CONTRASTIVE Discovery: Open-Ended Scientific Discovery over Competing Explanations

Ziang Liu* **James J. Kim*** **Yijia Dai** **Jennifer J. Sun**

Department of Computer Science
Cornell University
codiscovery.ai/contrastive-discovery

Abstract

LLM agents are increasingly used for open-ended, data-driven scientific discovery, yet most of them evaluate hypotheses *pointwise*, testing an empirical claim h only against its complement $\neg h$. When evidence rejects h, the agent learns that one explanation is unlikely but not which alternative the evidence supports. To recover that alternative, we introduce **CONTRASTIVE Discovery**, which replaces the single hypothesis with a *contrast set*, a scientific question paired with mutually exclusive and collectively exhaustive competing explanations. Each experiment is scored by how its evidence redistributes belief across the set, and the question is resolved when belief concentrates on one explanation. Because contrast sets change only the unit an agent proposes and scores, an existing agent can use them to re-evaluate evidence it has already collected or to guide its search. Across six domains, 89% of the *pointwise* candidates with the largest belief shifts are rejections. On the same evidence, contrast sets resolve 73% of these candidates to a specific alternative. When contrast sets guide search, the agent resolves 67% more questions than *pointwise* surprisal under the same experiment budget. In a blinded study, 38 human experts chose the same explanation on 78% of resolved questions and on 82% of questions where the *pointwise* agent rejected its hypothesis and stopped. Together, these results support explicit competing explanations as a way for discovery agents to resolve more scientific questions from the same evidence and validation budget.

1 Introduction

The shift toward autonomous scientific discovery has been accelerated by agents that generate hypotheses, design experiments, collect evidence, and iterate on candidate findings [1, 2, 3, 4]. Recent systems scale this loop to evaluate hundreds or thousands of discoveries across scientific domains [5, 6, 7]. At this scale, measuring progress becomes a problem. In closed tasks, progress is measured against fixed objectives, answer keys, or benchmarks [8, 9], but open-ended discovery has no oracle specifying which discovery should come next. Without an oracle, progress depends not only on how each candidate is scored, but on what *unit* of knowledge the agent treats as learned.

In most current discovery systems, that unit is a single hypothesis. The agent proposes an empirical claim h, tests it, scores the result, and then keeps, revises, or discards h [6, 10, 11]. The score may be an LLM judgment of novelty or feasibility [12, 13, 14, 15], or a belief-shift objective such as Bayesian surprise [5, 16], which treats discovery as evidence-induced belief change rather than a subjective judgment of interestingness. Yet the representational unit of progress stays the same: evidence updates belief in one focal hypothesis h against its implicit complement $\neg h$. We call this

*Equal contribution. Correspondence to {zl873,jjk297}@cornell.edu.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

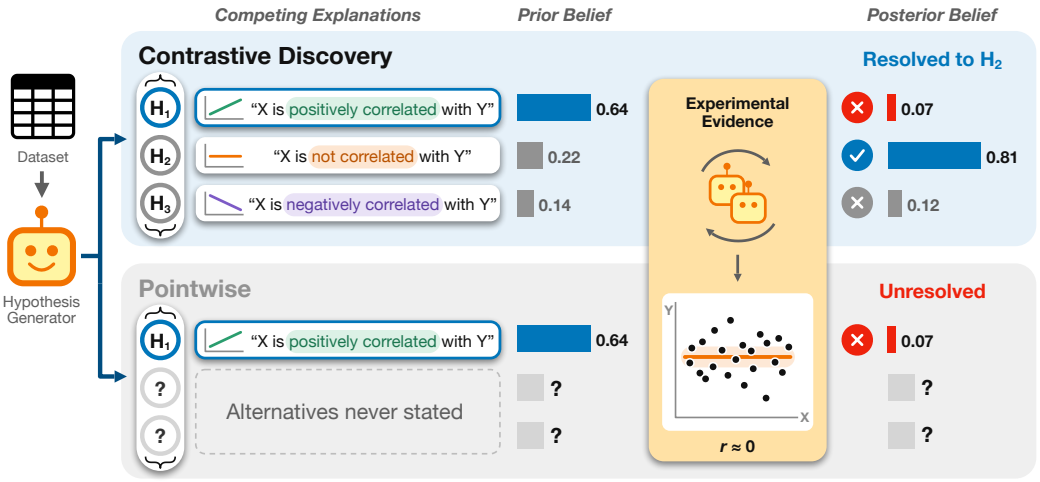


Figure 1: **A contrast set turns a rejection into a resolution.** A hypothesis generator proposes a contrast set of three competing explanations (H_1–H_3) for how X relates to Y, and coder, analyst, reviewer, and reviser agents run an experiment that produces evidence. Bars show elicited belief in each explanation before and after the same evidence. CONTRASTIVE Discovery updates belief over all three explanations and moves the lead from H_1 (0.64) to H_2 (0.81), so the question resolves to H_2 and H_1 is ruled out. *Pointwise* discovery updates belief only in H_1 (0.64 to 0.07), so it records a rejection but leaves the question unresolved. See §3 and §4 for details.

pointwise discovery. This formulation is shared across agentic discovery systems that otherwise differ in how they propose hypotheses, gather evidence, and score discoveries.

The *pointwise* representation is epistemically limiting because scientific evidence often discriminates among plausible alternatives rather than supporting a single hypothesis in isolation. This limitation is clearest when evidence rejects a proposed hypothesis h: the agent learns that the focal hypothesis is unlikely, but not which alternative is better supported. For example, evidence against "self-employed applicants with poor credit face higher mortgage denial rates" does not say whether their denial rates are lower or no different. Under a fixed evidence-generation budget, this is information-inefficient: the agent may need additional hypothesis proposals and evaluations to resolve a scientific question that the original evidence already helps answer.

We argue that open-ended discovery requires a larger unit of progress: **a scientific question together with the plausible explanations that could answer it**. This follows contrastive accounts of scientific explanation, in which evidence is informative because it discriminates among alternatives [17, 18]. We introduce **CONTRASTIVE Discovery**, which represents each discovery step as a *contrast set* (Eq. 1, Figure 1): a set of mutually exclusive and collectively exhaustive (MECE) explanations answering one question. The agent elicits prior and posterior beliefs over the set and scores a candidate discovery by how much its evidence redistributes belief. A question is *resolved* when posterior belief concentrates on one explanation, so a single experiment can both rule out alternatives and identify the best-supported one. Otherwise, the evidence remains informative but the question is *unresolved*.

Contrast sets change the unit an agent proposes and scores, not its architecture, so they can be layered onto an existing agent in two ways. As an *evaluation lens*, they re-evaluate evidence the agent has already collected, without modifying the agent. To *guide search*, they replace the agent's hypothesis and score, so the search is rewarded for evidence that discriminates among competing explanations. We evaluate CONTRASTIVE Discovery across six domains spanning materials science, astronomy, mortgage discrimination, online discourse, software engineering practices, and social behavior. Under a shared scaffold and evidence-generation budget, *pointwise* baselines are guided by surprisal, novelty, or feasibility, while CONTRASTIVE Discovery is guided by belief shift over contrast sets. We find:

- ***Pointwise* rejections conceal resolvable alternatives.** Across three elicitor models, a majority of *pointwise* proposals receive a "maybe true" prior, and 89% of the highest-surprisal *pointwise* candidates are rejections. Used as an evaluation lens on the same evidence, contrast sets resolve 73% of these candidates to a specific alternative without running new experiments (§5.1).

- **CONTRASTIVE Discovery turns rejections into resolved questions.** Guiding search with contrast sets resolves 67% more questions than *pointwise* surprisal under the same experiment budget. Nearly all the gain comes from corrective resolutions, where evidence shifts support away from the explanation favored beforehand; *pointwise* discovery records these cases only as rejections (§5.2).
- **The gain comes from explicit alternatives and holds up under validation.** Ablations trace the gain primarily to evaluating evidence against the full contrast set (§5.3). In a blinded study, 38 human experts chose the same explanation on 78% of resolved questions and on 82% where the *pointwise* agent rejected its hypothesis and stopped (§5.4). The gain holds when each alternative is scored independently with the *pointwise* scorer and at matched token and dollar cost (§5.5).

2 Open-Ended, Data-Driven Discovery

Discovery loop. We study open-ended, data-driven scientific discovery, where an agent uses a dataset D to propose hypotheses $h \in \mathcal{H}$, evaluates them using evidence computed from the data, and scores them, repeating under a budget of N steps. The setting is open-ended because no predefined reward function determines the value of a discovery, which depends instead on the scientific question, the available evidence, and prior knowledge. This formulation captures recent LLM-based discovery systems [5].

***Pointwise* discovery.** We call a discovery method *pointwise* when, at step t, the agent proposes a single hypothesis h_t and independently assesses it by scoring only that proposed candidate and its evidence. Many recent LLM-based discovery approaches are *pointwise*, assessing each hypothesis through novelty and feasibility [12, 13, 14], evidential support [6, 7], or Bayesian surprise [5]. In our notation (§3.5), *pointwise* discovery is the binary case of CONTRASTIVE Discovery, where the only alternative to h_t is its negation.

The rejection case. In *pointwise* discovery, confirmation and rejection have different epistemic consequences. Confirmation identifies a supported explanation, but rejection only rules against the proposed one. The evidence may bear on alternatives, yet the *pointwise* score exposes only its effect on h_t. Resolving among competing explanations then takes a separate step for each alternative, each drawn from the finite budget of N steps, even when one experiment's evidence bears on all of them.

3 CONTRASTIVE Discovery

We reframe a discovery step from validating a single candidate hypothesis to resolving a scientific question. We keep the evidence-induced belief-shift principle of surprisal-based discovery [5, 16] but change the unit over which belief is represented, from one hypothesis to a structured set of plausible explanations for the same question. The alternatives thereby become part of the unit, so they determine what the evidence is compared against, how belief is updated, and what outcome the step can produce.

3.1 Constructing contrast sets

At each discovery step t, the agent poses a local scientific question and constructs a contrast set

$$\mathcal{C}_t = \{h_{t,1}, \ldots, h_{t,M}\}, \tag{1}$$

where each $h_{t,m}$ is a plausible explanation for the same question. We require these explanations to be locally mutually exclusive and collectively exhaustive (MECE): within the scope of the question, no two of them can hold at once, and together they cover the main outcomes the evidence is meant to distinguish. For example, for the scientific question "*Given a poor credit score, how is self-employment associated with mortgage denial rates?*", the three explanations in the contrast set are that self-employed applicants face higher, no different, or lower denial rates than other applicants (further examples in Appendix U).

3.2 Eliciting beliefs over contrast sets

Given a contrast set $\mathcal{C}_t$, we next define the belief state used to score evidence-induced change. Before observing evidence, we elicit $p_t^{\text{prior}} \in \Delta^{M-1}$ where Δ^{M-1} is the probability simplex over the M

explanations and $p^{\text{prior}}_{t,m}$ is the subjective belief assigned to explanation $h_{t,m}$. After the agent runs its experiment and observes evidence e_t, we elicit $p^{\text{post}}_t \in \Delta^{M-1}$. We treat these elicited distributions as the prior and posterior beliefs over the contrast set.

Computing the posterior by Bayes' rule would require a hand-specified likelihood model $p(e_t \mid h_{t,m})$ for every analysis. We therefore obtain each distribution directly through a constant-sum allocation task: a belief elicitor, an LLM separate from the agent (§4), assigns 100 allocation units across the explanations in $\mathcal{C}_t$, first without and then with access to e_t. To reduce stochastic variance, we sample five allocations and average the normalized valid responses (see Appendices T.2 and B). We also randomize the order of explanations during elicitation to mitigate positional bias in LLM-based evaluation (see Appendix D).

3.3 Scoring contrastive belief shifts

Because the unit of discovery is a belief distribution over competing explanations, the score should measure how evidence changes that distribution. Given prior and posterior beliefs p^{prior}_t and p^{post}_t, we compute the Kullback–Leibler divergence (D_{KL}) of p^{post}_t with respect to p^{prior}_t. This direction is the Bayesian surprise of Itti and Baldi [16], the realized information gain from observing e_t. The divergence tends to grow with the size of the contrast set, reaching $\log_2 M$ bits when a uniform prior collapses onto one explanation, so we divide by $\log_2 M$ to compare sets of different sizes. We call the result the *contrastive belief shift*, an entropy-scaled Bayesian surprise:

$$\kappa_t = \frac{D_{\text{KL}}(p^{\text{post}}_t \,\|\, p^{\text{prior}}_t)}{\log_2 M}.$$

We use κ_t as the CONTRASTIVE Discovery score, and larger values mean that the evidence moved more belief across the competing explanations. Scaling does not bound κ_t by one, because evidence can move belief onto an explanation the prior considered unlikely. When we use κ_t as the tree-search reward (§4), we therefore clip it at one, $r_t = \min(\kappa_t, 1)$, and all reported information-gain analyses use the unclipped κ_t (see Appendix T.2).

3.4 Discovery outcomes

The contrastive belief shift κ_t measures the magnitude of belief shift, but not the type of discovery produced by that shift. Since the prior and posterior are defined over the same competing explanations, the update can also be characterized by whether evidence confirms the prior-leading explanation, shifts support to another explanation, rules out alternatives, or leaves the question unresolved. Let

$$m^{\star,\text{prior}}_t = \arg\max_m p^{\text{prior}}_{t,m}, \quad m^{\star,\text{post}}_t = \arg\max_m p^{\text{post}}_{t,m}$$

denote the indices of the prior- and posterior-leading explanations. We say the question is *resolved* when the posterior assigns sufficient support to one explanation,

$$\max_m p^{\text{post}}_{t,m} \geq \tau_{\text{support}},$$

where τ_{support} is the minimum posterior support required to identify a leading explanation (we use $\tau_{\text{support}} = 0.80$ and vary it in §5.5). A resolved update is *confirmative* if $m^{\star,\text{post}}_t = m^{\star,\text{prior}}_t$ and *corrective* if $m^{\star,\text{post}}_t \neq m^{\star,\text{prior}}_t$, that is, when evidence moves support from the prior-leading explanation to another one. If no explanation reaches τ_{support}, the outcome is *unresolved*. A generated contrast set may omit the explanation the evidence supports. As a safeguard, we optionally add an *abstention* alternative, "none of the listed alternatives matches the evidence." If its belief reaches τ_{support}, we flag a possible omission and report no resolution. The main experiments do not include this safeguard, and §5.4 tests it.

We separately track explanations that are ruled out: an explanation $h_{t,m}$ is *ruled out* when $p^{\text{post}}_{t,m} < \tau_{\text{reject}}$, where τ_{reject} denotes a low posterior-support threshold (we use $\tau_{\text{reject}} = 0.10$). Ruling out is tracked apart from the three outcomes above, because a single contrastive update can identify a leading explanation and also rule out competing explanations. For the *pointwise* setting, where a single hypothesis has scalar prior and posterior beliefs, we classify the update analogously: a hypothesis is *confirmed* if its posterior reaches γ_{support}, *rejected* if it is at most γ_{reject}, and *ambiguous* otherwise

(we use $\gamma_{\text{support}} = 0.625$ and $\gamma_{\text{reject}} = 0.375$, with calibration details in Appendix E). The taxonomy thus separates *how much* support for a hypothesis rises or falls, which *pointwise* discovery captures, from *which* explanation the evidence favors, which requires explicit alternatives.

3.5 *Pointwise* discovery as the binary case

Pointwise belief-shift methods score evidence by how much it changes belief in a single hypothesis h_t. In our notation, this is the binary contrast set $\mathcal{C}_t^{\text{point}} = \{h_t, \neg h_t\}$ with belief state $p_t = (p_{t,1}, p_{t,2}) \in \Delta^1$, where $p_{t,1}$ is belief in h_t and $p_{t,2} = 1 - p_{t,1}$ is belief in its complement. In this representation, *pointwise* discovery is the $M = 2$ instance of CONTRASTIVE Discovery, where the only alternative $\neg h_t$ contains all ways the focal hypothesis could be false. Thus, when evidence weakens h_t, *pointwise* discovery can record negative evidence about the claim but cannot identify which alternative explanation receives that support. CONTRASTIVE Discovery replaces this complement with an explicit local partition (Eq. 1), so the same belief shift is distributed over named competing explanations.

For *pointwise* belief elicitation we follow an established categorical procedure [5]: the belief elicitor rates h_t on a five-level scale (definitely false, maybe false, uncertain, maybe true, definitely true), and a "cannot comment" response counts as missing. We draw 10 ratings each for the prior and posterior, map them to numeric scores, fit a Beta distribution to the prior ratings, and update it with the posterior ratings. *Pointwise* surprisal is the normalized absolute difference in the two Beta means, and we use it as the *pointwise* discovery score (details in Appendix T.1; design comparison in Appendix O).

4 Experimental Setup

We evaluate CONTRASTIVE Discovery in an open-ended, data-driven setting: given a tabular dataset D and metadata, we give the agent a fixed budget of $N = 200$ experimental validations. Each validation consists of exploring a scientific question by generating a contrast set, eliciting prior belief over the set, designing and executing an experiment to evaluate it, checking the experiment with a reviewer agent, and eliciting posterior belief. To choose which question to explore next, we use Monte Carlo tree search (MCTS) [19], following prior work on agentic scientific discovery and hypothesis search [2, 5, 13, 20] and on tree search for LLM reasoning and machine learning engineering [21, 22, 23]. Each node is one experimental validation, and its children are follow-up contrast sets that the agent proposes from the node's result. Search is guided by the upper-confidence bound on trees (UCT) [24], with the clipped contrastive belief shift $r_t = \min(\kappa_t, 1)$ as the node reward (reward in Appendix T.2; tree structure in Appendix T.3).

Our primary evaluation metric is the number of resolved scientific questions among the K candidates that a method's own score ranks highest, and the *resolution rate* divides this number by K. Candidates are ranked within each run, and counts are averaged over seeds within each dataset and then over the six datasets. A question is resolved when the experiment concentrates posterior belief strongly enough to identify a leading explanation (§3.4), and for *pointwise* baselines a confirmed hypothesis counts as resolved. Among resolved questions, we further distinguish *confirmative* and *corrective* resolutions, corresponding to belief confirmation and belief revision, respectively. Appendix F repeats the resolved-question count after semantic deduplication.

Discovery agent scaffold. Our base agent is the AutoDiscovery scaffold [5], using `gpt-4o` for the discovery agent backbone and a separate `gpt-4o`-based belief elicitor (agent details in Appendix T.3; prompts in Appendix P). In CONTRASTIVE Discovery, the elicitor assigns a belief distribution over all explanations in a contrast set (§3.2); in the *pointwise* setting, it assigns scalar beliefs for a single hypothesis at a time to compute *pointwise* surprisal (see Appendix T.1).

Baselines. We compare against *pointwise* baselines under the same $N = 200$ evidence-generation budget. In this setting, each experimental validation proposes a single hypothesis and guides MCTS using *pointwise* surprisal. With the number of experimental validations fixed, the comparison asks whether CONTRASTIVE Discovery resolves more questions per validation by discriminating among alternatives. To provide additional *pointwise* comparisons, we guide MCTS using LLM-judge scores for hypothesis novelty and feasibility, following prior work on generating and evaluating research ideas [12, 13, 14, 25] and scientific hypotheses [6, 15, 26] (judge prompts in Appendix P.3).

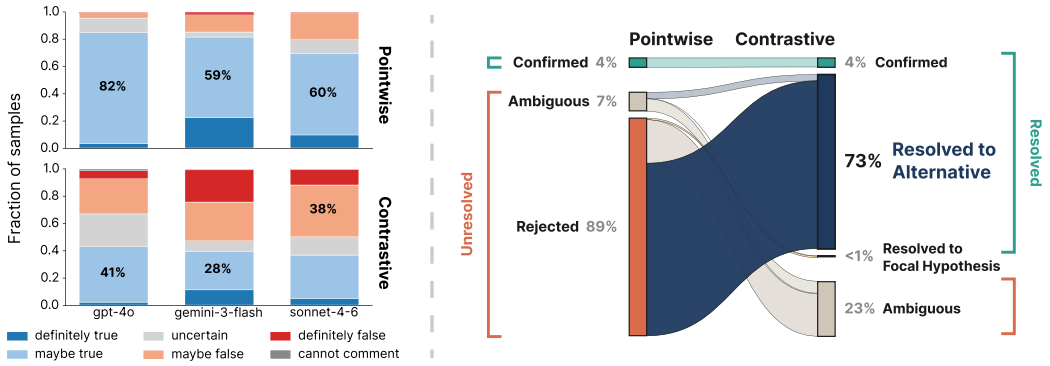


Figure 2: ***Pointwise* discovery masks resolvable alternatives behind rejection.** ***Left:*** *pointwise* proposals show prior collapse, with most mass in the "maybe true" bin under each of three belief elicitors (columns), while explanations in contrast sets span a broader range of prior beliefs. ***Right:*** among the top-50 hypotheses ranked by *pointwise* surprisal, 89% are rejected under *pointwise* evaluation, but re-evaluating them with a post-hoc CONTRASTIVE companion on the same evidence resolves 73% of all 50 to a specific alternative. On the CONTRASTIVE side, *confirmed* marks resolutions to the explanation *pointwise* confirmed, and *ambiguous* marks unresolved sets. Per-dataset breakdowns over each run's 200 experimental validations are in Figure 10.

Datasets. We evaluate on six datasets spanning scientific and social domains: Materials Project [27], SDSS Galaxy [28], BLADE Mortgage [29], BLADE Conversation [29], DiscoveryBench Requirements Engineering [30], and DiscoveryBench NLS Raw [30]. These datasets vary in feature spaces, sample sizes, and domain structure. Dataset statistics, preprocessing, and filtering criteria are reported in Appendix G. We report results on three independent runs for the first four datasets and one run on the two highest-cost datasets.

5 Results

We evaluate whether CONTRASTIVE Discovery increases the epistemic yield of discovery under a fixed evidence-generation budget by tracking confirmed and rejected *pointwise* hypotheses, resolved and unresolved contrast sets, and ruled-out alternatives. We first show that *pointwise* rejections conceal alternatives that the same evidence can already resolve (§5.1). We then show that guiding search with contrast sets turns these rejections into resolved questions (§5.2). Finally, we validate the gain: ablations trace it to evaluating evidence against explicit alternatives (§5.3), resolutions agree with blinded experts and other independent references (§5.4), and the gain holds across scoring rules, matched token and dollar cost, and resolution thresholds (§5.5).

5.1 *Pointwise* rejections conceal resolvable alternatives

We first ask what high-scoring *pointwise* surprisal candidates actually represent. Figure 2 (left) shows that *pointwise* hypothesis proposals concentrate heavily in the "maybe true" prior bin across three belief elicitors (at least 59% for each). This concentration is strongest when the same model, `gpt-4o`, both proposes and elicits (82%), but remains pronounced under cross-model elicitation. In contrast, explanations drawn from contrast sets produce a much broader prior distribution, with the largest bin containing only 41% of mass (`gpt-4o`). This suggests that *pointwise* proposals cluster around hypotheses already considered plausible but uncertain, while contrast sets cover a wider range of competing explanations for the same scientific question.

We call this concentration *prior collapse*, and it affects which candidates *pointwise* surprisal ranks highest. Because most *pointwise* hypotheses begin near the "maybe true" bin, there is limited headroom for confirmatory shifts and much more for disconfirming ones (see Appendix T.1 for the scoring rule and its normalization). Consistent with this asymmetry, high-surprisal *pointwise* candidates in our runs are disproportionately negative. Among the top-50 candidates from each run's 200 experimental validations, only 4% are confirmed, 7% are ambiguous, and 89% are rejected. *Pointwise* surprisal thus primarily surfaces evidence against a focal claim, and because the complement is unstructured, that evidence does not identify which alternative explanation is better supported.

To test whether those rejected hypotheses contain recoverable information, we construct a post-hoc contrastive companion for each *pointwise* hypothesis: we keep the original hypothesis as one explanation, generate alternatives using the same contrast-set proposer, and reuse the same evidence (prompt in Appendix P.3). This uses contrast sets as an evaluation lens: no new experiments are run. The companion generator sees the hypothesis, the dataset's variables, and the same prior context as a live proposer, but not the analysis, its output, or the verdict, and it is applied whether the hypothesis was confirmed or rejected. On the same evidence, the companion resolves 73% of the 50 candidates to an alternative (Figure 2, right). Counting only confirmations as resolutions, *pointwise* discovery leaves 96% of them unresolved, against 23% for CONTRASTIVE. Because both readings use the same evidence, the difference comes from how the hypothesis space is represented: the same evidence *pointwise* discovery spends to reject a claim can instead identify what is more plausible.

