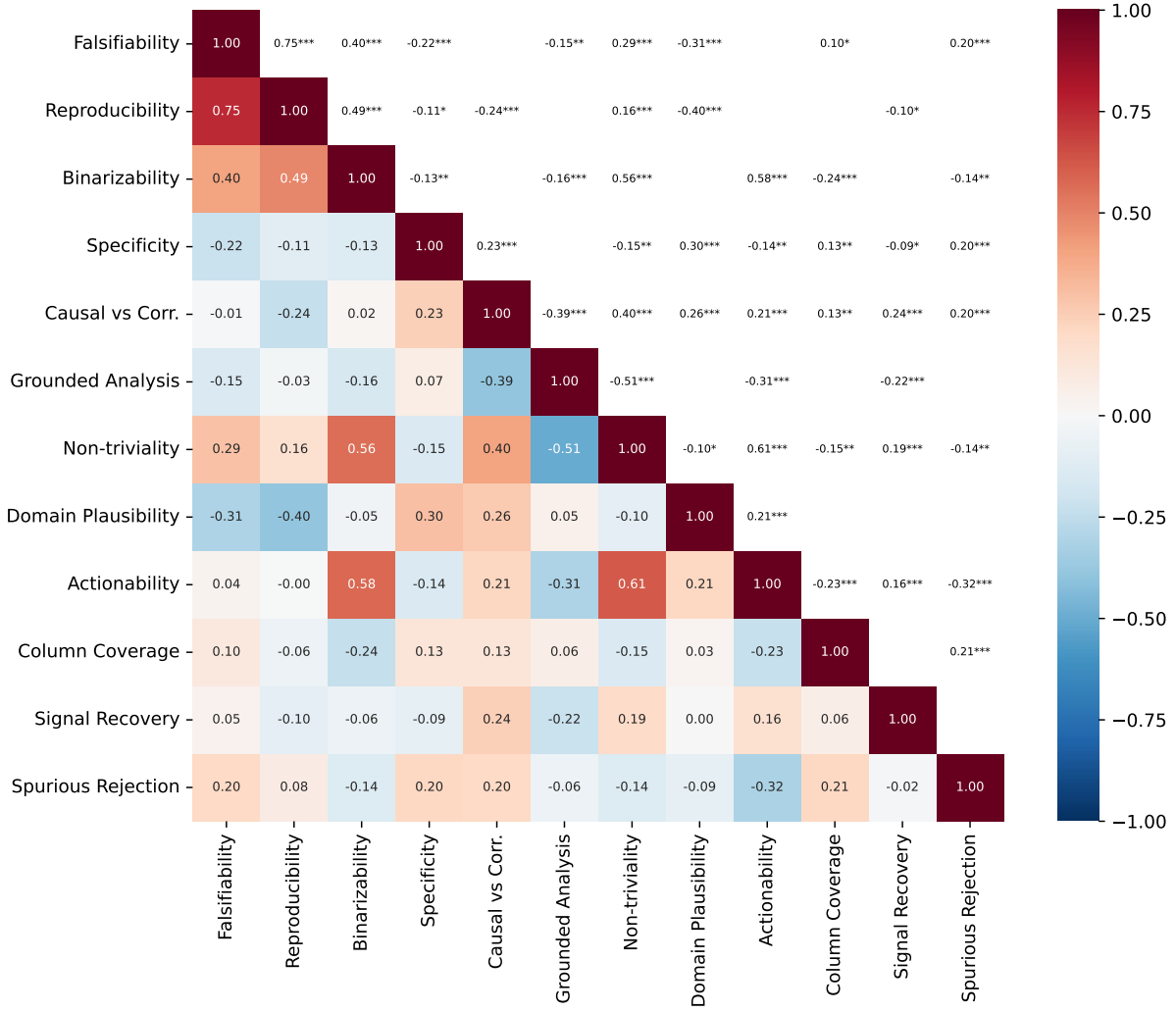


Figure 15: Spearman rank-correlation matrix of the 12 non-degenerate attributes over all 460 runs (upper triangle: ρ values; lower triangle: significance stars). Off-diagonal clusters confirm that the rubric captures several partially independent dimensions of hypothesis quality, which is what makes the 5/2/2 sign-split reported in Section 3.1 structurally interpretable rather than an artifact of correlated axes.

“validation” would misrepresent the relationship between the applied work and the benchmark.

