DO LANGUAGE MODELS KNOW WHAT COMPELLING SCIENTIFIC EVIDENCE LOOKS LIKE?

Anonymous authors
Paper under double-blind review

ABSTRACT

Expert scientists excel at weighing empirical findings from multiple sources that vary in both statistical reliability and epistemic relevance. To what extent are language models, which are increasingly driving autonomous discovery systems, capable of such *evidence deliberation*? We introduce **EDB-Bio**, a scientific reasoning benchmark curated from real-world data in three biological domains: drug-perturbation transcriptomics, CRISPR screens, and human genome-wide association studies (GWAS). Each of EDB-Bio's 96k episodes presents a sequence of experimental results from published studies and asks the model to predict the probability that a target hypothesis is true. Out-of-the-box, open-weight and frontier models alike demonstrate systematic underconfidence and attain Brier scores comparable to a trivial base-rate forecaster. Post-training against strictly proper scoring rules significantly improves both calibration and discrimination. This calibration pays off downstream: in a sequential experiment selection task, our Ibis0 model is $2.5\times$ more efficient than the best frontier model. Throughout, quantifying evidence weights as *Bayes factors* reveals how models update their beliefs over time, whether they agree with expert biologists on relevance, and where they exhibit inconsistencies and biases. Our work extends calibration training and uncertainty analysis to a realistic scientific forecasting setting with applications to human cancer biology.

1 INTRODUCTION

Experienced scientists are good at weighing evidence. Shown the result of an experiment, they can often judge whether it is compelling and likely to generalize; or whether it is weak, confounded, or irrelevant to the question being studied. Yet today, research is increasingly driven by language models (Boiko et al., 2023; Lu et al., 2024; Yamada et al., 2025; Mitchener et al., 2025b; Gottweis et al., 2026). How reliably can these models exercise epistemic judgment? This question is particularly important for autonomous workflows in which a model's output feeds decisions such as which hypothesis to pursue, what experiments to run, or how confidently to report conclusions.

Prior calibration work has concentrated on general-purpose domains such as mathematics and commonsense reasoning rather than on empirical scientific evidence (Damani et al., 2026; Puri et al., 2026; Xu et al., 2024). Agentic research systems weigh large and heterogeneous bodies of evidence of variable quality, which can be very different from observations in other domains. To this end, we introduce EvidenceDeliberationBench-Biology, EDB-Bio, a benchmark of real biological hypotheses, each paired with a controlled, relevance-graded pool of experimental evidence. The model is tasked to predict whether the hypothesis is true or not, conditioned on progressively more evidence from the pool. Because the evidence is controlled, we can measure both whether a model is right and how its belief tracks the strength and relevance of the evidence.

On EDB-Bio, language models can often differentiate more likely hypotheses from less likely ones, yet in all three domains their stated probabilities are poorly calibrated. Training an open model against a strictly proper scoring rule closes the calibration gap and matches or exceeds frontier discrimination. A model with reliable probabilistic judgments can also *choose* the experiment whose result would most change its belief. In a sequential Bayesian experimental design (Rainforth et al., 2024) setting, our trained Ibis0 extracts about $2.5\times$ more information per experiment than a prompted frontier model.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

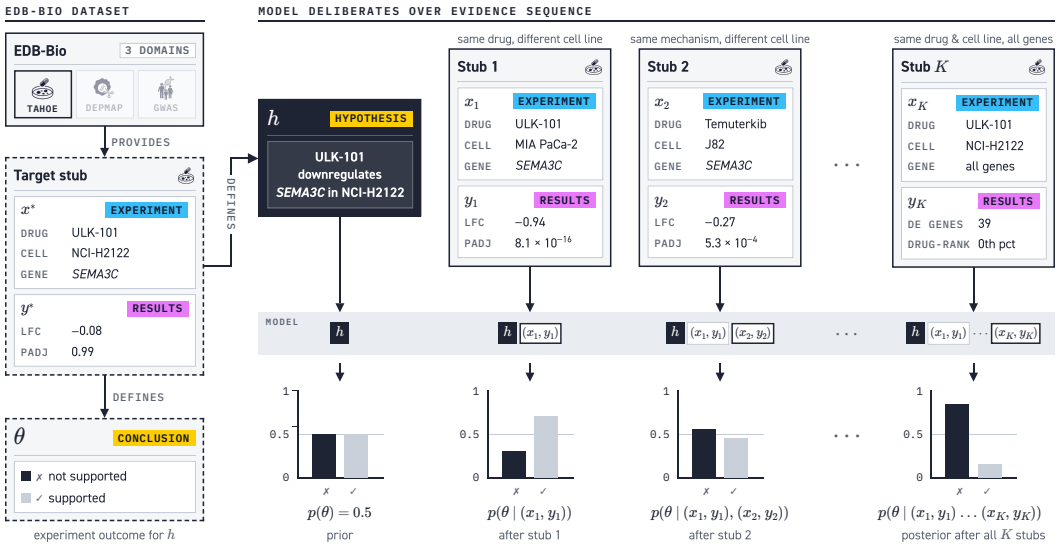

Figure 1: **Evidence deliberation task.** EDB-Bio asks models to reason about a scientific hypothesis h given a sequence of empirical evidence. Each *evidence stub* is derived from published data in one of three biological domains (Table 1). In the example here, each experiment x is an RNA-Seq study from Tahoe-100M (Zhang et al., 2025b) in which a drug perturbs a gene in a particular cell line. The result y reports a DESeq2 assay readout, including an *effect* ($\log_2$ fold-change) and its *significance* (adjusted p-value). *Left:* In each episode, the hypothesis h and binary support label θ are defined by a held-out *target stub*, which is not shown to the model. *Right:* Given h, the model sees a sequence of supporting stubs from studies on related drugs, cell lines, and genes and must verbalize its belief $p(\theta)$ after each. By comparing sequential belief states, we can measure which stubs models find compelling based on the strength of their updates.

1.1 Contributions

1. **A novel scientific forecasting task.** We formalize evidence deliberation as forecasting whether a target hypothesis is true or not, given a pool of prior *evidence stubs* revealed one at a time. The task structure uses a domain-agnostic format, is scored with proper scoring rules, and enables principled analysis of belief revision and sequential experiment selection.
2. **EDB-Bio dataset.** We instantiate the task as 96,000 directly tested biological hypotheses across three domains: drug-perturbation transcriptomics (Tahoe), CRISPR gene-dependency screens (DepMap), and genome-wide association studies (GWAS). Each domain is built using five design principles that ensure class balance/diversity and provide evidence relevance labels based on a tiered expert hierarchy.
3. **A calibration study of frontier and open models.** Frontier models rank supported hypotheses above unsupported ones, but their probabilities are poorly calibrated. Post-training Qwen3-30B-A3B against a strictly proper scoring rule yields Ibis0, which closes the calibration gap while outperforming frontier discrimination.
4. **Sequential analysis of belief revision in LMs.** We quantify the *weight of evidence* using log Bayes factors, which measure how each new stub causally influences the model's belief about the target hypothesis. This allows us to test if updates move in the right direction, track expert relevance, have the right magnitude, and are invariant to order. Our analysis reveals interesting biases; for instance, models show ordering effects and overreact to evidence, making updates larger than the outcomes would justify.
5. **An EIG estimator for sequential experiment selection.** We use EDB-Bio to instantiate a Bayesian experimental design setting and introduce a novel factored estimator of expected information gain (EIG) that is more sample-efficient than standard Monte Carlo methods. Under a well-calibrated belief model (Ibis0), it extracts roughly $2.5\times$ more information per experiment than a prompted frontier model. We also find that training an LLM to *simulate* experimental outcomes outperforms empirical samples when used for experiment selection with our EIG estimator (§6).
6. **Open release.** We release the EDB-Bio datasets, weights for all Ibis0 variants, our training and evaluation code, and the logs from our training runs.

We introduce EDB-Bio in §2 and the models and metrics in §3. We then evaluate the models' ability to forecast target hypotheses' validity on EDB-Bio (§4). Next, we examine how model beliefs update as evidence is progressively disclosed; in particular focusing on prior-independent analyses (§5). Finally, we use the models to select which experiment to run next and compare their efficiency against prompted frontier models (§6).

2 Evidence Deliberation

Our benchmark models the task of *evidence deliberation*, in which a scientist is trying to forecast the likelihood of a *target hypothesis* given a mixed pool of empirical evidence from prior studies. Evidence is organized into episodes (Fig. 1), which each contain:

- **Target hypothesis** h: A statement of a scientific hypothesis, which for each domain follows a consistent format; e.g., *Drug D downregulates gene G in cell line C.*
- **Label** $\theta \in \{0, 1\}$: Whether direct scientific evidence supports h ($\theta = 1$) or not ($\theta = 0$), as determined by a domain-specific criterion (Table 1).
- **Evidence pool** $I_k = \{(x_1, y_1), \ldots, (x_k, y_k)\}$: A variable-length sequence of $k \in \{1, \ldots, 16\}$ tuples each comprised of an experiment x_j and the outcome y_j of conducting the experiment. Each entry in the sequence is formatted consistently and referred to as a stub.

Model beliefs & weights of evidence. The model's task is to predict the probability that the target hypothesis h is true conditioned on the available evidence, i.e. $p(\theta = 1|I_k)$. For notational convenience, we define $p_k = p(\theta = 1 \mid I_k)$. The evolution of p_k as a function of k is a useful setting in which to study evidence that language models deem persuasive in favor of or against the target hypothesis. We also study belief updates using prior-invariant metrics that isolate the effect of scientific evidence on belief revisions, independent of model priors on EDB-Bio tasks (App. E). In particular, we leverage the *weight of evidence* (Good, 1950) of stub k, $w_k = \text{logit}(p_k) - \text{logit}(p_{k-1})$, where $\text{logit}(p) = \log \frac{p}{1-p}$. We report it in *decibans* (dB), $10 \log_{10}$ of the Bayes factor.

2.1 EDB-Bio Dataset

Domains. The same task format is instantiated in three biological domains: transcriptional responses to drug perturbation in cancer cell lines (Tahoe-100M; Zhang et al., 2025b), CRISPR gene-dependency screens (DepMap, Broad, 2024)), and published genetic associations from the GWAS Catalog (Sollis et al., 2023). The domains differ in what is measured, in the support criterion, and in how related evidence is defined, allowing us to test one form of evidence deliberation across interventional and observational data (Table 1A).

Hypothesis and label construction. Each hypothesis in EDB-Bio is derived from an experimental stub (x^*, y^*) from the source dataset. Specifically, the target hypothesis is defined to be the hypothesis that x^* directly tests. In stating the hypothesis h, we include details about x^* and how the truth label θ is determined as a function of y^* (Table 1). This minimizes ambiguity about the formulation of the hypothesis and its label.

Design principles. Because the source datasets are imbalanced and contain assay-specific artifacts, every domain is built under five design principles (Table 1B) that control for these issues: targets with artifact or trivial labels are excluded (P1); targets are diversified and balanced 50:50 across claim axes (P2); evidence pools follow a graded relevance hierarchy defined by domain experts (P3) and vary in sufficiency from evidence-rich to evidence-poor (P4); and the target experiment and outcome are never part of the pool for the corresponding hypothesis (P5). App. A reports domain-specific construction details, and App. I gives a complete prompt from each domain.

Dataset. Each domain contains 32,000 training, 256 validation, and 2,000 test episodes, with evidence pools of up to $k = 16$ stubs. We evaluate two test configurations: the *primary set* shows every test episode once at a single context length, and the *sensitivity set* shows a 256-episode subset at every $k \in \{0, \ldots, 16\}$ for the belief revision analyses (§5). App. A.6 gives the full configuration and provenance, together with domain-specific details.

(A) DOMAINS

	Tahoe *Pharmaco-transcriptomics*	DepMap *Functional genomics*	GWAS *Population genetics*
LEVEL OF BIOLOGY	Molecular	Cellular	Organism / population
SUBJECT	Cancer cell lines	Cancer cell lines	Human cohorts
HYPOTHESIS	*Drug D up/down-regulates gene G in cell line C*	*Gene G is a dependency in cell line C*	*SNP S (allele A) shifts trait T in direction D*
DATA SOURCE	Tahoe-100M	Broad Institute	NHGRI-EBI
ASSAY TYPE	Pseudobulk RNA-seq	CRISPR KO screen	Assoc. study
EVIDENCE MODALITY	Single assay	Multi-modal	Catalog associations
READOUT	$\log_2$ fold-change	gene-effect score	p-value + direction
SUPPORTED IFF	$\lvert\log_2 FC\rvert > 0.5$, $p_{adj} < 0.05$	Chronos ≤ -0.5, $p_{dep} \geq 0.5$	Direction match, $p < 5\times10^{-8}$

(B) DESIGN PRINCIPLES

P1 **No artifact or trivial labels.** Targets dominated by technical artifacts or with near-constant labels are excluded.

P2 **Diversity across claim axes.** Targets span entities and contexts; each domain balanced 50:50 supported vs. not.

P3 **A graded relevance hierarchy.** Evidence ranges from closely related measurements to unrelated controls, ordered by experts.

P4 **Diversity in evidence sufficiency.** Episodes split evenly among evidence-rich, -moderate, and -poor strata.

P5 **No target leakage.** The held-out target is excluded from its own pool; every domain audited for exact matches.

Table 1: **(A)** Comparison of the three EDB-Bio domains: each instantiates the same evidence-deliberation task with a distinct assay, hypothesis, support criterion, and evidence modality. **(B)** Five overarching design principles, which are applied to each domain.

3 METHODS

3.1 MODELS AND TRAINING

Models and baselines. We consider three kinds of LLMs: an open-weight model evaluated as released, our post-trained variants of it, and frontier models. Our starting checkpoint is `Qwen3-30B-A3B`, which we refer to throughout as **Qwen3**. We evaluate Qwen3 in two modes: *non-thinking* prompts the model to state a probability immediately, while *thinking* generates a reasoning trace before stating the probability.

Calibrating LLM uncertainty. We post-train our models against two objective functions: Brier score and binary cross-entropy, which are both strictly proper scoring rules (Gneiting & Raftery, 2007). This approach is a promising method for calibrating LLM uncertainty (Damani et al., 2026; Puri et al., 2026; Bani-Harouni et al., 2026; Leng et al., 2025; Xu et al., 2024), but prior work has applied it mainly to factual question answering and mathematics rather than to scientific judgment (§7). In total, we present three post-trained model variants, which we refer to collectively as **Ibis0**:

- **Ibis0 (non-thinking)** trains with a group-relative policy-gradient objective (App. B) on a negative-Brier reward $r = 1 - (p - \theta)^2$, where the model directly states a probability p that the target hypothesis is supported, with no intermediate reasoning.
- **Ibis0 (thinking)** trains using the same objective but reasons in text before stating the probability.
- **Ibis0 (classifier)** replaces the language-model output head with a scalar linear head, applies it to the final-token representation, and minimizes binary cross-entropy (BCE).

All models were trained on a multi-node cluster with 8x NVIDIA B200 GPUs per node using `prime-rl` (Prime Intellect, 2025). Training details are provided in App. B.

Frontier models. As a strong zero-shot baseline, we evaluate two frontier models: `Claude Opus 5` (Anthropic) and `GPT-5.6-Sol` (OpenAI). Both models were fed identical prompts and evaluated at high reasoning effort with a 24,000 token budget. Further details are provided in App. D.

3.2 EVALUATION METRICS

Aggregate forecast quality. For beliefs $p_i = p(\theta_i = 1 \mid I_i)$ and binary labels $\theta_i \in \{0, 1\}$ we report the *Brier score* $\frac{1}{N}\sum_i (p_i - \theta_i)^2$ (Brier, 1950); lower is better, and because EDB-Bio is balanced 50:50 by construction (P2), a constant forecast at the base rate scores 0.25.

Calibration vs. discrimination. We report *sMECE*, a binning-free calibration error in the smooth-ECE family of Błasiok & Nakkiran (2024): the density-weighted mean gap between a stated probability and the empirical frequency of support at that probability (App. C). *AUROC* measures whether supported targets are ranked above not-supported ones, independent of calibration.

Calibration of weights of evidence. If a model's probabilities are calibrated, its weights of evidence w should correlate with the label θ. Specifically, across episodes in which the model starts from belief p_{k-1}, an update of w dB should change the odds that the target hypothesis turns out to be supported by w dB, compared with near-zero updates. To test this, we bin each model's updates into five tiers by w. Each tier's *odds ratio of support* ψ is its odds of support divided by the odds after near-zero updates, in decibans like w. It is computed within bins of starting belief, so it does not depend on the model's prior (App. C). Plotting ψ against w gives a reliability diagram for belief updates (Fig. 4, left), which we summarize by its slope β: the change in the log-odds of support per deciban of weight of evidence. Calibrated updates give $\beta = 1$. When $\beta < 1$, the model *overreacts*: its updates are larger than the outcomes justify, and multiplying each update by β would calibrate it on average. When $\beta > 1$, the model *underreacts*, and its updates are too small. The quantity $\beta - 1$ is the log-odds analog of the coefficient that economists use to measure over- and underreaction to news, obtained by regressing forecast errors on forecast revisions (Coibion & Gorodnichenko, 2015; Bordalo et al., 2020).

All statistics are equal-weight macro-averages over the three domains, with 95% intervals from a bootstrap that resamples whole episodes within each domain (App. C).

4 Calibration and Discrimination

A useful forecast must rank hypotheses correctly and assign probabilities of the right magnitude, and the two properties need not coincide. Fig. 2 reports Brier score as the overall measure and then separates calibration from discrimination. Fig. 3 shows why a well-ranked forecast can still score poorly: out of the box, models state probabilities biased away from the 0.5 base rate and drawn from only a handful of distinct values.

Post-training outperforms zero-shot models. GPT-5.6-sol and Claude Opus 5 rank supported above not-supported hypotheses (AUROC 0.737 and 0.752), yet their Brier scores are not significantly better than a base-rate forecaster. Both Qwen3 modes show the same mismatch. All three Ibis0 models obtain Brier below 0.25 in every domain. The classifier has the best Brier score (0.177) and AUROC (0.820), whereas the RL policies are the best calibrated, with sMECE 0.032 against 0.066 for the classifier. Table 6 reports per-domain results with intervals.

Prior-aware prompting and post-hoc recalibration. We test two methods to improve calibration without post-training: (1) explicitly instructing models to use a prior of 0.5, and (2) fitting a monotone transform post-hoc (Platt, 1999). We find prompting modestly improves calibration but hurts discrimination (Fig. 2). Meanwhile, Platt rescaling approaches Ibis0 on calibration, but (by construction) does not affect discrimination (App. H). However, such post-hoc recalibration requires access to ground truth labels, so it is not a substitute for training.

Calibration is amortizable without reasoning. While the zero-shot thinking variant of Qwen3 outperforms non-thinking, after post-training, this gap disappears. In fact, Ibis0 (thinking) marginally under-performs non-thinking, though it does provide better interpretability through its reasoning traces (App. J). This aligns with recent work showing that calibration can be effectively *amortized* in a single forward pass (Zhao et al., 2025; Hollmann et al., 2023; 2025; Almeida, 2026), particularly for tasks that do not require mathematical or symbolic reasoning (Sprague et al., 2025).

Cross-domain transfer. To test generalization across domains, we train Ibis0 (classifier) on two domains and hold out the third. App. F reports the results of this transfer experiment for each held-out domain in terms of *retention*: the fraction of full-model performance preserved by the cross-evaluated classifier. Brier retention ranges by domain from 70.0% on Tahoe to 26.3% on GWAS (Fig. 11). By further decomposing Brier (Murphy, 1973), we observe that calibration transfers across domains, but discrimination does not. In other words, the model retains the right probability scale, but not the knowledge needed to rank hypotheses in an unseen domain.

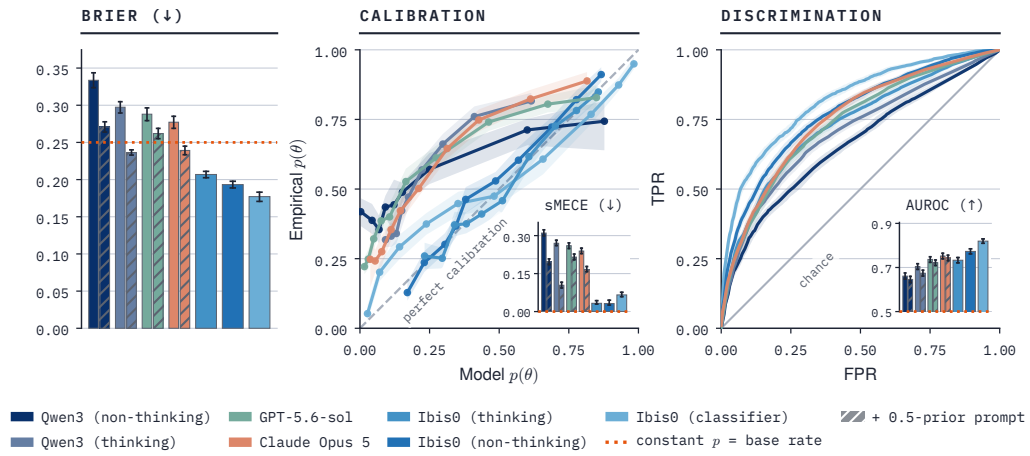

Figure 2: **Do language models produce calibrated & accurate beliefs when presented with evidence?** Zero-shot LLMs (Qwen3, GPT-5.6-sol, and Opus 5) underperform a constant $p = 0.5$ forecast on Brier score (dotted line, *left*). This is primarily due to poor calibration: Models that sit above the diagonal (*middle*) are systematically *underconfident*, despite demonstrating above-chance discrimination (*right*). Explicitly instructing models to use a prior of 0.5, matching the base rate, helps somewhat (dashed bars), but post-training on proper-score objectives (Ibis0 variants) improves both calibration and discrimination substantially.



Figure 3: **How do language models verbalize beliefs?** Histograms of every stated $p(\theta)$ produced as evidence accrues on the sensitivity set ($N = 13{,}056$ per model). Out of the box, Qwen3 (thinking) and the frontier models are underconfident; post-training recenters the mean near the 0.5 support rate. Models that verbalize a probability in text produce fewer than 100 distinct values, whereas the scalar head of Ibis0 (classifier) produces a smooth distribution with 18–50× finer resolution.

5 Belief Revision

The model's stated probabilities alone do not reveal whether a model effectively used the evidence provided or relied on its prior about the target hypothesis. We answer this question by presenting each episode sequentially at every evidence count $k \in \{0, \ldots, 16\}$: the value at $k = 0$ measures target-only prediction (prior), and the change in belief as k increases measures how the model uses the evidence. Every tested model increases its confidence in the ground-truth label as k increases (App. E). In the Tahoe and DepMap domains, the Ibis0 classifier has both the highest magnitude of belief revision and the highest ground-truth probability at $k = 16$.

Direction and relevance of belief updates. Sorting stubs by their direction relative to h (App. C), every model's mean update is positive on supporting stubs, negative on refuting stubs, and small on inconclusive ones (Fig. 9). Ordering stubs by the curated relevance hierarchy (App. A.2), Ibis0 (non-thinking) aligns most closely in every domain (Spearman $\rho = 0.18$–0.37 across domains; App. E.3), despite trailing the classifier on Brier and AUROC (Fig. 10). Where models disagree with the expert ordering, they do so together: on GWAS every model over-weights indirect same-SNP evidence, and on Tahoe, Claude Opus 5 prefers aggregate summaries to specific assay readouts (Fig. 4, *right*). App. E.2 reports both analyses in full.

Calibration of the weights of evidence. As observed in Fig. 4, every model has $\beta < 1$, which indicates overreaction (§ 3.2). Values range from 0.13 for untrained Qwen3 (non-thinking) to 0.89 for Ibis0 (non-thinking), the only model whose interval includes 1; the frontier models fall between.

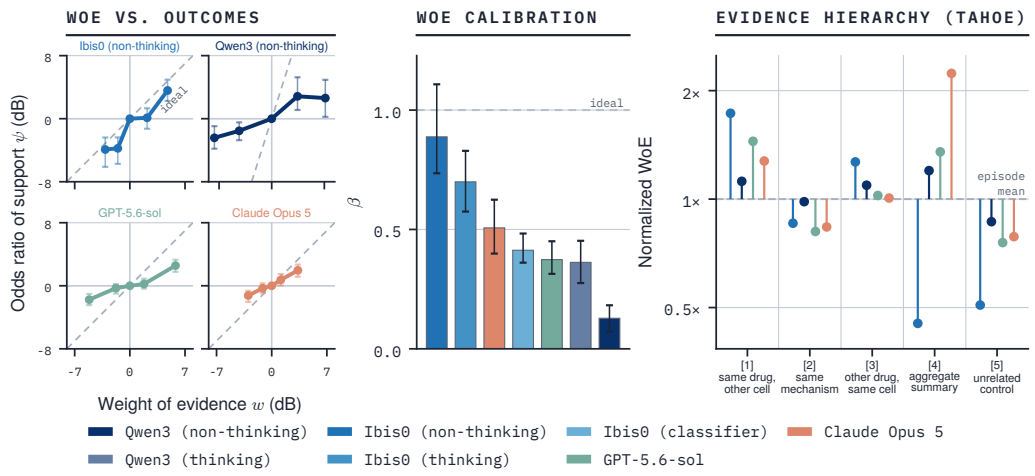

Figure 4: **Do models weigh each piece of evidence at the right magnitude?** *Left:* Odds ratio of support ψ against weight of evidence w; calibrated updates lie on the diagonal ($\beta = 1$). *Middle:* the slope β of each curve; $\beta < 1$ indicates overreaction, meaning updates are larger than justified. Post-training brings Ibis0 (non-thinking) closest to calibrated. *Right:* Evidence weights w on Tahoe, normalized by episode mean and grouped by biological relevance. A model that follows the expert hierarchy should show monotonic reduction from left to right. Instead, all models under-weight evidence from non-matching drugs that share a mechanism of action with the target hypothesis [2], and all but Ibis0 over-weight genome-wide aggregate summaries [4], which are only indirectly-related to the target.

At $\beta = 0.13$, the updates of Qwen3 (non-thinking) are about eight times larger than the outcomes justify. Ibis0 (classifier) has the best Brier score and AUROC of any model (§4) but a lower β (0.41) than either RL policy: accurate aggregate forecasts do not imply correctly sized updates.

