Foresight-Phys: A Benchmark for Forecasting the Results of Physical Experiments

Nikita Kazeev [1] **Ian Babich** [1]

Abstract

Forecasting has emerged as a hallmark aspiration of general-purpose AI, with a super-human performance ceiling: it pairs extraordinary difficulty, where the best human experts fail regularly, with rigorous real-world validation. Scientific research is itself a forecasting problem—finite resources force an exploration/exploitation tradeoff that relies on prophesying the results of studies before conducting them. We introduce **Foresight-Phys**, a benchmark that asks AI to predict the results of physical experiments from the newest arXiv preprints. We present the framework along with results for 135 experiments and 661 typed result fields extracted from 26 physics preprints first released in 2026, after the GPT-5.x training cutoff. The extraction is automated, suiting an online benchmark fed by regularly pulling fresh papers.

GPT-5.5 attains a mean prediction quality of 94.5% on the headline results papers state in their abstracts—an impressive, but inflated score, as published preprints report overwhelmingly successful outcomes, so a model steeped in the literature can anticipate a central result without reasoning about the specific experiment. A name-only baseline that hides the setup and reveals only the target quantity already scores 84.6%, leaving the full experimental description worth under ten points. On numeric quantities GPT-5.5 reaches a relative CRPS of 0.41 yet places only 52.4% of its predictions within one standard deviation of the truth: genuine intuition for physical scale, but marked overconfidence against an ideal Gaussian observer's 68.3%.

Code and data are available at `https://github.com/kazeevn/Foresight-Phys`.

[1]Institute for Functional Intelligent Materials, National University of Singapore. Correspondence to: Nikita Kazeev <kna@nus.edu.sg>, Ian Babich <ian.babich@u.nus.edu>.

Forecasting as a New Frontier of Intelligence Workshop at ICML, Seoul, South Korea. 2026.

1. Introduction

Forecasting is one of the most difficult problems facing general-purpose AI, demanding the ability to internalize patterns in dynamic environments and reason about consequences in the noisy real world. A large slice of AI benchmarks consists of ever-harder expert-level problem solving (Rein et al., 2023; He et al., 2024; Wang et al., 2026; Zhu et al., 2026). Such benchmarks are gated by what humans can pose and solve, so difficulty saturates as models surpass the graders. Forecasting offers a natural escape: the ground truth is set by reality, not by an examiner. Halawi et al. (2024); Schoenegger et al. (2024); Karger et al. (2024) have emphasized this view for prediction markets and geopolitical events; we apply it to physical experiments, where the answer is determined by nature and lies at the edge of our collective understanding. The argument is sharpest where intuition pays: a researcher bets a year of grant funding on whether a tweak to a sample, beamline, or circuit will move a quantity in the expected direction and by the expected order of magnitude, balancing novelty against a reasonable chance that the experiment nevertheless succeeds.

We contribute 1. a paper-derived benchmark of 135 experiments and 661 result fields across five typed answer formats 2. an LLM pipeline for extraction verification, masking, prediction, and scoring—numeric, discrete, and symbolic, the last via an LLM-as-judge 3. results for four frontier OpenAI models evaluated against their respective physical observables 4. a model-independent annotation of each field's centrality, *ex-ante* surprise, and leakage-sufficiency, together with a name-only baseline and a decision-usefulness evaluation (selection, direction, and order of magnitude) that separate genuine foresight from observable-type priors

2. Benchmark Design

Paper sourcing. We collected candidate arXiv physics preprints first released between January and May 2026, after the GPT-5.x training cutoff. Domains span condensed-matter ARPES, quantum optics, quantum key distribution, cold atoms, top-quark physics, and topological matter. We use preprints rather than journal versions because the multi-month peer-review lag carries the risk of the results leaking

into training corpora. After suitability filtering (below), 26 papers carry usable benchmark questions, yielding 135 experiments and 661 result fields.

