FROM PERFORMANCE TO CAPABILITY: A THREE-LEVEL FRAMEWORK FOR BEHAVIORAL LLM CAPABILITY SCIENCE WITH LARGE-SCALE EMPIRICAL STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

Behavioral LLM capability science asks how evidence from observable behavior can justify increasingly strong claims about what an LLM system can do and how its behavior is organized. We propose a three-level framework that distinguishes (1) Observable Performance, (2) Conditional Response and Capability Validity, and (3) Fundamental Organization, and makes the evidential transitions between them explicit and testable. We evaluate this framework across 34 heterogeneous released source collections spanning reasoning, feedback, memory, prompting, planning, coding, retrieval, and related settings, comprising over one million task-level responses. Conditional Response emerges as a substantive empirical object: response profiles reproduce across compatible roots (rank correlation 0.586), 73/311 eligible groups exhibit held-root nonconstant regularities, and multiple designs show non-additive effects. Stronger capability claims require connected evidence involving reliability, transfer, convergence across operationalizations and sources, specificity, discrimination, measurement quality, controls, external prediction, and compatible common support. Four of nine candidate constructs obtain local reliability and task-transfer support, and one obtains cross-operationalization support, but none completes connected validity because key links in the released evidence topology remain missing or conflicting. Human-readable semantic and conditional structure is predictively useful on held-out data, yet predictive or low-dimensional structure alone does not establish Fundamental Organization. The resulting evidence frontier distinguishes observed performance, established response organization, supported capability evidence, and unresolved deeper organization, providing a concrete foundation for cumulative LLM capability science.

1 INTRODUCTION

What permits an observable behavior to be called a capability? A benchmark score describes performance; a change under added information or feedback describes a response; a compact representation raises a question about common organization. Establishing one does not establish the next. We define *Behavioral LLM Capability Science* as the study of what externally observable behavior, together with controlled or systematic variation, warrants attributing to an LLM system. The relevant entity is the deployed system, including model version, scaffold, and operating regime. This behavioral scope complements the study of internal mechanisms and training dynamics.

Performance evaluation, behavioral testing, construct validity, and latent modeling have developed rich but differently focused traditions (Section 3). Their achievements leave a central theoretical problem: which scientific objects do they concern, how are those objects related, and which unresolved transitions determine progress beyond performance? We propose a three-level framework that answers these questions explicitly. It defines response as an object distinct from performance, specifies when response structure warrants capability attribution, and formulates the stronger hypothesis of Fundamental Organization. Connected validity and five organizational questions turn those relations into an empirical research program rather than a taxonomy of labels.

We then independently analyze a large collection of released measurements to investigate those questions without new LLM execution. Our contributions are: (i) the framework and its cross-level

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.

question architecture; (ii) reproducible, locally lawful and non-additive Conditional Response; (iii) a capability-validity frontier that preserves local and semantic predictive success while identifying missing evidence connections and a discriminant conflict; and (iv) five tests distinguishing predictive organization from fundamentality. To our knowledge, this is the first unified three-level empirical framework for behavioral LLM capability science to distinguish these objects and evaluate their transitions in one large-scale evidence program. The resulting frontier concerns the assembled collection, without claiming to exhaust all existing assets or possible analyses.

2 A THREE-LEVEL FRAMEWORK

Level 1: Observable Performance. Let Y_{src} be a direction-aligned outcome for deployed system s, canonical task root r, and condition c. Level 1 concerns its distribution or a specified summary. Mathematical accuracy or coding correctness measures what a system does well; a domain label does not identify whether information use, reasoning, familiarity, context handling, or their interaction enabled success. Mature performance measurement is indispensable, while its capability interpretation remains a separate question.

Level 2: Conditional Response and Capability Validity. Demands and conditions specify what a task requires or supplies: information, feedback, memory, context, constraints, horizon, reasoning perturbations, or surface form. These begin as design variables. Their variation exposes *conditional response*; in a same-root paired design,

$$R_{sr}(c, c_0) = Y_{src} - Y_{src_0}. \tag{1}$$

The changed condition and source-defined baseline c_0 are part of the estimand. A system's profile comprises these design-qualified responses. Similar average performance can coexist with different profiles; high performance need not imply a large benefit from added information. Factorial departures from additivity, structured-unpaired population relationships, and baseline-free surface dispersion retain their own estimands. Equation 1 does not turn unpaired relationships into same-task causal effects.

The inferential sequence is demand/condition variation $\rightarrow$ response $\rightarrow$ stable response structure $\rightarrow$ candidate construct $\rightarrow$ validated capability (Figure 1). It generates substantive questions: does response reproduce across compatible roots and transfer to held-out tasks; does intensity produce lawful variation with stable functional form; do conditions interact; and does ordering survive condition changes? A *candidate construct* is a hypothesis about what organizes this structure, not an already discovered capability.

A *validated capability*, under the interpretation tested here, requires connected evidence for reliability, task transfer, convergence across operationalizations and sources, specificity against matched nuisance and neighboring demands, minimality under ablation or merging, discrimination from neighbors, measurement quality, independent controls, external prediction, and compatible common support. Convergence asks whether different measurements support the same interpretation. Specificity separates the candidate from generic sensitivity; minimality asks whether deleting, merging or coarsening it loses predictive information. Discrimination requires positive evidence of distinction, rather than failure to detect correlation. Controls challenge alternative explanations, while external prediction tests the interpretation against independent same-target outcomes. Large mean response cannot substitute for these relations. These requirements operationalize the proposed interpretation rather than define every legitimate use of "capability."

Level 3: Fundamental Organization. Do many responses admit a smaller, stable, global and transferable organization? Local, modular and global explanations can predict the same limited panel; low-dimensional prediction alone cannot decide among them. Five questions define the stronger claim: *recoverability*, distinguishing known structural worlds under the observed topology; *stable geometry*, persisting under root uncertainty, missingness and held-out cells/intervention blocks; *structural competition*, separating competing explanations under matched information and loss; *exact cross-domain transfer*, predicting target-domain responses for linked deployed systems; and *diversity scaling*, gaining stability or recoverability as within-design response and intervention diversity grows. Organizational hypotheses can be tested before every semantic candidate is validated, but the interpretation then concerns responses rather than an established capability ontology.

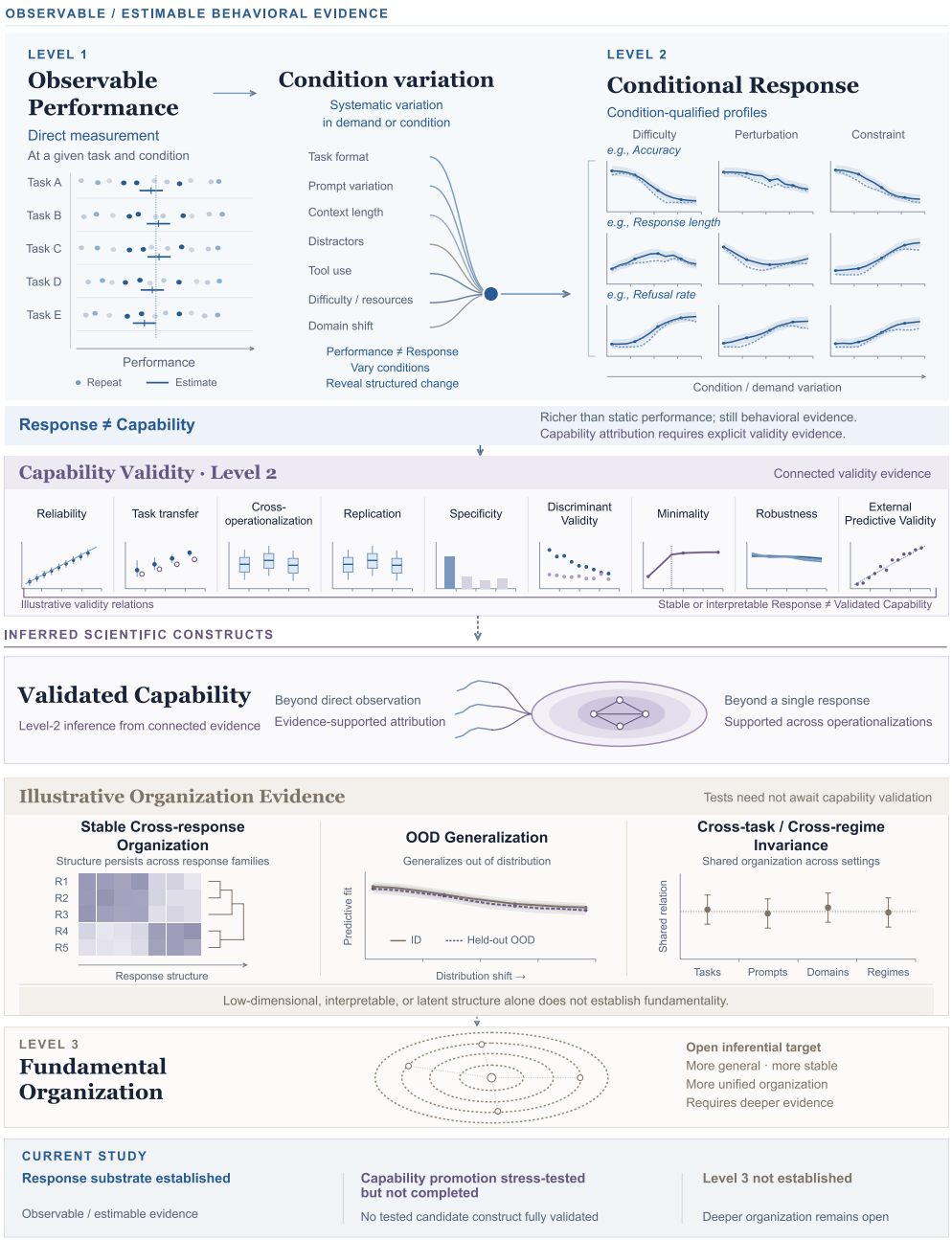

Figure 1: A cross-level framework for behavioral capability inference. Level 2 includes Conditional Response and Capability Validity. Mini-plots illustrate relations, not measurements or exhaustive criteria. Capability attribution requires the connected contract in Section 2, including measurement quality, controls, and common support. Response organization can be tested before capability validation; its five questions are defined in Section 2 and summarized in Figure 4D. Low-dimensional structure alone does not establish fundamentality.

Evidence topology and maturity. Evidence topology is the connection pattern among candidates, measurements, systems, operationalizations, sources, controls and external outcomes. More rows strengthen existing links without creating missing ones. Validity cannot be assembled from favorable criteria on incompatible populations.

We distinguish *supported*, a criterion met in its stated population; *estimable but not supported* (ENS), an identifiable target failing its positive criterion; *not identifiable* (NI), unavailable identifying infor-

mation or unresolved discrimination between explanations; and *conflicting*, evidence contradicting a required validity condition. Neither NI nor ENS proves absence. These states locate empirical transitions without making the three levels a presupposed natural ontology.

3 Related work: explanatory coverage of the framework

The framework connects four research traditions through their scientific objects. GLUE, SuperGLUE, MMLU, BIG-Bench and HELM establish mature Level-1 infrastructure: broad task coverage, standardized measurements and comparisons across systems (Wang et al., 2019a;b; Hendrycks et al., 2021; Srivastava et al., 2023; Liang et al., 2023). Their achievement is performance science; the further capability interpretation requires evidence about what produces or organizes those observations.

Behavioral testing provides important Level-2 components. CheckList organizes controlled tests (Ribeiro et al., 2020); counterfactual reasoning, shared invariances and benchmark sensitivity reveal behavior under altered rules or surfaces (Wu et al., 2024; Rawal & Toneva, 2024; Alzahrani et al., 2024). FLASK, Skill-Mix, CHAMP, CrossEval and SCAN supply finer skills, combinations and taxonomies (Ye et al., 2024; Yu et al., 2024; Mao et al., 2024; Zhong et al., 2025; Wang et al., 2026b). In our map, these descriptions specify demands and operations, while measured variation exposes response. This explains their shared contribution beyond aggregate accuracy and isolates a further question: when does reproducible response justify a stable capability interpretation?

Construct-validity theory supplies foundations for that transition through convergent and discriminant relations and arguments connecting scores to meaning (Cronbach & Meehl, 1955; Campbell & Fiske, 1959; Messick, 1995; Kane, 2013). Harding and Sharadin analyze capability attribution (Harding & Sharadin, 2024); Bean et al. document benchmark validity problems (Bean et al., 2025); Jo and Wilson explicitly derive capability inference from theory (Jo & Wilson, 2026); Mitchell emphasizes controls, variation and generalization (Mitchell, 2026); Freiesleben and Kearns develop networked and quantitative validity approaches (Freiesleben, 2026; Kearns, 2025). Our contribution extends these foundations to connected, candidate-level validity of conditional responses. Neither performance–capability distinctions nor the need for multiple validity relations originate here.

Latent and interpretable organization motivates Level-3-like questions. Factor analysis reveals performance structure (Burnell et al., 2023); observational scaling and Sloth predict through low-dimensional skills (Ruan et al., 2024; Polo et al., 2025); hierarchical latent modeling studies dependencies (Jin et al., 2026); BenchMIRT examines item-level latent signals (Allen Institute for AI, 2026). General Scales is a close positive precedent: interpretable demand dimensions explain and predict performance across tasks (Zhou et al., 2026a). These achievements concern informative organization of performance. Whether conditional responses support recoverable, stable, competitively explanatory and transferable organization is a distinct inference, captured by our five joint questions.

Thus the field already supplies substantial components at every level. The missing synthesis is a common map of their objects and relations that makes the transitions themselves testable. Our theoretical contribution specifies that map, the connected validity architecture and the organizational questions; our empirical contribution adjudicates them together. Explanatory coverage of these traditions supports the framework's usefulness without establishing it as the uniquely correct account of the field.

4 Empirical design

Evidence and measurement. We integrate 34 released source collections; 16 supply core reconstruction with 10,446,502 normalized observations, 1,000,452 task-level responses and 6,798 response cells. Other sources supply controls, external targets, observational structure or request ecology. These are processing counts, not independent sample sizes. Canonical task roots are the principal uncertainty unit. We use released measurements without new inference, training or benchmark execution. Valid zeros and evaluator-confirmed model failures remain outcomes; measurement/environment failures and missing observations do not become zeros. Paired responses require common roots; unpaired and surface designs retain their distinct meanings. Source-explicit stochastic repeats propagate where identifiable; singleton cells have unavailable standard errors. Objective outcomes and supporting proxies retain separate roles (Appendix A).

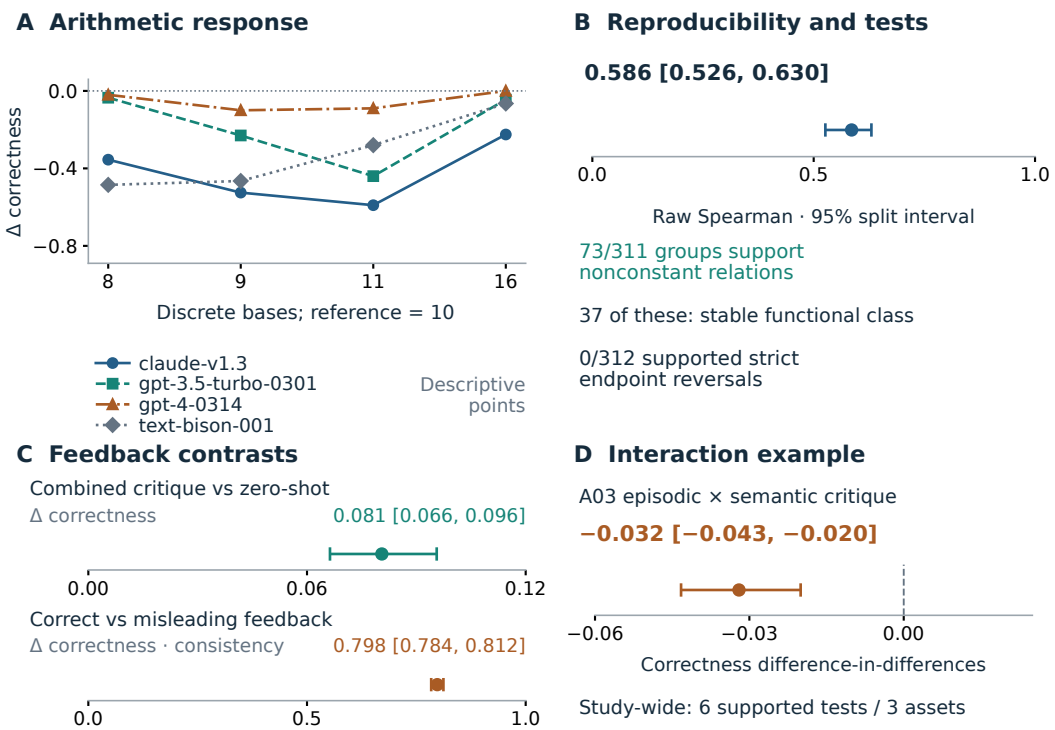

Figure 2: Reproducible response, local structure, and supported interactions. (A) All four systems in the illustrated A01 arithmetic configuration (ccc, unknown-CoT, zero-shot), relative to base 10; connectors are not fitted laws. Aggregate tests have separate eligible populations. (B) Asset-equal raw-Spearman summary with a 95% repeated-split interval; counts are separate tests. (C) A03 correctness contrasts on separate scales. (D) An A03 interaction, separate from study-wide counts. C–D: root-cluster 95% intervals.

Response and validity. Reproducibility uses 500 outcome-unconditioned common-root splits with raw Spearman correlation; transfer holds out roots. Local laws must improve exact leave-one-root-out prediction over a constant, with root uncertainty, coherent empirical-null tests and global multiplicity correction. Interactions and endpoint reversals use design-specific common-root inference. Nine prespecified candidates use positive/supporting memberships for positive evidence; nuisance/control roles can challenge but cannot promote. Validity axes are re-estimated on candidate-specific compatible cohorts. Cross-source and external tests require exact model/version and scaffold/regime identity, not names or families. Synthetic observations pass through the full production response and validity procedure; disjoint worlds select thresholds and independently qualify its null rejection and strong-signal detection. Released topology cannot choose thresholds.

Representation and organization. Semantic sparse features and masked low-rank prediction share held-out rows, system/source-context information, outer root assignments and training-only tuning. A sensitivity also matches training cells, weights and objective. Strength comparisons exclude target roots when estimating strength. Level 3 uses topology-faithful known-truth worlds, root-bootstrap geometry, held-cell and held intervention-family × held-system prediction, matched competitors, exact domain bridges and refitted within-design diversity subsets. Every required criterion must be supported for the joint claim. Across the registered 78-row robustness summary covering nine claim families, full family-exclusion and scale refits and evaluation reweighting change no major conclusion; effect sizes and support counts need not be invariant. Full methods, measurement-quality/taxonomy variants and robustness matrices remain in Appendices A and H.

5 Conditional Response is a Structured Scientific Object

Response profiles reproduce across roots: across 1,009 eligible demand groups from nine assets, with at least five systems per group, asset-equal median raw Spearman correlation is 0.586 [0.526,

0.630]. The 95% interval describes repeated partitions of the released population. Of 311 eligible system-specific law groups, 73 across four assets exhibit held-root nonconstant relationships, and 37 additionally support a stable exact functional class. These include both paired response laws and explicitly structured-unpaired horizon relations; they establish local lawful variation, not universal causal coefficients. Figure 2A illustrates how the same systems respond differently to changed arithmetic rules.

Joint conditions reveal non-additivity: six multiplicity-adjusted common-root interaction tests are supported across three assets. The A03 episodic-by-semantic critique interaction is -0.0320 $[-0.0433, -0.0200]$; the negative departure from additivity does not mean combined feedback is harmful. In that six-system panel, helpful combined critique improves over zero-shot by 0.08053 [0.06630, 0.09559], while correct versus misleading consistency feedback differs by 0.79781 [0.78400, 0.81190]. Feedback type therefore changes the response question and consequence, rather than contributing one universal coefficient.

Surface sensitivity supplies a necessary interpretive contrast: observed within-root surface dispersion averages 0.24699 across PromptSuite and DOVE-Lite, so large response need not identify a capability-relevant demand. Meanwhile, 0/312 strict endpoint-reversal tests are supported. This establishes neither widespread robust reversals nor universal system ordering. Reproducibility, laws and interactions nevertheless establish Conditional Response as substantive Level-2 science; subsequent validity boundaries do not erase it. Appendix D retains the complete catalogs and source-specific qualifications.

6 From response structure to capability validity

Four of nine candidates have supported local reproducibility and held-root transfer; information utilization additionally supports cross-operationalization convergence (Table 1). Reasoning integrity and misleading-information resistance also have supporting predictive-taxonomic specificity. Thus labels capture local structure, and some organization survives changed operations. The next question is whether the validity dimensions converge on a connected capability interpretation.