Order and composition of evidence. If stated probabilities behave like beliefs, rearranging a fixed evidence set should not change the result, and the joint weight of a matched supporting and refuting pair should equal the sum of the two single-stub weights. We test both properties by holding evidence content fixed while changing its position or composition (App. K, Fig. 12). Consistency does not follow from aggregate forecast quality: the classifier has the best Brier score of the Ibis0 models but the largest primacy bias (+1.01 dB) and composition residual (4.16 dB), whereas the two RL policies reduce the composition residual of Qwen3 (non-thinking) from 7.54 dB to below 1 dB. Enabling a reasoning trace reduces primacy bias by about 0.5 dB both before and after post-training.

6 Bayesian Experimental Design

We now turn to a key application of calibrated beliefs: a Bayesian experimental design setting in which the model selects *which experiment to run next*.

Experiment setup. We configure EDB-Bio as a sequential-selection task (App. G) on 512 held-out episodes in each of the three domains. Starting from a target hypothesis h and a pool of $k = 16$ *unrevealed* candidate stubs $\mathcal{C} = \{(x_1, _), \ldots, (x_k, _)\}$, a *policy* π chooses one experiment x_t at a time to reveal the corresponding outcome y_t. We measure performance in two ways:

(1) Belief distribution sharpening: An efficient policy should increase the probability assigned to the ground truth label, $\theta\, p(\theta = 1 \mid I) + (1 - \theta)\, p(\theta = 0 \mid I)$, in as few timesteps as possible.

(2) Information gain (IG): The reduction in *information entropy* (Shannon, 1948; MacKay, 1992) from experiment x is $\mathbb{H}\big(p(\theta \mid I)\big) - \mathbb{H}\big(p(\theta \mid I, x, y)\big)$. To isolate the effect of experiment selection, we compute IG with the same belief model, Ibis0 (classifier), for all policies.

Policies. The policies differ in how they use a language model.

- *Random*: π_{random} ignores the LM and chooses a uniformly random experiment from $\mathcal{C}$.
- *Naive prompting*: π_{naive} prompts a model to directly select a candidate based on h (App. G.1).

- *EIG*: π_{EIG} uses the model's verbalized belief $p(\theta \mid I)$ to select the experiment that maximizes *expected information gain*, $\text{EIG}(x) = \mathbb{H}\big(p(\theta \mid I)\big) - \mathbb{E}_{y\sim p(y|x)}\big[\mathbb{H}\big(p(\theta \mid I, x, y)\big)\big]$ (1), where the expectation arises from uncertainty over possible outcomes y. We compare two approaches to sampling outcomes: (i) *EIG-exemplar* draws empirical samples from the training set, and (ii) *EIG-simulate* (described below) trains a simulator $p(y \mid x, \theta)$ to sample outcomes for each candidate.

Experiment simulator and a factored EIG estimator. The standard Monte Carlo estimator for Eq. (1) requires a *simulator* $p(y \mid x, \theta)$ that can "imagine" possible outcomes for a candidate x. A simulator is easy to train from (x, y) pairs by standard SFT, but conditioning it on θ requires counterfactual reasoning about how a different experiment bears on the outcome of x. We instead marginalize over the label θ' of the *local* hypothesis h' that the candidate x directly tests, which factors EIG into a supportive and a non-supportive branch:

$$\text{EIG}(x) \approx \underbrace{\mathbb{H}\big(p(\theta \mid I)\big)}_{\text{current uncertainty}} - \underbrace{p(\theta' \mid I)\ \mathbb{E}_{y^+\sim p(y|x,\theta'=1)}\big[\mathbb{H}\big(p(\theta \mid I, x, y^+)\big)\big]}_{\text{supportive branch}} - \underbrace{\big(1 - p(\theta' \mid I)\big)\ \mathbb{E}_{y^-\sim p(y|x,\theta'=0)}\big[\mathbb{H}\big(p(\theta \mid I, x, y^-)\big)\big]}_{\text{non-supportive branch}}. \quad (2)$$

Our **factored Monte Carlo EIG estimator (Eq. (2))** uses two trained models: (i) a belief model (Ibis0) $M \approx p(\theta \mid I)$ that returns the branch weight $p(\theta' \mid I)$ and the revised belief $p(\theta \mid I, x, y)$, and (ii) an outcome simulator $S \approx p(y \mid x, \theta')$ that generates a candidate's outcome y from its design x and the value of θ'. We train S by supervised fine-tuning a separate Qwen3-30B-A3B on the EDB-Bio training split. The factorization is efficient. In practice, one outcome sample per branch gives a stable EIG estimate, and the two branches can be evaluated in parallel. Each $\text{EIG}(x)$ therefore costs three belief calls (the branch weight and the two posteriors) and two simulator calls (one outcome per branch); the prior entropy $\mathbb{H}(p(\theta \mid I))$ is shared across candidates and cancels in the argmax. A standard nested Monte Carlo estimate instead draws many outcomes from $p(y \mid x)$ per candidate and recomputes a posterior for each.

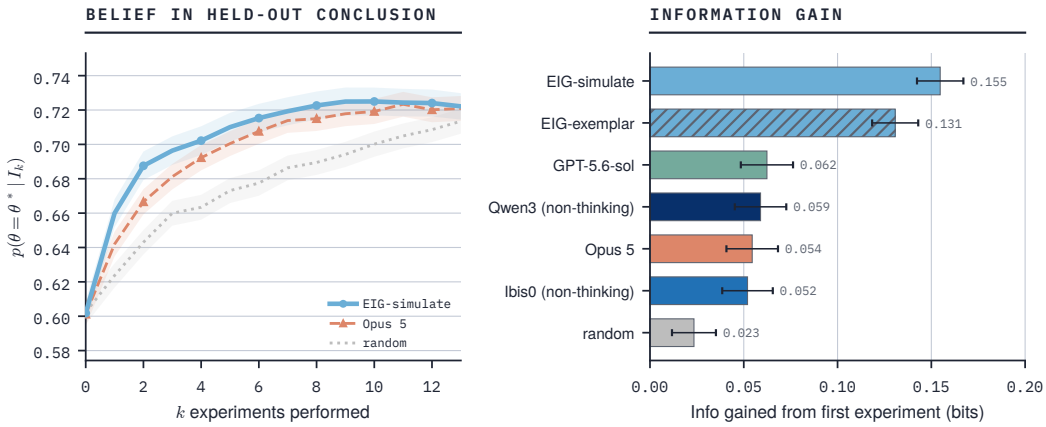

Figure 5: **Experiment selection under the EIG objective.** *Left:* Belief in the held-out conclusion, $p(\theta = \theta^* \mid I_k)$, versus experiments performed, pooled over domains (±s.e.m.). EIG-simulate, which uses Ibis0, converges more quickly than both Opus 5 (under naive prompting) and a random policy. *Right:* Information gained from the first experiment, $\mathbb{H}[p(\theta)] - \mathbb{H}[p(\theta \mid I_1)]$ (bits, 95% CI), pooled over domains. EIG-simulate and EIG-exemplar extract ∼2.5× the information of the best prompted model.

Ibis0 + EIG extracts more information per experiment. Fig. 5 (right) shows the entropy reduction for the first experiment selected by each policy, averaged over domains. EIG-simulate and EIG-exemplar reduce 0.155 and 0.131 bits respectively, versus at most 0.062 bits for any naive prompting policy (Opus 5, GPT-5.6-sol, Qwen3, Ibis0) and 0.023 for random. The EIG policies extract roughly 2.5× as much information as the best frontier policy, with non-overlapping 95% CIs. The advantage is strongest on GWAS and DepMap; on Tahoe, the top information-gain and frontier policies are statistically tied. Prompted directly, Ibis0 (non-thinking) does no better than the other naive policies. In this setting, calibration improved experiment selection only when EIG was estimated and optimized explicitly at test time.

Our approach converges more quickly to the truth. On belief in the held-out conclusion (Fig. 5, left), EIG-simulate reaches a given belief in roughly 30–40% fewer experiments than random and slightly outperforms prompted Opus 5.

A trained simulator matches empirical exemplar outcomes. The trained simulator (EIG-simulate) matches or slightly exceeds EIG-exemplar, which draws *actual outcomes from the training set*. A model can therefore learn to simulate outcomes of unseen experiments well enough to select informative ones.

7 Related work

Scientific reasoning benchmarks. Benchmarks for scientific AI span expert question answering (Rein et al., 2024), research coding and data-driven discovery (Tian et al., 2024; Majumder et al., 2025; Chen et al., 2025), simulated environments for scientific reasoning and discovery (Wang et al., 2022; Jansen et al., 2024), and biology research workflows (Laurent et al., 2026; Koch et al., 2026). Closest to our task is claim verification against retrieved abstracts (Wadden et al., 2020; Jansen et al., 2025), which scores a categorical verdict on existing evidence; EDB-Bio scores a calibrated probability that a hypothesis will be supported by an experiment whose result is not yet observed.

Calibration and proper-scoring training. Work on LLM calibration has centered on confidence in answer correctness (Kadavath et al., 2022; Geng et al., 2024), and alignment tends to leave models overconfident (OpenAI, 2023; Zhu et al., 2023). Recent methods post-train against a proper scoring rule to calibrate that confidence (Damani et al., 2026; Puri et al., 2026; Bani-Harouni et al., 2026; Leng et al., 2025; Xu et al., 2024); we extend these methods to scientific forecasting.

Bayesian experimental design with LLMs. Bayesian experimental design selects experiments by expected information gain (Lindley, 1956; Rainforth et al., 2024), whose estimation is costly (Rainforth et al., 2018; Foster et al., 2021). A growing line drives selection from a language model's own beliefs, for information gathering (Hu et al., 2024; Kobalczyk et al., 2025; Choudhury et al., 2026; Grand et al., 2026) and for planning perturbation screens (Huang et al., 2024; Hao et al., 2025). Such selection presupposes coherent belief updating, yet language models show primacy and anchoring effects (Wang et al., 2023b; Liu et al., 2024a; Lou & Sun, 2024) and depart from Bayesian updating on sequential evidence (Gupta et al., 2025; Kim et al., 2025). We measure updates as log Bayes factors in decibans (Good, 1950; Kass & Raftery, 1995) and test whether they over- or underreact to the evidence.

8 Discussion, limitations, and future work

Scientists routinely say that something is *known*. Epistemology and meta-epidemiology researchers (Kuhn, 1997; Ioannidis, 2005; Schulz et al., 1995) have fruitfully studied the various methodological, social and institutional forces behind that phrase. Our work is motivated by a complementary focus, which we term *agent epistemology*: what compels an AI system to shift its belief in a scientific claim? Framing evidence deliberation as calibrated probabilistic forecasting makes this question measurable. The same approach outlined in this paper could be used to more directly assess methodological notes (Sterne et al., 2019), absence of detail (Chan et al., 2004), or other meta-scientific information such as conflicts of interest or institutional affiliations (Lundh et al., 2017), questions that today largely rest on human intuition when weighing scientific evidence. In contrast, a well-calibrated model trained across a sufficiently broad scientific literature may help scientists surface systematic credibility issues with studies.

Limitations and future work. While constructed from real biological data sources, EDB-Bio is synthetic in several respects: evidence stubs are structured summaries of the underlying studies, hypotheses are binary, and every domain is deliberately balanced at $p(\theta) = 0.5$ (P2), which may account for the cross-domain calibration transfer (§4). We post-train one model family (Qwen3) at one scale, and the factored EIG estimator, while cheaper than nested Monte Carlo (Rainforth et al., 2018), still costs more than naive prompting. Finally, we treat a scientific belief as a single probability, but belief in ordinary usage is richer (Hawthorne et al., 2016); we adopt the probabilistic framing because it makes deliberation measurable via Bayesian analysis. We assume that experiments are uniform and have fixed cost, whereas real research programs face budgets and timelines; extending the selection objective to cost-aware acquisition is a natural next step, as is moving from structured evidence stubs to free-text evidence drawn from the literature.

AI USE STATEMENT

This research was carried out by computer scientists and biologists, with the assistance of AI agents and tools. In this work, we used generative AI tools to clean and reformat datasets, to provide feedback on research methodology and experimental design, to implement methods, and to support qualitative analysis of model reasoning traces. We have not used generative AI tools to generate data: AI tools helped write the scripts that construct EDB-Bio, but the data itself is sourced from published scientific studies (App. A). Formulating mathematical claims, providing ingredients for or writing proofs, and translation are not applicable to this work. Additionally, we used generative AI tools to draft and edit parts of the paper for readability, to suggest its structure, to summarize and identify related literature, to create and edit software code, to generate figures with matplotlib, and to draft schematic figures with Claude Design. The tools used were Claude Code, Claude Design, OpenAI Codex, and the models Claude Opus 4.8, Claude Opus 5, Claude Opus 5.5, Claude Sonnet 5, Claude Fable 5.1, GPT-5.6-sol, and GPT-6 Astra.

We have reviewed all AI-assisted work. All AI-drafted text was reviewed and revised by the authors; reported numbers were checked against the evaluation outputs; every citation was checked against its source; and AI-assisted code was reviewed and tested by the authors. Because Claude Opus 5 and GPT-5.6-sol are also evaluated as baselines, we note that AI assistance did not produce or alter any reported result. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no human subjects and no private or personally identifiable data. EDB-Bio is built entirely from public, curated scientific resources used under their respective licenses; the GWAS data are aggregate summary statistics, not individual genotypes.

Advancing the scientific capabilities of language models is dual-use: the same calibrated hypothesis evaluation that can accelerate beneficial research could in principle inform harmful research. However, EDB-Bio does not include examples of clearly harmful scientific hypotheses, and instead focuses on pre-existing hypotheses of relevance to cancer therapeutic design.

Finally, the Ibis0 models described in this work produce probabilistic judgments that remain imperfect and have not been exhaustively evaluated across other areas of life science; they are shared in the spirit of open-science research, not as a substitute for experimental validation and expert review.

REPRODUCIBILITY STATEMENT

The release will include the EDB-Bio dataset, Ibis0 model weights, model-evaluation harness, calibration analyses, and training configurations in a pinned dependency environment. The EDB-Bio artifacts will include commands for regenerating each domain from its source data. Repository and dataset URLs, license information, and frontier-model snapshot identifiers will be added before publication.

REFERENCES

Abhinav K. Adduri, Dhruv Gautam, Beatrice Bevilacqua, Alishba Imran, Rohan Shah, Mohsen Naghipourfar, Noam Teyssier, Rajesh Ilango, Sanjay Nagaraj, Mingze Dong, et al. Predicting cellular responses to perturbation across diverse contexts with State. *bioRxiv*, 2025. doi: 10.1101/2025.06.26.661135.

Constantin Ahlmann-Eltze, Wolfgang Huber, and Simon Anders. Deep-learning-based gene perturbation effect prediction does not yet outperform simple linear baselines. *Nature Methods*, 22(8): 1657–1661, 2025. doi: 10.1038/s41592-025-02772-6.

Diogo Almeida. Introducing System One Models & Jev. TypeSafe AI, https://typesafe.ai/blog/introducing-system-one-models-and-jev, 2026.

David Bani-Harouni, Chantal Pellegrini, Paul Stangel, Ege Özsoy, Kamilia Zaripova, Nassir Navab, and Matthias Keicher. Rewarding doubt: A reinforcement learning approach to calibrated confi-

dence expression of large language models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2503.02623.

Jarosław Błasiok and Preetum Nakkiran. Smooth ECE: Principled reliability diagrams via kernel smoothing. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2309.12236.

Tom Blau, Edwin V. Bonilla, Iadine Chades, and Amir Dezfouli. Optimizing sequential experimental design with deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 2107–2128, 2022. arXiv:2202.00821.

Daniil A. Boiko, Robert MacKnight, Ben Kline, and Gabe Gomes. Autonomous chemical research with large language models. *Nature*, 624(7992):570–578, 2023. doi: 10.1038/s41586-023-06792-0.

Pedro Bordalo, Nicola Gennaioli, Yueran Ma, and Andrei Shleifer. Overreaction in macroeconomic expectations. *American Economic Review*, 110(9):2748–2782, 2020.

Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950. doi: 10.1175/1520-0493(1950)078%3C0001:VOFEIT%3E2.0.CO;2.

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

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

An-Wen Chan, Asbjørn Hróbjartsson, Mette T Haahr, Peter C Gøtzsche, and Douglas G Altman. Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. *Jama*, 291(20):2457–2465, 2004.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5186–5200, 2024. doi: 10.18653/v1/2024.acl-long.283. arXiv:2308.16175.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward rigorous assessment of language agents for data-driven scientific discovery. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.05080.

Deepro Choudhury, Sinead Williamson, Adam Goliński, Ning Miao, Freddie Bickford Smith, Michael Kirchhof, Yizhe Zhang, and Tom Rainforth. BED-LLM: Intelligent information gathering with LLMs and bayesian experimental design. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2508.21184.

Olivier Coibion and Yuriy Gorodnichenko. Information rigidity and the expectations formation process: A simple framework and new facts. *American Economic Review*, 105(8):2644–2678, 2015.

Steven M. Corsello et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature Cancer*, 1(2):235–248, 2020. doi: 10.1038/s43018-019-0018-6.

Hao Cui, Zahra Shamsi, Gowoon Cheon, Xuejian Ma, Shutong Li, Maria Tikhanovskaya, Peter Norgaard, Nayantara Mudur, Martyna Plomecka, Paul Raccuglia, et al. CURIE: Evaluating LLMs on multitask scientific long context understanding and reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2503.13517.

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

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. Beyond binary rewards: Training LMs to reason about their uncertainty. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2507.16806.

Morris H. DeGroot and Stephen E. Fienberg. The comparison and evaluation of forecasters. *Journal of the Royal Statistical Society, Series D (The Statistician)*, 32(1-2):12–22, 1983. doi: 10.2307/2987588.

Joshua M. Dempster, Isabella Boyle, Francisca Vazquez, David E. Root, Jesse S. Boehm, William C. Hahn, Aviad Tsherniak, and James M. McFarland. Chronos: a cell population dynamics model of CRISPR experiments that improves inference of gene fitness effects. *Genome Biology*, 22:343, 2021. doi: 10.1186/s13059-021-02540-7.

DepMap, Broad. DepMap 24Q4 Public. Figshare+ dataset, https://doi.org/10.25452/figshare.plus.27993248.v1, 2024.

Shrey Desai and Greg Durrett. Calibration of pre-trained transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 295–302, 2020. doi: 10.18653/v1/2020.emnlp-main.21.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 5050–5063, 2024. doi: 10.18653/v1/2024.acl-long.276. arXiv:2307.01379.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0.

Adam Foster, Martin Jankowiak, Eli Bingham, Paul Horsfall, Yee Whye Teh, Tom Rainforth, and Noah Goodman. Variational bayesian optimal experimental design. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. arXiv:1903.05480.

Adam Foster, Desi R. Ivanova, Ilyas Malik, and Tom Rainforth. Deep adaptive design: Amortizing sequential bayesian experimental design. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pp. 3384–3395, 2021. arXiv:2103.02438.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 6577–6595, 2024. doi: 10.18653/v1/2024.naacl-long.366. arXiv:2311.08298.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007.

I. J. Good. Studies in the history of probability and statistics. XXXVII A. M. Turing's statistical work in World War II. *Biometrika*, 66(2):393–396, 1979. doi: 10.1093/biomet/66.2.393.

Irving John Good. *Probability and the Weighing of Evidence*. Charles Griffin, 1950.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, et al. Accelerating scientific discovery with co-scientist. *Nature*, 655(8122):487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Gabriel Grand, Valerio Pepe, Joshua B. Tenenbaum, and Jacob Andreas. Shoot first, ask questions later? building rational agents that explore and act like people. In *International Conference on Learning Representations (ICLR)*, 2026.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1321–1330, 2017.

Ritwik Gupta, Rodolfo Corona, Jiaxin Ge, Eric Wang, Dan Klein, Trevor Darrell, and David M. Chan. Enough coin flips can make LLMs act Bayesian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 7634–7655, 2025. doi: 10.18653/v1/2025.acl-long.377. arXiv:2503.04722.

Danny Halawi, Fred Zhang, Chen Yueh-Han, and Jacob Steinhardt. Approaching human-level forecasting with language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.18563.

Kunal Handa, Yarin Gal, Ellie Pavlick, Noah Goodman, Jacob Andreas, Alex Tamkin, and Belinda Z. Li. Bayesian preference elicitation with language models. *arXiv preprint arXiv:2403.05534*, 2024.

Minsheng Hao, Yongju Lee, Hanchen Wang, Gabriele Scalia, and Aviv Regev. PerTurboAgent: An LLM-based agent for designing iterative Perturb-Seq experiments. In *Proceedings of Machine Learning in Computational Biology (MLCB)*, volume 311 of *PMLR*, pp. 44–64, 2025.

John Hawthorne, Daniel Rothschild, and Levi Spectre. Belief is weak. *Philosophical Studies*, 173 (5):1393–1404, 2016.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. TabPFN: A transformer that solves small tabular classification problems in a second. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2207.01848.

Noah Hollmann, Samuel Müller, Lennart Purucker, Arjun Krishnakumar, Max Körfer, Shi Bin Hoo, Robin Tibor Schirrmeister, and Frank Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637(8045):319–326, 2025. doi: 10.1038/s41586-024-08328-6.

Zhiyuan Hu, Chumin Liu, Xidong Feng, Yilun Zhao, See-Kiong Ng, Anh Tuan Luu, Junxian He, Pang Wei Koh, and Bryan Hooi. Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2402.03271.

Kexin Huang, Romain Lopez, Jan-Christian Hütter, Takamasa Kudo, Antonio Rios, and Aviv Regev. Sequential optimal experimental design of perturbation screens guided by multi-modal priors. In *Research in Computational Molecular Biology (RECOMB)*, Lecture Notes in Computer Science, pp. 17–37, 2024. doi: 10.1007/978-1-0716-3989-4_2.

John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

Peter Jansen, Marc-Alexandre Côté, Tushar Khot, Erin Bransom, Bhavana Dalvi Mishra, Bodhisattwa Prasad Majumder, Oyvind Tafjord, and Peter Clark. DiscoveryWorld: A virtual environment for developing and evaluating automated scientific discovery agents. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2406.06769.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4090–4102, 2025. doi: 10.18653/v1/2025.emnlp-main.203. arXiv:2506.04410.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics (TACL)*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, et al. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022.

Robert E. Kass and Adrian E. Raftery. Bayes factors. *Journal of the American Statistical Association*, 90(430):773–795, 1995. doi: 10.1080/01621459.1995.10476572.

Minsu Kim, Sangryul Kim, and James Thorne. From evidence to belief: A Bayesian epistemology approach to language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*, pp. 10578–10611, 2025. doi: 10.18653/v1/2025.naacl-long.531. arXiv:2504.19622.

Kimi Team. Kimi K2.5: Visual agentic intelligence. *arXiv preprint arXiv:2602.02276*, 2026.

Katarzyna Kobalczyk, Nicolás Astorga, Tennison Liu, and Mihaela van der Schaar. Active task disambiguation with LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2502.04485.

Zane Koch, Asmamaw T. Wassie, Javier Valdes-Aleman, Jason Lee, Michaela M. Hinks, Samuel G. Rodriques, Andrew D. White, and Jon M. Laurent. BixBench3: Benchmarking AI agents on research-study-scale computational biology tasks. *arXiv preprint arXiv:2608.25286*, 2026.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *International Conference on Learning Representations (ICLR)*, 2023. arXiv:2302.09664.

Thomas S Kuhn. *The structure of scientific revolutions*, volume 962. University of Chicago press Chicago, 1997.

Jon M. Laurent, Joseph D. Janizek, Michael Ruzo, Michaela M. Hinks, Michael J. Hammerling, Siddharth Narayanan, Manvitha Ponnapati, Andrew D. White, and Samuel G. Rodriques. LAB-Bench: Measuring capabilities of language models for biology research. *arXiv preprint arXiv:2407.10362*, 2024.

Jon M. Laurent, Albert Bou, Michael Pieler, Conor Igoe, Alex Andonian, Siddharth Narayanan, James Braza, Alexandros Sanchez Vassopoulos, Jacob L. Steenwyk, Blake Lash, Andrew D. White, and Samuel G. Rodriques. LABBench2: An improved benchmark for AI systems performing biology research. *arXiv preprint arXiv:2604.09554*, 2026.

Jixuan Leng, Chengsong Huang, Banghua Zhu, and Jiaxin Huang. Taming overconfidence in LLMs: Reward calibration in RLHF. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2410.09724.

Belinda Z. Li, Alex Tamkin, Noah Goodman, and Jacob Andreas. Eliciting human preferences with language models. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2310.11589.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan S. Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. MediQ: Question-asking llms and a benchmark for reliable interactive clinical reasoning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. arXiv:2406.00922.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research (TMLR)*, 2022. arXiv:2205.14334.

Dennis V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics (TACL)*, 12:157–173, 2024a. doi: 10.1162/tacl_a_00638. arXiv:2307.03172.

Tennison Liu, Nicolás Astorga, Nabeel Seedat, and Mihaela van der Schaar. Large language models to enhance bayesian optimization. In *International Conference on Learning Representations (ICLR)*, 2024b. arXiv:2402.03921.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding R1-zero-like training: A critical perspective. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2503.20783.

Jiaxu Lou and Yifan Sun. Anchoring bias in large language models: An experimental study. *arXiv preprint arXiv:2412.06593*, 2024.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024.

Andreas Lundh, Joel Lexchin, Barbara Mintzes, Jeppe B Schroll, and Lisa Bero. Industry sponsorship and research outcome. *Cochrane Database of Systematic Reviews*, 2017(2):MR000033, 2017. doi: 10.1002/14651858.MR000033.pub3.

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

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 *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2407.01725.

Nathan Mantel and William Haenszel. Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4):719–748, 1959.

Davide Mazzaccara, Alberto Testoni, and Raffaella Bernardi. Learning to ask informative questions: Enhancing llms with preference optimization and expected information gain. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5064–5074, 2024. doi: 10.18653/v1/2024.findings-emnlp.291. arXiv:2406.17453.

Zhiting Mei, Christina Zhang, Tenny Yin, Justin Lidard, Ola Shorinwa, and Anirudha Majumdar. Reasoning about uncertainty: Do reasoning models know when they don't know? In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 3408–3458, 2026. doi: 10.18653/v1/2026.findings-eacl.178. arXiv:2506.18183.

Ludovico Mitchener, Jon M. Laurent, Alex Andonian, Benjamin Tenmann, Siddharth Narayanan, Geemi P. Wellawatte, Andrew White, Lorenzo Sani, and Samuel G. Rodriques. BixBench: a comprehensive benchmark for LLM-based agents in computational biology. *arXiv preprint arXiv:2503.00096*, 2025a.

Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C. Landsness, Daniel L. Barabasi, Siddharth Narayanan, Nicky Evans, Shriya Reddy, Martha Foiani, Aizad Kamal, Leah P. Shriver, Fang Cao, Asmamaw T. Wassie, Jon M. Laurent, Edwin Melville-Green, Mayk Caldas, Albert Bou, Kaleigh F. Roberts, Sladjana Zagorac, Timothy C. Orr, Miranda E. Orr, Kevin J. Zwezdaryk, Ali E. Ghareeb, Laurie McCoy, Bruna Gomes, Euan A. Ashley, Karen E. Duff, Tonio Buonassisi, Tom Rainforth, Randall J. Bateman, Michael Skarlinski, Samuel G. Rodriques, Michaela M. Hinks, and Andrew D. White. Kosmos: An AI scientist for autonomous discovery. *arXiv preprint arXiv:2511.02824*, 2025b.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022. arXiv:2112.10510.

Allan H. Murphy. A new vector partition of the probability score. *Journal of Applied Meteorology*, 12(4):595–600, 1973.

Allan H. Murphy and Robert L. Winkler. Reliability of subjective probability forecasts of precipitation and temperature. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 26 (1):41–47, 1977. doi: 10.2307/2346866.

Mahdi Pakdaman Naeini, Gregory F. Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

OpenAI. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

John C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large Margin Classifiers*, pp. 61–74. MIT Press, 1999.

Ben Poole, Sherjil Ozair, Aaron van den Oord, Alexander A. Alemi, and George Tucker. On variational bounds of mutual information. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 5171–5180, 2019. arXiv:1905.06922.

Prime Intellect. PRIME-RL. https://github.com/PrimeIntellect-ai/prime-rl, 2025. Software.

Isha Puri, Mehul Damani, Idan Shenfeld, Marzyeh Ghassemi, Jacob Andreas, and Yoon Kim. Reaching beyond the mode: RL for distributional reasoning in language models. *arXiv preprint arXiv:2603.24844*, 2026.

Penghui Qi, Xiangxin Zhou, Zichen Liu, Tianyu Pang, Chao Du, Min Lin, and Wee Sun Lee. Rethinking the trust region in LLM reinforcement learning. *arXiv preprint arXiv:2602.04879*, 2026.

Bianca Raimondi and Maurizio Gabbrielli. Exploiting primacy effect to improve large language models. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP)*, pp. 989–997, 2025. arXiv:2507.13949.

Tom Rainforth, Robert Cornish, Hongseok Yang, Andrew Warrington, and Frank Wood. On nesting monte carlo estimators. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 4267–4276, 2018. arXiv:1709.06181.

Tom Rainforth, Adam Foster, Desi R Ivanova, and Freddie Bickford Smith. Modern bayesian experimental design. *Statistical Science*, 39(1):100–114, 2024.

Mayk Caldas Ramos, Shane S. Michtavy, Andrew D. White, and Marc D. Porosoff. Bayesian optimization of catalysis with in-context learning. *ACS Central Science*, 12(5):599–615, 2026. doi: 10.1021/acscentsci.5c02418. arXiv:2304.05341.

Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. Humans overrely on overconfident language models, across languages. In *Conference on Language Modeling (COLM)*, 2025. arXiv:2507.06306.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. GPQA: A graduate-level Google-proof Q&A benchmark. In *Conference on Language Modeling (COLM)*, 2024. arXiv:2311.12022.

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

Yusuf H. Roohani, Tony J. Hua, Po-Yuan Tung, Lexi R. Bounds, Feiqiao B. Yu, Alexander Dobin, Noam Teyssier, Abhinav Adduri, Alden Woodrow, Brian S. Plosky, et al. Virtual cell challenge: Toward a Turing test for the virtual cell. *Cell*, 188(13):3370–3374, 2025. doi: 10.1016/j.cell.2025.06.008.

Philipp Schoenegger, Indre Tuminauskaite, Peter S. Park, Rafael Valdece Sousa Bastos, and Philip E. Tetlock. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy. *Science Advances*, 10(45):eadp1528, 2024. doi: 10.1126/sciadv.adp1528. arXiv:2402.19379.

Kenneth F Schulz, Iain Chalmers, Richard J Hayes, and Douglas G Altman. Empirical evidence of bias: dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *JAMA*, 273(5):408–412, 1995.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27 (3):379–423, 1948. ISSN 0005-8580. doi: 10.1002/j.1538-7305.1948.tb01338.x. URL https://ieeexplore.ieee.org/document/6773024.

Maohao Shen, Subhro Das, Kristjan Greenewald, Prasanna Sattigeri, Gregory Wornell, and Soumya Ghosh. Thermometer: Towards universal calibration for large language models. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 44687–44711, 2024. arXiv:2403.08819.

Michael D. Skarlinski, Sam Cox, Jon M. Laurent, James D. Braza, Michaela Hinks, Michael J. Hammerling, Manvitha Ponnapati, Samuel G. Rodriques, and Andrew D. White. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*, 2024.

Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, James Hayhurst, Arwa Ibrahim, Yue Ji, Sajo John, Elizabeth Lewis, Jacqueline A. L. MacArthur, Aoife McMahon, David Osumi-Sutherland, Kalliope Panoutsopoulou, Zoë Pendlington, Santhi Ramachandran, Ray Stefancsik, Jonathan Stewart, Patricia Whetzel, Robert Wilson, Lucia Hindorff, Fiona Cunningham, Samuel A. Lambert, Michael Inouye, Helen Parkinson, and Laura W. Harris. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Research*, 51(D1):D977–D985, 2023.

Zayne Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning. In *International Conference on Learning Representations (ICLR)*, 2025. arXiv:2409.12183.

Jonathan AC Sterne, Jelena Savović, Matthew J Page, Roy G Elbers, Natalie S Blencowe, Isabelle Boutron, Christopher J Cates, Hung-Yuan Cheng, Mark S Corbett, Sandra M Eldridge, et al. Rob 2: a revised tool for assessing risk of bias in randomised trials. *bmj*, 366, 2019.

Yasuaki Sumita, Koh Takeuchi, and Hisashi Kashima. Cognitive biases in large language models: A survey and mitigation experiments. *arXiv preprint arXiv:2412.00323*, 2024.

Valentine Svensson, Umair Khan, Hamed Heydari, Airol A Ubas, Nicole Thomas, Daniele Merico, Hani Goodarzi, John Yu, Nima Alidoust, and Shreshth Gandhi. Back to basics: Observed statistics are sufficient to predict drug responses. *bioRxiv*, 2026. doi: 10.64898/2026.06.09.731197.

Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of AI agents designs new SARS-CoV-2 nanobodies. *Nature*, 646(8085):716–723, 2025. doi: 10.1038/s41586-025-09442-9.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, 2023. doi: 10.1038/s41586-023-06139-9.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5433–5442, 2023. doi: 10.18653/v1/2023.emnlp-main.330. arXiv:2305.14975.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, et al. SciCode: A research coding benchmark curated by scientists. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. arXiv:2407.13168.

Amos Tversky and Derek J. Koehler. Support theory: A nonextensional representation of subjective probability. *Psychological Review*, 101(4):547–567, 1994.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Joon Oh. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 15440–15459, 2024. doi: 10.18653/v1/2024.acl-long.824. arXiv:2403.05973.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7534–7550, 2020. doi: 10.18653/v1/2020.emnlp-main.609. arXiv:2004.14974.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. SciFact-Open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 4719–4734, 2022. doi: 10.18653/v1/2022.findings-emnlp.347. arXiv:2210.13777.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. ScienceWorld: Is your agent smarter than a 5th grader? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11279–11298, 2022. doi: 10.18653/v1/2022.emnlp-main.775. arXiv:2203.07540.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations (ICLR)*, 2023a. arXiv:2203.11171.

Yiwei Wang, Yujun Cai, Muhao Chen, Yuxuan Liang, and Bryan Hooi. Primacy effect of ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 108–115, 2023b. doi: 10.18653/v1/2023.emnlp-main.8. arXiv:2310.13206.

Zhiting Wei, Yiheng Wang, Yicheng Gao, Shuguang Wang, Ping Li, Duanmiao Si, Yuli Gao, Siqi Wu, Danlu Li, Kejing Dong, Xingbo Yang, Chen Tang, Shaliu Fu, Xiaohan Chen, Wannian Li, Yuzhou You, Chen Zhang, Aibin Liang, Guohui Chuai, and Qi Liu. Benchmarking algorithms for generalizable single-cell perturbation response prediction. *Nature Methods*, 23(2):451–464, 2026. doi: 10.1038/s41592-025-02980-0.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5985–5998, 2024. doi: 10.18653/v1/2024.emnlp-main.343. arXiv:2405.20974.

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. *arXiv preprint arXiv:2504.08066*, 2025.

Jensen Zhang, Jing Yang, and Keze Wang. Large language models as discounted Bayesian filters. *arXiv preprint arXiv:2512.18489*, 2025a.

Jesse Zhang, Airol A. Ubas, Richard de Borja, Valentine Svensson, Nicole Thomas, Neha Thakar, Ian Lai, et al. Tahoe-100M: A giga-scale single-cell perturbation atlas for context-dependent gene function and cellular modeling. *bioRxiv*, 2025b. doi: 10.1101/2025.02.20.639398.

Yicong Zhao, King Yeung Tsang, Harshil Vejendla, Haizhou Shi, Zhuohang Li, Zhigang Hua, Qi Xu, Tunyu Zhang, Yi Wang, Ligong Han, Bradley A. Malin, and Hao Wang. Dist2ill: Distributional distillation for one-pass uncertainty estimation in large language models. *arXiv preprint arXiv:2505.11731*, 2025.

Chiwei Zhu, Benfeng Xu, Quan Wang, Yongdong Zhang, and Zhendong Mao. On the calibration of large language models and alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9778–9795, 2023. doi: 10.18653/v1/2023.findings-emnlp.654. arXiv:2311.13240.

APPENDIX CONTENTS

A EDB-Bio construction details **21**

A.1 Domains . . . 21

A.2 Design principles . . . 21

A.3 Tahoe: drug perturbation of cancer cell lines . . . 22

A.4 DepMap: cancer gene dependency . . . 23

A.5 GWAS: genetic associations . . . 24

A.6 Dataset splits and test-set provenance . . . 25

A.7 Known limitations of dataset construction . . . 26

B Model variants and training details **28**

C Evaluation protocol **30**

D Frontier baseline settings **34**

E Evidence accumulation and per-update belief revision **34**

E.1 Cumulative evidence accumulation . . . 34

E.2 Direction of belief revision . . . 35

E.3 Evidence by relevance tier . . . 36

F Full cross-domain transfer results **37**

G Experimental design details **39**

G.1 Naive-prompting selection prompt . . . 39

H Post-hoc recalibration **39**

I Prompts: one worked example per domain **43**

I.1 Tahoe: drug perturbation of cancer cell lines . . . 43

I.2 DepMap: cancer gene dependency . . . 45

I.3 GWAS: genetic associations . . . 46

J A thinking-mode reasoning trace **47**

J.1 Prompt . . . 48

J.2 Trace . . . 51

K Additional analyses of belief revision **52**

K.1 Composition of supporting and refuting evidence . . . 53

K.2 Dependence of evidence updates on sequence position . . . 54

L Structured and numeric-only baselines on Tahoe **56**

M Extended related work **56**

A EDB-BIO CONSTRUCTION DETAILS

This appendix expands §2: an overview of the three domains (App. A.1), the design principles applied to all of them (App. A.2), the construction and audit of each domain (Apps. A.3 to A.5), the dataset splits and test-set provenance (App. A.6), and a known limitation of the dataset construction (App. A.7). Each domain is built by a single seeded script that emits EDB-Bio together with a manifest recording every parameter; the scripts are released with the code.

A.1 DOMAINS

In each domain, the target measurement is excluded from the evidence pool, and the amount, relevance, and strength of the remaining evidence are controlled during construction. Each stub presents the measurement with its scientific context. In Tahoe, this is the perturbing drug's mechanism, approval status, and chemical structure, and the cell line's lineage and driver mutations; in DepMap and GWAS, it is analogous assay annotations and study provenance, such as gene function, variant consequence, cohort ancestry, and sample size. Table 1 in the main text compares the domains, Apps. A.3 to A.5 give construction details, and App. I gives a complete prompt from each domain.

Tahoe: drug perturbation of cancer cell lines. The Tahoe-100M atlas (Zhang et al., 2025b) measures transcriptional responses to small-molecule perturbations across cancer cell lines, supporting the study of drug mechanism and of how response varies across cellular contexts. A hypothesis states that "drug D up/down-regulates gene G in cell line C"; its held-out DESeq2 measurement supports the claim when $\mathrm{padj} < 0.05$, $|\log_2 \mathrm{FC}| > 0.5$, and the direction matches. Evidence consists of differential-expression measurements of G under other drug and cell-line conditions, ordered by relevance from the same drug in another cell line, through related mechanisms and other drugs, to a genome-wide aggregate summary and unrelated controls.

DepMap: cancer gene dependency. The Broad Institute Cancer Dependency Map (DepMap) (DepMap, Broad, 2024) uses pooled CRISPR screens to identify genes that cancer cells require for growth, with applications to therapeutic-target discovery and to mapping context-specific cancer vulnerabilities. A hypothesis states that "gene G is a dependency in cell line C"; the held-out screen supports the hypothesis when the Chronos gene-effect score is at most -0.5 and the calibrated dependency probability is at least 0.5 (Dempster et al., 2021). Evidence is multi-modal and ordered by how directly it bears on G's dependency in C: the dependency of G in related cell lines, then co-essentiality (genes whose across-cell-line dependency profiles correlate with G's, implying shared function), pharmacological response (the sensitivity of C to drugs that target G or its pathway (Corsello et al., 2020)), and the molecular state of G in C (its expression, mutation, and copy-number status). The dependencies of unrelated genes serve as irrelevant controls.

GWAS: genetic associations. The NHGRI-EBI GWAS Catalog (Sollis et al., 2023) curates published associations between human genetic variants and traits, providing a resource for mapping inherited risk and trait biology. A hypothesis states that "SNP S (risk allele A) influences trait T in a stated direction"; the held-out association supports the hypothesis when $p < 5 \times 10^{-8}$ and the effect direction matches. Evidence consists of other catalog associations, ordered by relevance from the same gene or variant and a related trait, through pathway-neighbor associations, to unrelated loci and traits.

A.2 DESIGN PRINCIPLES

The source datasets are imbalanced and contain assay-specific shortcuts. We apply five principles to remove shortcuts based on label prevalence or assay artifacts. Table 2 records how each principle is instantiated in each domain; the domain sections below give the corresponding selection rules and audits.

P1: No artifact or trivial labels. Target measurements dominated by technical artifacts are excluded, as are entities whose label is nearly constant across contexts. The restriction applies to targets; excluded entities may still appear as evidence.

P2: Diversity across claim axes. Targets span the biological entities and contexts on which we evaluate generalization, with caps or stratification to prevent any one group from dominating. Each domain is balanced $50{:}50$ between supported and not-supported targets.

P3: A graded relevance hierarchy. Evidence pools range from measurements closely related to the target to deliberately unrelated controls. The ordering within each domain's hierarchy was defined by PhD-level bioinformaticians with peer-reviewed publications in the relevant biological fields, rather than by an automated heuristic. In Tahoe, for example, the hierarchy runs from the same drug in another cell line, through drugs with the same mechanism and other drugs in the same cell line, to unrelated controls.

P4: Diversity in evidence sufficiency. Episodes are divided approximately equally among evidence-rich, -moderate, and -poor strata. A calibrated model should become confident when evidence is conclusive and remain uncertain when it is not.

P5: No target leakage. The held-out target measurement is excluded from its own evidence pool, and every domain is audited for exact matches. Domain-specific near-duplicate risks, such as variants in high linkage disequilibrium, are discussed in the domain sections below. Overlap between episodes across splits is a separate matter, treated in App. A.7.

	Tahoe	**DepMap**	**GWAS**
P1 Targets	Mitochondrial, ribosomal, pseudogene and housekeeping-lncRNA genes excluded, as are near-zero base-mean measurements	Pan-essential and non-expressed genes excluded; targets drawn from the 2,480 *selective*-dependency genes (essential in 2–50% of lines)	MHC/HLA variants excluded (long-range linkage disequilibrium makes single-variant labels unreliable), as are heavily replicated "textbook" pairs
P2 Diversity	49 cell lines, 179 drugs across mechanisms, 13,083 genes; balanced within every cell line	30 lineages, thousands of genes, at most 16 episodes per gene	Thousands of genes and traits across EFO parent categories, with per-gene and per-trait caps; one episode per unique (SNP, trait) pair
P3 Hierarchy	Within-assay: same drug, other line $\succ$ same mechanism $\succ$ other drug, same line $\succ$ aggregate summary $\succ$ unrelated control	Cross-modal: same gene, same lineage $\succ$ same gene, other lineage $\succ$ co-essential partner $\succ$ drug targeting G $\succ$ molecular state of G in C $\succ$ unrelated control	Locus and trait proximity: other SNP in gene, same trait $\succ$ same gene, related trait $\succ$ target SNP, related trait $\succ$ target SNP, unrelated trait $\succ$ pathway neighbor $\succ$ unrelated control
P4 Sufficiency	Significance-rich (≥ 4 significant stubs), -moderate (1–3), or -poor (none)	Rich, moderate or poor in dependency-supporting evidence: dependent related lines, over-expression, amplification, damaging mutation, drug sensitivity	Rich, moderate or poor by capping significant related stubs. Non-significant related evidence is scarce in a catalog of significant hits, so poor pools fill largely with controls
P5 Leakage	No pool stub shares the target's exact (drug, cell line) pair	Same-gene stubs come only from other lines; the molecular state of G in C is a different assay from the held-out fitness readout, so it is evidence rather than leakage	The exact (SNP, trait) pair is excluded. A different SNP in high linkage disequilibrium carries nearly the same signal; filtering it needs an LD reference panel and is deferred

Table 2: How each EDB-Bio design principle is instantiated in the three domains. Full criteria and audits are in Apps. A.3 to A.5.

A.3 TAHOE: DRUG PERTURBATION OF CANCER CELL LINES

We use the Tahoe-100M pseudobulk differential-expression tables, which report one DESeq2 result for each drug $\times$ cell-line $\times$ gene combination and are stored one cell line per file. Rows with failed

DESeq2 fits are discarded, and a result is *significant* iff $\text{padj} < 0.05 \wedge |\log_2 \text{FC}| > 0.5$. Significant results are rare (roughly 9% of measurements). Principles P1–P5 (App. A.2) are applied as follows.

Target principles

P1 (No artifact labels). Mitochondrial, ribosomal, and pseudogene genes, and measurements with near-zero base mean, are excluded as targets because their differential-expression calls reflect technical properties of the assay rather than drug action. Both remain admissible as evidence. The realized artifact-gene target rate is 0%.

P2 (Diversity across claim axes). Targets span 49 cell lines (no line contributes more than 2.0% of episodes), 179 drugs across mechanisms of action, and $13{,}083$ genes. All available cell lines and a shared set of drugs are retained so that every target has same-drug, different-cell-line evidence. Because significant results are rare, targets are drawn with fixed per-cell-line quotas from pre-classified candidate sets. A significant measurement yields a supported target (claimed direction equal to the observed one) or a refuted target (opposite direction); a non-significant measurement yields an inconclusive target. The supported rate is 0.50 within every cell line and 49.9% overall.

Evidence principles

P3 (Relevance diversity). Unlike those of DepMap and GWAS, the Tahoe hierarchy is *within-assay*: every stub is a differential-expression measurement of the target gene under another $(\text{drug}, \text{cell line})$ condition. We rank a stub's relevance to the target, most to least, as: (1) the same drug in another cell line; (2) a drug with the same mechanism of action; (3) a different drug in the same cell line; (4) a genome-wide aggregate summary for the target drug and cell line; and (5) an unrelated control. Collapsing (1)–(4) into *related* and (5) into *control*, we fill each pool to an $\approx 60{:}40$ ratio, drawing from the top of the hierarchy first. The realized split is 62.5% related and 37.5% control (same-drug 24%, same-mechanism 18%, different-drug 18% of stubs), and same-drug evidence is present in 100% of episodes. Each stub is tagged with its relation to the target and carries the drug and cell-line context listed in App. A.1.

P4 (Sufficiency diversity). Episodes vary in how much conclusive (significant) related evidence their pool contains: rich (≥ 4 significant related stubs), moderate (1–3), or poor (none), in equal thirds. Conclusive related stubs are added from highest to lowest relevance until the stratum's quota is met; inconclusive related stubs are then added round-robin across relevance levels, followed by one genome-wide aggregate summary and unrelated controls. Rich episodes carry 5.7 conclusive stubs on average against 0.8 for poor episodes, and 88% of all stubs are inconclusive.

P5 (No leakage). A target's own held-out measurement never appears in its pool, and no pool stub shares the target's exact $(\text{drug}, \text{cell line})$ pair; both are enforced by construction.

Class balance and the negative split. Supported and not-supported targets are balanced $50{:}50$. A Tahoe *refuted* outcome (a significant change opposite to the claimed direction) is comparatively common, so negatives are sampled at approximately 67% inconclusive and 33% refuted.

A.4 DepMap: cancer gene dependency

Dependencies comprise 9.7% of valid DepMap gene $\times$ cell-line pairs. Principles P1–P5 (App. A.2) are applied as follows.

Target principles

P1 (No artifact labels). Pan-essential (common-essential) genes, which are dependencies in nearly every line, and non-expressed genes, which cannot be dependencies, are excluded as targets. Both may still appear as (uninformative) evidence.

P2 (Diversity across claim axes). Targets span the 30 lineages and thousands of genes over which we make claims, with no stratum dominating (at most 16 episodes per gene). Unlike Tahoe, DepMap has no target-drug or mechanism axis: the perturbation is always a knockout, so drugs enter only as evidence. Because dependencies are rare (close to Tahoe's $\sim 9\%$ significant rate), targets are sampled to obtain a $50{:}50$ supported-to-not-supported balance.

Targets are drawn from the $2{,}480$ *selective*-dependency genes, which are essential in some cell lines but not others (2–50% of lines), so that the support label varies across cell-line contexts.

Evidence principles

P3 (Relevance diversity). We rank a stub's relevance to the target, most to least, as: (1) the same gene knocked out in another cell line of the same lineage; (2) the same gene in a different lineage; (3) a co-essential pathway partner of G knocked out in C; (4) a drug targeting G or its pathway profiled in C; (5) the molecular state of G in C (expression, copy number, mutation); and (6) an unrelated gene's fitness in C. Collapsing (1)–(5) into *related* and (6) into *irrelevant*, we fill each pool to an $\approx 60{:}40$ ratio, drawing from the top of the hierarchy first and guaranteeing multi-modal coverage. This hierarchy is *cross-modal*, spanning genetic, pharmacological and molecular evidence. As in Tahoe, each stub is tagged with its relation to the target (e.g. *target gene knocked out in a different cell line of the same lineage*), which keeps the prompt structure common across domains. Each stub also carries its context: the gene (name, family, identifiers), the cell line (lineage, disease, subtype, demographics, driver mutations), and for pharmacological stubs the drug (targets, mechanism, SMILES).

P4 (Sufficiency diversity). Episodes vary in how much evidence *supporting* a dependency their pool contains (related cell lines that are themselves dependencies, over-expression, amplification, damaging mutation, drug sensitivity): pools are built to be sufficiency-rich, -moderate, or -poor.

P5 (No leakage). A target's own held-out CRISPR measurement never appears in its pool; same-gene stubs are drawn only from *other* cell lines. The target gene's *molecular* state in C (expression, copy number, mutation) is measured by a different assay from the held-out fitness readout and is retained as evidence.

Class balance and the negative split. Supported and not-supported targets are balanced $50{:}50$, as in Tahoe. DepMap's *refuted* outcome, a knockout that *speeds* growth, is biologically rare (only 2.7% of measurements have $\mathrm{Chronos} > 0.25$), so negatives are sampled at approximately 85% inconclusive and 15% refuted.

A.5 GWAS: GENETIC ASSOCIATIONS

The GWAS Catalog is a curated collection of published genome-wide-significant associations and contains few natural negatives. Principles P1–P5 (App. A.2) are applied as follows.

Target principles

P1 (No artifact/trivial labels). Variants in the MHC/HLA region (chr6, $\sim$25–34 Mb) are excluded as targets because their extreme long-range linkage disequilibrium makes single-variant labels unreliable. Heavily replicated "textbook" pairs (those studied in dozens of prior GWAS) are also excluded to reduce dependence on frequently reported associations. Both classes remain admissible as (uninformative or trivially informative) evidence.

P2 (Diversity across claim axes). Targets span thousands of genes and traits across the EFO parent categories over which we make claims, with per-gene *and* per-trait caps so that no stratum dominates; one episode corresponds to one unique (SNP, trait) pair. Unlike Tahoe, GWAS has no perturbation: it is observational, so the "experiment" is the held-out association study itself. Because positives (genome-wide-significant, direction-consistent associations) are abundant and negatives scarce, targets are sampled across outcome classes to obtain a $50{:}50$ supported-to-not-supported balance.

Evidence principles

P3 (Relevance diversity). We rank a stub's relevance to the target, most to least, as: (1) a different SNP in the target gene associated with the same trait; (2) a different SNP in the target gene associated with a trait in the same EFO parent (related trait); (3) the target SNP associated with a different trait in the same EFO parent; (4) the target SNP associated with a trait in a different EFO parent; (5) a SNP in a pathway-neighbor gene associated with the target trait; and (6) an unrelated SNP and trait (different gene, different EFO parent).

This hierarchy is graded by both *genomic* proximity (same SNP → same gene → pathway neighbor → unrelated locus) and *phenotypic* proximity (same trait → related trait → unrelated trait). Collapsing (1)–(5) into *related* and (6) into *control*, we fill each pool toward a related:control target of approximately 60:40, drawing from the top of the hierarchy first. Each stub is tagged with its relation to the target.

P4 (Sufficiency diversity). Episodes vary in how much genome-wide-significant related evidence their pool contains; pools are built in roughly equal thirds to be sufficiency-rich, -moderate, or -poor by capping the number of significant related stubs. Because the catalog contains only significant hits, non-significant *related* evidence is scarce, so the "poor" pools fill largely with unrelated controls and the realized related:control ratio falls below its target.

P5 (No leakage). A target's own held-out (SNP, trait) association never appears in its pool; same-gene and same-SNP stubs are drawn only from *other* SNPs or *other* traits, and the exact-pair exclusion is audited to zero. This check does not detect variants in high linkage disequilibrium with the target, which may convey nearly the same signal. Filtering such variants requires an LD reference panel and is not performed.