Two-pass extraction. Both extraction passes use GPT-5.5 via the OpenAI API with structured output. Pass 1 emits a list of `(experiment_description, experiment_results)` tuples per paper. The system prompt requires every sweep limit, apparatus parameter, environmental condition, and unit convention to live in the description, and pushes only *intrinsic*, predictable observables into the targets—explicitly excluding extrinsic artifacts (e.g., the random thickness of a mechanically exfoliated flake) and qualitative claims. Pass 2 re-reads each extracted record and discards experiments unsuitable for blind prediction: descriptions that leak the result, tautological setups, or insufficient context.

Typed answer schema. Each result carries one of five types: `float` (501 fields), `integer` (21), `bool` (73), `categorical` (57, with explicit allowed-value lists), or `formula` (9, symbolic). The typed schema lets us mix numerical proximity, discrete accuracy, and symbolic equivalence in a principled way, and lets one experiment expose multiple independently scorable predictions.

Prediction. For each experiment, every `result` field is recursively replaced with the literal string `"TO_PREDICT"`, while the experiment description, result descriptions, types, units encoded in keys, and allowed categorical values remain visible. The masked JSON is sent with a fixed system prompt and a Pydantic structured-output schema. Numeric fields must return a distribution; boolean fields return `result` and `prob_true`; categorical fields return `result` plus a probability distribution over the allowed values; and formula fields return a symbolic expression plus a confidence.

Field annotation. Different measurements have different importance. To separate technical details from a paper's core contribution, a separate pass labels every result field with its scientific *centrality* (headline, key-supporting, secondary, or setup/control), its *ex-ante surprise* (whether the value follows from generic physical priors before the experiment), and whether the masked description alone already determines it (*leakage-sufficient*). The same pass groups fields into *comparison sets*—a swept series or a baseline-versus-intervention contrast—tagged with the control that varies and the decision the set encodes: which condition is best, the direction of a trend, the sign of an effect, or an order of magnitude. Because a model's own sense of importance is itself a judgement, we ground the top of the centrality scale in a separate factual check—whether the abstract actually states this field's specific result—and let the annotated levels apply only beneath it.

3. Metrics

The five answer types call for modality-specific proper scoring rules, which we map onto a single *quality* axis in $[0, 1]$ so heterogeneous fields can be compared and aggregated; full definitions, constants, and baselines are in Section A. Numeric (`float`, `integer`) forecasts are elicited as (p_{10}, p_{50}, p_{90}) quantiles under a `normal` or `log_normal` fit and scored by the Gaussian CRPS—relative to $|y|$ for `normal` and in dex for `log_normal` targets—with calibration read off the empirical $|z| < 1, 2, 3$ fractions against the ideal Gaussian observer ($68.3\%/95.4\%/99.7\%$). Boolean and formula fields use the Brier score, the latter judged for symbolic equivalence by an LLM-as-judge (GPT-5.4-nano, intentionally the weakest model in our lineup); categorical fields use a multiclass Brier over an explicit allowed-value list. The quality map is $1 - \widetilde{\mathrm{CRPS}}/3$ for numeric, $1 - \mathrm{Brier}$ for boolean and formula, and $\max(0, 1 - \frac{1}{2}\mathrm{Brier})$ for categorical fields; because it rewards calibrated uncertainty, a correct but overconfident forecast can score below a modest, honest one. All headline numbers are the paper-macro of the experiment-macro—per-field scores averaged within each experiment, then each paper, then across papers—so no field-rich paper dominates. The annotation further lets us report quality on *headline*, centrality, and *surprise* strata; a *name-only* baseline exposing only the typed key isolates the *foresight lift*; and each comparison set is scored as a decision—selection accuracy and normalised regret against a no-model random pick, directional accuracy and sign Brier, and order-of-magnitude coverage.

4. Results

Aggregate prediction quality separates the four models (Fig. 1): GPT-5.5 leads at 87.2%, GPT-5.4 follows closely at 85.0%, and the mini and nano variants trail at 78.7% and 76.9% respectively. As the difficulty profile below shows, however, these aggregates partly reflect benchmark composition—most fields are solved by every model—rather than ceiling performance. The overall ranking is consistent with the numeric proper scores and the boolean/categorical quality scores, while the symbolic subset is small enough that we treat it as a diagnostic rather than a stable model-ranking axis.