Table 1: Candidate maturity in the production validity evaluation. Local support requires both rank reproducibility and held-root transfer. NI denotes missing identifying evidence, not absence of a capability. The full criterion matrix is in the appendix.

Candidate	Furthest supported local evidence	Overall validity state
Semantic adaptation	Local response and task transfer	Not identifiable
Reasoning integrity	Local response and task transfer	Conflicting: discrimination
Planning / procedural control	Local validity not identifiable	Not identifiable
Information utilization	Cross-operationalization	Not identifiable
Feedback-guided repair	Local validity not identifiable	Not identifiable
Misleading-information resistance	Local response and task transfer	Not identifiable
Context-load robustness	Local validity not identifiable	Not identifiable
Constraint handling	Local validity not identifiable	Not identifiable
Surface robustness	Local validity not identifiable	Not identifiable

They do not yet converge fully: zero candidates satisfy the complete contract, eight are NI, and reasoning integrity is conflicting. Matched specificity is NI for all nine: predictive usefulness of labels cannot replace same-source/system/root comparisons against registered neighbors and nuisance. Minimality is estimable but unsupported for six and NI for three. Reasoning integrity has a sharper discriminant conflict: the adjusted neighbor-correlation interval exceeds the permitted $|\rho| = 0.8$ boundary. The semantic-adaptation/reasoning-integrity diagnostic has $\rho = 1$ over six common systems, with interval [1,1]; final states use candidate-specific populations rather than mechanically assigning this conflict to both candidates.

Evidence topology limits maturity independently of row count. All five required control sources have usable data: 46,735 rows and 40 model labels. Yet zero candidate-control predictive links are estimable under exact deployed-system identity; this is not evidence that controls are clean. Same-target external validity has maximum exact overlap zero, and cross-source evidence supplies

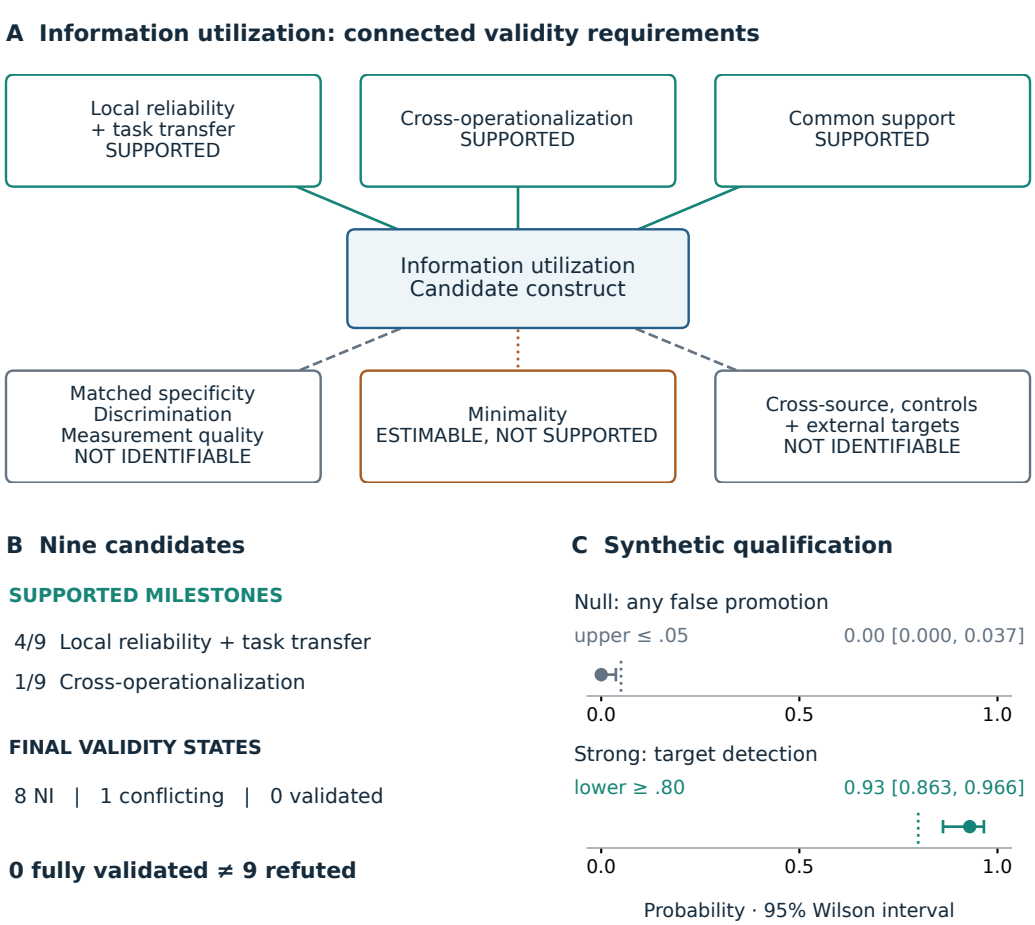

Figure 3: Local evidence does not complete capability validity. (A) Grouped requirements for information utilization: solid links supported, dashed NI, dotted ENS. Common support is candidate-specific. (B) Supported milestones and final validity states are separate summaries. (C) Independent synthetic qualification, 100 worlds per signal, with 95% Wilson intervals; thresholds apply to the null upper and strong lower bounds. Qualification cannot supply missing released-data bridges.

no supported candidate conclusion. Figure 3 shows how local convergence coexists with these missing connections. More observations within disconnected sources can improve precision without identifying the missing relation. Appendix E preserves the full validity matrix.

The procedure has power in identifiable worlds. On untouched synthetic qualification worlds, null false promotion is 0 (Wilson upper 0.03699), and mean strong-signal detection is 0.93 (Wilson lower 0.86250); calibration veto is false. Eighteen conflict worlds have zero fail-open cases. The procedure can therefore promote identifiable strong signals; zero released-data promotions cannot be explained simply by a rule that never promotes. This qualification does not establish equal power for every real candidate or weaker signal. The observed frontier combines missing bridges, unsupported minimality and a specific conflict, while preserving the positive local science.

7 What organizes response beyond overall performance?

Human semantics predicts held-out responses. On primary matched task-OOD, semantic-minus-masked-low-rank MSE is -0.01150 (95% repeated-split interval $[-0.01208, -0.01105]$), favoring semantics (Figure 4A). The equal-cell matched-objective sensitivity retains the direction: -0.00128 $[-0.00162, -0.00103]$. Human-readable structure thus generalizes beyond fitted roots and outperforms the evaluated compressed comparator. Other OOD targets remove different information and do

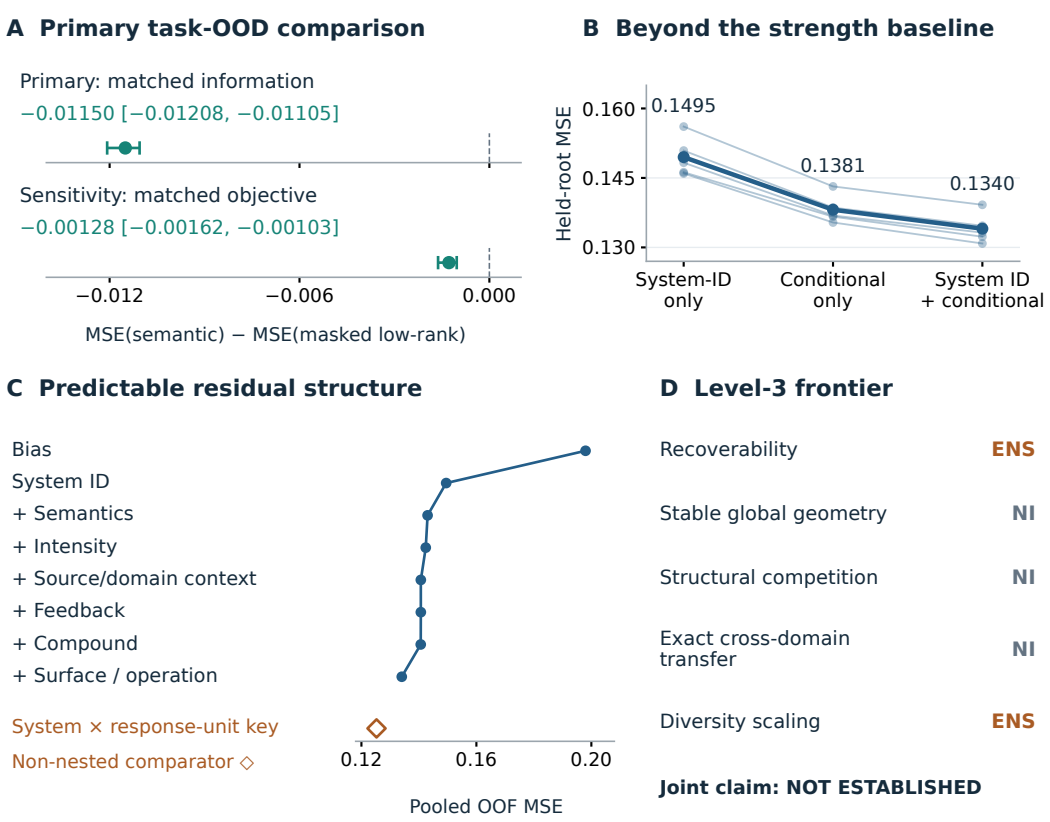

Figure 4: Predictive organization without established fundamentality. (A) Task-OOD semantic-minus-masked-low-rank MSE: asset-equal primary and equal-cell matched-objective sensitivity; 95% repeated-split intervals, negative favors semantics. (B) Five folds (thin), equal-fold mean (thick); strength is the cross-fitted system-ID-only predictor. (C) Ordered additions, pooled OOF-row MSE; diamond: non-nested system–response-unit key comparator. (D) ENS: estimable but unsupported; NI: not identifiable. All five criteria must be supported for the joint claim.

not imply a universal representation ranking. Predictive semantics remains distinct from necessary, discriminant and externally validated semantic axes.

Conditional structure adds to general strength. Five-fold held-root MSE is approximately 0.1495 for cross-fitted strength alone, 0.1381 for conditional structure without strength, and 0.1340 for both; the ordering holds in every fold (Figure 4B). Conditional Response is therefore more than a recoding of the tested overall-strength predictor. This comparison addresses the released task population, without excluding every conceivable strength model.

Predictable residual structure remains. The additive OOF path improves from bias-only MSE 0.198 to 0.149 with system identity, 0.143 with semantics, and approximately 0.141 with intensity and source/domain context; surface/operationalization brings it to 0.134. Feedback and compound blocks add little in this sequence. A separately fitted system–response-unit key comparator reaches 0.125 (Figure 4C), showing that the measured additive blocks leave predictable detail. It is a non-nested comparator, not another additive stage. Feature order changes incremental attribution: these are predictive gains, not a causal variance partition. Rich conditional organization motivates, but does not establish, a global explanation (Appendix G).

8 The Frontier of Fundamental Organization

The five questions produce the distinct vector (ENS, NI, NI, NI, ENS), leaving the joint claim *not established* (Figure 4D). Each state has a scientific reason.

Recoverability: ENS. Topology-faithful strong-signal structural recovery is 0.2917 (Wilson interval [0.1824, 0.4318]), with ambiguity 0.4167; recovery, ambiguity and negative-world bounds fail the joint criterion. Unlike candidate calibration in a connected identifiable topology, this test must distinguish competing structural worlds under the released response topology. **Stable global geometry: NI.** Of 15 within-asset panels, two identify the full test and one supports it. The aggregate requires at least three identifiable and three supported panels with 75% support, including root stability and both held-cell and held intervention-family $\times$ held-system prediction. Local geometric support survives the broader boundary.

Structural competition: NI. Hybrid global-plus-modular structure has the lowest mean MSE, 0.01764, and decisively improves over five alternatives, but does not separate from local-condition prediction. Useful hybrid structure leaves the explanatory competition unresolved. **Exact cross-domain transfer: NI.** None of 56 directional pairs is identifiable; maximum exact deployed-system overlap is zero. Family/name matching would change the target. **Diversity scaling: ENS.** Refitted 25/50/75/100% subsets improve subspace similarity by 0.150 [0.080, 0.223] along response-unit diversity and 0.168 [0.074, 0.335] along intervention-family diversity, but neither meets the joint requirements including rank stability and positive full-fraction prediction. Thus additional within-design diversity does not establish the required organizational improvement. Appendix K supplies full specifications and diagnostics.

9 Discussion: a research program for capability science

The framework makes uneven scientific maturity intelligible. Level 1 has mature measurement infrastructure; Level 2 supports reproducible, lawful, interaction-rich response science, while connected capability validity remains incomplete in this collection. Level-3-like prediction is informative before fundamental organization is established. These are different achievements with different unresolved questions. Human-readable semantics can be useful without already being a validated ontology; global prediction can improve without uniquely displacing local explanations. The framework provides objects and relations with which to formulate these distinctions and design the next empirical transitions.

Evidence topology makes the next design concrete. Share exact model versions and documented scaffolds across domains; measure multiple candidate operationalizations and matched neighboring/nuisance demands on compatible roots and scales; connect those systems to independent controls and same-target external outcomes. Preserve evaluator semantics, failure statuses and stochastic-repeat identifiers. Organizational tests additionally need shared systems across intervention families, genuine held-block transfer targets, topology-matched known-truth calibration and repeated within-design diversity refits. These connections change what can be identified, beyond increasing precision within an existing source.

Our independent released-asset program exposes a substantial empirical frontier without claiming a census of the field or exhaustion of its analytical potential. Conclusions remain conditional on the assembled systems, operations, populations and measurement fidelity; proxy outcomes remain proxies. The appendix preserves observational positives in BenchMIRT/ADeLe, nonsupport in MedHELM/OpenEval and NI in HELM, alongside response atlases, contextual modulation, self-repair proxies and request ecology. These distinct objects broaden the scientific picture without supplying missing capability-validity links.

10 Conclusion

The three-level framework turns performance, response, capability validity and fundamental organization into distinct objects with testable relations. Large-scale released evidence establishes rich Conditional Response and semantic predictive information while locating specific validity and organizational boundaries. Further progress requires connected capability-science designs that can support the next inference.

AI USE STATEMENT

The research idea and conceptual framework were developed without generative AI assistance. Generative AI tools (ChatGPT and Codex) subsequently assisted throughout the research and writing process, including retrieval and discovery of relevant literature and related research materials, research methodology and experimental design, implementation, analysis and interpretation of experimental results, and manuscript preparation, including drafting and editing. AI-assisted retrieval and discovery were used to identify potentially relevant materials, which were independently checked against the original sources before being incorporated into the research. All AI-assisted work incorporated into the final submission was subject to human review. The authors take responsibility for the final content of this work, including its text, claims, analyses, citations, and artifacts.

ETHICS STATEMENT

This study analyzes previously released model outputs, trajectories, benchmark results, evaluation records, and related research artifacts. It did not recruit human participants, conduct new human-subject interventions, or perform new LLM, VLM, or agent inference or deep-learning training. All datasets and released research artifacts used in this study were accessed and used in accordance with the official access requirements, licenses or terms of use, and usage and redistribution conditions applicable to each source. Where access was gated or otherwise restricted, the corresponding requirements were followed. Restricted raw data, access credentials, and source data whose redistribution is not permitted are not redistributed in our supplementary material.

Some source collections contain human-origin interactions or concern medical, safety, bias, fairness, or other potentially sensitive settings. We use only released information required for the reported analyses, do not attempt to identify or re-identify individuals, and do not use the source materials for purposes outside the scope permitted by their providers. Results are reported as aggregate or derived analyses, and our conclusions remain explicitly limited to the released populations, systems, measurements, operationalizations, and evaluator fidelity represented by the source materials. Any limitations or biases present in the original datasets and evaluation procedures may therefore also affect the corresponding analyses in this study.

To support responsible reproducibility, the supplementary material documents the provenance, pinned versions where available, required inputs, access conditions, and organization procedures for the source collections without relaxing the restrictions imposed by the original data providers. The authors have read and adhered to the ICLR Code of Ethics and take responsibility for the ethical use, analysis, and reporting of all source materials in this work.

REPRODUCIBILITY STATEMENT

We provide an anonymous reproduction package containing the experimental and analysis code, dependency and environment specifications, configuration files, and detailed instructions required to reproduce the empirical results reported in this paper. The study is based entirely on previously released model outputs, trajectories, evaluation results, and related research artifacts, and requires no new LLM, VLM, or agent inference and no deep-learning training. For the 34 source assets used in the study, the package documents the required data, their provenance and versions where applicable, and the procedures for obtaining and organizing them for reproduction. Randomized statistical procedures use fixed seeds and explicitly specified analysis settings. Together with the methodological and statistical details provided in the paper and supplementary material, these resources are intended to enable independent reproduction of the reported empirical analyses and results from the released source artifacts.

REFERENCES

Allen Institute for AI. BenchMIRT: Disentangling Safety and General Capabilities in LLM Evaluation. Technical report, Allen Institute for AI, 2026. URL https://allenai.org/papers/benchmirt. Prepared by Malia Morgan, Vasudha Varadarajan, Faeze Brahman, and Maarten Sap.

Norah A. Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf M. Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. When Benchmarks are Targets: Revealing the Sensitivity of Large Language Model Leaderboards. *Proceedings of ACL*, 2024. URL https://aclanthology.org/2024.acl-long.744/.

Anonymous. Replication Package for Probing Coding Benchmark Memorization in Black-box LLMs with Vague Inputs, 2026a. URL https://zenodo.org/records/20823175.

Anonymous. Replication Package for Fairness Bugs Detection Empirical Study, 2026b. URL https://zenodo.org/records/21032824.

Anonymous. SEAL Release V1: Semantically Altered Coding Tasks and Model Evaluations for LLM Code Generation, 2026c. URL https://zenodo.org/records/21058658.

Ashwath Vaithinathan Aravindan and Mayank Kejriwal. Fragile Thoughts: How Large Language Models Handle Chain-of-Thought Perturbations, 2026. URL https://arxiv.org/abs/2603.03332.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program Synthesis with Large Language Models. arXiv:2108.07732, 2021. URL https://arxiv.org/abs/2108.07732.

Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan Eghlidi, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, María Grandury, Sophia Han, Valentin Hofmann, Lujain Ibrahim, Hazel Kim, Hannah Rose Kirk, Fangru Lin, Gabrielle Liu, Lennart Luettgau, Jabez Magomere, Jonathan Rystrøm, Anna Sotnikova, Yushi Yang, Yilun Zhao, Adel Bibi, Antoine Bosselut, Ronald Clark, Arman Cohan, Jakob Foerster, Yarin Gal, Scott A. Hale, Deborah Raji, Christopher Summerfield, Philip H. S. Torr, Cozmin Ududec, Luc Rocher, and Adam Mahdi. Measuring what Matters: Construct Validity in Large Language Model Benchmarks. *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025. doi: 10.52202/085713-0590. URL https://papers.nips.cc/paper_files/paper/2025/hash/1967e0fc3aa6cbbace562f5cb8e3954e-Abstract-Datasets_and_Benchmarks_Track.html.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, et al. Holistic evaluation of large language models for medical tasks with MedHELM. *Nature Medicine*, 32(3):943–951, 2026. doi: 10.1038/s41591-025-04151-2. URL https://www.nature.com/articles/s41591-025-04151-2.

Ivan Bercovich, Ivgeni Segal, Kexun Zhang, Shashwat Saxena, Aditi Raghunathan, and Ziqian Zhong. Terminal Wrench: A Dataset of 331 Reward-Hackable Environments and 3,632 Exploit Trajectories, 2026. URL https://arxiv.org/abs/2604.17596.

Ryan Burnell, Han Hao, Andrew R. A. Conway, and José Hernández-Orallo. Revealing the structure of language model capabilities. *arXiv preprint arXiv:2306.10062*, 2023. URL https://arxiv.org/abs/2306.10062.

Donald T. Campbell and Donald W. Fiske. Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi: 10.1037/h0046016.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob

McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. Evaluating Large Language Models Trained on Code. arXiv:2107.03374, 2021. URL https://arxiv.org/abs/2107.03374.

Yick Chung. Replication package for: How model choice and memory shape preference consistency in large language models, 2026. URL https://zenodo.org/records/19004452.

Lee J. Cronbach and Paul E. Meehl. Construct Validity in Psychological Tests. *Psychological Bulletin*, 52(4):281–302, 1955. doi: 10.1037/h0040957.

J. de Curtò and I. de Zarzà. Metamorphic Testing for Semantic Invariance in Large Language Models. *IEEE Access*, 13:214772–214791, 2025. doi: 10.1109/ACCESS.2025.3646270.