Class balance and the negative split. Supported and not-supported targets are balanced 50:50, as in Tahoe and DepMap. A GWAS *direction-inconsistent* outcome is a genome-wide-significant association whose effect points in the opposite direction. It is rare relative to a *sub-threshold* association, so negatives are sampled at $\approx 85\%$ sub-threshold and 15% direction-inconsistent. This split concerns the *target* outcome. At the *evidence* level, GWAS has no refuting-stub category: once the target association is held out, other catalog associations can corroborate the hypothesis or be uninformative, but do not directly refute it. The directional analyses therefore omit GWAS (App. A.2).

A.6 Dataset splits and test-set provenance

Dataset splits. Each domain contains $32{,}000$ training episodes, 256 separate validation episodes reserved for model selection, and $2{,}000$ test episodes; both test configurations below are drawn from the test split. Within each domain, labels are balanced approximately 50:50, and evidence order is randomized independently of relevance. During training, the context length is sampled as $k \sim \mathrm{Uniform}\{2, \ldots, 16\}$. The sensitivity set has the same per-domain size as the validation split but is disjoint from it.

Differences between domains. All three domains use evidence pools of at most 16 stubs. They differ in the composition of the not-supported class: approximately 33% of not-supported targets are refuted in Tahoe and 15% in DepMap and GWAS (direction-inconsistent associations), with the remainder inconclusive or sub-threshold. GWAS also has no refuting *evidence* stubs (App. A.5).

Test-set configurations. The primary and sensitivity test sets are statically materialized and published as one versioned artifact. Each row holds one episode with its evidence pool truncated to the specified context length and corresponds to one forward pass. All models receive identical rendered inputs, including context length and evidence order. A domain column supports both per-domain and macro-averaged analyses from the same tables.

In the *primary* set, each episode appears once at a context length within the training range. It is used for the master table, reliability and ROC panels, and leave-one-out transfer matrix. In the *sensitivity* set, each episode appears at every context length from 0 to 16. The resulting nested sweep is used for dose–response and weight-of-evidence analyses.

Primary set. The primary set contains $2{,}000$ episodes per domain and $6{,}000$ in total, comprising the complete held-out TEST split of each domain. Each episode is assigned one context length $k \in \{2, \ldots, 16\}$, matching the `context_size_range` $= [2, 16]$ under which every trained model saw its episodes.

Context lengths are allocated in fixed counts. Each of the 15 values receives $\lfloor 2000/15 \rfloor = 133$ episodes, and $k \in \{12, \ldots, 16\}$ receives one additional episode, for 134. Uniform random assignment would have a multinomial standard deviation near 11 episodes per group; fixed allocation removes that variation.

Labels are close to balanced by construction: the supported rate is 50.0% for both GWAS and DepMap and 49.9% for Tahoe.

	Primary	Sensitivity
Episodes per domain	2,000	256
Episodes, all three domains	6,000	768
Context lengths per episode	1, from $k \in \{2, \ldots, 16\}$	17, all of $k \in \{0, \ldots, 16\}$
Rows per domain	2,000	4,352
Rows, all three domains	6,000	13,056
Models evaluated	10	7
Forward passes	60,000	91,392

Table 3: Primary and sensitivity test-set configurations. Both are published under [repository withheld for anonymous review] at revision `48389e59`, as six configs (two sets × three domains).

Sensitivity set. The sensitivity set contains 256 episodes per domain, each drawn from the corresponding primary set and scored at all 17 context lengths $k \in \{0, 1, \ldots, 16\}$: 4,352 rows per domain, 13,056 per model. The context at k is the first k stubs of the episode's fixed, pre-shuffled evidence pool, so the sweep is nested. The stub revealed at step j is the same for every model and at every $k \geq j$. This permits per-stub attribution in the weight-of-evidence analysis, where each increment is a paired difference between consecutive context lengths within an episode.

The $k = 0$ and $k = 1$ conditions are below the training range. At $k = 0$, the model receives the hypothesis without evidence, providing the reference forecast used by the cumulative weight-of-evidence metric. The $k = 1$ condition measures the initial response to evidence.

Deterministic construction. Within each domain, TEST episodes are ordered by the SHA-256 hash of the episode identifier. Context lengths are assigned in balanced blocks over this order, and the first 256 episodes form the sensitivity set. Selection is therefore independent of source-row order. Rebuilding with seed 0 reproduces all six parquet files byte for byte.

Provenance and release. Both sets are versioned together as the dataset [repository withheld for anonymous review]. Every number in this paper was computed against revision `48389e59`, in full `48389e5909bbac19c1aa6f442d6130bda2c4bdd7`. It was built from EDB-Bio, published as [repository withheld for anonymous review] at revision `98042211`. Reproductions should pin this revision because results from different revisions are not directly comparable; the revision is also part of the on-disk path of each result file. Release terms for the test sets, builder, and EDB-Bio are given in the code and data availability statement.

A.7 Known limitations of dataset construction

EDB-Bio assigns episodes to the training, validation, and test splits by i.i.d. sampling (App. A.6), so test targets can overlap with training evidence. Principle P5 excludes a target's own measurement from its own evidence pool but does not constrain other episodes: the held-out target of one episode may serve as an evidence stub in another. When the two episodes fall on opposite sides of the split, a model trained on EDB-Bio has seen the test target's readout during training. We measure the extent of this overlap and its effect on the reported results, which is small, and leave its removal to a future revision of the dataset.

Extent of the overlap. For each test episode we check whether its held-out measurement, identified by the domain's target key, appears as an evidence stub within the first 16 positions of any training or validation pool, which is the portion of a pool that training renders. Table 4 gives the result. Overlap affects 0.6% of test episodes in Tahoe, 10.1% in DepMap, and 53.2% in GWAS. A stricter check, whether a test hypothesis itself recurs as a training target, finds no cases in DepMap or GWAS and 11 in Tahoe, all of them one measurement claimed in opposite directions.

The spread across domains follows from the size of each domain's measurement space (the set of possible target keys) relative to the number of stubs that the training pools draw from it. Tahoe targets come from a product of 179 drugs, 49 cell lines, and 13,083 genes, about 115 million combinations, and the training pools expose 0.39% of them. DepMap's space has two axes and about 2.4 million pairs, and the restriction to selective-dependency genes under P1 concentrates about 16 episodes on each of 2,047 genes, so the pools expose 15%. GWAS is not a product space: a (SNP, trait) pair exists only if a study reported it, and the relevance tiers enumerate sibling records in the same gene or at the same SNP rather than sampling them. The training pools cover 92.5% of the associations

that the dataset touches. Under uniform sampling the overlap rate would equal this coverage, and Tahoe and DepMap fall within a factor of 1.5 of it. The construction choices that limit identity shortcuts, a narrow target set under P1 and relevance-first pool filling under P3, are the same choices that raise the overlap rate.

Domain	Measurement key	Measurement space	Exposed as stubs	Coverage	Test overlap
Tahoe	(drug, cell line, gene)	114,750,993	450,307	0.39%	11 (0.6%)
DepMap	(gene, cell line)	2,411,366	365,141	15.1%	202 (10.1%)
GWAS	(SNP, trait)	198,001*	183,166	92.5%	1,065 (53.2%)

Table 4: Overlap between test targets and training evidence. *Measurement space* is the number of possible target keys; *exposed as stubs* is the number of distinct measurements that appear as an evidence stub in the first 16 positions of any training or validation pool; *coverage* is their ratio; *test overlap* is the number of the 2,000 test episodes per domain whose held-out measurement is among the exposed stubs. *GWAS has no product space, so the entry is the number of distinct (SNP, trait) associations that appear anywhere in the dataset, and coverage is relative to that set.

Effect on the reported metrics. We recompute Brier score, AUROC, and sMECE on the primary test set with the overlapping episodes removed, for all seven models (Fig. 6). GWAS moves the most, as its overlap rate predicts. The absolute changes are small: across the three Ibis0 variants, macro-averaged Brier changes by at most 0.005 and macro-averaged AUROC by 0.006. The shifts are similar in sign and size across models, and no ranking changes. The zero-shot models, which were never trained on EDB-Bio, also score better on the overlapping episodes than on the remainder, and on GWAS their change on removal exceeds that of the Ibis0 RL variants. The overlapping episodes are therefore easier in their own right, and most of the movement is a selection effect rather than memorization. The one pattern consistent with memorization is on DepMap, where Brier score improves on the overlapping subset for the trained models and worsens for the untrained ones; it amounts to about 0.004 in the full DepMap Brier score.

Addressing the overlap. A split grouped by target entity closes the overlap for Tahoe and DepMap, whose pools are gene-scoped, but not for GWAS, whose pools cross gene, SNP, and trait. Alternatively, pools can be built after the split with test measurements excluded, which keeps the in-distribution framing at the cost of a rebuild. We intend to adopt one of these in a future revision.

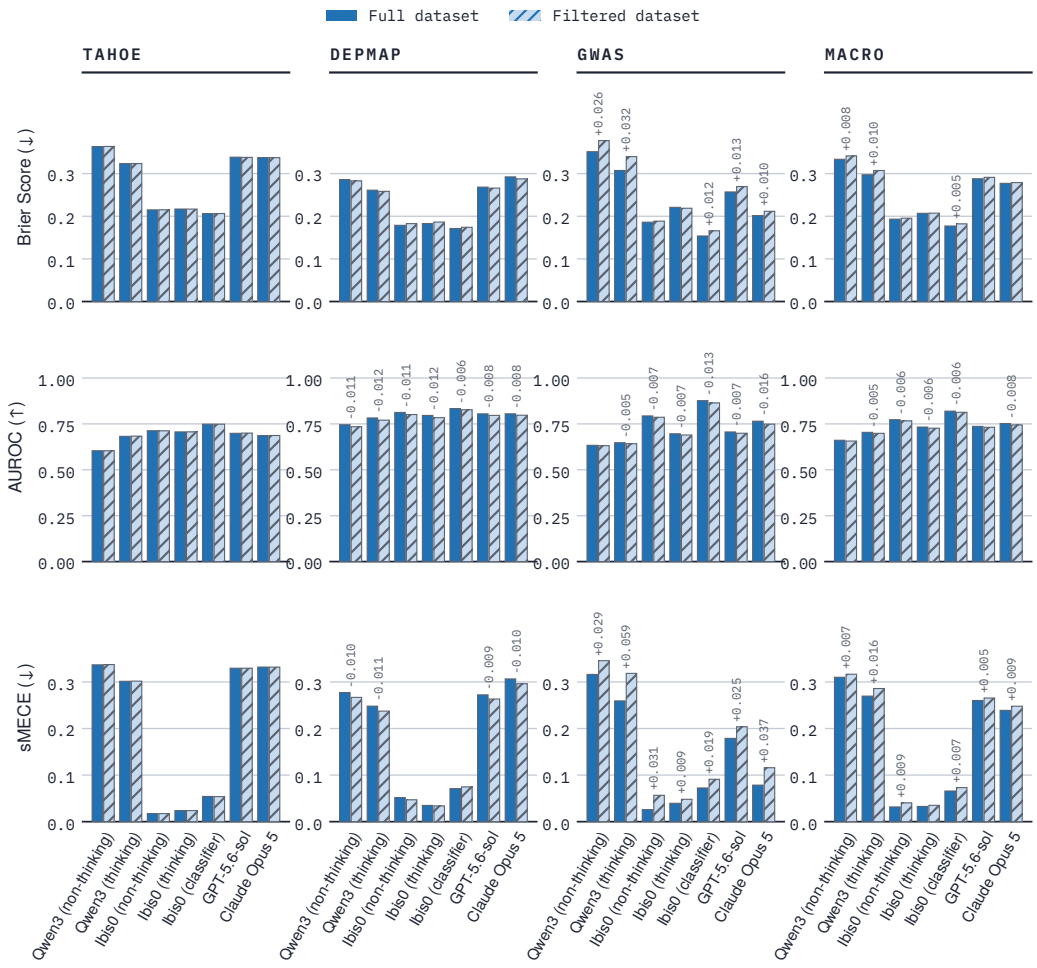

Figure 6: **Metrics with and without the overlapping test episodes.** Solid bars use the full 2,000-episode primary test set in each domain; hatched bars use the filtered dataset, which removes the episodes whose held-out measurement appears in a training or validation pool (Table 4). Rows are metrics and columns are domains, with the equal-weight macro-average on the right. Labels above a bar pair give the filtered value minus the full value where the difference reaches 0.005.

B Model variants and training details

Shared setup. All post-training runs start from `Qwen3-30B-A3B`. The runs comprise one classifier and two RL policies trained on all three domains and three classifiers trained on leave-one-domain-out mixtures (Table 5). They use the same training split, assign equal weight to each included domain, and render each episode with $K \sim \mathrm{Uniform}\{2, \ldots, 16\}$ evidence stubs. K is determined from the episode identifier, so a given episode has the same length across runs.

Optimization is AdamW at learning rate 9×10^{-6}, weight decay 0.01 and gradient-norm clipping at 1.0, with 225 linear warmup steps and no subsequent decay. These settings are shared by all training mixtures, including the leave-one-out runs.

Checkpoint selection. Each domain reserves a 256-episode validation split that is separate from the test episodes and from the equally sized sensitivity subset of the test set. Each run is evaluated on this held-out split every 100 steps (200 for Ibis0 (thinking), whose evaluation dominated its wall-clock time) and stops when its selection metric has gone 1,000 steps without improving (minimum improvement 10^{-4} for the classifiers, 10^{-3} for RL). No run reached the 9,000-step safety ceiling. Five runs stopped on this patience criterion; the GWAS-held-out classifier terminated on

a checkpoint-write error 799 steps into its patience window, after its selected checkpoint had been written.

Classifier runs are selected by validation cross-entropy and RL runs by mean validation reward over three evaluation environments (256 episodes each for Ibis0 (non-thinking), 128 for Ibis0 (thinking), 4 rollouts per episode). Table 5 reports the selected checkpoints used in all evaluations.

Run	Training domains	Selection metric	Best value	Step
Ibis0 (classifier)	Tahoe, DepMap, GWAS	in-domain BCE ↓	0.4785	2400
Ibis0 (non-thinking)	Tahoe, DepMap, GWAS	eval reward ↑	0.8304	4200
Ibis0 (thinking)	Tahoe, DepMap, GWAS	eval reward ↑	0.8168	2400
Ibis0 (classifier), Tahoe held out	DepMap, GWAS	in-domain BCE ↓	0.4390	1700
Ibis0 (classifier), DepMap held out	Tahoe, GWAS	in-domain BCE ↓	0.4794	1300
Ibis0 (classifier), GWAS held out	Tahoe, DepMap	in-domain BCE ↓	0.5096	1800

Table 5: Training runs and selected checkpoints. All runs start from `Qwen3-30B-A3B` and use the same optimizer settings; they differ in objective and training mixture. In-domain BCE is averaged over the trained domains only, so it is not comparable across rows with different mixtures.

Classifier. We replace the language-modeling head with a single linear unit ($\mathbb{R}^d \to \mathbb{R}$ for hidden size d) on the last-token hidden state, whose sigmoid is the predicted probability of support, and fine-tune the whole transformer with this head (no frozen backbone) under binary cross-entropy against the support label.

All classification runs use a sequence length of 16,384 tokens, which accommodates the longest render in any domain. One episode is processed per data-parallel rank per micro-step, with 16 gradient-accumulation steps on 8 ranks, giving 128 episodes per optimizer step. The classifier produces a continuous scalar forecast (Fig. 3) rather than a verbalized probability.

Leave-one-out classifiers. The transfer results in §4 come from three additional classification runs that use the same configuration but train on two domains. All three are evaluated on every domain. As for the all-domain classifier, checkpoints are selected by validation cross-entropy, here macro-averaged over the two *training* domains and excluding the omitted one.

RL runs. The two RL runs use `prime-rl` (Prime Intellect, 2025) under the reward $r = 1 - (p - y)^2$, where $y \in \{0, 1\}$ is the observed binary support label for the target hypothesis (§2) and p is the probability the model states given the revealed evidence. A completion with no parseable probability receives $r = -1$. Advantages are group-relative: each rollout's reward minus the mean reward of the rollouts sampled for its prompt, with *no* standard-deviation normalization. This is the unbiased, group-mean-baselined advantage of Dr. GRPO (Liu et al., 2025) rather than the standardized advantage of GRPO. Groups whose advantages are all zero are dropped before the trainer sees them. In place of PPO-style ratio clipping, the policy is updated with `prime-rl`'s default trust-region objective, the binary total-variation variant of DPPO (Qi et al., 2026). It masks a token out of the update when its sampling probability moves by more than 0.2 in the trust-region-violating direction (conditioned on the advantage sign) and leaves the importance ratio unclipped on the kept tokens. A squared-log-ratio KL penalty (coefficient 10^{-3}, following Kimi-K2.5 (Kimi Team, 2026)) regularizes the update. The policies are trained, validated and evaluated at sampling temperature 0.7.

Direct and thinking. The two RL runs share objective, reward, data and optimizer, and differ in Qwen3's thinking mode. Ibis0 (non-thinking) has thinking disabled and a 512-token completion budget: the model emits its answer immediately as a `<probability>`p`</probability>` tag, with no chain of thought. It trains at sequence length 9,216, with 32 rollouts per optimizer step and 8 rollouts per example. Ibis0 (thinking) has thinking enabled and a 24,576-token budget, which forces sequence length 32,768; it trains with 16 rollouts per step and 4 per example. The reward reads only the stated probability in both cases.

Use in subsequent analyses. The classifier is included in the weight-of-evidence and dose–response analyses. The models used as policies and as belief models in the multi-domain experimental-design evaluation will be reported with those results (§6).

C EVALUATION PROTOCOL

This appendix specifies the estimators, resampling procedures, and decoding rules used for the metrics defined in § 3.2.

Macro-averaging. Each aggregate statistic is computed separately within Tahoe, DepMap, and GWAS and then averaged with equal weight. This procedure prevents domain-specific metric distributions from changing the effective domain weights.

If a statistic is undefined for a resampled domain, as can occur for AUROC when a resample contains one class, that domain is omitted from the resampled macro-average. This behavior is implemented for the macro bootstrap and transfer-matrix statistic; the master-table implementation raises an error. No such case occurs in the 2,000-episode point estimates.

Within-domain stratified bootstrap. Each bootstrap draw resamples episodes with replacement within each domain, holding the per-domain count fixed, and recomputes the per-domain statistics and their macro-average. For the aggregate results on the primary set (the master table, headline figure, grouped-metric panels, transfer matrix and ROC AUROC table), intervals are the 2.5th and 97.5th percentiles over 2,000 draws. Other figures use fewer draws: 1,000 for the dose–response and weight-of-evidence curves, 500 for the per-bin reliability and ROC bands, and 250 for the per-position and signed-verdict intervals. Point estimates are computed on the complete evaluation sample.

Resampling within domains keeps domain weights fixed across draws. Pooling episodes before resampling would instead estimate uncertainty for a micro-average with variable domain proportions.

The sMECE estimator is $O(n^2)$ in the number of episodes and costs about 35 times as much per draw as the other aggregate metrics, so its intervals use 200 draws. Its point estimate uses the complete sample.

Episode-level resampling. Each episode contributes one forecast to the aggregate metrics and up to 16 updates to the per-stub analyses. Updates from the same episode share a target hypothesis, an evidence pool and an ordering, so they are strongly correlated. The bootstrap for per-stub quantities is therefore *episode-clustered*: it samples complete episodes with all of their stubs, preserving within-episode dependence.

Relevance-hierarchy quantities are macro-averaged over domains using episode clustering. Signed weight-of-evidence results are reported separately by domain because verdict categories differ across domains; their intervals use within-domain episode-clustered bootstraps.

Calibration of weights of evidence. This paragraph gives the estimators of the odds ratio of support ψ and the slope β (§ 3.2), which Analysis 2 of §5 reports. The unit of analysis is a single belief update. On the sensitivity set each model states a probability p_k for every episode at each evidence count $k \in \{0, \dots, 16\}$, and each consecutive pair of forecasts gives one update (ℓ_i, w_i, θ_i): the incoming belief $\ell_i = \operatorname{logit}(p_{k-1})$, the weight of evidence

$$w_i = \frac{10}{\ln 10}\left[\operatorname{logit}(p_k) - \operatorname{logit}(p_{k-1})\right]$$

in decibans (§2), and the label $\theta_i \in \{0, 1\}$ of the episode, which all of its updates share. Each model contributes 256×16 updates per domain. Probabilities are clipped to $[10^{-3},\ 1 - 10^{-3}]$ before the logit. Updates with $w_i = 0$, where the stated probability does not change, are excluded. They range from 1% of updates for Ibis0 (classifier) to 68% for Ibis0 (non-thinking); models that state probabilities in text often repeat the same value (§4).

Tiers and belief strata. Separately for each model and domain, the remaining updates are cut into five quantile bins of w, the *tiers*, and five quantile bins of ℓ, the *belief strata*; tied quantile edges are merged. The *reference tier* is the tier whose mean w, averaged over the three domains, is closest to zero. It is chosen once per model and shared by all three domains.

Odds ratio of support. For tier t and belief stratum s, let a_s and b_s count the supported ($\theta = 1$) and unsupported ($\theta = 0$) updates in tier t, let c_s and d_s count the same for the reference tier, and let $n_s = a_s + b_s + c_s + d_s$. The odds ratio of support of tier t is the Mantel–Haenszel estimate pooled

over belief strata (Mantel & Haenszel, 1959), in decibans:

$$\psi_t = \frac{10}{\ln 10} \ln \frac{\sum_s a_s d_s / n_s}{\sum_s b_s c_s / n_s}.$$

This odds ratio compares the odds that the hypothesis is supported after an update in tier t with the same odds after a reference update from the same incoming belief. Pooling within belief strata makes ψ independent of where the belief started. When either sum is zero, as happens when a tier separates the labels perfectly, 0.5 is added to every cell before pooling. Each tier is plotted at $x_t = \bar{w}_t - \bar{w}_{\mathrm{ref}}$, its mean weight of evidence relative to the reference tier, so the reference tier sits at the origin with $\psi = 0$. The points are then averaged over the three domains by tier rank.

Slope. If a model's probabilities are calibrated, $\psi_t = x_t$ for every tier, up to sampling error. We summarize the points by their least-squares slope through the origin,

$$\beta = \frac{\sum_t x_t \, \psi_t}{\sum_t x_t^2} = \frac{\sum_t x_t^2 \, (\psi_t / x_t)}{\sum_t x_t^2},$$

where the sums run over the non-reference tiers. The slope is a weighted average of the per-tier ratios ψ_t / x_t with weights x_t^2, so the tiers with the largest updates count most. A value of $\beta = 1$ indicates calibrated updates, $\beta < 1$ indicates overreaction (updates larger than the outcomes justify), and $\beta > 1$ indicates underreaction (updates smaller than the outcomes justify). The quantity $\beta - 1$ is the log-odds analog of the coefficient in a regression of forecast errors on forecast revisions, which is negative when forecasters overreact to news and positive when they underreact (Coibion & Gorodnichenko, 2015; Bordalo et al., 2020).

Uncertainty. Each episode contributes up to 16 updates that share its label, so the updates are not independent. Intervals come from the episode-clustered bootstrap described above: each draw resamples episodes within each domain, holds the bin edges and the reference tier fixed, and recomputes the points and the slope.

Sampling noise. Each p_k of a generative model is a single sample (see *Decoding* below), so part of the variation in w is sampling noise that is unrelated to the outcome. Binning on a noisy w biases β toward zero and therefore overstates overreaction for these models. The classifier is deterministic and is unaffected.

Stub verdicts. Each evidence stub is assigned a verdict relative to the episode's target hypothesis h^*, judged against h^*'s claim rather than the stub's own effect. In **Tahoe** the verdict is directional: a stub is *supported* when its differential-expression measurement is significant ($\mathrm{padj} < 0.05$, $|\log_2 \mathrm{FC}| > 0.5$) in the direction h^* claims, *refuted* when significant in the opposite direction, and *inconclusive* otherwise. In **DepMap**, where h^* ("G is a dependency in C") is directionless, a stub is *supported* when it indicates a dependency ($\mathrm{Chronos} \leq -0.5$, $P(\mathrm{dep}) \geq 0.5$), *refuted* when it indicates a suppressor ($\mathrm{Chronos} > 0.25$), and *inconclusive* otherwise. **GWAS** has no refuting category: a stub is *significant* when $p < 5 \times 10^{-8}$ and *sub-threshold* otherwise. Because effect-direction concordance across different variants and traits is not biologically guaranteed, a direction-flipped association is not treated as refuting.

Paired comparisons. Comparisons reported as differences use a paired bootstrap over models evaluated on identical episodes. Each iteration applies one episode resample to every model in the comparison. Leave-one-out deltas and retention intervals use this procedure. Intervals for individual rows in the master table are unpaired.

Calibration: sMECE in place of ECE. Expected calibration error (ECE) (Guo et al., 2017) averages the absolute difference between each bin's mean forecast and outcome rate, weighted by bin occupancy. Our implementation computes ECE over the same ten equal-width bins used for the Murphy terms, but the reported calibration statistic is sMECE because ECE depends on the number and placement of bin edges.

For sMECE, we estimate the empirical frequency $r(p) = \mathbb{E}[y \mid p]$ by Nadaraya–Watson regression of the outcomes on the forecasts,

$$\hat{r}(p_i) \;=\; \frac{\sum_j K(p_i, p_j)\, y_j}{\sum_j K(p_i, p_j)}, \qquad K(p_i, p_j) \;=\; \phi\Big(\tfrac{p_i - p_j}{h}\Big) + \phi\Big(\tfrac{p_i + p_j}{h}\Big) + \phi\Big(\tfrac{2 - p_i - p_j}{h}\Big),$$

with ϕ the standard Gaussian density, and summarize the fit as the mean of $|\hat{r}(p_i) - p_i|$ over the evaluated episodes. Averaging at the observed forecasts weights the error by forecast density.

The second and third kernel terms reflect the Gaussian across 0 and 1 to preserve boundary mass. The bandwidth follows Silverman's rule on the forecasts, $h = 1.06\,\hat{\sigma} n^{-1/5}$, with a minimum value of 0.02 to prevent a nearly degenerate forecast distribution from producing a near-zero bandwidth.

This is an estimator in the smooth-ECE family introduced by Błasiok & Nakkiran (2024), not their estimator: the bandwidth rule and the boundary handling are ours. We therefore refer to the statistic as sMECE.