H Sanity Check

Sanity check on real-world datasets. We add two real-world tabular datasets to the synthetic panel: R1 (891-row labelled Titanic manifest, four sex/class/age/fare signals and three weak decoys) and R2 (344-row Palmer Penguins morphology, four species/body-mass and within-species sex signals and two metadata decoys). Figure 20 plots the per-dataset Spearman ρ between disclosure level and metric value, with a 90% pooled within-block permutation interval, across fourteen attributes spanning hypothesis quality, batch-level coverage/diversity/parsimony, per-signal recovery, and per-decoy rejection. In every subplot R1 and R2 lie

inside the spread already produced by G1–G4: their closest-G neighbour shifts with the attribute, with R1 tracking G1 (linear-trend) on specificity and domain plausibility but G3 (mediation) on causal-vs-correlational and G2 (Simpson’s paradox) on per-signal HMS, and R2 splitting almost evenly between G2 and G4 across the eleven informative attributes. Three batch-level metrics saturate at the LLM-judge ceiling and contribute $\rho = 0$ for every dataset; the eleven non-degenerate attributes show no case in which a real-world dataset behaves systematically unlike every synthetic dataset. The real-world panel sits inside the envelope already covered by G1–G4 rather than indicating a regime our metrics fail to capture.

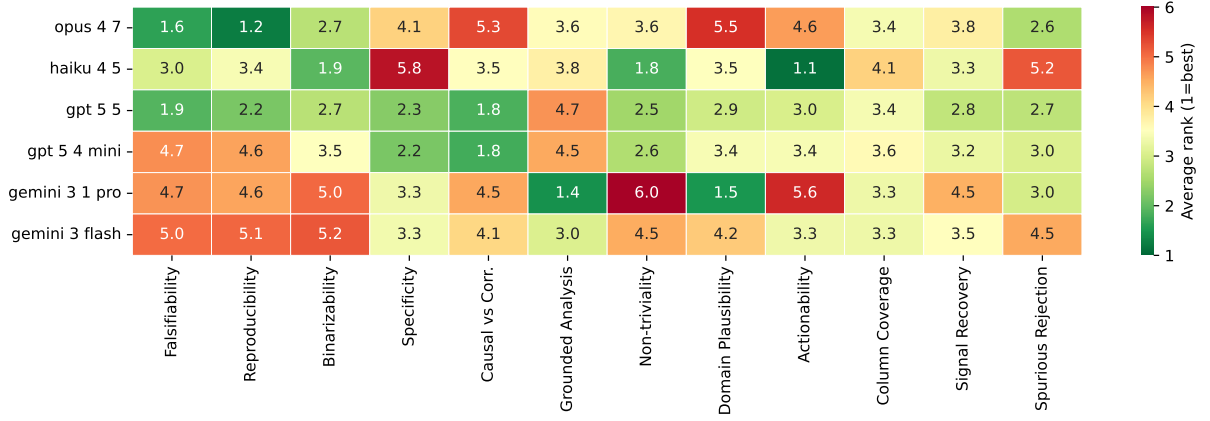


Figure 16: Model $\times$ attribute rank heatmap (rank 1 = best, 6 = worst) averaged over the 28 (dataset $\times$ level) cells in the core panel. No model dominates across all attributes, and visible row-level rank inversions confirm that models trade off different aspects of hypothesis quality rather than sitting on a single quality dimension.

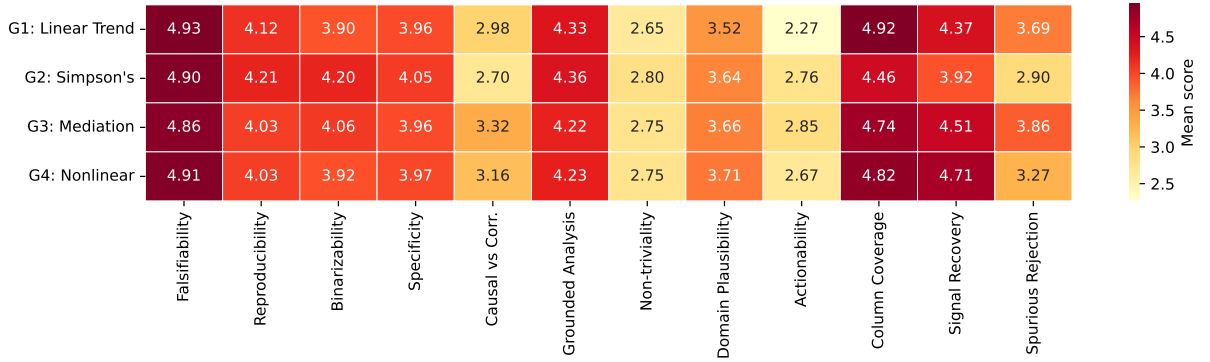


Figure 17: Dataset $\times$ attribute mean scores averaged over models, levels, and seeds. Datasets vary non-uniformly across attributes, justifying the multi-dataset panel design: no single dataset is uniformly easy or hard, and datasets stress different rubric axes depending on their planted-signal structure.