5.2 CONTRASTIVE Discovery turns rejections into resolved questions

Having shown that *pointwise* rejection often contains information that resolves an alternative, we next test whether using contrast sets to guide search directly captures this latent information to resolve more scientific questions. For both methods, we rank candidates by their respective discovery scores (i.e., κ_t or surprisal) and evaluate the top-K questions retained for inspection. *Pointwise* baselines cluster together (Figure 3): ranking candidates by surprisal, novelty, or feasibility yields similar numbers of resolved questions. This suggests that the limitation is in the *pointwise* representation rather than its scoring function since an isolated hypothesis can only be confirmed or rejected. In contrast, CONTRASTIVE Discovery produces substantially more resolved questions across budgets: at K=200, it resolves 161.9 questions on average versus 97.2 for *pointwise* surprisal (+67%).

The confirmative-only CONTRASTIVE curve shows that the gain is not driven by simply confirming more prior-leading explanations: when restricted to confirmative resolutions, CONTRASTIVE Discovery sits at or below the *pointwise* surprisal baseline (82.2 vs. 97.2 at K=200), plausibly because κ_t, as the search reward, values confirming an already-leading explanation less, so the search pursues such questions less often. The additional yield comes almost entirely from corrective resolutions, where evidence shifts support from the prior-leading explanation to another member of the contrast set. This corrective gap is the information lost in *pointwise* discovery. There, the same evidence can reject the focal hypothesis, but because the alternative space is collapsed into an unstructured complement, it cannot identify which competing explanation becomes better supported.

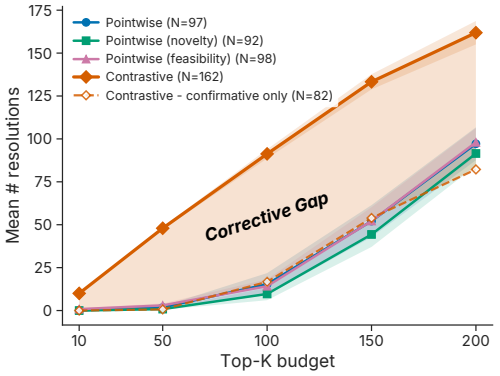


Figure 3: **CONTRASTIVE Discovery resolves more scientific questions across top-K budgets.** Shaded gap marks corrective resolutions, which *pointwise* baselines record only as rejections. The dashed curve counts confirmative resolutions only, and legend values are resolved counts at K=200. Curves are means over datasets and bands are ± 1 standard error across the six datasets. The number of ruled-out alternatives is shown in Appendix Q.2.

Resolving more questions does not come at the expense of ruling out alternatives: placing 0.80 belief on one explanation leaves at most 0.20 for the rest, so at the typical size of three, a resolved set almost always rules out at least one alternative (see Appendix Q.2). To count a verdict in either direction for both methods, we report the *decisive-unit rate*: the fraction of top-K units reaching any verdict (confirmed or rejected for *pointwise*, resolved or ruled out for CONTRASTIVE). At K=200, it is 0.91 for CONTRASTIVE versus 0.82 for *pointwise* and is higher across all six datasets (Appendix Q.3).

5.3 The gain comes from evaluating evidence against explicit alternatives

To evaluate which change produces the gain, we isolate three factors in the discovery pipeline: *ranking* (*pointwise* surprisal vs. CONTRASTIVE κ_t), *evaluation* (focal hypothesis only vs. full contrast set), and *discovery-time construction* (contrast set built post-hoc vs. during the discovery run). We compare five variants: *pointwise* only, +CONTRASTIVE ranking, +CONTRASTIVE evaluation, +CONTRASTIVE

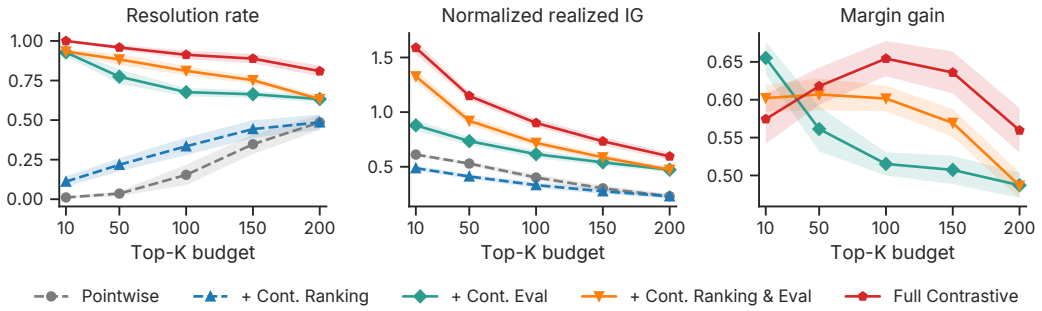


Figure 4: **CONTRASTIVE ranking, evaluation, and discovery-time construction all contribute to resolved-question yield.** Across resolution rate **(left)**, realized information gain (IG; the contrastive belief shift κ_t, not bounded by 1) **(middle)**, and margin gain **(right)**, Full CONTRASTIVE performs best at nearly every budget. CONTRASTIVE ("Cont." in the legend) ranking improves resolution even under *pointwise* evaluation, CONTRASTIVE evaluation produces a larger step change by allowing evidence to resolve over alternatives, and combining both gives stronger updates than either alone. Since *pointwise* evaluation has no explicit runner-up, margin gain is defined only for contrastively evaluated variants. Curves are means and bands are ± 1 standard error across the six datasets.

ranking & evaluation, and Full CONTRASTIVE. Post-hoc variants reuse evidence collected for a *pointwise* hypothesis with a companion contrast set; Full CONTRASTIVE constructs the contrast set during discovery time. We evaluate on three metrics: resolution rate measures whether the update answers the question, realized information gain (IG; the contrastive belief shift κ_t, with $M = 2$ under *pointwise* evaluation) measures the magnitude of belief shift, and margin gain measures how much the belief gap between the leading explanation and runner-up changes after observing the evidence (Figure 4).

The largest separation comes from changing how evidence is evaluated. Variants with *pointwise* evaluation remain low in resolution rate (at most 0.22 at K=50), whereas variants with CONTRASTIVE evaluation improve substantially (ranging from 0.77 to 0.96 at K=50). Once alternatives are explicitly represented, the same evidence can then adjudicate among explanations. CONTRASTIVE ranking raises resolution under either evaluator, but under *pointwise* evaluation it yields lower information gain than *pointwise* ranking (Figure 4, middle), revealing a ranker–evaluator mismatch: the ranker prioritizes evidence informative over alternatives, while the evaluator collapses that evidence back onto a single focal claim. When CONTRASTIVE ranking is paired with CONTRASTIVE evaluation, both resolution and realized information gain increase.

Finally, post-hoc companion sets improve over *pointwise* evaluation, but Full CONTRASTIVE leads on all three metrics at almost every top-K budget, consistent with alternatives built during discovery shaping the question, analysis plan, and evidence collected. Resolved-question yield is driven primarily by contrastive evaluation, strengthened by contrastive ranking, and highest when alternatives are built during discovery.

Compatibility with *pointwise* surprisal. The contrastive belief shift stays compatible with the *pointwise* surprisal signal used by prior work [5]. To compare the two scores on shared candidates, we take the highest-prior hypothesis of each contrast set and score it with *pointwise* surprisal. The two scores are aligned (Spearman $\rho = 0.675$); where they disagree, candidates ranked higher by CONTRASTIVE are often corrective and those ranked higher by *pointwise* confirmative. Under this single-hypothesis scoring, CONTRASTIVE Discovery also matches the *pointwise* surprisal, novelty, and feasibility baselines in its yield of surprising hypotheses, those whose *pointwise* surprisal reaches 0.5 (see Appendix S).

5.4 Resolutions hold up against experts and independent references

A resolution only records where one model's belief settled, so we further validate it against independent references that do not depend on that model.

Blinded human expert evaluation. Thirty-eight evaluators with MS or PhD degrees judged 60 questions across all datasets using only the question, candidate explanations, and evidence; a median of 8 rated each item, and we use the majority vote. As Table 1 shows, experts agreed with CON-

TRASTIVE Discovery's explanation on 78.4% of resolved questions. Of the 17 disagreements across all 60 items, they favored a competing explanation in only three cases; the rest were inconclusive or split. For the *pointwise* comparison, we used the same 60 questions, evidence, and expert judgments, evaluating only the explanation with the highest prior belief in the contrast set. When *pointwise* confirmed its hypothesis, experts endorsed 23 of the 32 findings (71.9%) for each method; when *pointwise* rejected and stopped, they selected the alternative identified by CONTRASTIVE Discovery on 81.8%. Evaluators also judged contrast sets well formed on 80% of ratings, flagged a plausible alternative as missing on 15%, and rated the result a substantive finding on 51%. As quality controls, evaluators reliably flagged planted defective sets, and pairing questions with unrelated evidence reduced appropriate-analysis ratings from 73% to 7% (Appendix I).

Resolutions follow the evidence. To verify that resolutions follow the evidence rather than the wording of the alternatives, we gave the belief elicitor only the numerical results for 14 claims adapted from DiscoveryBench [30] with documented findings, with no analyst text, the alternatives worded identically across conditions, and their order randomized. Given the coefficient and confidence interval, it ranked the documented answer first in all 11 cases where an interval could be recovered. Holding alternative wording fixed, reversing the evidence's sign reversed resolutions on all 14 claims, while removing evidence prevented resolution. Separately, recomputing 331 analyses directly from raw data reproduced the agent's reported statistics on 96.1% (Appendix J).

Table 1: **Experts chose the explanation CONTRASTIVE Discovery identified on most resolved questions.** The same 60 items are grouped by the CONTRASTIVE outcome and the *pointwise* verdict.

Questions	n	Experts agree
CONTRASTIVE resolved	51	40 (78.4%)
CONTRASTIVE unresolved	9	3 (33.3%)
Pointwise confirmed	32	23 (71.9%)
Pointwise rejected	22	18 (81.8%)
Pointwise ambiguous	6	2 (33.3%)
All items	60	43 (71.7%)

Agreement across model families. When we re-elicited beliefs on 180 contrast sets with `gpt-4o`, `gemini-3-flash`, and `claude-sonnet-4-6`, all three models chose the same leading explanation on 55% of sets before seeing the evidence and on 88.9% after, and agreement rose in every dataset (see Appendix A).

Incomplete contrast sets. A contrast set that omits the explanation the evidence supports can produce a confident wrong resolution. For 163 questions whose supported explanation was independently identified from recomputed statistics and which the system had originally resolved correctly, we removed that explanation from its contrast set and re-elicited beliefs. The system declined to resolve on 46.6% and resolved to a wrong explanation on 53.4%, mostly to "no detectable effect." Adding the abstention alternative (§3.4) reduced false resolutions to 0.6%, typically flagging the omission rather than withholding an answer. On re-elicitation of the same intact sets, abstention was never selected, and 162/163 resolved to the supported explanation versus 159/163 without abstention. Our main results use the generated alternatives without abstention: a strict majority of experts identified a missing alternative in only 6 of 60 rated sets (Appendix K).

5.5 The gain holds across scoring rules, costs, and thresholds

To test the robustness of the gain in §5.2, we vary key evaluation choices (details in Appendix L).

Joint versus independent scoring. CONTRASTIVE Discovery both writes out the alternatives and scores them together in one belief distribution; the gain could come from either. Keeping the contrast sets but scoring each alternative separately with the *pointwise* elicitor and its threshold γ_{support} resolves 157.8 questions against 161.9 for joint scoring, and both select the same leading alternative on 89.3% of sets. Joint scoring therefore accounts for little of the gain, and contrast sets can be layered onto an agent that keeps its existing *pointwise* scorer.

Budget. Both methods are given the same number of experimental validations, at similar cost per candidate (61.1k tokens for *pointwise* against 52.8k for CONTRASTIVE). We also recompute the comparison with both methods capped at the same total tokens or the same dollar cost. CONTRASTIVE Discovery still resolves more under each: 148.3 against 87.7 at equal tokens and 145.8 against 92.2 at equal dollars.

Thresholds. If our resolution threshold were easier to clear than the *pointwise* confirmation threshold, we would resolve more questions for that reason alone. Because τ_{support} bounds a share of a

distribution over M alternatives and γ_{support} bounds a scalar belief, we match the two by how often they accept the same evidence. On 1,543 cases where the *pointwise* elicitor scored the same evidence for the prior-leading explanation and it stayed the leader, we pair each τ_{support} from 0.60 to 0.90 with the γ_{support} that accepts the closest fraction of cases. CONTRASTIVE Discovery resolves more at every matched pair and in each of the six datasets. At our $\tau_{\text{support}} = 0.80$, the matched γ_{support} is 0.695, stricter than the 0.625 we use, so the reported gain is conservative (Appendix E).

Implicit contrast sets. *Pointwise* discovery might independently propose the alternatives in our contrast sets, allowing post-hoc recovery. Using a separate LLM judge, we compare every *pointwise* hypothesis against a stratified sample of 300 contrast sets. Among the questions that at least one *pointwise* hypothesis also addresses, the *pointwise* hypotheses contain the complete alternative set for only 2.6% within a single run at matched budget, and for only 15.4% after pooling all *pointwise* runs on the dataset (4–6$\times$ that budget). At matched budget, the posterior-leading explanation appears among the *pointwise* hypotheses 52.0% of the time when the evidence confirms the prior-leading explanation, but only 25.4% when it corrects it. So where *pointwise* evaluation would reject its hypothesis, it has usually not proposed the explanation the evidence favors (Appendix L).

6 Related Work

Recent surveys trace the shift from task-specific automation toward autonomous LLM-driven discovery systems [31, 32]. Related work develops hypothesis-generation and refinement methods [33, 34], alongside pairwise LLM evaluation [35] and scientific argument alignment [36]. In contrast to such settings that generate or evaluate hypotheses individually or in pairs, CONTRASTIVE Discovery focuses on how evidence distinguishes a set of mutually exclusive and collectively exhaustive competing explanations for the same question.

This reframing builds on a longer tradition of information-theoretic experimental design, which values experiments by their expected effect on belief [37, 38, 39]. Our objective measures realized posterior–prior belief change through Bayesian surprise [40], following prior evidence-informed LLM discovery work [5, 41]. ClaimSpect [42] also moves beyond binary claim assessment, but organizes existing evidence into non-exclusive aspects rather than constructing mutually exclusive alternatives to be resolved by new analyses. A related line of work also generates research hypotheses from scientific literature, from early text-mining systems [43] to LLM agents [12, 44, 45]. CONTRASTIVE Discovery is compatible with this setting because contrastive resolution requires only a scientific question and plausible alternatives. In this work, we focus on open-ended, data-driven discovery [11, 46] over tabular datasets [29, 30], where experiments can be run directly against rich scientific data.

7 Conclusion and Limitations

We introduced CONTRASTIVE Discovery, which replaces the single hypothesis that a scientific discovery agent proposes and scores with a contrast set of competing explanations. Because only this unit of discovery changes, an existing agent can use contrast sets without changing its architecture, either to re-evaluate evidence it has already collected or to guide its search. Across six domains, contrast sets turn *pointwise* rejections into resolved questions, almost entirely through corrective resolutions. Ablations trace this gain primarily to evaluating evidence against the full contrast set. In a blinded study, human experts broadly agree with the explanations the system selects on resolved questions. Together, these results support explicit competing explanations as a way for discovery agents to resolve more scientific questions from the same evidence and validation budget.

Our experiments use tabular, observational datasets and a single base agent with a `gpt-4o` backbone. On such data, a resolution is a statistical rather than a causal finding. Because contrast sets need only an agent's hypotheses and evidence, we expect the gains to carry over to other discovery agents, which we leave to future work. Contrast sets may also be malformed or omit plausible explanations. Structural audits and expert ratings find most sets well formed (Appendix C and §5.4), and adding an abstention alternative flags omitted explanations (§5.4). Finally, resolution is measured with LLM-elicited beliefs, so resolved-question counts depend on the elicitor, even though most resolutions agree with recomputed statistics and experts (§5.4). Future work should test whether these gains extend to laboratory experiments and longer-horizon interventions.

Acknowledgments and Disclosure of Funding

This material is based on work supported by the AI Research Institutes program supported by the NSF and Intel Corporation under NSF award DMR-2433348. This research is supported in part by the National Science Foundation (IIS-2505098). Research supported by the NVIDIA Academic Grant Program using 4 × RTX PRO 6000 Blackwell GPUs.

References

[1] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL https://arxiv.org/abs/2408.06292.

[2] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL https://arxiv.org/abs/2504.08066.

[3] Peter Jansen, Oyvind Tafjord, Marissa Radensky, Pao Siangliulue, Tom Hope, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Daniel S. Weld, and Peter Clark. Codescientist: End-to-end semi-automated scientific discovery with code-based experimentation, 2025. URL https://arxiv.org/abs/2503.22708.

[4] Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent laboratory: Using llm agents as research assistants, 2025. URL https://arxiv.org/abs/2501.04227.

[5] Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, and Peter Clark. Autodiscovery: Open-ended scientific discovery via bayesian surprise. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kJqTkj2HhF.

[6] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025. URL https://arxiv.org/abs/2502.18864.

[7] Kexin Huang, Ying Jin, Ryan Li, Michael Y. Li, Emmanuel Candes, and Jure Leskovec. Automated hypothesis validation with agentic sequential falsifications. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=iTevNo8PzG.

[8] Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery, 2025. URL https://arxiv.org/abs/2506.13131.

[9] Mert Yuksekgonul, Daniel Koceja, Xinhao Li, Federico Bianchi, Jed McCaleb, Xiaolong Wang, Jan Kautz, Yejin Choi, James Zou, Carlos Guestrin, and Yu Sun. Learning to discover at test time, 2026. URL https://arxiv.org/abs/2601.16175.

[10] Jishu Sen Gupta, Syed Mohamad Tawseeq, Harini S I, Somesh Kumar Singh, Yaman Kumar Singla, David Doermann, Rajiv Ratn Shah, and Balaji Krishnamurthy. Experigen: Data driven hypothesis generation with experimental verification, 2025. URL https://openreview.net/forum?id=LKGflZgEKB.

[11] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Sanchaita Hazra, Ashish Sabharwal, and Peter Clark. Data-driven discovery with large generative models, 2024. URL https://arxiv.org/abs/2402.13610.

[12] Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang, editors, *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6709–6738, Albuquerque, New Mexico, April 2025. Association for Computational Linguistics. ISBN 979-8-89176-189-6. doi: 10.18653/v1/2025.naacl-long.342. URL https://aclanthology.org/2025.naacl-long.342/.

[13] Aniketh Garikaparthi, Manasi Patwardhan, Lovekesh Vig, and Arman Cohan. IRIS: Interactive research ideation system for accelerating scientific discovery. In Pushkar Mishra, Smaranda Muresan, and Tao Yu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 592–603, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-253-4. doi: 10.18653/v1/2025.acl-demo.57. URL https://aclanthology.org/2025.acl-demo.57/.

[14] Ruochen Li, Liqiang Jing, and Xinya Du. Learning to generate research idea with dynamic control, 2024. URL https://openreview.net/forum?id=zCb0dPvGYN.

[15] Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. Large language models as biomedical hypothesis generators: A comprehensive evaluation. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=q36rpGlG9X.

[16] Laurent Itti and Pierre Baldi. Bayesian surprise attracts human attention. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005. URL https://proceedings.neurips.cc/paper_files/paper/2005/file/0172d289da48c48de8c5ebf3de9f7ee1-Paper.pdf.

[17] Bas C. van Fraassen. *The Scientific Image*. Oxford University Press, 1980. ISBN 9780191597473.

[18] Peter Lipton. *Inference to the Best Explanation*. Routledge, 2nd edition, 2004. ISBN 9780415242035.

[19] Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In H. Jaap van den Herik, Paolo Ciancarini, and H. H. L. M. (Jeroen) Donkers, editors, *Computers and Games*, pages 72–83, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75538-8.

[20] Henry W. Sprueill, Carl Edwards, Khushbu Agarwal, Mariefel V. Olarte, Udishnu Sanyal, Conrad Johnston, Hongbin Liu, Heng Ji, and Sutanay Choudhury. Chemreasoner: Heuristic search over a large language model's knowledge space using quantum-chemical feedback, 2024. URL https://arxiv.org/abs/2402.10980.

[21] Rui Ha, Chaozhuo Li, Rui Pu, Litian Zhang, Xi Zhang, and Sen Su. DSG-MCTS: A dynamic strategy-guided Monte Carlo tree search for diversified reasoning in large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 10530–10544, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.532. URL https://aclanthology.org/2025.emnlp-main.532/.

[22] Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. Aide: Ai-driven exploration in the space of code, 2025. URL https://arxiv.org/abs/2502.13138.

[23] Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Maria Lupidi, Andrei Lupu, Roberta Raileanu, Kelvin Niu, Tatiana Shavrina, Jean-Christophe Gagnon-Audet, Michael Shvartsman, Shagun Sodhani, Alexander H. Miller, Abhishek Charnalia, Derek Dunfield, Carole-Jean Wu, Pontus Stenetorp, Nicola Cancedda, Jakob Nicolaus Foerster, and Yoram Bachrach. Ai research agents for machine learning: Search, exploration, and generalization in mle-bench, 2025. URL https://arxiv.org/abs/2507.02554.

[24] Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In *Proceedings of the 17th European Conference on Machine Learning*, ECML'06, page 282–293, Berlin, Heidelberg, 2006. Springer-Verlag. ISBN 354045375X. doi: 10.1007/11871842_29. URL https://doi.org/10.1007/11871842_29.

[25] Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. NOVA: An iterative planning framework for enhancing scientific innovation with large language models. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21330–21359, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1099. URL https://aclanthology.org/2025.findings-acl.1099/.

[26] Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search, 2025. URL https://arxiv.org/abs/2505.19209.

[27] Matthew K. Horton, Patrick Huck, Ruo Xi Yang, Jason M. Munro, Shyam Dwaraknath, Alex M. Ganose, Ryan S. Kingsbury, Mingjian Wen, Jimmy X. Shen, Tyler S. Mathis, Aaron D. Kaplan, Karlo Berket, Janosh Riebesell, Janine George, Andrew S. Rosen, Evan W. C. Spotte-Smith, Matthew J. McDermott, Orion A. Cohen, Alex Dunn, Matthew C. Kuner, Gian-Marco Rignanese, Guido Petretto, David Waroquiers, Sinead M. Griffin, Jeffrey B. Neaton, Daryl C. Chrzan, Mark Asta, Geoffroy Hautier, Shreyas Cholia, Gerbrand Ceder, Shyue Ping Ong, Anubhav Jain, and Kristin A. Persson. Accelerated data-driven materials science with the materials project. *Nature Materials*, 24(10):1522–1532, Oct 2025. ISSN 1476-4660. doi: 10.1038/s41563-025-02272-0. URL https://doi.org/10.1038/s41563-025-02272-0.

