PROVER: A CLAIM-CENTRIC FRAMEWORK FOR AUDITING REASONING TRACES IN CELLULAR PERTURBATION ANALYSIS

Anonymous authors
Paper under double-blind review

ABSTRACT

In the context of AI for Virtual Cells (AIVC), cellular perturbation models increasingly generate natural-language analyses alongside predictions of cellular responses. While these analyses promise biologically grounded interpretability, their validity is rarely examined beyond final prediction accuracy. We introduce **PROVER** (**P**erturbation **R**easoning **O**utput **VER**ifier), a claim-centric hierarchical framework for auditing reasoning traces in cellular perturbation analysis. PROVER reconstructs each trace as a structured argument graph centered on Claims and evaluates it along two dimensions: *Cogency* and *Completeness*. Cogency assesses whether premises are supportable, evidence is relevant to target Claims, and inferential connections provide sufficient support for their conclusions, while Completeness measures coverage of task-relevant analytical perspectives. We apply PROVER to a 600-case benchmark spanning eight perturbation-analysis systems, examine three LLM audit backbones and holistic-judge baselines, and complement these analyses with expert feedback. The audits flag unsupported assumptions, context-misaligned evidence, and insufficiently supported inferences that answer-level metrics do not assess. Our results show that evaluating biological reasoning requires assessing not only prediction outcomes but also the evidence and reasoning used to justify them. We release the framework, auditing protocol, benchmark metadata, and audit artifacts.

1 INTRODUCTION

Predicting how a cell responds to genetic perturbation is a cornerstone of the virtual-cell agenda: models that anticipate the consequences of a knockdown promise to replace thousands of wet-lab experiments with computation (Bunne et al., 2024). Deep discriminative models (Roohani et al., 2024; Lotfollahi et al., 2019; Hao et al., 2024; Hetzel et al., 2022) have made striking progress on this task, and their black-box predictions hold legitimate value for large-scale screening and prioritization. Yet a prediction is not an insight. As the authors of a recent language-based system themselves observe, "most existing methods are black-box, revealing little about the learned biology without post-hoc probing" (Wu et al., 2025). These models tell us *what* will happen, not *why*: their rationale remains implicit in learned representations, offering no evidence chain that a scientist can interrogate, learn from, or follow up experimentally.

Large language models, prompted or fine-tuned to reason step by step, appear to close this gap. If language is "a natural medium for traversing the structured biological knowledge relevant to perturbation experiments" (Wu et al., 2025), then a reasoning trace is precisely the missing artifact: an inspectable, falsifiable hypothesis. Systems such as Wu et al. (2025); Istrate et al. (2025); Phillips et al. (2025); Wang et al. (2026) now produce putative insights at scale—chains that cite analogous perturbations, invoke pathways, and commit to directional claims. For the first time, the reasoning behind perturbation prediction is available for scrutiny.

Scrutiny, however, reveals that fluent reasoning is not necessarily faithful reasoning. Our preliminary analysis of public traces surfaces two recurring patterns. The first is *decoupling*: models trained on perturbation data often reach correct answers while their verbalized rationale is disconnected from whatever evidence drove the prediction—the text reads causally, but functions as post-hoc narra-

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.

tion. The second is *unverified confidence*: knowledge-guided chains are riddled with fabrication and over-generalization—expanding a gene symbol into a nonexistent protein name, relocating a protein across compartments to force an analogy, or elevating a single co-expression observation into a class-level regulatory law. We argue that an unverified rationale is worse than none, for it manufactures false insight under the appearance of understanding. Yet these systems are still measured by the outcome yardstick inherited from the prediction era—answer accuracy—and general-purpose reasoning metrics (Golovneva et al., 2023) capture surface coherence but encode no domain knowledge against which content can be verified.To our knowledge, no existing framework systematically evaluates the content validity of perturbation-reasoning traces themselves.

We present **PROVER** (**P**erturbation **R**easoning **O**utput **VER**ifier). The approach emphasizes that the object of audit is the model-presented reasoning output, rather than the model's hidden computation, the final prediction alone, or a complete proof of biological truth. A central design principle of PROVER is that reasoning decomposition and evaluation decomposition are aligned. Reasoning traces are decomposed into claims, supporting evidence, and inferential links, and evaluations are attached directly to these same units. Rather than treating a trace as a monolithic piece of text, PROVER evaluates the reliability of evidence, the sufficiency of support, and the validity of inference at the locations where they occur. By preserving these local relationships, PROVER can explain not only whether a reasoning trace succeeds or fails, but where and why. The same structured representation also provides a basis for studying reasoning patterns themselves, enabling the discovery of recurring reasoning paradigms and their characteristic failure modes.

The overall contributions are summarized as follows.

(1) We introduce a hierarchical evaluation framework that jointly decomposes reasoning traces and evaluation criteria, establishing explicit correspondences between them to enable structured claim-level assessment. This design allows fine-grained auditing of complex reasoning processes beyond conventional answer-level evaluation.

(2) We conduct a systematic benchmark and audit of eight perturbation-analysis systems. Our study reveals a range of reasoning failures that remain largely invisible under answer-level metrics, highlighting the limitations of outcome-based evaluation for scientific reasoning tasks.

(3) Beyond evaluation, we demonstrate that the proposed decomposition serves as a lens for analyzing reasoning structure itself. Using this framework, we identify recurring reasoning paradigms in perturbation analysis, characterize their associated failure modes, and derive broader insights for AI for Virtual Cells and evidence-based biomedical reasoning.

2 Related Work

Perturbation-response prediction: outcomes without explicit reasoning. Predicting cellular responses to genetic perturbations has become a central task in AI for Virtual Cells. Deep discriminative models, including graph-based architectures, variational models, and recent single-cell foundation models, have achieved increasingly strong performance on perturbation-response prediction (Roohani et al., 2024; Lotfollahi et al., 2019; Hao et al., 2024; Cui et al., 2024; Hetzel et al., 2022). These systems are valuable for large-scale screening and prioritization because they predict what cellular state is likely to emerge after perturbation. However, the knowledge captured by these models remains largely implicit. Their internal representations may encode rich perturbation structure, but the resulting predictions are not accompanied by an inspectable chain of evidence that explains why a particular outcome is expected. Consequently, evaluation in this line of work is necessarily outcome-based, focusing on prediction accuracy and distributional fidelity rather than the validity of reasoning. Although these models learn perturbation-response regularities implicitly, they do not expose those regularities as scientific arguments that can be examined, challenged, or reused.

Language-based perturbation reasoning: explanations become visible. Recent work has begun to augment perturbation analysis with large language models, making biological reasoning an explicit artifact. Systems such as Wu et al. (2025) retrieve structured biological knowledge and generate explanatory rationales; Istrate et al. (2025) incorporates reasoning into perturbation prediction pipelines; Phillips et al. (2025) generates perturbation reasoning traces for model training; and Wang et al. (2026) explores agentic biological reasoning frameworks. These systems represent an

important shift. For the first time, perturbation predictions are accompanied by textual analyses that reference biological entities, pathways, prior observations, and analogous perturbations. Yet evaluation remains largely inherited from the prediction era: systems are primarily judged by answer-level correctness, while the generated reasoning is treated as supporting narrative rather than a first-class evaluation target. As a result, the field lacks a systematic methodology for assessing whether a reasoning trace is biologically supported, whether its evidence is sufficient, or whether its inferential steps are justified. More fundamentally, existing work treats reasoning traces primarily as objects to be generated. We instead argue that they should also be treated as objects of analysis. Understanding recurring reasoning patterns and their failure modes is essential if language-based systems are to contribute scientific insight rather than merely produce plausible explanations.

Process-level evaluation of reasoning. A growing body of work evaluates reasoning processes rather than final answers alone. Existing approaches can be roughly grouped into four categories: executable verification in domains such as mathematics and programming (Lightman et al., 2024; Zheng et al., 2025; Song et al., 2025; Ling et al., 2023; Lyu et al., 2023), coherence- and logic-based evaluation (Golovneva et al., 2023), faithfulness analyses that compare rationales against model decision-making processes (Turpin et al., 2023; Lanham et al., 2023), and expert-driven assessment in scientific and medical settings (Zhang et al., 2026; Ostmeier et al., 2024; Yu et al., 2023). These directions establish that reasoning should be evaluated beyond answer correctness, but they rely on executability, surface reasoning quality, faithfulness, or expert judgment as their primary source of supervision. Perturbation biology poses a different challenge: reasoning traces make explicit claims about genes, proteins, pathways, perturbations, and their relationships, many of which can be checked against external biological knowledge. Our work therefore focuses on *content validity*: whether the evidence cited by a reasoning trace is reliable, whether that evidence is sufficient to support the claim being made, and whether the inferential connection between them is justified.

Claim verification and reasoning analysis. Claim-level verification has become a mature research area, particularly in factuality evaluation. Existing methods verify generated statements against reference corpora or knowledge sources, including work on scientific claim verification, factual precision, retrieval-grounded verification, and citation-supported generation (Wadden et al., 2020; Min et al., 2023; Hu et al., 2024; Gao et al., 2023). These approaches establish that generated content can be decomposed into verifiable units and evaluated systematically. Our work builds on this intuition but extends it in two directions. First, perturbation reasoning requires evaluating more than factual claims alone. The reliability of evidence, the sufficiency of support, and the validity of inferential links all contribute to whether a biological conclusion is credible. Second, we view decomposition not only as a mechanism for evaluation but also as a mechanism for analysis. By representing reasoning traces as structured collections of claims, evidence, and inferential connections, we can localize errors, characterize recurring reasoning paradigms, and study how different forms of biological reasoning succeed or fail across systems. This perspective shifts the objective from merely grading reasoning outputs to understanding the structure of reasoning itself.

3 PROVER DESIGN

PROVER aligns a claim-centered argument graph with evaluation criteria attached to its constituent units. Figure 1 connects this representation, the evaluation hierarchy, and the execution workflow.

3.1 AUDIT OBJECTIVE AND DESIGN PRINCIPLES

Let q denote the task and its cellular context, x the model-presented analysis, and D a frozen dossier of observations and domain references. PROVER assesses whether the account in x adequately supports its claims:

$$\mathcal{A}(x; q, D) \longrightarrow (G,\ J,\ H), \tag{1}$$

where G is an argument graph, J contains local judgments, and H links them to source spans, reference records, or structural checks. The audit concerns the presented account, not inaccessible internal computation: it does not establish hidden-reasoning faithfulness, certify prediction correctness, or reconstruct the cell's complete causal state.

Four principles organize the framework.

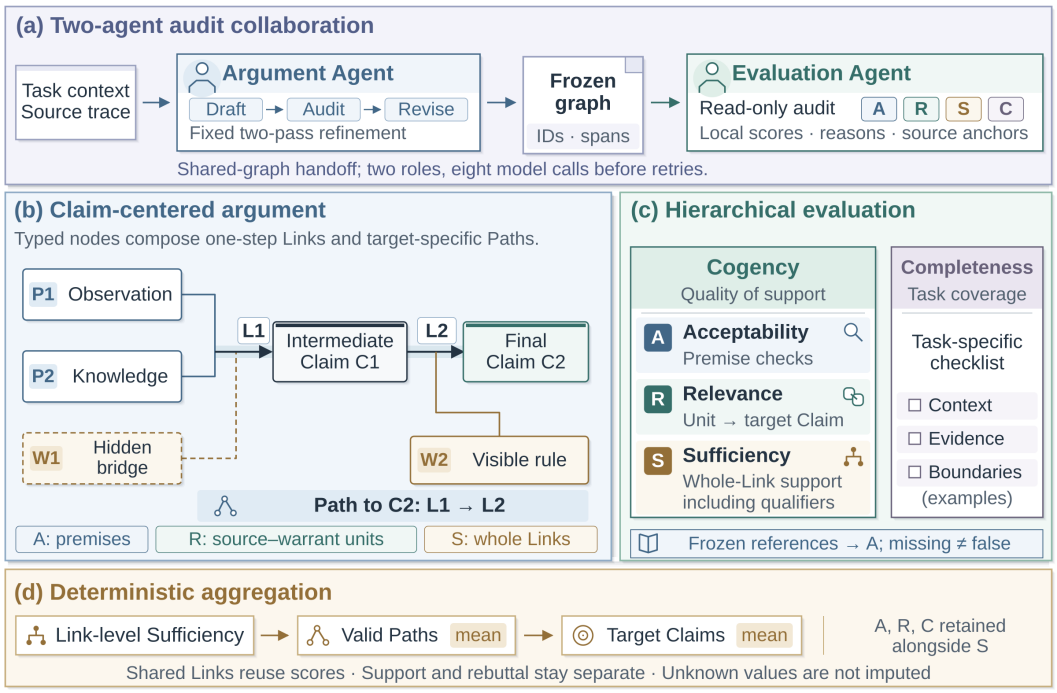

Figure 1: PROVER design. The Argument Agent hands a frozen graph to the Evaluation Agent for a read-only audit. Cogency contains A/R/S; Completeness is its sibling. Only Link-level Sufficiency propagates through Paths to target Claims. The graph is schematic, not a benchmark example.