I Human Agreement on the Rated Attributes

Scope. Eleven of the fourteen attributes are computed mechanically - (i) five deterministically, and (ii) six as aggregations of four binary atoms each, with fourteen of the twenty-four atoms being mechanical checks on schema-validated JSON (Tables 5 and 6). Hence these eleven attributes are reproducible by construction. No human study is needed there. What the study below validates is the three LLM-rated attributes alone - non-triviality, domain plausibility and actionability.

Protocol. We draw 60 hypotheses at random and without replacement from the 16,110 scored hypotheses in the panel, and collect 1–5 ratings on all three attributes from 18 annotators - three authors of this article and fifteen colleagues from outside the project, none of whom had seen the LLM ratings. Annotators receive the rubric text *verbatim*

as given to the LLM raters, and see the hypothesis text alone. We randomise the presentation order once, at sampling time, and then hold it fixed across annotators.

Two humans agree barely better than chance, while two LLMs agree modestly and rank consistently. Table 9 reports mean pairwise agreement within and across rater types. Observe that two typical *human* raters sit at $\alpha = -0.027$ and $\rho = 0.040$ pooled, while two LLM raters reach $\alpha = 0.238$ and $\rho = 0.482$ on the same items. And what the gap between the two ρ values gives us is – *the models rank hypotheses similarly even where they disagree on the absolute level*.

The two series agree in aggregate, which is the level at which we use them. Table 10 compares the mean over every annotator against the three-rater LLM mean, and measures agreement between the two series. Spearman ρ is 0.560–0.620 on all

Table 4: Argument schemas for the 25 tools available in Phase (b), grouped by category.

Tool	Arguments
<i>Descriptives</i>	
describe	cols: list[str]
value_counts	col: str; top_k: int = 10
groupby_agg	by: str list[str]; target: str; agg: Lit["mean", "median", "std", "count"]
crosstab	row: str; col: str; normalize: bool = False
corr_matrix	cols: list[str]; method: Lit["pearson", "spearman"] = "pearson"
quantiles	col: str; qs: list[float]
<i>Statsmodels primitives</i>	
gls	formula: str; sigma: list[list[float]] None = None
glm_logit	formula: str
glm_poisson	formula: str
anova	formula: str
ttest	group_col: str; target: str
chi_square	row: str; col: str
<i>Added primitives</i>	
histogram	col: str; num_bins: int = 10
mnlogit	formula: str
<i>Diagnostics</i>	
cohens_d_or_eta_squared	group_col: str; target: str
vif	formula: str
regression_diagnostics	model_spec: str
<i>Compound</i>	
stratified_correlation	x: str; y: str; by: str; method: Lit["pearson", "spearman"] = "pearson"
binning_chi_square	x: str; y: str; bins: int = 10
ttest_per_group	group_col: str; target: str; by: str
simpsons_diagnostic	x: str; y: str; by: str
outlier_robust_regression	formula: str; method: Lit["rlm_huber", "quantile_median"] = "rlm_huber"
<i>Misc.</i>	
partial_correlation	col_a: str; col_b: str; control_col: str list[str]; method: Lit["pearson"] = "pearson"
binning_then_mean	x: str; y: str; bins: int = 6
plot_hint	col_a: str; col_b: str; kind: Lit["scatter", "bar", "line", "box", "hex", "heatmap"]

three attributes, and domain plausibility reaches $\alpha = 0.535$. So the conclusion is - *individual raters, human or LLM, may disagree, but their means agree modestly*. Note that this is also the procedure we adopt in this article: every score in Section 3 is a mean over raters, turns and (model, dataset) cells.

Open Questions: A larger study - stratified by disclosure level, with a calibration round and duplicated items to establish an intra-annotator noise floor - would support per-level human means and a direct test of whether rater bias varies with disclosure. We wish to pursue this as future work.