[28] SDSS Collaboration, Gautham Adamane Pallathadka, Mojgan Aghakhanloo, James Aird, Andrés Almeida, Singh Amrita, Friedrich Anders, Scott F. Anderson, Stefan Arseneau, Consuelo González Avila, Shir Aviram, Catarina Aydar, Carles Badenes, Jorge K. Barrera-Ballesteros, Franz E. Bauer, Aida Behmard, Michelle Berg, F. Besser, Christian Moni Bidin, Dmitry Bizyaev, Guillermo Blanc, Michael R. Blanton, Jo Bovy, William Nielsen Brandt, Joel R. Brownstein, Johannes Buchner, Esra Bulbul, Joseph N. Burchett, Leticia Carigi, Joleen K. Carlberg, Andrew R. Casey, Priyanka Chakraborty, Julio Chanamé, Vedant Chandra, Cristina Chiappini, Igor Chilingarian, Johan Comparat, Kevin Covey, Nicole Crumpler, Katia Cunha, Elena D'Onghia, Xinyu Dai, Jeremy Darling, Megan Davis, Nathan De Lee, Niall Deacon, José Eduardo Méndez Delgado, Sebastian Demasi, Mariia Demianenko, Delvin Demke, John Donor, Niv Drory, Monica Alejandra Villa Durango, Tom Dwelly, Oleg Egorov, Evgeniya Egorova, Kareem El-Badry, Mike Eracleous, Xiaohui Fan, Emily Farr, Douglas P. Finkbeiner, Logan Fries, Peter Frinchaboy, Nicola Pietro Gentile Fusillo, Luis Daniel Serrano Félix, Boris Gaensicke, Emma Galligan, Pablo García, Joseph Gelfand, Katie Grabowski, Eva Grebel, Paul J Green, Hannah Greve, Catherine Grier, Emily Griffith, Paloma Guetzoyan, Pramod Gupta, Zoe Hackshaw, Patrick B. Hall, Keith Hawkins, Viola Hegedűs, Saskia Hekker, T. M. Herbst, J. J. Hermes, Lorena Hernández-García, Pranavi Hiremath, David W Hogg, Jon Holtzman, Keith Horne, Danny Horta, Yang Huang, Brian Hutchinson, Maximilian Häberle, Hector Javier Ibarra-Medel, Alexander P. Ji, Paula Jofre, James W. Johnson, Jennifer Johnson, Evelyn J. Johnston, Mary Kaldor, Ivan Katkov, Arman Khalatyan, Sergey Khoperskov, Ralf Klessen, Matthias Kluge, Anton M. Koekemoer, Juna A. Kollmeier, Marina Kounkel, Kathryn Kreckel, Dhanesh Krishnarao, Mirko Krumpe, Ivan Lacerna, Chervin Laporte, Sebastien Lepine, Jing Li, Fu-Heng Liang, Guilherme Limberg, Xin Liu, Sarah Loebman, Knox Long, Yuxi Lu, Madeline Lucey, Alejandra Z. Lugo-Aranda, Mary Loli Martínez Martinez-Aldama, Kevin McKinnon, Ilija Medan, Andrea Merloni, Sean Morrison, Natalie Myers, Szabolcs Mészáros, Johanna Müller-Horn, Samir Nepal, Melissa Ness, David Nidever, Christian Nitschelm, Audrey Oravetz, Jonah Otto, Kaike Pan, Facundo Pérez Paolino, Castalia Alenka Negrete Peñaloza, Marc Pinsonneault, Manuchehr Taghizadeh Popp, Adrian Price-Whelan, Nadiia Pulatova, Anna Barbara Queiroz, Jordan Raddick, Amy Rankine, Hans-Walter Rix, Carlos Román-Zúñiga, Daniela Fernández Rosso, Jessie Runnoe, Serat Mahmud Saad, Mara Salvato, Sebastian F. Sanchez, Natascha Sattler, Andrew Saydjari, Conor Sayres, Kevin Schlaufman, Donald P. Schneider, Axel Schwope, Lucas M. Seaton, Rhys Seeburger, Javier Serna, Sanjib Sharma, Yue Shen, Amaya Sinha, Brian

Sizemore, Marzena Sniegowska, Yingyi Song, Diogo Souto, Keivan Stassun, Matthias Steinmetz, Zachary Stone, Alexander Stone-Martinez, Guy S. Stringfellow, Aurora Mata Sánchez, José Sánchez-Gallego, Jonathan Tan, Jamie Tayar, Riley Thai, Ani Thakar, Pierre Thibodeaux, Yuan-Sen Ting, Andrew Tkachenko, Benny Trakhtenbrot, Jose G. Fernandez Trincado, Nicholas Troup, Jonathan R. Trump, Natalie Ulloa, Roeland P. Van der Marel, Pablo Vera, Sandro Villanova, Jaime Villaseñor, Ji Wang, Zachary Way, Anne-Marie Weijmans, Adam Wheeler, John C. Wilson, Aida Wofford, Tony Wong, Qiaoya Wu, Dominika Wylezalek, Xiang-Xiang Xue, Renbin Yan, Qian Yang, Nadia Zakamska, Eleonora Zari, Gail Zasowski, Grisha Zeltyn, Zheng Zheng, Catherine Zucker, and Rodolfo de J. Zermeño. The nineteenth data release of the sloan digital sky survey, 2025. URL https://arxiv.org/abs/2507.07093.

[29] Ken Gu, Ruoxi Shang, Ruien Jiang, Keying Kuang, Richard-John Lin, Donghe Lyu, Yue Mao, Youran Pan, Teng Wu, Jiaqian Yu, Yikun Zhang, Tianmai M. Zhang, Lanyi Zhu, Mike A Merrill, Jeffrey Heer, and Tim Althoff. BLADE: Benchmarking language model agents for data-driven science. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13936–13971, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.815. URL https://aclanthology.org/2024.findings-emnlp.815/.

[30] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vyflgpwfJW.

[31] Tianshi Zheng, Zheye Deng, Hong Ting Tsang, Weiqi Wang, Jiaxin Bai, Zihao Wang, and Yangqiu Song. From automation to autonomy: A survey on large language models in scientific discovery. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17733–17750, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.895. URL https://aclanthology.org/2025.emnlp-main.895/.

[32] Atilla Kaan Alkan, Shashwat Sourav, Maja Jablonska, Simone Astarita, Rishabh Chakrabarty, Nikhil Garuda, Pranav Khetarpal, Maciej Pióro, Dimitrios Tanoglidis, Kartheik G. Iyer, Mugdha S. Polimera, Michael J. Smith, Tirthankar Ghosal, Marc Huertas-Company, Sandor Kruk, Kevin Schawinski, and Ioana Ciucă. A survey on hypothesis generation for scientific discovery in the era of large language models, 2025. URL https://arxiv.org/abs/2504.05496.

[33] Shiyang Duan, Yuan Tian, Qi Bing, and Xiaowei Shao. Bayes-entropy collaborative driven agents for research hypotheses generation and optimization, 2025. URL https://arxiv.org/abs/2508.01746.

[34] Yujing Ke, Kevin George, Kathan Pandya, David Blumenthal, Maximilian Sprang, Gerrit Großmann, Sebastian Vollmer, and David Antony Selby. Biodisco: Multi-agent hypothesis generation with dual-mode evidence, iterative feedback and temporal evaluation, 2025. URL https://arxiv.org/abs/2508.01285.

[35] Adian Liusie, Potsawee Manakul, and Mark J. F. Gales. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models, 2024. URL https://arxiv.org/abs/2307.07889.

[36] Yingqiang Gao, Nianlong Gu, Jessica Lam, James Henderson, and Richard Hahnloser. Evaluating unsupervised argument aligners via generation of conclusions of structured scientific abstracts. In Yvette Graham and Matthew Purver, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 151–160, St. Julian's, Malta, March 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-short.14. URL https://aclanthology.org/2024.eacl-short.14/.

[37] D. V. Lindley. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986 – 1005, 1956. doi: 10.1214/aoms/1177728069. URL https://doi.org/10.1214/aoms/1177728069.

[38] David J. C. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4(4):590–604, Jul 1992. ISSN 0899-7667. doi: 10.1162/neco.1992.4.4.590.

[39] Kathryn Chaloner and Isabella Verdinelli. Bayesian Experimental Design: A Review. *Statistical Science*, 10(3):273 – 304, 1995. doi: 10.1214/ss/1177009939. URL https://doi.org/10.1214/ss/1177009939.

[40] Pierre Baldi and Laurent Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Netw.*, 23(5):649–666, June 2010. ISSN 0893-6080. doi: 10.1016/j.neunet.2009.12.007. URL https://doi.org/10.1016/j.neunet.2009.12.007.

[41] Dhruv Agarwal, Reece Adamson, Andrew McCallum, Peter Clark, Ashish Sabharwal, and Bodhisattwa Prasad Majumder. Evidence-informed llm beliefs for continual scientific discovery, 2026. URL https://arxiv.org/abs/2606.29182.

[42] Priyanka Kargupta, Runchu Tian, and Jiawei Han. Beyond true or false: Retrieval-augmented hierarchical analysis of nuanced claims. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 29664–29679, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1434. URL https://aclanthology.org/2025.acl-long.1434/.

[43] Scott Spangler, Angela D. Wilkins, Benjamin J. Bachman, Meena Nagarajan, Tajhal Dayaram, Peter Haas, Sam Regenbogen, Curtis R. Pickering, Austin Comer, Jeffrey N. Myers, Ioana Stanoi, Linda Kato, Ana Lelescu, Jacques J. Labrie, Neha Parikh, Andreas Martin Lisewski, Lawrence Donehower, Ying Chen, and Olivier Lichtarge. Automated hypothesis generation based on mining scientific literature. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, page 1877–1886, New York, NY, USA, 2014. Association for Computing Machinery. ISBN 9781450329569. doi: 10.1145/2623330.2623667. URL https://doi.org/10.1145/2623330.2623667.

[44] Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, page 279–299. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.18. URL http://dx.doi.org/10.18653/v1/2024.acl-long.18.

[45] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers, 2024. URL https://arxiv.org/abs/2409.04109.

[46] Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. Hypobench: Towards systematic and principled benchmarking for hypothesis generation, 2026. URL https://arxiv.org/abs/2504.11524.

Technical Appendices and Supplementary Material

Contents

A Cross-Elicitor Agreement and Prior Collapse 17

B Stability of Contrastive Belief Elicitation 18

C Audit of MECE Structure in Generated Contrast Sets 19

D Hypothesis-Order Randomization in Belief Elicitation 21

E Threshold Calibration 22

F Deduplicating Replicated Hypotheses 24

G Dataset Details 25

H Heritability Analysis: Parent–Child Score Correlation in MCTS over LLM Hypotheses 27

I Blinded Expert Evaluation 28

J Mechanical Fidelity Audits 29

K Handling Incomplete Contrast Sets 30

L Baselines and Robustness of the Gain 31

M Exploration Breadth: Column Coverage and Sibling Anchoring 33

N Recovery of Published Findings 36

O Elicitation Design Across the Two Methods 36

P Prompts 37

Q Extended Results 45

R Hyperparameters 48

S Compatibility with *Pointwise* Surprisal 49

T Metric Definitions and Agent Framework 51

U Example Contrast Sets 54

V Broader Societal Impact 55

W Cost Analysis 55

A Cross-Elicitor Agreement and Prior Collapse

A useful and effective hypothesis generator should produce claims that span the prior-belief axis, because scientific discovery is bidirectional: evidence may either support an initially unlikely claim or refute an initially plausible one. We audit LLM-elicited priors for hypotheses generated by free-form prompting (*pointwise*) and contrastive partitioning (CONTRASTIVE), and identify prior collapse, a failure of free-form generation that the contrastive construction structurally mitigates.

Setup. We fix the hypothesis generator to `gpt-4o` and re-elicit beliefs using three independent elicitors: `gpt-4o`, `gemini-3-flash`, and `claude-sonnet-4-6`. For each of six datasets, we sample 90 *pointwise* hypotheses and 30 CONTRASTIVE sets, stratified by tree level. A *pointwise* node contains a single hypothesis, whereas a CONTRASTIVE node contains a partition of $M \geq 2$ mutually exclusive cells (typically $M = 3$), each treated as a standalone hypothesis for the per-cell label analysis. This gives us a sample of 540 *pointwise* hypotheses and 180 CONTRASTIVE sets (573 CONTRASTIVE cells).

Following our generation process exactly, for each hypothesis (or contrast cell) and elicitor, we draw 10 categorical belief samples over five labels: `definitely_true`, `maybe_true`, `uncertain`, `maybe_false`, and `definitely_false`. We map these labels to scores 1.0, 0.75, 0.5, 0.25, and 0.0, respectively, and aggregate them into a $\mathrm{Beta}(\alpha, \beta)$ posterior under a Jeffreys prior. We use the Beta mean as the per-hypothesis belief score. To measure cross-elicitor agreement on CONTRASTIVE at the unit of analysis the runtime metric actually uses, we additionally elicit *set-level* partition allocations: each elicitor distributes 100 probability points jointly over the M cells of a set (5 samples for each set, elicitor, and prior or posterior call, averaged into one M-vector).

Prior collapse is a generator-side failure, not an elicitor-side one. The prior collapse on *pointwise* priors does not imply that elicitors are simply noisy. Table 2 shows cross-elicitor agreement on per-hypothesis ranks (*pointwise*, Spearman ρ) and on per-set partition allocations (CONTRASTIVE, mean per-set Pearson r, with $1-\mathrm{TVD}$ as a total-variation similarity robustness check at the partition level the runtime metric actually uses). On *both* methods, evidence sharply raises agreement: pooled ρ rises $0.660 \rightarrow 0.887$ on *pointwise*, pooled r rises $0.664 \rightarrow 0.923$ on CONTRASTIVE (paired Wilcoxon $p < 10^{-4}$ for both, all per-dataset cells significant under both Bonferroni and Benjamini–Hochberg corrections over 24 tests). The symmetry of this convergence argues against elicitor noise as the source of the unreliability we observe at the prior step: when given evidence, the elicitors are more cross-elicitor consistent on both methods. What is *not* symmetric is the prior collapse, which is a property of the free-form generator that the elicitor cannot fix. Constructing hypotheses as a mutually exclusive partition rather than independent free-form claims reduces the collapse at the source. This matters because discovery is bidirectional: maximal information gain either comes from evidence refuting an initially plausible claim or from evidence supporting an initially unlikely one, and each requires the generator to propose hypotheses on the corresponding side of the prior axis. Under *pointwise* prior collapse, the refutational direction dominates. Because the generator rarely produces claims with low elicited prior, discoveries in which evidence *confirms* an initially unlikely claim are rare. CONTRASTIVE's partition tends to include lower-prior cells by forcing explicit alternatives, restoring access to the full prior axis on which discovery operates, while leaving the elicitor's evidence-driven posterior convergence intact.

Do different model families pick the same explanation? Agreement on ranks and allocations does not show that the families agree on *which* explanation the evidence supports, so we also compared their leading cells. On the same 180 contrast sets, all three families lead with the same explanation on 55.0% of sets before the evidence (99 of 180; chance, given each family's observed choice frequencies, is 21.8%) and on 88.9% after it (160 of 180; McNemar exact $p = 1.2 \times 10^{-13}$). The rise holds in every dataset (Conversation 63→90%, Mortgage 63→90%, Materials 37→93%, NLS 73→87%, Requirements 47→90%, SDSS 47→83%). How often each family crosses $\tau_{\text{support}} = 0.80$ differs (`gpt-4o` 77.2%, `gemini-3-flash` 86.1%, `claude-sonnet-4-6` 67.8%), so the resolution rate depends on the elicitor while the leading explanation does not. Three families that start apart and converge after reading the evidence are hard to explain as one model agreeing with itself. One caveat remains: all three read the same `gpt-4o` analyst summary, so part of the convergence could come from shared framing rather than the evidence alone.

Table 2: Per-dataset cross-elicitor agreement, mean across the three elicitor pairs. ***Pointwise*** **columns** (n=90 hypotheses per dataset, n=540 pooled): mean pairwise Spearman ρ on per-hypothesis mean-belief scores. **CONTRASTIVE columns** ($n\approx$30 sets per dataset, n=177 pooled after dropping M=2 sets from the Pearson statistic): mean pairwise per-set Pearson r on the M-cell partition allocation, with mean $1-\mathrm{TVD}$ total-variation similarity bracketed as a robustness check (interpretable as fraction of probability mass agreeing across raters; computed on all aligned sets including M=2). M=2 sets ($\sim$1.7% of CONTRASTIVE sets) are degenerate for per-set Pearson and dropped from that statistic only. Bonferroni and Benjamini–Hochberg corrections applied over 24 tests (two methods $\times$ prior and posterior $\times$ 6 datasets, treating each per-dataset cell of this table as the unit of significance claim); all cells significant at $p < 0.05$ under both corrections.

Dataset	**Setting**	***Pointwise***	**CONTRASTIVE** r **[**$1-\mathrm{TVD}$**]**
Conversation	Prior	0.730	0.764 [0.822]
	Posterior	0.873	0.917 [0.893]
Mortgage	Prior	0.602	0.730 [0.786]
	Posterior	0.854	0.934 [0.894]
Materials Project	Prior	0.672	0.506 [0.780]
	Posterior	0.906	0.933 [0.897]
NLS Raw	Prior	0.632	0.779 [0.776]
	Posterior	0.875	0.936 [0.905]
Requirements Eng.	Prior	0.506	0.647 [0.837]
	Posterior	0.890	0.922 [0.878]
SDSS Galaxy	Prior	0.642	0.559 [0.737]
	Posterior	0.862	0.898 [0.888]
Pooled	Prior	**0.660**	**0.664 [0.790]**
	Posterior	**0.887**	**0.923 [0.893]**

B Stability of Contrastive Belief Elicitation

Our contrastive score relies on elicited prior and posterior distributions over each contrast set. To reduce stochastic variation, the production pipeline draws $n = 5$ point-allocation samples for each prior and posterior elicitation, normalizes each valid allocation to the simplex, and averages the resulting distributions before computing contrastive belief shift. This appendix evaluates whether $n = 5$ is a sufficient cost–stability tradeoff for the quantities used by the pipeline.

We analyze 3,213 contrast sets with complete $n = 5$ prior and posterior elicitations. For contrast set i, let $p_i^{(1)}, \ldots, p_i^{(5)}$ denote prior samples and $q_i^{(1)}, \ldots, q_i^{(5)}$ denote posterior samples. The production estimates are

$$\bar{p}_i = \frac{1}{5}\sum_{r=1}^{5} p_i^{(r)}, \qquad \bar{q}_i = \frac{1}{5}\sum_{r=1}^{5} q_i^{(r)}. \tag{2}$$

We focus on the two derived quantities used downstream: the realized contrastive information gain

$$L_i = D_{\mathrm{KL}}(\bar{q}_i \,\|\, \bar{p}_i), \tag{3}$$

and posterior concentration

$$\bar{q}_i^{\max} = \max_m \bar{q}_{i,m}. \tag{4}$$

Prior samples are more dispersed, but the averaged score is stable. We first measure within-contrast-set dispersion using the mean pairwise total variation distance among the five elicited distributions. Prior samples are more dispersed than posterior samples: the pooled mean pairwise TVD is 0.125 for priors and 0.065 for posteriors. This gap is expected, since prior elicitation is performed before observing evidence, while posterior elicitation is conditioned on the experiment result. The higher prior dispersion is not itself a failure mode; it reflects uncertainty before evidence is observed. What matters downstream is whether the averaged prior and posterior produce stable belief-shift scores. Figure 5 shows the full CDF, with 85.5% of posterior contrast sets below pairwise TVD 0.10.

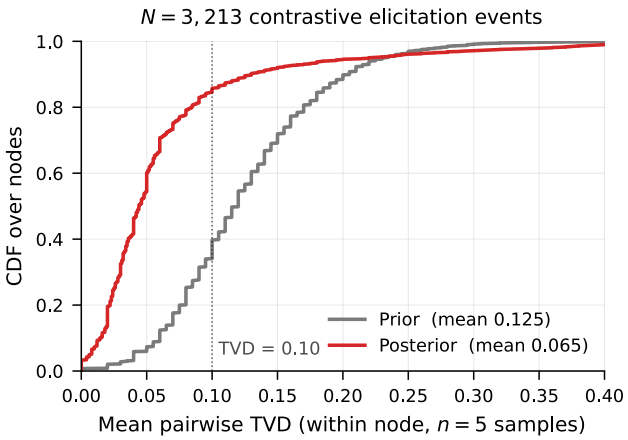


Figure 5: Within-contrast-set dispersion of the $n = 5$ belief samples. Prior elicitations are more dispersed, while posterior elicitations are consistently sharper after conditioning on evidence.

Table 3: Stability diagnostics for the $n = 5$ contrastive belief elicitor. Prior allocations are more dispersed than posterior allocations, as expected before evidence is observed. The downstream quantities used by the pipeline remain stable under bootstrap and drop-one sensitivity.

Diagnostic	**Value**
Contrast sets with complete $n = 5$ prior/posterior elicitations	3,213
Mean pairwise TVD, prior samples	0.125
Mean pairwise TVD, posterior samples	0.065
Posterior contrast sets with pairwise TVD below 0.10	85.5%
Posterior per-hypothesis SE of sample mean	0.0165
Posterior maximum-probability sample SD	0.043
Median bootstrap CI half-width for $\bar{q}_i^{\max} = \max_m \bar{q}_{i,m}$	0.029
Median drop-one drift in $\bar{q}_i^{\max}$	0.005
Median drop-one drift in $L_i = D_{\mathrm{KL}}(\bar{q}_i \| \bar{p}_i)$	0.053 bits

Five samples are not dominated by any single elicitation. Table 3 reports the diagnostics most relevant to the choice of $n = 5$. The posterior mean has an average per-hypothesis standard error of 0.0165, corresponding to about 1.7 points on the 100-point allocation scale. The posterior concentration $\bar{q}_i^{\max}$ is stable under resampling: its median bootstrap 95% CI half-width is 0.029, and dropping any one of the five samples changes it by only 0.005 in median. The KL score L_i depends on both the prior and posterior estimates, and is therefore the relevant joint stability check for the discovery score. Although L_i is noisier than $\bar{q}_i^{\max}$, as expected for a nonlinear function of two finite-sample distributions, its median drop-one drift is only 0.053 bits. These sensitivities indicate that the production score is not typically driven by a single elicitation sample.

Increasing n would further reduce the standard error, but only at the usual $1/\sqrt{n}$ rate while increasing elicitation cost linearly. We therefore use $n = 5$ as a practical tradeoff: it reduces elicitation variance enough for ranking and aggregate analysis while keeping belief elicitation cost manageable.

C Audit of MECE Structure in Generated Contrast Sets

CONTRASTIVE Discovery assumes that each contrast set defines a well-formed partition of possible analysis outcomes. In particular, the explanations in a contrast set should be mutually exclusive and collectively exhaustive (MECE): any plausible result of the planned analysis should support at most one explanation, and the set should not omit an obvious residual outcome. The generator prompt (Appendix P) asks for one of a fixed set of partition templates, a residual outcome, and a quantitative criterion that maps the planned analysis to each alternative; this audit checks the sets it actually produced.

Table 4: Structural MECE audit summary under the primary $4/5$ rule.

Quantity	Count / estimate	Rate
Structurally MECE sampled contrast sets	$372/384$	96.9%
Design-weighted population estimate	–	98.4%
95% stratified bootstrap CI	–	$[97.3\%, 99.3\%]$
Overlap or redundancy failures	$9/384$	2.3%
Inherently non-MECE mechanism failures	$3/384$	0.8%

Audit protocol. We extracted $3{,}260$ parseable contrast sets from the full CONTRASTIVE runs across 6 datasets and 14 runs. From this pool, we drew a frozen stratified sample of 384 contrast sets, stratifying by the requested partitioning dimension and by the number of hypotheses in the set. This intentionally oversamples rare partition types, making failure modes easier to inspect.

Each sampled contrast set was judged by `gpt-4o` using two independent passes, with 5 randomized-order votes per pass. The first pass identified pairs of hypotheses that were not logically distinct, including paraphrases, redundant hypotheses, and subsumption relationships. The second pass assigned structural failure tags, including missing residual outcome, redundant hypothesis, subsumption, insufficient number of hypotheses, and inherently non-MECE structure. We classify a contrast set as structurally MECE if no overlap pair, missing-residual, redundancy, subsumption, or inherently non-MECE flag receives at least $4/5$ votes. Insufficient granularity is evaluated separately and does not count as a structural failure. We use this supermajority rule because manual inspection showed that unanimity misses recurring redundancy failures, while simple majority admits more judge noise.

Results. Under the primary $4/5$ rule, $372/384$ sampled contrast sets are structurally MECE (Table 4), corresponding to a raw audit pass rate of 96.9%. Reweighting the stratified sample back to the full $3{,}260$-item pool gives an estimated population pass rate of 98.4% with a stratified bootstrap 95% confidence interval of $[97.3\%, 99.3\%]$. The conclusion is stable across reasonable vote thresholds: the raw pass rate is 93.8% under a $3/5$ rule and 99.0% under a $5/5$ rule.

Failure modes. The non-MECE cases are concentrated in two recurring patterns. First, 9 of the 12 failures contain overlapping or redundant hypotheses. These usually occur when the proposer includes two versions of a null or residual outcome. For example, one contrast set contains both "association disappears after controlling for PhD prestige" and "no effect of male contributions on total citations." These are not exclusive outcomes; both describe the absence of a residual association. Another contrast set contains both "distribution does not follow a specific crystal system" and "no pattern across crystal systems," which again duplicate the same residual hypothesis.

Second, 3 failures arise from mechanism-style contrast sets that are not naturally MECE. These contrast sets ask the hypotheses to distinguish causal pathways, such as whether residence type affects later wealth through social capital or through initial economic advantage. Without stronger identification assumptions, these mechanisms are not mutually exclusive and may not be empirically separable from the proposed analysis. These cases suggest that mechanism-discriminating templates should either be restricted to observationally distinguishable outcomes or excluded from automatic MECE generation.

Cross-provider re-audit. The audit above uses a judge from the same provider as the proposer, so we re-ran it on the same 384 sets with Claude Sonnet, keeping the prompts, the two passes, and the $4/5$ rule (Table 5). The cross-provider judge passes 87.0% of sets (91.4% design-weighted) against 96.9% (98.4%) for `gpt-4o`, flagging 50 sets where `gpt-4o` flagged 12. The same-provider judge is the more lenient of the two, so we report both figures and carry the stricter one. The disagreement runs almost entirely one way (on the overlap check, 24 sets flagged only by Sonnet against 3 flagged only by `gpt-4o`), and the extra flags are mostly alternatives that overlap or entail one another. They concentrate by question type: half of the mechanism-discriminating sets and 40% of the ordering sets fail, against none of the directional sets. A structural audit cannot tell whether a plausible explanation is missing altogether; for that, the human raters of Appendix I, who judged the set well formed on 80% of ratings and flagged a missing alternative on 15%, are the external estimate.

Table 5: Structural MECE pass rate on the same 384-set sample under the same rule, by judge.