Murphy decomposition. The Murphy decomposition separates the Brier score into reliability, resolution, and uncertainty terms (Murphy, 1973; Murphy & Winkler, 1977; DeGroot & Fienberg, 1983):

$$\text{Brier Score} = \frac{1}{N}\sum_i (p_i - \theta_i)^2 = \underbrace{\sum_k w_k(\bar{p}_k - \bar{\theta}_k)^2}_{\text{Reliability}} - \underbrace{\sum_k w_k(\bar{\theta}_k - \bar{\theta})^2}_{\text{Resolution}} + \underbrace{\bar{\theta}(1-\bar{\theta})}_{\text{Uncertainty}}, \quad (3)$$

where k indexes ten equal-width bins, w_k is the bin's share of episodes, $\bar{p}_k$ and $\bar{\theta}_k$ are its mean forecast and observed outcome rate, and $\bar{\theta}$ is the base rate. Bins are right-closed, with the first bin including 0.

Reliability measures within-bin calibration error and is lower when mean forecasts match observed frequencies. Resolution measures variation in observed outcome rates across forecast bins and is higher when forecasts separate outcomes. Uncertainty depends only on the label distribution and is identical across models evaluated on a fixed test set; for the balanced test set it is 0.25.

The identity is exact for binned forecasts. The tables report unbinned mean squared error, so the right-hand side of Eq. (3) differs from the reported Brier score by within-bin dispersion. We use the decomposition to interpret differences among models.

Reading intervals near zero. Reliability and sMECE are non-negative and positively biased: resampling injects noise that a squared or absolute deviation can only turn into a larger value. For a row whose true calibration error is close to zero, the bootstrap distribution can lie above the point estimate, leaving the point estimate outside its own interval. Such rows are flagged in the tables. These intervals support comparisons among rows but should not be interpreted as coverage intervals for a near-zero point estimate.

Decoding. The generative rows are sampled with a fixed seed at temperature 0.7, at which the RL policies are also trained and validated (App. B). On the primary set each episode is sampled four times and the reported probability is the mean of parsed draws. On the sensitivity set, where every episode is rendered at all 17 context lengths, each episode is drawn once. Both estimators target the same quantity; the four-draw mean has lower sampling variance. The classifier is deterministic: its probability is the sigmoid of a scalar head applied to the last-token representation. Frontier models use provider-default sampling settings as described in App. D. All evaluated completions produced a readable probability, and no episodes are excluded.

Complete results table. Table 6 reports the macro-average for every model with its per-domain breakdown, the number of episodes scored, and a bootstrap interval on every value.

Model	n	Brier Score (↓)	sMECE (↓)	AUROC (↑)	Reliability (↓)	Resolution (↑)
Qwen3 (non-thinking)	6,000	0.3336 [0.3237, 0.3436]	0.3101 [0.2990, 0.3229]	0.6608 [0.6471, 0.6752]	0.1096 [0.1024, 0.1189]	0.0250 [0.0226, 0.0295]
Tahoe	2,000	0.3634 [0.3459, 0.3827]	0.3367 [0.3201, 0.3589]	0.6039 [0.5805, 0.6279]	0.1277 [0.1140, 0.1452]	0.0147 [0.0111, 0.0204]
DepMap	2,000	0.2860 [0.2715, 0.2995]	0.2773 [0.2575, 0.2978]	0.7454 [0.7239, 0.7671]	0.0854 [0.0754, 0.0977]	0.0485 [0.0414, 0.0577]
GWAS	2,000	0.3513 [0.3333, 0.3694]	0.3162 [0.2966, 0.3391]	0.6333 [0.6080, 0.6571]	0.1156 [0.1022, 0.1314]	0.0120 [0.0089, 0.0176]
Qwen3 (thinking)	6,000	0.2973 [0.2896, 0.3047]	0.2696 [0.2607, 0.2818]	0.7040 [0.6914, 0.7170]	0.0829 [0.0772, 0.0907]	0.0339 [0.0307, 0.0387]
Tahoe	2,000	0.3235 [0.3103, 0.3373]	0.3014 [0.2770, 0.3244]	0.6822 [0.6586, 0.7049]	0.0988 [0.0877, 0.1131]	0.0224 [0.0172, 0.0288]
DepMap	2,000	0.2611 [0.2492, 0.2725]	0.2481 [0.2301, 0.2678]	0.7824 [0.7630, 0.8029]	0.0751 [0.0659, 0.0859]	0.0626 [0.0546, 0.0722]
GWAS	2,000	0.3074 [0.2933, 0.3213]	0.2593 [0.2385, 0.2822]	0.6474 [0.6228, 0.6721]	0.0748 [0.0644, 0.0877]	0.0165 [0.0124, 0.0231]
Ibis0 (non-thinking)	6,000	0.1933 [0.1888, 0.1977]	0.0316 [0.0258, 0.0450]	0.7729 [0.7616, 0.7846]	0.0020 [0.0017, 0.0038]	0.0579 [0.0540, 0.0631]
Tahoe	2,000	0.2149 [0.2083, 0.2215]	0.0174 [0.0141, 0.0389]	0.7128 [0.6902, 0.7348]	0.0007 [0.0006, 0.0028]	0.0347 [0.0289, 0.0424]
DepMap	2,000	0.1787 [0.1693, 0.1881]	0.0517 [0.0340, 0.0677]	0.8125 [0.7931, 0.8314]	0.0042 [0.0026, 0.0078]	0.0747 [0.0663, 0.0845]
GWAS	2,000	0.1861 [0.1781, 0.1944]	0.0258 [0.0155, 0.0431]	0.7934 [0.7731, 0.8127]	0.0010 [0.0006, 0.0035]	0.0644 [0.0559, 0.0739]
Ibis0 (thinking)	6,000	0.2069 [0.2025, 0.2111]	0.0327 [0.0296, 0.0431]	0.7329 [0.7209, 0.7450]	0.0023 [0.0021, 0.0044]	0.0448 [0.0413, 0.0500]
Tahoe	2,000	0.2169 [0.2102, 0.2235]	0.0238 [0.0172, 0.0456]	0.7064 [0.6839, 0.7286]	0.0016 [0.0009, 0.0042]	0.0347 [0.0287, 0.0424]
DepMap	2,000	0.1828 [0.1747, 0.1907]	0.0348 [0.0287, 0.0522]	0.7961 [0.7765, 0.8160]	0.0028 [0.0020, 0.0055]	0.0689 [0.0607, 0.0787]
GWAS	2,000	0.2209 [0.2128, 0.2289]	0.0394 [0.0314, 0.0558]	0.6961 [0.6730, 0.7198]	0.0024 [0.0015, 0.0059]	0.0307 [0.0249, 0.0392]
Ibis0 (classifier)	6,000	0.1771 [0.1707, 0.1831]	0.0658 [0.0588, 0.0762]	0.8195 [0.8099, 0.8295]	0.0075 [0.0065, 0.0104]	0.0790 [0.0753, 0.0849]
Tahoe	2,000	0.2063 [0.1971, 0.2157]	0.0541 [0.0389, 0.0707]	0.7485 [0.7275, 0.7683]	0.0042 [0.0028, 0.0082]	0.0470 [0.0407, 0.0556]
DepMap	2,000	0.1713 [0.1602, 0.1821]	0.0710 [0.0589, 0.0886]	0.8334 [0.8163, 0.8500]	0.0076 [0.0054, 0.0124]	0.0855 [0.0780, 0.0952]
GWAS	2,000	0.1537 [0.1416, 0.1654]	0.0724 [0.0584, 0.0882]	0.8766 [0.8617, 0.8915]	0.0107 [0.0078, 0.0160]	0.1046 [0.0964, 0.1153]
GPT-5.6-sol	6,000	0.2879 [0.2793, 0.2964]	0.2602 [0.2490, 0.2712]	0.7365 [0.7235, 0.7485]	0.0830 [0.0775, 0.0908]	0.0430 [0.0395, 0.0484]
Tahoe	2,000	0.3385 [0.3228, 0.3549]	0.3295 [0.3095, 0.3479]	0.6987 [0.6769, 0.7208]	0.1204 [0.1072, 0.1360]	0.0304 [0.0247, 0.0377]
DepMap	2,000	0.2683 [0.2547, 0.2829]	0.2723 [0.2533, 0.2888]	0.8046 [0.7846, 0.8237]	0.0880 [0.0773, 0.1003]	0.0665 [0.0581, 0.0765]
GWAS	2,000	0.2570 [0.2426, 0.2720]	0.1788 [0.1623, 0.2005]	0.7061 [0.6828, 0.7287]	0.0408 [0.0339, 0.0509]	0.0321 [0.0261, 0.0407]
Claude Opus 5	6,000	0.2772 [0.2689, 0.2851]	0.2390 [0.2291, 0.2505]	0.7519 [0.7392, 0.7646]	0.0786 [0.0730, 0.0857]	0.0502 [0.0461, 0.0556]
Tahoe	2,000	0.3376 [0.3229, 0.3525]	0.3320 [0.3114, 0.3498]	0.6863 [0.6640, 0.7082]	0.1142 [0.1017, 0.1291]	0.0254 [0.0205, 0.0320]
DepMap	2,000	0.2923 [0.2776, 0.3068]	0.3065 [0.2880, 0.3244]	0.8048 [0.7848, 0.8240]	0.1132 [0.1010, 0.1269]	0.0683 [0.0598, 0.0780]
GWAS	2,000	0.2016 [0.1907, 0.2124]	0.0785 [0.0646, 0.0951]	0.7647 [0.7435, 0.7862]	0.0083 [0.0059, 0.0136]	0.0569 [0.0495, 0.0669]
constant p = base rate	6,000	0.2500 [0.2500, 0.2500]	0.0000† [0.0027, 0.0187]	0.5000 [0.5000, 0.5000]	0.0000† [0.0000, 0.0004]	0.0000 [0.0000, 0.0000]
Tahoe	2,000	0.2500 [0.2500, 0.2500]	0.0000 [0.0000, 0.0250]	0.5000 [0.5000, 0.5000]	0.0000† [0.0000, 0.0006]	0.0000 [0.0000, 0.0000]
DepMap	2,000	0.2500 [0.2500, 0.2500]	0.0000† [0.0005, 0.0276]	0.5000 [0.5000, 0.5000]	0.0000† [0.0000, 0.0006]	0.0000 [0.0000, 0.0000]
GWAS	2,000	0.2500 [0.2500, 0.2500]	0.0000† [0.0005, 0.0265]	0.5000 [0.5000, 0.5000]	0.0000† [0.0000, 0.0007]	0.0000 [0.0000, 0.0000]

Macro-average over Tahoe, DepMap, GWAS; 95% CI from a within-domain bootstrap resampling episodes (2,000 draws; sMECE 200, it is 35x costlier per draw). The unindented row of each block is that macro-average; the indented rows below it are the single domains, bootstrapped over that domain's episodes alone. The constant row ignores all evidence; the test set is balanced 50:50 by construction, so it scores ≈0.25 Brier. Rows above that line do not beat guessing. † point estimate outside its bootstrap interval, as expected for the positively-biased binned metrics (Reliability, sMECE) on rows near the zero floor. Compare rows by these intervals; do not read them as covering the point.

Table 6: Calibration and discrimination on the held-out multi-domain test set, in full.

D FRONTIER BASELINE SETTINGS

The frontier baselines, GPT-5.6-sol (OpenAI) and Claude Opus 5 (Anthropic), are evaluated on the same episodes as the other models, with reasoning effort and output budget set explicitly for each provider. Reasoning effort is `high` for both, mapped to `output_config.effort` on Anthropic and `reasoning_effort` on OpenAI. The output budget is 24,000 tokens for both, mapped to `max_tokens` and `max_completion_tokens`, respectively; the local thinking-mode models use the same budget. On Anthropic, this limit includes both reasoning and response tokens, and truncation before the final response produces an unreadable probability (App. C).

The local generative models are decoded at temperature 0.7 with a fixed seed, matching their training and validation temperature (App. B). The classifier does not sample: a scalar head reads its probability off the last-token representation. Claude Opus 5 rejects temperature, top-p and top-k, so the frontier rows use provider-default sampling. Both frontier models receive the system prompt of the local thinking-mode models.

Declines and API errors are recorded separately from unreadable completions, so policy refusals are not counted as parse failures. Remaining per-provider API details are in the released evaluation code.

E EVIDENCE ACCUMULATION AND PER-UPDATE BELIEF REVISION

This appendix reports the evidence-accumulation and per-stub analyses summarized in §5. All use the sensitivity set, which presents each of 256 episodes per domain at every evidence count $k \in \{0, \ldots, 16\}$.

E.1 CUMULATIVE EVIDENCE ACCUMULATION

Fig. 7 reports the mean probability assigned to the true label (supported or not supported) as evidence accumulates,

$$p_k^{\text{true}} = \theta p_k + (1-\theta)(1-p_k).$$

Every model assigns more probability to the true label at $k = 16$ than at $k = 0$ (Fig. 7); the paired gain excludes zero in 19 of the 21 model–domain curves. The exceptions are the two frontier models on DepMap, whose mean changes are small relative to their uncertainty. Starting points vary widely: Ibis0 (classifier) begins at 0.729 on GWAS before seeing any evidence and gains only 0.050, whereas on DepMap it begins near 0.5 and gains 0.181.

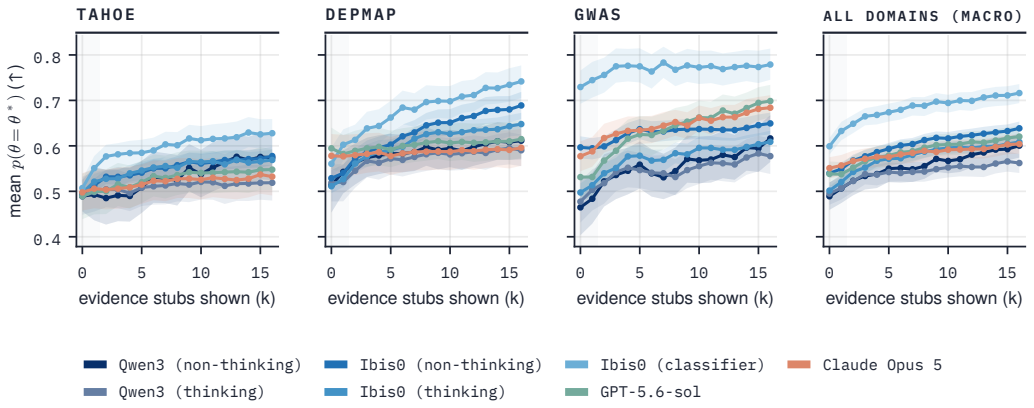

Figure 7: **Confidence in the true label increases as evidence accumulates.** Mean probability each model assigns to the true label vs. the number of evidence stubs shown (256 episodes per domain). The slope reflects integration of new evidence (up to $k = 16$ stubs; shading marks $k < 2$, below the range seen in training). The intercept reflects prior knowledge: $p > 0.5$ at $k = 0$ means that the model predicts the target hypothesis *with no evidence*. Above-chance *a priori* predictions on GWAS, most notably by Ibis0 (classifier), reflect non-random association of alleles in high linkage disequilibrium in EDB-Bio (see App. A.5).

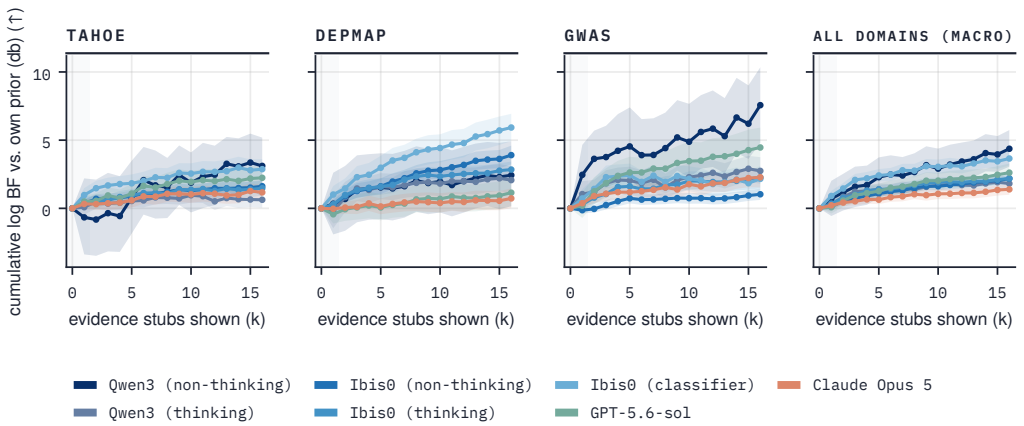

Figure 8: **Cumulative evidence-induced log-odds change by domain.** Cumulative log Bayes factor toward the observed outcome, $s\,[\mathrm{logit}(p_k) - \mathrm{logit}(p_0)]$ in decibans, against the number of stubs shown. Each curve is relative to that model's own $k = 0$ forecast. Panels show the three domains and their macro-average on a shared y axis; bands are 95% intervals from within-domain episode resampling.

This value depends on the prior p_0, the model's belief in the hypothesis at the start of the episode, and several models in Fig. 7 start at $p_0^{\text{true}} > 0.5$. We therefore also measure how far the evidence moves each model from its k=0 belief. The *cumulative weight of evidence* is the running sum of the per-stub updates w_j (§2) and, for an ideal Bayesian, the log Bayes factor of the revealed evidence. We orient it toward the ground truth:

$$w_k^{\text{cum}} = s\,[\mathrm{logit}(p_k) - \mathrm{logit}(p_0)] \cdot \frac{10}{\log 10}, \quad s = \begin{cases} +1, & \theta = 1 \\ -1, & \theta = 0 \end{cases}$$

so that w_k^{cum}, in decibans, is positive when evidence has pushed belief toward the true outcome (Fig. 8). The metric is *prior-agnostic*: it is zero at k=0 by construction and measures how far a model updates, independent of its starting belief. It is *cumulative*: differencing against step $k-1$ gives the marginal counterpart, $\log BF_k - \log BF_{k-1}$. It is also *direction-independent*: the per-episode sign s swaps the numerator and denominator of BF_k on episodes that are not supported, so a positive value means that the stubs moved belief in the correct direction whether or not the target hypothesis is supported.

By k=16, models differ roughly threefold in how far the evidence moves them: Qwen3 (non-thinking) moves farthest toward the truth ($+4.37$ dB $[+3.05, +5.74]$), then Ibis0 (classifier) ($+3.65$ $[+3.14, +4.17]$), while Claude Opus 5 moves least ($+1.39$ $[+0.99, +1.76]$). The most responsive model, Qwen3 (non-thinking), also has the worst Brier score. Movement this large can reflect strong response to useful evidence or indiscriminate overreaction; the per-stub analyses in Apps. E.2 and E.3 and the calibration of weights of evidence in §5 separate the two.

E.2 Direction of belief revision

Direction of belief revision. We test whether models revise belief in the expected direction on evidence that plausibly supports or refutes the target hypothesis h^*. In each domain, stubs are sorted into heuristic *supporting* and *refuting* categories by their direction relative to h^*. For example, when h^* says drug D up-regulates gene G, a related drug that significantly up-regulates G counts as supporting and one that significantly *down*-regulates it as refuting; a stub with no statistically significant effect is *inconclusive*. Full per-domain definitions are given in App. C.

Belief revisions have the expected signs (Fig. 9): every model's mean update is positive on supporting stubs, negative on refuting stubs, and small on inconclusive ones. Magnitudes differ substantially across models. Because all models see the same stubs in the same order, the difference reflects how strongly each model weighs the same category of evidence. Qwen3 (non-thinking) produces the largest updates in both directions, and §5 shows that they overreact; its reinforcement-learned counterpart Ibis0 (non-thinking) responds more selectively (App. E.3).

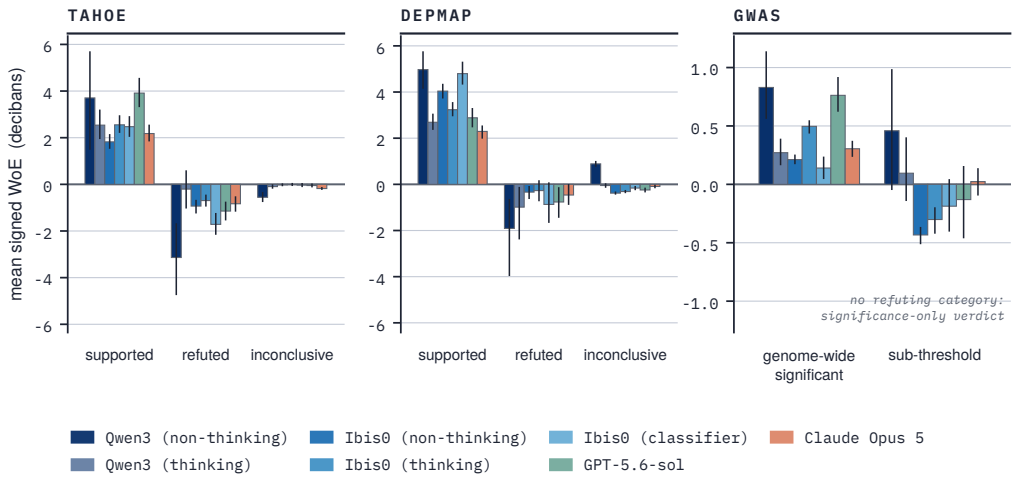

Figure 9: **Do models distinguish supporting vs. refuting evidence?** Mean signed w_j in decibans; positive raises belief in the target hypothesis and negative lowers it. Every model's mean update has the expected sign for supporting and refuting evidence, and most models place little weight on inconclusive evidence. (GWAS admits no refuting evidence (App. A.2), so its panel contrasts genome-wide-significant with sub-threshold associations.)

E.3 Evidence by relevance tier

Evidence by relevance tier. A model that deliberates well should reserve its largest updates for the evidence that bears most directly on the target. Fig. 10 plots each model's mean $|w_j|$ by relevance tier and summarizes its alignment with the curated hierarchy (App. A.2) as a per-domain Spearman correlation ρ between $|w_j|$ and tier rank (positive when more-relevant stubs draw larger updates). Among the trained models, Ibis0 (non-thinking) aligns most closely in every domain ($\rho = 0.22$, 0.37, 0.18 on Tahoe, DepMap, and GWAS), despite trailing the classifier on final Brier and AUROC (the classifier's correlations are only 0.09, 0.19, 0.09). Aggregate forecast quality and the structure of individual updates therefore capture different behaviors. Selectivity is also distinct from responsiveness, the overall magnitude of belief change: Qwen3 (non-thinking) produces large updates on average yet aligns only weakly with the hierarchy ($\rho = 0.07$ on Tahoe).

The cross-model ordering in Fig. 10 also tests the hierarchy itself. Agreement is strongest on DepMap. On GWAS, models over-weight indirect same-SNP evidence relative to the curated top tier, and on Tahoe, Claude Opus 5 places disproportionate weight on aggregate summary evidence over more-specific assay readouts. These discrepancies may reflect model failures, limitations of the expert ordering, or both.

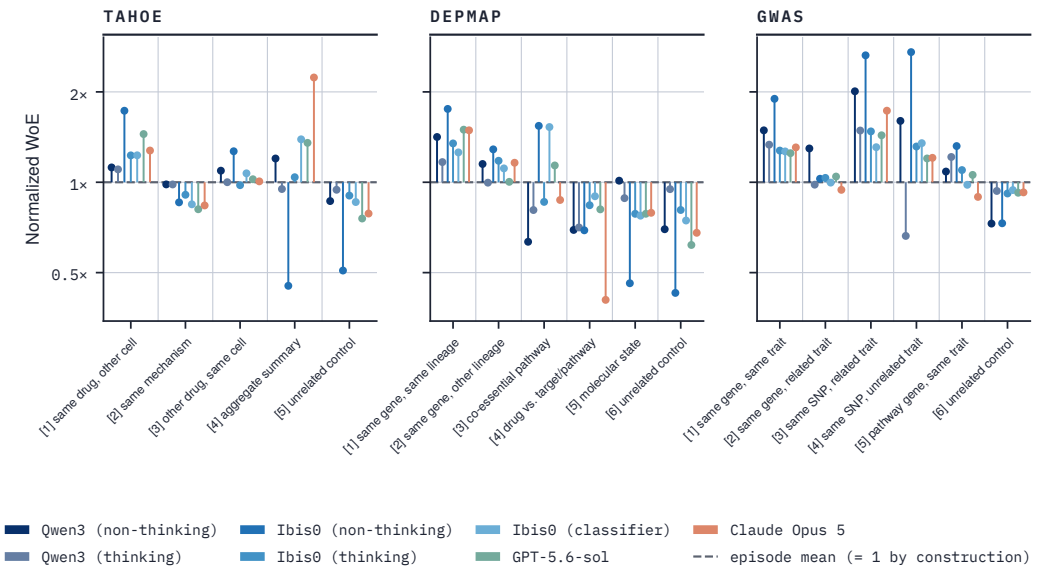

Figure 10: **Do models assign higher weight to more relevant evidence in an expert biological knowledge hierarchy?** Each stem shows how strongly a tier of evidence moves a model, relative to that model's average stub in the same episode ($1\times$ is average). Tiers run left to right in the expert-ranked order, so a model that follows the hierarchy shows heights falling from left to right. On DepMap, the models largely follow the expert ordering. On GWAS, models assign more weight to indirect evidence (same SNP) than to more direct evidence (same gene, same trait). On Tahoe, Claude gives disproportionately high weight to aggregate summary evidence over more-specific assay readouts.

F Full cross-domain transfer results

Raw performance on the omitted domain is hard to interpret because Qwen3 and the all-domain classifier set a different floor and ceiling for each metric. We therefore normalize the leave-one-domain-out result between these two references. For domain d and metric m, we report

$$\text{retention} = \frac{m_{\text{leave-out-}d} - m_{\text{untrained}}}{m_{\text{all domains}} - m_{\text{untrained}}} \qquad \text{on domain } d, \tag{4}$$

where 100% denotes the full improvement from the untrained model (Qwen3) to the all-domain classifier and 0% denotes no improvement over Qwen3. For higher-is-better metrics, both differences are reversed. Paired bootstrap intervals use identical episodes for all three models.

Table 7 reports the complete numerical results underlying Fig. 11 and §4. Block A adds the all-domain reference and macro-averages omitted from the main figure. Block B reports the held-out-domain decomposition in absolute units, as a paired loss relative to the all-domain classifier, and on the normalized retention scale.