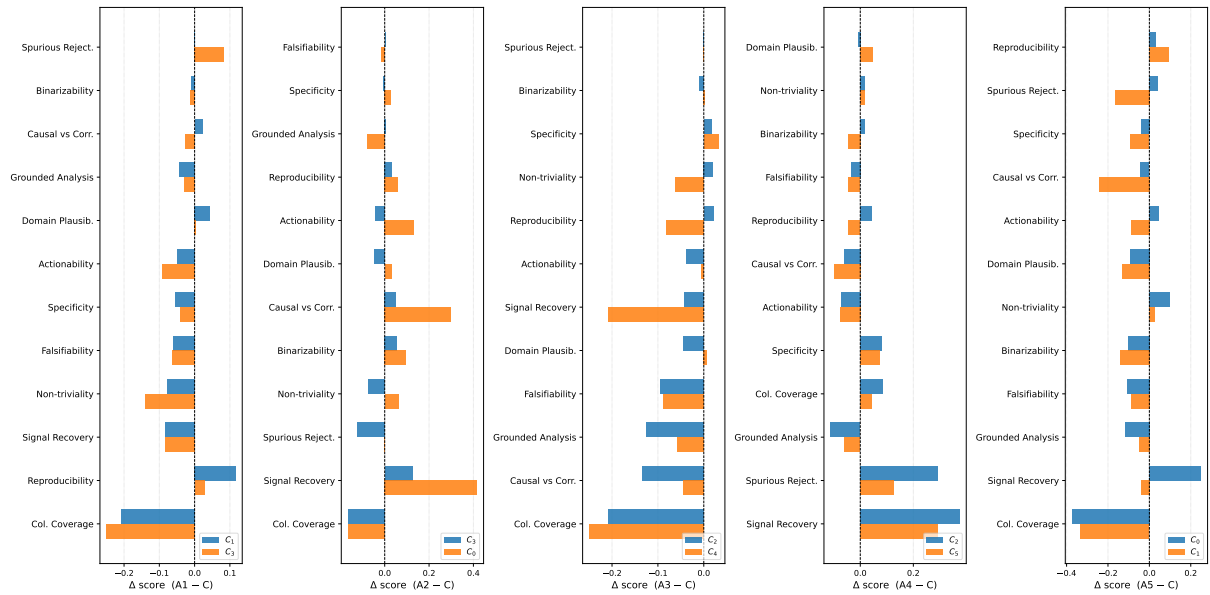


Figure 18: Per-attribute score difference between matched A-level (subtractive) and C-level (additive) disclosure conditions. Directional consistency across the two manipulation types supports the robustness of the attribute-level disclosure effects reported in Section 3.1.

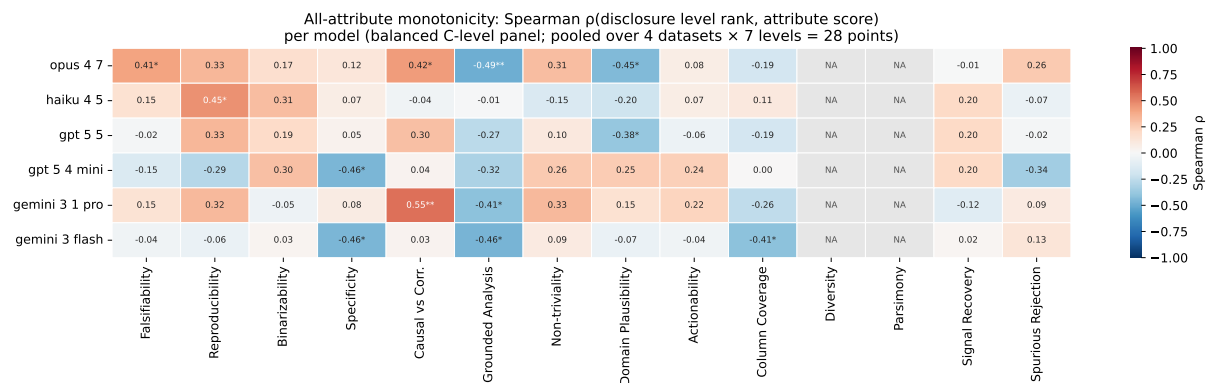


Figure 19: Per-model Spearman rank correlation between disclosure level and each rubric attribute, showing which attributes increase, decrease, or remain flat as more information is disclosed.