Shuangrui Ding, Xuanlang Dai, Long Xing, Shengyuan Ding, Ziyu Liu, Jingyi Yang, Penghui Yang, Zhixiong Zhang, Xilin Wei, Xinyu Fang, Yubo Ma, Haodong Duan, Jing Shao, Jiaqi Wang, Dahua Lin, Kai Chen, and Yuhang Zang. WildClawBench: A Benchmark for Real-World, Long-Horizon Agent Evaluation, 2026. URL https://arxiv.org/abs/2605.10912.

Guoliang Duan, Mingwei Liu, Yanlin Wang, Chong Wang, Xin Peng, and Zibin Zheng. A Hierarchical and Evolvable Benchmark for Fine-Grained Code Instruction Following with Multi-Turn Feedback, 2025. URL https://arxiv.org/abs/2507.00699.

Timo Freiesleben. Establishing Construct Validity in LLM Capability Benchmarks Requires Nomological Networks. *arXiv preprint arXiv:2603.15121*, 2026.

Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. DOVE: A Large-Scale Multi-Dimensional Predictions Dataset Towards Meaningful LLM Evaluation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 11744–11763, 2025a. doi: 10.18653/v1/2025.findings-acl.611. URL https://aclanthology.org/2025.findings-acl.611/.

Eliya Habba, Noam Dahan, Gili Lior, and Gabriel Stanovsky. PromptSuite: A Task-Agnostic Framework for Multi-Prompt Generation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 254–263, 2025b. doi: 10.18653/v1/2025.emnlp-demos.19. URL https://aclanthology.org/2025.emnlp-demos.19/.

Jacqueline Harding and Nathaniel Sharadin. What Is It for a Machine Learning Model to Have a Capability? *The British Journal for the Philosophy of Science*, 2024. doi: 10.1086/732153. URL https://doi.org/10.1086/732153. Accepted 2024.

Jackson Hassell, Dan Zhang, Hannah Kim, Tom Mitchell, and Estevam Hruschka. Learning from Supervision with Semantic and Episodic Memory: A Reflective Approach to Agent Adaptation. In *Proceedings of the ACM Conference on AI and Agentic Systems*, pp. 276–291, 2026. doi: 10.1145/3786335.3813154. URL https://doi.org/10.1145/3786335.3813154.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring Massive Multitask Language Understanding. *International Conference on Learning Representations*, 2021. URL https://arxiv.org/abs/2009.03300.

Mengya Hu, Qiong Wei, and Sandeep Atluri. From Prompt Risk to Response Risk: Paired Analysis of Safety Behavior of Large Language Models, 2026. URL https://arxiv.org/abs/2604.26052v4. Analyzed release: https://github.com/microsoft/PairedSafety/tree/3428a08235a827e591871f6272a3801bcd9eb0b3.

Jikai Jin, Vasilis Syrgkanis, Sham Kakade, and Hanlin Zhang. Discovering Hierarchical Latent Capabilities of Language Models via Causal Representation Learning. In *Conference on Language Modeling*, 2026. URL https://arxiv.org/abs/2506.10378.

Nathanael Jo and Ashia Wilson. Position: AI Evaluations Should be Grounded on a Theory of Capability. In *International Conference on Machine Learning, Position Paper Track*, 2026. URL https://arxiv.org/abs/2509.19590v2.

Michael T. Kane. Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1):1–73, 2013. doi: 10.1111/jedm.12000.

Ivan Kartáč, Mateusz Lango, and Ondrej Dusek. Reasoning Gets Harder for LLMs Inside A Dialogue. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 12263–12303, 2026. doi: 10.18653/v1/2026.acl-long.560. URL https://aclanthology.org/2026.acl-long.560/.

Ryan Othniel Kearns. Quantifying construct validity in large language model evaluations. MSc thesis, Oxford Internet Institute, University of Oxford, 2025. URL https://arxiv.org/abs/2602.15532.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI Ability to Complete Long Tasks, 2025. URL https://arxiv.org/abs/2503.14499v1. Analyzed release: METR Time Horizon 1.1, https://github.com/METR/eval-analysis-public/tree/52cb829c7a2efb2d659285c4b1768d191d97f8d2/reports/time-horizon-1-1.

Yu Li, Haoyu Luo, Yuejin Xie, Yuqian Fu, Zhonghao Yang, Shuai Shao, Qihan Ren, Wanying Qu, Yanwei Fu, Yujiu Yang, Jing Shao, Xia Hu, and Dongrui Liu. AT-Bench: A Diverse and Realistic Agent Trajectory Benchmark for Safety Evaluation and Diagnosis, 2026. URL https://arxiv.org/abs/2604.02022v4. Analyzed release: https://huggingface.co/datasets/AI45Research/ATBench/tree/4476ef92ed8f85c8d58d8a5b9dfdf55aa7893138.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic Evaluation of Language Models. *Transactions on Machine Learning Research*, 2023. URL https://arxiv.org/abs/2211.09110.

Shuyang Liu, Saman Dehghan, Jatin Ganhotra, Martin Hirzel, and Reyhaneh Jabbarvand. From Plan to Action: How Well Do Agents Follow the Plan? In *IEEE/ACM International Conference on Automated Software Engineering (ASE), Research Papers*, 2026. URL https://arxiv.org/abs/2604.12147. Analyzed release: https://zenodo.org/records/22047054.

Xiang Long, Li Du, Yilong Xu, RongJian Xu, Qiyanhui Lu, Ying Gao, Qinhua Xie, Fangcheng Liu, Ning Ding, Haoqing Wang, Ziheng Li, Changjiang Zhou, Jianyuan Guo, and Yehui Tang. LiveClawBench: Benchmarking LLM Agents on Complex, Real-World Assistant Tasks, 2026. URL https://arxiv.org/abs/2604.13072v2. Analyzed release: v0.2.1, https://huggingface.co/datasets/Mosi-AI/LiveClawbench-trajectories/tree/c3cfa229a90a8d540c88b9c490ac3dd88db1425a.

Yujun Mao, Yoon Kim, and Yilun Zhou. CHAMP: A Competition-level Dataset for Fine-Grained Analyses of LLMs' Mathematical Reasoning Capabilities. *Findings of ACL*, 2024. URL https://aclanthology.org/2024.findings-acl.785/.

Samuel Messick. 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. doi: 10.1037/0003-066X.50.9.741.

Melanie Mitchell. Six principles for evaluating cognitive capabilities in AI models. *AI Magazine*, 47 (2):e70061, 2026. doi: 10.1002/aaai.70061.

Open Eval Commons. OpenEval. Hugging Face dataset snapshot, 2026. URL https://huggingface.co/datasets/Open-Eval-Commons/OpenEval/tree/23a1ded985b3c2cdaaddb27581a35bfabe0ad7e7.

Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. Sloth: scaling laws for LLM skills to predict multi-benchmark performance across families. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1318. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/3834d037cd6e29ac7649df42611c5d66-Abstract-Conference.html.

Haritz Puerto, Haonan Li, Xudong Han, Timothy Baldwin, and Iryna Gurevych. From Leaky Thoughts to Private Reasoning: Controlling What LRMs Say to Themselves. Accepted at EMNLP 2026 (main), 2026. URL https://arxiv.org/abs/2602.24210v4.

Ruchit Rawal and Mariya Toneva. Perturbed examples reveal invariances shared by language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11564–11584, 2024. doi: 10.18653/v1/2024.findings-acl.687.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. *Proceedings of ACL*, 2020. URL https://aclanthology.org/2020.acl-main.442/.

Yangjun Ruan, Chris J. Maddison, and Tatsunori Hashimoto. Observational Scaling Laws and the Predictability of Language Model Performance. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-0506.

Pranav Singh. Failure Modes in Perturbation-Based Measurement of Language Model Reliability. GitHub repository snapshot, 2026. URL https://github.com/Pranavsingh431/Perturbation-Based-Measurements/tree/d7fa58215bf6966df70d05f5433f1238c05d2ea0.

Specific Labs. HalBench: A Sycophancy and Hallucination Benchmark for Large Language Models. Hugging Face dataset snapshot, 2026. URL https://huggingface.co/datasets/Specific-Labs/halbench/tree/4697df3b0e9b768f1a955ad587b531918b80244d. Analyzed asset: per-model scores.jsonl files in the responses directory.

Aarohi Srivastava et al. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *Transactions on Machine Learning Research*, 2023.

Hyewon Suh, Binfei Ji, Seojune Lee, Rishi Khare, Basit Khan, Hyunjun Kim, Tianyi Zhang, Venkat Krishna Srinivasan, Peter Belcak, Shizhe Diao, Pavlo Molchanov, Yingyan Celine Lin, and Zhen Dong. AgentSuite: Toward More Reliable Agent Evaluation with a Component-Based Benchmark Auditing Pipeline. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=2Exmr1eIKZ.

The mathlib Community. The Lean Mathematical Library. In *Proceedings of the 9th ACM SIGPLAN International Conference on Certified Programs and Proofs*, CPP 2020, pp. 367–381. Association for Computing Machinery, 2020. doi: 10.1145/3372885.3373824. URL https://doi.org/10.1145/3372885.3373824.

Nishal Thomas and Noel Thomas. FormInv: A Measurement Protocol for Semantic Invariance in Mathematical Reasoning Benchmarks, 2026. URL https://arxiv.org/abs/2605.29001.

Alex Wang et al. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*, 2019a.

Alex Wang et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In *Advances in Neural Information Processing Systems*, 2019b.

Jiaming Wang, Yunke Zhao, Peng Ding, Jun Kuang, Yibin Shen, Zhe Tang, Yilin Jin, ZongYu Wang, Xiaoyu Li, and Xuezhi Cao. Meeseeks: A Feedback-Driven, Iterative Self-Correction Benchmark evaluating LLMs' Instruction Following Capability. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 14745–14773, 2026a. doi: 10.18653/v1/2026.findings-acl.725. URL https://aclanthology.org/2026.findings-acl.725/.

Zongqi Wang, Tianle Gu, Chen Gong, Xin Tian, Siqi Bao, and Yujiu Yang. SCAN: Structured Capability Assessment and Navigation for LLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics*, pp. 2760–2799, 2026b. doi: 10.18653/v1/2026.acl-long.126.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or Reciting? Exploring the Capabilities and Limitations of Language Models Through Counterfactual Tasks. *Proceedings of NAACL*, 2024. URL https://aclanthology.org/2024.naacl-long.102/.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained Language Model Evaluation based on Alignment Skill Sets. *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2307.10928.

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. Skill-Mix: a Flexible and Expandable Family of Evaluations for AI models. *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.17567.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/5f9bfdfe3685e4ccdbc0e7fb29cccf2a-Abstract-Conference.html.

Ming Zhong, Aston Zhang, Xuewei Wang, Rui Hou, Wenhan Xiong, Chenguang Zhu, Zhengxing Chen, Liang Tan, Chloe Bi, Mike Lewis, Sravya Popuri, Sharan Narang, Melanie Kambadur, Dhruv Mahajan, Sergey Edunov, Jiawei Han, and Laurens van der Maaten. Law of the Weakest Link: Cross Capabilities of Large Language Models. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/63ea6a7d01a34a2c7334dcf1a2c3b03d-Abstract-Conference.html.

Lexin Zhou, Lorenzo Pacchiardi, Fernando Martínez-Plumed, Katherine M. Collins, Yael Moros-Daval, Seraphina Zhang, Qinlin Zhao, Yitian Huang, Luning Sun, Jonathan E. Prunty, Zongqian Li, Pablo Sánchez-García, Kexin Jiang-Chen, Pablo A. M. Casares, Jiyun Zu, John Burden, Behzad Mehrbakhsh, David Stillwell, Manuel Cebrian, Jindong Wang, Peter Henderson, Sherry Tongshuang Wu, Patrick C. Kyllonen, Lucy Cheke, Xing Xie, and José Hernández-Orallo. General scales unlock AI evaluation with explanatory and predictive power. *Nature*, 652:58–67, 2026a. doi: 10.1038/s41586-026-10303-2. URL https://www.nature.com/articles/s41586-026-10303-2.

Yongxin Zhou, Philippe Mulhem, and Didier Schwab. TempPerturb-Eval: On the Joint Effects of Internal Temperature and External Perturbations in RAG Robustness. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pp. 117–127, 2026b. doi: 10.63317/3v28zfbo6agu. URL https://aclanthology.org/2026.lrec-1.9/.

A Complete methods and statistical specification

A.1 Released-source documentation

The normalized response program draws on released resources for counterfactual tasks, plan adherence, dialogue reasoning, critique and adaptation, iterative correction, prompt-surface variation, mathematical reasoning, memory conditions, private reasoning, coding robustness, chain-of-thought perturbation, evolving instruction following, and temperature–retrieval perturbation (Wu et al., 2024; Liu et al., 2026; Kartáč et al., 2026; Hassell et al., 2026; Wang et al., 2026a; Habba et al., 2025a;b; Mao et al., 2024; Chung, 2026; Puerto et al., 2026; Anonymous, 2026a;c; Aravindan & Kejriwal, 2026; Duan et al., 2025; Zhou et al., 2026b). Independent control resources include the released MedBias, semantic-invariance, FormInv, and perturbation-measurement packages (Anonymous, 2026b; de Curtò & de Zarzà, 2025; Thomas & Thomas, 2026; The mathlib Community, 2020; Singh, 2026). Supporting external, observational, and request-ecology analyses use AgentSuite, LiveClawBench, WildClawBench, METR long-task data, BenchMIRT, LMSYS-Chat-1M, HalBench, paired safety responses, ATBench, MedHELM, Terminal Wrench, CrossEval, and OpenEval where their released schemas identify the relevant estimand (Suh et al., 2026; Long et al., 2026; Ding et al., 2026; Kwa et al., 2025; Allen Institute for AI, 2026; Zheng et al., 2024; Specific Labs, 2026; Hu et al., 2026; Li et al., 2026; Bedi et al., 2026; Bercovich et al., 2026; Zhong et al., 2025; Open Eval Commons, 2026). HELM and the ADeLe demand scales provide additional observational context (Liang et al., 2023; Zhou et al., 2026a). Each empirical claim remains limited to the exact released version and fields actually analyzed.

A.2 Response estimands and direction

Every released outcome is examined for whether larger values denote better behavior; error- or harm-like outcomes are explicitly sign-reversed. Frozen raw-source fixtures verify selected raw-item-to-observation mappings without consulting active processed results, and each source-system-qualified selector must match exactly one normalized observation. These fixtures certify the mapping they exercise, not an unreleased upstream evaluator; C05, C16, and C17 remain explicitly proxy-only where applicable, and C14 evaluator fidelity is not independently verifiable. For a paired design, responses are computed only on canonical roots observed in both baseline and treatment. Structured-unpaired designs retain distinct population labels and never enter a paired causal interpretation. Surface assets use the root-level standard deviation across observed variants as their primary response. Signed leave-one-condition-out deviations are retained only as secondary diagnostics.

Canonical root identifiers are asset-qualified unless a source exposes an explicit HumanEval/MBPP benchmark task identity (Chen et al., 2021; Austin et al., 2021). Response-unit identifiers include the asset. Exact repeats are removed only after one global scientific-key conflict review; conflicting values or provenance hard-fail. Generic model aliases are not used for cross-source validation. A deployed-system identity is exact only when the release supplies a consistent model–scaffold pair; unknown regimes remain source-local. Measurement status distinguishes `VALID`, `VALID_ZERO`, evaluator-confirmed `MODEL_FAILURE`, `POLICY_REJECTION`, `ENVIRONMENT_FAILURE`, `TEST_DEFINITION_FAILURE`, `MEASUREMENT_FAILURE`, and `MISSING`. Only the first three enter scientific outcomes. In C14, the released entry point is resolved before test execution; assertion or missing-entry-point failure under an executable released checker is a model failure, while local safety rejection, resource/environment failure, absent test, and invalid test definition remain missing outcomes rather than zeros.

A.3 Pairing, missingness, and uncertainty

Pairing coverage is measured for each asset, system, intervention family, and treatment condition against its common baseline. The resulting table records paired roots, roots present on only one side, and the baseline means of retained and dropped roots. System–asset root coverage is reported directly; no missing outcome is imputed. A response cell with fewer than two roots has unavailable sampling standard error and cannot receive precision weight; 426 such cells are present and none is assigned a nonmissing standard error. Fixed design cells and stochastic repeats are never pooled into one resampling level. C17's three aligned response indices are sampled coherently by root and intervention family in task-response modules that admit C17. C06's documented independent runs are

sampled within horizon cells in law and overall-strength modules that admit the structured-unpaired source. Each method states its actual propagation scope; no claim is made that either source enters an inapplicable module. Sources whose released row IDs do not establish stochastic repetition remain conditional fixed-design analyses rather than receiving invented replicate uncertainty.

A.4 Reliability and task transfer

Within each construct, asset, domain, intervention family, and condition, the common system–root panel is frozen before splitting. Candidate systems require at least six roots; systems are pruned only to obtain a common panel. Five hundred deterministic random streams assign roots without conditioning on observed outcomes; outcome-stratified balanced halves are reported only as a sensitivity, and odd extras alternate sides. Raw Spearman rank correlation is primary. Spearman–Brown length correction is secondary and never replaces the primary estimate; Pearson and Lin concordance are separate. Repeated five-fold task-OOD partitions likewise report raw Spearman correlation. Eligibility, half sizes, fixed-cohort identity, and finite-estimate fractions are disclosed.

A.5 Local relationships and rank crossings

Paired response laws and structured-unpaired performance–demand relations are separate estimands. The confirmatory nonconstancy statistic is a prespecified exact leave-one-root-out saturated-intensity gain over a constant model. Observed and null worlds run this identical estimator. Paired nulls use independent seeded permutations of registered intensity labels within every root, with each root/world transform shared across systems; the structured-unpaired horizon program independently permutes horizon labels across released task roots, again sharing a panel transform across systems. The registry states the permutation, quantities held fixed, and source-specific exchangeability rationale. One thousand disjoint reference and evaluation null worlds are default. Every evaluation-world supported count replays the root confidence bound, global BH rule, and the source-identifiable C17/C06 first-stage stochastic uncertainty criterion used by observed support. Complete decision-bearing outputs are compared at 10^{-12} tolerance after actual 1-, 2-, and 14-shard executions. Functional-family selection among monotone, threshold, linear/cardinal, and spline alternatives uses 20 training-only repeated grouped-fold assignments and is secondary; exact class requires at least 70% stability. BY and Holm are reported as sensitivities.

Intensity is never inferred from an arbitrary number embedded in a condition name. A source-explicit registry admits only released arithmetic-base, memory-example, METR log-time, constraint-count, repair-round, and temperature designs, with units and cardinal/ordinal semantics recorded. Rotation labels, hint IDs, task IDs, and unregistered numeric strings fail closed. Metric slopes, breakpoints, and threshold locations are reported only for registered cardinal scales; ordinal repair rounds use endpoint, monotonic/threshold, and ordered-boundary summaries without a fictitious metric slope. Root-cluster parameter intervals propagate C17 paired response indices and C06 independent runs where those stochastic semantics are released.

All eligible system pairs receive identical endpoint inference. Paired designs first intersect system A/B at both endpoints, then jointly resample the resulting four-way common-root vector. A shared sign vector supplies the joint randomization null. Two reversal orientations are evaluated by an intersection–union test with orientation multiplicity correction and global BH. Structured-unpaired horizon relations independently resample their explicitly labeled endpoint populations. Endpoint reversal is not interpreted as full-function crossing.

A.6 Interactions, feedback, and surface designs

Natural interactions use exact released cell names and source-appropriate common-root contrasts for A03, C05, C16, and C17. C05 is restricted to its registered concept-by-hint factorial before label matching; its common-root four-cell contrasts are confirmatory, while any disjoint-root population contrasts remain secondary structured-unpaired evidence. Root-cluster sign flips supply randomization p-values, and all confirmatory natural-scale tests share one multiplicity family. The two feedback contrasts are root-cluster means with system-equal sensitivities and apply only to the released critique operationalization. Surface models combine every physical parquet partition belonging to a scientific

asset–domain–system group before fitting, use every eligible row without a cap, and predict raw outcomes after training-only root centering under condition- and task-OOD.

A.7 CONSTRUCT VALIDITY

Nine candidates and prespecified neighbors use an exploded primary/secondary multi-label view. Every candidate–operationalization membership maps to exactly one evidence role. Only `POSITIVE` and `SUPPORTING` roles enter positive candidate estimators; `NUISANCE_NEGATIVE` and `CONTROL` roles are retained for challenge analyses and cannot promote. One shared evaluator receives local reliability, raw task transfer, cross-operationalization/source transfer, an all-neighbor same-source–system–root matched specificity criterion, discriminant evidence, minimality, measurement-quality stability, negative controls, exact-target external validity, and common support. Small executable state tests cover positive, boundary-overlap, conflicting, missing, non-conflict, NaN-control, and calibration-veto cases. Held-root predictive specificity against the registered coarse taxonomy is supporting evidence and cannot rescue a failed matched criterion. Candidate-specific compatible cohorts are frozen before evaluation across the complete nine-candidate universe; missing components remain missing rather than dropping candidates. Cross-source and external tests accept only source-explicit exact deployed-system identities. Required cross-source pairs are registered before that exact-identity filter; a source with no exact rows therefore remains a missing required pair and forces non-identifiability rather than disappearing from the aggregate. The asset-equal demand–nuisance normalized-magnitude difference is 0.05415 (13 demand, three nuisance assets): descriptive, not construct specificity.