1. **Joint decomposition.** Evaluation criteria correspond to explicit argument objects, preventing an aggregate score from concealing which premise or inferential move caused a failure.
2. **Claim-centered traceability.** Claims are the targets of reasoning; premises and warrants explain how they are reached, and local connections expose their dependencies.
3. **Support quality versus coverage.** *Cogency* assesses the offered support; *Completeness* assesses task-required perspectives. A broad but unsupported account differs from a sound but narrow one.
4. **Auditable uncertainty.** Missing reference records remain unknown rather than false. Qualifiers and provenance are preserved, and deterministic aggregation is separated from model judgment.

3.2 Claim-Centered Decomposition

Following the argument components of Toulmin (1958), we represent a trace as $G = (V, L)$ with three parent node types: **Premise** (experimental result or background knowledge), **Warrant** (visible or hidden), and **Claim** (intermediate or final). Evidence is a premise subtype, not a separate edge type. A normalized statement makes each node readable, while exact spans preserve its source. Repeated claims retain their occurrences and qualifiers without becoming new claims solely because wording changes. A hidden warrant records an unstated assumption required by an inference; it receives no invented quotation and is not treated as verified evidence. This source-aware decomposition builds on fine-grained claim verification (Min et al., 2023; Hu et al., 2024) while additionally representing inferential dependencies.

A **Link** is one complete inferential connection:

$$\ell = (F_\ell, W_\ell, t_\ell, \rho_\ell), \tag{2}$$

where F_ℓ contains premise or earlier-claim sources, W_ℓ their associated warrants, t_ℓ a target Claim, and $\rho_\ell \in \{\text{support}, \text{rebuttal}\}$. Its target is always a Claim, not a warrant. This one-hop edge can join several jointly necessary sources; it denotes an argumentative dependency, not necessarily a direct biological causal relation. Within a Link, a **Relation unit** groups one or more jointly necessary

Table 1: Correspondence between argument objects and evaluation criteria.

Graph unit	Judgment	Output
Premise	Acceptability	rubric state and score; unknown retained
Relation unit	Relevance	source–warrant support (0–10)
Complete Link	Sufficiency	joint support (0–10), qualifier audit
Valid Path	derived Sufficiency	mean of constituent Link scores
Target Claim	derived Sufficiency	separate support/rebuttal summaries
Task checklist	Completeness	present / partial / absent / inapplicable

sources with their warrants and target, allowing local contributions to be assessed without separating evidence from the rule that makes it relevant.

A **Path** is a target-conditioned argument subgraph whose Links reach a specified Claim. Paths may branch, merge, or share prefixes; their order follows actual Claim dependencies rather than sentence or identifier order. A single Link is a valid Path, but disconnected Links do not become a Path by being listed together. Support and rebuttal are Path types, not new node types. Each Path retains both its formal dependencies and a human-readable summary that introduces no additional facts. Distinct Paths reaching the same Claim form its support or rebuttal Path family.

3.3 Cogency and Completeness

Cogency is operationalized through three complementary judgments:

- **Acceptability (A):** whether each premise is supported by the supplied observations or references. A trace's assertion that an experiment occurred is not itself independent confirmation.
- **Relevance (R):** how strongly a Relation unit supports its target. Evidence and warrant are assessed jointly where inseparable; this is neither a standalone truth score for a warrant nor a second complete-Link score.
- **Sufficiency (S):** whether a complete Link's sources and warrants jointly justify its target at the degree of certainty expressed by its qualifier. Path and Claim summaries are derived from these local judgments.

Completeness (C) measures coverage of a frozen, task-specific checklist, such as cellular context, signed evidence for a directional claim, or the scope of a generalization. It does not require a complete biological network, and external references cannot fill omissions in the trace itself.

3.4 Two-Stage Audit Workflow

An Argument Reconstruction Agent recovers the argument; an Evaluation Agent judges the frozen graph. These two logical roles schedule eight calls: nodes, Paths, audit, revise, and four dimension judgments, before retries (Appendix B).

Reconstruction. The first pass extracts typed nodes, Links, and Paths from the task and source response. The second checks the draft against the full trace and revises it. Revision may recover hidden warrants, separate compound claims, or correct dependencies, but cannot invent observations, strengthen qualifiers, or improve the source argument. Both passes have fixed budgets within each evaluator configuration. The graph is then frozen so evaluation cannot rewrite its target.

Evaluation. Four dimension-specific calls return local scores or states, reasons, and source anchors without modifying the graph. Frozen domain references can include HGNC, UniProt, Gene Ontology, and CORUM (Seal et al., 2023; The UniProt Consortium, 2023; Gene Ontology Consortium, 2021; Tsitsiridis et al., 2023). Identity or function records do not establish context-specific effects, and failed lookups are not refutations. Backend-specific source and dossier exposure is documented in Appendix B.

Quality control. Runs retain inputs, raw outputs, configuration versions, and repair provenance. Unresolved failures remain missing. Structural checks do not certify judgment correctness, and

extraction errors are not source-system reasoning failures. Source anchors and local reasons make judgments inspectable. Evaluator configurations remain fixed across source systems; alternative backbones produce separate runs, not pooled votes or an additional arbitration stage. Retry and recovery rules are given in Appendix B.

3.5 Hierarchical Aggregation

Models return local judgments; aggregate scores follow explicit rules. The runtime validates types, identifiers, source spans, Relation coverage, and Claim dependencies before aggregation. Path and Claim summaries derive from Link scores without an additional Link-to-Link continuity score. For a valid Path p with distinct Link set L_p, and distinct valid Paths $\mathcal{P}_\rho(c)$ of type ρ reaching Claim c:

$$S_{\text{path}}(p) = \frac{1}{|L_p|} \sum_{\ell \in L_p} S_{\text{link}}(\ell), \tag{3}$$

$$S_\rho(c) = \frac{1}{|\mathcal{P}_\rho(c)|} \sum_{p \in \mathcal{P}_\rho(c)} S_{\text{path}}(p). \tag{4}$$

Shared Links reuse the same local score, with their dependencies retained; overlapping Paths are not assumed to be independent evidence. Support and rebuttal remain separate rather than being subtracted. Missing assessments or invalid Paths are flagged, not assigned ordinary zero scores. These arithmetic summaries measure average assessed support, not proof probabilities or weakest-link guarantees.

The current rubric uses 0–3 for A and C: Acceptability ranges from directly contradicted (0) through weak (1) and partial support (2) to supported (3); Completeness ranges from absent (0) through partial coverage (1 or 2) to present (3). Unknown and inapplicable remain explicit. Scores are normalized by 3 for A, C and by 10 for R, S. Dimension summaries average applicable local judgments, except that $\widehat{S}$ averages derived final-Claim support. A provisional scalar is

$$\text{Cogency} = \widehat{A}\,\widehat{R}\,\widehat{S}, \qquad \text{Overall} = \widehat{C}\,\text{Cogency}. \tag{5}$$

Unknown inputs yield bounds rather than an imputed midpoint; a point score requires determinate inputs. The judgment vector and its provenance remain primary. Neither the scalar nor its bounds are calibrated probabilities, and the bounds are not statistical confidence intervals.

4 Evaluation

Following our three contributions, we examine audit sensitivity and expert feedback (RQ1), metric profiles and local concerns across systems (RQ2), and the roles of generated text revealed by claim-centered decomposition (RQ3).

4.1 Benchmark and Audit Protocol

Benchmark. The corpus contains 600 analyses from eight systems (Table 2), retaining tasks, responses, and provenance. Task heterogeneity precludes a shared predictive leaderboard. Correctness-conditioned analyses use 382 records with normalized correctness flags. Outcome-balanced sampling does not represent deployment frequencies; quota and provenance details are in Appendix A.

Audit protocol. We run Qwen3.6-35B-A3B (Qwen Team, 2026) and Llama-3.3-70B-Instruct (Meta, 2024) locally, and DeepSeek-v4-pro (DeepSeek-AI, 2026) through its official API. All three reconstruct and evaluate the same 600 traces. Qwen supplies the primary profiles; Llama and DeepSeek provide sensitivity runs. Structural repairs preserve source propositions and refresh affected judgments. Prompt and dossier differences are recorded in Appendix B.

Analysis. We report determinate-score denominators and unknown-state coverage. SUMMER weights its three task means equally; other systems weight traces equally. Trace-bootstrap intervals (2,000 draws, seed 42) accompany the Appendix results.

Table 2: Audited systems and task coverage (600 cases). DE: differential expression; Dir.: direction; GSE: the release's gene-set description/naming task.

System	Cases	Audited task or component
rbio-1 (Istrate et al., 2025)	60	DE
SUMMER (Wu et al., 2025)	180	DE (71), Dir. (71), GSE (38)
AROMA (Wang et al., 2026)	60	Dir.
SynthPert (Phillips et al., 2025)	60	Dir.
PerturbAgent (Pei et al., 2026)	60	Directional probabilities
CellReasoner (Cao et al., 2025)	60	Cell-type annotation (extension)
VCWorld (Wei et al., 2026)	60	DE (26), Dir. (34)
LangPert (Märtens et al., 2025)	60	Candidate selection for effect transfer
Total	600	

4.2 Audit Sensitivity and Expert Feedback

All three pipelines pass structural checks on 600 records. On matched traces, DeepSeek gives higher R than Qwen (7.55 vs. 4.70), but lower C (6.52 vs. 9.43); S means ($n = 532$) are 3.95/6.61/5.55 for Qwen/Llama/DeepSeek. These dimension-dependent shifts motivate separate profiles. Graphs, prompts, and reference exposure differ, so this is end-to-end sensitivity, not a controlled model-only comparison (Appendix F.3).

Holistic baseline and outcome separation. A single-call judge using each backbone scores all 600 traces (Table 3). Its correlations with decomposed dimensions are weak ($|\rho| \leq 0.33$), and 59–93 cases per backbone receive opposing holistic and S assessments. On the 382 outcome-labeled traces, decomposed scores also show little positive correctness separation (AUC 0.438–0.524). These results show that the scores are not interchangeable, not that either audit is more valid or that reasoning quality is independent of correctness. Input and budget differences prevent a pure decomposition ablation (Appendix G).

Table 3: Holistic baseline (h, 0–10; 600 traces per backbone). Opposing: $h \geq 7, S \leq 3$ or vice versa, among determinate S. AUC measures answer-correctness separation on 382 traces, not audit validity.

Backbone	h mean$\pm$SD	$\rho(h, S)$	Opposing/n_S	AUC(h)
Qwen	2.86 ± 2.25	0.06	80/579	0.502
Llama	7.51 ± 1.76	0.32	59/555	0.531
DeepSeek	2.55 ± 1.72	0.20	93/583	0.479

Expert feedback. Initial expert reviews flag overclaim alongside concerns about evidence relevance, inference, and coverage. They provide complementary trace-level judgments, not gold graphs or confirmation of automatic diagnoses. The returned sample, exact-text matching, and exploratory score associations are detailed in Appendix J.

4.3 Metric-Level Audit Findings

RQ2 connects the primary Qwen profiles (Table 4) to source-linked concerns. The four criteria follow Section 3.3; these task-dependent profiles are not universal rankings. Cases in Appendix H ground the observations below; means alone do not establish failure-mode frequencies.

Completeness: coverage is not support. High C means (8.29–10.00) record checklist coverage, not support. SUMMER-0 answers all five questions but leaves its cross-compartment transfer unjustified (Appendix H.1). GSE additionally requires caution: the frozen checklist is broader than the released naming task (Appendix D.4).

Table 4: Primary metric profile by system under the fixed Qwen3.6-35B-A3B evaluator (0–10). N is the full sample count; n_A and n_S count traces with determinate Acceptability and Sufficiency scores. Relevance and Completeness cover all N traces. SUMMER macro-averages its three tasks.

System	N	Acceptability		Relevance	Sufficiency		Completeness
		score	n_A	score	score	n_S	score
rbio	60	8.90	21	4.98	3.84	59	9.69
SUMMER	180	9.20	128	4.25	3.59	179	9.28
AROMA	60	9.31	35	4.71	3.48	60	10.00
SynthPert	60	9.32	55	5.71	5.27	47	8.29
PerturbAgent	60	10.00	55	2.90	2.79	60	9.83
CellReasoner	60	9.76	42	7.09	6.00	57	9.87
VCWorld	60	9.44	9	4.40	3.92	59	9.64
LangPert	60	9.75	58	5.18	4.14	58	8.87

Acceptability: premise quality is conditional. A is high where determinate, but known/applicable premise coverage ranges from 11.8% for VCWorld to 86.3% for PerturbAgent. These means describe checkable subsets. PerturbAgent-0's $A = 10$ verifies retrieval provenance, not a biological null effect (Appendix H.3). Unresolved premises remain unknown, not false; both A and S retain missing-score denominators.