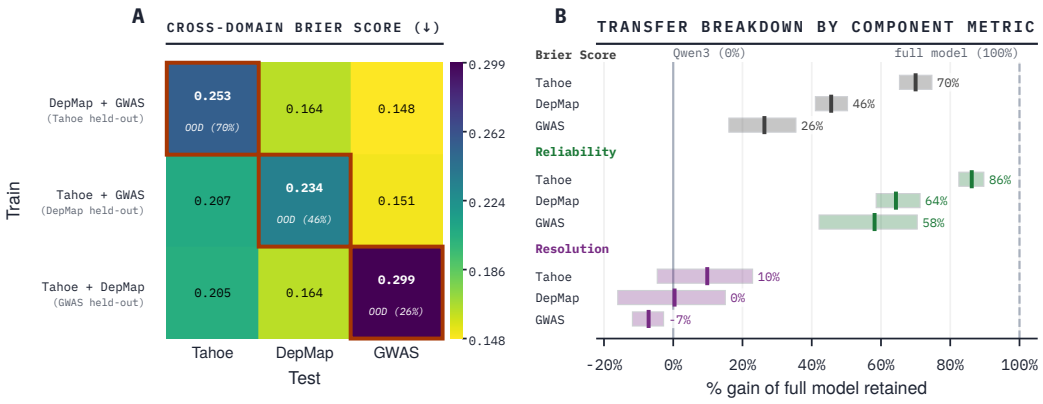

Figure 11: **Transfer to domains excluded from training. A:** We train Ibis0 (classifier) on two domains (rows) and evaluate on a third held-out domain (columns). Outlined diagonal cells are o.o.d. evaluations on the excluded domain and report the raw Brier score and the retention percentage (Eq. (4)). **B:** Retention on the excluded domains, decomposed into Brier score and its Murphy reliability and resolution terms. Boxes show paired-bootstrap 95% intervals and internal lines show point estimates; vertical lines mark Qwen3 (0%) and the all-domain classifier (100%). Reliability retains 58–86% of the full-model improvement, whereas effectively no resolution improvement is retained.

Table 7: **Full cross-domain transfer results. A:** Brier score for every classifier training mix and test domain, with 95% bootstrap intervals. Bold cells evaluate a domain excluded from training; other per-domain cells are descriptive in-domain evaluations. Macro is the equal-domain average and is bootstrapped directly as a stratified macro statistic. **B:** Exact results on the excluded domain for Brier score and its Murphy reliability and resolution terms. The Qwen3 and full-model references denote the released checkpoint and the all-domain classifier, respectively. Loss vs. full is signed so that positive values indicate worse performance: leave-one-out minus full for Brier and reliability, and full minus leave-one-out for resolution. Gain retained places the leave-one-out classifier between Qwen3 (0%) and the full model (100%). Loss and retention intervals use paired bootstraps over identical episodes.

A. Absolute Brier performance

Training mix	Tahoe	DepMap	GWAS	Macro
DepMap + GWAS (Tahoe held out)	**0.2535** [0.2426, 0.2641]	0.1641 [0.1539, 0.1739]	0.1483 [0.1367, 0.1599]	0.1886 [0.1825, 0.1945]
Tahoe + GWAS (DepMap held out)	0.2072 [0.1994, 0.2151]	**0.2337** [0.2253, 0.2417]	0.1509 [0.1400, 0.1613]	0.1972 [0.1919, 0.2025]
Tahoe + DepMap (GWAS held out)	0.2047 [0.1955, 0.2139]	0.1639 [0.1534, 0.1743]	**0.2993** [0.2870, 0.3114]	0.2226 [0.2165, 0.2287]
All domains	0.2063 [0.1971, 0.2157]	0.1713 [0.1602, 0.1821]	0.1537 [0.1416, 0.1654]	0.1771 [0.1707, 0.1831]

B. Held-out-domain transfer decomposition

Held-out	Metric	Qwen3	Leave-one-out	Full model	Loss vs. full	Gain retained
Tahoe	Brier Score	0.3634 [0.3459, 0.3827]	0.2535 [0.2426, 0.2641]	0.2063 [0.1971, 0.2157]	0.0471 [0.0378, 0.0571]	70.0% [65.3, 74.7]%
	Reliability	0.1277 [0.1140, 0.1452]	0.0213 [0.0161, 0.0290]	0.0042 [0.0028, 0.0082]	0.0170 [0.0115, 0.0229]	86.2% [82.5, 89.7]%
	Resolution	0.0147 [0.0111, 0.0204]	0.0178 [0.0137, 0.0245]	0.0470 [0.0407, 0.0556]	0.0291 [0.0229, 0.0356]	9.8% [-4.6, 22.9]%
DepMap	Brier Score	0.2860 [0.2715, 0.2995]	0.2337 [0.2253, 0.2417]	0.1713 [0.1602, 0.1821]	0.0624 [0.0528, 0.0722]	45.6% [41.0, 50.3]%
	Reliability	0.0854 [0.0754, 0.0977]	0.0354 [0.0290, 0.0436]	0.0076 [0.0054, 0.0124]	0.0278 [0.0202, 0.0348]	64.3% [58.6, 71.3]%
	Resolution	0.0485 [0.0414, 0.0577]	0.0486 [0.0415, 0.0573]	0.0855 [0.0780, 0.0952]	0.0369 [0.0298, 0.0449]	0.4% [-16.0, 15.0]%
GWAS	Brier Score	0.3513 [0.3333, 0.3694]	0.2993 [0.2870, 0.3114]	0.1537 [0.1416, 0.1654]	0.1456 [0.1296, 0.1625]	26.3% [16.1, 35.4]%
	Reliability	0.1156 [0.1022, 0.1314]	0.0546 [0.0459, 0.0666]	0.0107 [0.0078, 0.0160]	0.0439 [0.0329, 0.0557]	58.1% [42.1, 70.5]%
	Resolution	0.0120 [0.0089, 0.0176]	0.0054 [0.0035, 0.0097]	0.1046 [0.0964, 0.1153]	0.0992 [0.0898, 0.1086]	-7.1% [-11.8, -2.8]%

G EXPERIMENTAL DESIGN DETAILS

This appendix collects material supporting §6.

G.1 NAIVE-PROMPTING SELECTION PROMPT

The naive-prompting policy π_{naive} (§6) asks a model to name the single most informative candidate directly, given the target hypothesis h^* and the evidence revealed so far. Because this is a selection task rather than the probability elicitation of the belief model, it uses a distinct system prompt that forbids a probability and fixes the answer format to `<choice>i</choice>`: reusing the belief prompt caused format-strict models to emit `<probability>` and silently fall back to random selection. The system prompt is

System prompt

```
You are an expert scientist selecting which experiment to run next. You are
shown a target hypothesis, the evidence gathered so far, and a numbered list
of candidate experiments whose results are hidden. Do NOT output a probability.
Your only task is to choose the single candidate whose hidden result would most
change your belief about the target hypothesis. Respond with just its index in
the exact form <choice>i</choice>.
```

The user message reuses the belief model's rendering of the target hypothesis and the revealed evidence—the `CONTEXT EVIDENCE` and `TARGET HYPOTHESIS` blocks documented in App. I—then appends the numbered candidate experiments, each rendered with its result hidden, followed by the selection instruction:

User prompt template

```
<CONTEXT EVIDENCE and TARGET HYPOTHESIS, rendered as in the prompts appendix>

CANDIDATE NEXT EXPERIMENTS (results hidden):
[0] <candidate experiment setup, result hidden>
[1] <candidate experiment setup, result hidden>
...

Which single candidate experiment, once its result is revealed, would most
change your belief about the target hypothesis? Answer with just its index as
<choice>i</choice>.
```

The same prompt is used for every naive policy (Qwen3, Ibis0, GPT-5.6-sol, Opus 5); only the answering model changes. The belief tracker that scores each policy's trajectory is held fixed across policies, so the comparison isolates selection quality.

H POST-HOC RECALIBRATION

We test how much of each model's Brier deficit lies in its probability scale rather than its ranking, and whether the trained models keep their advantage when every model is recalibrated.

Procedure. For each model and each domain we fit Platt scaling (Platt, 1999) to the model's own reported probabilities: a two-parameter affine map on the logit, $\hat{p} = \sigma(a \cdot \text{logit}(p) + b)$, estimated by logistic regression of the observed outcome on $\text{logit}(p)$ with negligible regularization ($C = 10^6$). The map is *cross-fitted* over two random folds: a and b are estimated on one half of a domain's episodes and applied to the other, so no episode is scored under a map to which it contributed. Refitting within each bootstrap draw propagates fit variability into the intervals in Table 8. We recalibrate all seven models, trained and untrained alike.

Calibration and ranking after scaling. sMECE decreases for every model, by a factor of 5.1 to 10.9 for the four untrained models and 1.3 to 3.1 for the three trained ones. Murphy reliability follows the same split, falling by a factor of 26 to 80 against 1.4 to 5.4. The trained models gain less because they start better calibrated.

Post-hoc scaling reduces calibration error without materially changing rank discrimination. The largest change in AUROC is 0.003, for Qwen3 (non-thinking); every other change is below 0.001, and the model ordering by AUROC is unchanged. A single affine map on the logit is monotone and preserves ranking. The small changes arise because the two cross-fitting folds use different maps, which can reorder observations across the fold boundary.

Model	n	Brier Score (↓)	sMECE (↓)	AUROC (↑)	Reliability (↓)	Resolution (↑)
As reported						
Qwen3 (non-thinking)	6,000	0.3336 [0.3237, 0.3436]	0.3101 [0.2990, 0.3229]	0.6608 [0.6471, 0.6752]	0.1096 [0.1024, 0.1189]	0.0250 [0.0226, 0.0295]
Qwen3 (thinking)	6,000	0.2973 [0.2896, 0.3047]	0.2696 [0.2607, 0.2818]	0.7040 [0.6914, 0.7170]	0.0829 [0.0772, 0.0907]	0.0339 [0.0307, 0.0387]
Ibis0 (non-thinking)	6,000	0.1933 [0.1888, 0.1977]	0.0316 [0.0258, 0.0450]	0.7729 [0.7616, 0.7846]	0.0020 [0.0017, 0.0038]	0.0579 [0.0540, 0.0631]
Ibis0 (thinking)	6,000	0.2069 [0.2025, 0.2111]	0.0327 [0.0296, 0.0431]	0.7329 [0.7209, 0.7450]	0.0023 [0.0021, 0.0044]	0.0448 [0.0413, 0.0500]
Ibis0 (classifier)	6,000	0.1771 [0.1707, 0.1831]	0.0658 [0.0588, 0.0762]	0.8195 [0.8099, 0.8295]	0.0075 [0.0065, 0.0104]	0.0790 [0.0753, 0.0849]
GPT-5.6-sol	6,000	0.2879 [0.2793, 0.2964]	0.2602 [0.2490, 0.2712]	0.7365 [0.7235, 0.7485]	0.0830 [0.0775, 0.0908]	0.0430 [0.0395, 0.0484]
Claude Opus 5	6,000	0.2772 [0.2689, 0.2851]	0.2390 [0.2291, 0.2505]	0.7519 [0.7392, 0.7646]	0.0786 [0.0730, 0.0857]	0.0502 [0.0461, 0.0556]
After cross-fit Platt scaling						
Qwen3 (non-thinking)	6,000	0.2315 [0.2274, 0.2353]	0.0603 [0.0417, 0.0657]	0.6580 [0.6363, 0.6698]	0.0043 [0.0027, 0.0075]	0.0214 [0.0180, 0.0266]
Qwen3 (thinking)	6,000	0.2148 [0.2109, 0.2192]	0.0281 [0.0239, 0.0380]	0.7033 [0.6880, 0.7152]	0.0018 [0.0013, 0.0034]	0.0363 [0.0324, 0.0407]
Ibis0 (non-thinking)	6,000	0.1920 [0.1874, 0.1969]	0.0172 [0.0162, 0.0283]	0.7723 [0.7591, 0.7828]	0.0006† [0.0007, 0.0020]	0.0574 [0.0533, 0.0628]
Ibis0 (thinking)	6,000	0.2062 [0.2017, 0.2105]	0.0257 [0.0247, 0.0363]	0.7327 [0.7185, 0.7437]	0.0016 [0.0012, 0.0031]	0.0442 [0.0406, 0.0496]
Ibis0 (classifier)	6,000	0.1706 [0.1657, 0.1754]	0.0212 [0.0184, 0.0292]	0.8193 [0.8087, 0.8285]	0.0014 [0.0011, 0.0029]	0.0799 [0.0754, 0.0852]
GPT-5.6-sol	6,000	0.2055 [0.2011, 0.2103]	0.0239 [0.0205, 0.0331]	0.7360 [0.7219, 0.7472]	0.0010 [0.0008, 0.0026]	0.0450 [0.0404, 0.0499]
Claude Opus 5	6,000	0.1997 [0.1950, 0.2046]	0.0314 [0.0263, 0.0406]	0.7512 [0.7377, 0.7633]	0.0017 [0.0014, 0.0034]	0.0511 [0.0468, 0.0569]

Macro-average over Tahoe, DepMap, GWAS; 95% CI from a within-domain bootstrap resampling episodes (2,000 draws; sMECE 200, it is 35x costlier per draw). † point estimate outside its bootstrap interval — expected for the positively-biased binned metrics (Reliability, sMECE) on rows near the zero floor. Compare rows by these intervals; do not read them as covering the point.

Table 8: Cross-fit Platt recalibration of every evaluated system: a two-parameter affine map on the logit, fitted out-of-fold within each domain. Because it requires labelled target-domain data, it bounds how much of each gap a post-hoc fix can close and is not a deployable method. The trained systems are also recalibrated so that the two blocks are comparable.

Forecast metrics after scaling. Recalibration removes most of the untrained models' Brier deficit: Brier falls by 0.077 to 0.102 for the four untrained models, against 0.001 to 0.007 for the three trained ones. Ibis0 (classifier) still has the lowest recalibrated Brier at 0.171, ahead of Claude Opus 5 at 0.200, with non-overlapping intervals.

Matching the treatment reverses the sMECE comparison. Recalibrated GPT-5.6-sol reaches 0.024, against 0.017 for recalibrated Ibis0 (non-thinking) and 0.021 for Ibis0 (classifier), with overlapping intervals. Comparing recalibrated frontier models with unscaled trained ones understates the trained models on this metric, because the classifier's as-reported sMECE of 0.066 is the one value in its row with appreciable headroom.

Recalibration by domain. Table 9 fits a and b separately within Tahoe, DepMap and GWAS. The gains are not uniform across domains. Tahoe remains the hardest domain for every model after scaling, with recalibrated Brier between 0.203 and 0.245, against 0.165 to 0.209 on DepMap. Ibis0 (classifier) has the widest spread, from 0.144 on GWAS to 0.203 on Tahoe, a range that the macro-average conceals.

Requirement for representative labels. Because a and b are estimated from other labeled episodes in the same domain, the procedure applies only when representative historical outcomes exist in the deployment domain. It is unavailable for a new assay with no calibration sample and may be invalid under distribution shift.

Resolution after scaling. Resolution is computed over ten equal-width forecast bins, so a monotone rescaling changes it only by shifting episodes across a bin edge. The seven resolutions change by at most 0.004 in either direction. After scaling, Qwen3 (non-thinking) is at 0.021 and Qwen3 (thinking) at 0.036, compared with 0.080 for the classifier. Coarse quantization of the verbalized

forecasts may contribute to this difference, but the current comparison does not isolate that mechanism.

Model	n	Brier Score (↓)	sMECE (↓)	AUROC (↑)	Reliability (↓)	Resolution (↑)
Qwen3 (non-thinking)	6,000	0.2315 [0.2274, 0.2353]	0.0603 [0.0417, 0.0657]	0.6580 [0.6363, 0.6698]	0.0043 [0.0027, 0.0075]	0.0214 [0.0180, 0.0266]
Tahoe	2,000	0.2451 [0.2419, 0.2487]	0.0656 [0.0383, 0.0743]	0.6034 [0.5508, 0.6223]	0.0029 [0.0006, 0.0078]	0.0048 [0.0007, 0.0138]
DepMap	2,000	0.2092 [0.1987, 0.2186]	0.0703 [0.0469, 0.0968]	0.7431 [0.7218, 0.7654]	0.0082 [0.0041, 0.0154]	0.0486 [0.0410, 0.0588]
GWAS	2,000	0.2401 [0.2348, 0.2443]	0.0450 [0.0165, 0.0634]	0.6274 [0.5936, 0.6502]	0.0018 [0.0003, 0.0051]	0.0109 [0.0051, 0.0177]
Qwen3 (thinking)	6,000	0.2148 [0.2109, 0.2192]	0.0281 [0.0239, 0.0380]	0.7033 [0.6880, 0.7152]	0.0018 [0.0013, 0.0034]	0.0363 [0.0324, 0.0407]
Tahoe	2,000	0.2245 [0.2182, 0.2312]	0.0302 [0.0204, 0.0455]	0.6814 [0.6540, 0.7027]	0.0023 [0.0010, 0.0050]	0.0278 [0.0204, 0.0338]
DepMap	2,000	0.1863 [0.1773, 0.1951]	0.0275 [0.0212, 0.0428]	0.7829 [0.7614, 0.8020]	0.0011 [0.0007, 0.0046]	0.0645 [0.0568, 0.0745]
GWAS	2,000	0.2337 [0.2277, 0.2391]	0.0265 [0.0172, 0.0444]	0.6456 [0.6190, 0.6702]	0.0019 [0.0005, 0.0037]	0.0166 [0.0113, 0.0233]
Ibis0 (non-thinking)	6,000	0.1920 [0.1874, 0.1969]	0.0172 [0.0162, 0.0283]	0.7723 [0.7591, 0.7828]	0.0006† [0.0007, 0.0020]	0.0574 [0.0533, 0.0628]
Tahoe	2,000	0.2149 [0.2082, 0.2220]	0.0197 [0.0146, 0.0373]	0.7131 [0.6875, 0.7333]	0.0004 [0.0003, 0.0026]	0.0349 [0.0282, 0.0424]
DepMap	2,000	0.1754 [0.1665, 0.1844]	0.0119 [0.0114, 0.0240]	0.8116 [0.7909, 0.8297]	0.0005 [0.0003, 0.0025]	0.0737 [0.0653, 0.0841]
GWAS	2,000	0.1858 [0.1770, 0.1952]	0.0200 [0.0142, 0.0348]	0.7921 [0.7703, 0.8113]	0.0008 [0.0004, 0.0030]	0.0635 [0.0553, 0.0732]
Ibis0 (thinking)	6,000	0.2062 [0.2017, 0.2105]	0.0257 [0.0247, 0.0363]	0.7327 [0.7185, 0.7437]	0.0016 [0.0012, 0.0031]	0.0442 [0.0406, 0.0496]
Tahoe	2,000	0.2164 [0.2095, 0.2234]	0.0176 [0.0135, 0.0325]	0.7065 [0.6816, 0.7278]	0.0007 [0.0003, 0.0027]	0.0325 [0.0264, 0.0404]
DepMap	2,000	0.1819 [0.1726, 0.1908]	0.0274 [0.0226, 0.0424]	0.7951 [0.7748, 0.8152]	0.0019 [0.0007, 0.0043]	0.0692 [0.0601, 0.0793]
GWAS	2,000	0.2202 [0.2127, 0.2272]	0.0321 [0.0227, 0.0499]	0.6966 [0.6701, 0.7179]	0.0023 [0.0010, 0.0053]	0.0310 [0.0248, 0.0397]
Ibis0 (classifier)	6,000	0.1706 [0.1657, 0.1754]	0.0212 [0.0184, 0.0292]	0.8193 [0.8087, 0.8285]	0.0014 [0.0011, 0.0029]	0.0799 [0.0754, 0.0852]
Tahoe	2,000	0.2032 [0.1960, 0.2108]	0.0270 [0.0178, 0.0436]	0.7486 [0.7262, 0.7675]	0.0019 [0.0007, 0.0041]	0.0477 [0.0401, 0.0553]
DepMap	2,000	0.1646 [0.1559, 0.1730]	0.0181 [0.0133, 0.0323]	0.8331 [0.8152, 0.8489]	0.0007 [0.0005, 0.0032]	0.0851 [0.0773, 0.0947]
GWAS	2,000	0.1440 [0.1345, 0.1531]	0.0187 [0.0134, 0.0291]	0.8762 [0.8606, 0.8909]	0.0016 [0.0008, 0.0042]	0.1069 [0.0984, 0.1164]
GPT-5.6-sol	6,000	0.2055 [0.2011, 0.2103]	0.0239 [0.0205, 0.0331]	0.7360 [0.7219, 0.7472]	0.0010 [0.0008, 0.0026]	0.0450 [0.0404, 0.0499]
Tahoe	2,000	0.2192 [0.2125, 0.2260]	0.0203 [0.0143, 0.0351]	0.6983 [0.6737, 0.7198]	0.0009 [0.0003, 0.0033]	0.0310 [0.0241, 0.0387]
DepMap	2,000	0.1801 [0.1711, 0.1895]	0.0259 [0.0175, 0.0393]	0.8042 [0.7830, 0.8222]	0.0014 [0.0007, 0.0036]	0.0703 [0.0613, 0.0803]
GWAS	2,000	0.2173 [0.2100, 0.2249]	0.0254 [0.0166, 0.0426]	0.7054 [0.6801, 0.7273]	0.0008 [0.0003, 0.0031]	0.0337 [0.0258, 0.0415]
Claude Opus 5	6,000	0.1997 [0.1950, 0.2046]	0.0314 [0.0263, 0.0406]	0.7512 [0.7377, 0.7633]	0.0017 [0.0014, 0.0034]	0.0511 [0.0468, 0.0569]
Tahoe	2,000	0.2232 [0.2169, 0.2296]	0.0157 [0.0120, 0.0303]	0.6854 [0.6603, 0.7069]	0.0002† [0.0002, 0.0019]	0.0259 [0.0203, 0.0337]
DepMap	2,000	0.1788 [0.1689, 0.1878]	0.0253 [0.0156, 0.0362]	0.8035 [0.7836, 0.8230]	0.0009 [0.0004, 0.0026]	0.0714 [0.0627, 0.0813]
GWAS	2,000	0.1971 [0.1881, 0.2060]	0.0532 [0.0391, 0.0691]	0.7647 [0.7420, 0.7854]	0.0041 [0.0025, 0.0077]	0.0561 [0.0476, 0.0664]

Macro-average over Tahoe, DepMap, GWAS; 95% CI from a within-domain bootstrap resampling episodes (2,000 draws; sMECE 200, it is 35x costlier per draw). The unindented row of each block is that macro-average; the indented rows below it are the single domains, bootstrapped over that domain's episodes alone. † point estimate outside its bootstrap interval — expected for the positively-biased binned metrics (Reliability, sMECE) on rows near the zero floor. Compare rows by these intervals; do not read them as covering the point.

Table 9: Cross-fit Platt recalibration by domain, with the map fitted separately within each of Tahoe, DepMap and GWAS. The as-reported breakdown is in the complete master table.

I PROMPTS: ONE WORKED EXAMPLE PER DOMAIN

A shared function renders every episode as a prompt. The prompt contains a `CONTEXT EVIDENCE` header, one numbered block per evidence stub beginning with its relevance tag, a `TARGET HYPOTHESIS` header, the target block, and a fixed answer-format instruction. Domain-specific functions render individual stubs and targets. §2 defines the episode structure; this appendix provides one rendered example per domain.

Each system prompt describes the assay and evidence-generation procedure, defines the reported quantities, states the support-label rule, and supplies an answer-format instruction. The *direct* instruction requests only a probability; the *thinking* instruction requests a reasoning trace followed by the same probability tag. Each evaluated model receives the same instruction in every domain (Apps. B and D); the examples use the direct instruction.

Direct instruction

```
Respond with ONLY a single line in this exact format — no reasoning, no explanation, no other text before or
    ↪ after:
<probability>0.XX</probability>
```

Thinking instruction

```
Think through the evidence carefully, then commit to a single probability in the closed interval [0, 1].
  After your reasoning, output your probability in this exact format:
<probability>0.XX</probability>
```

The examples are drawn from the `primary` test-set configuration at revision `48389e59`. Long lines wrap at the margin, with each continuation marked by ↪, and the sole `!=` in the Tahoe system prompt replaces a ≠ glyph the document font cannot set. Evidence pools are truncated after the first two stubs, and the number of omitted stubs is marked in place. The displayed text is otherwise unchanged.

I.1 TAHOE: DRUG PERTURBATION OF CANCER CELL LINES

Episode `eb33e117-0ab` asks whether encorafenib upregulates `LINC00342` in `SNU-1`. The episode contains six stubs, and the held-out DESeq2 result supports the hypothesis.

The two displayed stubs both measure the target drug and gene and are statistically significant; one reports a decrease in a lung line and the other an increase of similar magnitude in a skin line. Of the four omitted stubs, one is the target cell line perturbed by a different drug (a null result), and three are unrelated drug × cell-line baselines. No stub reports encorafenib in `SNU-1`, so the pool constrains the answer without containing it.