Minimality predicts held-root responses with system, construct, and system–construct effects. For each candidate, the full coding is compared separately with a target-profile ablation that retains its population main effect, a system-only ablation that removes all construct information, two prespecified neighbor merges, and coarse coding over 20 repeated five-fold root assignments. Every comparator chooses its own regularization inside the same training fold. No baseline-outcome predictor enters this model. Mean response is reported on a separate axis and never determines promotion. Common-support validity requires the relevant axes to coexist rather than assembling a positive label from unrelated subsets. For information utilization, local proxy-based support does not complete the required measurement-quality criterion: its compatible cohort has 348,778 supporting-proxy rows and no objective-measurement rows, so measurement quality remains non-identifiable.

Discriminant inference uses only the ten prespecified neighbor pairs for the primary family. Bootstrap intervals receive Bonferroni familywise control, and a conflict requires the adjusted interval—not merely a correlation test against zero—to lie beyond $|\rho| = .8$. Exploratory all-pair comparisons remain separately labeled. Independent controls are executed from C09 MedBias, C10 Semantic Invariance, C18 FormInv, C19 Perturbation-Based Measurements, and S01 Controlled Reasoning Outputs. Their released negative/control outcomes form a model-by-control matrix; only explicit deployed-system links with matching released model/version and scaffold/regime can support held-system contamination prediction and a regularized control-subspace diagnostic. Otherwise the source–candidate link is explicitly non-identifiable.

A.8 END-TO-END CONSTRUCT CALIBRATION

Calibration uses a generated identifiable topology that is separate from released observations: 24 synthetic systems, 16 roots, two imbalanced sources, two operationalizations, stable incomplete cells, registered primary/secondary memberships, exact cross-source bridges, negative controls, and independent external targets. Null, weak, medium, and strong signals are generated at the observation level and passed through the same normalization, response construction, reliability/task-transfer, profile-transfer, matched and predictive specificity, minimality, discriminant, control, external, common-support, and production-evaluator code—with the same bootstrap and repeat configuration—used for real evidence. Separate conflict worlds inject specificity, discriminant, or control failure. Deterministically disjoint simulation worlds are used for threshold search and independent qualification. A prespecified threshold grid targets null any-promotion probability at most 0.05 and strong-signal detection at least 0.80 on the search half; the selected rule can support promotion only when Wilson bounds on the untouched qualification half satisfy both limits. Monte Carlo evaluates the complete prespecified bank of 200 total worlds per signal before either qualification or veto; no interim outcome

can stop the procedure or permit promotion. Otherwise calibration veto prevents promotion. Released evidence topology is evaluated only after selection and cannot choose thresholds. Calibration therefore tests an executable production rule, not whether constructs ontologically exist.

A.9 Matched representation comparisons

Four comparisons progressively align model identity, system identity, source context, and exact response-unit information. The semantic model uses exact train-fitted sparse indicators with training-only Ridge selection. The low-rank comparator uses alternating least squares optimized only on observed entries, with training-only rank selection constrained by the semantic design-width parameter budget (rank-one relaxation is flagged). No row cap is used. Bias-only predictions, identity/unit cold-start coverage, and a richer measured comparator are retained. Twenty repeated grouped partitions use shared outer assignments across representations for task, intervention, operationalization, source, and system OOD. Root-, asset-, and system-equal headlines reweight identical predictions. A matched-training-objective sensitivity collapses both models to the same analysis cells, equal cell weights, common training population, comparable inner validation target, and common headline weighting.

A.10 Sensitivity families

Natural signed response remains the sole canonical outcome. The scale sensitivity does not create a universal prestandardized dataset. Instead, every predictive or OOD estimator fits a positive scale only on its outer-training roots (with a prespecified floor), applies that scale to the corresponding test rows without centering, and then fully refits the claim-specific estimator. Law null reference and evaluation banks remain disjoint, and standardized-law supported counts replay the same scaled root bound, BH correction, and scaled C06/C17 first-stage uncertainty criterion as the observed rule. Interaction cells belonging to one registered contrast share the same held-root-excluded scale so A03 retains its four-cell difference-in-differences estimand, C05 retains its registered common-root concept-by-hint difference-in-differences estimand, C16 retains its adjacent-level second difference, and C17 retains its temperature-by-evidence difference-in-differences estimand. Reliability, laws, crossing, interactions, construct validity, and representation are all re-evaluated. The exact candidate evaluator is also rerun for the full 3×2 design of all-eligible, objective-only, and supporting-proxy measurements crossed with primary-only and canonical primary/secondary multi-label taxonomies.

Model-family robustness labels the operation, rather than calling every deletion a refit. Family-equal weighting is evaluation composition. Leave-one-family-out reliability rebuilds the common panel and every split; laws and interactions recompute retained estimators/sufficient results, null adjudication, multiplicity, and support; representation retrains and retunes both competitors; and candidate validity refits local, transfer, cross-operationalization, cross-source, matched/predictive specificity, minimality, discriminant, measurement-quality, control, external, common-support, and calibrated-evaluator stages. Post-adjudication exclusion is a distinct operation and is not used as evidence for a full-refit claim. Additional families cover root-, asset-, and system-equal weights; leave-one-asset and intervention intervals; a hierarchy that retains every fixed condition while resampling only source-documented stochastic repeats; BH, BY, and Holm multiplicity; and six source-root-OOF additive feature-block orders (canonical, reverse, and four deterministic permutations). These analyses bound claims but do not manufacture missing exact identity or external targets.

A.11 Restored residual, atlas, and supporting science

The failure program predicts held canonical roots in five folds while cumulatively adding deployed-system ID, semantic, registered intensity, source/domain context, feedback, compound-demand, surface, and operationalization blocks. The stage key `model` uses `system_id`; its cross-fitted prediction is the system-ID-only strength baseline. Semantics uses primary, secondary, and coarse construct indicators, not text embeddings. Source/domain context is the asset–domain combination. The feedback block groups intervention-family names matching feedback, crit, hint, or repair; the compound block groups primary/secondary construct pairs. The surface block combines demand role with intervention family, followed by operationalization. These are predictive feature encodings, not separately measured capabilities. Figure 4B averages five fold MSEs equally; panel C pools their OOF rows. The non-nested system–response-unit key comparator shrinks each training key

mean toward the training grand mean with weight $n/(n+5)$ and falls back to that grand mean for unseen keys. The failure program reports stagewise OOF MSE, incremental gains, final row-level residuals, and six prespecified/reordered block sequences; these are predictive decompositions, not explained-variance identities. Singular spectra, entropy effective rank, and endpoint residual-crossing diagnostics are recomputed at every cumulative OOF stage; asset-level system-bootstrap intervals additionally characterize the final residual panel. All are explicitly bounded as geometry descriptors rather than estimates of true latent dimensionality.

The descriptive atlas retains root, system, model-family, intervention-family, sign, source, and baseline-response heterogeneity. Candidate–asset forest estimates use canonical-root uncertainty; between-asset τ^2/I^2 are produced only with at least three estimable asset effects. Baseline-response correlations are descriptive and never treated as mechanism or promotion evidence. The context/feedback/memory map covers BOULDER, Critique Learning, Meeseeks, LiveClaw, CHAMP, Memory Condition, Coding Hint, and MultiCodeIF, with per-system distributions. MultiCodeIF additionally reports constraint family/type, language, and root-level first-to-last-round lexical/proxy-score change; that proxy is not correctness.

LMSYS ecology streams all released conversations and reports model/language, turn/message depth, request length, redaction, URL, and multi-user-turn structure. ATBench reports label, risk source, failure mode, harm, tool multiplicity, trajectories, and dialogue depth. These are within-source request-ecology summaries. For BenchMIRT, ADeLe, HELM, MedHELM, and OpenEval, released observational matrices retain source-local geometry. Where matrix support permits, training-system-only nested rank selection is followed by exact-system leave-one-out/task-fold prediction; unsupported matrix shapes are marked non-identifiable. ADeLe separately uses all 18 released demand dimensions, training-only preprocessing, task- and benchmark-OOD splits, prevalence baselines, and held-group bootstrap intervals.

A.12 Executable Level-3 program

Ten known-truth world families—global ordering; global low rank at ranks one, two, and three; local; modular; hybrid; high-dimensional; nuisance; and misspecified/no-structure—are generated at weak, medium, and strong signal on released system–unit masks with released root incidence, replicate patterns, and cell sizes. Every world runs through the production structural pipeline; the earlier spectral-world classifier is retained only as a secondary diagnostic. Within each released asset we report systems, response units, observed cells, density, and missingness. Complete/common-support geometry is primary when at least five systems and eight complete units exist; masked low-rank reconstruction is the missingness-aware sensitivity and column-mean imputation is only a simple sensitivity.

Geometry uncertainty uses a canonical-root exponential-weight cluster bootstrap that rebuilds the response matrix without duplicating system rows. We report bootstrap bias, the original estimate and percentile interval, interval inclusion, and subspace stability. Predictive stability has two required tests. The first is matched held-cell low-rank improvement over an additive comparator. The second partitions systems within each intervention family, removes the complete intervention-family $\times$ held-system test block, learns both the intervention-specific response units from other systems and the held systems from other interventions, and compares selected masked low rank with the additive model only on cells whose system and unit were both observed in training. A panel is non-identifiable when either required predictive test cannot be estimated; it is supported only when both are supported together with the geometry criteria. Structural competitors share outer folds, observed test cells, and squared-error loss. Prediction-equivalent comparators are first labeled `NOT_DISTINGUISHABLE_UNDER_CURRENT_DESIGN`; the remaining classes receive matched pairwise ΔMSE intervals and fold-win summaries, and no model is crowned from mean MSE alone. Cross-domain transfer learns structural coordinates only in source domain A and tests target-domain B conditional-response cells under exact released model/version plus scaffold/regime identity and held-system OOD; cross-asset response-strength concordance is secondary. Diversity scaling repeatedly subsets response units and intervention families at 25/50/75/100%, refits the estimator, and recomputes geometry and held-cell prediction within the same panel; cross-asset ecology is secondary. The final output retains all five criterion states, and the full Level-3 claim is established only when every criterion is supported.

A.13 EXTERNAL TOPOLOGY

External predictive validation requires at least five exact source-explicit model overlaps on a construct-mapped independent target. The source inventory includes AgentSuite, WildClaw, METR, HalBench, PairedSafety, Terminal Wrench, and anonymous CrossEval ratings, while BenchMIRT, ADeLe, HELM, MedHELM, and OpenEval enter a separate observational-geometry program. CrossEval response slots are never inferred to be model identities. WildClaw's released coexistence of failed and successful tool results is not called temporal recovery. Descriptive trajectory, exploit, composition, ecology, or observational analyses cannot substitute for same-target linkage.

B READING THE COMPLETE EVIDENCE LAYER

The appendix follows the scientific questions, while preserving the distinct populations and estimands needed to interpret them. Tables use S for supported, ENS for estimable but not supported, NI for not identifiable, C for conflicting, and D for descriptive only. These are not ordered scores. A blank numerical estimate is shown as an em dash and is never interpreted as zero. Candidate states are taken from the production construct evaluation; the secondary released-topology diagnostic is not allowed to override them.

The paper presents all central estimates, candidate criteria, source populations, calibration outcomes, OOD comparisons, sensitivity families, and five organizational questions below. Long row-level catalogs accompany the manuscript in `appendix/catalogs/`: these preserve the full response-law and parameter catalogs, endpoint tests, measurement contracts and fixtures, pair retention, response atlas, source-specific secondary results, residual spectra, and structural transfer/diversity diagnostics. They are copies of completed results, not newly estimated outputs. The PDF explains their scientific interpretation so a file name is never the sole statement of a finding.

B.1 COMPLETE SOURCE UNIVERSE

The 34 collections have different roles. Sixteen reconstruct conditional responses; the others supply independent challenges, external targets, observational matrices, or request ecology. A source's presence in the universe does not imply that it identifies every desired link. In particular, anonymous slots and source-local model labels remain source-local even when their names resemble a system elsewhere. The following table also provides the source-code key used in compact tables and figures.

Table 2: All 34 released source collections. Zero core observations denotes a supporting, external, control, observational or ecology role, not absence of released data.

Source identifier and collection	Role	Core observations
A01 Counterfactual Evaluation	core response	134120
A02 BOULDER	core response	19200
A03 Critique Learning	core response	103473
A04 Meeseeks	core response	7693
A05 PromptSuite	core response	36500
A06 DOVE Lite Selected	core response	9022704
C01 Planning Analysis	core response	16577
C02 AgentSuite	supporting or external	0
C03 LiveClawBench	core response	6747
C04 WildClawBench Trajectories	supporting or external	0
C05 CHAMP	core response	19980
C06 METR Time Horizon 1 1	core response	24008
C07 BenchMIRT	supporting or external	0
C08 ADeLe	supporting or external	0
C09 MedBias	supporting or external	0
C10 Semantic Invariance	supporting or external	0
C11 Memory Condition	core response	3960
C12 HELM Capabilities Selected	supporting or external	0
C13 SEAL	core response	144192
C14 Coding Hint	core response	73520
C15 CoTPerturbation	core response	23400
C16 MultiCodeIF	core response	18560
C17 TempPerturb Eval	core response	791868
C18 FormInv	supporting or external	0
C19 Perturbation Based Measurements	supporting or external	0
S01 Controlled Reasoning Outputs	supporting or external	0
S02 LMSYS-Chat-1M	supporting or external	0
S03 HalBench	supporting or external	0

Continued from previous page

Source identifier and collection	Role	Core observations
S04 PairedSafety	supporting or external	0
S05 ATBench	supporting or external	0
S06 MedHELM Public Selected	supporting or external	0
S07 Terminal Wrench	supporting or external	0
S08 CrossEval	supporting or external	0
S09 OpenEval	supporting or external	0

C MEASUREMENT, FAILURE SEMANTICS, AND COMMON-ROOT RESPONSE

Across 3,099 treatment-baseline pairing groups, median pairing retention is 1.0; this median does not remove source- or group-specific losses. All 426 singleton cells have unavailable sampling standard errors. The measurement record contains 12 uniquely selected raw-source fixtures with no failed mapping and no ambiguous selector. These fixtures verify the exercised raw-to-normalized mapping; they do not independently certify unreleased evaluator fidelity. The distinction is especially relevant to C05, C16, and C17 supporting proxies and C14's restricted released-test execution.

C14 contains 64,330 released scoring records: 30,144 pass, 30,967 evaluator-confirmed model failures, 2,448 missing, 657 policy rejections, and 114 environment failures. The first two categories provide outcomes. Policy rejection and environment failure do not become model errors, and missing tests do not become valid zeros. Test-definition and measurement failures remain separate possible status categories even when a particular released summary has no such rows. This measurement distinction changes the meaning of a response and is therefore part of the scientific object, not a clerical preprocessing detail.

Table 3: Measurement-quality coverage of the frozen response cells. Source IDs are expanded in the source guide; root counts are source-scoped.

Source	Quality	Pairing	Cells	Systems	Roots
A01 Counterfactual Evaluation	objective	paired	220	50	76966
A02 BOULDER	objective	paired	127	8	11886
A03 Critique Learning	objective	paired	1677	6	68210
A04 Meeseeks	admissible released	structured unpaired	1	1	11
A04 Meeseeks	objective	structured unpaired	66	1	1838
A04 Meeseeks	supporting proxy	structured unpaired	6	1	1212
A05 PromptSuite	objective	surface special	44	8	1080
A06 DOVE Lite Selected	objective	surface special	26	4	2600
C01 Planning Analysis	objective	paired	28	4	10851
C03 LiveClawBench	admissible released	structured unpaired	713	17	1270
C05 CHAMP	supporting proxy	paired	340	4	17550
C11 Memory Condition	objective	paired	22	11	2640
C13 SEAL	objective	paired	126	6	10278
C14 Coding Hint	supporting proxy	paired	168	7	34910
C15 CoTPerturbation	admissible released	paired	117	13	11700
C16 MultiCodeIF	supporting proxy	paired	927	5	14683
C17 TempPerturb Eval	supporting proxy	paired	2190	5	218941

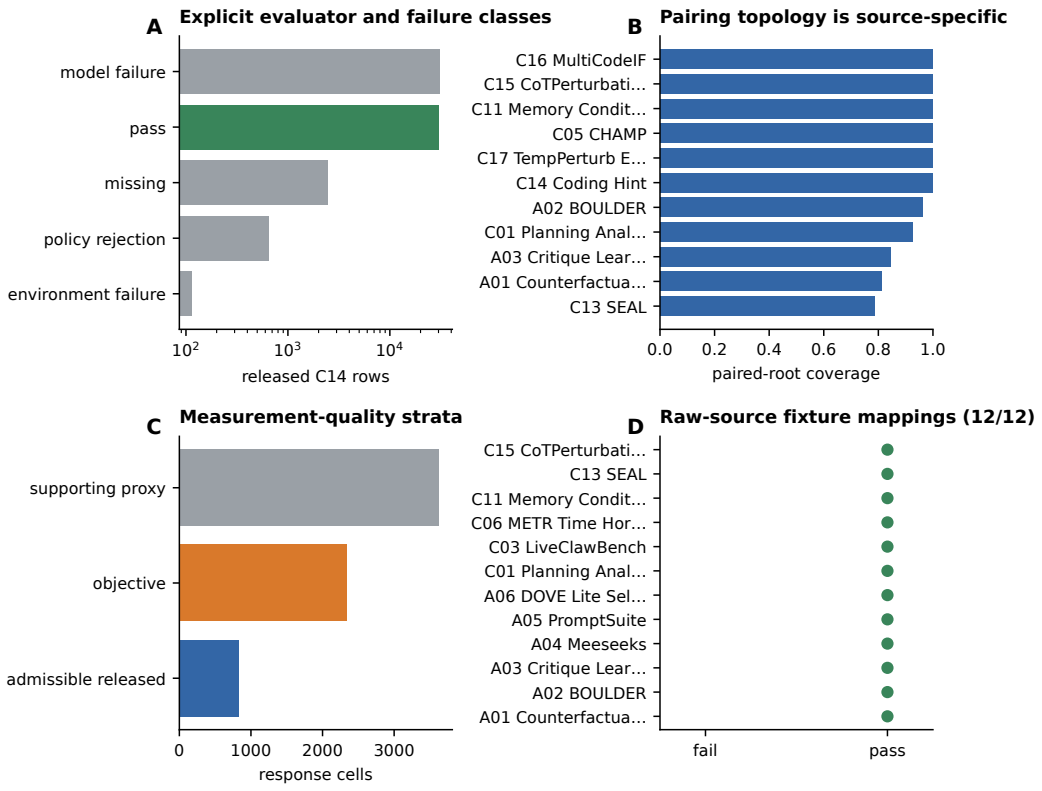

Figure 5: Measurement meaning and support. Released C14 statuses separate outcomes from non-outcomes (A); pairing support is source-specific (B); objective, admissible released and supporting-proxy strata remain distinct (C); uniquely selected fixtures verify raw-to-normalized mappings (D). A fixture success does not certify an unreleased evaluator.

D Complete Conditional Response evidence

D.1 Reliability and held-root transfer

Raw Spearman is the primary relative-profile statistic. Its 0.586 primary estimate is not a Spearman–Brown correction; the latter is 0.683 and is secondary. The 0.608 Lin concordance describes absolute agreement on the same eligible population. The three-system sensitivity expands to 1,227 groups across 13 assets and therefore changes eligibility as well as precision. The common cohort is fixed before random splitting, and finite or degenerate estimates are not silently selected after seeing the split.

Task-OOD transfer is separately estimated over repeated grouped folds. Candidate-local support joins interval-qualified rank reproducibility and held-root transfer on the candidate-compatible cohort. The four supported candidates have respective point summaries of (rank, task transfer): semantic adaptation (0.543, 0.486), reasoning integrity (0.498, 0.437), information utilization (0.685, 0.613), and misleading-information resistance (0.470, 0.445). These candidate summaries answer a narrower question than the all-demand-group headline and should not be substituted for it.

D.2 Local laws, functional families, and endpoint tests

The 311 eligible law groups use six registered intensity designs: arithmetic base, memory examples, METR log-time, constraint count, repair round, and temperature. The intensity's cardinal or ordinal meaning is inherited from the source. Arbitrary numeric condition labels do not create a scientific dose. Nonconstancy is confirmed by held-root gain over a constant, its root interval, a coherent empirical-null BH result, and any applicable first-stage stochastic criterion. The frozen null evaluation bank has mean supported count zero under the complete replayed rule; this does not turn observed relationships into universal laws.