Relevance: related evidence need not support the target. R ranges from 2.90 (PerturbAgent) to 7.09 (CellReasoner). SUMMER-0 transfers effects across mitochondrial and cytosolic systems through the category "ribosome" (Appendix H.1); a shared description does not establish effect transfer. SUMMER's DE/direction/GSE means are 3.82/3.41/5.53, motivating task-specific inspection rather than a universal failure explanation.

Sufficiency: conclusions can exceed their support. S means range from 2.79 to 6.00. PerturbAgent-0 uses failed retrieval and interaction evidence to infer no effect; both supporting Links receive $S = 3$ (Appendix H.3). The gap is the move from not finding a response to asserting its absence. Qualifier assessment similarly asks whether certainty exceeds support, not whether a hedging word is present.

4.4 Reasoning Structures and Domain Insights

RQ3 focuses on reasoning about *cellular perturbation effects*, the core task family in our eight-system audit. Decomposition reveals two broad forms in the inspected arguments: *Mechanism reasoning* connects biological knowledge through intermediate molecular Claims; *Evidence-based reasoning* joins observed responses to mechanistic or functional warrants (Figure 2). These characterize visible Paths, not mutually exclusive system types or hidden computation.

The empirical-trace gap. Numerical predictors can learn high-dimensional response regularities without expressing them as biological propositions (Lotfollahi et al., 2019). Optimizing predictive fit need not produce a textual account of the relationships learned from data. AROMA's low sensitivity to the tested explanation-context changes and rbio's trace–answer mismatch illustrate why successful prediction does not establish such an account; neither identifies hidden computation (Appendices H, I). The missing explanatory capability is to select the most informative response relationships for a target, explain their joint implications, and delimit their applicability. This is more demanding than citing an isolated association or attaching a mechanistic story to a prediction.

Mechanisms should generate testable hypotheses. SynthPert's damage–compensation chains and VCWorld's proposed signaling cascades introduce protein activities and interactions that expression measurements alone do not establish. Their value is to propose *how* an effect could arise, but plausibility does not determine the observed direction: SynthPert-0 invokes stress-induced HSPD1 compensation after PIK3C3 knockdown yet predicts the wrong direction. Useful traces should specify testable intermediate states, competing amplification and compensation, and context-specific Links. Positive or negative feedback requires an explicit return relation, not merely an increasing or decreasing chain. The goal is an experimentally discriminable explanation, not an unqualified causal narrative.

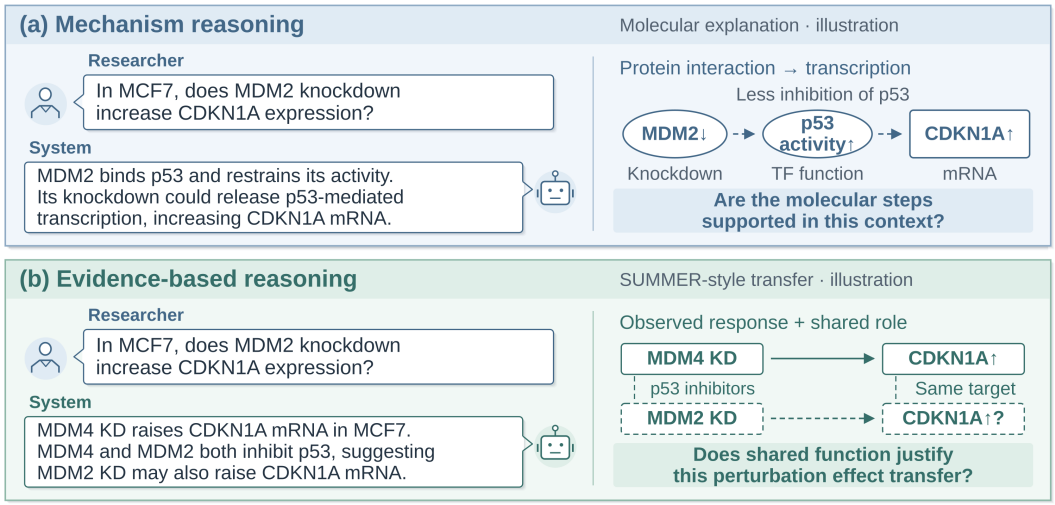

Figure 2: Two reasoning forms for the same MDM2–CDKN1A query in MCF7 cells (wild-type p53). These author-constructed dialogues illustrate (a) molecular explanation and (b) SUMMER-style transfer, not measured system outputs. Biological premises follow Heminger et al. (2009); dashed arrows indicate proposed support. KD: knockdown; TF: transcription factor.

Evidence-based traces need a claim-specific warrant. SUMMER, LangPert, and PerturbAgent bridge observations to target decisions, but their warrants can be too coarse. SUMMER-0 crosses compartments through "ribosome," while PerturbAgent-0 infers no effect from unsuccessful retrieval. Auditable traces must identify the transferred observation, licensing relation, cellular context, counterevidence, and qualified conclusion. The question is not whether a bridge is explicit, but whether it warrants this transfer.

From prediction to experimental questions in AI for Virtual Cells. CellReasoner marker convergence and SUMMER GSE set abstraction extend this perspective beyond directional prediction. For AIVC, our contribution is an inspectable interface between predicted responses and proposed explanations: it distinguishes a missing observation, an unjustified transfer, and an unsupported mechanism. These distinctions suggest different follow-ups—obtain context-matched evidence, test the transfer assumption, or measure a proposed intermediate state—rather than treating every failure as a prediction error. They motivate, but do not yet demonstrate, improved experimental decisions. An LLM can act as an explanation analyst: select relevant response relations, connect them to knowledge, distinguish competing damage and compensation accounts, and state what would challenge each one. Text complements numerical and graphical representations; it need not reproduce the full cellular network or be the causal basis of prediction. Observations can motivate a mechanism that experiments then test, as in the MDM4–CDKN1A/p53 example (Heminger et al., 2009). PROVER makes that proposed connection auditable, not biologically proven.

5 Conclusion

PROVER aligns argument decomposition with evaluation criteria to audit cellular analyses beyond prediction accuracy. Across eight systems, it distinguishes evidence gaps, questionable transfer assumptions, and unsupported inferences. Holistic scores do not reproduce these profiles; preliminary expert feedback complements them without establishing diagnostic validity.

For cellular perturbation research and AIVC, PROVER separates response prediction from mechanistic justification and organizes candidate explanations into testable experimental questions. Language should expose consequential assumptions and competing mechanisms, not replace evidence. Scores remain evaluator-dependent and uncalibrated; reference gaps are not refutations. Graph fidelity, diagnostic accuracy, and downstream utility require broader independent validation.

AI USE STATEMENT

Generative AI tools assisted discussions of research ideas and refinement of the evaluation framework, implementation and execution of local and API-based experiments, benchmark selection and preparation in consultation with the authors, and interpretation of audit outputs. They also assisted manuscript drafting and revision, literature and reference checks, and figure preparation. The authors directed these activities, made the final methodological and editorial decisions, and take responsibility for the submitted text, claims, and artifacts. The LLMs used within PROVER's audit pipeline are documented in Section 4.1 and Appendix B.

REFERENCES

Charlotte Bunne, Yusuf Roohani, Yanay Rosen, Ankit Gupta, Xikun Zhang, Marcel Roed, Theo Alexandrov, Mohammed AlQuraishi, Patricia Brennan, Daniel B. Burkhardt, et al. How to build the virtual cell with artificial intelligence: Priorities and opportunities. *Cell*, 187(25):7045–7063, 2024. doi: 10.1016/j.cell.2024.11.015.

Guangshuo Cao, Yi Shen, Jianghong Wu, Haoyu Chao, Ming Chen, and Dijun Chen. Cellreasoner: A reasoning-enhanced large language model for cell type annotation. *bioRxiv*, 2025. doi: 10.1101/2025.05.20.655112. URL https://www.biorxiv.org/content/10.1101/2025.05.20.655112v1.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

DeepSeek-AI. DeepSeek-V4: Towards highly efficient million-token context intelligence. *arXiv preprint arXiv:2606.19348*, 2026. doi: 10.48550/arXiv.2606.19348. URL https://arxiv.org/abs/2606.19348.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6465–6488, 2023. doi: 10.18653/v1/2023.emnlp-main.398. URL https://aclanthology.org/2023.emnlp-main.398/.

Gene Ontology Consortium. The gene ontology resource: enriching a GOld mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021. doi: 10.1093/nar/gkaa1113.

Olga Golovneva, Moya Chen, Spencer Poff, Martin Corredor, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. ROSCOE: A suite of metrics for scoring step-by-step reasoning. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.07919.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, 21(8):1481–1491, 2024. doi: 10.1038/s41592-024-02305-7.

Katherine Heminger, Michael Markey, Meldrick Mpagi, and Steven J. Berberich. Alterations in gene expression and sensitivity to genotoxic stress following HdmX or Hdm2 knockdown in human tumor cells harboring wild-type p53. *Aging*, 1(1):89–108, 2009. doi: 10.18632/aging.100008. URL https://www.aging-us.com/article/100008/text.

Leon Hetzel, Simon Böhm, Niki Kilbertus, Stephan Günnemann, Mohammad Lotfollahi, and Fabian J. Theis. Predicting cellular responses to novel drug perturbations at a single-cell resolution. In *Advances in Neural Information Processing Systems*, volume 35, pp. 26711–26722, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/aa933b5abc1be30baece1d230ec575a7-Abstract-Conference.html.

Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 6953–6975, 2024. doi: 10.18653/v1/2024.emnlp-main.395. URL https://aclanthology.org/2024.emnlp-main.395/.

Ana-Maria Istrate, Fausto Milletari, Fabrizio Castrotorres, Jakub M. Tomczak, Michaela Torkar, Donghui Li, and Theofanis Karaletsos. rbio1: Training scientific reasoning LLMs with biological world models as soft verifiers. *bioRxiv*, 2025. doi: 10.1101/2025.08.18.670981. URL https://www.biorxiv.org/content/10.1101/2025.08.18.670981v1.full.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*, 2023. doi: 10.48550/arXiv.2307.13702. URL https://arxiv.org/abs/2307.13702.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=v8L0pN6EOi.

Zhan Ling, Yunhao Fang, Xuanlin Li, Zhiao Huang, Mingu Lee, Roland Memisevic, and Hao Su. Deductive verification of chain-of-thought reasoning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://arxiv.org/abs/2306.03872.

Mohammad Lotfollahi, F. Alexander Wolf, and Fabian J. Theis. scgen predicts single-cell perturbation responses. *Nature Methods*, 16:715–721, 2019. doi: 10.1038/s41592-019-0494-8.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 305–329, 2023. doi: 10.18653/v1/2023.ijcnlp-main.20. URL https://aclanthology.org/2023.ijcnlp-main.20/.

Kaspar Märtens, Marc Boubnovski Martell, Cesar A. Prada-Medina, and Rory Donovan-Maiye. Langpert: LLM-driven contextual synthesis for unseen perturbation prediction. In *ICLR 2025 Workshop on Machine Learning for Genomics Explorations*, 2025. URL https://openreview.net/forum?id=Tmx4o3Jg55.

Meta. Llama-3.3-70B-Instruct model card. Hugging Face, December 2024. URL https://huggingface.co/meta-llama/Llama-3.3-70B-Instruct.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Sophie Ostmeier, Justin Xu, Zhihong Chen, Maya Varma, Louis Blankemeier, Christian Bluethgen, Arne Edward Michalson, Michael Moseley, Curtis Langlotz, Akshay S. Chaudhari, and Jean-Benoit Delbrouck. GREEN: Generative radiology report evaluation and error notation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 374–390, 2024. doi: 10.18653/v1/2024.findings-emnlp.21. URL https://aclanthology.org/2024.findings-emnlp.21/.

Kanglu Pei, Sisi Qu, Philip H. S. Torr, Jonathan G. Hedley, and Christian Schroeder de Witt. Perturbagent: An agentic AI system for analysis and prediction of genetic perturbations. AAAI 2026 Workshop on Explainable AI for Science, 2026. URL https://openreview.net/forum?id=FfHATDCkCQ.

Lawrence Phillips, Marc Boubnovski Martell, Aditya Misra, Josefa Lia Stoisser, Cesar A. Prada-Medina, Rory Donovan-Maiye, and Kaspar Märtens. Synthpert: Enhancing LLM biological reasoning via synthetic reasoning traces for cellular perturbation prediction. *arXiv preprint*

arXiv:2509.25346, 2025. doi: 10.48550/arXiv.2509.25346. URL https://arxiv.org/abs/2509.25346.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Yusuf Roohani, Kexin Huang, and Jure Leskovec. Predicting transcriptional outcomes of novel multigene perturbations with GEARS. *Nature Biotechnology*, 42(6):927–935, 2024. doi: 10.1038/s41587-023-01905-6.

Ruth L. Seal, Bryony Braschi, Kristian Gray, Tamsin E. M. Jones, Susan Tweedie, Liora Haim-Vilmovsky, and Elspeth A. Bruford. Genenames.org: the HGNC resources in 2023. *Nucleic Acids Research*, 51(D1):D1003–D1009, 2023. doi: 10.1093/nar/gkac888.