Attribute	Scope	Method	What it measures, and how the 1–5 score is computed
<i>Atomised: score = clamp($1 + \sum_{i=1}^4 a_i$, 1, 5) over the four binary atoms named below (Table 6)</i>			
Falsifiability-H	H	atomised	Could the run show the claim false? Atoms: declarative assertion; argument columns non-empty and inside the disclosed schema; a numeric leaf in some error-free result; a threshold in the text or a method-selecting argument key.
Reproducibility-H	H	atomised	Could another analyst re-run the stated check? Atoms: tools called and none of them sample-based (binning or quantiles); a method-selecting argument key; a threshold, comparison or named method in the text; no subjective hedging.
Binarizability-H	H	atomised	Does the claim resolve to a yes/no verdict? Atoms: threshold or boolean comparator in the text; a scalar, boolean or headline-keyed tool result; arguments that aggregate, or a scalar-output tool; no open-ended phrasing.
Specificity-H	H	atomised	How tightly is the claim pinned down? Atoms: a context phrase restricting the population; a disclosed column named in the text; a relationship verb; no loose quantifiers.
Causal-vs-correlational-H	H	atomised	Does the claimed strength match the evidence? Atoms: a tool tagged <i>supports-causal</i> ; a non-trivial conditioning set in the arguments; a partial-effect tool invoked; the text’s verb within the capability of the tools called.
Grounded-analysis-HA	HA	atomised	Do the text and the executed analysis refer to the same thing? Atoms: tool family matches the relationship type; columns named in the text present in the arguments; a stated threshold resolved by a numeric result; no angle drift across later hypotheses of the turn that share a column.
<i>Rated: $K = 3$ small cross-provider raters score each hypothesis at temperature 0 against the 1–5 anchors below; score = their mean</i>			
Non-triviality-H	H	rated	Is the claim already readable from what was disclosed? Rated from the disclosed block, the hypothesis and the tool names. 1 = readable off the column names or metadata alone; 2 = plausible, but needing no analytical insight beyond basic domain knowledge; 3 = a non-obvious relationship, or a quantitative threshold that requires analysis; 4 = a meaningful pattern that would take real domain or analytical expertise to see; 5 = a counter-intuitive finding - a reversal, an interaction, a sign switch under conditioning.
Domain-plausibility-H	H	rated	Is the claim consistent with domain knowledge? Rated from the hypothesis and the tool names alone. 1 = contradicts a known empirical or logical constraint (a negative price or count, an impossible direction of effect); 2 = logically possible but implausible under standard domain knowledge; 3 = plausible, with nothing obviously ruling it out; 4 = aligned with prior literature or strong domain priors; 5 = tightly aligned with a known mechanism or a well-replicated finding.
Actionability-H	H	rated	Would a decision change if the verdict were known? 1 = none would - the claim is purely academic; 2 = the verdict could inform a future study, but no operational decision; 3 = a specific practitioner decision would plausibly change; 4 = a concrete decision-maker (person, policy, system) and the decision they would change are both named; 5 = a specific intervention with a named beneficiary and a measurable outcome.
<i>Mechanical: fixed formulae over the structured report and the ground-truth planted-signal record</i>			
Column-coverage-Batch	Batch	mechanical	Fraction r of the disclosed columns (15 in G1–G4) appearing as a variable, formula term, or conditioning argument in at least one hypothesis of the run. Score 5 only at $r = 1$, else $\max(1, \lceil 4r \rceil)$; 4 above 75%, 3 above 50%, 2 above 25%, 1 at or below 25%.
Diversity-Batch	Batch	mechanical	Shannon entropy in bits of the run’s tool calls over the seven declared inferential angles (univariate, bivariate, conditional, distributional, threshold, group-structure, temporal), of which the tool registry realises five. Bands ≥ 1.0 , 0.75, 0.50, 0.25 $\rightarrow$ 5, 4, 3, 2; else 1.
Parsimony-Batch	Batch	mechanical	Fraction of the run’s hypotheses that call a complex-tier tool (gls, the two GLMs, mnlogit, outlier_robust_regression) with no conditioning argument. Bands 0 $\rightarrow$ 5, $\leq 0.10 \rightarrow$ 4, $\leq 0.25 \rightarrow$ 3, $\leq 0.50 \rightarrow$ 2, else 1; 5 thus means complex tools were reached for only once a conditioning set was in hand.
Signal-recovery-HA	Batch	mechanical	Per planted signal, the best-matching hypothesis of the run scores the DiscoveryBench triple $\text{ctxF1} \times \text{varF1} \times \text{relacc}$ - F1 of the routed variable and conditioning columns against the signal’s gold columns ($\text{ctxF1} = 1$ when it names none), and $\text{relacc} \in \{0, 1\}$ from direction-keyword overlap (positive, negative, reversal). HMS is the mean over the dataset’s signals, an untouched signal contributing 0; bands > 0.75 , 0.50, 0.25, 0 $\rightarrow$ 5, 4, 3, 2, else 1.
Spurious-rejection-HA	Batch	hybrid	One state per planted decoy column: named in the final summary, it is labelled <i>flagged-null</i> , <i>accepted-as-signal</i> or <i>tested-inconclusive</i> ; otherwise <i>probed</i> (present in tool arguments only) or <i>absent</i> . A decoy counts as correctly handled if it is flagged as a non-signal or kept out of the summary, and score = clamp($1 + \text{round}(4c/n)$, 1, 5) over n decoys. A dataset with no decoys scores 5.

Table 5: The fourteen high-level evaluation attributes, grouped by extraction method. Scope tags are *H* (per hypothesis), *HA* (per hypothesis, against its executed analysis) and *Batch* (once per run; the *-HA* suffix on the last two names is historical - both are computed at run level). The eleven atomised and mechanical rows are fixed formulae over the structured report rather than holistic judgements, so their reliability is a property of the construction rather than of a rater; only the three *rated* attributes ask for a judgement - each scored by haiku-4-5, gemini-3-flash and gpt-5-4-mini, whose mean is the reported (non-integer) score - and Appendix I reports a human-agreement study on exactly those three. Spurious-rejection is the one row of the last block with a model in the loop: its per-decoy states are mechanical, but the polarity of a decoy that reaches the final summary is read off by the same temperature-0 structured-output extractor that supplies the text atoms. The atoms behind the six atomised rows, and which of them are mechanical, are listed in Table 6.