The 37 supported exact classes require at least 70% stability under repeated training-only functional-family selection, in addition to supported nonconstancy. The complete electronic law parameter catalog contains 311 groups, with root-cluster endpoint-change, threshold, and available slope/breakpoint intervals. Metric slope and breakpoint interpretations apply only to registered cardinal scales; ordinal repair rounds retain ordered-boundary summaries. All groups, including unsupported relationships, remain in the catalogs. This avoids presenting selected illustrative curves as the entire evidence universe.

The 312 strict crossing tests jointly require endpoint separation and reversal stability; none is supported. Four-way common-root pairing is required for paired reversal, whereas structured-unpaired horizon endpoints retain their population-comparison interpretation. Descriptive residual crossings later in the appendix are different estimands and cannot overturn this strict original-response result.

D.3 Interaction, feedback, and surface structure

The complete interaction table retains all confirmatory and secondary rows, including nonsupported CHAMP contrasts. A03's critique difference-in-differences, C16's adjacent-constraint second difference, and C17's temperature-by-evidence contrasts are not pooled onto a common effect scale. Six confirmatory tests are supported after their full uncertainty requirements. A C17 row can have a nominally positive multiplicity result and still not be supported if the first-stage stochastic interval fails the required condition.

The feedback contrasts describe 1,750 roots and six released systems. System-equal sensitivity estimates are 0.08051 [0.05409, 0.11070] for helpful feedback and 0.79781 [0.69905, 0.85962] for correct versus misleading feedback. The wider system-equal intervals emphasize that roots and systems define different uncertainty questions. The system-level association between helpful utility and misleading-feedback loss is descriptive ($\rho = 0.771$, interval $[-0.091, 1]$); six systems do not establish a general feedback tradeoff or a causal relation.

Surface dispersion is baseline-free. A leave-one-condition-out signed response exists as a secondary diagnostic, but a deviation from a mean containing the same condition is not used as the primary object.

The frozen pairwise-versus-additive surface prediction table is specific to the eligible DOVE-Lite designs and the named condition- or task-OOD target.

Table 4: Response-profile rank reproducibility under the primary five-system threshold and the three-system sensitivity. Bounds are repeated-split percentiles. Five systems is primary.

Min. systems	Raw rank	Lower	Upper	Groups	Assets
3	0.6262	0.5813	0.6626	1,227	13
5	0.5857	0.5264	0.6304	1,009	9

Table 5: Complete local-law universe by source and estimand. Stable class is an additional requirement, not synonymous with nonconstancy.

Source	Relationship type	Eligible	Nonconstant	Stable class
A01	paired response law	19	13	11
A03	paired response law	43	0	0
C06	structured unpaired performance demand relation	36	3	1
C16	paired response law	93	3	3
C17	paired response law	120	54	22

Table 6: Complete natural-interaction universe. Each source retains its own natural outcome scale. Confirmatory common-root tests and secondary structured-unpaired descriptions are separated. A supported C17 test must also pass its first-stage stochastic bound.

Source	Contrast	Role	Effect	Lower	Upper	Support
A03	episodic x semantic critique	confirmatory	-0.03202	-0.04328	-0.02003	Yes
C05	concept x hint / Direct C	confirmatory	0.01574	-0.01759	0.05093	No
C05	concept x hint / Direct C	secondary descriptive	0.01775	-0.02314	0.06127	No
C05	concept x hint / Example C	confirmatory	0.006481	-0.02685	0.03889	No
C05	concept x hint / Example C	secondary descriptive	0.007125	-0.03476	0.04725	No
C05	concept x hint / Misleading C	confirmatory	-0.01389	-0.04632	0.01944	No
C05	concept x hint / Misleading C	secondary descriptive	-0.01375	-0.04977	0.02115	No
C05	concept x hint / Name C	confirmatory	0.01944	-0.01574	0.05463	No
C05	concept x hint / Name C	secondary descriptive	0.02	-0.01738	0.06239	No
C05	concept x hint / Problem C	confirmatory	0.01759	-0.01574	0.05	No
C05	concept x hint / Problem C	secondary descriptive	0.01413	-0.01163	0.04163	No
C05	concept x hint / Root C	confirmatory	0.01759	-0.01574	0.05185	No
C05	concept x hint / Root C	secondary descriptive	0.02087	-0.01813	0.05713	No
C16	constraint second difference at 1	confirmatory	-0.1261	-0.1928	-0.06906	Yes
C16	constraint second difference at 2	confirmatory	0.01212	-0.05458	0.08485	No
C17	temperature x evidence / bridge 2facts / ner	confirmatory	-0.078	-0.1087	-0.04667	Yes
C17	temperature x evidence / bridge 2facts / remove half	confirmatory	-0.046	-0.08333	-0.009983	No

Continued from previous page

Source	Contrast	Role	Effect	Lower	Upper	Support
C17	temperature x evidence / bridge 2facts / replace half	confirmatory	-0.024	-0.05867	0.014	No
C17	temperature x evidence / bridge 3facts / ner	confirmatory	-0.04867	-0.08733	-0.01067	No
C17	temperature x evidence / bridge 3facts / remove half	confirmatory	-0.028	-0.06667	0.01335	No
C17	temperature x evidence / bridge 3facts / replace half	confirmatory	-0.01733	-0.06002	0.026	No
C17	temperature x evidence / bridge 4facts / ner	confirmatory	-0.02733	-0.06068	0.00335	No
C17	temperature x evidence / bridge 4facts / remove half	confirmatory	-0.08067	-0.132	-0.03197	Yes
C17	temperature x evidence / bridge 4facts / replace half	confirmatory	-0.048	-0.09202	-0.003983	No
C17	temperature x evidence / comparison 2facts / ner	confirmatory	-0.01683	-0.04636	0.01401	No
C17	temperature x evidence / comparison 2facts / remove half	confirmatory	-0.03967	-0.07901	-0.001325	No
C17	temperature x evidence / comparison 2facts / replace half	confirmatory	0.0045	-0.02935	0.04001	No
C17	temperature x evidence / comparison 3facts / ner	confirmatory	-0.04067	-0.06667	-0.016	Yes
C17	temperature x evidence / comparison 3facts / remove half	confirmatory	-0.038	-0.06533	-0.007983	Yes
C17	temperature x evidence / comparison 3facts / replace half	confirmatory	-0.03333	-0.05802	-0.01	No
C17	temperature x evidence / comparison 4facts / ner	confirmatory	-0.02333	-0.04867	0.00335	No
C17	temperature x evidence / comparison 4facts / remove half	confirmatory	-0.03867	-0.07067	-0.008	No
C17	temperature x evidence / comparison 4facts / replace half	confirmatory	-0.04	-0.07803	-0.00465	No

Table 7: Surface factorial prediction: RMSE gain of pairwise over additive surface features on the frozen evaluated DOVE-Lite groups. Other surface sources still contribute descriptive dispersion.

Source	OOD target	Mean gain	Median gain	Positive fraction	Groups
A06 DOVE Lite Selected	condition OOD	0.003152	0.002058	1	26
A06 DOVE Lite Selected	task OOD	0.002356	0.0009847	0.7385	26

E CANDIDATE CAPABILITY VALIDITY AND EVIDENCE POPULATIONS

The nine candidates are prespecified hypotheses, not latent discoveries. The complete state matrix is split into three readable tables instead of compressed into a single small heatmap. It includes the positive local and cross-operationalization findings, unsupported minimality, missing matched specificity, candidate measurement-quality states, independent links, common support, and the final conflict. Predictive specificity remains explicitly supporting evidence.

Information utilization illustrates why validity cannot be reduced to a count of positive cells: local and cross-operation support coexist with missing specificity, discrimination, quality, controls, independent-source and external components. Reasoning integrity illustrates a different boundary: it has supported local behavior and predictive specificity but affirmative discriminant conflict. Missing evidence and conflicting evidence call for different follow-up designs.

The cohort map documents the population attached to each candidate and criterion. Common support is not the smallest intersection across the repository. Each within-source candidate is evaluated on its compatible population, while independent links require their own source-specific rows under a consistent construct and exact-system contract. Pairwise diagnostic correlations are reported without overriding the resulting candidate-specific production states.

Table 8: Raw-rank local reliability on the frozen candidate-compatible production population. This population differs from broader exploratory candidate diagnostics.

Candidate	Estimate	Lower	Upper
semantic adaptation	0.5429	0.3714	0.6571
reasoning integrity	0.4981	0.4	0.5966
planning procedural control	—	—	—
information utilization	0.6851	0.6243	0.7459
feedback guided repair	—	—	—
misleading information resistance	0.4699	0.3307	0.6384
context load robustness	—	—	—
constraint handling	—	—	—
surface robustness	—	—	—

Table 9: Held-root task transfer on the frozen candidate-compatible production population. This population differs from broader exploratory candidate diagnostics.

Candidate	Estimate	Lower	Upper
semantic adaptation	0.4857	0.3765	0.5429
reasoning integrity	0.4365	0.3694	0.4916
planning procedural control	—	—	—
information utilization	0.6133	0.5999	0.669
feedback guided repair	—	—	—
misleading information resistance	0.4452	0.3937	0.5155
context load robustness	—	—	—
constraint handling	—	—	—
surface robustness	—	—	—

Table 10: Complete candidate evidence matrix: local. S = supported; ENS = estimable but not supported; NI = not identifiable; C = conflicting. Predictive specificity is supporting evidence, not a promotion conjunct.

Candidate	Local	Cross-operation	Cross-source	External
semantic adaptation	S	NI	NI	NI
reasoning integrity	S	NI	NI	NI
planning procedural control	NI	NI	NI	NI
information utilization	S	S	NI	NI
feedback guided repair	NI	NI	NI	NI
misleading information resistance	S	NI	NI	NI
context load robustness	NI	NI	NI	NI
constraint handling	NI	NI	NI	NI
surface robustness	NI	NI	NI	NI

Table 11: Complete candidate evidence matrix: specificity. S = supported; ENS = estimable but not supported; NI = not identifiable; C = conflicting. Predictive specificity is supporting evidence, not a promotion conjunct.

Candidate	Matched specificity	Predictive specificity	Minimality	Discrimination
semantic adaptation	NI	NI	ENS	NI
reasoning integrity	NI	S	ENS	C
planning procedural control	NI	NI	NI	NI
information utilization	NI	NI	ENS	NI
feedback guided repair	NI	NI	ENS	NI
misleading information resistance	NI	S	ENS	NI
context load robustness	NI	NI	NI	NI
constraint handling	NI	NI	ENS	NI
surface robustness	NI	NI	NI	NI

Table 12: Complete candidate evidence matrix: connection. S = supported; ENS = estimable but not supported; NI = not identifiable; C = conflicting. Predictive specificity is supporting evidence, not a promotion conjunct.

Candidate	Controls	Quality	Common support	Overall
semantic adaptation	NI	S	S	NI
reasoning integrity	NI	S	S	C
planning procedural control	NI	NI	NI	NI
information utilization	NI	NI	S	NI
feedback guided repair	NI	NI	NI	NI
misleading information resistance	NI	NI	S	NI
context load robustness	NI	NI	NI	NI
constraint handling	NI	NI	NI	NI
surface robustness	NI	NI	NI	NI

Table 13: Production cross-operationalization evidence. All required operation pairs remain in the target question; one candidate, information utilization, is supported.

Candidate	Estimate	Lower	Upper	Required pairs	Estimable pairs	State
semantic adaptation	—	—	—	399	21	NI
reasoning integrity	—	—	—	2179	2176	NI
planning procedural control	—	—	—	0	0	NI
information utilization	0.4714	0.0993	0.7141	2146	2146	S
feedback guided repair	—	—	—	2697	0	NI
misleading information resistance	—	—	—	7	6	NI
context load robustness	—	—	—	0	0	NI
constraint handling	—	—	—	378	0	NI
surface robustness	—	—	—	0	0	NI

Table 14: Complete estimable minimality comparisons. Candidate-level support requires all required comparisons on compatible support; a positive row alone is insufficient.

Candidate	Ablation / comparator	Gain	Lower	Upper	Support
constraint handling	coarse	2.545e-08	-1.006e-07	1.505e-07	No
constraint handling	full information removed	8.008e-05	-0.0001277	0.0002701	No
constraint handling	merge / planning procedural control	0	0	0	No
constraint handling	merge / reasoning integrity	-1.243e-08	-8.84e-07	8.244e-07	No
constraint handling	profile removed	1.507e-05	-4.354e-05	7.735e-05	No
feedback guided repair	coarse	9.293e-08	-3.602e-07	5.392e-07	No
feedback guided repair	full information removed	3.225e-05	-3.915e-05	0.000102	No
feedback guided repair	merge / information utilization	1.936e-08	-3.482e-07	4.143e-07	No
feedback guided repair	merge / misleading information resistance	-6.548e-08	-7.023e-07	6.461e-07	No
feedback guided repair	profile removed	2.633e-05	-5.286e-05	0.0001075	No
information utilization	coarse	4.496e-08	-2.422e-08	1.165e-07	No
information utilization	full information removed	6.419e-08	-9.169e-08	2.207e-07	No
information utilization	merge / feedback guided repair	1.067e-08	-4.464e-08	6.839e-08	No
information utilization	merge / misleading information resistance	5.458e-08	-1.579e-07	2.877e-07	No
information utilization	profile removed	-3.07e-07	-1.8e-05	1.76e-05	No
misleading information resistance	coarse	0.00221	0.001371	0.003014	Yes
misleading information resistance	full information removed	0.00221	0.001412	0.003021	Yes
misleading information resistance	merge / information utilization	5.675e-06	-8.871e-06	1.929e-05	No
misleading information resistance	merge / reasoning integrity	0.00221	0.001403	0.003048	Yes
misleading information resistance	profile removed	0.002145	0.001304	0.002993	Yes
reasoning integrity	coarse	0.0004342	0.0002439	0.0006435	Yes
reasoning integrity	full information removed	0.0004343	0.0002641	0.000631	Yes
reasoning integrity	merge / misleading information resistance	0.0004342	0.0002476	0.0006413	Yes
reasoning integrity	merge / semantic adaptation	-6.348e-08	-2.278e-07	8.341e-08	No
reasoning integrity	profile removed	0.0004223	0.0002394	0.0006288	Yes
semantic adaptation	coarse	-4.072e-08	-5.408e-07	4.628e-07	No
semantic adaptation	full information removed	-2.924e-07	-8.744e-07	2.833e-07	No
semantic adaptation	merge / reasoning integrity	-2.033e-09	-5.712e-08	5.066e-08	No
semantic adaptation	merge / surface robustness	0	0	0	No
semantic adaptation	profile removed	1.453e-06	-1.914e-05	1.959e-05	No

Table 15: Complete registered matched-specificity comparisons. Candidate support requires all registered neighbors plus candidate-level joint uncertainty; an isolated estimable pair cannot replace a missing neighbor.

Target	Neighbor	Contrast	Lower	Upper	Pair state
misleading information resistance	reasoning integrity	-0.001143	-0.01267	0.009084	ENS
reasoning integrity	misleading information resistance	0.001143	-0.009682	0.01194	ENS

Continued from previous page

Target	Neighbor	Contrast	Lower	Upper	Pair state
reasoning integrity	semantic adaptation	—	—	—	NI
semantic adaptation	reasoning integrity	—	—	—	NI

Table 16: Held-root predictive specificity against the registered coarse taxonomy, retaining the measurement-quality scope of each row. This is supporting semantic evidence and cannot substitute for matched specificity.

Candidate	Quality scope	Predictive gain	Lower	Upper	State
constraint handling	all admissible	—	—	—	NI
constraint handling	non objective	—	—	—	NI
feedback guided repair	all admissible	—	—	—	NI
feedback guided repair	non objective	—	—	—	NI
feedback guided repair	objective only	—	—	—	NI
information utilization	all admissible	—	—	—	NI
information utilization	non objective	—	—	—	NI
misleading information resistance	all admissible	0.001964	0.0009775	0.002976	S
misleading information resistance	non objective	0.001875	0.0009107	0.002887	S
reasoning integrity	all admissible	0.001939	0.0009944	0.002933	S
reasoning integrity	non objective	0.001954	0.0009413	0.00298	S
reasoning integrity	objective only	—	—	—	NI
semantic adaptation	all admissible	—	—	—	NI
semantic adaptation	objective only	—	—	—	NI

Table 17: Prespecified discriminant-neighbor family. Adjusted intervals test the absolute-correlation boundary 0.8, not correlation against zero. No detected conflict is not positive discriminant support. Pairwise rows do not replace candidate-specific production states.

Candidate	Neighbor	Systems	Rank	Lower	Upper	State
semantic adaptation	reasoning integrity	6	1	1	1	C
semantic adaptation	surface robustness	0	—	—	—	NI
reasoning integrity	misleading information resistance	17	-0.4047	-0.9185	0.5644	No conflict detected
reasoning integrity	constraint handling	0	—	—	—	NI
planning procedural control	context load robustness	0	—	—	—	NI
planning procedural control	constraint handling	0	—	—	—	NI
information utilization	feedback guided repair	0	—	—	—	NI
information utilization	misleading information resistance	0	—	—	—	NI
feedback guided repair	misleading information resistance	0	—	—	—	NI
context load robustness	surface robustness	0	—	—	—	NI

Table 18: Candidate measurement-quality evidence. Supporting proxy observations cannot independently establish the required criterion.

Candidate	Objective rows	Admissible rows	Proxy rows	Common systems	State
semantic adaptation	106467	0	0	17	S
reasoning integrity	69270	6500	314466	6	S
planning procedural control	0	0	0	0	NI
information utilization	0	0	348778	0	NI
feedback guided repair	1852	0	13417	1	NI
misleading information resistance	0	5200	14310	0	NI
context load robustness	0	0	0	0	NI
constraint handling	0	0	13198	0	NI
surface robustness	0	0	0	0	NI

Table 19: Complete candidate-specific cohort map. Within-source axes reuse the candidate-compatible population; independent source, control and external axes retain their own explicit bridge requirements.