Mingyang Song, Zhaochen Su, Xiaoye Qu, Jiawei Zhou, and Yu Cheng. PRMBench: A fine-grained and challenging benchmark for process-level reward models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25299–25346, 2025. doi: 10.18653/v1/2025.acl-long.1230. URL https://aclanthology.org/2025.acl-long.1230/.

The UniProt Consortium. UniProt: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Stephen E. Toulmin. *The Uses of Argument*. Cambridge University Press, 1958.

George Tsitsiridis, Ralph Steinkamp, Madalina Giurgiu, Barbara Brauner, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. CORUM: the comprehensive resource of mammalian protein complexes–2022. *Nucleic Acids Research*, 51(D1):D539–D545, 2023. doi: 10.1093/nar/gkac1015.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*, 2023.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. URL https://aclanthology.org/2020.emnlp-main.609/.

Zhenyu Wang, Geyan Ye, Wei Liu, and Man Tat Alexander Ng. AROMA: Augmented reasoning over a multimodal architecture for virtual cell genetic perturbation modeling. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 31855–31881, 2026. doi: 10.18653/v1/2026.findings-acl.1594. URL https://aclanthology.org/2026.findings-acl.1594/.

Zhijian Wei, Runze Ma, Zichen Wang, Zhongmin Li, Shuotong Song, and Shuangjia Zheng. VC-WORLD: A biological world model for virtual cell simulation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=hhq89Hs7T3.

Menghua Wu, Russell Littman, Jacob Levine, Lin Qiu, Tommaso Biancalani, David Richmond, and Jan-Christian Huetter. Contextualizing biological perturbation experiments through language. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://github.com/Genentech/PerturbQA.

Feiyang Yu, Mark Endo, Rayan Krishnan, Ian Pan, Andy Tsai, Eduardo Pontes Reis, Eduardo Kaiser Ururahy Nunes Fonseca, Henrique Min Ho Lee, Zahra Shakeri Hossein Abad, Andrew Y. Ng, Curtis P. Langlotz, Vasantha Kumar Venugopal, and Pranav Rajpurkar. Evaluating progress in automatic chest x-ray radiology report generation. *Patterns*, 4(9):100802, 2023. doi: 10.1016/j.patter.2023.100802. URL https://doi.org/10.1016/j.patter.2023.100802.

Zhiyun Zhang, Liwen Sun, Xiang Qian, and Chenyan Xiong. Faithmed: Training LLMs for faithful evidence-based medical reasoning. *arXiv preprint arXiv:2607.01440*, 2026. doi: 10.48550/arXiv.2607.01440. URL https://arxiv.org/abs/2607.01440.

Chujie Zheng, Zhenru Zhang, Beichen Zhang, Runji Lin, Keming Lu, Bowen Yu, Dayiheng Liu, Jingren Zhou, and Junyang Lin. ProcessBench: Identifying process errors in mathematical reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1009–1024, 2025. doi: 10.18653/v1/2025.acl-long.50. URL https://aclanthology.org/2025.acl-long.50/.

APPENDIX

A BENCHMARK CONSTRUCTION AND PROVENANCE

Design. The benchmark (`benchmark-600-v1`, random seed 42) pools 600 complete model-generated traces from eight systems. All three evaluator pipelines audit the same 600 records; cross-evaluator statistics additionally require a determinate score from each compared evaluator for the dimension in question. Table 5 gives the stratification.

Table 5: Stratification of the 600-trace benchmark. SUMMER retains all 38 available GSE traces rather than duplicating them, so its per-task counts are uneven. No traces are duplicated to equalize task quotas.

System	Cases	Task coverage
rbio	60	DE
SUMMER	180	DE (71) / Direction (71) / Gene-set interpretation (38)
AROMA	60	Direction
SynthPert	60	Direction
PerturbAgent	60	Direction-probability
CellReasoner	60	Cell-type annotation
VCWorld	60	DE (26) / Direction (34)
LangPert	60	Candidate selection for expression-effect transfer
Total	600	six task types

Selection principles. (1) *Original sources first*: every record keeps its complete original trace, task question, answer field, source file, and source index; traces are never rewritten. (2) *No truncation to fit*: records with explicit truncation markers, empty or `nan` fields, or below per-task quality thresholds are excluded. (3) *Traceable repairs*: the CellReasoner empty responses and 8,000-character hard truncations found during sampling were replaced with the complete responses from the same original workbook; pre-/post-replacement hashes and lengths are recorded in the quality report. (4) *Deduplication*: no repeated `trace_sha256` within a system, and the same text is not double-counted across tasks. (5) *Labels are local-only*: gold labels, correctness flags, and source identity are never sent to an evaluator; requests contain only the task, context, trace, and the frozen evidence dossier. (6) *Label balance is not a natural distribution*: for systems with reliable labels, sampling keeps correct and incorrect outcomes close to balanced; unlabeled systems are sampled randomly and reported as unlabeled. (7) *Task adaptation does not change the common framework*: the completeness checklist varies by task, but the definitions of Claim, Premise, Warrant, Link, Path, the cogency dimensions, and Completeness are task-invariant.

Label availability. Of the 600 traces, 382 carry a normalized correctness flag (`is_correct`) and 467 carry a non-empty `gold_label`. These are different counts: a reference cell type or gene-set name is not automatically a normalized binary correctness judgment. The current correctness-conditioned analyses use only the 382 existing flags. Other traces remain eligible for content auditing; missing flags are not replaced with pseudo-labels.

Quality gates. Each regeneration must pass: total/system/task/split count checks; unique `sample_id`; within-system trace-hash uniqueness; non-empty, non-truncated traces; existence of the original source file and a resolvable source index; reportable length, label, and task distributions for both splits; and a fixed random seed for the generator script.

Files. System names and case IDs use publication-facing aliases, with mappings to unchanged canonical identifiers in `data/benchmark/analysis/paper_display_aliases.json`. `data/benchmark/canonical/samples_600.jsonl` (full 600), `samples_600_core300.jsonl` and `samples_600_expansion300.jsonl` (historical sampling partitions, both now included in every full run), `benchmark_600_quality.json` (distributions, replacements, quality report), and `scripts/prepare_benchmark_600.py` (reproducible generator).

B AUDIT PROTOCOL AND REPRODUCIBILITY

Pipeline. All reported pipelines use `D_dimension_split`: *nodes* reconstructs the node inventory and conclusion pointer; *paths* constructs Links and Paths over those nodes; *audit* compares the draft with the source; *revise* returns a revised graph. Four independent *dimension* calls then assess Acceptability, Relevance, Sufficiency, and Completeness on that frozen graph, without seeing each other's judgments. This is eight scheduled calls before retries and structural recovery, with two logical agent roles rather than eight distinct agents.

Table 6: Configured completion-token ceilings for the initial 600-case runs. These are per-call upper bounds, not token usage or guaranteed output lengths. Context limits may reduce a local call's available completion budget.

Stage	Qwen / Llama	DeepSeek API	Output contract
nodes	8,000	16,000	Nodes and conclusion pointer
paths	6,000	16,000	Links and target-conditioned Paths
audit	8,000	12,000	Findings and repair instructions
revise	12,000	16,000	Complete revised graph
dimension ($\times 4$)	6,000	16,000	One dimension per call

Shared semantics, non-identical requests. Qwen3.6-35B-A3B (Qwen Team, 2026) and Llama-3.3-70B-Instruct (Meta, 2024) run locally; DeepSeek-v4-pro (DeepSeek-AI, 2026) uses the official API. They share the graph/evaluation types, task checklists, dimension definitions, stage order, and deterministic aggregation. Their realized prompts are *not* identical. The runs use the compact prompt profile; local models additionally receive graph-layout, decomposition, and dimension guards, including dimension-specific scoring anchors. Compact calls omit the reference-example set that is present in the full prompt variant. Appendix C records these differences.

Local dimension calls receive graph spans instead of the full source trace. Their Acceptability call receives at most 12 frozen dossier records; Relevance, Sufficiency, and Completeness receive no dossier. The official DeepSeek dimension calls retain the full source payload and dossier. Local node extraction omits the dossier, whereas the API retains the supplied base payload. Consequently, cross-evaluator differences combine model, reconstruction, prompt, and reference exposure: they are end-to-end pipeline sensitivity, not a controlled model-only or fixed-graph comparison.

Blinding and reference scope. Model-visible task payloads exclude source-system metadata, gold labels, correctness flags, and source-file paths; identifiers in orchestration logs are not added to the task text. Explicit self-identification in an original trace is not silently rewritten. Dossiers are frozen before evaluation. An identity or function record supports only the annotation it contains, not an unobserved perturbation response. A visible tool log can establish what the trace reports retrieving, not the exhaustiveness of its database.

Structural recovery and eligibility. Per-call format retries have a fixed limit and do not add scheduled substantive refinement passes. Structural recovery is versioned separately. A reportable record has status `ok`, an `eligible_for_reporting` flag, and zero hard errors in both graph and evaluation audits. Recovery uses the same worker model where new judgments are required. Deterministic repairs resolve unambiguous identifier collisions, reconcile final-Claim role labels with the declared endpoint, recover exact quotations where possible, and split malformed Paths according to existing Link dependencies. Unrecoverable quotations retain their proposition as explicitly ungrounded reconstruction, with no invented span. Links and biological propositions are not improved to make the source argument more convincing.

Changed graphs receive fresh dimension assessments; missing dimension coverage can be repaired without rerunning successful dimensions. Original results, intermediate repairs, and refreshed judgments remain separately versioned. Warnings and unknown scores are retained: a structurally valid 600-case run is not 600 biologically verified analyses. No-path cases may remain structurally admissible with missing support, rather than being given an invented supporting Path. These rules follow the same audit policy used to freeze the two local runs.

Transport or parsing failures are distinguished from scientific uncertainty: unknown placeholders left by failed dimension calls are retried, not counted as completed judgments. The final DeepSeek snapshot has no unresolved dimension-call fallbacks; naturally undetermined assessments remain unknown.

The API recovery uses the current question-grounding clarification in the compact instruction. The scoring configuration and four dimension schemas match the source run, but the global prompt/schema fingerprints differ; we retain both versions rather than describe the repair requests as byte-identical to the original protocol.

Aggregation and dependence. The orchestrator validates connectivity, averages distinct complete-Link scores within valid Paths, and averages distinct Paths separately by target and support/rebuttal polarity. Shared Links reuse their local score; exact duplicate Paths are deduplicated. Partial overlap is retained but not given a learned dependence correction, so the mean is not an estimate of independent corroboration. Qualifier preservation, weakening, strengthening, or conflict is recorded at Link, Path, and final-Claim levels as a Sufficiency diagnostic, not an additional top-level dimension or a second Link-to-Link score. Acceptability averages known premise assessments; Completeness averages applicable checklist items. Unknown values are excluded and counted, never zero-imputed. Reported dimension means use the saved normalized summaries on a 0–10 display scale.

Reproducibility. Run manifests and saved requests retain model, prompt, schema, rubric, token budget, and retry provenance. Frozen results and sample SHA-256 digests accompany the tables produced by `scripts/summarize_paper_system_scores.py`. Bootstrap intervals use 2,000 trace-level draws (seed 42), conditional on available scores. System summaries use equal trace weights except SUMMER, whose three task means receive equal weight. Matched sensitivity tables use the same trace IDs across evaluators separately for each dimension and state their own weighting and denominators.

C PROMPT LIBRARY

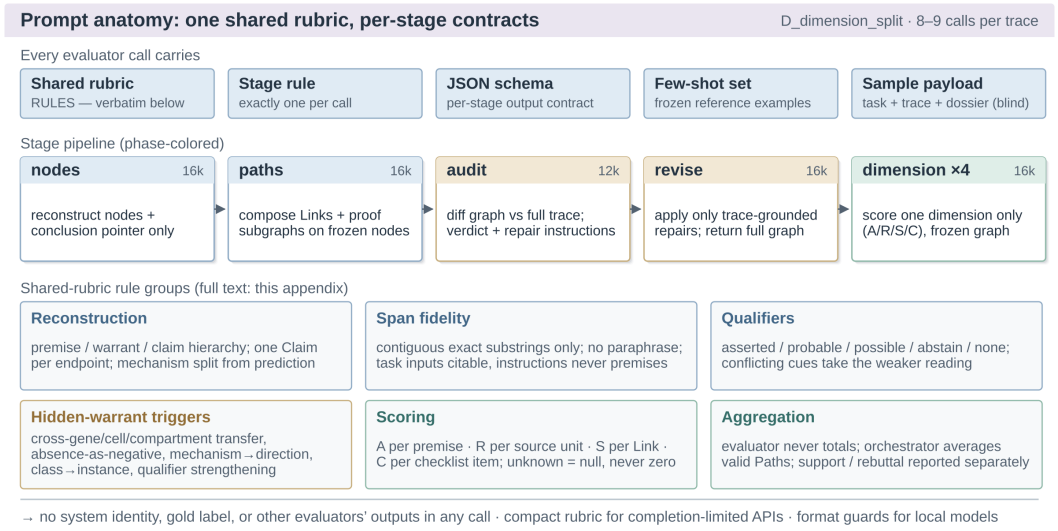

Figure 3: Prompt components and stage contracts. The three 600-case runs use the compact profile with eight scheduled calls before retries. Local and API requests differ in guards, scoring-anchor exposure, and reference payloads (Appendix B). Few-shot examples belong to the full variant, not these compact calls.