Atom key	Parent	Method	Yes/no question answered
<i>Falsifiability-H</i>			
claim_is_assertion	Fal-H	LLM	Is the hypothesis a declarative assertion rather than an open-ended question or exploratory statement?
args_cols_in_disclosed	Fal-H	mech.	Do the tool calls name at least one column, and do all of the columns they name lie within the agent's disclosed schema?
result_is_typed	Fal-H	mech.	Does at least one error-free tool result contain a numeric (non-boolean) leaf anywhere in its payload?
threshold_or_method_present	Fal-H	hybrid	Does the text state a numeric threshold, comparison or named method, <i>or</i> does any tool call carry one of the method-selecting argument keys listed under method_named_in_args below?
<i>Reproducibility-H</i>			
tool_operates_on_population	Rep-H	mech.	Was at least one tool called, and none of the three tools tagged as operating on a sample rather than the population (quantiles, binning_then_mean, binning_chi_square)?
method_named_in_args	Rep-H	mech.	Does any tool call carry a method-selecting argument key (method, bins, qs, agg, agg_fn, formula, kind)?
threshold_present	Rep-H	LLM	Does the text state a numeric threshold, a comparison, or a named method (<i>exceeds 50</i> , $p < 0.05$, <i>Spearman</i>)?
no_subjective_hedge	Rep-H	LLM	Is the text free of subjective hedges (<i>seems, appears, looks, striking, roughly, possibly, perhaps</i>)?
<i>Binarizability-H</i>			
claim_has_threshold_or_boolean	Bin-H	LLM	Does the text contain a threshold or boolean comparator - a numeric comparison, <i>differs from, greater/less than, equals, does not</i> ?
result_type_is_scalar_or_bool	Bin-H	mech.	Does some error-free tool result reduce to a headline value - a bare scalar or boolean, a single-entry mapping, or a mapping carrying a key such as <i>r, p_value, statistic</i> or <i>coef</i> ?
args_specify_aggregation	Bin-H	mech.	Do the tool args invoke an aggregation (by, group_col, formula, agg, agg_fn), or is the tool itself one of the scalar-output tests?
no_open_ended_phrasing	Bin-H	LLM	Is the text free of open-ended phrasing (<i>explore, investigate, look into, examine ... relationship</i>) offered without a directional claim?
<i>Specificity-H</i>			
context_phrase_present	Spec-H	LLM	Does the text restrict the population - a prepositional restriction (<i>among ... , within the ... group, for ... = X</i>), a time window, or a named stratum?
schema_column_named	Spec-H	mech.	Does at least one disclosed column name occur in the text, matched case-insensitively on word boundaries?
relationship_verb_present	Spec-H	LLM	Does the text contain an associative or causal relationship verb (<i>associated with, correlated, predicts, causes, drives, leads to</i>)?
no_loose_quantifier	Spec-H	LLM	Is the text free of loose quantifiers (<i>some, many, often, sometimes, a few, tend to, roughly</i>)?
<i>Causal-vs-correlational-H</i>			
tool_supports_causal_claim	CvC-H	mech.	Is at least one tool called marked <i>supports-causal</i> in the tool metadata (the conditional correlations, the three diagnostics, anova, and the four regressions)?
conditioning_set_named	CvC-H	mech.	Do the args name a non-trivial conditioning set - a non-empty control_col, a by list of length ≥ 2 , or a formula whose right-hand side has ≥ 2 terms?
partial_or_stratified_test_used	CvC-H	mech.	Was one of the five partial-effect tools invoked (partial_correlation, stratified_correlation, vif, simpsons_diagnostic, regression_diagnostics)?
verb_matches_tool_capability	CvC-H	LLM	If the text uses a causal verb, do the tools called support causal inference? (True by default when the text is only correlational.)
<i>Grounded-analysis-HA</i>			
tool_family_matches_relationship_type	GA-HA	LLM	Does the tool family match the relationship type implied by the text - a correlation claim with corr_matrix, a group difference with ttest, but not a causal claim with corr_matrix alone?
args_columns_match_claim_columns	GA-HA	mech.	Do all disclosed columns named in the text also appear in the tool arguments? (True by default when the text names none.)
result_addresses_claim_threshold	GA-HA	mech.	If the text states a threshold, does some tool result carry a numeric value that resolves it? (True by default when it states none.)
analysis_does_not_drift	GA-HA	mech.	Do the later hypotheses of the same turn that share an argument column with this one keep at least one tool angle in common with it?