Candidate	Evidence axis	Sources	Systems	Roots	Operations
semantic adaptation	controls	A01, C13	17	9047	27
semantic adaptation	cross operationalization	A01, C13	17	9047	27
semantic adaptation	cross source	A01, C13	17	9047	27
semantic adaptation	discriminant	A01, C13	17	9047	27
semantic adaptation	external	A01, C13	17	9047	27
semantic adaptation	local	A01, C13	17	9047	27
semantic adaptation	matched specificity	A01, C13	17	9047	27
semantic adaptation	measurement quality	A01, C13	17	9047	27
semantic adaptation	minimality	A01, C13	17	9047	27
semantic adaptation	predictive specificity	A01, C13	17	9047	27
reasoning integrity	controls	C05, C13, C15, C17	28	1518	81
reasoning integrity	cross operationalization	C05, C13, C15, C17	28	1518	81
reasoning integrity	cross source	C05, C13, C15, C17	28	1518	81
reasoning integrity	discriminant	C05, C13, C15, C17	28	1518	81
reasoning integrity	external	C05, C13, C15, C17	28	1518	81
reasoning integrity	local	C05, C13, C15, C17	28	1518	81
reasoning integrity	matched specificity	C05, C13, C15, C17	28	1518	81
reasoning integrity	measurement quality	C05, C13, C15, C17	28	1518	81
reasoning integrity	minimality	C05, C13, C15, C17	28	1518	81
reasoning integrity	predictive specificity	C05, C13, C15, C17	28	1518	81
planning procedural control	controls		0	0	0
planning procedural control	cross operationalization		0	0	0
planning procedural control	cross source		0	0	0
planning procedural control	discriminant		0	0	0
planning procedural control	external		0	0	0

Continued from previous page

Candidate	Evidence axis	Sources	Systems	Roots	Operations
planning procedural control	local		0	0	0
planning procedural control	matched specificity		0	0	0
planning procedural control	measurement quality		0	0	0
planning procedural control	minimality		0	0	0
planning procedural control	predictive specificity		0	0	0
information utilization	controls	C14, C17	12	1442	68
information utilization	cross operationalization	C14, C17	12	1442	68
information utilization	cross source	C14, C17	12	1442	68
information utilization	discriminant	C14, C17	12	1442	68
information utilization	external	C14, C17	12	1442	68
information utilization	local	C14, C17	12	1442	68
information utilization	matched specificity	C14, C17	12	1442	68
information utilization	measurement quality	C14, C17	12	1442	68
information utilization	minimality	C14, C17	12	1442	68
information utilization	predictive specificity	C14, C17	12	1442	68
feedback guided repair	controls	A04, C16	3	4587	96
feedback guided repair	cross operationalization	A04, C16	3	4587	96
feedback guided repair	cross source	A04, C16	3	4587	96
feedback guided repair	discriminant	A04, C16	3	4587	96
feedback guided repair	external	A04, C16	3	4587	96
feedback guided repair	local	A04, C16	3	4587	96
feedback guided repair	matched specificity	A04, C16	3	4587	96
feedback guided repair	measurement quality	A04, C16	3	4587	96
feedback guided repair	minimality	A04, C16	3	4587	96
feedback guided repair	predictive specificity	A04, C16	3	4587	96
misleading information resistance	controls	C05, C15	17	471	6
misleading information resistance	cross operationalization	C05, C15	17	471	6
misleading information resistance	cross source	C05, C15	17	471	6
misleading information resistance	discriminant	C05, C15	17	471	6
misleading information resistance	external	C05, C15	17	471	6
misleading information resistance	local	C05, C15	17	471	6
misleading information resistance	matched specificity	C05, C15	17	471	6
misleading information resistance	measurement quality	C05, C15	17	471	6
misleading information resistance	minimality	C05, C15	17	471	6
misleading information resistance	predictive specificity	C05, C15	17	471	6
context load robustness	controls		0	0	0
context load robustness	cross operationalization		0	0	0
context load robustness	cross source		0	0	0
context load robustness	discriminant		0	0	0
context load robustness	external		0	0	0
context load robustness	local		0	0	0
context load robustness	matched specificity		0	0	0
context load robustness	measurement quality		0	0	0
context load robustness	minimality		0	0	0
context load robustness	predictive specificity		0	0	0

Continued from previous page

Candidate	Evidence axis	Sources	Systems	Roots	Operations
constraint handling	controls	C16	2	1691	28
constraint handling	cross operationalization	C16	2	1691	28
constraint handling	cross source	C16	2	1691	28
constraint handling	discriminant	C16	2	1691	28
constraint handling	external	C16	2	1691	28
constraint handling	local	C16	2	1691	28
constraint handling	matched specificity	C16	2	1691	28
constraint handling	measurement quality	C16	2	1691	28
constraint handling	minimality	C16	2	1691	28
constraint handling	predictive specificity	C16	2	1691	28
surface robustness	controls		0	0	0
surface robustness	cross operationalization		0	0	0
surface robustness	cross source		0	0	0
surface robustness	discriminant		0	0	0
surface robustness	external		0	0	0
surface robustness	local		0	0	0
surface robustness	matched specificity		0	0	0
surface robustness	measurement quality		0	0	0
surface robustness	minimality		0	0	0
surface robustness	predictive specificity		0	0	0

F CALIBRATION, INDEPENDENT CONTROLS, AND EXTERNAL TARGETS

The calibration topology contains 24 synthetic systems, 16 roots, two imbalanced sources, two operationalizations, incomplete cells, and explicit cross-source, control, and external links. It has a missing fraction of approximately 0.159 and 20 exact cross-source overlapping systems. These designed connections make the validity target identifiable; they are not inferred connections in the released collection.

For each signal condition, 100 worlds select thresholds and 100 disjoint worlds qualify them. The selected local reliability threshold is 0.1, task and cross-operation/source thresholds are zero, specificity and minimality gains must exceed zero, the discriminant absolute-correlation boundary is 0.8, and common support requires five systems. The uncertainty-aware conjunction, not a point estimate alone, determines support. Qualification evaluates the complete predetermined world bank before permitting promotion. The released-data topology contributes neither threshold selection nor qualification.

The independent qualification results and the 18 zero-fail-open conflict tests establish that the candidate procedure can detect a connected strong signal while rejecting these null and conflicting worlds. They do not resolve the released candidate populations' missing bridges. Secondary observed-topology diagnostics use their own identifiability summaries; production information utilization remains cross-operationalization supported regardless of a differently scoped diagnostic cell.

Table 20: Independent qualification of the production candidate procedure after disjoint-world selection. Null measures any false promotion; non-null rows summarize detection. Released topology was excluded from selection and qualification.

Signal	Worlds	Rate	Wilson lower	Wilson upper
Null	100	0	0	0.03699
weak	100	0	0	0.03699
medium	100	0.93	0.8625	0.9657
strong	100	0.93	0.8625	0.9657

F.1 INDEPENDENT CONTROL AND EXTERNAL COVERAGE

Control-source availability is complete for the required five sources, but exact candidate-control linkage is not. Forty control model labels should not be read as forty linked deployed systems. The external inventory likewise records the released evidence without equating source presence with same-target prediction. AgentSuite, WildClaw, METR, HalBench, paired safety, Terminal Wrench and CrossEval contribute different supporting targets; observational and request-ecology sources have separate roles. Anonymous CrossEval response slots cannot be reverse-engineered into model identities. WildClaw coexistence of successful and failed tool results is not a measured temporal-recovery process.

Table 21: All five required independent control sources are represented. Source availability does not establish candidate-control predictive linkage: the latter is NI under exact deployed-system identity.

Source	Rows	Model labels	Roots	Control units	Source state
C09	41364	18	383	6	Evaluated
C10	1264	2	79	8	Evaluated
C18	281	9	6	14	Evaluated
C19	2956	6	354	3	Evaluated
S01	870	7	262	178	Evaluated

Table 22: Complete supporting-source inventory. Observational and ecology measurements are retained in their own programs and do not substitute for same-target external validity.

Source	Program	Rows	Systems	Roots	State
C02 AgentSuite	external or secondary	435,309	236	15,125	Available
C04 WildClawBench Trajectories	external or secondary	720	12	60	Available
C06 METR Time Horizon 1 1	external or secondary	24,008	95	228	Available
S03 HalBench	external or secondary	12,236	4	3,196	Available
S04 PairedSafety	external or secondary	3,768	6	628	Available
S07 Terminal Wrench	external or secondary	2,387	3	331	Available
S08 CrossEval	external or secondary	4,200	3	1,400	Available
C07 BenchMIRT	observational geometry	—	—	—	Source-local analysis
C08 ADeLe	observational geometry	—	—	—	Source-local analysis
C12 HELM Capabilities Selected	observational geometry	—	—	—	Source-local analysis
S06 MedHELM Public Selected	observational geometry	—	—	—	Source-local analysis
S09 OpenEval	observational geometry	—	—	—	Source-local analysis
S05 ATBench	ecology	1,000	0	—	Task labels only
S02 LMSYS-Chat-1M	ecology	1,000,000	—	—	Ecology coverage

G REPRESENTATION, OVERALL STRENGTH, AND RESIDUAL ORGANIZATION

The primary task-OOD paired headline is asset-equal and uses shared out-of-fold rows. The OOD hierarchy table below reports mean fold losses under each information condition and target; its task-OOD difference therefore need not numerically equal the asset-equal paired headline. This distinction is part of the estimand. Conditions A, B, C and D respectively compare model identity,

deployed-system identity, matched source context, and exact-unit ablation. Task, intervention, operationalization, source and system OOD are different generalization questions.

The primary paired advantage of semantics is -0.01150497 $[-0.01207774, -0.01105286]$. Under the equal-cell matched-objective sensitivity it is -0.00127666 $[-0.00162095, -0.00103168]$, across 20 repeats. The matched-source-context hierarchy also shows that source OOD has a positive mean semantic-minus-low-rank difference, whereas the primary task-OOD difference is negative. This does not license one generic winner across targets, nor does a cold-start fallback demonstrate transferred latent interactions.

The complete strength table shows the same predictor ordering in each of five folds. The additive residual path uses 486,626 rows and 20,948 canonical roots. Surface/operationalization contributes information after semantic and context blocks; the separate system–response-unit key comparator remains more predictive. Six block orders, eight residual stages, 1,341 spectrum components, 1,536 residual crossing panels, and 14 final geometry-uncertainty panels characterize the remaining structure. Their complete electronic spectra and crossing catalogs retain stage and source identifiers. Rank and entropy describe the estimated residual panel; neither estimates a true number of capabilities.

Table 23: Matched-information OOD hierarchy: mean fold MSE on the frozen comparisons. These fold means differ in weighting from the primary asset-equal paired headline; their task-OOD difference must not replace that headline.

Information condition	OOD target	Semantic	Low rank	Richer measured	Bias only
A model identity	task OOD	0.1328	0.1522	0.1333	0.1702
B system identity	task OOD	0.1327	0.1467	0.1333	0.1575
C matched source context	intervention OOD	0.1554	0.1767	0.2059	0.16
C matched source context	operationalization OOD	0.1707	0.196	0.1991	0.1952
C matched source context	source OOD	0.2462	0.232	0.2395	0.2395
C matched source context	system OOD	0.1688	0.2001	0.2272	0.2272
C matched source context	task OOD	0.1328	0.1467	0.1333	0.1575
D exact unit ablated	intervention OOD	0.1537	0.1749	0.2059	0.16
D exact unit ablated	operationalization OOD	0.1698	0.1882	0.1991	0.1952
D exact unit ablated	source OOD	0.2402	0.2473	0.2395	0.2395
D exact unit ablated	system OOD	0.1794	0.2037	0.2272	0.2272
D exact unit ablated	task OOD	0.1483	0.1692	0.1333	0.1575

Table 24: Cold-start coverage. Identity and response-unit absence remove different predictive information; target-specific results must retain these distinctions.

Information condition	OOD target	Unseen identity	Unseen unit	Bias MSE
A model identity	task OOD	0	0.001093	0.1702
B system identity	task OOD	1.627e-05	0.001093	0.1575
C matched source context	intervention OOD	0.003384	1	0.16
C matched source context	operationalization OOD	0.2711	1	0.1952
C matched source context	source OOD	1	1	0.2395
C matched source context	system OOD	1	0.08472	0.2272
C matched source context	task OOD	1.627e-05	0.001093	0.1575
D exact unit ablated	intervention OOD	0.003384	1	0.16
D exact unit ablated	operationalization OOD	0.2711	1	0.1952
D exact unit ablated	source OOD	1	1	0.2395
D exact unit ablated	system OOD	1	0.08472	0.2272
D exact unit ablated	task OOD	1.627e-05	0.001093	0.1575

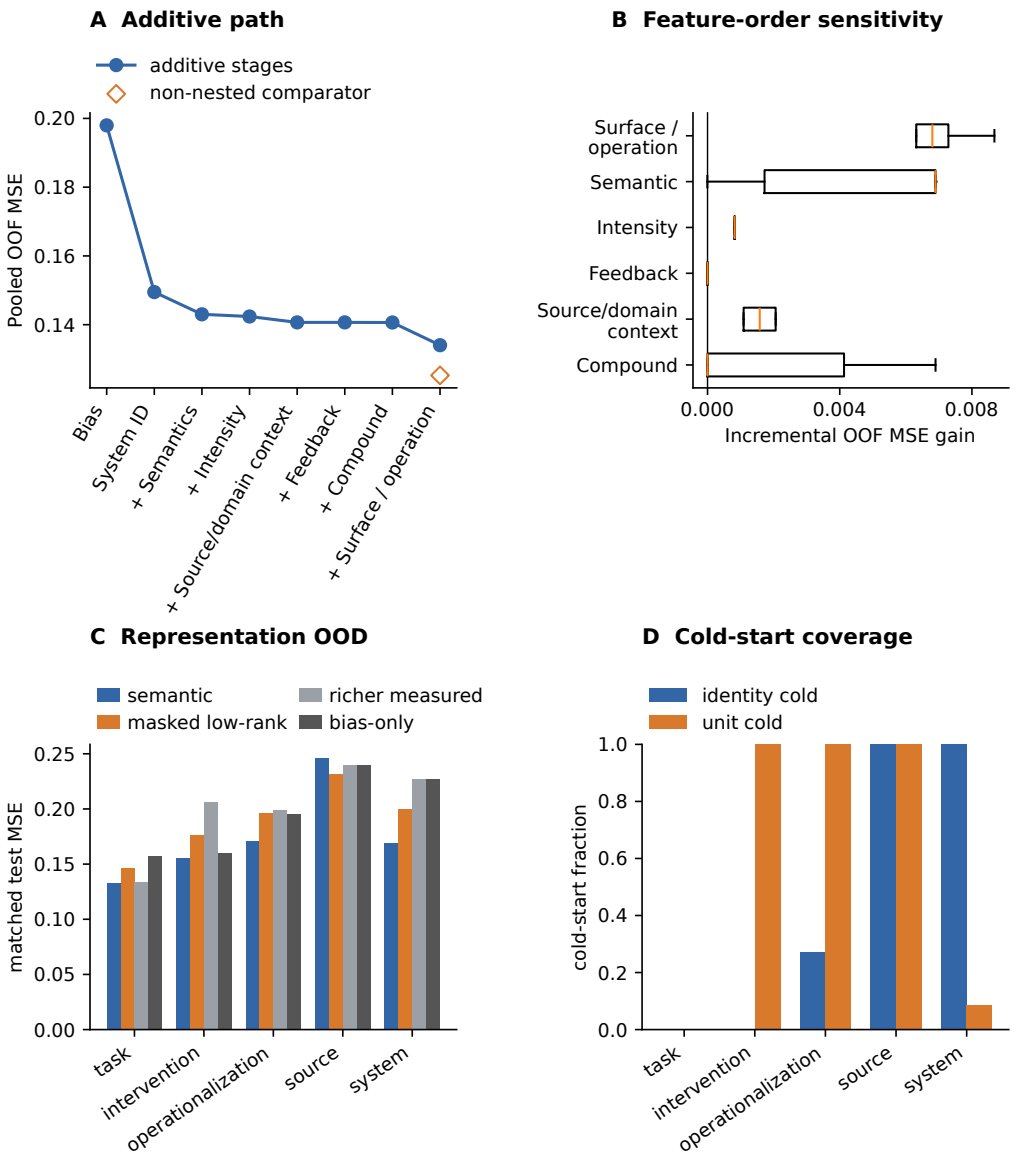

Figure 6: Predictive residual and OOD structure. The additive sequence and separately fitted system–response-unit key comparator are distinct (A); feature-block order changes incremental attribution (B); OOD losses are reported by target rather than pooled (C); unseen identities and response units restrict what transfer can mean (D). The plotted hierarchy uses mean-fold summaries, while the primary main-text task-OOD difference is asset-equal over matched paired predictions.

Table 25: Complete five-fold overall-strength comparison. Target roots are excluded from strength estimation in every row. Main-text means are simple means of the five displayed MSEs.

Fold	Predictor	MSE	MAE	Test rows
0	System-ID only	0.1483	0.241	100795
0	Conditional only	0.1369	0.2348	100795
0	System ID + conditional	0.1332	0.2293	100795
1	System-ID only	0.1462	0.2392	94206
1	Conditional only	0.1366	0.2346	94206
1	System ID + conditional	0.1323	0.2284	94206
2	System-ID only	0.1561	0.2458	97261
2	Conditional only	0.1432	0.2426	97261
2	System ID + conditional	0.1392	0.237	97261
3	System-ID only	0.1509	0.2439	95275
3	Conditional only	0.1386	0.2396	95275
3	System ID + conditional	0.1347	0.2339	95275
4	System-ID only	0.146	0.2361	99089
4	Conditional only	0.1354	0.2347	99089
4	System ID + conditional	0.1308	0.2281	99089

Table 26: Held-root residual decomposition with pooled OOF-row MSE. Rows 0–7 form the ordered additive sequence; row 8 is a separately fitted non-nested system–response-unit key comparator. All use 486,626 rows and 20,948 roots. Display names map in order to the unchanged stage keys in the released decomposition; the model key denotes system ID. Feature encodings and key shrinkage are defined in Appendix A.

Step	Predictive information	OOF MSE	Incremental gain
0	Bias only	0.198	—
1	System ID	0.1495	0.04849
2	+ Semantics	0.143	0.006467
3	+ Intensity	0.1424	0.0006369
4	+ Source/domain context	0.1407	0.001716
5	+ Feedback	0.1407	-2.93e-06
6	+ Compound	0.1406	3.36e-05
7	+ Surface / operationalization	0.134	0.0066
8	System × response-unit key	0.1253	—

Table 27: Companion demand-role additive specification: predictive gains across six feature-block orders. These blocks differ from the residual sequence displayed in the main figure; neither specification is a causal variance partition.

Block	Median gain	Lower quartile	Upper quartile
asset domain	-2.154e-06	-2.376e-06	0.001434
demand role	2.525e-05	1.049e-07	0.001418
domain	0.001923	0.0004788	0.001943
feedback block	2.36e-08	7.647e-09	6.971e-07
intensity block	0.0007377	0.0007377	0.0007745
interaction block	0.002693	0.0006701	0.002704
primary construct	6.449e-07	6.449e-07	0.003398

Table 28: Complete six-order residual-block sensitivity underlying the appendix residual figure. Each order uses 4,190 held-out roots and 97,575 test rows; this sensitivity population is distinct from the five-fold pooled main decomposition.

Order	Position	Added block	MSE before	MSE after	Gain
0	0	semantic	0.1556	0.1487	0.006898
0	1	intensity	0.1487	0.1478	0.0008217
0	2	context	0.1478	0.1458	0.002063
0	3	feedback	0.1458	0.1458	6.999e-07
0	4	compound	0.1458	0.1458	1.619e-07
0	5	surface	0.1458	0.1395	0.006315
1	0	surface	0.1556	0.1469	0.00868
1	1	compound	0.1469	0.1414	0.005507
1	2	feedback	0.1414	0.1414	5.224e-08
1	3	context	0.1414	0.1403	0.001098
1	4	intensity	0.1403	0.1395	0.0008185
1	5	semantic	0.1395	0.1395	-5.443e-06
2	0	compound	0.1556	0.1487	0.006899
2	1	semantic	0.1487	0.1487	4.918e-08
2	2	feedback	0.1487	0.1487	8.452e-09
2	3	context	0.1487	0.1466	0.002061
2	4	intensity	0.1466	0.1458	0.0008234
2	5	surface	0.1458	0.1395	0.006315
3	0	semantic	0.1556	0.1487	0.006898
3	1	intensity	0.1487	0.1478	0.0008217
3	2	compound	0.1478	0.1478	-4.895e-07
3	3	feedback	0.1478	0.1478	1.165e-08
3	4	surface	0.1478	0.1405	0.007289
3	5	context	0.1405	0.1395	0.00109
4	0	semantic	0.1556	0.1487	0.006898
4	1	intensity	0.1487	0.1478	0.0008217
4	2	surface	0.1478	0.1406	0.007286
4	3	compound	0.1406	0.1405	1.627e-06
4	4	context	0.1405	0.1395	0.00109
4	5	feedback	0.1395	0.1395	4.185e-07
5	0	semantic	0.1556	0.1487	0.006898
5	1	compound	0.1487	0.1487	4.331e-07
5	2	context	0.1487	0.1466	0.002061
5	3	surface	0.1466	0.1403	0.006319
5	4	feedback	0.1403	0.1403	3.969e-07
5	5	intensity	0.1403	0.1395	0.0008187

Table 29: Final residual geometry under system-bootstrap uncertainty. These intervals describe the final residual panel, whereas the Level-3 primary geometry uses root/task bootstrap reconstruction.

Source	Systems	Units	Effective rank	Lower	Upper
A01	50	84	4.307	1.702	9.511
A02	8	16	4.098	1.826	4.063
A03	6	387	4.517	1.674	3.736
A05	8	22	1.607	1	1.996
A06	4	10	1.771	1	1.981
C01	4	7	2.342	1	2.342
C03	17	42	9.548	4.671	8.027
C05	4	85	2.53	1	2.53

Continued from previous page

Source	Systems	Units	Effective rank	Lower	Upper
C11	11	2	1.443	1.069	1.735
C13	6	21	2.999	1	2.851
C14	7	24	2.874	1.444	2.889
C15	13	9	2.102	1.244	3.089
C16	5	459	1.236	1	1.858
C17	5	438	3.192	1	3.192

H Robustness and alternative scientific specifications

The claim robustness summary contains 78 rows across nine claim families: 60 full leave-one-family-out refits, 12 evaluation-composition changes, and six alternative-scale full refits. None changes the registered major conclusion in that summary. A changed population, effect magnitude, eligible denominator, or supported group count is nevertheless not asserted to be invariant. Outcome-scale refits learn scales on training roots only and preserve every contrast's shared scale.

Measurement-quality and taxonomy robustness is a separate complete 3×2 program: all eligible, objective-only and supporting-proxy scopes crossed with primary-only and canonical multi-label membership. These variants can change local support while retaining the overall zero-promotion frontier. Supporting-proxy analysis remains a sensitivity and does not become independent evidence of measurement fidelity.

Weighting changes answer different population questions. Asset-equal headlines avoid equating a large source with a representative field population; system- and family-equal summaries change the target without changing the meaning of exact deployed-system linkage. BY and Holm are additional law multiplicity sensitivities, while coherent null banks retain the same observed estimator and support conjunction. No robustness operation supplies an unobserved external target.