The operative prompt is assembled from a shared instruction, a stage rule, a JSON output schema, and the blinded stage payload. The blocks below quote the compact-profile constants with whitespace reflowed and an em dash rendered as two hyphens. Local backends append the indicated guards and receive dimension-specific rubric anchors. Additional per-sample ID contracts enumerate the objects that must be covered; they are generated from the frozen graph and checklist, not from gold outcomes. Saved request files are the authoritative record of what each call actually received. The blocks show the current recovery version; the original API prompt lacked the later clarification that factual task inputs can provide exact question spans. The four dimension output schemas and scoring configuration are unchanged. Source and recovery manifests retain their different prompt/schema fingerprints.

C.1 SHARED COMPACT INSTRUCTION

Shared compact instruction — issued with every stage call

```
Evaluate the supplied trace, not whether its biological answer is true.
Return ONLY one JSON object matching the supplied schema; never return a
private reasoning transcript or Markdown. Decide quickly and emit the
artifact immediately; do not spend tokens deliberating. Preserve the
author's claims and uncertainty; do not invent facts or repair the
trace. Use exactly these field names: node_class, node_subclass,
normalized_statement, span, source_field, origin, qualifier,
qualifier_span, occurrences; link fields id, from_ids, warrant_ids,
to_id, relation; path fields id, path_type, step_ids, target_id, origin,
summary, dependency_pairs. Nodes are premise
(experimental_result/knowledge), warrant (visible/hidden), or claim
(intermediate/final). Explicit spans must be contiguous substrings of
trace or answer; a task-supplied factual input may be a Premise with an
exact question span, but task instructions are not Premises. Hidden
warrants have span='', source_field='none',
origin='implicit_reconstruction', qualifier='none'. Claims with the same
endpoint are one final Claim with occurrences, not qualifier-only
duplicate Claims. A Link is a one-hop edge from premise/claim IDs plus
warrant IDs to one claim target; relation is support or rebuttal.
Rebuttal is a path, never a node. A Path is target-conditioned;
dependency_pairs list only actual Claim-fed link dependencies, and
```

```
summary must be human-readable. Add hidden warrants for
cross-gene/context/abstraction transfers, absence-as-negative evidence,
or qualifier strengthening. Preserve every numbered task component. For
evaluation, assess each premise, each Link source component (Relation),
each complete Link (Sufficiency), qualifier loyalty at
link/path/final_claim scope, and each task requirement; scores are
anchored integers 0..10 where the schema says so, unknown uses null. Do
not compute totals that the schema does not request.
```

C.2 COMPACT STAGE RULES

Stage rule — nodes

```
Return nodes and conclusion_id only; do not create links or paths.
```

Stage rule — paths

```
Using frozen_nodes only, return links and paths. Do not add or change
nodes; record actual dependency_pairs for converging branches.
```

Stage rule — audit

```
Audit the complete trace against frozen_graph. List omissions,
over-merges, missing links/warrants, qualifier changes, and every
numbered subtask; give precise trace-grounded repairs only.
```

Stage rule — revise

```
Apply only the audit repairs. Return a complete graph, preserve
unaffected IDs/spans, merge qualifier-only duplicate endpoint Claims,
restore omitted branches, and do not invent facts or rebuttals.
```

Stage rule — dimension

```
Assess only the requested dimension on frozen_graph. Keep Relation
source-component scores separate from complete-Link Sufficiency scores;
runtime computes Path/Claim aggregates. Be extremely concise: one item
per required target, reason <=12 words, no optional narrative, no
duplicate items. Obey null/zero rules exactly (refuted score=0;
unknown/not_applicable score=null).
```

C.3 LOCAL-BACKEND ADDITIONS

These instructions can affect reconstruction and scoring, not merely JSON formatting. They are documented as protocol differences rather than evidence that only the endpoint changed.

Local guard — all local calls

```
LOCAL FORMAT GUARD (mandatory): hidden warrants are reconstructed
metadata, not authorial utterances. For every node with
node_class=warrant and node_subclass=hidden, set qualifier="none",
qualifier_span="", span="", source_field="none", and
origin="implicit_reconstruction". Do not encode the logical strength of
a hidden warrant as an authorial qualifier. Scan all hidden warrants
before returning. IDs are globally unique across the whole artifact: use
p*/w*/c* for nodes, l* for Links, and s*/r* for Paths; never reuse a
node ID for a Path. Path.step_ids contain Link IDs only,
dependency_pairs contain only [Link ID, Link ID] pairs, and a one-Link
Path has an empty dependency_pairs list. In each Link, from_ids may
```

```
contain only p*/c* node IDs and warrant_ids may contain only w* node
IDs; never put a warrant in from_ids. Set
Path.origin="implicit_reconstruction" unless the trace explicitly states
the complete formal path. Qualifier is exactly one of asserted,
probable, possible, abstain, or none; never emit labels such as
uncertain, likely, or unknown (map a weak/uncertain stance to possible).
Never upgrade a trace qualifier because the answer label is binary:
preserve a cautious phrase such as "does not necessarily", "may", or
"could" as possible/probable and record any answer mismatch for the
evaluator. Every explicit, decision-relevant premise must be consumed by
at least one Link; an alternative-factor premise must be connected as a
support/rebuttal branch rather than left orphaned. For Completeness
checklist Items, encode present as status="supported", absent as
status="refuted", and partial coverage as status="partial"; the shared
Item schema does not accept the alias "absent". Return raw JSON without
``` fences. The occurrences array is Claim-only: premise and warrant
nodes must use occurrences=[]. If the answer is a bare label such as
"no", use an exact trace sentence as the final Claim span, not a
fabricated sentence such as "The answer is no."
```

Graph output and decomposition. The output guard applies to graph-producing stages; the decomposition guard applies to nodes, graph, and revise stages.

Local guard — graph output

```
Graph and revision outputs contain only graph fields (nodes, links,
paths, conclusion_id). Never attach score, reason, or evaluation fields
to a node, Link, or Path.
```

Local guard — graph decomposition

```
LOCAL GRAPH DECOMPOSITION (mandatory): distinguish a serial mechanism
from independent support. If the trace has the form A causes B, B causes
C, and C supports endpoint D, do not create separate A->D, B->D, and
C->D Links. Reconstruct the applied intermediate conclusions that are
outputs of earlier steps and inputs to later steps, then connect them
topologically. For example, premises 'A knockdown impairs X' and
'impaired X causes stress' support an implicit intermediate Claim 'A
knockdown causes stress'; that Claim plus 'stress induces Y' can support
the next Claim. An implicit Claim has span='', source_field='none', and
origin='implicit_reconstruction'; it must not pretend to be quoted.
Conversely, do not manufacture intermediate Claims when sources are
genuinely parallel and jointly support one target. A Path summary that
narrates a serial chain must be represented by Claim-fed Link
dependencies, not by a flat list of unrelated Links all targeting the
final Claim. Every ID used by a Link or Path must exist in frozen_nodes;
never refer to an uncreated warrant or Claim.
```

Local guard — dimension calls

```
LOCAL DIMENSION FORMAT (mandatory): The requested dimension is the only
dimension being scored. For acceptability and completeness, each
item.score is an integer 0..3; for relevance RelationAssessment.score
and sufficiency LinkSufficiencyAssessment.score/qualifier scores are
integers 0..10. Use null only with status unknown or not_applicable. For
acceptability, Item.basis must be database, trace, inference, or none
and database citations must be supplied record IDs; never use
evidence_dossier as a basis. For completeness, assess exactly the
supplied requirements and use only status supported, partial, unknown,
or not_applicable; use partial with a low score for an omitted but
applicable requirement and never emit the word absent. Use
basis=inference or trace as appropriate. For relevance, cover every Link
source component exactly once or group only jointly necessary sources,
preserve the exact link/source/target/warrant IDs, and write a
source-specific reason. For sufficiency, cover every Graph Link and
return qualifier assessments at link, path, and final_claim scopes when
those targets exist; do not invent Path or Claim totals. Hidden warrants
have no authorial qualifier, and a cautious source qualifier must not be
upgraded to asserted because the benchmark answer is binary. Keep
reasons concise and return only the schema object, with no Markdown or
```
```

```
commentary. The arrays must be emitted as one JSON array, for example
"link_assessments":[{"id":"la1","link_id":"l1"},
{"id":"la2","link_id":"l2"}]
--never as two adjacent objects or two repeated top-level keys.
```

Additional audit and revision guards. The local audit checks for unconsumed decision-relevant premises and serial mechanisms flattened into independent premise-to-final Links. Local revision preserves cautious qualifiers and adds only source-grounded applied Claims needed to represent an already narrated serial chain. These dynamically appended guards are implemented in `scripts/compare_prover_architectures.py`, method `Experiment.invoke`; they and the per-sample ID contract are retained in the saved requests.

C.4 Schemas, anchors, and prompt variants

Schemas are defined in `src/evidence_scoring/prover_architecture.py`. Reconstruction returns `nodes`, `conclusion_id`, `links`, and `paths`. Acceptability and Completeness use premise/checklist `items`; Relevance uses `relation_assessments`; Sufficiency uses `link_assessments` and `qualifier_assessments`. Derived Path and Claim collections are computed by the runtime, not requested as independent model totals.

The frozen rubric in `framework/framework.yaml` uses 0–3 for Acceptability and Completeness and 0–10 for Relevance and Sufficiency. For Relevance, local rubric anchors distinguish topical overlap (1–2), indirect or partial help (3–5), specific but limited support (6–8), and directly decisive support (9–10), with 0 for no support. Explicit null states remain distinct from numeric zero.

The full prompt variant uses `RULES`, `STAGE_RULES`, the full rubric, and `configs/prover_reference_examples.json`. It is retained as a reproducibility artifact but is not the realized prompt for the three compact 600-case pipelines. In particular, the compact API dimension call retains the schema but does not insert the local dimension-specific rubric block; this is one reason the reported comparison cannot isolate model choice. No new few-shot examples or retrospective gold labels are added during structural recovery.

D Task-Specific Reasoning Schemas

The audit framework is task-invariant, but each task type induces a characteristic argument schema: a typical final-Claim shape, the premise types a complete analysis is expected to draw on, and the reasoning forms that legitimately connect them. This appendix states, for each of the six audited task types, (i) the abstract schema, (ii) the task-conditioned Completeness checklist, and (iii) the characteristic audit risks the schema exposes. The checklists below describe the frozen criteria and their task interpretation; the legacy GSE wording requires the caveat given below. Scoring anchors are *present* $= 3$, *partial* $= 1$–2, *absent* $= 0$, and *not applicable* $=$ null, where a requirement that is applicable but undiscussed must be scored absent or partial, never not-applicable.

Three requirements are common to all tasks:

- **context** — preserve task-relevant entities, background, and scope of applicability;
- **provenance** — distinguish empirical observations, knowledge, and mechanistic inference, and state which sources were available or missing;
- **boundaries** — express uncertainty, alternative explanations, or applicability boundaries when decision-relevant.

D.1 Differential expression (DE)

Schema. Final Claim: *perturbing gene g_p changes the expression of gene g_t in context k* (polarity: change / no-change). Expected premises: an instance-level perturbation–response observation for the pair or a close analogue (experimental_result), gene-function knowledge, and, when retrieval is used, an explicit record of retrieval scope. Legitimate forms: response-grounded abstraction (transfer an observed response under a stated similarity warrant) and mechanism-grounded explanation (connect g_p to g_t through intermediate molecular Claims).

Task checklist. *relationship* — provide instance-level relational evidence, or an explicitly bounded retrieval together with the limits of its conclusion.

Characteristic risks. Cross-compartment or cross-gene analogy under an over-coarse shared category (e.g., transferring via "ribosome" across mitochondrial and cytosolic systems); upgrading an empty retrieval into evidence of no effect; promoting class-level facts into instance-level edges. Each requires a hidden warrant at audit time, which is where the support typically fails.

D.2 Direction (Dir.)

Schema. Final Claim: g_t *is upregulated / downregulated upon perturbing* g_p *in* k, conditioned on a stated DE premise. Expected premises: as for DE, plus at least one *signed* relation (activates/inhibits, or a response observation with direction) on the supporting path.

Task checklist. *signed_path* — present signed evidence or a signed mechanism and explain how it supports the chosen direction.