Table 6: The twenty-four atoms, grouped by their parent atomised attribute. Each attribute aggregates its four atoms via $\text{score} = \text{clamp}(1 + \sum_{i=1}^4 a_i, 1, 5)$ (Equation (1)). The method tag says how each boolean is obtained. *mech.* (13 atoms): a set-membership, type, or literal-string check on the schema-validated JSON report - arguments, results, and the disclosed column list, the last also matched against the hypothesis text for schema_column_named - with no model in the loop. *LLM* (10 atoms): a boolean field returned by a shared Haiku-4.5 structured-output extractor at temperature 0, which reads the hypothesis text and the names of the tools called and returns all ten fields in one call per hypothesis, so no atom is produced by free generation. *hybrid* (1 atom): the logical OR of a text-side LLM field with a mechanical argument-key check, firing when either source supplies the evidence; the field it reads is the one also reported as threshold_present under Reproducibility-H. Atoms with a precondition - the last three of Grounded-analysis-HA and verb_matches_tool_capability - return true when it does not fire. Fourteen of the twenty-four - the mechanical and hybrid atoms - therefore require no judgement of any kind (Appendix I).

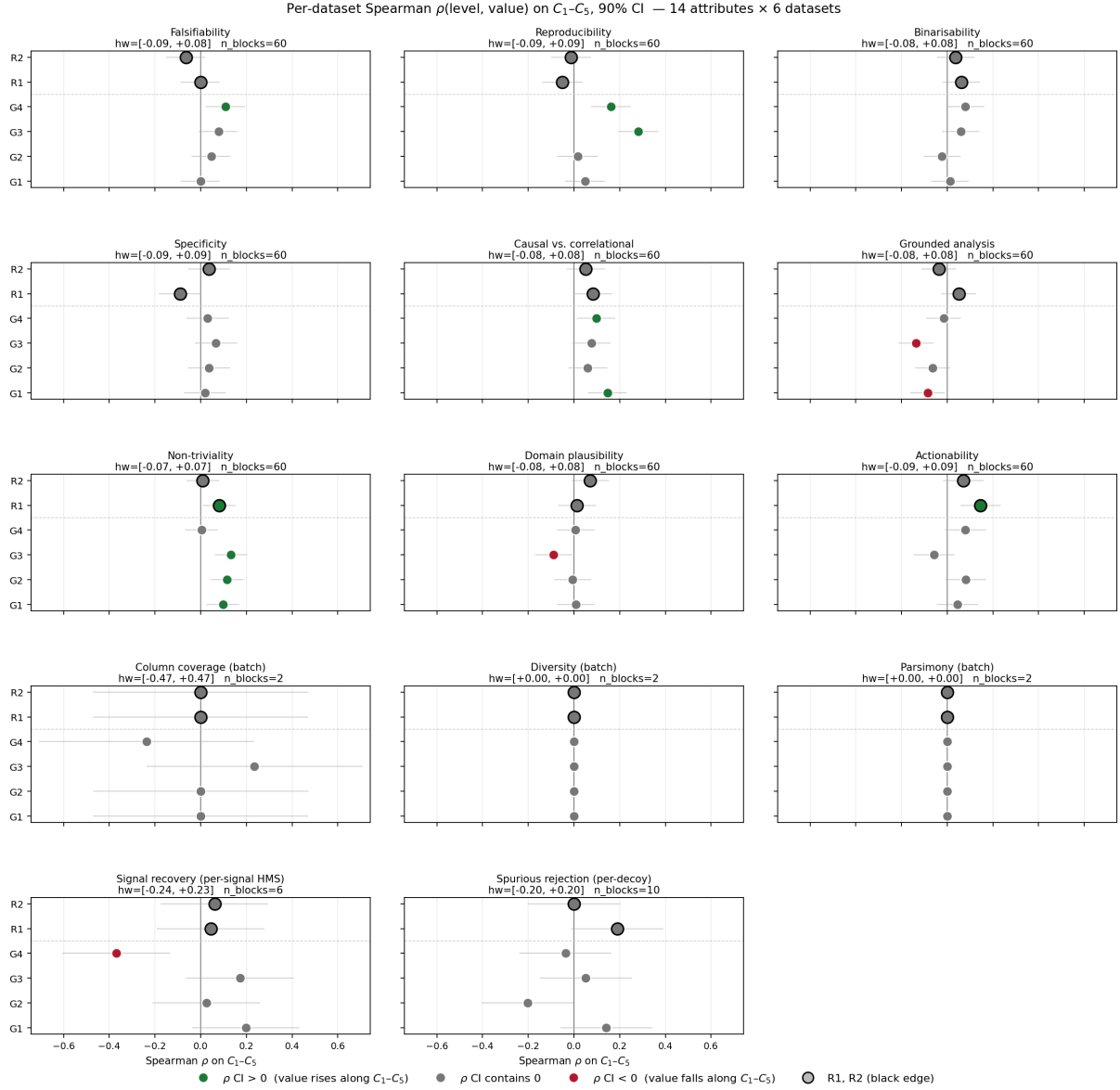


Figure 20: Per-model Spearman rank correlation between disclosure level and each rubric attribute, showing which attributes increase, decrease, or remain flat as more information is disclosed.