Table 30: Full claim-specific robustness results. Refits and reweighting remain distinct. Construct-promotion rows report the aggregate count, not negative candidate-validity evidence; the production candidate states remain eight NI and one conflict. Other rows retain their frozen claim state.

Claim family	Specification	Operation	Estimate	State	Change
response profile reliability	family equal	EVALUATION COMPOSITION ONLY	0.6073	S	No
response profile reliability	leave family out / alibaba qwen	FULL REFIT LOFO	0.5979	S	No
response profile reliability	leave family out / allenai	FULL REFIT LOFO	0.5922	S	No
response profile reliability	leave family out / anthropic	FULL REFIT LOFO	0.606	S	No
response profile reliability	leave family out / command	FULL REFIT LOFO	0.5922	S	No
response profile reliability	leave family out / deepseek	FULL REFIT LOFO	0.6508	S	No
response profile reliability	leave family out / devstral	FULL REFIT LOFO	0.5922	S	No
response profile reliability	leave family out / glm	FULL REFIT LOFO	0.6104	S	No
response profile reliability	leave family out / google	FULL REFIT LOFO	0.6218	S	No
response profile reliability	leave family out / kimi	FULL REFIT LOFO	0.5993	S	No
response profile reliability	leave family out / meta	FULL REFIT LOFO	0.6527	S	No

Continued from previous page

Claim family	Specification	Operation	Estimate	State	Change
response profile reliability	leave family out / mimo	FULL REFIT LOFO	0.5922	S	No
response profile reliability	leave family out / minimax	FULL REFIT LOFO	0.5685	S	No
response profile reliability	leave family out / mistral	FULL REFIT LOFO	0.5306	S	No
response profile reliability	leave family out / models	FULL REFIT LOFO	0.5922	S	No
response profile reliability	leave family out / openai	FULL REFIT LOFO	0.6131	S	No
response profile reliability	leave family out / released	FULL REFIT LOFO	0.5922	S	No
construct promotion	leave family out / alibaba qwen	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / allenai	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / anthropic	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / command	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / deepseek	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / devstral	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / glm	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / google	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / kimi	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / meta	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / mimo	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / minimax	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / mistral	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / models	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / openai	FULL REFIT LOFO	0	0 promoted	No
construct promotion	leave family out / released	FULL REFIT LOFO	0	0 promoted	No
held out nonconstant laws	family equal	EVALUATION COMPOSITION ONLY	0.2612	S	No
held out nonconstant laws	leave family out / alibaba qwen	FULL REFIT LOFO	73	S	No
held out nonconstant laws	leave family out / anthropic	FULL REFIT LOFO	69	S	No
held out nonconstant laws	leave family out / deepseek	FULL REFIT LOFO	70	S	No
held out nonconstant laws	leave family out / meta	FULL REFIT LOFO	52	S	No
held out nonconstant laws	leave family out / models	FULL REFIT LOFO	71	S	No
held out nonconstant laws	leave family out / openai	FULL REFIT LOFO	33	S	No

Continued from previous page

Claim family	Specification	Operation	Estimate	State	Change
matched representation task ood	family equal	EVALUATION COMPOSITION ONLY	-0.01275	S	No
matched representation task ood	leave family out / alibaba qwen	FULL REFIT LOFO	-0.0118	S	No
matched representation task ood	leave family out / allenai	FULL REFIT LOFO	-0.01171	S	No
matched representation task ood	leave family out / anthropic	FULL REFIT LOFO	-0.01254	S	No
matched representation task ood	leave family out / command	FULL REFIT LOFO	-0.01164	S	No
matched representation task ood	leave family out / deepseek	FULL REFIT LOFO	-0.01165	S	No
matched representation task ood	leave family out / devstral	FULL REFIT LOFO	-0.01136	S	No
matched representation task ood	leave family out / glm	FULL REFIT LOFO	-0.0117	S	No
matched representation task ood	leave family out / google	FULL REFIT LOFO	-0.01115	S	No
matched representation task ood	leave family out / kimi	FULL REFIT LOFO	-0.01164	S	No
matched representation task ood	leave family out / meta	FULL REFIT LOFO	-0.01201	S	No
matched representation task ood	leave family out / mimo	FULL REFIT LOFO	-0.01151	S	No
matched representation task ood	leave family out / minimax	FULL REFIT LOFO	-0.01144	S	No
matched representation task ood	leave family out / mistral	FULL REFIT LOFO	-0.01179	S	No
matched representation task ood	leave family out / models	FULL REFIT LOFO	-0.01187	S	No
matched representation task ood	leave family out / openai	FULL REFIT LOFO	-0.01445	S	No
matched representation task ood	leave family out / released	FULL REFIT LOFO	-0.0122	S	No
feedback / helpful feedback	family equal	EVALUATION COMPOSITION ONLY	0.07149	S	No
feedback / helpful feedback	leave family out / alibaba qwen	EVALUATION COMPOSITION ONLY	0.07257	S	No
feedback / helpful feedback	leave family out / meta	EVALUATION COMPOSITION ONLY	0.08952	S	No
feedback / helpful feedback	leave family out / openai	EVALUATION COMPOSITION ONLY	0.07492	S	No
feedback / misleading feedback loss	family equal	EVALUATION COMPOSITION ONLY	0.751	S	No
feedback / misleading feedback loss	leave family out / alibaba qwen	EVALUATION COMPOSITION ONLY	0.7729	S	No
feedback / misleading feedback loss	leave family out / meta	EVALUATION COMPOSITION ONLY	0.8462	S	No

Continued from previous page

Claim family	Specification	Operation	Estimate	State	Change
feedback / misleading feedback loss	leave family out / openai	EVALUATION COMPOSITION ONLY	0.7526	S	No
common root interactions	family equal	EVALUATION COMPOSITION ONLY	7	S	No
common root interactions	leave family out / alibaba qwen	FULL REFIT LOFO	6	S	No
common root interactions	leave family out / anthropic	FULL REFIT LOFO	6	S	No
common root interactions	leave family out / deepseek	FULL REFIT LOFO	2	S	No
common root interactions	leave family out / google	FULL REFIT LOFO	6	S	No
common root interactions	leave family out / meta	FULL REFIT LOFO	10	S	No
common root interactions	leave family out / openai	FULL REFIT LOFO	2	S	No
response profile reliability	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	0.5036	S	No
held out nonconstant laws	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	73	S	No
rank crossing	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	0	ENS	No
interactions	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	9	S	No
construct promotion	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	0	0 promoted	No
matched representation task ood	outer fold train scale standardized response	FULL REFIT ALTERNATIVE SCALE	-0.1921	S	No

Table 31: All six measurement-quality by taxonomy candidate-chain variants use the same production evaluator and preserve the zero-promotion frontier.

Quality scope	Taxonomy	Local support	Cross-operation support	Promoted
all eligible	canonical multilabel	4	1	0
all eligible	primary only	4	2	0
objective only	canonical multilabel	2	0	0
objective only	primary only	1	0	0
proxy supporting only	canonical multilabel	2	1	0
proxy supporting only	primary only	1	1	0

I OBSERVATIONAL CAPABILITY SPACES AND INTERPRETABLE TASK DEMANDS

Five independent observational panels provide context for the response-based program. BenchMIRT has a 100×16 matrix, ADeLe 15×18, HELM 3×5, MedHELM 9×13, and OpenEval 456×85

before predictive eligibility. Geometry is source-local and missingness-aware. In the nested prediction program, the eligible system count can be smaller than the raw matrix count, as it is for OpenEval (431). The matrices therefore should not be merged by name or interpreted as a shared capability space.

Nested held-system/task prediction is supported in BenchMIRT and ADeLe; it is estimable but not supported in MedHELM and OpenEval, and not identifiable in HELM. The gains have source-specific scales and are tabulated separately. ADeLe's additional 18-dimensional demand analysis covers 15 systems under both task- and benchmark-OOD, with 483,240 prediction rows and 1,000 held-group bootstrap draws. These are positive supporting investigations of interpretable demand prediction, not replacements for response-based capability validation.

Table 32: Five source-internal observational prediction spaces. Gains are interpreted within each source, not compared as a common outcome scale.

Source	Systems	MSE gain	Lower	Upper	State
C07 BenchMIRT	100	0.02456	0.02092	0.02852	S
C08 ADeLe 18D	15	0.0002172	0.0001172	0.0003414	S
S06 MedHELM	9	-0.002419	-0.00489	-0.0004709	ENS
S09 OpenEval	431	-3.021	-7.544	0.1983	ENS
C12 HELM	3	—	—	—	NI

Table 33: Complete ADeLe 18-dimensional predictive results across 15 systems and both OOD targets. Gains and intervals are held-group-equal relative to the training-prevalence baseline.

Released system	OOD target	Held groups	Brier gain	Lower	Upper
DK-R1-Dist-Qwen-1.5B	benchmark OOD	20	0.03417	0.007601	0.06102
DK-R1-Dist-Qwen-1.5B	task OOD	63	0.04198	0.02624	0.05643
DK-R1-Dist-Qwen-14B	benchmark OOD	20	0.01515	-0.002915	0.03336
DK-R1-Dist-Qwen-14B	task OOD	63	0.03123	0.01267	0.05135
DK-R1-Dist-Qwen-32B	benchmark OOD	20	0.01608	0.002859	0.02897
DK-R1-Dist-Qwen-32B	task OOD	63	0.02565	0.006014	0.04612
DK-R1-Dist-Qwen-7B	benchmark OOD	20	0.02855	0.004377	0.05252
DK-R1-Dist-Qwen-7B	task OOD	63	0.04395	0.02626	0.06222
babbage-002	benchmark OOD	20	0.003161	-0.002073	0.008819
babbage-002	task OOD	63	0.004805	0.001681	0.007717
davinci-002	benchmark OOD	20	0.009206	0.0009563	0.01953
davinci-002	task OOD	63	0.01109	0.006201	0.01796
gpt-3.5-turbo	benchmark OOD	20	0.05006	0.03019	0.07484
gpt-3.5-turbo	task OOD	63	0.04241	0.0293	0.05655
gpt-4o	benchmark OOD	20	0.04737	0.03119	0.06615
gpt-4o	task OOD	63	0.04764	0.02017	0.07255
llama3d1-405b	benchmark OOD	20	0.05001	0.035	0.06802
llama3d1-405b	task OOD	63	0.04542	0.01646	0.06964
llama3d2-11b	benchmark OOD	20	0.05685	0.03518	0.07838
llama3d2-11b	task OOD	63	0.04498	0.02469	0.06439
llama3d2-90b	benchmark OOD	20	0.05305	0.03429	0.07257
llama3d2-90b	task OOD	63	0.04592	0.01918	0.06931
meta-llama/Llama-3.2-1B-Instruct	benchmark OOD	20	0.006733	-0.01864	0.0252
meta-llama/Llama-3.2-1B-Instruct	task OOD	63	0.01213	0.002289	0.01993
meta-llama/Llama-3.2-3B-Instruct	benchmark OOD	20	0.04565	0.02173	0.07091

Continued from previous page

Released system	OOD target	Held groups	Brier gain	Lower	Upper
meta-llama/Llama-3.2-3B-Instruct	task OOD	63	0.03919	0.02461	0.05435
o1-mini	benchmark OOD	20	0.01466	-0.002207	0.03168
o1-mini	task OOD	63	0.0229	0.005184	0.04241
o1 re=low	benchmark OOD	20	-0.0003661	-0.01243	0.009245
o1 re=low	task OOD	63	0.007177	-0.004223	0.01773

Table 34: Descriptive source-local observational geometry. Matrix dimensions, effective rank and top singular share are not validated capability counts.

Source	Systems	Dimensions	Effective rank	Top share
C07 BenchMIRT	100	16	5.066	0.4918
C12 HELM	3	5	1.789	0.7315
S06 MedHELM	9	13	4.619	0.455
S09 OpenEval	456	85	20.37	0.2057
C08 ADeLe 18D	15	18	1.378	0.9366

J Response atlas, contextual modulation, and request ecology

J.1 Atlas and cross-source heterogeneity

The descriptive atlas contains 2,830 rows, 206 system-candidate profiles, and 24 candidate-by-asset forest estimates. Positive candidate profiles use positive/supporting roles, whereas the full role atlas retains nuisance and control evidence. Its 1,467,246 expanded membership rows are accounted for as role-expanded observations; they are not additional independent responses beyond the unexpanded task-response count. Between-source heterogeneity is estimated only where at least three asset effects are available. The resulting large heterogeneity values reinforce the need to preserve source meaning, not an inference to universal capability axes.

Baseline-response association, task heterogeneity, model-family variation, and signed source distributions remain descriptive. Mathematical coupling to a baseline or differences in task population prevent a correlation from identifying a mechanism. The complete electronic atlas and system/task heterogeneity catalogs retain those distinctions.

J.2 Context, feedback, memory, and MultiCodeIF

The context/feedback/memory map spans eight released assets and contains 776 condition-level rows, 59 system distribution rows, and 33 interaction rows. It preserves BOULDER dialogue conditions, critique and iterative correction, LiveClaw context, CHAMP information conditions, memory settings, coding hints, and MultiCodeIF structure. The complete interaction rows appear in the conditional-response table; per-system distributions appear below, with quantiles explicitly distinguished from confidence intervals.

MultiCodeIF contributes 570 structural summary rows and 3,052 root-level self-repair rows. The typeset first-to-last-round summaries retain constraint type, language, roots and model counts. A positive lexical/proxy change is not execution correctness or a validated repair capability. This boundary permits the source-specific response science to remain visible without allowing it to establish the candidate's missing validity components.

J.3 Released request ecology

LMSYS contributes 1,000,000 conversations, 25 model labels, and 154 language labels. Mean recorded turn count is 2.015645; 266,260 rows are redacted, 6,984 contain request URLs, and

334,319 contain multiple user turns. The retained model/language, message-depth, turn-depth and request-length distributions characterize that released request population. They do not identify a probability distribution over every future user or deployment.

ATBench contains 1,000 tasks; 958 have compound tool demand. The release supplies label, risk-source, failure-mode, harm, tool multiplicity, trajectory and dialogue-depth fields, but the analyzed ecology panel contains zero system predictions. Its contribution is the structure of requests and risks, not evidence about which deployed system possesses a capability. These ecology findings explain why broad behavioral coverage and connected construct measurements are complementary design goals.

Table 35: Cross-asset response heterogeneity. Variance summaries require at least three estimable asset effects; missing entries remain unavailable.

Candidate	Assets	Positive	Negative	Tau squared	I squared
constraint handling	1	1	0	—	—
context load robustness	3	1	2	0.09424	0.9992
feedback guided repair	3	3	0	0.001438	0.9764
information utilization	5	4	1	0.00896	0.9962
misleading information resistance	3	2	1	0.06952	0.9991
planning procedural control	1	0	1	—	—
reasoning integrity	4	1	3	0.01091	0.9925
semantic adaptation	2	0	2	—	—
surface robustness	2	2	0	—	—

Table 36: Complete candidate-by-source descriptive forest estimates with canonical-root intervals. Distinct source estimands retain their native scales; these are not capability scores.

Candidate	Source	Response	Lower	Upper	Roots
constraint handling	C16	0.06382	0.05295	0.07468	1691
context load robustness	A02	-0.3822	-0.3985	-0.3658	800
context load robustness	C03	-0.2425	-0.3114	-0.1737	75
context load robustness	C11	0.05991	0.05383	0.066	120
feedback guided repair	A03	0.1104	0.09904	0.1218	1750
feedback guided repair	A04	0.0326	0.0205	0.0447	3061
feedback guided repair	C16	0.07815	0.06692	0.08937	1526
information utilization	A03	0.02594	0.01471	0.03716	1750
information utilization	C05	0.04537	0.02248	0.06826	270
information utilization	C11	0.05991	0.05383	0.066	120
information utilization	C14	0.187	0.1746	0.1994	421
information utilization	C17	-0.1175	-0.1313	-0.1038	600
misleading information resistance	A03	-0.4386	-0.4546	-0.4225	1750
misleading information resistance	C05	0.01239	-0.008559	0.03334	270
misleading information resistance	C15	0.01378	0.00319	0.02436	201
planning procedural control	C01	-0.09309	-0.106	-0.08015	500
reasoning integrity	C05	0.01848	-0.001736	0.03869	270
reasoning integrity	C13	-0.2989	-0.3244	-0.2735	447
reasoning integrity	C15	-0.05645	-0.07172	-0.04118	201
reasoning integrity	C17	-0.1175	-0.1313	-0.1038	600
semantic adaptation	A01	-0.239	-0.2467	-0.2312	9051
semantic adaptation	C13	-0.2989	-0.3244	-0.2735	447
surface robustness	A05	0.08708	0.07758	0.09659	540
surface robustness	A06	0.3624	0.3559	0.3688	1000

Table 37: Complete MultiCodeIF first-to-last-round self-repair summaries. The outcome is a lexical/proxy score, not released execution correctness.

Constraint type	Language	Mean change	Positive rate	Roots	Models
algorithm space complexity	mixed or unspecified	0.1091	0.1091	55	2
algorithm time complexity	mixed or unspecified	0.1455	0.1545	55	2
algorithm type	mixed or unspecified	0.04545	0.04545	55	2
code contextual	mixed or unspecified	0.05	0.05	60	2
code convention brace block formatting	cpp	0	0	9	2
code convention brace block formatting	java	0	0	10	2
code convention brace block formatting	javascript	0	0	9	2
code convention brace block formatting	mixed or unspecified	0	0	8	2
code convention brace block formatting	python	0	0	9	2
code convention brace block formatting	rust	0	0	10	2
code convention comment formatting	c++	0	0	3	2
code convention comment formatting	go	0	0	10	2
code convention comment formatting	java	0	0	10	2
code convention comment formatting	javascript	0	0	3	2
code convention comment formatting	kotlin	0	0	8	2
code convention comment formatting	mixed or unspecified	0.125	0.125	4	2
code convention comment formatting	python	0	0	9	2
code convention comment formatting	rust	0	0	8	2
code convention declaration style	cpp	0	0	7	2
code convention declaration style	go	0	0	1	2
code convention declaration style	java	0	0	9	2
code convention declaration style	javascript	0	0	8	2
code convention declaration style	mixed or unspecified	0	0	8	2
code convention declaration style	python	0	0	14	2
code convention declaration style	rust	0	0	8	2
code convention indentation spacing	c++	0	0	10	2
code convention indentation spacing	cpp	0	0	1	2

Continued from previous page

Constraint type	Language	Mean change	Positive rate	Roots	Models
code convention indentation spacing	go	0	0	12	2
code convention indentation spacing	java	0	0	7	2
code convention indentation spacing	javascript	0	0	4	2
code convention indentation spacing	kotlin	0	0	1	2
code convention indentation spacing	mixed or unspecified	0	0	4	2
code convention indentation spacing	python	0	0	8	2
code convention indentation spacing	rust	0	0	8	2
code convention naming convention	cpp	0	0	5	2
code convention naming convention	go	0	0	7	2
code convention naming convention	java	0	0	4	2
code convention naming convention	javascript	0	0	3	2
code convention naming convention	kotlin	0	0	5	2
code convention naming convention	mixed or unspecified	0	0	12	2
code convention naming convention	python	0	0	12	2
code convention naming convention	rust	0	0	6	2
data structure operation	cpp	0	0	11	2
data structure operation	go	0.0625	0.0625	8	2
data structure operation	java	0	0	14	2
data structure operation	python	0.04545	0.04545	11	2
data structure operation	rust	0	0	11	2
data structure size	cpp	0	0	13	2
data structure size	go	0.05556	0.05556	9	2
data structure size	java	0	0	9	2
data structure size	python	0.07692	0.07692	13	2
data structure size	rust	0	0	11	2
data structure type	cpp	0.1154	0.1154	13	2
data structure type	go	0.125	0.125	8	2
data structure type	java	0	0	13	2
data structure type	python	0.2	0.2	10	2
data structure type	rust	0.04545	0.04545	11	2
example	mixed or unspecified	0.03077	0.03077	65	2
input output input order	cpp	0.03846	0.03846	13	2
input output input order	java	0.03571	0.03571	14	2
input output input order	python	0.1333	0.1333	15	2
input output input order	rust	0.03846	0.03846	13	2
input output input range	cpp	0.1	0.1	10	2
input output input range	java	0.05	0.1	10	2
input output input range	python	0.05	0.05	10	2
input output input range	rust	0.125	0.125	8	2

Continued from previous page

Constraint type	Language	Mean change	Positive rate	Roots	Models
input output input type	cpp	0	0	13	2
input output input type	java	0	0	14	2
input output input type	python	0.03333	0.03333	15	2
input output input type	rust	0.1923	0.1923	13	2
input output output order	go	0	0	14	2
input output output order	python	0.2381	0.2619	21	2
input output output order	rust	0	0	20	2
input output output range	cpp	-0.0625	0	8	2
input output output range	java	0.1667	0.1667	9	2
input output output range	python	0.04167	0.04167	12	2
input output output range	rust	0.05556	0.05556	9	2
input output output type	cpp	0	0	13	2
input output output type	java	0.03571	0.03571	14	2
input output output type	python	0	0	15	2
input output output type	rust	0.1154	0.1154	13	2
non functional requirements	mixed or unspecified	0.615	0.62	100	2
situation	mixed or unspecified	0.3	0.3083	60	2
syntax advanced syntax	go	0.0625	0.0625	8	2
syntax advanced syntax	javascript	0	0	12	2
syntax advanced syntax	kotlin	0	0	10	2
syntax advanced syntax	python	0	0.03125	16	2
syntax advanced syntax	rust	0	0	9	2
syntax framework	cpp	0	0	6	2
syntax framework	java	0	0	14	2
syntax framework	javascript	0	0	14	2
syntax framework	python	0	0	14	2
syntax framework	rust	0	0	7	2
syntax function	cpp	0	0	4	2
syntax function	go	0	0	7	2
syntax function	java	0	0	4	2
syntax function	javascript	0	0	11	2
syntax function	mixed or unspecified	0.04545	0.04545	11	2
syntax function	python	0	0	8	2
syntax function	rust	0	0	10	2
syntax language type	c++	0	0	4	2
syntax language type	java	0	0	10	2
syntax language type	javascript	0	0	6	2
syntax language type	mixed or unspecified	-0.02083	0	24	2
syntax language type	python	0	0	12	2
syntax language type	rust	0	0	7	2
syntax language version	java	0	0	22	2
syntax language version	python	0	0	14	2
syntax language version	rust	0.1136	0.1136	22	2
syntax library	go	0	0	12	2
syntax library	javascript	0	0	12	2
syntax library	mixed or unspecified	0	0	10	2
syntax library	python	0	0	10	2
syntax library	rust	0.04545	0.04545	11	2

Table 38: Complete source-specific context, feedback and memory system distributions. Quantiles describe available conditions; they are not uncertainty intervals or cross-source capability comparisons.