Characteristic risks. Unsigned mechanism transferred to a direction (the path is real but carries no sign); qualifier escalation from *possible* to *asserted* between trace and answer; a trace–answer sign conflict in which the narrative concludes "essentially unchanged" while the submitted label is *upregulated* (Appendix H).

D.3 Direction with probabilities (direction-probability)

Schema. Final Claim: a probability assignment over {up, down, no-change}, or an abstention. Expected premises: retrieval results *with their scope*, functional annotations, and an explicit link between evidence mass and the reported numbers.

Task checklist. *signed_path* — directional evidence, or explicit retrieval scope and the reason direction remains undetermined; *probability* — reported probabilities, no-effect outcomes, and abstention semantics consistent with the evidence limits.

Characteristic risks. An empty search upgraded to a confident $0/0$ or uniform assignment (absence treated as no-effect evidence); "label bridging", where a confident assignment is stated without adequate supporting evidence. Explicitly describing an unsuccessful search is good provenance practice, but does not by itself justify either no effect or abstention. PerturbAgent-0 illustrates this distinction (Appendix H.3); its $0/0$ output accompanies an explicit no-effect decision, not an abstention.

D.4 Gene-set interpretation (release label: GSE)

Schema. The 38 released questions ask the model to *describe the function of a supplied Perturb-seq gene set and name it*; they do not ask for a binary enrichment decision. Expected premises are the supplied gene membership and functional annotations. The conclusion is a functional description or name, with upstream processes discussed in the released response template.

Task checklist. The frozen *gene_set* criterion refers to set members, set-level enrichment evidence, and conclusion scope. Its enrichment wording is broader than the released naming task. We retain that wording and the existing scores for reproducibility, but flag this task–rubric mismatch: unrequested enrichment statistics are not a demonstrated omission in these traces. GSE Completeness therefore needs task-specific human calibration before being interpreted as naming-task coverage.

Characteristic risks. Overgeneralizing a subset's function to every member; confusing a functional label with a demonstrated perturbation response; or treating a possible upstream influence as an observed experimental finding.

D.5 CELL-TYPE ANNOTATION

Schema. Final Claim: *the cell belongs to type t.* Expected premises: observed markers (experimental_result), the knowledge warrant mapping markers to types, and exclusion evidence for confusable alternatives. This task extends the framework beyond perturbation response; the same node types apply, with the final Claim attached to an identity rather than a response.

Task checklist. *markers* — distinguish observed markers from the knowledge warrant that maps markers to cell types; *alternatives* — provide discriminating evidence for confusable cell types and the boundary of the negative evidence.

Characteristic risks. Lineage-level support presented as subtype-level proof, if the trace actually claims that subtype; exclusion of alternatives without bounded negative evidence. A finer-grained reference label must not silently replace the Claim being audited. CellReasoner-x132 concludes only T-cell, whereas its held-out label is Naive CD4 T (Appendix H.5).

D.6 EXPRESSION-PROFILE REGRESSION (CANDIDATE SELECTION)

Schema. Final Claim: *a candidate set of perturbations is submitted for effect transfer*. Expected premises: the actually submitted candidate list, the selection rationale, and the source and applicability of each candidate's effect estimate.

Task checklist. *candidates* — the actually submitted candidate set and its selection rationale; *effect_transfer* — the source and applicability of candidate perturbation effects, marked unobservable when the external computation is not exposed; *aggregation* — distinguish textual candidate selection from downstream numeric aggregation, without fabricating unpresented numeric predictions.

Characteristic risks. Narrating numeric computations the trace never shows; treating all listed candidates as equally grounded when the list itself mixes direct matches and peripheral candidates (Relation-level assessment preserves this heterogeneity).

E PER-SYSTEM PROFILES

Each profile states the audited task, the primary-evaluator (Qwen3.6-35B-A3B) metric profile from Table 4, illustrative reasoning forms observed in the decomposition, and one representative audit finding. Values are descriptive single-evaluator profiles, not independently validated quality rankings. A coverage differs sharply across systems (Table 7), so A means are not equally informative. Mechanism and Evidence-based reasoning follow the Path-level distinctions in Section 4.4; they do not label a system's training procedure or imply that all of its traces use one form. The empirical-trace gap discussed there concerns explicit accounts of learned response patterns, not the absence of empirical learning in these systems.

rbio (DE; N=60). Profile: A 8.90 (21 traces), R 4.98, S 3.84 (59), C 9.69. Traces are short, template-consistent functional arguments. Representative finding: answer–trace inconsistency is observable even when the submitted answer is correct (rbio-4: the trace predicts that SOS1 knockdown likely changes ARL5B expression while the submitted, correctly labeled answer is *no*; Appendix H). The low A denominator reflects traces whose premises lack determinate verification under the supplied dossier; this is not evidence that those premises are false.

SUMMER (DE/Dir./GSE; N=180). Profile: A 9.20 (128), R 4.25, S 3.59 (179), C 9.28 (equal-weight macro over tasks). Task templates organize five DE questions, seven direction questions, or three GSE prompts. Coverage is high under the frozen checklists, but support lags: DE and direction traces score lower on R/S (3.82/3.27 and 3.41/3.39) than GSE (5.53/4.10). Representative finding: SUMMER-0 fills every question slot, yet bridges CMTR2 observations to RPS11 through the over-coarse category "ribosome" across compartments — the canonical high-coverage, weak-bridge pattern. SUMMER-66 (GSE) is a scope-matched contrast: it describes and names the supplied set, rather than answering an unasked binary enrichment question. The legacy GSE checklist caveat applies (Appendix D.4).

AROMA (Dir.; N=60**).** Profile: A 9.31 (35), R 4.71, S 3.48 (60), C 10.00. Traces are mechanism-shaped rationales assembled from retrieved gene descriptions, cell-line context, and associative pathway paths. Representative finding: AROMA-0 argues from functional dissimilarity and absence of direct interaction to *no significant change*; the exclusion warrant does not address indirect or compensatory paths. The separate 100-case input-ablation audit (Appendix I) finds low prediction sensitivity to the tested context changes, without proving that the rationale has no causal role.

SynthPert (Dir.; N=60**).** Profile: A 9.32 (55), R 5.71, S 5.27 (47), C 8.29. Fluent mechanistic narratives; the strongest R/S among the direction-prediction systems under the primary evaluator. Representative finding: SynthPert-0 presents a damage–compensation narrative for PIK3C3→HSPD1 with an incorrect direction and no instance-specific experiment cited. Its high automatic scores also illustrate the need for evaluator calibration, not a confirmed detection success. SynthPert-x071 adds the trace–answer sign conflict (Appendix H).

PerturbAgent (direction-probability; N=60**).** Profile: A 10.00 (55), R 2.90, S 2.79 (60), C 9.83; highest A coverage (86.2% known premises), since its premises are concrete retrieval outcomes. Representative finding: PerturbAgent-0 documents unsuccessful retrieval and a PPI search, but still infers no significant effect. Verifiable search provenance and adequate support for a no-effect conclusion are different questions (Appendix H.3); the trace is not an abstention.

CellReasoner (cell-type annotation; N=60**).** Profile: A 9.76 (42), R 7.09, S 6.00 (57), C 9.87 — the strongest R/S profile under the primary evaluator. Traces use marker evidence and explicit exclusion warrants. Representative finding: CellReasoner-x132 concludes T-cell while the held-out annotation is Naive CD4 T. The latter may matter for answer granularity, but it must not be introduced as an unstated target of the Sufficiency audit. Marker assertions and alternative exclusions still need premise-level checking.

VCWorld (DE/Dir.; N=60**).** Profile: A 9.44 (9), R 4.40, S 3.92 (59), C 9.64; lowest A coverage (11.8%), so its A mean rests on very few determinate premises and should not be compared with high-coverage systems. The released DE and direction traces combine gene-function descriptions with proposed downstream effects; when an effect is explained mechanistically rather than supported by a visible response observation, reconstruction retains that distinction.

LangPert (expression-profile regression; N=60**).** Profile: A 9.75 (58), R 5.18, S 4.14 (58), C 8.87. Traces are decision-bearing candidate-selection rationales. Representative finding: LangPert-10's candidate list mixes direct matches with peripheral candidates; Relation-level assessment preserves this intra-list heterogeneity that a holistic score would flatten.

F Complete Metric Results

F.1 System-by-task profiles (primary evaluator)

Table 7 expands the primary Qwen profile by task. All dimension summaries are displayed on 0–10. Acceptability coverage is the trace-average known/applicable premise fraction, not verification accuracy. The score mean conditions on traces with a determinate value; unknown values are excluded and counted, never imputed as zero. Bootstrap intervals and exact sample IDs accompany the per-worker CSV/JSON artifacts under `data/benchmark/analysis/`. GSE retains the task–checklist caveat in Appendix D.4.

Table 7: Primary Qwen system-by-task results. N is cohort size; parentheses give determinate trace counts for A and S; cov is Acceptability coverage. R and C have values for every trace. SUMMER macro averages its three task means equally; VCWorld pooled uses equal trace weights.

System / task	N	A (n)	cov	R	S (n)	C
rbio / DE	60	8.90 (21)	19.7%	4.98	3.84 (59)	9.69
SUMMER / DE	71	9.58 (53)	42.2%	3.82	3.27 (70)	9.40
SUMMER / Dir.	71	9.88 (48)	37.7%	3.41	3.39 (71)	9.62
SUMMER / GSE	38	8.15 (27)	43.4%	5.53	4.10 (38)	8.82
SUMMER macro	180	9.20 (128)	41.1%	4.25	3.59 (179)	9.28
AROMA / Dir.	60	9.31 (35)	31.7%	4.71	3.48 (60)	10.00
SynthPert / Dir.	60	9.32 (55)	43.9%	5.71	5.27 (47)	8.29
PerturbAgent / Dir.-proba	60	10.00 (55)	86.3%	2.90	2.79 (60)	9.83
CellReasoner / cell-type	60	9.76 (42)	33.1%	7.09	6.00 (57)	9.87
VCWorld / DE	26	10.00 (4)	11.7%	3.98	3.27 (25)	9.42
VCWorld / Dir.	34	9.00 (5)	12.0%	4.71	4.40 (34)	9.80
VCWorld pooled	60	9.44 (9)	11.8%	4.40	3.92 (59)	9.64
LangPert / candidates	60	9.75 (58)	76.6%	5.18	4.14 (58)	8.87

F.2 Complete sensitivity-worker profiles

Both sensitivity workers cover the same 600 canonical records as Qwen. The official DeepSeek recovery resolved the original 56 ineligible records and 18 additional multi-final-Claim violations exposed by the current common audit contract. In total, 64 graphs required representation changes; six further cases needed missing dimension coverage and four were revalidated without new model calls. No model-generated graph revision was needed: propositions and Links were preserved, while identifiers, source grounding, Claim roles, and Path serialization were normalized. All 600 records now pass structural checks; this does not imply determinate values in every dimension.

F.3 Matched evaluator sensitivity

Cross-worker contrasts use exactly the same traces within each dimension, requiring a determinate value from every compared worker. The three-way cohort therefore differs across dimensions. Table 9 weights traces equally, including SUMMER; it is not an average of the equal-task system means above. Graphs, prompts, and reference exposure differ between pipelines (Appendix B), so this is not a fixed-graph or controlled model-only experiment.

On the maximal Qwen–Llama Sufficiency pairing ($n = 538$), means are 3.96 and 6.60; the paired Llama-minus-Qwen difference is 2.64 (95% paired trace-bootstrap interval $[2.47, 2.81]$). All pairwise cohorts, differences, and intervals are recorded in `data/benchmark/analysis/evaluator_sensitivity_600.json`.

The shifts are dimension-dependent, not a uniform leniency offset. DeepSeek assigns higher Relevance but lower Acceptability and Completeness than Qwen on their three-way matched sets; Llama gives higher Sufficiency but lower Completeness. These comparisons do not establish which worker is correct. The primary table remains single-worker and descriptive, not an independently validated quality ranking. Agreement on individual failure diagnoses is not measured by this score-level comparison; structural completion likewise does not certify substantive judgments.

Table 8: Llama and official DeepSeek profiles. The weighting and missing-value policy match the primary table; parentheses give determinate trace counts. Each block contains 600 traces, with 180 SUMMER cases and 60 per other system.