Attribute	GPT-5.4-mini	GPT-5.5	Gemini 3 Flash	Gemini 3.1 Pro	Haiku 4.5	Opus 4.7	Overall
Grounded analysis	-0.36	<i>-0.14</i>	-0.66	-0.41	<i>-0.05</i>	-0.54	-0.36
Domain plausibility	+0.21	-0.61	<i>-0.10</i>	+0.13	<i>-0.25</i>	-0.56	<i>-0.19</i>
Specificity	-0.41	+0.07	-0.45	+0.07	+0.03	+0.20	<i>-0.08</i>
Falsifiability	<i>-0.17</i>	+0.08	+0.10	+0.17	+0.16	+0.36	+0.11
Reproducibility	-0.33	+0.37	<i>-0.16</i>	+0.40	+0.37	+0.25	+0.15
Causal vs. correlational	<i>-0.04</i>	+0.36	+0.05	+0.57	-0.31	+0.39	+0.17
Non-triviality	+0.22	+0.18	+0.20	+0.45	<i>-0.26</i>	+0.32	+0.19
Actionability	+0.34	<i>-0.07</i>	+0.06	+0.51	+0.21	+0.33	+0.23
Binarisability	+0.28	+0.21	+0.02	<i>-0.02</i>	+0.44	+0.44	+0.23

Table 7: Per-model mean within-block Spearman ρ between ordinal disclosure level and mean attribute score, averaged over the four datasets. Rows are sorted by the *Overall* column (left to right: lowest to highest ρ , matching the forest plot in Figure 4). Negative values are italicised; values with $|\rho| \geq 0.30$ are bold. *Overall* is the unweighted mean across all six models.

Attribute	G1 (linear)	G2 (Simpson’s)	G3 (mediation)	G4 (nonlinear)	Overall
Grounded analysis	-0.47	-0.29	-0.21	-0.47	-0.36
Domain plausibility	+0.07	-0.44	-0.27	-0.14	-0.19
Specificity	+0.10	-0.20	-0.23	-0.00	-0.08
Falsifiability	+0.02	-0.05	+0.37	+0.12	+0.11
Reproducibility	+0.06	-0.05	+0.48	+0.12	+0.15
Causal vs. correlational	+0.15	+0.18	-0.10	+0.46	+0.17
Non-triviality	+0.39	-0.06	+0.14	+0.28	+0.19
Actionability	+0.35	+0.16	-0.02	+0.42	+0.23
Binarisability	+0.21	-0.11	+0.30	+0.52	+0.23

Table 8: Per-dataset mean within-block Spearman ρ between ordinal disclosure level and mean attribute score, averaged over the six models. Row order and formatting conventions are identical to Table 7. *Overall* is the unweighted mean across all four datasets.

Attribute	inter-human (153 pairs)	inter-LLM (3 pairs)	human-LLM (54 pairs)
<i>Krippendorff α, ordinal: mean over rater pairs, [min, max] across pairs</i>			
Non-triviality	-0.041 [-0.58, 0.72]	0.275 [0.13, 0.44]	0.046 [-0.61, 0.63]
Domain plausibility	0.027 [-0.43, 0.64]	0.286 [0.22, 0.35]	0.105 [-0.32, 0.70]
Actionability	-0.068 [-0.58, 0.60]	0.151 [-0.02, 0.36]	-0.033 [-0.61, 0.56]
Pooled	-0.027	0.238	0.039
<i>Spearman ρ: rank agreement, insensitive to a constant offset</i>			
Non-triviality	0.039 [-0.55, 0.72]	0.513 [0.48, 0.53]	0.168 [-0.23, 0.71]
Domain plausibility	0.061 [-0.45, 0.65]	0.444 [0.35, 0.55]	0.156 [-0.22, 0.80]
Actionability	0.021 [-0.47, 0.63]	0.489 [0.41, 0.54]	0.147 [-0.29, 0.62]
Pooled	0.040	0.482	0.157

Table 9: **Pairwise rater agreement on the three rated attributes.** Each cell is the mean statistic over all eligible rater pairs drawn one from each named group, computed on the items both members rated; a pair must share at least 20 of the 60 items to be eligible. *Pooled* averages every defined pair value across the three attributes. Inter-LLM agreement exceeds inter-human agreement on every attribute and on both statistics, under identical rubric text.

Attribute	α	ρ	QWK	human-LLM
Non-triviality	0.388	0.620	0.367	+0.360
Domain plausibility	0.535	0.609	0.405	-0.101
Actionability	0.140	0.560	0.201	+0.570

Table 10: **Aggregate human-LLM agreement** over the $n = 60$ items. Each side is collapsed to one series first: the human score is the mean over every annotator who rated the item, the LLM score the mean over the three raters. α is ordinal Krippendorff’s; QWK rounds both series to the nearest integer (halves up); the last column is the mean signed difference, a calibration offset rather than an error.