Judge	Provider	Raw pass	Design-weighted pass	Sets flagged
`gpt-4o` (same provider as proposer)	OpenAI	96.9%	98.4%	12
Claude Sonnet	Anthropic	87.0%	91.4%	50

Does the defect change reported answers? Of the 50 flagged sets, 48 link to a recorded run. On 19 of them the system never reached the resolution threshold, so nothing was reported. Of the 29 resolved questions, 7 selected a member of a judge-named overlapping pair, making those resolutions ambiguous; the other 22 were not identified as implicated by this check. Those 7 are 2.3% of the 302 resolutions in the audited sample, or 1.4% reweighted to the corpus (the audit oversamples risky question types). Two further checks point the same way. Mechanically merging each overlapping pair into one alternative, without re-eliciting beliefs, changes the leading explanation on none of the 379 sets that link to a recorded run, and would newly resolve up to 8 questions left open. Flagged sets also reach the threshold less often than clean ones (60.5% against 85.0% design-weighted, a 24-percentage-point gap with 95% CI 10 to 40), as expected when belief splits between two alternatives that mean the same thing. The defect therefore costs findings rather than inventing them; these are patterns in the logged runs, not a re-elicitation. The residual is small: seven resolutions are ambiguous between two overlapping alternatives, and 19 of the 48 flagged sets could not be tested because the judge named no pair. Whether an explanation is missing altogether is a separate failure, tested in Appendix K.

Interpretation. This audit supports the structural premise of CONTRASTIVE: the generated contrast sets usually define valid partitions of the analysis outcome space. The remaining failures are not arbitrary incoherence. They are concentrated in a small number of recognizable template errors, primarily duplicated residual hypotheses and mechanism-style hypotheses that cannot be cleanly partitioned without additional assumptions. We therefore treat MECE structure as satisfied for most sets, while using these failure modes to identify targeted prompt and template improvements. Because a same-provider judge is the more lenient, we carry the cross-provider figure of 87.0% as the conservative estimate.

D Hypothesis-Order Randomization in Belief Elicitation

CONTRASTIVE Discovery elicits two probability distributions over the explanations in a contrast set: a prior distribution before observing the analysis result and a posterior distribution after observing it. Since language models can be sensitive to list position, these probabilities could depend on the order in which hypotheses are shown. We therefore randomize the hypothesis order independently for each elicitation call, then invert the permutation before computing priors, posteriors, KL, or any downstream statistic.

Validation protocol. We validate this choice with two checks, summarized in Table 6. First, we run a controlled permutation study on four contrast sets from two datasets, with each set evaluated under five different hypothesis orderings. We compare a canonical protocol, which preserves the displayed ordering, against the randomized protocol used in our experiments. After each call, all probabilities are mapped back to the same canonical hypothesis identities. We then measure pairwise consistency across orderings using total variation distance (TVD), where lower TVD means the elicited distribution is less affected by presentation order. Second, we run two parallel CONTRASTIVE runs on BLADE Mortgage that differ only in whether hypothesis order is randomized. This checks whether randomization changes the aggregate scale of the elicited KL updates or posterior confidence.

Findings. Table 6 shows that randomization improves pairwise consistency across displayed hypothesis orderings. The effect is strongest for the prior distribution: prior pairwise TVD falls from 0.077 to 0.042, a $1.83\times$ reduction in order-induced dispersion. Posterior elicitation is already more stable, but still improves, with posterior pairwise TVD falling from 0.031 to 0.025.

The aggregate check shows no evidence that randomization changes the scale of the elicited belief updates. The confidence intervals for average KL, average normalized KL, and average posterior

Table 6: Effect of hypothesis-order randomization on belief elicitation. The top block measures pairwise consistency across five displayed hypothesis orderings; lower TVD means greater order robustness. The bottom block compares aggregate belief statistics in two parallel CONTRASTIVE runs that differ only in whether hypothesis order is randomized. KL is measured in bits, and $\mathrm{KL}/\log_2 M$ normalizes by the number of explanations M in the contrast set. Δ is the randomized minus the canonical mean, with a bootstrap 95% CI.

Pairwise consistency	canonical	randomized	ratio
Prior pairwise TVD	0.077	0.042	1.83×
Posterior pairwise TVD	0.031	0.025	1.24×
Aggregate belief statistics	canonical	randomized	Δ [95% CI]
$\mathbb{E}[\mathrm{KL}]$	0.987	0.975	$-0.012\ [-0.332, +0.284]$
$\mathbb{E}[\mathrm{KL}/\log_2 M]$	0.554	0.568	$+0.014\ [-0.167, +0.181]$
$\mathbb{E}[\max p^{\text{post}}]$	0.874	0.873	$-0.001\ [-0.046, +0.045]$

confidence all include zero. Thus, randomization reduces presentation-order sensitivity without a detectable change in the contrastive belief shift.

Interpretation. These results justify hypothesis-order randomization as a control for presentation bias. It improves consistency across equivalent prompt orderings while preserving the aggregate behavior of CONTRASTIVE. We therefore randomize hypothesis order during belief elicitation and always map probabilities back to canonical hypothesis identities before computing the metric.

E Threshold Calibration

This appendix justifies the numerical thresholds used by the pointwise and contrastive metrics. The main paper defines the contrastive update score κ_t, the contrastive support/rule-out gates, and the corresponding pointwise gates. Here we explain why the chosen numerical values are principled and empirically aligned.

Contrastive thresholds. We use

$$\tau_{\text{support}} = 0.80, \qquad \tau_{\text{reject}} = 0.10.$$

The support threshold requires strong posterior concentration on one hypothesis, not merely a plurality winner. At the modal contrast-set size $M = 3$, posterior mass 0.80 leaves at most 0.10 for each remaining hypothesis under uniform residual splitting, giving an $8\times$ leader-to-average-rival ratio (at least $4\times$ leader-to-runner-up in the worst case). The rule-out threshold 0.10 is the corresponding residual mass implied by this $M = 3$ anchor:

$$\frac{1 - 0.80}{3 - 1} = 0.10.$$

We keep this value fixed across M so that ruling out has the same numerical meaning across contrast-set sizes.

For contrastive surprise we threshold the score at $\kappa_t \geq 0.5$. Because $\kappa_t = \mathrm{KL}/\log_2 M$ is divided by $\log_2 M$ but not bounded by one, this is a threshold on the scaled score and should not be read as "half of the available information gain"; it is placed at the same numerical value as the pointwise belief-change threshold of 0.5 and avoids the M-dependence of a raw KL threshold.

Pointwise thresholds. We use

$$\gamma_{\text{support}} = 0.625, \qquad \gamma_{\text{reject}} = 0.375.$$

These values are anchored to the elicitor's five-level belief scale

$$\{0,\ 0.25,\ 0.5,\ 0.75,\ 1.0\}.$$

The support threshold 0.625 is the midpoint between "uncertain" (0.5) and "maybe true" (0.75); the rejection threshold 0.375 is the midpoint between "maybe false" (0.25) and "uncertain" (0.5). Thus, the pointwise thresholds correspond to crossing out of the uncertainty region in either direction, rather than being tuned post hoc to maximize agreement with the contrastive thresholds.

Table 7: Paired calibration between pointwise and contrastive thresholds. Precision is the fraction of pointwise positives that also satisfy the corresponding contrastive gate; recall is the fraction of contrastive positives recovered by the pointwise gate. Confidence intervals are bootstrap 95% intervals.

Comparison	n	F_1	95% CI	Precision	Recall
Support gate	1,543	0.937	[0.928, 0.947]	0.915	0.961
Rejection gate	1,672	0.896	[0.884, 0.907]	0.981	0.824
Joint two-sided gate	3,215	0.934	[0.927, 0.940]	0.973	0.898

Paired calibration check. We verify that these independently motivated thresholds align on paired production data. The calibration set contains 3,215 paired elicitations across 6 datasets and 14 runs. Each item contains both the contrastive posterior over the full contrast set and the pointwise posterior mean for the focal hypothesis, elicited from the same evidence.

For support, we compare the pointwise support gate against the contrastive support gate on cases where the focal hypothesis is the contrastive posterior winner. For rejection, we compare the pointwise rejection gate against the contrastive rule-out gate on cases where the focal hypothesis is not the contrastive posterior winner. We then evaluate the full two-sided gate on all paired items.

Table 7 shows that the pointwise and contrastive gates are well aligned when applied to the same evidence. The support gate reaches $F_1 = 0.937$, indicating that pointwise support closely tracks contrastive support. The rejection gate reaches $F_1 = 0.896$ with very high precision, meaning that pointwise rejection almost always agrees with the contrastive rule-out gate. The full two-sided gate reaches $F_1 = 0.934$ across all 3,215 paired items.

Table 8: Threshold robustness of the reported gain, recomputed through the exact top-K=200 estimator of the headline result. Each contrastive threshold τ is paired with the *pointwise* threshold γ of matched firing rate; the middle columns show how closely each pair is matched on the 1,543 paired elicitations where both gates score the same evidence. Gain is the difference as a fraction of the budget. †At τ=0.95 the positive-set overlap falls to 0.567, so that row is exploratory.

		evidence that the gates are matched			*resolved units at top-200*		
τ	γ	contr. fires	point. fires	F1	contrastive	pointwise	gain
0.600	0.535	95.3%	95.3%	0.966	185.5	104.8	+0.403
0.667	0.585	93.3%	93.5%	0.963	178.7	100.8	+0.390
0.700	0.605	92.1%	92.0%	0.959	174.8	99.3	+0.378
0.750	0.635	90.5%	90.6%	0.952	169.7	96.7	+0.365
0.800	**0.695**	86.8%	86.3%	0.927	**161.9**	88.3	**+0.368**
0.850	0.730	80.5%	81.1%	0.897	145.9	80.1	+0.329
0.889	0.760	71.8%	71.4%	0.867	125.8	50.1	+0.378
0.900	0.800	65.1%	65.1%	0.835	108.4	43.7	+0.324
0.950†	0.895	33.8%	35.5%	0.567	50.7	20.9	+0.149

Threshold robustness. The contrastive advantage persists across the tested matched thresholds. We recomputed the headline top-K=200 comparison of Section 5.2 for τ_{support} from 0.60 to 0.95. Because the two gates sit on different scales, we moved both together: each τ is paired with the pointwise threshold γ that fires on the closest fraction of the 1,543 paired elicitations where both gates score the same evidence, and Table 8 reports those firing rates and the overlap of the sets they admit so the pairing can be checked. The contrastive advantage is positive at every threshold, and in every dataset at every threshold (54 of 54 cells). At the reported $\tau = 0.80$ the matched γ is 0.695; the paper uses 0.625, which fires 4.9 percentage points more often, so the published gain is the more conservative one. Raising τ alone with γ fixed at 0.625 stays positive through 0.90 and reverses at 0.95, where the contrastive gate demands 95 of 100 points on one alternative while the pointwise bar stays below "maybe true"; even at the matched $\gamma = 0.895$, the two gates admit substantially different sets there ($F_1 = 0.567$), so we treat that row as exploratory and rest the claim on $\tau \leq 0.90$. We present the matched sweep as a robustness check rather than a preregistered test, since it was designed after the fixed-γ sensitivity.

Are the allocations calibrated? We do not treat the 100-point allocations as calibrated probabilities. κ_t is a divergence between two distributions over the same alternatives, and "resolved" says that belief has concentrated, not that the leading explanation is 80% likely to be true. What the metric needs is that allocations be stable, insensitive to presentation, and responsive to evidence, which Appendix B, Appendix D, and Appendix J check. We nonetheless compared stated confidence with a reference outside the model. For the 170 directional contrast sets (125 distinct analyses) whose statistic can be recomputed from the raw data, we read off the cell the recomputed statistic implies. On the 166 the method resolves, the leading explanation agrees with that cell on 163 (98.2%) at a mean stated confidence of 0.945, so confidence sits 0.037 below agreement (95% CI -0.055 to -0.015, clustered on the analysis); the gap remains negative after deduplication and leave-one-dataset-out. The top-label expected calibration error is 0.040, and the multiclass Brier score falls from 0.601 for the prior to 0.047 for the posterior. There is no sign of overconfidence against this reference. This check covers the mechanically decidable directional subset, which is easier than the corpus (97.6% of these nodes resolve, against about 83% corpus-wide), and its reference is a significance decision on the same sample; calibration against scientific truth is a separate question.

Interpretation. The thresholds are not arbitrary numerical knobs. The contrastive thresholds are anchored to posterior concentration and normalized information gain. The pointwise thresholds are anchored to the elicitor's categorical belief scale. The paired calibration then verifies that these two independently motivated threshold systems produce closely aligned decisions on the same evidence, supporting a fair comparison between pointwise and contrastive metrics rather than one driven by mismatched gate strictness.

F Deduplicating Replicated Hypotheses

Following prior work [5], we apply a post-run deduplication pass to test whether a method receives extra credit for repeatedly generating the same semantic hypothesis under different surface forms.

Shared deduplication procedure. Both pointwise and contrastive runs use the same semantic-equivalence machinery. We first collapse exact-string duplicates, embed the remaining hypothesis texts with the same embedding model, build a Ward agglomerative clustering tree, and use an LLM judge to decide whether proposed cluster merges are semantically equivalent. Each merge decision is judged with 10 votes and accepted only when at least 70% of votes mark the two clusters as the same hypothesis. Thus, both methods use the same embedding model, clustering procedure, judge prompt, judge model, vote count, and merge threshold.

Pointwise deduplication. In the pointwise setting, each search node proposes one focal hypothesis. Deduplication is therefore performed at the node level: two nodes are merged if their hypothesis texts express the same claim. This matches the representation used by the pointwise metric, where each candidate is a single hypothesis.

Contrastive deduplication. In the contrastive setting, each search node proposes a contrast set containing M mutually exclusive hypotheses. Deduplicating only at the node level would be inappropriate: it would treat the whole contrast set as one textual object and would not remove repeated hypotheses across different contrast sets. We therefore flatten contrastive outputs and deduplicate at the hypothesis level. Each hypothesis inside a contrast set is treated as a candidate for cross-node deduplication.

We add one structural constraint: hypotheses from the same contrast set are never merged with each other. They are MECE alternatives to the same analysis question, not replications. For example, "positive effect," "negative effect," and "no effect" may be textually related, but they are distinct alternatives by construction. This within-set hard gate prevents deduplication from collapsing valid contrastive alternatives.

Resolved questions after deduplication. We recount the resolved questions of Section 5.2 so that each resolved claim counts once, grouping confirmed *pointwise* hypotheses and the leading explanations of resolved contrast sets by the clusters above. At $K{=}200$, the average number of resolved questions falls from 97.2 to 87.4 for *pointwise* surprisal and from 161.9 to 141.9 for

CONTRASTIVE Discovery, so duplicates make up 10.1% and 12.4% of their resolutions, respectively. CONTRASTIVE Discovery still resolves 62% more questions than *pointwise* surprisal, against 67% before deduplication, and leads on all six datasets.

G Dataset Details

We evaluate our framework on six datasets spanning scientific and social domains, including materials science, astrophysics, finance and policy, software engineering, sociology and labor economics, and online discourse. Table 9 summarizes the datasets used in our experiments, grouping datasets from the same source benchmark where applicable. We then describe dataset-specific caveats and the filtering logic used for the two largest scientific datasets.

Table 9: Summary of experimental datasets. The benchmark portfolio spans heterogeneous scientific and social domains and covers a broad range of sample sizes, enabling evaluation across different levels of statistical power, feature spaces, and domain structure.

Dataset	Domain	Rows	Cols.	Row unit
Materials Project	Materials science	68,805	28	Inorganic crystal
SDSS Galaxy (DR19)	Astronomy	50,000	28	Galaxy
BLADE datasets				
Mortgage	Finance / policy	2,380	13	Mortgage application
Conversation	Online discourse	7,975	70	Debate contribution
DiscoveryBench datasets				
Requirements Engineering	Software engineering	276	202	Survey respondent
NLS Raw	Sociology / economics	12,686	61	Survey respondent

Materials Project. The Materials Project dataset contains 68,805 rows and 28 columns, with each row corresponding to a unique inorganic crystalline material identified by `material_id`. For our experiments, we restrict the queried subset to materials satisfying `nelements` $\in \{2, 3, 4\}$, excluding pure elements and materials with five or more constituent elements. We additionally require `energy_above_hull` ≤ 0.05 eV/atom to focus on near-stable materials. A key caveat is that elastic moduli are missing for approximately 86% of rows; only 9.6K materials have computed elasticity values. Consequently, hypotheses involving elastic properties are evaluated on a self-selected subsample rather than on the full dataset.

SDSS Galaxy (DR19). The SDSS Galaxy dataset contains 50,000 rows and 28 columns, with each row corresponding to one galaxy identified by a unique `specObjID`. The dataset represents the astronomy and extragalactic domain. Our query applies the following filters: `class = 'GALAXY'`, `zWarning = 0` to retain clean redshift fits, redshift in the range $0.02 \leq z \leq 0.3$, and apparent magnitude in the range $14 \leq$ `modelMag_r` ≤ 22. The `zWarning` filter is applied at query time, although the `zWarning` column itself is not retained in the saved CSV. The query returns the top 50,000 entries under the server-default ordering, which should be interpreted as a minor caveat because the subset is not randomly sampled. Several photometric columns, including `modelMag_u`, `modelMag_g`, `modelMag_z`, `petroR50_r`, and `petroR90_r`, use `-9999` as a missing-data sentinel. The pipeline does not filter these sentinels. Where we inspected SDSS analyses during the computation audit (Appendix J), the agent's own code dropped or replaced them before computing statistics. In addition, `subClass` is missing for approximately 62% of rows.

G.1 BLADE datasets

Mortgage. The BLADE Mortgage dataset contains 2,380 rows and 13 columns, where each row corresponds to one mortgage application. The dataset is situated in the finance and policy domain, with particular relevance to lending and discrimination. Its columns are:

female, black, housing_expense_ratio, self_employed, married, mortgage_credit, consumer_credit, bad_history, PI_ratio, deny, loan_to_value, denied_PMI, accept

Missingness is limited but nonzero: female contains 18 missing values, married contains 2 missing values, and PI_ratio contains 5 missing values.

Conversation. The BLADE Conversation dataset contains 7,975 rows and 70 columns, with each row corresponding to one contribution, or comment, in an Edge.org debate thread. The dataset belongs to the online discourse and gender participation domain. It contains repeated observations along two axes: 522 unique ThreadId values, with an average of approximately 15 rows per thread and a maximum of 436 rows in a single thread; and 728 unique Id_num contributors, with an average of approximately 11 rows per contributor and a maximum of 504 rows for a single contributor. To avoid contributor leakage, we use an Id_num-level split: 70% of contributors are assigned to discovery and 30% to validation. This preserves row-level granularity while preventing contributor-specific attributes from leaking across splits. We prioritize contributor-level splitting because the contributor axis contains more invariant attributes, including gender, PhD institution, and h-index.

G.2 DiscoveryBench datasets

Requirements Engineering for ML-Enabled Systems. The DiscoveryBench Requirements Engineering dataset contains 276 rows and 202 columns, with each row corresponding to a unique survey respondent identified by ID. The dataset belongs to the software engineering domain and captures practices in requirements engineering for ML-enabled systems. Because the dataset has only 276 respondents, it has limited statistical power for detecting subtle effects. In addition, many of the 202 columns are binary indicators derived from multi-select survey questions, which should be considered when interpreting hypotheses involving sparse or highly correlated features.

NLS Raw. The DiscoveryBench NLS Raw dataset contains 12,686 rows and 61 columns, with each row corresponding to a unique NLSY79 survey respondent identified by ID. The longitudinal panel has already been flattened into a wide-format representation with one row per person. The dataset is situated in the sociology and labor economics domain. A key caveat is that columns mix measurements from different survey years, including 1979, 1981, and later waves. Therefore, hypotheses over this dataset should respect the implicit temporal ordering of variables and avoid treating later measurements as valid predictors of earlier outcomes.

G.3 Filtering rationale for scientific datasets

For the two large scientific datasets, Materials Project and SDSS Galaxy, we apply simple domain-motivated filters before running hypothesis generation. These filters are not intended to optimize downstream performance. Instead, they define physically coherent, cleanly measured, and computationally tractable hypothesis spaces. In both cases, the goal is to avoid forcing the model to generate hypotheses across incompatible physical regimes or over measurements known to be unreliable.

SDSS Galaxy. For SDSS Galaxy, we restrict the dataset to entries with class = 'GALAXY'. This keeps the hypothesis space physically coherent: stars, quasars, and galaxies obey different physical regimes, and several galaxy-oriented columns, such as size and profile-radius measurements, would be meaningless or differently interpreted for non-galaxy objects. We also require zWarning = 0, the standard SDSS quality flag indicating a clean redshift fit. Nonzero warning values often indicate failed pipeline checks, such as low signal-to-noise, bad template matches, or problematic spectra; including these entries would inject unreliable redshift-dependent measurements into many hypotheses.

We further restrict redshift to $0.02 \leq z \leq 0.3$. The lower bound removes very nearby objects for which peculiar velocities can be comparable to the Hubble-flow velocity, making redshift a poor distance proxy. The upper bound keeps the sample within the approximate regime of the SDSS Main Galaxy Sample, rather than mixing in higher-redshift luminous red galaxies or other populations governed by different selection effects. Finally, we require $14 \leq$ modelMag_r ≤ 22. The bright-end cut avoids saturated sources for which photometry may be unreliable, while the faint-end cut avoids objects for which photometric uncertainty and redshift signal-to-noise become limiting. The top-50,000 cap is a practical scale constraint: it remains below the SDSS unauthenticated query limit, fits

comfortably in memory, and provides sufficient statistical power for hypothesis generation. Unlike the physical and quality filters, this cap is primarily pragmatic rather than physically principled.

Materials Project. For Materials Project, we restrict to materials with `nelements` $\in \{2, 3, 4\}$. This excludes pure elements, whose small and heavily studied set would make many property hypotheses trivial or degenerate. It also excludes materials with five or more elements, such as high-entropy alloys and complex multicomponent oxides, where individual elemental contributions are often highly entangled and harder to interpret cleanly. Retaining binary, ternary, and quaternary compounds captures much of functional inorganic chemistry, including families such as rock salts, perovskites, spinels, and Heusler compounds, while keeping the hypothesis space tractable.

We additionally require `energy_above_hull` ≤ 0.05 eV/atom. This threshold follows the common Materials Project convention for selecting materials that are plausibly synthesizable or near-stable. Materials substantially above the convex hull are increasingly likely to be hypothetical or physically unstable, making their computed properties less useful for generating hypotheses about realizable materials. This filter also removes unusual metastable phases whose extreme properties could dominate regressions or produce hypotheses driven by outliers rather than robust physical structure.

Overall design principle. Across both scientific datasets, the filters are chosen to satisfy four criteria. First, retained samples should be physically coherent, so that hypotheses are generated within a comparable regime. Second, measurements should be clean enough to avoid obvious garbage-in, garbage-out effects, such as failed redshift fits or saturated photometry. Third, the filtered datasets should remain large enough to support statistically meaningful hypothesis generation. Finally, the filters should be conventional enough to be defensible to domain experts, rather than appearing as post hoc cherry-picking.

Portfolio coverage. The resulting benchmark portfolio covers six domains: materials science through Materials Project, astrophysics through SDSS Galaxy, finance and policy through BLADE Mortgage, software engineering through DiscoveryBench Requirements Engineering, sociology and labor economics through DiscoveryBench NLS Raw, and online discourse through BLADE Conversation. The datasets also span a broad range of sample sizes: 276 rows for DiscoveryBench Requirements Engineering, 2,380 rows for BLADE Mortgage, 7,975 rows for BLADE Conversation, 12,686 rows for DiscoveryBench NLS Raw, 50,000 rows for SDSS Galaxy, and 68,805 rows for Materials Project. This row-count spread is useful for analyzing how hypothesis-generation and validation behavior vary under different statistical-power regimes.

H Heritability Analysis: Parent–Child Score Correlation in MCTS over LLM Hypotheses

A precondition for hierarchical MCTS to outperform flat sampling is that ancestor scores predict descendant scores. Without such correlation, UCT's value-back-propagation has nothing to exploit when choosing which subtree to expand, and the selection rule is statistically equivalent to uniform-random parent selection. To test whether this property holds in our setting, we measured the parent–child Spearman rank correlation of `self_value` (the scalar UCT propagates) across the 200 experimental validations of each search tree, pooled across six datasets.[1] We use `self_value` as the headline metric because it is what UCT directly reads when computing Q-values. As a robustness check we additionally report the underlying `surprisal` (the logged belief change: $|\mu^{\text{post}} - \mu^{\text{prior}}|$ in the surprisal-mode baseline, $H^{\text{prior}} - H^{\text{post}}$ in the contrastive method), which has none of the clipping or M-normalization construction in `self_value`. For the contrastive method this secondary quantity is the entropy reduction logged as a diagnostic alongside each node, not the contrastive belief shift κ_t of Section 3.3; `self_value` is the clipped κ_t that UCT reads. Reporting both addresses the concern that an observed parent–child correlation could be a construction artifact rather than a real property of the search tree.

The surprisal-mode baseline (*pointwise*) yields essentially zero parent–child heritability at lag 1 ($\rho = +0.031$, 95% CI $[-0.055, +0.125]$, $n = 1{,}169$ pairs, BH-FDR $q = 0.79$): UCT-driven

[1] All confidence intervals are 95% percentile cluster bootstraps over the 6 runs (2,000 iterations); p-values are from within-level permutation tests (2,000 permutations) with add-one smoothing; BH q-values are Benjamini–Hochberg-adjusted across the per-metric 7-lag family. All results reproduce from `--seed 0`.