Numeric. GPT-5.5 places 52.4% of numeric predictions within 1σ, 76.3% within 2σ, and 88.7% within 3σ of ground truth, achieving a mean relative CRPS of 0.413 (raw 3.28 in mixed units) and numeric quality of 86.2%. The strongest two models are close on relative CRPS—GPT-5.5 at 0.413 and GPT-5.4 at 0.423—while mini and nano trail at 0.650

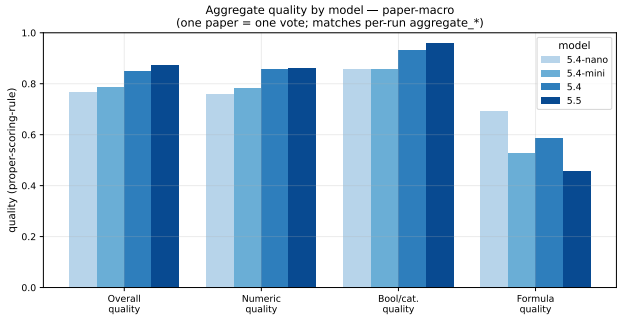

Figure 1. Paper-macro prediction quality by model, split by modality (overall, numeric, bool/categorical, formula). The four models separate in the overall and numeric panels, while bool/categorical quality stays uniformly high.

and 0.719. Coverage, however, shows all models remain overconfident (Figure 2): GPT-5.5's 1σ coverage is well below the 68.3% expected for a calibrated Gaussian observer, and GPT-5.4, mini, and nano fall further behind at 42.0%, 35.6%, and 29.8%. The models often find the right physical scale, but their uncertainty intervals are too narrow. Read as a point forecast, GPT-5.5's median lands within a factor of three of the truth on 77.0% of numeric fields and within a decade on 91.7%, falling to 59.5% and 79.5% for nano—an order-of-magnitude accuracy that, unlike relative CRPS, admits the name-only baseline below.

Discrete and symbolic. Boolean and categorical exact accuracy is high across all models, ranging from 83.5% to 95.8%—comfortably above the 29.3% categorical uniform baseline—with Brier-derived quality from 85.6% to 95.9%. The models thus often identify the correct qualitative physical regime (topological vs. trivial, gapped vs. gapless, monotonic increase vs. decrease). Formula fields are noisier: across only nine symbolic targets, exact accuracy ranges from 14.3% to 28.6% and quality from 45.8% to 69.4%, so we use the symbolic subset mainly to exercise the extraction and judging machinery rather than to draw scaling conclusions.

Difficulty profile. Under the analysis definition of "correct" (quality at least 0.5), the benchmark has a substantial solved mass and a smaller but important tail. Across all 661 fields, 483 (73.1%) are *trivial* (every model correct), 90 (13.6%) *easy* (one model wrong), 52 (7.9%) *discriminative* (split), 18 (2.7%) *hard* (only one right), and 18 (2.7%) *intractable* (none right). Restricting quality to the 178 non-trivial fields deflates the headline and sharpens the ranking: GPT-5.5 falls to 71.3%, GPT-5.4 to 65.6%, mini to 52.6%, and nano to 49.1%, widening the top-to-bottom gap from 10.3 to 22.2 points. The aggregate 87.2% is therefore best read as a benchmark-composition figure: most of it is trivial mass, and the discriminating fields remain genuinely hard. That tail is dominated by numeric floats (41 of 52 discriminative, 15 of 18 intractable): qualitative regime recognition is largely solved, whereas calibrated continuous quantities still expose differences between systems.