System prompt

```
You are a scientific reasoning assistant analyzing drug-perturbation evidence from Tahoe-100M, a large-scale
    ↪ pseudobulk RNA-seq dataset measuring drug × cell-line × gene differential expression.

How the evidence was generated:
  - Cells from a given cell line are treated with a drug (or DMSO vehicle control).
  - Each cell's transcripts are captured by single-cell RNA-seq; counts are UMI-deduplicated per gene.
  - Single cells from the same (drug, cell_line, plate, replicate) bucket are *pseudobulked* by summing
      ↪ per-gene UMI counts.
  - For each gene, DESeq2 fits a negative-binomial GLM comparing treated vs control pseudobulk samples.

Column definitions (DESeq2 conventions):
  - log2FoldChange: shrunken log2(treated / control) of normalized counts. Positive = upregulated; negative =
      ↪ downregulated.
  - lfcSE:          standard error of the log2FoldChange estimate (uncertainty).
  - pvalue:         Wald test p-value for log2FoldChange != 0.
  - padj:           Benjamini-Hochberg adjusted p-value (FDR-controlled across all genes in that drug ×
      ↪ cell_line comparison).
  - baseMean:       mean of normalized counts across both treated AND control samples combined (an
      ↪ expression-level indicator, NOT a control baseline).
  - n_cells_trt / n_cells_ctrl: number of single cells pooled into the treated / control pseudobulk.

You will be shown:
  1. CONTEXT — prior experimental results: drug × cell-line × gene differential-expression measurements, plus
      ↪ aggregate perturbation summaries.
  2. A TARGET HYPOTHESIS — a claim of the form "drug D upregulates/downregulates gene G in cell line C".
```

```
Weigh the evidence and report your degree of belief that a held-out experiment would *support* the target
    ↪ hypothesis. Supported means padj < 0.05 AND |log2FoldChange| > 0.5 in the claimed direction.

Respond with ONLY a single line in this exact format — no reasoning, no explanation, no other text before or
    ↪ after:
<probability>0.XX</probability>
```

User prompt (episode `eb33e117-0ab`, 6 stubs, label = supported)

```
CONTEXT EVIDENCE
================

[1] same drug as target, different cell line, target gene
    Drug: Encorafenib
      MOA (broad)          inhibitor/antagonist
      MOA (fine)           RAF inhibitor
      Targets              BRAF
      Human-approved       yes
      In clinical trials   yes
      Notes                Approved for use in BRAF-mutant melanoma.
      Canonical SMILES     CC(C)N1C=C(C(=N1)C2=C(C(=CC(=C2)Cl)NS(=O)(=O)C)F)C3=NC(=NC=C3)NCC(C)NC(=O)OC
      PubChem CID          50922675
    Cell line: NCI-H2347
      DepMap ID                   ACH-000875
      Cellosaurus ID              CVCL_1550
      Organ                       Lung
      Driver gene                 B2M
      Driver variant type         Stopgain
      Driver zygosity             Hom
      Driver protein/cDNA effect  p.Q28*
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: LINC00342
      Ensembl ID  ENSG00000232931
    Experiment:
      Plate                                          3
      n_cells treated                                2747
      n_cells control                                3106
      baseMean (avg normalized counts across both arms)  90.74
    Result:
      log2FoldChange      -0.473 ± 0.153  (downregulates)
      Wald statistic      -3.101
      p-value (raw)       0.00193
      p-value (adjusted)  0.01023

[2] same drug as target, different cell line, target gene
    Drug: Encorafenib
      MOA (broad)          inhibitor/antagonist
      MOA (fine)           RAF inhibitor
      Targets              BRAF
      Human-approved       yes
      In clinical trials   yes
      Notes                Approved for use in BRAF-mutant melanoma.
      Canonical SMILES     CC(C)N1C=C(C(=N1)C2=C(C(=CC(=C2)Cl)NS(=O)(=O)C)F)C3=NC(=NC=C3)NCC(C)NC(=O)OC
      PubChem CID          50922675
    Cell line: LOX-IMVI
      DepMap ID                   ACH-000750
      Cellosaurus ID              CVCL_1381
      Organ                       Skin
      Driver gene                 ARID1A
      Driver variant type         Stopgain
      Driver zygosity             Het
      Driver protein/cDNA effect  p.Q1212*
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: LINC00342
      Ensembl ID  ENSG00000232931
    Experiment:
      Plate                                          3
      n_cells treated                                2163
      n_cells control                                2021
      baseMean (avg normalized counts across both arms)  154.3
    Result:
      log2FoldChange      +0.535 ± 0.136  (upregulates)
      Wald statistic      3.925
      p-value (raw)       8.66e-05
      p-value (adjusted)  0.00113

[... 4 further evidence stubs omitted ...]
```

```
TARGET HYPOTHESIS
=================
Drug Encorafenib upregulates gene LINC00342 in cell line SNU-1.

    Drug: Encorafenib
      MOA (broad)         inhibitor/antagonist
      MOA (fine)          RAF inhibitor
      Targets             BRAF
      Human-approved      yes
      In clinical trials  yes
      Notes               Approved for use in BRAF-mutant melanoma.
      Canonical SMILES    CC(C)N1C=C(C(=N1)C2=C(C(=CC(=C2)Cl)NS(=O)(=O)C)F)C3=NC(=NC=C3)NCC(C)NC(=O)OC
      PubChem CID         50922675
    Cell line: SNU-1
      DepMap ID                   ACH-000932
      Cellosaurus ID              CVCL_0099
      Organ                       Esophagus/Stomach
      Driver gene                 ARID1A
      Driver variant type         Frameshift
      Driver zygosity             Het
      Driver protein/cDNA effect  p.D1633fs
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: LINC00342
      Ensembl ID  ENSG00000232931

Respond using the format specified in the system message:
  <probability>0.XX</probability>
```

I.2 DepMap: cancer gene dependency

Episode `65b4265e-9f1` asks whether `DIS3L` is a dependency in the mesothelioma line `MERO25`. The held-out screen does not support the hypothesis.

The first displayed stub reports the target gene's mutation status in the target line rather than a knockout result, an example of the multi-modal evidence pool. The four omitted stubs are two unrelated knockouts in `MERO25` and two knockouts of `DIS3L` in other pleural mesothelioma lines, the latter scoring -0.25 and -0.45 against a threshold of -0.5.

System prompt

```
You are a scientific reasoning assistant analyzing cancer gene-dependency evidence from DepMap, a large-scale
    ↪ project of genome-wide CRISPR knockout screens of cancer cell lines with matched molecular and
    ↪ pharmacological profiles.
How the evidence was generated:
  - Each cell line is screened by pooled genome-wide CRISPR: every gene is knocked out (one gene per cell),
      ↪ the culture grows ~3 weeks, and cells whose knocked-out gene was required for proliferation drop out
      ↪ of the pool.
  - The depletion of each gene is summarized by the Chronos gene-effect score; a calibrated P(dependency) is
      ↪ derived from it.
  - Matched profiles, measured independently of the screen, give each line's baseline mRNA expression, copy
      ↪ number, somatic mutations, and (for some) drug sensitivity.
Quantity definitions:
  - Chronos gene-effect: a CRISPR-knockout fitness score, scaled so 0 = no effect on growth and -1 = as lethal
      ↪ as a typical core-essential gene (more negative = more essential; positive = knockout speeds
      ↪ growth). Analogous to a log2 fold-change.
  - P(dependency): the calibrated probability in [0,1] that the gene is a true dependency in that cell line,
      ↪ inferred from the Chronos score. Analogous to statistical significance.
You will be shown:
  1. CONTEXT — prior measurements, each labeled by its relation to the target.
  2. A TARGET HYPOTHESIS — a claim of the form "gene G is a dependency in cell line C".
Weigh the evidence and report your degree of belief that a held-out genome-wide CRISPR screen would support
    ↪ the target hypothesis. Supported means the screen scores G at Chronos gene-effect <= -0.5 AND
    ↪ P(dependency) >= 0.5.
Respond with ONLY a single line in this exact format — no reasoning, no explanation, no other text before or
    ↪ after:
<probability>0.XX</probability>
```

User prompt (episode `65b4265e-9f1`, 6 stubs, label = not supported)

```
CONTEXT EVIDENCE
================

[1] somatic mutation status of the target gene in the target cell line
Somatic mutation status of DIS3L in MERO25
      Gene: DIS3L — DIS3 like exosome 3'-5' exoribonuclease (family: Exosome complex)
```

```
      Cell line: MERO25 (DepMap ACH-001556) — Pleura / Pleural Mesothelioma / Pleural Mesothelioma, Biphasic
          ↪ Type; Male; Primary
      Result: wild-type

[2] an unrelated gene knocked out in the target cell line
CRISPR knockout of KRT25 in MERO25
      Gene: KRT25 — keratin 25 (family: Keratins, type I)
      Cell line: MERO25 (DepMap ACH-001556) — Pleura / Pleural Mesothelioma / Pleural Mesothelioma, Biphasic
          ↪ Type; Male; Primary
      Result:
        Chronos gene-effect  -0.02  (no clear fitness effect)
        P(dependency)        0.03   (almost certainly not a dependency)

[... 4 further evidence stubs omitted ...]

TARGET HYPOTHESIS
=================
Gene DIS3L is a dependency (essential for fitness) in cell line MERO25: a held-out genome-wide CRISPR knockout
    ↪ screen of MERO25 would score gene DIS3L at Chronos gene-effect <= -0.5 AND P(dependency) >= 0.5.
  Gene: DIS3L — DIS3 like exosome 3'-5' exoribonuclease (family: Exosome complex)
  Cell line: MERO25 (DepMap ACH-001556) — Pleura / Pleural Mesothelioma / Pleural Mesothelioma, Biphasic
      ↪ Type; Male; Primary

Respond using the format specified in the system message:
  <probability>0.XX</probability>
```

I.3 GWAS: GENETIC ASSOCIATIONS

Episode `6b7db1d8-f2b` asks whether `rs2238689` (T allele) decreases a phospholipid ratio measurement. The held-out study supports the hypothesis at genome-wide significance.

GWAS relevance depends on locus and trait similarity: a stub may share the target locus, trait category, both, or neither. The omitted stubs include a missense variant 2,731 bp away in the same gene, associated at $p = 4 \times 10^{-15}$ with a trait in the target's own EFO parent, and one association reported at $p = 8 \times 10^{-8}$, below the genome-wide significance threshold.

System prompt

```
You are an expert geneticist and epidemiologist analyzing genetic-association evidence from the NHGRI-EBI
    ↪ GWAS Catalog, a curated collection of published genome-wide association studies (GWAS) of human traits
    ↪ and diseases.
How the evidence was generated:
  - A GWAS genotypes many individuals and tests each common variant (SNP) for association with a trait,
      ↪ reporting an effect size and a p-value per SNP.
  - An association is called genome-wide significant at p < 5e-8, the threshold that corrects for ~1 million
      ↪ independent tests across the genome.
  - Each study reports, for the lead SNP, the risk allele whose effect is measured, its frequency, and the
      ↪ direction/size of effect (odds ratio for disease traits, beta for quantitative traits).
Quantity definitions:
  - p-value: statistical significance of the SNP-trait association; < 5e-8 is genome-wide significant.
  - Risk allele: the allele whose effect is reported; different studies may report opposite alleles for the
      ↪ same SNP, so direction is always stated relative to the named allele.
  - OR (odds ratio, binary/disease traits): OR > 1 means the risk allele increases disease odds, OR < 1 means
      ↪ it decreases them (protective).
  - Beta (quantitative traits): the named allele raises the trait when the direction is 'increase', lowers it
      ↪ when 'decrease'.
  - RAF: risk-allele frequency in the study population.
  - Mapped gene: gene assigned to the SNP by genomic position (not a causal claim); EFO parent: the
      ↪ Experimental Factor Ontology category grouping related traits.
You will be shown:
  1. CONTEXT — prior GWAS associations, each labeled by its relation to the target.
  2. A TARGET HYPOTHESIS — a claim that a SNP influences a trait in a stated direction.
Weigh the evidence and report your degree of belief that a held-out GWAS would support the target hypothesis.
    ↪ Supported means the held-out study reports p < 5e-8 for this SNP-trait pair with a consistent
    ↪ direction of effect.
Respond with ONLY a single line in this exact format — no reasoning, no explanation, no other text before or
    ↪ after:
<probability>0.XX</probability>
```

User prompt (episode `6b7db1d8-f2b`, 6 stubs, label = supported)

```
CONTEXT EVIDENCE
================

[1] the target SNP, associated with a trait in a different EFO parent
rs2238689 for type 2 diabetes mellitus
```

```
      SNP: rs2238689 — chr19:45675403
        Risk allele: C
        RAF: 0.4177
        Mapped gene: GIPR
        Functional class: intron_variant
      Gene context:
        Reactome pathways: G alpha (s) signalling events, Glucagon-type ligand receptors
        GTEx expression: Cells - EBV-transformed lymphocytes 23.8 TPM, Adipose - Visceral (Omentum) 10.8 TPM,
            ↪ Heart - Atrial Appendage 9.3 TPM
      Study: GCST009379
        Ancestry: European
        Sample size: 898,130
        Platform: Affymetrix, Illumina [~ 27000000] (imputed)
      Result: p = 5.0e-09 (genome-wide significant); OR/beta = 1.04
        95% CI: [1.03-1.05]
        Direction: C allele increases type 2 diabetes mellitus

[2] an unrelated SNP and trait (different gene, different EFO parent)
rs12324089 for testosterone measurement
      SNP: rs12324089 — chr15:64025657
        Risk allele: G
        RAF: 0.409
        Mapped gene: DAPK2
        Functional class: intron_variant
      Gene context:
        Reactome pathways: Caspase activation via Dependence Receptors in the absence of ligand
        GTEx expression: Thyroid 42.0 TPM, Muscle - Skeletal 19.2 TPM, Lung 16.1 TPM
      Study: GCST90483487
        Ancestry: European
        Sample size: 235,096
        Platform: NR [23899339] (imputed)
      Result: p = 4.0e-08 (genome-wide significant); OR/beta = 0.0152
        95% CI: [0.0097-0.0207] unit decrease
        Direction: G allele decreases testosterone measurement

[... 4 further evidence stubs omitted ...]

TARGET HYPOTHESIS
=================
SNP rs2238689 (T allele) decreases phospholipids:total lipids ratio, high density lipoprotein cholesterol
    ↪ measurement.
A held-out GWAS would report p < 5e-8 for this SNP-trait pair with a consistent direction of effect.
  SNP: rs2238689 — chr19:45675403
    Risk allele: T
    RAF: 0.600605
    Mapped gene: GIPR
    Functional class: intron_variant
  Gene context:
    Reactome pathways: G alpha (s) signalling events, Glucagon-type ligand receptors
    GTEx expression: Cells - EBV-transformed lymphocytes 23.8 TPM, Adipose - Visceral (Omentum) 10.8 TPM,
        ↪ Heart - Atrial Appendage 9.3 TPM
  Trait: phospholipids:total lipids ratio, high density lipoprotein cholesterol measurement
    EFO parent: Other measurement

Respond using the format specified in the system message:
  <probability>0.XX</probability>
```

J A THINKING-MODE REASONING TRACE

Ibis0 (thinking) emits a reasoning trace before committing to a probability. This appendix reproduces one complete trace and the episode that produced it. The rollout below is a single sample from the trained policy, on an episode from the Tahoe validation split rather than the test set.

Episode `6a6866fd-49a` asks whether belumosudil, a ROCK1/ROCK2 inhibitor, downregulates `SGMS2` in the colorectal line `HCT15`. The prompt renders seven evidence stubs. Two measure the same drug and gene in other cell lines, and both report *up*regulation: $+0.833$ at padj $= 1.7\times10^{-17}$ in `HOP62`, and $+0.146$ at padj $= 0.045$ in `HT-29`. One measures the target cell line under a different drug and is null, as are the four remaining stubs. The held-out measurement refutes the hypothesis rather than leaving it unresolved: belumosudil significantly *up*regulates `SGMS2` in `HCT15` ($+0.554$, padj $= 0.0056$). The trace reads the same-drug stubs as evidence against the claimed direction and settles at 0.20.

The trace also questions one of the prompt's relevance tags. Stub [4] is labeled as sharing the target drug's mechanism, a match that holds on the coarse MOA field, where both drugs read `inhibitor/antagonist`, but not below it: artesunate's fine MOA is a JAK/STAT inhibitor,

and belumosudil's listed targets are ROCK1 and ROCK2. The trace registers the doubt ("same mechanism as Belumosudil? Or not sure") and sets the stub aside as null.

The trace is highly redundant. The conclusion that the same-drug stubs argue against a downregulation claim is restated at least four times, and the answer is approached, qualified and revisited before it is committed.

Rendering follows the conventions of App. I, except that the completion is set as prose with its paragraph breaks preserved. One drug name is transliterated (`18beta-Glycyrrhetinic acid`) because the document font cannot set its Greek letter. The `<think>` tags enclosing the trace are dropped. The text is otherwise unchanged, and the evidence pool is shown in full.

J.1 Prompt

The system prompt is the Tahoe system prompt of App. I carrying the thinking instruction in place of the direct one, with a further paragraph inserted before it that names considerations to weigh:

System prompt (addition over App. I)

```
Reason about what the context implies — for example: the drug's transcriptional footprint in this and nearby
    ↪ cell lines, whether mechanistically similar drugs show consistent effects on the target gene, whether
    ↪ the gene appears broadly responsive or quiet, the drug's aggregate perturbation strength in the target
    ↪ cell line, and any contradictions between evidence pieces. Then commit to a single probability in the
    ↪ closed interval [0, 1].
```

User prompt (episode `6a6866fd-49a`, 7 stubs, label = not supported)

```
CONTEXT EVIDENCE
================

[1] unrelated drug × cell line, included as a baseline
    Drug: Anethole trithione
      MOA (broad)          unclear
      MOA (fine)           unclear
      Human-approved       no
      In clinical trials   yes
      Notes                Investigated for dry mouth in clinical trials.
      Canonical SMILES     COC1=CC=C(C=C1)C2=CC(=S)SS2
      PubChem CID          2194
    Cell line: hTERT-HPNE
      Cellosaurus ID       CVCL_C466
      Organ                Pancreas
      Driver gene          TERT
      Driver variant type  Transgene
      Driver gene type     Oncogene
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                              4
      n_cells treated                                    2002
      n_cells control                                    2917
      baseMean (avg normalized counts across both arms)  26.75
    Result:
      log2FoldChange      +0.115 ± 0.334  (upregulates)
      Wald statistic      0.3436
      p-value (raw)       0.7312
      p-value (adjusted)  0.9005

[2] unrelated drug × cell line, included as a baseline
    Drug: DT-061
      MOA (broad)          activator/agonist
      MOA (fine)           unclear
      Human-approved       no
      In clinical trials   yes
      Notes                Investigated for oncology, no approval yet for human use.
      Canonical SMILES     C1CC(C(C(C1)N2C3=CC=CC=C3OC4=CC=CC=C42)O)NS(=O)(=O)C5=CC=C(C=C5)OC(F)(F)F
      PubChem CID          91885558
    Cell line: CHP-212
      DepMap ID                  ACH-000120
      Cellosaurus ID             CVCL_1125
      Organ                      Peripheral Nervous System
      Driver gene                NRAS
      Driver variant type        Missense
      Driver zygosity            Het
      Driver protein/cDNA effect  p.Q61K
      Driver mechanism (LoF/GoF)  GoF
```

```
      Driver gene type            Oncogene
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                          2
      n_cells treated                                524
      n_cells control                                917
      baseMean (avg normalized counts across both arms)  6.928
    Result:
      log2FoldChange      +0.285 ± 0.552  (upregulates)
      Wald statistic      0.5167
      p-value (raw)       0.6053
      p-value (adjusted)  0.9021

[3] same drug as target, different cell line, target gene
    Drug: Belumosudil (mesylate)
      MOA (broad)         inhibitor/antagonist
      MOA (fine)          unclear
      Targets             ROCK2, ROCK1
      Human-approved      yes
      In clinical trials  yes
      Notes               Approved for chronic graft-versus-host disease.
      Canonical SMILES    CC(C)NC(=O)COC1=CC=CC(=C1)C2=NC3=CC=CC=C3C(=N2)NC4=CC5=C(C=C4)NN=C5.CS(=O)(=O)O
      PubChem CID         146681181
    Cell line: HOP62
      DepMap ID                   ACH-000861
      Cellosaurus ID              CVCL_1285
      Organ                       Lung
      Driver gene                 B2M
      Driver variant type         Missense
      Driver zygosity             Het
      Driver protein/cDNA effect  p.E70Q
      Driver gene type            Suppressor
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                          2
      n_cells treated                                2743
      n_cells control                                8616
      baseMean (avg normalized counts across both arms)  292.5
    Result:
      log2FoldChange      +0.833 ± 0.093  (upregulates)
      Wald statistic      8.991
      p-value (raw)       2.45e-19
      p-value (adjusted)  1.67e-17

[4] drug with same mechanism of action as the target drug
    Drug: Artesunate
      MOA (broad)         inhibitor/antagonist
      MOA (fine)          JAK/STAT inhibitor
      Human-approved      yes
      In clinical trials  yes
      Notes               Used for severe malaria, often partnered with other antimalarial drugs.
      Canonical SMILES    CC1CCC2C(C(OC3C24C1CCC(O3)(OO4)C)OC(=O)CCC(=O)O)C
      PubChem CID         6917864
    Cell line: RPMI-7951
      DepMap ID                   ACH-000348
      Cellosaurus ID              CVCL_1666
      Organ                       Skin
      Driver gene                 BRAF
      Driver variant type         Missense
      Driver zygosity             Het
      Driver protein/cDNA effect  p.V600E
      Driver mechanism (LoF/GoF)  GoF
      Driver gene type            Oncogene
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                          4
      n_cells treated                                1577
      n_cells control                                2477
      baseMean (avg normalized counts across both arms)  22.73
    Result:
      log2FoldChange      +0.200 ± 0.361  (upregulates)
      Wald statistic      0.555
      p-value (raw)       0.5789
      p-value (adjusted)  0.8409

[5] unrelated drug × cell line, included as a baseline
    Drug: 18beta-Glycyrrhetinic acid
      MOA (broad)         unclear
      MOA (fine)          unclear
```

```
      Human-approved      no
      In clinical trials  yes
      Notes               Used in cosmetics, derived from licorice root, no approval for human drugs.
      Canonical SMILES C[C@]12CC[C@](C[C@H]1C3=CC(=O)[C@@H]4[C@]5(CC[C@@H](C([C@@H]5CC[C@]4([C@@]3(CC2)C)C)(
          ↪ C)C)O)C)(C)C(=O)O
      PubChem CID         10114
    Cell line: HS-578T
      DepMap ID                   ACH-000148
      Cellosaurus ID              CVCL_0332
      Organ                       Breast
      Driver gene                 CDKN2A
      Driver variant type         Deletion
      Driver zygosity             Hom
      Driver protein/cDNA effect  DEL
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                           4
      n_cells treated                                 857
      n_cells control                                 2654
      baseMean (avg normalized counts across both arms)  43.66
    Result:
      log2FoldChange      +0.255 ± 0.218  (upregulates)
      Wald statistic      1.17
      p-value (raw)       0.242
      p-value (adjusted)  0.6025

[6] same drug as target, different cell line, target gene
    Drug: Belumosudil (mesylate)
      MOA (broad)         inhibitor/antagonist
      MOA (fine)          unclear
      Targets             ROCK2, ROCK1
      Human-approved      yes
      In clinical trials  yes
      Notes               Approved for chronic graft-versus-host disease.
      Canonical SMILES    CC(C)NC(=O)COC1=CC=CC(=C1)C2=NC3=CC=CC=C3C(=N2)NC4=CC5=C(C=C4)NN=C5.CS(=O)(=O)O
      PubChem CID         146681181
    Cell line: HT-29
      DepMap ID                   ACH-000552
      Cellosaurus ID              CVCL_0320
      Organ                       Bowel
      Driver gene                 APC
      Driver variant type         Stopgain
      Driver zygosity             Het
      Driver protein/cDNA effect  p.E853*
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                           2
      n_cells treated                                 2179
      n_cells control                                 5453
      baseMean (avg normalized counts across both arms)  1.18e+03
    Result:
      log2FoldChange      +0.146 ± 0.056  (upregulates)
      Wald statistic      2.584
      p-value (raw)       0.00976
      p-value (adjusted)  0.04521

[7] different drug, same cell line as target, target gene
    Drug: Adenosine
      MOA (broad)         unclear
      MOA (fine)          unclear
      Targets             ADORA1, ADORA2A, ADORA2B, ADORA3
      Human-approved      yes
      In clinical trials  yes
      Notes               Used for certain arrhythmias; approved as a drug.
      Canonical SMILES    C1=NC(=C2C(=N1)N(C=N2)C3C(C(C(O3)CO)O)O)N
      PubChem CID         60961
    Cell line: HCT15
      DepMap ID                   ACH-000997
      Cellosaurus ID              CVCL_0292
      Organ                       Bowel
      Driver gene                 ACVR2A
      Driver variant type         Frameshift
      Driver zygosity             Hom
      Driver protein/cDNA effect  p.K437fs
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
```

```
    Gene: SGMS2
      Ensembl ID  ENSG00000164023
    Experiment:
      Plate                                          4
      n_cells treated                                925
      n_cells control                                2226
      baseMean (avg normalized counts across both arms)  83.5
    Result:
      log2FoldChange      +0.142 ± 0.205  (upregulates)
      Wald statistic      0.6909
      p-value (raw)       0.4897
      p-value (adjusted)  0.7822

TARGET HYPOTHESIS
=================
Drug Belumosudil (mesylate) downregulates gene SGMS2 in cell line HCT15.

    Drug: Belumosudil (mesylate)
      MOA (broad)         inhibitor/antagonist
      MOA (fine)          unclear
      Targets             ROCK2, ROCK1
      Human-approved      yes
      In clinical trials  yes
      Notes               Approved for chronic graft-versus-host disease.
      Canonical SMILES    CC(C)NC(=O)COC1=CC=CC(=C1)C2=NC3=CC=CC=C3C(=N2)NC4=CC5=C(C=C4)NN=C5.CS(=O)(=O)O
      PubChem CID         146681181
    Cell line: HCT15
      DepMap ID                   ACH-000997
      Cellosaurus ID              CVCL_0292
      Organ                       Bowel
      Driver gene                 ACVR2A
      Driver variant type         Frameshift
      Driver zygosity             Hom
      Driver protein/cDNA effect  p.K437fs
      Driver mechanism (LoF/GoF)  LoF
      Driver gene type            Suppressor
    Gene: SGMS2
      Ensembl ID  ENSG00000164023

Respond using the format specified in the system message:
  <probability>0.XX</probability>
```

J.2 TRACE

Completion (reasoning trace, verbatim)

Okay, let's tackle this step by step. So the task is to determine the probability that the target hypothesis (Belumosudil downregulates SGMS2 in HCT15) is supported based on the context provided. The context has seven previous experiments, and I need to analyze each to see how they relate to the target hypothesis.

First, let's understand the target hypothesis. The drug is Belumosudil, which is an inhibitor/antagonist targeting ROCK2 and ROCK1. The cell line is HCT15, which has a frameshift mutation in ACVR2A, a suppressor gene. The gene in question is SGMS2, and the hypothesis is that Belumosudil downregulates it. To support this, the experiment needs to have a padj < 0.05 and |log2FC| > 0.5 in the direction of downregulation.

Now, looking at the context. There are several experiments with Belumosudil (same drug) in different cell lines. Let's check those first.

Context [3] is Belumosudil in HOP62 cell line. The result here is a log2FC of +0.833, which is a strong upregulation (padj is very low, 1.67e-17). So this is a clear upregulation of SGMS2. But the target hypothesis here is a downregulation. That's a contradiction. However, the cell line here is HOP62, which is different from HCT15.

Context [6] is Belumosudil in HT-29, another cell line. The log2FC here is +0.146, which is a slight upregulation, and the padj is 0.04521 (just under 0.05, so it's significant). But again, it's an upregulation, not downregulation. So again, the same drug in different cell lines shows upregulation.