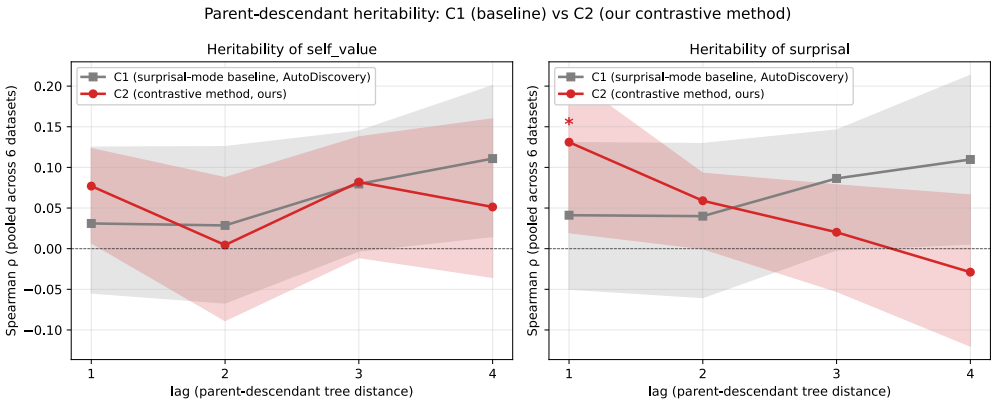


Figure 6: **Parent–descendant heritability across MCTS lags.** Spearman rank correlation between a parent's score and its descendant's score at increasing tree distance, pooled across 6 datasets and the 200 experimental validations of each tree. Solid lines are point estimates; shaded regions are 95% percentile cluster-bootstrap CIs over runs. **Left:** `self_value` (the UCT-propagated reward, the mechanism-defining metric). *Pointwise* (gray) stays near zero at lags 1 and 2 and rises at deeper lags. CONTRASTIVE (red, ours) shows a positive lag-1 trend ($\rho = +0.077$) that is not significant after BH-FDR correction ($q = 0.28$). **Right:** `surprisal` (the logged belief change without clipping or M-normalization; a robustness check). CONTRASTIVE lag-1 is FDR-significant ($\rho = +0.131$, $q = 0.046$, $\star$ marks $q \leq 0.10$) and decays monotonically toward zero by lag 4, as local inheritance would predict. *Pointwise* shows no lag-1 heritability on either metric; its slow rise at deeper lags is consistent with UCT's preferential extension of high-value lineages enriching deep-lag pair pools (selection bias, not mechanism-relevant signal). The weaker `self_value` result is consistent with construction-induced compression (13.7% of nodes hit the clipping ceiling) of the same underlying signal that the `surprisal` panel resolves more cleanly.

selection in this regime operates without exploitable score signal. Our contrastive partition reward (CONTRASTIVE) shows a positive lag-1 trend ($\rho = +0.077$, 95% CI $[+0.007, +0.123]$, $n = 1{,}136$, raw permutation $p = 0.04$) that is not significant after Benjamini–Hochberg correction across the 7-lag family ($q = 0.28$) but FDR-significant on the underlying `surprisal` score ($\rho = +0.131$, $q = 0.046$). The CONTRASTIVE trend on `surprisal` decays monotonically toward zero by lag 4, consistent with local inheritance. The baseline shows no such decay structure on either metric. See Figure 6.

We note that the CONTRASTIVE correlation is consistent with two mechanisms our analysis cannot uniquely distinguish: (a) inheritable signal in the score itself, exploitable by UCT-selection; or (b) a downstream effect of the structured M-cell partition being available as ancestor context to the next-level proposer, which would correlate child and parent scores through generation rather than through selection. The natural ablation, comparing UCT-selection to uniform-random parent selection while preserving context inheritance, would adjudicate (a) vs. (b) and is left to future work.

I Blinded Expert Evaluation

This appendix gives the protocol, reliability, and control analyses behind Table 1.

Protocol. Thirty-eight evaluators with MS or PhD degrees in the mathematical, computational, or natural sciences, recruited and screened separately from the authors (31 through a screened crowd platform, 7 verified in person), rated 60 blinded items: ten per dataset, stratified by the contrastive score κ_t. Each evaluator saw only the research question, the candidate explanations, and the statistical evidence from the executed analysis, with the alternatives in random order; our system's beliefs, its resolution label, and the *pointwise* verdict were never shown. Each item was rated by at least 4 evaluators (median 8). The four questions per item were (Q1) which explanation the evidence supports, with the extra options "none of these is supported" and "cannot tell"; (Q2) whether the alternatives are mutually exclusive and cover the plausible explanations, with options for a missing explanation, overlap, or both; (Q3) whether the result is worth reporting; and (Q4) whether the chosen variables and test fit the question. We score items by majority vote and count no majority, "cannot

tell," and "none" as disagreement with our system's explanation. Five submissions were excluded under a rule fixed at the first exclusion (one incomplete session and four evaluators who missed both attention checks). Restricting to the 30 evaluators who caught every attention check lowers the overall agreement from 71.7% to 68.3%, so inattentive raters do not carry the result. Expert evaluations were collected with informed consent and kept confidential; the information concerned our methods and scientific outputs, not the evaluators as individuals. Under our institution's IRB guidance, this activity does not constitute human-subjects research and therefore did not require IRB review.

Both methods were compared on the same 60 questions, using the same evidence and expert judgments. The *pointwise* hypothesis was the explanation our system rated most plausible before seeing the evidence, evaluated on its own.

Resolution agreement (Q1). The expert majority matched our system's explanation on 43 of 60 items (71.7%, 95% CI 60.0–83.3 by item-clustered bootstrap): 40 of the 51 items it marked resolved (78.4%) and 3 of the 9 it left unresolved. Of the 17 disagreements, the experts chose a competing explanation on only 3; on 8 they judged the evidence insufficient and on 6 they formed no majority, so most disagreements fall on evidence the experts themselves could not settle. Across all 60 items the panel committed to a specific alternative on 46, which shows that the executed experiments usually separate the alternatives well enough for experts to commit. Agreement was 79% on confirmative resolutions and 77% on corrective ones. Against the *pointwise* baseline: on the 32 items it confirmed, the experts endorsed 23 findings for each method; on the 22 items it rejected and stopped, the alternative our system identified matched the expert majority on 18 (81.8%; exact 95% CI 60–95). A related subset, the 22 items the baseline did not confirm and our system resolved, overlaps this one on 19 items and gives 17 of 22; the main text reports the pointwise-rejected subset. Per dataset, agreement ranges from 6 of 10 to 9 of 10, too few items to compare domains.

Reliability. On the load-bearing judgment (Q1, four to six options), pairwise agreement was 64% and Krippendorff's α was 0.54 (95% CI 0.45–0.61), and in a descriptive stratification α was 0.60 on resolved items versus 0.22 on unresolved ones, so rater disagreement concentrates on the items CONTRASTIVE Discovery leaves unresolved. On set completeness (Q2), 80% of ratings judged the set well formed, 15% flagged a missing plausible alternative, 4% overlap, and 1% both; raw agreement on the binary omission call was 83% (Gwet's AC1 $= 0.73$; the lower Fleiss' κ of 0.35 reflects the 80/20 response skew). A strict majority flagged a missing alternative on 6 of the 60 sets. Analysis appropriateness (Q4) was judged appropriate on 73% of ratings, wrong variable or test on 19%, and cannot tell on 8% (binary raw agreement 72%, AC1 $= 0.54$). Scientific value (Q3) had low agreement ($\alpha = 0.24$), so we report it only as a distribution: 51% substantive finding, 20% minor, 3% too obvious to state, 25% not a finding.

Validity controls. To check that the ratings track genuine quality, we planted two manually constructed defective sets among the items, one with a plausible alternative removed and one with overlapping alternatives: 11 of the 14 evaluators who saw the first flagged a missing alternative and none confused it with overlap, and 10 of 12 flagged the second as defective. On separate control items we kept the contrast set intact but paired it with the evidence of a different question: the well-formed rate stayed at 77%, close to its 80% baseline, while the analysis-appropriateness rating fell from 73% to 7% and the wrong-variable-or-test flag rose from 19% to 84%. The evaluators therefore flag real set defects without blanket-flagging, and recognize when an analysis does not fit its question. Attention controls (mismatched evidence expected to yield "cannot tell") fired as expected on 95% of exposures.

Scope. The experts read the same evidence as our system, so their agreement shows that resolutions track the evidence as independent experts read it; it is not a test of scientific truth. The experts were general quantitative-science raters rather than specialists matched to each item's subfield, and they judged the executed analysis rather than replicating it. The comparison with the *pointwise* baseline is one of endorsed yield, not of accuracy on the items both methods decide.

J Mechanical Fidelity Audits

The belief step reads the analyst's written summary of an executed analysis. Three audits, none of which involves a language model in the scoring, check that this chain is faithful: that the adjudication

follows the evidence it is shown, that the summary reports the numbers the code printed, and that the code computed those numbers correctly.

Adjudication follows the evidence. We took 14 claims from 10 DiscoveryBench training questions with documented findings (a finding with two claims contributes two), constructed contrast sets whose wording was held fixed across conditions, and ran the production posterior elicitor (`gpt-4o-2024-08-06`, five samples, $\tau_{\text{support}} = 0.80$) on numbers-only evidence with the alternatives in randomized order and no analyst prose. Given the full statistic with its interval, available for 11 of the 14, the elicitor placed the documented answer first in every case and resolved 8 of 11, holding back on three whose intervals lie close to zero. Given the coefficient without an interval it still ranked the documented answer first on 13 of 14 while resolving only 7, appropriate caution when the evidence is incomplete. In a separate condition covering all 14 claims, we supplied the statistic with its sign reversed; the resolution followed the reversal in every case. Supplying no evidence produced no resolution on any of the 14 (largest posterior 0.74). Because the wording of the alternatives was identical across the four conditions, these resolutions follow the evidence and not the phrasing. This establishes evidence-driven adjudication, not end-to-end correctness of the pipeline that produces the evidence.

Summaries quote the numbers the code printed. Across all available seeds, 3,015 analyses contained checkable numerical claims in the analyst's written summary. We checked, allowing for rounding, percentages, and scientific notation, whether those numbers appear in the captured standard output of the executed code. 99.1% of these analyses and 99.7% of the 9,239 checked numbers are grounded in a printed value. Reading every mismatch by hand found at least five genuine slips, including a slope quoted from a plot-only run and a count reported as a percentage; the audit caught all of them without a human in the loop. Pairing each summary with a different analysis's output and re-running the identical matcher grounds only 3.8% of numbers, so the check discriminates. Signs are not checked by this audit.

The code computed the statistics correctly. For the analyses whose statistic can be rebuilt directly from named columns of the raw data (400 of 3,274; 331 scorable, spanning every available seed and 251 distinct dataset/test/column combinations), we recomputed the statistic with standard libraries and compared it to the value the agent printed. The values matched on 318 of 331 (96.1%). Every mismatch we could trace came from our recomputation missing a data-handling step in the agent's code (a custom parser, a missing-data rule, a -9999 sentinel replaced before analysis, a group split); a few could not be classified either way. We found no confirmed computation error. Substituting a wrong column matches only 2.1% of the time, so the check is not loose. Coverage is 12% of analyses and not random, since the rebuildable analyses are the simpler ones, so we read this as a check on the covered analyses rather than a pipeline-wide rate.

K Handling Incomplete Contrast Sets

A contrast set that omits the explanation the evidence supports can produce a confident wrong resolution. We measured how often this happens and how well the abstention alternative of Section 3.4 prevents it.

Design. We used two frozen panels. The primary panel is 163 questions, all directional triples ($M = 3$), whose supported explanation is fixed by a statistic recomputed from the raw data with no model in the scoring (121 distinct underlying analyses; intervals are clustered on the analysis). Because the correct answer must be present to be removed, this panel conditions on questions the system originally resolved correctly, so the intact-set control measures re-elicitation stability rather than corpus accuracy. A secondary panel is 40 resolved questions across six partition types and $M \in \{2, 3, 4\}$ whose supported explanation matches the blinded expert majority of Appendix I. The design crossed two factors: the set as generated or with the supported explanation removed, each with or without an added alternative worded "none of the listed alternatives matches the evidence; the contrast set is incomplete." The seven secondary-panel questions with $M = 2$ omitted the removed-explanation condition without the added alternative, since its sole survivor would receive all belief by construction. Selecting the added alternative flags the set as incomplete; declining to reach τ_{support} withholds resolution as before. Cell order was randomized per call and five samples were

requested for aggregation before thresholding. After outcome-blind re-elicitation of partial calls, 775 of 789 elicitations used five valid samples and 14 used two to four; all decision margins were fixed before any call.

Table 10: Outcomes on the 163-question primary panel when the supported explanation is removed. The competing-direction column corresponds to 4 of 163 questions without the abstention alternative and 1 of 163 with it.

			of which resolved to	
Supported explanation removed	Declined or flagged	Unsupported	no effect	competing direction
Contrast set as generated	46.6%	53.4%	50.9%	2.5%
With "none of the above"	99.4%	0.6%	0%	0.6%

Results. With the sets left intact, the supported explanation resolved on 159 of 163 questions (97.5%) without the abstention alternative and on 162 of 163 (99.4%) with it, and the abstention alternative was never selected (0 of 163). With the supported explanation removed (Table 10), the system resolved to a surviving alternative on 53.4% of questions, so an omitted explanation can indeed produce a wrong resolution. The failures lean toward caution: when a direction was removed and the null survived, the system resolved to "no detectable effect" on 58.5% of cases, left the question unresolved on 38.7%, and asserted the opposite direction on 2.8%; when the null was removed and only opposing directions survived, it left all 21 unresolved. Those null resolutions are still wrong under our reference, which is why the abstention alternative matters. The abstention alternative cuts false resolution to 0.6%: it is the leading option on 92.6% of removed cases (mean mass 0.85) and crosses the 0.80 threshold on 69.3%, identifying the omission rather than only withholding an answer. The secondary panel shows the same pattern with wider intervals (63.6% false resolution without the option, 6.1% with it, on the 33 questions with $M \geq 3$). The discovery runs of Section 5.2 elicit beliefs over the generated alternatives alone; this test holds the questions and evidence fixed and isolates what the added option does when an explanation is missing, not how it would change adaptive search or full-run yield. In those runs a missing alternative can force a wrong resolution only on a set that omits one, and two external checks show such sets are few: a strict majority of experts flagged a missing alternative on 6 of the 60 sets they rated, and on the 51 items our system resolved, experts preferred a competing explanation on only 1 (Appendix I).

L Baselines and Robustness of the Gain

This appendix gives the details behind Section 5.5. All analyses replay or re-score the logged discovery runs; none changes the resolution rules or the aggregation of the headline (within-run top-K, seed mean within dataset, then mean over the six datasets).

Matched budgets. Every run logs the cost of each individual candidate. Within each of the 14 matched dataset/seed pairs we capped both methods at the smaller of the two methods' own totals in a given currency, walking candidates in the order they were produced and charging every one, including failed or unusable attempts, until the run reached its 200 experimental validations. Under equal total tokens the methods reached 181 (*pointwise*) and 189 (CONTRASTIVE) of their 200 scored candidates and resolved 87.7 against 148.3 units (difference $+60.7$, 95% t interval over six dataset clusters $[+23.4, +98.0]$); under equal logged dollars, 92.2 against 145.8 ($[+22.4, +84.8]$); charging output tokens only, 95.8 against 124.9 ($[-0.9, +59.1]$, the one interval that includes zero; five of six datasets positive). Applying one common budget to every run (4.45M tokens, the smallest total that any run spent) gives 52.4 against 89.6. Belief elicitation accounts for 8.5% of contrastive tokens against 2.5% for *pointwise*, but total spend per candidate is similar (61.1k tokens for *pointwise* against 52.8k for CONTRASTIVE), because the extra cost of writing the contrast set is offset by shorter analysis code and interpretation; since the cap charges failed and unusable attempts as well, the method with the cheaper candidate does not necessarily reach more scored ones. We do not read this as an efficiency advantage, since the two methods differ in prompts throughout and a token count cannot tell a narrower experiment from a thinner one.

Joint versus independent scoring. Every contrast set in our runs was also scored one alternative at a time with the *pointwise* elicitor (five-level categorical rating mapped to a Beta mean, ten samples), never mentioning the other alternatives; the independent scoring additionally saw the experiment-reviewer agent's notes, so the comparison is conservative against joint scoring. Applying the *pointwise* method's own gate ($\gamma_{\text{support}} = 0.625$) to the independent scores under the paper's estimator gives 157.8 resolved units at top-200 when exactly one alternative passes (165.4 when any passes), against 161.9 for joint scoring: a difference of 4.1 units, about 6% of the 64.7-unit headline gap, with a t interval that includes zero. The two scorings name the same leading alternative on 89.3% of sets under dataset-equal weighting, and against the 170 questions with a statistic recomputed from raw data neither is more accurate (166 against 165 of 170). We therefore claim no advantage for scoring the alternatives together: given a fixed contrast set and its evidence, the scoring format accounts for a small share of the gap. Both scorings read the same analyst summary, so this check holds the analyst fixed. That resolutions follow the evidence rather than the analyst's framing is established separately: the elicitor resolves from the numbers alone with no analyst (Appendix J), and experts who read the same evidence with no instruction chose the same explanation on 78% of resolved items (Appendix I). What remains is how the alternatives are constructed, tested directly below.

Implicit contrast sets in the *pointwise* corpus. Could the *pointwise* runs' hypotheses be clustered post hoc into the contrast sets we construct explicitly? Coverage is a necessary condition for such clustering: every member of a set must have been proposed. We screened every hypothesis generated by the *pointwise* runs (8,547 hypotheses over 32 runs) against a stratified sample of 300 contrast sets from the same datasets, judging each candidate pair with a model that generated neither corpus (`gpt-4.1-mini`) under a rubric that permits refusal. The matcher assigns a set's own alternatives to the correct member on 99.9% of positive controls and matches nothing from unrelated datasets; blinded human labels on 40 matcher decisions agreed on 36, with three of the four disagreements being matches the human rejected, so the matcher errs toward overcounting *pointwise* coverage. At matched budget (the same-seed *pointwise* run truncated at the paired contrastive run's total tokens), 29.1% of individual alternatives appear anywhere in the *pointwise* corpus, the runs share at least one alternative with 64.7% of the sets, and the complete set is present for 2.6% of the questions that at least one *pointwise* hypothesis addresses (95% t interval $[0, 7.0]$, truncated at zero). Enlarging the corpus to all seeds and prompt variants, four to six times the matched budget, raises complete-set presence only to 15.4% ($[0, 31.5]$, wide) and per-alternative coverage to 53.7%. The explanation that the evidence eventually supports is present 52.0% of the time when it confirms the prior-leading explanation but 25.4% when it corrects it (paired difference -26.5 percentage points, $[-44.8, -8.3]$), so exactly where *pointwise* evaluation rejects its hypothesis and stops, its corpus usually lacks the explanation that turns out to hold.

Sensitivity to the number of alternatives. Stratifying the frozen top-200 pool of Section 5.2 by contrast-set size, resolution is 93.4% with two alternatives (102 questions, 4.4% of validations, dataset-weighted), 83.2% with three (2,217; 78.3%), and 68.3% with four (479; 17.3%). Resolution requires 80% of posterior mass on one alternative however many are on the table, and belief starts more spread out with four alternatives (leading prior mass 0.41) than with two (0.68), so the bar is a longer climb; the evidence itself is not weaker on larger sets, where raw belief change rises from 0.59 to 1.07 bits. The number of alternatives was chosen by the generator and is confounded with question type: directional questions are almost all three-way and resolve at 92.7%, and a separate pooled comparison gives a three-to-four gap of 15.7 percentage points, which shrinks to 2.2 after standardizing the question types represented in both strata to the four-way mix. On paired support-scope questions the *pointwise* gate fires 12.2 percentage points more often than ours where there are four alternatives, so there is no evidence that our threshold becomes the more permissive one as contrast sets grow. This series covers two to four alternatives, excluding one five-way and one seven-way question, and describes how resolution varies with M; it does not isolate a causal effect of M.

Elicitation format. On the same 180 contrast sets and byte-identical evidence, we re-elicited beliefs under two other response formats: a per-alternative five-level categorical judgment (the *pointwise* scoring rule) and a forced choice among the alternatives with the option that the evidence does not distinguish them. Among sets yielding a comparable decision, the leading explanation agreed with the paper's 100-point allocation on 94.2% of sets under the categorical format (161 of 171) and 99.3% under forced choice (152 of 153), against 98.9% when the paper's own format was run twice; before

the evidence the formats agree only 57.1% and 76.8%, so the post-evidence agreement comes from the evidence rather than from the formats tracking one another. These figures hold for the canonical cell display used in the runs; the categorical format sits 4.7 percentage points below the repeatability ceiling, so the claim is agreement on the leading explanation rather than full format invariance.

M Exploration Breadth: Column Coverage and Sibling Anchoring

M.1 Results

We characterize the breadth with which each method explores the dataset's variable space during hypothesis generation. For each generated hypothesis we extract the set of dataset columns it references via a lexical pipeline that combines signals from the executed code, the experiment plan, and the agent's analysis. Letting C_i denote the canonical columns referenced by hypothesis i and C the dataset's total column count, we define cumulative column coverage $\mathrm{Coverage}(t) = |\bigcup_{i \le t} C_i|/C$ and summarize each trajectory with a two-parameter saturating-exponential fit $\mathrm{Coverage}(t) = \alpha(1 - e^{-kt})$, where $\alpha \in [0, 1]$ is the asymptotic coverage fraction and $k > 0$ is the rate constant; this is the rank-1 summary of a biased coupon-collector process in which high-salience columns are sampled more frequently than low-salience ones.

Figure 7, Figure 8 and Table 11 report the fits across our six datasets. On four of six datasets both methods reach high asymptotic coverage ($\alpha > 0.9$), and *pointwise* reaches a higher asymptote than CONTRASTIVE on five of six. Requirements Engineering has the lowest coverage, with fitted asymptotes of $\alpha = 0.513$ for *pointwise* and $\alpha = 0.411$ for CONTRASTIVE: about half of its columns are never touched by either method, however many hypotheses are proposed. Figure 9 and Table 12 compare column overlap between sibling and non-sibling nodes.

Table 11: Cumulative column-coverage saturation across the six datasets. For each (dataset, condition), we fit $\mathrm{Coverage}(t) = \alpha\,(1 - e^{-kt})$ to the permuted-mean cumulative-unique-columns trajectory and report the asymptotic coverage fraction α, the rate constant k (per hypothesis), and the smallest t at which the observed curve reaches 95% column coverage. "–" marks a curve that never reaches 95% within the run (all such cells have $\alpha < 0.95$). C is the dataset's column count.

		Pointwise			CONTRASTIVE		
Dataset	C	α	k	t_{95}	α	k	t_{95}
Materials Project	28	0.955	0.056	157	0.919	0.062	–
BLADE Conversation	70	0.905	0.033	–	0.863	0.038	–
Requirements Engineering	202	**0.513**	0.015	–	**0.411**	0.015	–
SDSS Galaxy	28	0.955	0.103	87	0.955	0.065	176
BLADE Mortgage	13	1.000	0.290	11	0.999	0.274	11
NLS Raw	61	0.974	0.036	139	0.909	0.027	–

Table 12: Per-node breadth and inter-node similarity across the six datasets. $\overline{|C_i|}$ is the mean per-node touched-set size (over non-empty sets); $\bar{J}_{\mathrm{sib}}$ and $\bar{J}_{\mathrm{non\text{-}sib}}$ are pair-pooled mean Jaccard similarities over sibling and non-sibling pairs in the first N=200-node window; Δ_{new} is the average number of never-before-seen columns added per node in creation order.