Crucial versus trivial. We mark a field crucial when the abstract states its specific result, checked field-by-field by an LLM; reassuringly, this factual marker corroborates the model's own headline labels, which are $9.6\times$ likelier to appear in the abstract than the rest (55.2% versus 5.7%). On these 81 abstract-grounded crucial fields GPT-5.5 scores 94.5%, *above* its 87.2% aggregate, and every model clears its own aggregate: the worry that models nail only incidental quantities is misplaced—if anything, the most important results are the easiest. The reason is sobering. The name-only baseline—prediction from the typed key alone, experiment hidden—already reaches 84.6% on these fields for GPT-5.5 (and 73.0% for nano), so reading the experiment buys a foresight lift of only 9.9 points; overall, GPT-5.5 reading nothing but field names (78.6%) even outscores GPT-5.4-nano with the full setup (76.9%). The surprise axis agrees: fields flagged *surprising* are predicted about as well (90.4%) as conventionally *implied* ones (91.3%), with merely *uncertain* fields lowest (86.8%)—unsurprising once one notes that a *published* surprise is still a confirmed positive. Two checks rule out cheaper stories: excluding the $37/661$ leakage-sufficient fields barely moves the aggregate ($87.2\% \to 86.8\%$), and collapsing each comparison set to a single trend vote leaves it essentially unchanged (86.9%). The headline is therefore real but *prior-dominated*—most of it is recoverable without reading the experiment.

Decision usefulness. Recast as the decision a scientist faces—which knob to turn, in which direction, to what scale—the picture sharpens (Table 1). Across 83 scorable comparison sets the models are strong on the *direction* of an effect: GPT-5.5 calls the sign or trend correctly 97.9% of the time with a well-calibrated sign Brier of 0.05, and even nano reaches 81.2%. Selecting the single best condition from a sweep is far harder: top-1 accuracy is only about 50%, and while GPT-5.5 cuts normalised selection regret to 0.33 from a random pick's 0.50, GPT-5.4 adds *no* decision value at all—and nano, despite trailing on every quality axis, edges out the rest on selection regret (0.32). The single order-of-magnitude set is too few to score separately; field-level coverage above (77.0% within a factor of three for GPT-5.5) speaks to the same question.

5. Related Work

Knowledge-and-recall benchmarks such as ScienceQA (Lu et al., 2022), MMLU (Hendrycks et al., 2021), SciEval (Sun et al., 2024), and C-Eval (Huang et al., 2023) have been complemented by expert-level reasoning benchmarks:

Table 1. Decision usefulness over the 83 scorable comparison sets, each scored per model. Selection accuracy, regret, and decision value are for the 33 "which condition is best" sets; regret is normalised in $[0, 1]$ (lower is better) and the no-model random-pick baseline has regret 0.50, so the decision value is that baseline minus the model's regret. Directional accuracy and sign Brier are over the 49 direction/sign sets.

Model	Sel.↑	Regret↓	Value↑	Dir.↑	Brier↓
GPT-5.5	0.52	0.33	+0.16	0.98	0.05
GPT-5.4	0.45	0.50	+0.00	0.85	0.12
GPT-5.4-mini	0.45	0.39	+0.11	0.79	0.12
GPT-5.4-nano	0.49	0.32	+0.18	0.81	0.13

GPQA (Rein et al., 2023), OlympiadBench (He et al., 2024), FrontierScience (Wang et al., 2026), and CritPt (Zhu et al., 2026). All are bounded by what human experts can solve and grade; Foresight-Phys instead asks models to forecast outcomes whose ground truth is fixed by physical reality.

The closest prior work is BrainBench (Luo et al., 2025), in which LLMs predict outcomes of neuroscience experiments and exceed domain experts. We adapt the spirit to physics with typed answers, post-cutoff sourcing, and an LLM-driven extraction pipeline. TastyBench (Parv Mahajan, 2025) measures research taste via citation-velocity prediction, a delayed and noisy proxy; Lyu et al. (2025); Park et al. (2026) assess LLMs' ability to predict computational-experiment outcomes; we target physical observables.