Source	Deployed system	Mean	10th pct.	90th pct.	Cond.
A02	claude-sonnet-4.5 / boulder released parser t0	-0.2087	-1	0	2
A02	command-a:111b-03-2025-q4 K M / boulder released parser t0	-0.4029	-1	0	2
A02	deepseek-v3.2 / boulder released parser t0	-0.3115	-1	0	2
A02	gemini-2.5-flash / boulder released parser t0	-0.2119	-1	0	2
A02	mistral-small:24b-instruct-2501-q8 0 / boulder released parser t0	-0.4683	-1	0	2
A02	qwen3:235b-a22b-instruct-2507-q4 K M / boulder released parser t0	-0.3619	-1	0	2
A02	qwen3:30b-a3b-instruct-2507-q8 0 / boulder released parser t0	-0.4708	-1	0	2
A02	qwen3:4b-instruct-2507-fp16 / boulder released parser t0	-0.5229	-1	0	2
A03	gpt-4o-mini-2024-07-18 / critique learning released logs	0.01803	-1	1	9
A03	gpt-5-2025-08-07 / critique learning released logs	0.05162	-1	1	6
A03	gpt-oss-20b / critique learning released logs	0.007143	-1	1	6
A03	llama-4-scout / critique learning released logs	0.00981	-1	1	6
A03	qwen3-235b-a22b-instruct-2507 / critique learning released logs	0.03905	-1	1	6
A03	qwen3-vl-235b-a22b-thinking / critique learning released logs	0.03384	-1	1	6
A04	released system unknown / meeseeks released multilingual	0.0326	-0.4059	0.4776	1
C03	MiniMax-M2.7 / OpenClaw/Harbor	-0.2409	-0.78	0.1667	17
C03	deepseek-v4-flash / OpenClaw/Harbor	-0.3101	-0.8817	0.1652	17
C03	deepseek-v4-pro / OpenClaw/Harbor	-0.2599	-0.82	0.1367	17
C03	glm-5.1 / OpenClaw/Harbor	-0.2166	-0.7413	0.1234	17
C03	glm-5.2 / OpenClaw/Harbor	-0.2504	-0.77	0.1567	17
C03	gpt-5.5 / OpenClaw/Harbor	-0.2103	-0.87	0.1239	17
C03	kimi-k2.6 / OpenClaw/Harbor	-0.26	-0.8	0.2067	17
C03	kimi-k2.7-code / OpenClaw/Harbor	-0.218	-0.7067	0.083	17
C03	mimo-v2.5-pro / OpenClaw/Harbor	-0.2592	-0.8013	0.09333	17
C03	minimax-m3 / OpenClaw/Harbor	-0.2565	-0.7691	0.1257	17
C03	opus-4.8 / OpenClaw/Harbor	-0.2979	-0.9333	0.1224	17
C03	qwen3.5-27b / OpenClaw/Harbor	-0.05107	-0.5533	0.4483	17
C03	qwen3.5-397b-a17b / OpenClaw/Harbor	-0.2508	-0.754	0.2803	17
C03	qwen3.5-flash / OpenClaw/Harbor	-0.3331	-0.9	0.1815	17
C03	qwen3.6-27b / OpenClaw/Harbor	-0.2267	-0.7778	0.2222	17
C03	qwen3.6-flash / OpenClaw/Harbor	-0.2241	-0.8317	0.1683	17
C03	qwen3.6-plus / OpenClaw/Harbor	-0.2534	-0.9	0.1667	17
C05	google/palm-2-medium	0.009323	0	0	16
C05	openai/gpt-3.5-turbo	0.04086	-0.04581	1	16
C05	openai/gpt-4	0.009169	-1	1	16
C05	openai/gpt-4-turbo	0.01456	0	1	16
C11	claude-haiku-4-5 / memory-condition released API regime; temperature=0.0; other decoding unspecified	-0.004286	0	5.695e-14	2
C11	claude-haiku-4-5 / memory-condition released API regime; temperature=1.0; other decoding unspecified	-0.003571	0	5.695e-14	2

Continued from previous page

Source	Deployed system	Mean	10th pct.	90th pct.	Cond.
C11	deepseek-chat / memory-condition released API regime; temperature=0.0; other decoding unspecified	0.007224	-5.695e-14	0.1452	2
C11	deepseek-chat / memory-condition released API regime; temperature=1.0; other decoding unspecified	0.02852	-0.1667	0.2857	2
C11	deepseek-chat / memory-condition released API regime; temperature=1.5; other decoding unspecified	0.05875	-0.2619	0.3333	2
C11	gemini-2.5-flash / memory-condition released API regime; temperature=0.0; other decoding unspecified	0.2536	0	0.4286	2
C11	gemini-2.5-flash / memory-condition released API regime; temperature=1.0; other decoding unspecified	0.2008	-0.0005495	0.4286	2
C11	gemini-2.5-flash / memory-condition released API regime; temperature=1.5; other decoding unspecified	0.134	-0.1429	0.4286	2
C11	gpt-4.1-mini / memory-condition released API regime; temperature=0.0; other decoding unspecified	-0.0097	-0.05	5.695e-14	2
C11	gpt-4.1-mini / memory-condition released API regime; temperature=1.0; other decoding unspecified	-0.01292	-0.07692	5.695e-14	2
C11	gpt-4.1-mini / memory-condition released API regime; temperature=1.5; other decoding unspecified	0.006649	-0.07709	0.1111	2
C14	MiniMax-M2.7 / coding hint T5 unspecified decoding	0.2519	0	1	6
C14	claude-opus-4-6 / coding hint T5 unspecified decoding	0.1642	0	1	6
C14	deepseek-v3-2-251201 / coding hint T5 unspecified decoding	0.1856	0	1	6
C14	gemini-3.1-pro-preview / coding hint T5 unspecified decoding	0.04971	0	1	6
C14	glm-5.1 / coding hint T5 unspecified decoding	0.2214	0	1	6
C14	gpt-5.4 / coding hint T5 unspecified decoding	0.2408	0	1	6
C14	kimi-k2.5 / coding hint T5 unspecified decoding	0.2062	0	1	6
C16	Qwen3-1.7B / multicodeif released unknown decoding	-0.396	-1	0	3
C16	claude-3-7-sonnet-20250219 / multicodeif released unknown decoding	0.0694	0	0	7
C16	deepseek-v3-0324 / multicodeif released unknown decoding	0.4525	0	1	3
C16	gpt-4o / multicodeif released unknown decoding	0.06062	0	0	7
C16	meta-llama/Llama-3.2-3B-Instruct / multicodeif released unknown decoding	-0.2061	-1	0	3

K COMPLETE FUNDAMENTAL ORGANIZATION EVIDENCE

Table 39: Five-question Level-3 summary. The complete claim is not established; positive local or predictive results remain visible within that boundary.

Question	Empirical answer	State
Recoverability	Strong-world exact recovery 0.292; ambiguity 0.417	Estimable, not supported
Stable global geometry	15 panels: 2 identifiable, 1 supported	Not identifiable
Structural competition	Hybrid best mean error; no decisive separation from local condition	Not identifiable
Exact cross-domain transfer	0 identifiable of 56 directional pairs; no exact deployed-system overlap	Not identifiable
Diversity scaling	Neither response-unit nor intervention-family criterion met	Estimable, not supported

K.1 KNOWN-TRUTH RECOVERY AND GLOBAL GEOMETRY

Ten structural truth families at weak, medium and strong signal produce 180 production-recovery rows, with six world replicates and ten topology templates. Strong-signal exact recovery is 0.2917, ambiguity 0.4167, and negative-world false-positive rate 0.1667 (Wilson upper 0.3189). The rule requires exact-recovery Wilson lower at least 0.50, ambiguity Wilson upper at most 0.35, negative false-positive upper at most 0.20, and strong point recovery of at least 0.50 in every structural world. Those conditions are not met. The earlier spectral-world classifier remains a secondary diagnostic and does not adjudicate this criterion.

The primary geometry requires at least five systems and eight complete units, adequate root bootstrap support, a subspace lower bound at least 0.60, original rank inside its interval, and positive matched gains on both held-cell and held intervention-family $\times$ held-system targets. Fifteen panels yield two fully identifiable tests and one supported panel. The tables expose the two predictive targets separately, preventing a positive held-cell result from concealing an unavailable held-intervention test.

K.2 COMPETITION, EXACT TRANSFER, AND WITHIN-DESIGN DIVERSITY

Seven prediction-distinguishable competitor classes share 279 outer fold blocks. Hybrid global-plus-modular structure has the best mean error and positive separation from five alternatives, but not from local condition. Complete pairwise intervals are retained below. The absence of a decisive winner is not silently rewritten as proof that local or global organization is absent.

All 56 registered directional source-domain pairs have zero identifying exact deployed-system bridges for the primary structural transfer target. Their full electronic table preserves the source and target definitions, as well as the non-identifiable outcomes. Cross-asset response-strength concordance is secondary and cannot supply structural transfer. Likewise, an ecological association between source size and geometry is not within-design diversity scaling.

The diversity experiment uses 12 repeated subsets at each of 25/50/75/100% along each axis and contains 601 refit rows. Subspace, held-cell prediction and rank-variability changes are reported with their own intervals. Both axes are estimable but not supported. The full criterion vector remains (ENS, NI, NI, NI, ENS), and the joint claim remains *not established*.

Table 40: Complete production structural recovery results by truth and signal. These topology-faithful worlds answer a different question from candidate-promotion calibration.

Known world	Signal	Simulations	Exact recovery	False positive	Ambiguity
global	medium	6	0	0	0.8333
global	strong	6	0	0	0.3333
global	weak	6	0.1667	0	0.6667

Continued from previous page

Known world	Signal	Simulations	Exact recovery	False positive	Ambiguity
global low rank k1	medium	6	0.1667	0	0.5
global low rank k1	strong	6	0.1667	0	0.6667
global low rank k1	weak	6	0.3333	0	0.5
global low rank k2	medium	6	0	0	0.6667
global low rank k2	strong	6	0.5	0	0.5
global low rank k2	weak	6	0	0	0.6667
global low rank k3	medium	6	0	0	0.5
global low rank k3	strong	6	0.3333	0	0.5
global low rank k3	weak	6	0	0	0.5
high dimensional	medium	6	0	0	0.6667
high dimensional	strong	6	0.5	0	0.3333
high dimensional	weak	6	0.1667	0	0.3333
hybrid	medium	6	0.1667	0	0.6667
hybrid	strong	6	0.3333	0	0.6667
hybrid	weak	6	0.1667	0	0.5
local	medium	6	1	0	0
local	strong	6	0.5	0	0.1667
local	weak	6	0.6667	0	0.3333
misspecified no structure	medium	6	0	0	1
misspecified no structure	strong	6	0.3333	0	0.6667
misspecified no structure	weak	6	0.3333	0.1667	0.5
modular	medium	6	0.3333	0	0.6667
modular	strong	6	0	0	0.1667
modular	weak	6	0.5	0	0.5
nuisance	medium	6	0	0.6667	0.3333
nuisance	strong	6	0	0.1667	0.8333
nuisance	weak	6	0	0	1

Table 41: All 15 within-asset response panels: available cells and complete/common-support geometry. Missingness-aware and column-mean sensitivities do not replace the primary population.

Source	Systems	Units	Density	Common systems	Common units	State
A01	50	84	0.05238	0	0	NI
A02	8	16	0.9922	8	15	NI
A03	6	387	0.7222	6	258	NI
A04	1	73	1	0	0	NI
A05	8	22	0.25	0	0	NI
A06	4	10	0.65	0	0	NI
C01	4	7	1	0	0	NI
C03	17	42	0.9986	17	41	NI
C05	4	85	1	0	0	NI
C11	11	2	1	0	0	NI
C13	6	21	1	6	21	S
C14	7	24	1	7	24	NI
C15	13	9	1	13	9	NI
C16	5	459	0.4039	0	0	NI
C17	5	438	1	5	438	ENS

Table 42: Root/task bootstrap geometry: effective rank is a descriptor, not a count of true capabilities. The full panel rule also requires both held-cell and held-intervention-block prediction.

Source	Effective rank	Lower	Upper	Subspace lower	State
A01	—	—	—	—	NI
A02	3.46	3.279	3.847	0.9612	NI
A03	4.475	4.4	4.761	0.6889	NI
A04	—	—	—	—	NI
A05	—	—	—	—	NI
A06	—	—	—	—	NI
C01	—	—	—	—	NI
C03	9.375	9.172	9.858	0.7032	NI
C05	—	—	—	—	NI
C11	—	—	—	—	NI
C13	2.983	2.224	3.547	0.6915	S
C14	2.751	2.528	3.176	0.6796	NI
C15	2.321	2.129	2.977	0.6886	NI
C16	—	—	—	—	NI
C17	3.182	3.204	3.648	0.884	ENS

Table 43: All seven structural competitors after equivalence checking. Lowest mean error alone does not establish a decisive winner.

Structural explanation	Mean MSE	Fold blocks	Win fraction
hybrid global modular	0.01764	279	0.5591
high dimensional rank	0.01877	279	0.05018
global low rank selected	0.01881	279	0.08961
local condition	0.01919	279	0.1613
modular intervention	0.01941	279	0.03226
additive system unit	0.02093	279	0.05018
null scalar	0.03809	279	0.05735

Table 44: Complete matched pairwise structural competition. Negative differences favor the left explanation. Hybrid does not decisively separate from local-condition structure.

Left explanation	Right explanation	Left minus right	Lower	Upper
null scalar	additive system unit	0.0171	0.007283	0.02956
null scalar	global low rank selected	0.01921	0.007218	0.03262
null scalar	local condition	0.01883	0.0079	0.03325
null scalar	modular intervention	0.01862	0.0078	0.03223
null scalar	hybrid global modular	0.02038	0.008723	0.03363
null scalar	high dimensional rank	0.01925	0.007787	0.03226
additive system unit	global low rank selected	0.002112	0.000906	0.003577
additive system unit	local condition	0.001732	-0.0003229	0.005084
additive system unit	modular intervention	0.001521	0.0004387	0.003333
additive system unit	hybrid global modular	0.003282	0.001499	0.005725
additive system unit	high dimensional rank	0.002152	0.0009161	0.003691
global low rank selected	local condition	-0.0003792	-0.002395	0.001863
global low rank selected	modular intervention	-0.0005906	-0.002525	0.001131
global low rank selected	hybrid global modular	0.001171	0.0002609	0.002588
global low rank selected	high dimensional rank	4.038e-05	1.235e-14	0.0001211

Continued from previous page

Left explanation	Right explanation	Left minus right	Lower	Upper
local condition	modular intervention	-0.0002114	-0.004362	0.003389
local condition	hybrid global modular	0.00155	-0.001205	0.004872
local condition	high dimensional rank	0.0004196	-0.002071	0.002402
modular intervention	hybrid global modular	0.001761	0.0005934	0.003227
modular intervention	high dimensional rank	0.000631	-0.001017	0.002595
hybrid global modular	high dimensional rank	-0.00113	-0.00266	-0.0002112

Table 45: Within-design diversity scaling along both required axes. Full intervals and repeated subset diagnostics accompany the electronic result tables.

Diversity axis	Panels	Subspace change	Prediction change	Rank-SD reduction	State
response units	8	0.1495	5.591e-05	-0.006121	ENS
intervention families	4	0.1684	-0.0003737	-4.353e-15	ENS

Table 46: Held-cell predictive geometry. Positive gain favors selected masked low rank over additive prediction. Both predictive targets are required for geometry support.

Source	Gain	Lower	Upper	State
A01	0.005962	0.004832	0.007256	S
A02	0.002026	0.001571	0.002503	S
A03	0.009724	0.009264	0.0102	S
A04	—	—	—	NI
A05	0.0006514	0.0004493	0.0008399	S
A06	0.0002939	0.0001868	0.0003989	S
C01	0.000429	-4.263e-05	0.0009338	ENS
C03	0.002974	0.002438	0.003489	S
C05	-0.0001839	-0.0002338	-0.0001395	ENS
C11	0.000934	0.0007262	0.001175	S
C13	0.00101	0.0007799	0.001257	S
C14	0.004335	0.003402	0.005204	S
C15	-0.0002388	-0.0005467	-4.975e-05	ENS
C16	0.0006598	0.0005409	0.0007667	S
C17	0.0009856	0.0009218	0.001054	S

Table 47: Held-intervention block predictive geometry. Positive gain favors selected masked low rank over additive prediction. Both predictive targets are required for geometry support.

Source	Gain	Lower	Upper	State
A01	—	—	—	NI
A02	—	—	—	NI
A03	—	—	—	NI
A04	—	—	—	NI
A05	—	—	—	NI
A06	—	—	—	NI
C01	—	—	—	NI

Continued from previous page

Source	Gain	Lower	Upper	State
C03	—	—	—	NI
C05	—	—	—	NI
C11	—	—	—	NI
C13	0.0009272	0.0002166	0.001677	S
C14	—	—	—	NI
C15	—	—	—	NI
C16	0.0005247	—	—	NI
C17	0.00143	0.001074	0.001838	S

Table 48: Within-design diversity: subspace change and its frozen interval. Positive rank-SD reduction means greater rank stability.

Diversity axis	Change	Lower	Upper
response units	0.1495	0.07958	0.2226
intervention families	0.1684	0.07391	0.3349

Table 49: Within-design diversity: predictive change and its frozen interval. Positive rank-SD reduction means greater rank stability.

Diversity axis	Change	Lower	Upper
response units	5.591e-05	-0.000668	0.0007193
intervention families	-0.0003737	-0.0008745	0.000127

Table 50: Within-design diversity: rank change and its frozen interval. Positive rank-SD reduction means greater rank stability.

Diversity axis	Change	Lower	Upper
response units	-0.006121	-0.01836	-5.836e-16
intervention families	-4.353e-15	-1.306e-14	0

L THE SCIENTIFIC FRONTIER AND USE OF THE ACCOMPANYING CATALOGS

The complete evidence distinguishes three claims that should remain separate in future reuse: Conditional Response is a supported empirical substrate; semantic prediction and local candidate structure are informative; full connected capability validity and fundamental organization require further evidence. A new analysis of the same disconnected sources cannot manufacture exact identity or a missing external outcome. A newly connected design can change those identifying conditions, while additional within-source data can improve local precision and structural discrimination.

The electronic catalogs preserve long observational and response descriptions at their original precision. They support detailed inspection of individual laws, root retention, response heterogeneity and source-specific examples beyond the readable summaries in this PDF. Their interpretation remains bounded by the definitions in this manuscript, particularly the distinction between a descriptive row, an estimable nonsupport result, a missing identifying connection, and a supported cross-level inference.