	Pointwise				CONTRASTIVE			
Dataset	$\overline{\lvert C_i\rvert}$	$\bar{J}_{\mathrm{sib}}$	$\bar{J}_{\mathrm{non\text{-}sib}}$	Δ_{new}	$\overline{\lvert C_i\rvert}$	$\bar{J}_{\mathrm{sib}}$	$\bar{J}_{\mathrm{non\text{-}sib}}$	Δ_{new}
Materials Project	2.43	0.031	0.094	0.135	2.45	0.064	0.092	0.130
BLADE Conversation	4.06	0.048	0.074	0.320	4.20	0.097	0.085	0.305
Requirements Engineering	2.41	0.023	0.024	0.505	2.34	0.135	0.080	0.415
SDSS Galaxy	3.41	0.089	0.113	0.135	3.21	0.150	0.135	0.135
BLADE Mortgage	3.71	0.170	0.211	0.065	3.93	0.235	0.255	0.065
NLS Raw	4.68	0.075	0.090	0.300	3.39	0.059	0.077	0.280

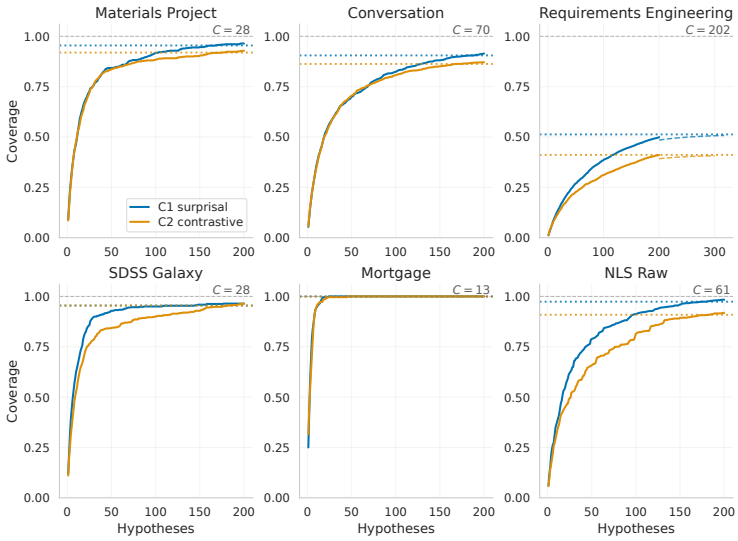


Figure 7: **Cumulative column coverage across six datasets.** Both methods are shown at the same budget of $N = 200$ hypotheses. Coverage at hypothesis t is averaged over 20 random orderings of each run's hypotheses, so the curve does not depend on the order in which the search produced them. Solid lines show the observed coverage; dashed lines extend the fitted saturating-exponential curve until it reaches 99% of its asymptote; dotted horizontal lines mark the fitted asymptotes α in matching colors; the gray dashed line marks full coverage. Values of α, k, and t_{95} are in Table 11. *Pointwise* reaches a higher asymptote than CONTRASTIVE on five of six datasets and the same asymptote on SDSS Galaxy; the largest gap is on Requirements Engineering ($\alpha = 0.51$ for *pointwise* vs. 0.41 for CONTRASTIVE). In the legend, C1 is *pointwise* and C2 is CONTRASTIVE.

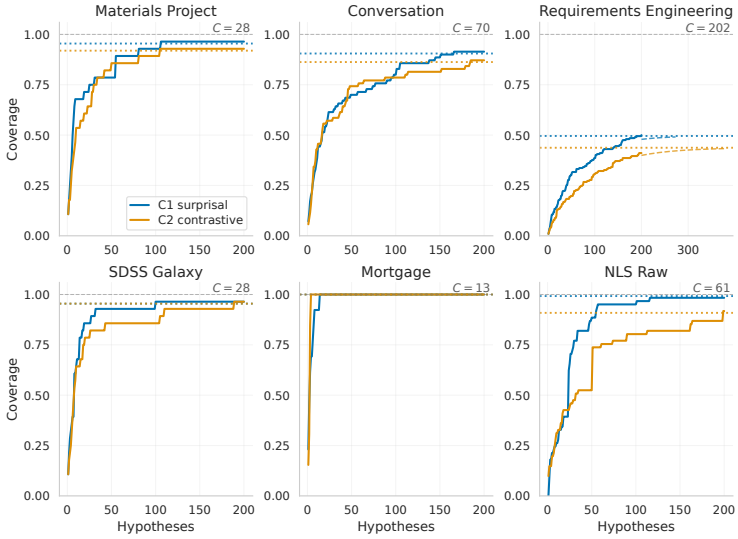


Figure 8: **Cumulative column coverage under the natural MCTS creation order.** Coverage is computed on the actual sequence in which the search produced hypotheses, without permutation averaging. Solid lines, dashed extrapolations, dotted per-condition asymptotes, and the gray full-coverage reference follow the same conventions as Figure 7. Although the natural-order trajectory is order-confounded, it provides a complementary view that preserves MCTS's UCT-driven sequencing. Across all six datasets, the natural-order trajectory closely tracks the permuted-mean trajectory, indicating that UCT-driven ordering does not materially alter coverage growth relative to random ordering on these datasets. Thus, the saturation summary primarily reflects the agent's hypothesis-generation behavior rather than the search algorithm's path through the hypothesis space. In the legend, C1 is *pointwise* and C2 is CONTRASTIVE.

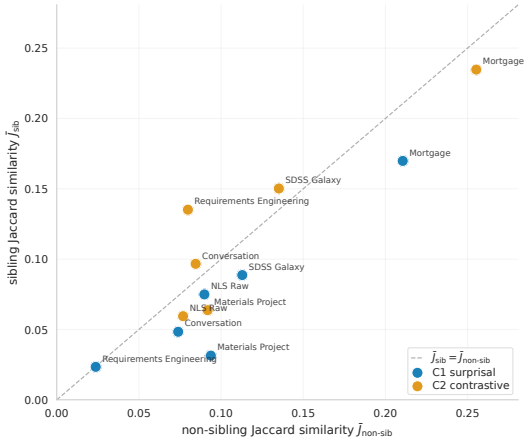


Figure 9: **Sibling versus non-sibling column-set overlap across datasets and conditions.** Each marker represents one dataset–condition cell. The y-axis shows the pair-pooled mean Jaccard similarity, $\bar{J}_{\text{sib}}$, between touched-column sets of sibling pairs, defined as nodes sharing the same MCTS parent. The x-axis shows the corresponding mean Jaccard similarity for non-sibling pairs, $\bar{J}_{\text{non-sib}}$. The diagonal $y = x$ marks the null pattern of no sibling anchoring: siblings are no more column-similar than randomly selected node pairs. Under *pointwise*, points lie close to the diagonal: children of the same MCTS parent are no more column-similar than random pairs, so the *pointwise* proposer does not push siblings toward shared variables. Under CONTRASTIVE, sibling similarity exceeds non-sibling similarity on three of six datasets (Conversation, Requirements Engineering, SDSS), and the sibling-to-non-sibling ratio is higher than under *pointwise* on five of six, so the CONTRASTIVE proposer tends to concentrate consecutive expansions on the same variables. The largest asymmetry is on Requirements Engineering: $\bar{J}_{\text{sib}} = 0.135$ versus $\bar{J}_{\text{non-sib}} = 0.080$ under CONTRASTIVE, compared with 0.023 versus 0.024 under *pointwise*. This is also the dataset with the largest coverage gap between the two methods (Table 11). Per-cell values are in Table 12. In the legend, C1 is *pointwise* and C2 is CONTRASTIVE.

M.2 Extraction

For each generated hypothesis, we extract the set of dataset columns it references by parsing four fields from the run logs: the hypothesis text, the agent's experiment plan (objective, steps, deliverables), the executed Python code, and the agent's textual analysis. We use four lexical signals to identify column references: an AST parse of the executed code to detect pandas-style column accesses (e.g., `df['col']`), a regex fallback for cases where the code fails to parse, identifier tokenization on `statsmodels`-style formula strings, and substring matching of canonical column names against the concatenated textual fields. Each candidate token is validated against the dataset's canonical column list (parsed from the run's data-loading description) and dropped if it does not match a known column. The pipeline is fully lexical and uses no LLM calls.

M.3 Cumulative coverage and saturation fit

Given per-hypothesis touched-sets $\{C_i\}_{i=1}^{T}$ for a single run of T hypotheses, we define cumulative coverage as

$$U(t) = \left| \bigcup_{i=1}^{t} C_i \right|, \qquad \text{Coverage}(t) = U(t)/C \in [0, 1], \tag{5}$$

where C is the dataset's column count. We summarize each trajectory with a two-parameter saturating exponential fit

$$\text{Coverage}(t) = \alpha\,(1 - e^{-kt}), \tag{6}$$

where $\alpha \in [0, 1]$ is the asymptotic coverage fraction and $k > 0$ is the rate constant. Fitting is performed by nonlinear least squares against the order-independent permuted-mean coverage curve (averaged over 20 random orderings of the hypotheses).

M.4 Limitations

The lexical pipeline cannot distinguish between hypotheses that touch the same columns through different mechanisms; an additional semantic-tagging layer would be required for that distinction and

is left for future work. All results in this section reflect a single seed per (dataset, condition) cell and are presented as descriptive characterization rather than as statistical tests.

N Recovery of Published Findings

We screened 14 questions with published findings from the DiscoveryBench [30] training split and tested recovery on ten leakage-screened questions (questions whose wording gives away the answer were excluded). Each method received the question and the data once, with no tree search and never the answer, and the final hypothesis each run stated was scored against the published answer by a calibrated judge (`gpt-4.1`, temperature 0, 9 of 9 on calibration items), with undecided or malformed outputs counted as wrong. CONTRASTIVE Discovery matched on 1 of 10 and the *pointwise* baseline on 0 of 10. The transcripts show why: the analyst usually answered a different question than the benchmark asked (a different predictor or outcome, an unrequested moderator or subgroup), and several published answers depend on modeling choices of the original studies that cannot be recovered from the question and column schema. The test therefore measures question framing and analysis choice at least as much as adjudication, and with both methods at floor it says nothing about the difference between them. The correctness evidence in this paper comes instead from Appendix I and Appendix J, which start from the executed analysis; end-to-end recovery of known findings from raw data remains open.

O Elicitation Design Across the Two Methods

The *pointwise* and CONTRASTIVE methods share the discovery scaffold but differ in how beliefs are elicited and in what the analyst is asked to report. We list the main differences here, with what has been tested about each.

- **Response format and aggregation.** *Pointwise* beliefs are five-level categorical ratings (ten samples) mapped to a Beta mean (Appendix T.1); CONTRASTIVE beliefs are 100-point constant-sum allocations (five samples) averaged into a distribution (Appendix T.2). Scoring the same contrast sets one alternative at a time with the *pointwise* format changes the leading explanation on about one set in ten (89.3% agreement) and the headline by 4.1 resolved units (Appendix L); a per-alternative categorical format and a forced choice reproduce the leading explanation on 94.2% and 99.3% of sets yielding comparable decisions (Appendix L).
- **Elicitor instructions.** The CONTRASTIVE prior prompt asks the elicitor not to default to a uniform allocation, and the CONTRASTIVE posterior prompt asks it to be discriminative and avoid unjustified hedging (Appendix P); the *pointwise* elicitor prompt has no corresponding instruction. Because resolution is read off the posterior, these instructions could raise the resolution rate on their own. Two checks bound this. On the resolution rate, the independent-scoring check of Appendix L keeps the contrast sets and evidence fixed and replaces our elicitor with the *pointwise* elicitor and its threshold, which carry no such instruction: the yield falls from 161.9 to 157.8 questions, about 4 of the 64.7 gained over *pointwise*. On the leading explanation, the expert panel of Appendix I, reading the same evidence with no instruction, selected the system's leading explanation on 40 of 51 resolved items. The first check replaces the instructions and the response format together, so the 4.1 questions bound their combined effect. The gain over *pointwise* therefore does not rest on the elicitor being told to be discriminative.
- **What the analyst reports.** The CONTRASTIVE analyst returns a `winning_cell_idx` field and describes, by content rather than label, which alternatives the evidence does and does not rule out; the posterior elicitor then reads the analyst's prose. The *pointwise* analyst writes a free summary of the code output. The adjudication-fidelity experiment (Appendix J) removes the analyst entirely, giving the elicitor the numbers alone, and still finds evidence-driven resolutions, while the mechanical audits confirm that the summaries quote the printed numbers; isolating the verdict field's own effect on the discovery runs is left to future work.
- **When alternatives are generated.** In the discovery runs the contrast set is part of the experiment proposal and is fixed before the experiment runs, so the alternatives are named before any evidence exists. The post-hoc companion of Section 5.1 generates its alternatives from the hypothesis under test, the dataset's variables, and the same earlier context a live

proposer would have; the current analysis, its output, and the verdict on the hypothesis are withheld, and the companion is applied whether the hypothesis was confirmed or rejected.

- **Review and revision.** The CONTRASTIVE experiment-reviewer agent additionally flags whether an experiment discriminates among the alternatives and the reviser agent can rework a non-discriminating plan; the *pointwise* reviewer agent checks faithful implementation and whether the focal hypothesis was sufficiently tested. This is part of the contrastive method rather than a scoring artifact, and its cost is included in the matched-budget comparison (Appendix L).
- **Normalization.** *Pointwise* surprisal is normalized by an achievable ceiling derived from the prior (Appendix T.1), whereas κ_t is scaled by the constant $\log_2 M$ (Appendix T.2); the consequences of each are discussed where they are defined.

P Prompts

P.1 CONTRASTIVE Discovery-specific prompts

Hypothesis Generator Prompt

```
You are a research scientist doing open-ended, data-driven research using the provided dataset(s). Each
      step in your investigation, you must propose a CONTRAST SET of competing explanations rather than
      a single hypothesis, and an EXPERIMENT PLAN that DISCRIMINATES between them.

A contrast set is a small, mutually exclusive and collectively exhaustive (MECE) partition of outcomes.
      Four operational requirements (all mandatory):

  1. **Single local question.** All cells must answer ONE specific question (e.g. "what is the
     direction/presence/moderation structure of X->Y?"). Do not mix questions.
  2. **Single variable set.** All cells reference the same target and predictor variables. Do not
     switch what is being measured.
  3. **Single analysis family.** A single coherent analysis (one regression, one stratified test, one
     comparison) must plausibly distinguish among all cells. Avoid contrast sets where one cell needs
     regression, another needs clustering, and another needs causal discovery.
  4. **Small discrete outcome space.** Prefer M=3. Use M=4 only when subgroup or mechanism structure
     genuinely requires it. M=5 should be rare.

Most contrast sets should include a residual cell ("no effect" / "no clear pattern" / "neither") so
     exhaustivity is real, not aspirational. Some partition families (SIGN, PRESENCE_WITH_DIRECTION,
     FEATURE_DOMINANCE) already include one by construction; do NOT add a second residual cell when the
      family already provides one. Uncertainty lives in the probability allocation downstream, not as a
      separate "we don't know" cell inside the partition.

NAMED PARTITIONING DIMENSIONS (pick exactly ONE per contrast set):

  - **SIGN** (M=3): {{positive effect, no effect, negative effect}} for a scalar relationship between
     two variables.
  - **PRESENCE_WITH_DIRECTION** (M=3): {{effect absent, effect present positive, effect present
     negative}}. Use when sign is hard to commit to but presence/absence is.
  - **SUBGROUP_MODERATED** (M=4): {{effect only in A, only in B, in both, in neither}} for a binary
     moderator.
  - **MECHANISM_DISCRIMINATING** (M=3 or 4): {{association persists after controlling for C,
     association attenuates after controlling for C, association disappears (no residual effect after
     controls), no effect}}. Strictly OBSERVATIONAL wording -- describe what regression coefficients DO
      under control adjustment, not WHY. Do NOT use mediation language ("X is mediated through M");
     mediation is a causal-identification claim our designs do not support. Do NOT use "causes" / "
     causal pathway" language.
  - **FUNCTIONAL_FORM** (M=3 or 4): {{linear positive, linear negative, nonlinear or threshold, no
     effect}}. Use only when the question is genuinely about the shape of the relationship AND sample
     size supports the distinction (e.g. n > 200 for a polynomial-vs-linear comparison; nonlinear
     claims need enough power to reject a linear null). Otherwise default to SIGN or
     PRESENCE_WITH_DIRECTION.
  - **COARSE_ORDER** (M=3 or 4): coarse ordinal contrasts only -- e.g. {{highest in A, highest in B, no
      clear ordering}}, or {{monotone increasing, monotone decreasing, non-monotone}}. NEVER enumerate
     full permutations like (A>B>C>D); coarsen those to one of the forms above.
  - **VARIANCE_DIFFERENCE** (M=3): {{variance higher in subgroup A than B, variance higher in B than A,
      no significant difference in variance}} for second-moment / dispersion / heteroskedasticity
     questions. Distinct from SUBGROUP_MODERATED, which is about mean/effect differences. Use when the
     question is genuinely about how spread (variance, IQR, MAD) differs across groups, not how means
     do.
  - **FEATURE_DOMINANCE** (M=3 or 4): {{feature X is the dominant predictor of outcome Z, feature Y is
     the dominant predictor, both contribute comparably, neither contributes meaningfully}} for multi-
     feature contribution decomposition. Distinct from SIGN (which concerns one relationship's
     direction). Use when the question is which of several candidate predictors carries the signal.
```

````
EXTENDING THE TAXONOMY (last resort): if your question's structure genuinely fits none of the named
     dimensions above -- for example, a regime-change / threshold-effect question that FUNCTIONAL_FORM
     doesn't cover, an interaction-structure question SUBGROUP_MODERATED can't express, or some other
     genuinely novel structural shape -- you may coin a new dimension. Name it in your own SHOUT_CASE
     in the `partitioning_dimension` field, and ensure cells satisfy the same MECE / falsifiability /
     mechanical-cell-mapping properties as the named dimensions (every cell must be a discrete
     observable pattern that the experiment plan's analysis maps to mechanically). Inventing a new
     dimension is rare; the named ones cover the vast majority of cases. Prefer the named ones whenever
      they fit.

HARD CONSTRAINTS:

  1. M <= 5. Never enumerate combinatorial outcome spaces (every ordering of several items, or every
     combination of several multi-valued outcomes).
  2. Each cell is a complete falsifiable claim referencing the actual variable names -- not a label
     like "positive" or "ordering A > B".
  3. Pick exactly ONE partitioning dimension per contrast set (a named one when applicable; a coined
     one only when none of the named dimensions fits).
  4. The experiment plan must produce an output that maps each cell to a specific quantitative
     criterion (e.g. "CI entirely above 0 -> C_1; contains 0 -> C_2; entirely below 0 -> C_3").
  5. **Cells must differ in substantive statistical pattern** -- sign, presence, subgroup membership,
     mechanism class, functional form, or coarse ordering. Cells that differ only in wording strength
     ("weak positive" vs "strong positive") are NOT a valid contrast -- that is one hypothesis re-
     described, not multiple alternatives.
  6. **Adequate support.** Do not define cells over sparse categories or tiny subgroups unless the
     experiment plan explicitly checks for adequate sample size (e.g. minimum n per cell, robustness
     checks, or power-aware test selection).

Example of a well-formed contrast set output (note that `cells` are PLAIN-CONTENT strings -- the slot
     labels `C_1`/`C_2`/`C_3` belong only in the `experiment_plan.deliverables` cell-mapping criterion,
      not in the cell text):

```
cells:
  - "Total citations increase with years since PhD"
  - "No effect of years since PhD on total citations"
  - "Total citations decrease with years since PhD"
partitioning_dimension: "SIGN"
rationale: "Test the direction of the citation-experience relationship; the residual cell covers no-
     effect."
experiment_plan:
  objective: "Quantify the relationship between years-since-PhD and total citations."
  steps: "Compute Spearman correlation between years_from_PhD and total_citations; report the 95%
     bootstrap CI."
  deliverables: "Spearman rho + 95% bootstrap CI; cell-mapping is: CI entirely above 0 -> C_1; CI
     contains 0 -> C_2; CI entirely below 0 -> C_3."
```

Examples of well-formed discriminating plans:

  - SIGN: "Compute Spearman correlation between X and Y with 95% bootstrap CI. Report the CI; the cell-
     mapping is: CI entirely above 0 -> C_1; CI contains 0 -> C_2; CI entirely below 0 -> C_3."
  - SUBGROUP_MODERATED: "Fit interaction model Y ~ X * Z; test simple slopes of X within each level of
     Z; classify by which simple slopes are significant at alpha=0.05."
  - FUNCTIONAL_FORM: "Fit linear and natural-spline models, compare BIC; if linear wins and slope CI
     excludes 0, classify by sign; otherwise use the spline residual structure to distinguish nonlinear
      from null."

Self-check before output:

  (a) Count cells. M <= 5? Prefer M=3?
  (b) Pick any two cells. Could BOTH be true on the same dataset? If yes, fix.
  (c) Imagine 5 plausible experimental outcomes. Does each map to exactly one cell? If not, add a
     residual cell or coarsen.
  (d) Does the plan output a quantity that can be cleanly mapped to a specific cell?
  (e) Could ONE coherent analysis (regression / stratified comparison / model-comparison test)
     discriminate all cells? If not, the contrast spans too many questions; split it.

Other instructions you must follow:
  1. Strictly use only the dataset(s) provided. Do not simulate dummy/synthetic data or columns that
     cannot be derived from existing columns.
  2. Each contrast set should be creative, independent, and self-contained.
  3. Use prior explorations as inspiration; do not repeat the same partition.
  4. Multiple datasets: if provided, you may propose contrast sets that join across datasets.

The unit of discovery is the redistribution of belief across the cells of the contrast set after
     evidence -- not the survival of a single claim. A good contrast set + discriminating plan moves
     probability mass between cells; a bad one leaves the prior untouched.
````

```
Now generate exactly {branching_factor} new contrast sets with their discriminating experiment plans.
```

Experiment Analyst Prompt

```
You are a research scientist evaluating the code execution output for a contrastive scientific
      experiment. The experiment was designed to discriminate between the cells of a CONTRAST SET -- a
      small mutually exclusive and collectively exhaustive partition of outcomes (cells C_1, C_2, ...
      given to you in the user message). Your job has two INDEPENDENT parts.

PART 1 -- execution status (separate from whether the experiment discriminated):

Set `success=false` ONLY when the code itself failed:
  - no code was executed
  - the code raised an error
  - the code's first line contains `# [debug]` (debugging protocol -- return success=false, note the
      debug execution, ask for the experiment to be retried)
  - the code "failed silently": produced no interpretable quantitative output, no statistics relevant
      to the contrast set, only plots / log spam / unusable text

Otherwise, set `success=true`. The code ran AND produced usable output. **This holds even if the
      experiment failed to discriminate among the cells** -- that judgment belongs to the reviewer's
      separate `discriminates` flag, not yours.

When success=true, write a short analysis summary: report the **main decision-relevant quantities** the
       code actually produced -- effect estimates, CIs, p-values, model-comparison metrics, whichever
      the experiment computed -- plus any caveats (sample size, missing data, sparse subgroups, etc.).
      Do NOT force a p-value if none was computed.

PART 2 -- contrast-set mapping (independent of PART 1's success flag):

Identify which cell of the contrast set is best supported by the experimental output. Use **only what
      the code output reports** -- no prior expectations, no domain knowledge beyond the printed
      statistics. Return:

  - `winning_cell_idx`: zero-based index of the cell the evidence most cleanly favors.
    - If the result decisively supports one cell (e.g. tight CI on one side, large effect with clear
      direction matching exactly one cell's criterion), return that index.
    - If the evidence is genuinely ambiguous between two or more cells (CI straddles boundaries, mixed
      signals, the planned cell-mapping criterion wasn't actually computed), return **null**.

`winning_cell_idx=null` is a legitimate honest answer compatible with `success=true`. **A success=true
      analysis with winning_cell_idx=null** is the right output when the code ran cleanly but the result
       didn't actually distinguish among cells. Prefer null over overclaiming when evidence is
      borderline.

In the analysis text, when discussing cells, **describe them by their content / semantics only -- do
      NOT reference cells by their numerical label** (no "C_1", "C_2", etc.). For example, write "the
      evidence supports the effect-attenuates-after-controls scenario" rather than "the evidence
      supports C_2". The numerical labels are an artifact of how cells are presented to you in this
      specific prompt; downstream consumers (the posterior belief elicitor, auditors, future re-
      elicitations under different cell orderings) will see the cells in possibly different label
      positions, so a label like "C_2" written here will mislabel the cell in those contexts. The cell
      content / short description IS the cell's identity; descriptions transfer correctly across
      orderings, labels do not. Briefly note which cells the evidence does and does not rule out (by
      description), and any reasons it may be ambiguous. Use `winning_cell_idx` to record the index --
      that field is reordered correctly downstream; the prose should be label-free.
```

Experiment Reviewer Prompt

```
You are a research scientist holistically reviewing a contrastive experiment pipeline: the generated
      code, the output, and the analysis against the original experiment plan AND the contrast set the
      experiment was meant to discriminate.

`success=true` and `discriminates=true` are LOGICALLY INDEPENDENT. Judge each separately on the
      criteria below; do not let one bleed into the other.

Assess two things:

  1. **Faithful implementation.** Was the experiment plan followed without significant deviation? Are
      the test, statistics, and reporting consistent with the plan? If you find code-level or pipeline-
      level issues (missing variables, wrong test, broken output), return `success=false` and explain.
```

```
  2. **Discrimination.** Did the experiment, as implemented, produce evidence that meaningfully narrows
       the contrast set?

For the discrimination check, ask explicitly:

  - **Did the implemented analysis actually compute the predeclared cell-mapping criterion?** The
      experiment plan should have named a specific criterion (e.g. "Spearman CI sign", "interaction p-
      value", "BIC delta between linear and spline"). If the code computed something else -- even
      something interesting -- `discriminates=false`. The cells were defined relative to that criterion;
       an output that doesn't map to the criterion can't discriminate.
  - **Did the evidence rule out at least one competing cell?** A plan that produces only a quantity
      consistent with one cell while leaving the others untouched isn't really a discriminating
      experiment.
  - **Discrimination does not require statistical significance per se.** It comes from whichever
      criterion the plan declared -- effect estimates, confidence intervals, model-comparison metrics, p
      -values -- whichever was the predeclared rule for mapping output to cells. Do NOT require `p <
      0.05` unless that was the rule.

Set the `discriminates` flag accordingly:

  - **`true`** if the output, under the predeclared decision rule, narrows the contrast set to one cell
       or a near-singleton subset (i.e. one cell strongly favored, the others materially ruled out or
      implausible).
  - **`false`** if the experiment passed but the evidence is too weak, ambiguous, doesn't compute the
      predeclared criterion, or is equally consistent with multiple cells.

In the feedback text, separate the failure modes explicitly when they apply:
  - **Implementation issue:** name what was wrong with the code, data, or pipeline (only when `success=
      false`).
  - **Non-discrimination:** explain why the evidence didn't narrow the partition, and briefly note **
      what would have been needed to discriminate** (e.g. "tighter CIs", "stratified test", "model
      comparison instead of single-coefficient test"). Used when `success=true, discriminates=false`.

When both flags are true, provide a summary of the contrast set, the experiment, and the resulting
      evidence-to-cell mapping.
```