System	A (n)	cov	R	S (n)	C (n)
Llama-3.3-70B-Instruct					
rbio	10.00 (40)	40.2%	5.41	6.58 (26)	6.62 (60)
SUMMER	9.27 (78)	33.9%	6.55	7.02 (180)	7.69 (180)
AROMA	10.00 (28)	28.0%	6.35	6.67 (59)	7.57 (60)
SynthPert	9.93 (49)	48.6%	6.11	6.94 (52)	7.14 (60)
PerturbAgent	9.59 (34)	44.1%	3.02	3.06 (60)	6.68 (60)
CellReasoner	9.81 (45)	46.2%	7.03	7.84 (60)	7.34 (60)
VCWorld	10.00 (3)	3.2%	6.49	6.95 (58)	7.74 (60)
LangPert	10.00 (58)	82.6%	7.28	7.44 (60)	8.38 (60)
DeepSeek-v4-pro (official API)					
rbio	6.49 (49)	40.5%	6.82	4.75 (57)	6.28 (60)
SUMMER	8.25 (139)	34.8%	7.39	5.25 (179)	5.51 (180)
AROMA	7.79 (51)	35.5%	7.69	5.47 (60)	7.42 (60)
SynthPert	6.86 (53)	40.0%	8.29	6.49 (47)	6.14 (60)
PerturbAgent	9.82 (60)	89.6%	7.35	5.14 (60)	6.48 (60)
CellReasoner	8.27 (42)	48.2%	7.77	6.74 (60)	7.06 (60)
VCWorld	6.13 (41)	18.2%	7.35	4.99 (60)	7.08 (60)
LangPert	7.41 (59)	60.2%	7.81	5.94 (60)	7.37 (60)

Table 9: Three-way matched 0–10 means. n is the common determinate cohort for that row, not the total number of structurally valid records.

Dimension	matched n	Qwen	Llama	DeepSeek
Acceptability	242	9.62	9.79	7.98
Relevance	600	4.70	6.11	7.55
Sufficiency	532	3.95	6.61	5.55
Completeness	600	9.43	7.43	6.52

G HOLISTIC-JUDGE BASELINE AND COGENCY–CORRECTNESS ANALYSIS

G.1 A SINGLE-CALL HOLISTIC-JUDGE BASELINE

Motivation and design. A natural alternative to the decomposed audit is the common LLM-as-a-judge pattern: hand the whole trace to a strong model and ask for one overall trustworthiness score. To isolate the contribution of *decomposition* rather than model choice, the baseline reuses the same three backbones (Qwen3.6-35B-A3B (Qwen Team, 2026), Llama-3.3-70B-Instruct (Meta, 2024), DeepSeek-v4-pro (DeepSeek-AI, 2026)) and the same inputs as the local pipeline: the task, the complete trace, the submitted answer, and the identical frozen evidence dossier (`dossier_for`, 40-gene / 900-character caps). The judge returns a single integer trustworthiness score (0–10) and a $\leq$40-word reason; there is no argument graph, no dimension split, and no checklist. Generation is temperature 0 and the run is resumable per sample (`scripts/run_holistic_judge.py`). Ten overlong prompts (6 Qwen, 4 Llama) were shortened dossier-first under a token-exact budget; the trace itself was never truncated before the dossier was exhausted.

Holistic-judge instruction (single call per trace)

```
You are auditing the trustworthiness of a model-generated reasoning trace
for a biological perturbation-analysis task. Judge the reasoning process,
not whether you independently believe the final biological answer. Consider
jointly: whether the stated premises look reliable and traceable to the
supplied evidence dossier or the trace itself; whether the cited evidence is
relevant to the conclusion; whether the reasoning steps sufficiently support
the final answer; whether important aspects of the task were omitted; and
whether uncertainty is expressed honestly. Return ONLY one JSON object of
the form {"trustworthiness_score": <integer 0-10>, "reason": "<=40 words"}.
0 means the reasoning is entirely untrustworthy (fabricated, irrelevant, or
self-contradictory); 10 means fully trustworthy reasoning for this task. Do
not output Markdown or any other text.
```

Results. All 600 traces were judged by each backbone (parse success 600/600 after a higher-token retry pass for 47 initially empty API completions). Table 10 compares the holistic score with the frozen decomposed pipeline of the *same* backbone on the same traces.

Table 10: Holistic-judge baseline vs. the decomposed pipeline on the same backbone and traces. ρ is the Spearman correlation between the holistic score and each decomposed dimension (0–10). Divergence counts traces where the holistic score and Sufficiency disagree by regime ($\geq$7 vs. $\leq$3). AUC is the rank separation of correct vs. incorrect submitted answers on the 382 labeled traces.

Backbone	mean$\pm$sd	ρ_A	ρ_R	ρ_S	ρ_C	h$\geq$7/S$\leq$3	h$\leq$3/S$\geq$7
Qwen3.6-35B-A3B	2.86 ± 2.25	0.15	0.06	0.06	0.15	44	36
Llama-3.3-70B-Instruct	7.51 ± 1.76	-0.03	0.31	0.32	0.08	46	13
DeepSeek-v4-pro (API)	2.55 ± 1.72	0.33	-0.02	0.20	0.08	0	93

Findings. (1) *The holistic score does not reproduce the structured audit.* Correlations with the decomposed dimensions are weak throughout ($|\rho| \leq 0.33$, most below 0.2), and 13–93 traces per backbone fall into opposite regimes of the holistic score and Sufficiency. (2) *The holistic scale is dominated by backbone temperament.* On identical inputs the three judges center at 2.55, 7.51, and 2.86 — a five-point spread that conflates the trace with the judge's strictness prior. The decomposed pipeline's cross-evaluator differences are instead dimension-specific and interpretable (Table 9). (3) *Holistic scoring sacrifices diagnosability without a validity gain.* Its correctness separation (AUC 0.479–0.531, Table 11) is no better than any decomposed dimension, while a single number cannot localize a failure to premises, evidence use, inference, or coverage. (4) Scope note: the baseline is one prompt form with one call per trace; adding rubrics, intermediate structure, or checklists to it would move it toward the decomposed design itself.

G.2 COGENCY–CORRECTNESS CROSS-ANALYSIS

Setup. 382 of the 600 traces carry a normalized answer-correctness flag (190 correct / 192 incorrect; rbio, SUMMER DE/Direction, AROMA, SynthPert, VCWorld; PerturbAgent, Cell-Reasoner, and LangPert are unlabeled). For each frozen evaluator we report, per dimension, the correct-vs-incorrect mean difference with a trace-level bootstrap interval (2,000 draws, seed 42) and the rank AUC $\mathrm{P}(\mathrm{score}_{\mathrm{correct}} > \mathrm{score}_{\mathrm{incorrect}})$, pooled and averaged within systems (`scripts/analyze_cogency_correctness.py`).

Table 11: Correctness separation of each decomposed dimension and of the holistic-judge score, on the 382 labeled traces. Δ is mean(correct)−mean(incorrect) with a 95% bootstrap interval; AUC is the rank separation (0.5 = none); *strat.* averages AUC within each system to control for cross-system difficulty and scale differences.

Evaluator	Measure	Δ [95% CI]	AUC	strat. AUC
Qwen3.6-35B-A3B	Acceptability	-0.19 [-0.58, $+0.16$]	0.488	0.485
	Relevance	-0.15 [-0.47, $+0.16$]	0.467	0.457
	Sufficiency	-0.23 [-0.57, $+0.11$]	0.464	0.444
	Completeness	-0.08 [-0.28, $+0.10$]	0.470	0.479
	holistic judge	—	0.502	—
Llama-3.3-70B-Instruct	Acceptability	$+0.22$ [-0.22, $+0.69$]	0.513	0.506
	Relevance	-0.01 [-0.24, $+0.22$]	0.503	0.494
	Sufficiency	-0.07 [-0.29, $+0.15$]	0.470	0.451
	Completeness	$+0.00$ [-0.17, $+0.17$]	0.506	0.496
	holistic judge	—	0.531	—
DeepSeek-v4-pro (API)	Acceptability	-0.71 [-1.27, -0.17]	0.438	0.415
	Relevance	-0.06 [-0.28, $+0.16$]	0.487	0.482
	Sufficiency	$+0.06$ [-0.25, $+0.39$]	0.516	0.495
	Completeness	$+0.11$ [-0.20, $+0.42$]	0.524	0.528
	holistic judge	—	0.479	—

Result: answer correctness and process cogency are near-orthogonal in this benchmark. No dimension, under any of the three evaluators, separates correct from incorrect submitted answers (pooled AUC 0.44–0.52; all but one bootstrap interval include zero; the single exception is DeepSeek Acceptability, which is *negatively* associated with correctness). The holistic judge does no better (0.479–0.531). We caution against reading this as a defect of the audit: the labeled systems include pipelines whose answers are often right while their rationales are broken (gene-name priors; Section H.4) and pipelines that abstain honestly under missing evidence (Section H.3). Outcome labels therefore cannot proxy reasoning fidelity in either direction — which is precisely the gap a process-level audit is meant to fill. Whether the decomposed scores track *expert* process judgments is the separate question addressed by the expert-calibration protocol (Section 4.2).

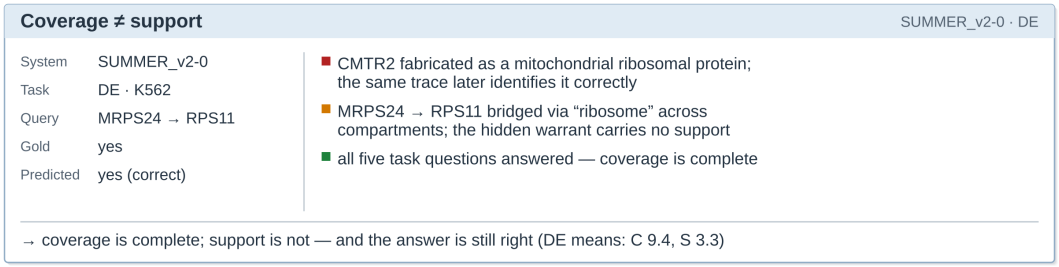

Figure 4: SUMMER_v2-0: all five task questions receive an answer, but the entity attribution and cross-compartment bridge require scrutiny. The submitted DE answer is correct; coverage does not establish support.

H Representative Cases and Traceable Graph Excerpts

These cases are selected for qualitative inspection, not to estimate failure rates or certify the automatic evaluator. The tags describe the researcher's case reading: **FAIL** a concrete inconsistency, **WARN** an inferential concern, and **PASS** an observable coverage or provenance feature. They are not expert gold annotations. All cases come from the canonical benchmark; scores below use the final frozen Qwen run, not historical holistic estimates.

H.1 Case A: correct answer, problematic support (SUMMER_v2-0)

Task. MRPS24 → RPS11 (K562, DE). Gold and prediction: *yes*.

Observable concern. The trace describes CMTR2 as an mt-SSU component, then later describes its role in mRNA cap methylation, without reconciling the two descriptions. Its conclusion transfers effects through the broad category *ribosome*, treating RPS11 as involved in mitochondrial translation. **WARN** The relevant audit questions are the correctness of these premise attributions and the warrant for transferring effects between the mitochondrial and cytosolic systems. **PASS** Answering every template question does not settle either question.

Graph excerpt. In the frozen Qwen artifact, `c1` is the Claim that MRPS24 perturbation disrupts mitochondrial ribosome function; `c2` links that disruption to expression of genes such as RPS11; `c3` is the final MRPS24–RPS11 DE Claim. Path `s1` traverses `l1`, `l2`, and `l4`. In ordinary language: a proposed ribosomal disruption is transferred to the target expression response. The other two Paths use the asserted CMTR2/MRPS35 observations (`l3`) and upstream-interaction statements (`l5`). These are reconstructions of the trace's propositions, not newly verified biological facts.

The five Link scores are $6, 5, 4, 5, 3$. Accordingly, $S(\texttt{s1}) = (6 + 5 + 5)/3 = 5.33$, while the other Paths score 4 and 3; the final-Claim summary is $(5.33 + 4 + 3)/3 = 4.11$. The case-level $C = 9.17$ differs from the DE-stratum average. This example makes the aggregation inspectable without treating the automatic scores as gold.

H.2 Case B: signed hypothesis without instance confirmation

Task. SynthPert-0: PIK3C3 → HSPD1 in RPE1. Prediction: upregulated; gold: downregulated.

Observable concern. The trace proposes a signed chain: PIK3C3 knockdown impairs autophagy/mTORC1 signaling, increases stress, activates an HSF1 response, and increases HSPD1. It is therefore *not* an unsigned mechanism. **WARN** The unestablished bridge is whether this proposed compensation operates and dominates in the specified cellular context; no instance-specific experiment is cited in the presented trace. **FAIL** The answer disagrees with the gold direction, but that alone does not identify which mechanistic proposition is false.

Evaluator limitation. Qwen assigns $R = 8.33$ and $S = 8.33$ here. This is a calibration challenge for the auditor, not evidence that PROVER reliably detects every unsupported mechanistic hypothesis.

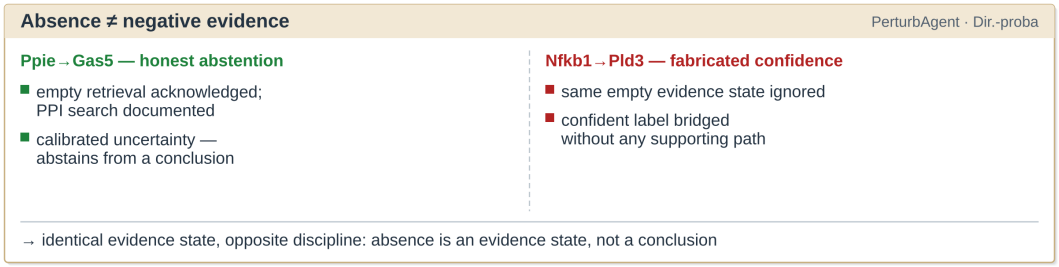

Figure 5: PerturbAgent-0: explicit retrieval provenance (left) does not establish the no-effect conclusion (right). The trace submits a no-effect decision, not an abstention.