Then there's context [7], which is a different drug (Adenosine) in the same HCT15 cell line. The log2FC here is +0.142, which is a slight upregulation, but the padj is 0.7822, which is not significant. So that doesn't support either direction.

Other contexts [1], [2], [4], [5] are unrelated drugs, but maybe they can give some general idea. For example, [4] is Artesunate (same mechanism as Belumosudil? Or not sure). The result here is +0.20, which is upregulation, but not significant (padj 0.8409). Not helpful here.

Now, the key is to see if there's any evidence in the context that supports downregulation of SGMS2 in HCT15 or similar cell lines. The context for Belumosudil in different cell lines (3 and 6) both show

upregulation, but the target hypothesis is about downregulation. So this might suggest that the drug is more likely to upregulate, which would make the target hypothesis less probable.

But wait, the context for HCT15 with Adenosine (context [7]) doesn't show significant upregulation. So maybe the gene is not always responsive, but the drug in other cell lines does show upregulation. However, the target hypothesis is about HCT15. Since the context for Belumosudil in other cell lines shows upregulation, but the target hypothesis is about a different direction (downregulation), that might suggest that the target hypothesis is less likely.

However, I should also check if there's any evidence in the context that directly contradicts the target hypothesis. For example, if other experiments in the same cell line or similar cell lines show downregulation, but in this case, none of the other contexts for HCT15 show that. The same drug in other cell lines shows upregulation, and the unrelated drug in HCT15 (context [7]) isn't significant.

But I need to consider the strength of the context. The context for the same drug (Belumosudil) in different cell lines (HOP62 and HT-29) both show upregulation. That's a pattern, but the target hypothesis here is for HCT15. The question is whether this pattern in other cell lines supports or contradicts the hypothesis here.

Another thing to consider is the gene's responsiveness. The gene SGMS2 is being tested in multiple contexts, and in the same drug's context, it's upregulated in different cell lines. That suggests that the gene might be a consistent target for this drug. However, the target hypothesis here is the opposite. So this would make the probability of the target hypothesis being supported lower.

But what about the other contexts? For example, context [1] and [2] are unrelated drugs, but they also show upregulation. However, those are not directly relevant. The key is the same drug in different cell lines.

Also, the base mean in the context experiments for the same drug (context [3] and [6]) are higher, which might indicate more reliable data. But the target hypothesis here is for HCT15.

So, considering the context, the same drug in different cell lines shows upregulation of SGMS2. The target hypothesis is that it's downregulated in HCT15. Therefore, the context suggests that the drug is more likely to upregulate, which would make the target hypothesis (downregulation) less probable. However, the context doesn't have any direct evidence for the same cell line (HCT15) and the same drug.

But the problem is to weigh the evidence. The fact that in two different cell lines, the same drug upregulates the gene might suggest that the gene is a target of the drug. If the target hypothesis here is in the opposite direction, that might lower the probability. However, there's no direct evidence against it.

But in the context, the same drug in different cell lines shows upregulation, which is a positive sign for the direction. But the target hypothesis here is for downregulation, which is opposite. So that would make the probability of the target hypothesis being supported lower.

But I need to be careful here. Are there any other factors? For example, the driver mutations in the cell lines. The HCT15 cell line has a LoF mutation in ACVR2A. But the other cell lines in context [3] (HOP62) has a B2M mutation (suppressor), and HT-29 (context [6]) has APC (LoF). So the cell lines vary in their driver mutations. But the same drug in different cell lines still shows upregulation.

Therefore, based on the context, the evidence for the same drug in other cell lines showing upregulation of SGMS2 would make the target hypothesis (downregulation) less likely. However, the context doesn't have any experiments that directly contradict it.

But how does that translate into a probability? If the context shows that in two different cell lines, the same drug upregulates the gene, that's a pattern. The target hypothesis is the opposite. So the probability might be low, say 0.2 or 0.3. But I need to think carefully.

Alternatively, maybe the context doesn't have enough evidence to make a strong conclusion. However, since the same drug in other cell lines shows upregulation, it's more likely that the gene is upregulated, not downregulated. Therefore, the probability that the target hypothesis (downregulation) is supported is low.

But I need to consider the possibility that the target hypothesis could still be correct. For example, if there's a cell line-specific effect. But the context doesn't have any evidence for that here.

So, based on the context provided, I would say that the probability is around 0.2 (20% chance that the target hypothesis is supported). But I need to be precise here.

Answer

```
<probability>0.20</probability>
```

K ADDITIONAL ANALYSES OF BELIEF REVISION

This appendix reports the two probes of belief revision summarized in §5, Analysis 3. Both use the multi-domain test set and the current roster, with 256 episodes per domain.

The preceding analyses ask whether individual updates have sensible signs and magnitudes; these probes ask whether a model combines the same evidence consistently. If the stated probabilities behave like beliefs, rearranging a fixed evidence set should not materially change the result, and the joint weight of evidence of a matched pair should be predictable from the two single-stub weights. Both probes hold evidence content fixed while changing its position or composition.

The **position probe** places one compelling stub at four locations in an otherwise identical ten-stub context. The reported position-1 bias is the belief when that stub appears first minus the mean belief when it appears later. The **composition probe** presents a supporting stub, a constructed equal-and-opposite refuting stub, and the pair; its residual measures departure from adding the two single-stub weights of evidence; it covers Tahoe and DepMap only. Apps. K.1 and K.2 give the construction and results of each probe, and Fig. 12 summarizes both.

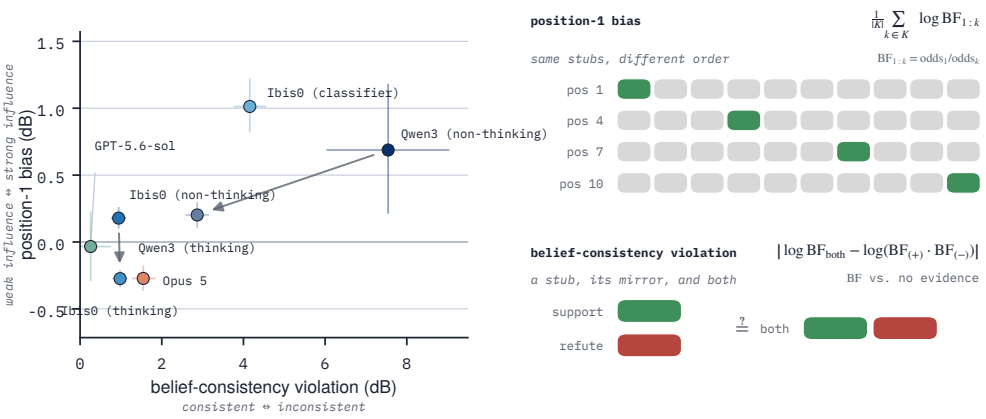

Figure 12: **Ordering effects: Given the same evidence, are models consistent in belief revision?** Each point is one model; the schematic at right shows the two manipulations. **Position-1 bias** (vertical): a single "compelling" evidence stub (supporting and highly-relevant) is placed at varying positions within an otherwise identical context. We report the stated belief when the stub appears first minus the mean stated belief when it appears in the later three positions, in decibans. Positive values indicate that the same stub moves the belief further when it appears first, a form of primacy bias. **Belief-consistency violation** (horizontal): a matched pair of supporting/refuting stubs is constructed by inverting the direction of effect while keeping all other details the same, and we measure how far the joint update departs from the product of the two single-stub updates. Zero on either axis is the unbiased value, so the lower left is best. Arrows join the two matched pairs that differ only in whether a reasoning trace is verbalized. Across both measures, thinking reduces bias. Ibis0 (classifier) is well calibrated yet shows pronounced bias on both axes, whereas RL post-training improves calibration *and* reduces the observed biases.

Consistency does not follow from aggregate forecast quality (Fig. 12): the classifier has the best Brier score of the Ibis0 models but the largest position-1 bias ($+1.01$ dB) and composition residual (4.16 dB). Post-training still improves consistency: all three trained variants reduce the Qwen3 (non-thinking) composition residual of 7.54 dB. The classifier reaches 4.16 dB, and the two RL policies reach 0.94 dB (Ibis0 non-thinking) and 0.98 dB (Ibis0 thinking), in the range of the frontier models (GPT-5.6-sol 0.26 dB, Claude Opus 5 1.55 dB).

Reasoning traces are associated with a separate shift in position dependence. Within both the Qwen3 and Ibis0 pairs, enabling a trace reduces position-1 bias by about 0.5 dB. Because the same shift appears before and after post-training, it is more naturally associated with the response process than with the proper-scoring objective alone.

K.1 Composition of supporting and refuting evidence

The conflict probe selects the clearest supporting stub in an episode and constructs a matched refuting stub of equal magnitude. In Tahoe this reverses the signs of the differential-expression fold change and Wald statistic, leaving the adjusted p-value unchanged. In DepMap it reflects the Chronos gene effect about zero and reads the accompanying dependency probability off the empirical relation between the two quantities in the test set, so that both numbers on the rendered card move together. GWAS is excluded because the catalog records published associations but not fail-

ures to associate, which a refuting stub would require. The same property restricts the directional analysis in §5.

The supporting stub, the refuting stub, and the pair are presented separately. The pair is presented in both orders and the two are averaged, so that order effects cancel and do not enter the residual. Departure from additive weight-of-evidence updating is measured as

$$\left|\log \mathrm{BF}_{\mathrm{both}} - \log\left(\mathrm{BF}_{\mathrm{support}}\mathrm{BF}_{\mathrm{refute}}\right)\right|,$$

where each Bayes factor is taken relative to the same episode with no evidence. Zero denotes exact agreement between the paired update and the sum of the two single-stub updates. The reference for each model is therefore its own single-stub responses.

Fig. 13 reports the result over 256 episodes in each of Tahoe and DepMap, grouped by training regime. Within Qwen3, the violation is 7.54 dB $[6.03, 9.04]$ in non-thinking mode and 2.87 dB $[2.58, 3.16]$ in thinking mode. Among the Ibis0 models it is 0.94 dB $[0.75, 1.13]$ for Ibis0 (non-thinking) and 0.98 dB $[0.80, 1.15]$ for Ibis0 (thinking), which differ from each other by less than their intervals, and 4.16 dB $[3.76, 4.55]$ for Ibis0 (classifier). The two frontier models give 0.26 dB $[0.01, 0.75]$ for GPT-5.6-sol and 1.55 dB $[1.27, 1.85]$ for Claude Opus 5.

The clearest effect in the figure is that both reinforcement-learning policies are far more consistent than the Qwen3 checkpoint they were trained from. The classifier is not: it violates consistency more than either frontier model and both policies, while achieving the lowest Brier score of any model in §4. Post-training against a proper scoring rule therefore does not improve this property on its own, and the output channel separates the models that share a training objective. GPT-5.6-sol is the most consistent model measured, so post-training on EDB-Bio is not required to obtain the property.

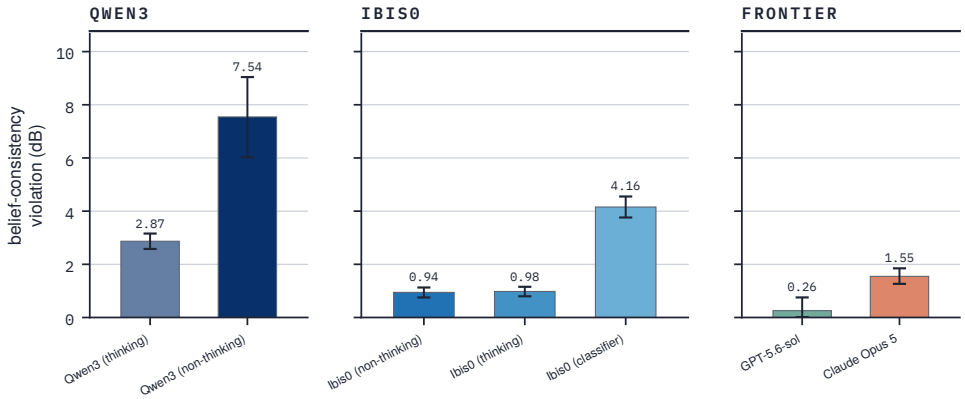

Figure 13: **Composition of conflicting evidence.** Absolute departure from additive weight-of-evidence updating for a supporting stub and its constructed equal-and-opposite counterpart, macro-averaged over Tahoe and DepMap. Panels group models by training regime and bars are ordered within each panel; the vertical scale is shared. Intervals are 95% episode bootstrap intervals with episodes resampled within domain, 256 episodes per domain. Lower values indicate closer agreement with multiplicative combination of Bayes factors.

K.2 Dependence of evidence updates on sequence position

The position probe inserts the same evidence stub at positions 1, 4, 7, and 10 of a ten-stub context and fills the remaining slots with the same neutral padding in the same order. The four conditions contain the same evidence in different arrangements, and an order-invariant model reports the same probability in all four. The context holds ten stubs rather than the sixteen of the earlier pilot because DepMap episodes carry about eleven stubs that can serve as padding; a sixteen-stub context can be constructed for only 14 of 2,000 episodes.

Reported values are the stated belief in decibans, centred within episode across the four positions to remove variation in how compelling each episode's probe stub is. The statistic of interest is the belief at the first position minus the mean over the later three. Centring cancels in any contrast between

Model	Position-1 bias (dB)	95% CI
Ibis0 (classifier)	**+1.01**	$[+0.82, +1.22]$
Qwen3 (non-thinking)	**+0.69**	$[+0.21, +1.18]$
Qwen3 (thinking)	**+0.20**	$[+0.11, +0.30]$
Ibis0 (non-thinking)	**+0.18**	$[+0.10, +0.26]$
GPT-5.6-sol	-0.03	$[-0.29, +0.23]$
Claude Opus 5	**-0.27**	$[-0.37, -0.18]$
Ibis0 (thinking)	**-0.27**	$[-0.34, -0.21]$

Table 10: **Dependence of the stated belief on the first evidence position.** Stated belief at position 1 minus the mean over positions 4, 7 and 10, in decibans, macro-averaged over the three domains with 256 episodes per domain. Positive values indicate that the same stub moves the belief further when it appears first. Intervals are 95% episode bootstrap intervals resampled within domain; bold marks an interval excluding zero.

positions, so this quantity is a raw within-episode difference in log-odds and does not depend on which positions were sampled. A contrast between the first and last positions alone would not suffice: for Ibis0 (non-thinking) the last position is as elevated as the first, and the two cancel.

Table 10 gives the statistic for each model. It is positive and separated from zero for Ibis0 (classifier), both modes of Qwen3, and Ibis0 (non-thinking), so in those four models the same stub moves the belief further when it appears first. It is negative and separated from zero for Ibis0 (thinking) and Claude Opus 5, and contains zero for GPT-5.6-sol.

The two matched pairs that differ only in whether the model verbalizes a reasoning trace move together. Paired within episode over 768 episodes, adding the trace changes first-position bias by -0.49 dB $[-0.96, -0.01]$ for Qwen3 and by -0.45 dB $[-0.54, -0.36]$ for the reinforcement-learning policy. The two estimates agree to within 0.04 dB although only one pair is trained, which suggests that deliberating before answering and training against a proper scoring rule act on this property separately. Training reduces the effect from $+0.69$ dB for Qwen3 (non-thinking) to $+0.18$ dB for Ibis0 (non-thinking), and the trace shifts it downward from that level; in the reinforcement-learning pair, the shift passes zero into a small preference for later positions. The Qwen3 interval only just excludes zero, so the paired result rests mainly on the policy pair.

The classifier, the only model that states no probability in text, shows the largest effect in the roster. It also achieves the lowest Brier score in §4.

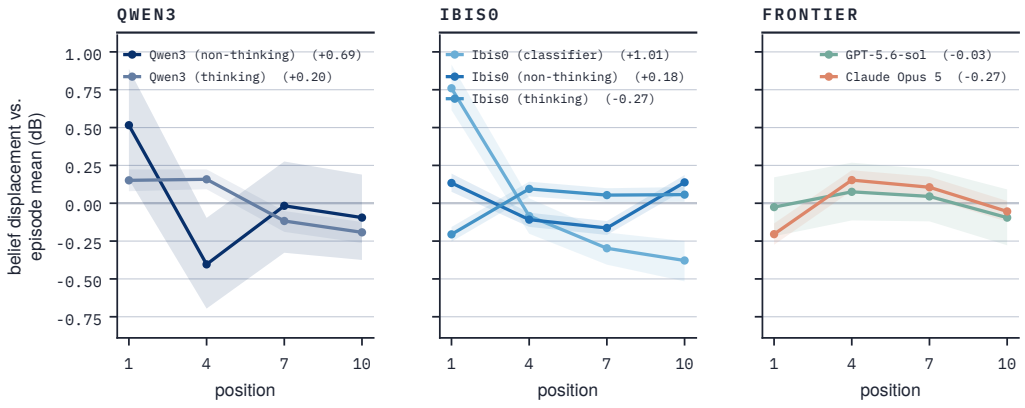

Figure 14: **Dependence of the stated belief on evidence position.** The same evidence stub placed at four positions of a ten-stub context, with the remaining slots held to identical neutral padding in identical order. Panels group models that differ in whether a reasoning trace is verbalized. Values are decibans, centred within episode, macro-averaged over the three domains with episodes resampled within domain; bands are 95% intervals, 256 episodes per domain. Legend entries give the position-1 bias of Table 10. A flat line at zero indicates a belief that does not depend on the order of the evidence.

L STRUCTURED AND NUMERIC-ONLY BASELINES ON TAHOE

Rhaister (Svensson et al., 2026) predicts a differential-expression outcome for a specified cell, treatment, and gene. We convert its predicted log fold change to a probability by assigning its sign relative to the target direction and applying a sigmoid, and also report a hard-thresholded version. The published episodes do not record drug concentration, so treatment identity is defined by drug. Rhaister can provide a forecast only when its fit data contain the target gene under a compatible treatment.

The *prompt-local* setting fits Rhaister only to the evidence stubs shown in the prompt. The *all-context* setting uses training-split target rows and all structured context evidence, with held-out target rows excluded. A numeric-only gradient-boosted tree is trained on the Tahoe training split using DESeq2 summary statistics and counts, without gene, drug or cell-line identities or text; it produces a forecast for every test episode.

Baseline	Fit data	Coverage	Brier (↓)	Hard Brier (↓)
Rhaister prompt-local	Prompt evidence only	85.1%	0.2310	0.4130
Rhaister all-context	Train targets + context	88.5%	0.2295	0.3968
Numeric-only GBM	DESeq2 numeric features	100.0%	0.2206	n/a
All-domain classifier	Three-domain training	100.0%	0.2063	n/a

Table 11: Tahoe-only structured baselines on the 2,000-episode test set. Coverage is the fraction of episodes for which a baseline produces a forecast.

M EXTENDED RELATED WORK

This appendix expands the summary in §7.

Scientific reasoning and evidence benchmarks. Benchmarks for scientific AI have moved from expert-written question answering (Rein et al., 2024) toward tasks drawn from research practice: long-context reasoning over scientific documents (Cui et al., 2025), research coding (Tian et al., 2024), data-driven discovery from datasets (Majumder et al., 2025; Chen et al., 2025), simulated environments in which an agent must form hypotheses and run experiments (Wang et al., 2022; Jansen et al., 2024), and biology research workflows spanning literature retrieval, protocol troubleshooting, and study-scale computational analysis (Laurent et al., 2024; 2026; Mitchener et al., 2025a; Koch et al., 2026). Literature agents answer questions by retrieving and synthesizing published evidence (Skarlinski et al., 2024). Closest in form to our task is claim verification, where SciFact and SciFact-Open ask whether a retrieved abstract supports or refutes a claim (Wadden et al., 2020; 2022) and Matter-of-Fact asks whether a literature-supported claim is feasible (Jansen et al., 2025). These tasks score a categorical verdict about existing evidence. EDB-Bio instead asks for a probability that a held-out measurement will support the hypothesis, scores it with a proper scoring rule, and controls the relevance and sufficiency of the evidence pool so that belief revision can be measured update by update.

Calibration and confidence estimation in LLMs. Work on LLM calibration has centered on confidence in answer *correctness* for factual or multiple-choice question answering, where large models can be reasonably calibrated when probabilities are read from token logits or elicited as $P(\text{true})$ (Kadavath et al., 2022; Jiang et al., 2021; Desai & Durrett, 2020; Geng et al., 2024). A recurring finding is that alignment (RLHF and instruction tuning) degrades this calibration and leaves models systematically overconfident (OpenAI, 2023; Zhu et al., 2023), with a corresponding risk that users over-trust confident-sounding output (Rathi et al., 2025). Proposed remedies span verbalized confidence elicitation (Lin et al., 2022; Tian et al., 2023), logit- and sequence-likelihood scores (Duan et al., 2024), sampling-consistency measures (Wang et al., 2023a; Chen & Mueller, 2024), semantic-entropy estimators (Kuhn et al., 2023; Farquhar et al., 2024), and auxiliary or post-hoc calibrators (Ulmer et al., 2024; Shen et al., 2024; Guo et al., 2017). Test-time reasoning does not reliably improve calibration and can worsen it (Mei et al., 2026), which motivates our direct-mode elicitation of a single probability. The task closest to ours is probabilistic forecasting scored by strictly proper rules (Gneiting & Raftery, 2007; Brier, 1950; Murphy, 1973), where LLMs now approach human-crowd accuracy on real-world events (Halawi et al., 2024; Schoenegger et al., 2024). We elicit calibrated probabilities for *novel scientific hypotheses* under a controlled, manipulable

evidence pool, and report discrimination (AUROC) and calibration (ECE (Naeini et al., 2015)) separately rather than accuracy alone.

Proper-scoring post-training. A recent line of work post-trains language models against strictly proper scoring rules (Gneiting & Raftery, 2007), using the Brier score or log loss of a verbalized confidence as a reinforcement-learning reward or a supervised objective (Damani et al., 2026; Puri et al., 2026; Bani-Harouni et al., 2026; Leng et al., 2025; Xu et al., 2024). These methods target confidence in the correctness of an answer or a reasoning chain on question-answering tasks, so the label is whether the model's own output is right. We apply the same family of objectives to a target external to the model: the outcome of a held-out experiment, conditioned on an evidence pool. The model must weigh evidence instead of assessing its own reliability, and the reward comes from resolved scientific outcomes. We compare a negative-Brier reward in direct and thinking modes with binary cross-entropy on a scalar head (§3).

Bayesian experimental design with LLMs. Classical Bayesian experimental design chooses the experiment that maximizes expected information gain (EIG) (Lindley, 1956; Chaloner & Verdinelli, 1995; Rainforth et al., 2024), but estimating EIG is costly, requiring nested Monte-Carlo integration (Rainforth et al., 2018), variational or contrastive bounds (Foster et al., 2019; Poole et al., 2019), or amortized design policies (Foster et al., 2021; Blau et al., 2022; Müller et al., 2022). A growing line instead drives information-theoretic selection from a language model's own beliefs: LLMs pick EIG-maximizing queries for interactive information gathering (Hu et al., 2024; Mazzaccara et al., 2024; Kobalczyk et al., 2025; Choudhury et al., 2026; Grand et al., 2026), elicit user intent by asking informative questions (Li et al., 2025; Handa et al., 2024), serve as uncertainty-bearing surrogates for Bayesian optimization in scientific domains (Ramos et al., 2026; Liu et al., 2024b), and plan iterative perturbation screens (Huang et al., 2024; Hao et al., 2025). Prior studies also report that direct prompting for information gathering can reduce performance (Li et al., 2024) and that off-the-shelf agents may ask uninformative questions (Grand et al., 2026). The multi-domain evaluation in §6 applies model-based EIG selection to evidence-deliberation episodes from Tahoe, DepMap, and GWAS.

Belief revision, order effects, and weights of evidence. Whether language models revise beliefs coherently has been studied through cognitive-bias probes and Bayesian-updating tasks. Models show primacy effects in multiple-choice and long-context settings (Wang et al., 2023b; Liu et al., 2024a; Raimondi & Gabbrielli, 2025) and anchor on initial numeric hints (Lou & Sun, 2024); surveys catalog these and other human-like biases (Sumita et al., 2024). On sequential evidence, models approach Bayesian posteriors given enough observations (Gupta et al., 2025) but discount earlier evidence relative to an ideal filter (Zhang et al., 2025a) and depart from Bayesian invariances even when instructed to follow them (Kim et al., 2025). We measure updates in decibans, the unit of weight of evidence that Good and Turing introduced for sequential evidence (Good, 1950; 1979), and summarize each update as a log Bayes factor (Kass & Raftery, 1995). Whether a model's weights of evidence are the right size, the question of §5, has to our knowledge not been asked of language models before; the nearest prior work is support theory's account of how human probability judgments depend on how the evidence is described (Tversky & Koehler, 1994).

Prediction of perturbation outcomes in biology. Predicting the transcriptional response to a perturbation is an active area. GEARS uses gene-interaction graphs to extrapolate to unseen combinations (Roohani et al., 2024), and single-cell foundation models such as Geneformer and scGPT are fine-tuned for the task (Theodoris et al., 2023; Cui et al., 2024). Systematic benchmarks find that these models do not yet outperform simple linear baselines (Ahlmann-Eltze et al., 2025; Wei et al., 2026). The virtual-cell program aims at models that predict perturbed cell state across contexts (Bunne et al., 2024; Roohani et al., 2025; Adduri et al., 2025), and perturbation atlases such as Tahoe-100M supply the training data (Zhang et al., 2025b). These models predict the measurement itself from molecular inputs. EDB-Bio asks a different question: given prior measurements as text, how confident should a reader be that the next one will support a hypothesis. The two are complementary, and a perturbation model could serve as the outcome simulator in our experimental-design loop (§6).

AI for science and autonomous discovery. Language-model agents now automate parts of the research cycle, and the systems differ mainly in which parts. One line couples a model to laboratory automation so that it plans and executes experiments (Boiko et al., 2023). A second generates and ranks hypotheses or research proposals, in biomedicine (Gottweis et al., 2026) and in machine

learning, where the loop extends to running the experiments and writing the paper (Lu et al., 2024; Yamada et al., 2025). A third organizes several agents into a research team that iterates between hypothesis, experiment, and analysis, with wet-lab validation of the resulting designs and candidate treatments (Swanson et al., 2025; Ghareeb et al., 2026), and recent systems extend this loop to long-horizon, data-driven discovery (Mitchener et al., 2025b). Each of these systems must decide which hypotheses to pursue and when the evidence suffices, decisions that depend on the model's probabilistic judgment. EDB-Bio measures that judgment in isolation, and the experimental-design results in §6 show one way to use calibrated belief for better decisions inside such a loop.