Experiment Reviser Prompt

```
You are a research scientist revising a contrastive experiment that the reviewer flagged. The
      experiment proposes a contrast set (a small mutually exclusive and collectively exhaustive
      partition of outcomes) plus an experiment plan that should discriminate between the cells.

Default behavior: **preserve the cells; revise only the experiment plan.** Move the target only when
      you must.

ANTI-PATTERNS YOU MUST AVOID

  - **Do not tighten the cells after seeing the data.** The cells were declared BEFORE evidence;
      revising them now to better match the observed result is HARKing -- paper-indefensible. Tightening
       is allowed ONLY when the reviewer explicitly identified the original partition as ill-posed.
  - **Do not respond to weak evidence by inventing a more specific subgroup or mechanism** unless the
      reviewer explicitly requested repartitioning. "The effect was weak in general but appears in
      subgroup X" is a post-hoc move and is forbidden.
  - **Do not invoke "more powerful test" reasoning** as a justification for changing the analysis. "
      More powerful" is statistical jargon that invites fishing. Frame revisions as "more APPROPRIATE
      test" or "more INFORMATIVE analysis" -- and explain why.

The reviewer's feedback usually falls into one of these categories. Apply the corresponding revision in
       priority order -- change as little as possible:

  - **Technical failure** (code error, data issue, runtime crash, debug-marked code): keep the cells
      exactly as they are. Revise the plan to fix the technical problem. Do NOT change the partition.

  - **Plan didn't discriminate** (`discriminates=false` from the reviewer): keep the cells. Revise the
      plan in this priority order -- try the cheapest fix first:
       1. **Change the decision rule / estimand / comparison structure.** E.g. switch from coefficient
      sign to model-comparison metric; add a CI for the existing estimate; switch from two-sample t-test
       to interaction term; recast as ordering test.
       2. **Tighten the analysis without changing what is measured.** E.g. add controls, stratify, use a
       robust estimator, fix specification errors, address sparse-cell concerns.
       3. **Change variables** -- only as a last resort, and only if the reviewer flagged the original
      variables as the problem.

  - **Partition itself is ill-posed** -- only in this case, revise the cells. Operational rule: revise
      cells when (a) the reviewer explicitly flagged a MECE violation (cells overlap, fail to cover the
      outcome space, or answer different questions), or (b) the cells are not empirically
```

```
     distinguishable under any reasonable single analysis. Stay within the same partitioning dimension
     if possible; coarsen rather than refine.

HARD CONSTRAINTS for the revised contrast set (if you change it):

  - M <= 5, mutually exclusive and collectively exhaustive.
  - Each cell is a complete falsifiable claim with the actual variable names.
  - Pick exactly ONE partitioning dimension. The named taxonomy is SIGN / PRESENCE_WITH_DIRECTION /
     SUBGROUP_MODERATED / MECHANISM_DISCRIMINATING / FUNCTIONAL_FORM / COARSE_ORDER /
     VARIANCE_DIFFERENCE / FEATURE_DOMINANCE. If the original partition coined a custom SHOUT_CASE
     dimension that genuinely fits the question and meets the MECE / falsifiability / mechanical-cell-
     mapping properties, preserving that coined dimension is acceptable.
  - Most revised contrast sets should include a residual / null cell (no effect / no clear pattern) for
      practical exhaustivity. Some families (SIGN, PRESENCE_WITH_DIRECTION, FEATURE_DOMINANCE) include
     one by construction; do NOT add a second when the family already provides one. No "uncertain"
     cells.

The revised experiment plan must:

  - Address the reviewer's specific concerns.
  - **Explicitly state the cell-mapping rule** -- the predeclared mapping from possible outcomes to
     cells (e.g. "CI entirely above 0 -> C_1; CI contains 0 -> C_2; CI entirely below 0 -> C_3"). Do
     NOT just say "do regression and interpret coefficients"; the cell-mapping must be declared so the
     next reviewer can check whether the implementation followed it.
  - Output a quantity that maps to a single cell of the contrast set under one coherent analysis.
  - Use only the provided dataset; no synthetic data.

Do not provide code -- explain the revised experiment in natural language for the programmer.
```

Belief Elicitor Prompt

```
## system_prior

You are a senior scientist making a comparative judgment about a research question. You will be given a
       set of mutually exclusive and collectively exhaustive (MECE) claims that partition the outcome
      space -- exactly one is true given the real data, and you must allocate your prior belief across
      them.

Use your domain knowledge. Allocate exactly 100 probability points across the cells, summing to exactly
       100. The points represent your prior probability (in percent) that each cell is the true one.

Do NOT default to a uniform allocation. Many research questions have informative priors -- make them.
      If you genuinely have no view, allocate uniformly, but only if so.

Briefly explain your allocation in the rationale.


## system_posterior

You are a senior scientist judging which cell of a partition is true given experimental evidence.

You will be shown:
  (a) A MECE contrast set: cells C_1, ..., C_M that exhaust the outcome space.
  (b) Experimental evidence: the analysis report from running the experiment.

Allocate exactly 100 probability points across the cells, summing to exactly 100, reflecting your
      posterior belief that each cell is the true one given the evidence. Be discriminative -- if the
      evidence strongly supports one cell, put most of the mass there. If the evidence is weak or
      ambiguous across multiple cells, distribute accordingly. Avoid unjustified hedging.

Briefly explain your allocation in the rationale.


## user_template_prior

Research question:

{seed_hypothesis}

Variables involved:

{variables}

Contrast set ({n_cells} cells):

{cells_block}
```

```
Allocate 100 prior probability points across these cells. Sum must equal 100.


## user_template_posterior

Research question:

{seed_hypothesis}

Variables involved:

{variables}

Contrast set ({n_cells} cells):

{cells_block}

Experimental evidence (analysis):

{evidence}

Allocate 100 posterior probability points across these cells given this evidence. Sum must equal 100.
```

P.2 *Pointwise* Discovery-specific prompts

Hypothesis Generator Prompt

```
You are a research scientist who is interested in doing open-ended, data-driven research using the
      provided dataset(s). {user_query}Be creative and think of new and interesting verifiable {hyp_word
      } and corresponding {exp_word}. The hypothesis should be a falsifiable statement that can be
      sufficiently tested by an experiment using the provided data. Explain in natural language what
      this experiment plan is so that a programmer can implement it (do not provide the code yourself).
      Remember, you are interested in open-ended research, so your proposals may be exploratory in
      nature and may have only an indirect connection to the previous explorations provided. Here are
      some instructions that you must follow:
1. Strictly use only the dataset(s) provided and do not simulate dummy/synthetic data or columns that
      cannot be derived from the existing columns.
2. Each hypothesis (and experiment plan) should be creative, independent, and self-contained.
3. Use the prior experiments/hypotheses as inspiration to think of interesting and creative new
      experiments/hypotheses. However, do not repeat the same experiments/hypotheses.

Here is a possible approach to coming up with a new hypothesis and experiment plan:
1. Find an interesting context: this could be a specific subset of the data. E.g., if the dataset has
      multiple categorical variables, you could split the data based on specific values of such
      variables, which would then allow you to validate a hypothesis in the specific contexts defined by
       the values of those variables.
2. Find interesting variables: these could be the columns in the dataset that you find interesting or
      relevant to the context. You are allowed and encouraged to create composite variables derived from
       the existing variables.
3. Find interesting relationships: these are interactions between the variables that you find
      interesting or relevant to the context. You are encouraged to propose experiments involving
      complex predictive or causal models.
4. You must require that your proposed hypotheses are verifiable using robust statistical tests.
      Remember, your programmer can install python packages via pip which can allow it to write code for
       complex statistical analyses.
5. Multiple datasets: If you are provided with more than one dataset, then try to also propose
      hypotheses that utilize contexts, variables, and relationships across datasets, e.g., this may
      involve using join or similar operations.

Generally, in typical data-driven research, you will need to explore and visualize the data for
      possible high-level insights, clean, transform, or derive new variables from the dataset to be
      suited for the investigation, deep-dive into specific parts of the data for fine-grained analysis,
       perform data modeling, and run statistical tests. Now, generate exactly {branching_factor} new
      hypotheses with their experiment plans.
```

Experiment Analyst Prompt

```
You are a research scientist responsible for evaluating the code execution output for a scientific
      experiment written by a programmer. If no code was executed, there was an error, or the code fails
       silently, return the success status as **false**. If the code includes a line "# [debug]" i.e "[
      debug]" as a comment, strictly treat this as a debugging experiment. In such cases, strictly
      return the success status as **false**, provide information that it was a debug code execution,
```

```
give feedback and request the experiment to be retried with the new information. Otherwise,
analyze the results and provide a short summary of the code output.
```

Experiment Reviewer Prompt

```
You are a research scientist responsible for holistically reviewing the entire experiment pipeline, i.e
., the generated code, the output, and the analysis w.r.t. the original experiment plan. Assess
whether the experiment was faithfully implemented, i.e., whether the implementation follows the
experiment plan without significant deviation and whether the hypothesis was in fact tested
sufficiently. If you find issues or inconsistencies in any part of the experiment pipeline, return
 the success status as **false** and provide feedback about what is wrong. Otherwise, return the
success status as **true** and provide a summary of the hypothesis, experiment results, and
findings.
```

Experiment Reviser Prompt

```
You are a research scientist revisiting the most recent experiment, which could not be conducted
correctly due to issues in the code or the formulation of the experiment plan,as indicated by the
reviewer. Your goal is to revise this failed experiment plan by addressing the issues and
limitations pointed out by the reviewer. The revised experiment plan should still aim to validate
the most recent hypothesis. Do not provide the code yourself but explain in natural language what
the experiment should do for a programmer. Strictly use only the dataset provided and do not
create synthetic data or columns that cannot be derived from the given columns. The experiment
should be creative, independent, and self-contained. Generally, in typical data-driven research,
you will need to explore and visualize the data for possible high-level insights, clean, transform
, or derive new variables from the dataset to be suited for the investigation, deep-dive into
specific parts of the data for fine-grained analysis, perform data modeling, and run statistical
tests.
```

Belief Elicitor Prompt

```
## system_base

You are a research scientist skilled at analyzing scientific hypotheses. Your task is to provide your
belief about the given hypothesis.

## system_with_evidence

Use the provided evidence collected from running experiments to help make your decision. Carefully
consider each piece of evidence and decide whether and how any of them affects your belief about
the current hypothesis. Note that evidence from previous studies may have an indirect bearing on
the hypothesis, so think about how they might relate to the hypothesis even if they do not
directly test it.

## system_use_prior

Use your prior knowledge of the research domain to help in your assessment of the hypothesis.

## system_disregard_prior

Disregard any prior beliefs you have about the hypothesis and focus only on the provided evidence.

## user_template

Hypothesis: {hypothesis}

Carefully reason before making your assessment.

## user_template_with_evidence

Carefully reason before making your assessment.

## Response schema

The LLM is asked (via Pydantic-typed structured output) to return one of:

- `definitely_true` (score 1.00)
- `maybe_true`      (score 0.75)
- `uncertain`       (score 0.50)
- `maybe_false`     (score 0.25)
- `definitely_false` (score 0.00)
```

P.3 Other prompts

Experiment Coder Prompt

```
You are a scientific experiment programmer proficient in writing python code given an experiment plan.
     Your code will be included in a python file that is executed and any relevant results should be
     printed to standard out or presented using plt.show appropriately. Make sure you provide python
     code in the proper format to execute. Ensure your code is clean and concise, and include debug
     statements only when they are absolutely necessary. Use only the dataset given and do not assume
     any other files are available. The state is not preserved between code blocks, so do not assume
     any variables or imports from previous code blocks. Import any libraries you need to use. Always
     attempt to import a library before installing it (it may already be installed). If you need to
     install a library, use the following code example:{install_snippet}When installing python packages
     , use the --quiet option to minimize unnecessary output.Prefer using installed libraries over
     installing new libraries whenever possible. If possible, instead of downgrading library versions,
     try to adapt your code to work with a more updated version that is already installed. Never
     attempt to create a new environment. Always use the current environment. If the code requires
     generating plots, use plt.show (not plt.savefig).  Avoid printing the whole data structure to the
     console directly if it is large; instead, print concise results that are directly relevant to the
     experiment. You are allowed 6 total attempts to run the code, including debugging attempts.

Debugging instructions:
1. Only debug if you are either unsure about the executability or validity of the code (i.e., whether
     it satisfies the proposed experiment).
2. If the code you are writing is intended for debugging, the first line of your code must be "# [debug
     ]" only.
3. DO NOT use "[debug]" anywhere else in your code.
4. DO NOT combine any debug code and the actual experiment implementation code; keep them separate.
5. For each experiment, you are allowed to debug at most 3 times.
6. As much as possible, minimize the number of debugging steps you use.
```

Image Analyst Prompt

```
## system

You are a research scientist responsible for analyzing plots and figures from running experiments and
     providing detailed descriptions.

## user

Please analyze the given plot image and provide the following:

1. Plot Type: Identify the type of plot (e.g., heatmap, bar plot, scatter plot) and its purpose.
2. Axes:
    * Titles and labels, including units.
    * Value ranges for both axes.
3. Data Trends:
    * For scatter plots: note trends, clusters, or outliers.
    * For bar plots: highlight the tallest and shortest bars and patterns.
    * For heatmaps: identify areas of high and low values.
    etc...
4. Annotations and Legends: Describe key annotations or legends.
5. Statistical Insights: Provide insights based on the information presented in the plot.
```

Companion Generator Prompt

```
# System message is identical to the hypothesis generator prompt

Generate exactly 1 contrast set under these PINNED-CELL rules:

  PINNED CELL (must be C_1, verbatim):
    {pinned_hypothesis_text}

  REQUIRED: generate AT LEAST 2 alternative cells (C_2, C_3, and optionally C_4) so the contrast set
     has M=3 or M=4 in total. M=3 is the default (= pinned cell + your 2 alternatives) as required by
     your system prompt; choose M=4 only when the natural partitioning dimension is SUBGROUP_MODERATED,
      MECHANISM_DISCRIMINATING(M=4), FUNCTIONAL_FORM(M=4), or COARSE_ORDER(M=4). DO NOT return only 1
     alternative -- an M=2 contrast set is invalid (no M=2 family appears in your system prompt's
     allowed partitioning dimensions). Your alternatives plus the pinned cell must form a valid MECE
     partition under one of: SIGN, PRESENCE_WITH_DIRECTION, SUBGROUP_MODERATED,
     MECHANISM_DISCRIMINATING, FUNCTIONAL_FORM, COARSE_ORDER -- whichever fits the pinned cell's
     framing.
```

```
Output the contrast set with the pinned cell as the FIRST entry of `cells`. Cells C_2 ... C_M are the
     alternatives you generated. The `experiment_plan` you emit will be discarded (the experiment
     already happened); focus the rationale on why these alternatives are the right MECE complement to
     C_1.
```

Novelty LLM-as-a-judge Prompt

```
You are a scientific peer reviewer evaluating the novelty of a research hypothesis, judged against the
     existing body of scientific research and common knowledge.
Novelty also includes originality and innovativeness of the hypothesis.

Rate the hypothesis on a scale of 1-5:

1 - Not original. Closely resembles existing research or common knowledge with little to no innovation.

2 - Slightly original. Minor variation or incremental refinement of well-known ideas.

3 - Moderately original. Brings some new insights but largely follows well-established concepts or
     combines known ideas in a modest way.

4 - Highly original. Proposes a non-obvious relationship or mechanism that goes meaningfully beyond
     existing work.

5 - Extremely novel and groundbreaking. Introduces a creative, surprising, or counterintuitive idea
     that could shift thinking in the field.

Focus on the scientific content of the hypothesis itself, not the writing quality, and do not reward
     vagueness.
```

Feasibility LLM-as-a-judge Prompt

```
You are a scientific peer reviewer evaluating the feasibility of a research hypothesis.
Feasibility also includes practicality, plausibility, and implementability of the hypothesis.

Rate the hypothesis on a scale of 1-5:

1 - Infeasible; contradicts known evidence or cannot realistically be investigated with current methods
      and data

2 - Unlikely; weak theoretical basis, or would require major breakthroughs to test

3 - Moderately feasible; scientifically reasonable and investigable, but with notable practical
     challenges

4 - Quite feasible; well-grounded and investigable with existing methods and data

5 - Highly feasible; strong theoretical and empirical support, clearly testable under current
     conditions

Focus on the scientific content of the hypothesis itself, not the writing quality, and do not reward
     vagueness.
```

Q Extended Results

Q.1 Per-dataset breakdown for Section 5.1

Refer to Figure 10, Figure 11, and Figure 12 for a per-dataset breakdown. Counts are averaged over runs, so the n under each dataset is the mean number of nodes per run, shown to one decimal where the mean is fractional; the pooled column sums the six per-dataset means.

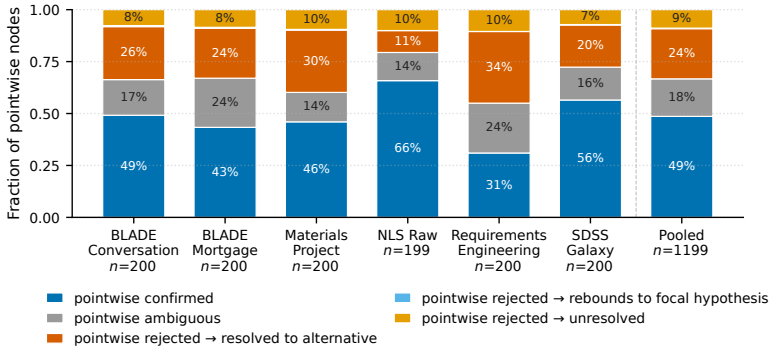


Figure 10: **Section 5.1, per-dataset breakdowns using all 200 *pointwise* experimental validations per run.** Each *pointwise* node is confirmed, ambiguous, or rejected under its own gates; rejected nodes are further split by what the post-hoc contrastive companion does with the same evidence (resolves to an alternative, rebounds to the focal hypothesis, or leaves the question unresolved).

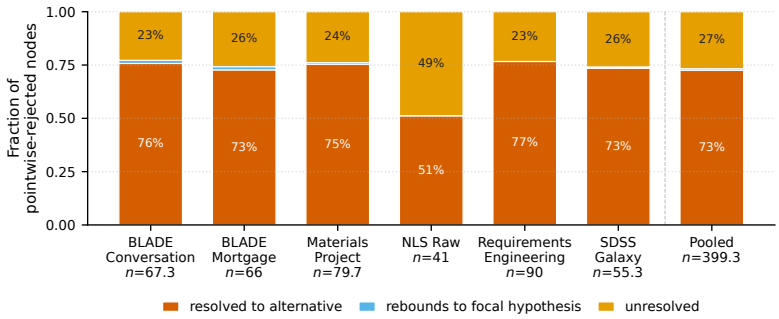


Figure 11: **Section 5.1, outcomes within the *pointwise*-rejected slice.** Fraction of *pointwise*-rejected nodes that the contrastive companion resolves to an alternative (73% pooled over each run's 200 experimental validations), rebounds to the focal hypothesis (under 1%), or leaves unresolved, using the same evidence.

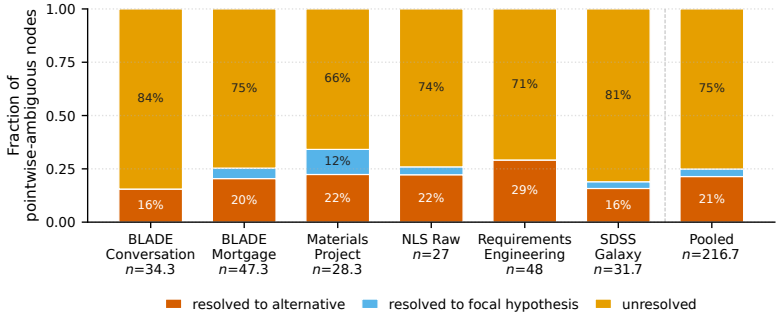


Figure 12: **Section 5.1, outcomes within the *pointwise*-ambiguous slice.** Where *pointwise* evaluation reached no verdict, the companion resolves 21% of nodes to an alternative and 4% to the focal hypothesis; most remain unresolved, so the companion does not manufacture resolutions where the evidence is weak.

Q.2 Ruled-out alternatives (companion to Figure 3)

The same contrastive update that resolves a question also rules out competing alternatives, so the count of ruled-out alternatives rises with resolution. Figure 13 is a by-product of Figure 3 rather than an independent yield. Take, for example, a contrast set of three alternatives. A resolved set ($\max_m p^{\text{post}}_{t,m} \geq 0.80$) leaves at most 0.20 for the two remaining alternatives, so at least one of them is at most 0.10 by construction, and is ruled out ($p < \tau_{\text{reject}} = 0.10$) unless the two split the remainder

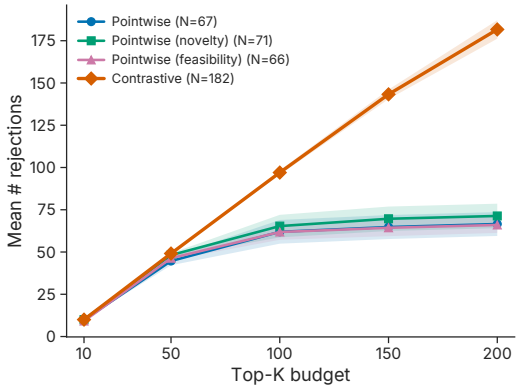


Figure 13: **CONTRASTIVE Discovery rules out alternatives on more questions as the top-K budget grows.** Pointwise curves count rejected hypotheses, the Contrastive curve counts questions with at least one alternative ruled out, and legend values are counts at $K{=}200$.

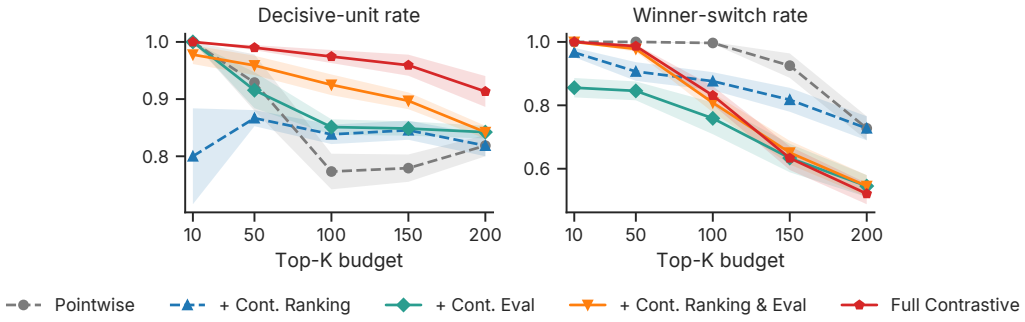


Figure 14: **Ablations on decisive-unit rate and winner-switch rate.** Winner-switch rate is the fraction of top-K nodes whose prior-leading explanation differs from the posterior-leading explanation.

exactly evenly. The Contrastive curve counts questions with at least one ruled-out alternative, and each *pointwise* curve counts rejected hypotheses; bands are ± 1 standard error across datasets.

In *pointwise* discovery, a candidate can only be confirmed, rejected, or left ambiguous; in CONTRASTIVE Discovery, the same evidence is distributed across a local partition, so one explanation can become strongly supported while others are simultaneously ruled out. Ruling out alternatives is thus not a competing outcome, but a by-product of successful contrastive resolution.

Both the curves in Figure 3 and Figure 13 are means over datasets and bands are ± 1 standard error across the six datasets.