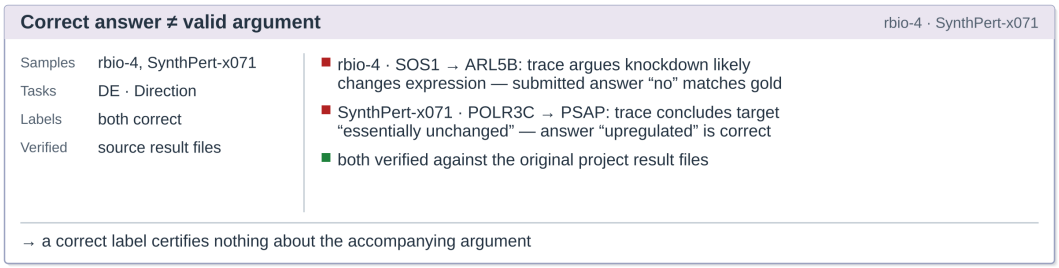

Figure 6: Two submitted answers are labeled correct while the accompanying traces argue a different outcome, as checked against source result files.

H.3 Case C: documented absence, unsupported no-effect inference

Task and behavior. PerturbAgent-0 predicts the effect of Ppie knockdown on Gas5. The logs report no training examples for either gene, provide functional annotations, and list Ppie interaction partners. The final thought states that the most supported conclusion is no significant effect; the tool submits option C and the saved answer is `up=0.0, down=0.0`. These fields jointly identify a no-effect decision, not calibrated abstention.

Traceable decomposition. In the frozen Qwen graph, `p1` records the failed training-example lookup and `c1` is the no-effect Claim. Link `l1` connects `p1`, through hidden warrant `w1`, to `c1`; Link `l2` connects the annotation and interaction premises through `w2` to the same Claim. Each Link forms a one-step support Path. In plain language, two forms of not finding a connection are used to infer that no effect occurs.

Scores and interpretation. Both Links receive $S = 3/10$, giving final-Claim $S = (3 + 3)/2 = 3/10$; the other summaries are $A = 10$, $R = 3.2$, and $C = 10$. High A reflects assessable retrieval records, not confirmation of a biological null effect. The earlier draft's holistic estimates and non-canonical Nfkb1–Pld3 comparator are not used in these results.

H.4 Trace–answer inconsistency with correct labels

rbio-4 (SOS1 → ARL5B, K562). The trace states that SOS1 knockdown is likely to change ARL5B expression, while the submitted answer and gold label are both *no*. The source is `k562_full_rbio1-TF_results.csv`, zero-based record 107518.

SynthPert-x071 (POLR3C → PSAP, RPE1). The task states that the target is differentially expressed and asks for direction; the trace concludes that PSAP mRNA remains essentially unchanged, while the submitted *upregulated* answer is marked correct by the source file's `correct_solution`. The source is line 388 of `rpe1_official_cot_direction_10pct/eval.jsonl`.

Interpretation. Task condition, textual conclusion, and submitted answer are independently checkable objects. A correct label does not establish support from the accompanying analysis.

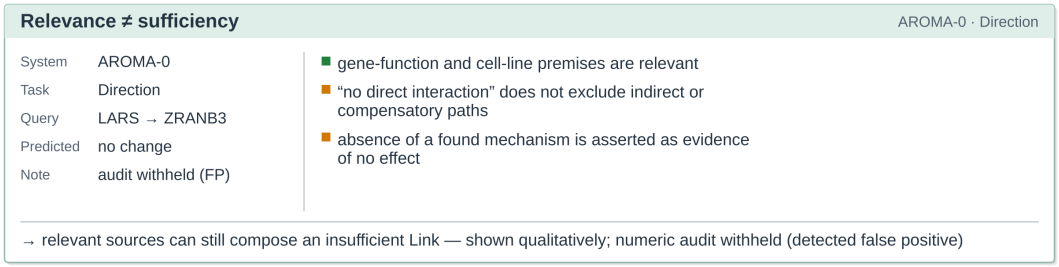

Figure 7: AROMA-0: function and cell-line descriptions provide background, but the no-effect conclusion additionally requires an exclusion warrant. The current frozen Qwen run includes this case: $R = 3.5$, $S = 3.0$, $C = 10$, with Acceptability undetermined.

H.5 Target granularity and task scope

SUMMER_v2-1 (PPP1R8 → CSTB). All five questions are answered, but the trace uses functional dissimilarity and missing retrieved evidence to infer no change; the canonical label is DE. This illustrates the distinction between topic coverage and the warrant needed to infer a negative effect claim.

CellReasoner-x132. The trace discusses T-cell markers and an NK alternative, then concludes T-cell. The question requests a cell-type annotation without specifying Naive CD4 T; that narrower label appears only in the held-out reference. Sufficiency must target the actual T-cell Claim, while any reference-label granularity mismatch is a separate outcome issue. Mentioning an alternative does not certify the correctness of its marker assertions.

SUMMER_v2-66 (GSE). The trace gives an overview, possible upstream processes, and the name *Ribosomal Protein Component Gene Set*. This responds to the supplied description-and-naming task; it is not a statistical enrichment result. Proposed upstream influences remain hypotheses. The frozen checklist's legacy enrichment wording is discussed in Appendix D.4.

LangPert-10. Candidates are justified using claimed interactions, shared mitochondrial translation, and broader pathway proximity. These are different transfer warrants, not interchangeable evidence of perturbation-effect similarity. Their biological assertions still require verification; a candidate called a direct match in the trace is not thereby established as one.

Artifact scope. Graphs, raw trace spans, and dimension objects for these cases are stored in `data/benchmark/runs/local_campaign_600_converged_v1/qwen36_full/results/`. The earlier AROMA-0 exclusion is obsolete; all 60 AROMA records enter the primary cohort. Unknown dimensions remain excluded from that dimension's mean and counted in its denominator.

I AROMA INPUT-CONTEXT SENSITIVITY AUDIT

Protocol. 100 AROMA evaluation traces (HepG2 / Jurkat / K562 / RPE1, 25 each, evenly spaced; gold distribution 90/2/8 over no-change/down/up) are re-decoded under greedy decoding (max 512 new tokens) with ten input variants that remove, corrupt, or replace individual evidence channels of the prompt and the fused embeddings (Table 12). The majority-class baseline (always predicting no-change) is 90.0%. We emphasize the *flip rate* relative to the intact control — how often the prediction changes when an evidence channel is ablated — rather than treating small accuracy differences as established gains.

Table 12: AROMA input-ablation results. *flip* is the fraction of predictions changed vs. the intact control (A). *comp.* counts outputs containing `compensatory`; *co-occ.* counts outputs containing both `stabilize` and `upregulat`, using the original script's literal substring checks. These are lexical diagnostics, not validated template or contradiction annotations. I/J strip context paragraphs; I retains normal embeddings and J uses mean embeddings.

Variant	Change vs. control	acc	flip	comp.	co-occ.
A	intact prompt + embeddings	94%	—	34%	0%
B	remove pathway context	97%	3%	41%	0%
C	remove target-gene information	97%	3%	45%	0%
D	C + B	98%	4%	54%	1%
E	swap cell line for a mismatched one	95%	5%	29%	0%
F	remove cell-line description	96%	4%	22%	0%
G	replace all fused embeddings with mean vectors	93%	1%	39%	0%
H	G + D	97%	5%	54%	1%
I	remove all context paragraphs (names only)	98%	5%	67%	2%
J	I + mean embeddings	97%	3%	60%	4%

Findings. Each tested input modification changes only 1–5 of the 100 predictions. Variants I and J retain accuracies of 98% and 97%, compared with 94% in the intact control and a 90% majority baseline. The frequency of `compensatory` changes from 34% in A to 67% in I, while the two-substring co-occurrence reaches 4% in J. These observations motivate inspection of how input context affects prediction and wording separately. They do not establish gene-name memorization, semantic contradictions, or a monotonic relation between context volume and template reliance.

Scope and caveats. There are only two downregulated and eight upregulated gold cases, and this is a separate 100-case ablation slice rather than the 60-case PROVER AROMA sample. The variants intervene on supplied context and embeddings, not on the generated reasoning text itself. Low prediction sensitivity does not prove that context is unused, identify the internal decision mechanism, or establish that explanations are causally irrelevant. We therefore report an observable input-sensitivity result, not a causal-faithfulness verdict. The source artifacts are the AROMA repository's `aroma/ablation_analyze.py`, `data/Knowledge_Graph/ablation_study_results.jsonl`, and `data/Knowledge_Graph/ablation_study2_results.jsonl`.

J PRELIMINARY EXPERT REVIEW

Returned sample. Approximately 40 traces received complete expert reviews. Requiring all five ratings, a failure-locus selection, and epistemic status defines the analysis sample, all from one expert. Three incomplete forms from another account were excluded; autosave revisions are not additional reviews. There are no overlapping complete reviews, so inter-rater agreement cannot be estimated. The 120-case packet is the candidate pool, not the returned sample size. The returned subset covers AROMA (4), CellReasoner (5), LangPert (4), PerturbAgent (11), SUMMER (8), rbio (6), and SynthPert (6). No complete VCWorld reviews were available. This availability subset cannot estimate population-level failure prevalence.

Protocol. The interface presents the question, source trace, task checklist, and reference/background context, with an optional Chinese reading aid. Source-system metadata, gold outcomes, correctness flags, and automated judgments are withheld. English remains the reference when translation is ambiguous. The original packet was stratified from the 300-case sampling partition, with 15 candidates per system and five per SUMMER task. Separate packet cleaning and replacement requires explicit version matching before comparison with automatic audits.

Five ordinal fields use 0–4 anchors and a separate unable-to-judge option: overall support, evidence Relevance, inferential-connection quality, evidence sufficiency, and Completeness. Scores increase with support or coverage. Evidence sufficiency concerns missing evidence or conditions; inference quality is the closest questionnaire counterpart of PROVER's Link-based Sufficiency, but neither is an identical measurement. Failure loci are multi-select (including no major issue); epistemic status is supported, underdetermined, or contradicted. Written reasons are optional.

Descriptive feedback. Thirty-one reviews select supported, eight underdetermined, and five contradicted. Twenty-eight select no major issue. Concern labels include overclaim (6), contradiction (5), Relevance (3), Completeness (2), and Sufficiency (1); multiple selections are permitted. No review selects Acceptability as an issue. All five ratings are numeric in all complete forms (Table 13). No per-item written reasons or separate free observations were submitted. We can report concern labels, but cannot claim experts verified particular automatic explanations or localized concerns to specific graph edges.

Table 13: Expert pilot ratings (0–4; 44 complete reviews). No rating is 0, 1, or unable-to-judge. Means descriptively summarize ordinal responses.

Field	Rating 2	Rating 3	Rating 4	Mean
Overall support	5	26	13	3.18
Evidence Relevance	6	26	12	3.14
Inference quality	5	26	13	3.18
Evidence sufficiency	5	27	12	3.16
Completeness	5	27	12	3.16

Exact-text pairing and exploratory associations. Thirty-three reviews match the benchmark by source file/index, question, and exact trace SHA-256. Eleven have different trace text (five CellReasoner and six rbio cases) and are excluded from paired comparisons, not the descriptive pilot. Table 14 pairs Relevance with R, inference quality with S, and Completeness with C. Missing automatic scores are excluded per dimension; scales are not equated numerically. Associations are mixed, including negative values: the pilot does not establish consistent expert alignment. These are exploratory estimates from a small, restricted-range, heterogeneous subset, without confirmatory tests.

Scope and reproducibility. Feedback is trace-level, not gold argument graphs, diagnostic precision/recall, or evidence of downstream utility. An overclaim label does not establish that the automatic qualifier diagnostic found the same issue. There is no direct numeric expert Acceptability field. The legacy GSE checklist wording also applies to the interface and warrants refinement (Appendix D.4). Anonymized records and statistics are frozen in `data/benchmark/analysis/expert_pilot_snapshot_v2.json` and `data/benchmark/analysis/`

Table 14: Exploratory expert–audit Spearman correlations; parentheses give exact-text paired counts. Comparisons are trace-level, not local-graph judgments.

Backbone	Relevance	Inference quality vs. S	Completeness
Qwen	-0.38 (33)	-0.20 (32)	0.27 (33)
Llama	-0.20 (33)	-0.34 (30)	-0.25 (33)
DeepSeek	-0.04 (33)	0.08 (30)	0.50 (33)

`expert_pilot_summary_v2.json`; `scripts/summarize_expert_pilot.py` with `-revision v2` reproduces the summary without changing the live database or calling a model.