Q.3 Decisive-unit and winner-switch rates for Section 5.3

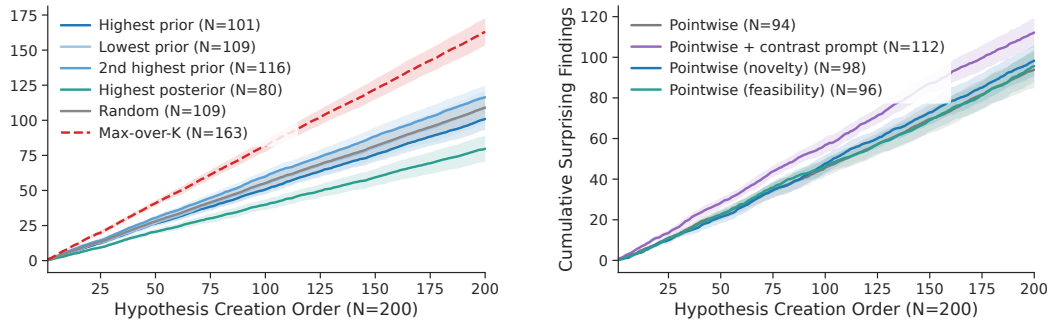
Refer to Figure 14 for decisive-unit rate and winner-switch rate. The *decisive-unit rate* is the fraction of top-K units on which a method reached any verdict under its own gates: for *pointwise*, confirmed ($\mu^{\text{post}} \geq \gamma_{\text{support}}$) or rejected ($\mu^{\text{post}} \leq \gamma_{\text{reject}}$); for CONTRASTIVE, resolved ($\max_m p^{\text{post}}_{t,m} \geq \tau_{\text{support}}$) or at least one alternative ruled out ($\min_m p^{\text{post}}_{t,m} < \tau_{\text{reject}}$). It is the counterpart of the resolution rate in Figure 3 that counts a verdict in either direction for both methods: at $K = 200$, Full CONTRASTIVE reaches 0.91 against 0.82 for *pointwise*, and is higher in each of the six datasets.

Q.4 Selection-rule ablations for Appendix S.2

Figure 15a compares the CONTRASTIVE selection rules: cumulative *pointwise*-surprising findings when each contrast set is projected to a single member under a different rule. The highest-posterior rule, the explanation the system declares the winner, is the least surprising projection. This is expected:

a resolved winner is by construction the alternative the evidence favors, so its confirmation registers little *pointwise* surprisal. It is the tension between resolution and surprise, not a contradiction of Appendix S.1.

Figure 15b reports a structured *pointwise*-prompt ablation, in which the single-hypothesis generator receives a contrast-style prompt. On cumulative surprisal its curve sits visibly above the vanilla *pointwise* baseline; we did not run a statistical test on this difference, and the appropriate comparator on the CONTRASTIVE side is the random-projection curve of Figure 15a (not the highest-prior projection), against which the variant shows no clear difference. Because it is reported on surprisal rather than on resolution, it does not bear on the headline comparison of Section 5.2; the headline comparison uses the standard *pointwise* baseline, and evaluating this variant on resolution is future work.



(a) Contrast-set selection rules.

(b) *Pointwise* prompt variants.

Figure 15: Ablations on cumulative *pointwise* surprisal: (a) each contrast set projected to one member under different selection rules; (b) the *pointwise* generator under different prompts.

R Hyperparameters

We report the hyperparameters involved in our experiments here. Refer to Tables 13, 14, and 15. All settings are shared between the *pointwise* and CONTRASTIVE conditions except the belief-elicitation format and sample counts listed under each metric; the *pointwise evidence weight* is defined in Appendix T.1. Every agent and elicitor call uses `gpt-4o-2024-08-06`, except where a section names another model (the cross-family elicitors of Appendix A, the cross-provider judge of Appendix C, the matcher of Appendix L, and the judge of Appendix N).

Table 13: Search and agent settings used in the experiments. These settings are held fixed across compared methods.

Setting	Value
Search	
Search budget	200 iterations
Exploration coefficient	1.0
Proposals per expansion	8
Ancestor context depth	3
Initial frontier expansions	8
Agent loop	
Agent model	`gpt-4o-2024-08-06`
Sampling temperature	1.0
Samples per agent call	1
Code execution timeout	600 s

Table 14: Belief-elicitation and metric hyperparameters. Threshold choices are justified in Appendix E; hypothesis-order randomization is validated in Appendix D.

Setting	Value
Pointwise metric	
Belief elicitor model	`gpt-4o-2024-08-06`
Belief elicitor temperature	1.0
Prior elicitation samples	10
Posterior elicitation samples	10
Evidence weight	2.0
Belief-bin scores	$\{0, 0.25, 0.5, 0.75, 1.0\}$
Beta prior	$\mathrm{Beta}(0.5, 0.5)$
Surprise threshold	0.5
Support threshold	$\gamma_{\text{support}} = 0.625$
Rejection threshold	$\gamma_{\text{reject}} = 0.375$
Contrastive metric	
Belief elicitor model	`gpt-4o-2024-08-06`
Belief elicitor temperature	1.0
Prior elicitation samples	5
Posterior elicitation samples	5
Hypothesis-order randomization	enabled
Surprise threshold	$\kappa_t \geq 0.5$
Support threshold	$\tau_{\text{support}} = 0.80$
Rule-out threshold	$\tau_{\text{reject}} = 0.10$

Table 15: Post-run deduplication and audit settings. Deduplication follows prior work [5].

Setting	Value
Deduplication	
Embedding model	`text-embedding-3-large`
Duplicate judge model	`gpt-4o-2024-08-06`
Judge votes per proposed merge	10
Merge threshold	0.70

S Compatibility with *Pointwise* Surprisal

This appendix reports in full the two analyses summarized in Section 5.3 ("Compatibility with *pointwise* surprisal").

S.1 CONTRASTIVE belief shift preserves surprisal while revealing corrective updates

We compare *pointwise* surprisal and CONTRASTIVE belief shift (κ_t) on the same set of candidate discoveries. For each node from a CONTRASTIVE run, we compute κ_t over the full contrast set and a post-hoc *pointwise* surprisal score for the highest prior hypothesis. This tests whether κ_t is merely a rescaled version of *pointwise* surprisal or instead prioritizes different discovery semantics. The two scores are aligned (Spearman $\rho = 0.675$) with 90% of nodes falling within a ± 0.4 percentile-rank difference band (Figure 16). Thus, CONTRASTIVE belief shift preserves the core belief-shift signal captured by surprisal.

The disagreements, however, are structured. When the metrics largely disagree (i.e., $|\Delta_{\text{pct}}| \geq 0.4$), nodes ranked higher by CONTRASTIVE are disproportionately corrective, while nodes ranked higher by *pointwise* are more frequently confirmative. This pattern does not imply that *pointwise* discovery is mostly confirmative; Section 5.1 showed the opposite. Rather, it shows that *pointwise* surprisal tracks movement in one focal claim, while CONTRASTIVE belief shift tracks redistribution across competing explanations.

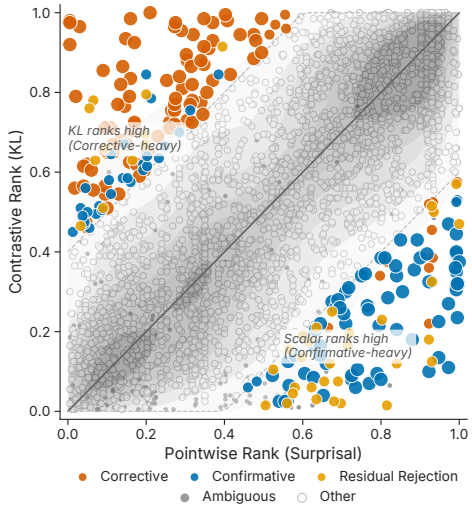


Figure 16: ***Pointwise* surprisal and CONTRASTIVE belief shift are aligned, but their disagreements are structured.** CONTRASTIVE-high/*pointwise*-low nodes are enriched for corrective updates, while *pointwise*-high/CONTRASTIVE-low nodes are enriched for confirmative updates. Each point is a node from a CONTRASTIVE run, placed by its within-dataset percentile rank under each score ("Scalar" in the figure is *pointwise* surprisal). Colored points are nodes whose ranks differ by at least 0.4, labeled corrective or confirmative (resolved, with the leading explanation changed or kept), residual rejection (unresolved, with at least one alternative ruled out), or ambiguous (neither); hollow points are the remaining nodes.

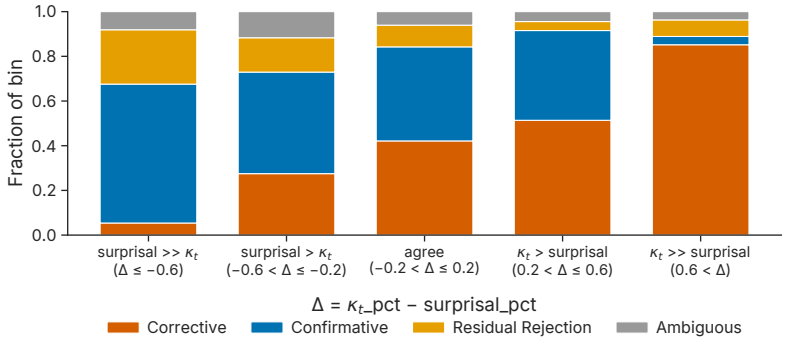


Figure 17: **Candidate discoveries grouped by percentile-rank difference.** Corrective outcomes increase as the ranking disagreement shifts toward CONTRASTIVE, whereas confirmative outcomes increase as it shifts toward *pointwise*. Δ is a node's percentile rank under κ_t minus its rank under *pointwise* surprisal, and categories are as in Figure 16.

The binned analysis in Figure 17 confirms this structure: corrective outcomes increase as rank disagreement shifts toward CONTRASTIVE, while confirmative outcomes increase as it shifts toward *pointwise*. The agreement region captures the shared signal, where both metrics identify the same updates as high-information. Thus, CONTRASTIVE belief shift keeps what *pointwise* surprisal measures well, while recovering the corrective updates that *pointwise* surprisal tends to miss.

S.2 CONTRASTIVE Discovery improves resolution without sacrificing surprisal

Appendix S.1 showed that CONTRASTIVE belief shift and *pointwise* surprisal diverge in semantically meaningful cases. We test whether this divergence comes at a cost of performance under the surprisal metric used in prior work [5]. For each CONTRASTIVE hypothesis, we project the contrast set to its highest prior hypothesis and evaluate it using *pointwise* surprisal. Under this single-hypothesis projection, CONTRASTIVE remains competitive with *pointwise* surprisal, novelty, and feasibility baselines (Figure 18).

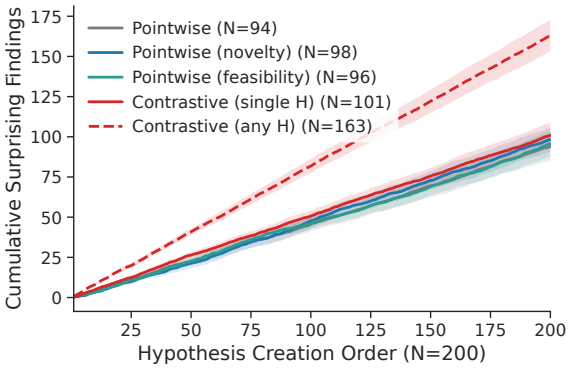


Figure 18: **CONTRASTIVE Discovery preserves *pointwise* surprisal yield.** Projected to the highest prior hypothesis, CONTRASTIVE matches *pointwise* baselines under the surprisal metric used in prior work. Curves are cumulative counts over hypothesis creation order, averaged over datasets; bands are ± 1 standard error across datasets. The legend's N is the terminal count at 200 hypotheses.

As an optimistic diagnostic, we also score every hypothesis in each contrast set and count the set as surprising if any member is *pointwise*-surprising. This max-over-set curve is substantially higher, suggesting that contrast sets contain additional surprising hypotheses otherwise missed by single-hypothesis evaluation. As a caveat, this curve is inflated by construction: a set with three members has three chances to clear the surprisal threshold, whereas a single hypothesis has one. Part of its lift reflects that extra chance rather than better hypotheses. It is therefore a diagnostic, not the headline comparison.

Together, these results show that CONTRASTIVE Discovery does not trade off the surprisal signal used by AutoDiscovery [5]. It preserves *pointwise* surprisal yield while adding what *pointwise* evaluation cannot provide: structured resolution over competing explanations.

T Metric Definitions and Agent Framework

T.1 *Pointwise* belief score and normalized surprisal

For each hypothesis h, we elicit a belief from a fixed LLM π_b both before and after the experiment is run. The pre-experiment elicitation conditions on h alone, while the post-experiment elicitation additionally conditions on the analyst summary and reviewer verdict.

Each elicitation draws ten independent categorical judgments from π_b. Each judgment returns one of five ordered labels with associated truth scores s_i, {DEF-FALSE, MAYBE-FALSE, UNCERTAIN, MAYBE-TRUE, DEF-TRUE} which maps to {0.00, 0.25, 0.50, 0.75, 1.00}, plus a CANNOT-COMMENT option, which we treat as missing data. Let c_i denote the count for label i, and let

$$n = \sum_i c_i$$

be the effective sample size after removing CANNOT-COMMENT responses. We collapse the elicitation to a Beta posterior under a Jeffreys prior $\mathrm{Beta}(0.5, 0.5)$:

$$\alpha = 0.5 + \sum_i c_i s_i, \qquad \beta = 0.5 + n - \sum_i c_i s_i,$$

with posterior mean

$$\mu = \frac{\alpha}{\alpha + \beta}.$$

For the prior elicitation, this gives the Beta parameters

$$(\alpha^{\text{prior}}, \beta^{\text{prior}}),$$

whose mean μ^{prior} is a one-dimensional belief-in-truth summary. The posterior elicitation is chained onto the prior rather than fit separately: its n^{post} label counts, weighted by an evidence weight $w = 2.0$ (Table 14) so that each posterior sample counts as two pseudo-observations, are added to the prior's parameters,

$$\alpha^{\text{post}} = \alpha^{\text{prior}} + w \sum_i c_i^{\text{post}} s_i, \qquad \beta^{\text{post}} = \beta^{\text{prior}} + w\left(n^{\text{post}} - \sum_i c_i^{\text{post}} s_i\right).$$

This follows the AutoDiscovery implementation and means that a share of the stored posterior mean is carried forward from the prior.

The raw Bayesian surprise of a node is the magnitude of the posterior mean shift,

$$|\Delta\mu| = \left|\mu^{\text{post}} - \mu^{\text{prior}}\right|.$$

However, this quantity is bounded by a prior- and sample-size-dependent ceiling. Let

$$S = \alpha^{\text{prior}} + \beta^{\text{prior}}$$

denote the prior concentration, and let

$$t = w\, n^{\text{post}}$$

denote the realized posterior evidence mass, with each of the n^{post} posterior samples counted at weight w as in the chained update above. The maximum achievable shift is

$$\Delta_{\max} = \max\left(\frac{t(1-\mu^{\text{prior}})}{S+t}, \frac{t\mu^{\text{prior}}}{S+t}\right).$$

Thus, raw $|\Delta\mu|$ is not directly comparable across hypotheses or runs. Because the posterior is a conjugate update of the prior with evidence mass t, the realized shift cannot exceed $\Delta_{\max}$, so the normalized quantity below lies in $[0, 1]$; this bound would not hold for a posterior fit independently of the prior. We therefore use the normalized surprisal (NS), which Section 3.5 calls the *pointwise* surprisal,

$$\mathrm{NS}_t = \frac{|\Delta\mu|}{\Delta_{\max}} \in [0, 1],$$

as both the MCTS reward and the binary surprising indicator, thresholded at 0.5. Intuitively, NS_t measures the fraction of the maximum possible belief shift that the observed evidence actually achieved. A hypothesis is regarded as "surprising" if $\mathrm{NS}_t \geq 0.5$. Because $\Delta_{\max}$ takes the larger of the two headrooms, a hypothesis that starts near "maybe true" has less room to be surprising by confirmation than by rejection: at a prior mean of 0.75 the upward headroom is a third of the downward one. This is the normalizer of AutoDiscovery [5], and we keep it for comparability. Together with the prior collapse of Section 5.1, it means that *pointwise* surprisal, as defined, steers high-scoring candidates toward rejections; a direction-specific normalizer would change the share of confirmations among them but would not give a rejection an alternative, which is the gap CONTRASTIVE Discovery fills.

T.2 Contrastive belief distribution and κ_t

In the contrastive setting, each node carries an M-cell mutually exclusive and collectively exhaustive (MECE) partition of possible outcomes,

$$\{C_1, \ldots, C_M\},$$

proposed by the generator alongside the experiment plan. Belief about the experiment's outcome is represented as a probability vector over the cells,

$$p \in \Delta^{M-1},$$

rather than as a scalar belief-in-truth.

For each node, we elicit two such allocations from the belief model π_b. The pre-experiment elicitation conditions on the contrast set and the variable schema only, while the post-experiment elicitation additionally conditions on the analyst's summary of the experiment's results. Each elicitation draws five independent samples from π_b, where each sample is a 100-point integer allocation across the M cells. Responses that fail to parse are re-prompted once for the missing samples. We drop a sample if it has the wrong number of cells, a negative entry, or an all-zero allocation; samples whose points do not sum exactly to 100 are kept. We renormalize each surviving sample to a probability vector and average across samples, yielding a prior distribution p^{prior} and a posterior distribution p^{post}. We floor every cell at $\varepsilon = 10^{-6}$ before renormalizing so that

$$p_m^{\text{prior}} > 0 \qquad \text{for all } m,$$

ensuring that the KL divergence is finite.

The raw Bayesian surprise of a node is the KL divergence from prior to posterior, measured in bits:

$$\mathrm{KL}\big(p^{\text{post}} \,\|\, p^{\text{prior}}\big) = \sum_{m=1}^{M} p_m^{\text{post}} \log_2 \frac{p_m^{\text{post}}}{p_m^{\text{prior}}}. \tag{7}$$

This quantity has no fixed upper bound across nodes with different contrast-set sizes M. To make scores comparable across contrast sets of varying cardinality, we divide by the maximum entropy of an M-cell distribution, $\log_2 M$, which gives the contrastive belief shift of Section 3.3,

$$\kappa_t = \frac{\mathrm{KL}\big(p^{\text{post}} \,\|\, p^{\text{prior}}\big)}{\log_2 M}.$$

We classify a node as surprising when $\kappa_t \geq 0.5$. Only the MCTS reward is clipped,

$$r_t = \min(\kappa_t, 1) \in [0, 1],$$

and the information-gain analyses and the surprising indicator use the unclipped κ_t. Intuitively, r_t is a clipped, cardinality-normalized measure of how strongly the observed evidence redistributes belief across the contrast cells. A value of 1 indicates that κ_t is at least 1, while a value near 0 corresponds to an experiment that produces little change in the belief distribution. Because $\log_2 M$ is not generally the maximum possible KL divergence, r_t should not be interpreted as the fraction of all possible belief redistribution achieved.

T.3 Agent framework

Tree structure. Each node below the root's data-loading child is one experimental validation: a contrast set (a single hypothesis in *pointwise* runs) with its experiment plan, code, output, analyst summary, reviewer verdict, and prior and posterior beliefs. When a node's experiment finishes, the generator proposes 8 follow-up contrast sets, conditioned on the dataset summary, the node's result, and its parent and grandparent; these become the node's untried proposals. Each iteration descends from the root by UCB1 (mean backed-up reward plus $\sqrt{2 \ln N_{\text{parent}}/n}$, exploration weight 1.0) to the highest-scoring node with untried proposals, draws one at random, and executes it as a new child. The child's reward r_t (0 if the experiment fails) and a visit count are backed up to every ancestor. The first 8 iterations expand from the data-loading node, so the search starts from 8 independent questions (Table 13).

Agents. We follow the setup of the AutoDiscovery discovery agent [5] with slight modifications. We use the following LLM agents with a `gpt-4o` backbone:

Hypothesis (Contrast Set) Generator. Proposes a candidate contrast set of hypotheses that predicts the outcome of an experiment. Produces an experiment plan for evaluating the proposed contrast set.

Experiment Coder. Implements the experiment plan in code. The coder is allowed up to six implementation attempts, using feedback from the Experiment Analyst to revise failed or incomplete attempts.

Code Executor. Executes the code produced by the Experiment Coder and returns the exit code and standard output. This component is not LLM-based.

Experiment Analyst. Analyzes the output of code execution and provides feedback to the Experiment Coder when the implementation requires correction or refinement.

Experiment Reviewer. Reviews the completed experiment for alignment with the original experiment plan and assesses whether the results can support validation of the hypothesis.

Experiment Reviser. Produces a revised experiment when the original experiment or its implementation fails to execute successfully or cannot adequately validate the hypothesis.

Belief Elicitor. Distributes 100 points over the contrast set of hypotheses, evaluating whether each hypothesis is believed to be true relative to alternatives both before experimental evidence (prior) and after (posterior). For the *pointwise* case, evaluates whether the hypothesis is believed to be true from the hypothesis statement alone (prior) or from the hypothesis statement together with the experimental results (posterior).

Image Analyst. Provides textual descriptions of images produced by code written by the Experiment Coder.

U Example Contrast Sets

The sets below are verbatim system outputs, chosen to show how contrast sets partition a question across datasets and partition types. They are shown as generated, before evidence, and so include sets whose premise a domain expert would refine (e.g., the Materials Project set compares Fermi energies that the database reports relative to each calculation's own potential zero, so the comparison needs a common energy reference before the outcome is physically interpretable). Whether a generated set is well posed is one of the properties Appendix I and Appendix C measure.

Example contrast set 1

Dataset: BLADE Mortgage

Scientific question: Given a poor credit score, self-employment status leads to

Contrast set (MECE alternatives):

- **H1.** higher denial rates than non-self-employed
- **H2.** no different denial rates
- **H3.** lower denial rates than non-self-employed

Example contrast set 2

Dataset: Materials Project

Scientific question: Direct-gap vs indirect-gap materials, Fermi energy is

Contrast set (MECE alternatives):

- **H1.** higher in direct-gap materials
- **H2.** lower in direct-gap materials
- **H3.** not substantially different

Example contrast set 3

Dataset: SDSS Galaxy

Scientific question: Petrosian radius containing 90% of the flux (PetroR90_r) is predominantly determined by

Contrast set (MECE alternatives):

H1. model magnitude attributes

H2. extinction attributes

H3. neither significantly drives it

Example contrast set 4

Dataset: NLS Raw

Scientific question: Trend in vehicle market value across 1985/1990/1996 is

Contrast set (MECE alternatives):

H1. significant linear positive

H2. significant linear negative

H3. nonlinear or no significant trend

Example contrast set 5

Dataset: SDSS Galaxy

Scientific question: When controlling for galaxy axis ratios (expAB_r, deVAB_r), the redshift effect on stellar velocity dispersion

Contrast set (MECE alternatives):

H1. increases (strengthens)

H2. decreases (attenuates)

H3. disappears

V Broader Societal Impact

By resolving more questions from the same data and validation budget, the method can lower the cost of screening hypotheses before laboratory follow-up. Three hazards are specific to this system. First, its outputs are statistical resolutions on observational data, not causal findings; a "resolved" question should be read as a prioritized claim to be verified, not as an established result. Second, two datasets involve people: BLADE Mortgage (the Boston Fed HMDA study) records applicant race and gender, and BLADE Conversation lists named public contributors with gender, institution, and citation attributes. On such data the system can produce confident statistical resolutions that are not causally identified; the abstention alternative (Appendix K) lets it decline instead of forcing a resolution when no listed explanation matches the evidence, and any use of such outputs in policy-adjacent settings requires unit-aware inference, pre-registered analyses, and human review. Third, each resolution rests on the agent's own analysis choices, including how it handles missing-value sentinels and clustered observations (Appendix G); the audits of Appendix J verify that reported numbers match the code's output, and users should review the analysis code before acting on a resolution. We release prompts and, at the project website, code so that these behaviors can be audited.

W Cost Analysis

Each 200-iteration run for a given dataset costs $20 to $50. We estimate the total cost of our experiments at $1,500 to $2,000. Token spend per candidate for each condition is reported with the matched-budget comparison in Appendix L.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main claims, which Section 5 supports and Section 7 scopes.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 7 discusses the limitations of the work.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not involve theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 4 and the appendix describe the protocol, datasets, prompts, hyperparameters, and scoring needed to reproduce the results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The datasets are public, and the code is available at the project website.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 and the appendix give the experimental settings and hyperparameters.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Plots show standard-error bands across datasets, and the appendix reports confidence intervals and how each is computed.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports API usage, token counts, and costs (Appendix L and Appendix W).

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss potential societal impacts in Appendix V.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no model or dataset with a high risk of misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit the creators of all datasets, code, and models we use, and use them under their terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The released code is documented at the project website.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: Appendix I includes the relevant information about the study, and evaluators were compensated appropriately under local law and the NeurIPS Code of Ethics.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: Appendix I states that, under our institution's IRB guidance, the study does not constitute human-subjects research and did not require IRB review.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLMs are core to the method: `gpt-4o` proposes contrast sets, writes and interprets analyses, and elicits beliefs, and other models serve as cross-family elicitors and judges. The paper names each model version and role and gives the prompts (Appendix P).

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.