6. Limitations and Future Work

Positivity bias. Published preprints overwhelmingly over-represent successful experiments, so the benchmark inherits a survivorship-biased view of what physics is forecastable. This is the most significant limitation behind the headline 94.5% quality on abstract-stated results, and the reason that number rises rather than falls as we tighten the marker: the more central a result, the more it is a confirmed, publishable positive that a model steeped in the literature can anticipate. It is mitigated somewhat because we predict not *whether an experiment worked* but, where available, *which value was measured*. Still, the 80 setup/control fields—chiefly "did the process work?" success flags—are near-constant positives predicted at 90.2% by GPT-5.5, and the name-only baseline points the same way: a model that has internalised the literature's positivity reaches 84.6% on the crucial fields without reading the experiment. The bias is real and inflates precisely the headline that stratification deflates; we surface it rather than hide it by reporting the headline against the name-only baseline and reading quality on non-trivial subsets net of setup/control fields, while repairing the corpus remains future work (Section B).

Human baseline. We frame the super-human ceiling as a property of the forecasting task—reality, not an examiner, fixes the answer—rather than a measured margin over experts. We do not yet quantify how domain physicists who have read the masked method section fare on the same fields; doing so would turn the aggregate and hard-field scores into a calibrated human-relative ranking.

Data leakage. Post-cutoff sourcing mitigates but cannot fully eliminate leakage: some works may have been discussed in public forums before model release, though fresh preprints make this less likely than mature benchmark questions. Embargoed experimental data would be the cleanest solution; a weaker but practical check is post-factum web-search detection over titles, arXiv identifiers, and distinctive target phrases.

Paper parsing verification. Parsing must keep descriptions complete enough to specify the physical system and controls without leaking measured values through data-dependent phrasing. We combine manual spot checks with an automated audit: the annotation flags each field's leakage-sufficiency and whether the abstract states its result, a lexical term-overlap diagnostic cross-checks the latter, and a literal-and-normalized numeric scan flags values restated in the descriptions. The high trivial fraction $(483/661)$ is itself an audit signal rather than direct evidence of leakage—many regime questions are genuinely easy—and independent human verification of a random subset of records and annotations is the natural next step.

Scope. We evaluate a single frontier model family (GPT-5.$\{4,5\}$) under chain-of-thought alone, without computational tools or external retrieval. Standardizing on one family suffices to show the benchmark differentiates model tiers; both restrictions and their rationale are detailed in Section B.

7. Conclusion

Forecasting unobserved physical outcomes is a clean testbed whose ceiling is naturally super-human, pushing foundation models' forecasting and world-modelling abilities to their limits. Frontier models achieve meaningful performance but remain far from that ceiling: stripping the trivial tasks drops GPT-5.5 to 71.3% on the model-separating fields, and numeric forecasts stay systematically overconfident—a super-human bar for the task, not super-human performance for today's models. We controld the positivity bias uaing a name-only baseline, which shows that most of the performance is recoverable from observable-type priors. The pipeline, dataset, and per-paper reports are released to grow with each new arXiv month and model.

Acknowledgements

We thank Kostya Novoselov for fruitful discussions and feedback on this idea.

This research is supported by the Ministry of Education, Singapore, under its Research Centre of Excellence award to the Institute for Functional Intelligent Materials (I-FIM, project No. EDUNC-33-18-279-V12). This research/project is supported by the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award No: AISG3-RP-2022-028).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Halawi, D., Zhang, F., Yueh-Han, C., and Steinhardt, J. Approaching human-level forecasting with language models, 2024. URL https://arxiv.org/abs/2402.18563.

He, C., Luo, R., Bai, Y., Hu, S., Thai, Z. L., Shen, J., Hu, J., Han, X., Huang, Y., Zhang, Y., Liu, J., Qi, L., Liu, Z., and Sun, M. Olympiadbench: A challenging benchmark for promoting agi with olympiad-level bilingual multimodal scientific problems, 2024. URL https://arxiv.org/abs/2402.14008.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. Measuring massive multitask language understanding, 2021. URL https://arxiv.org/abs/2009.03300.

Huang, Y., Bai, Y., Zhu, Z., Zhang, J., Zhang, J., Su, T., Liu, J., Lv, C., Zhang, Y., Fu, Y., et al. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in neural information processing systems*, 36:62991–63010, 2023.

Karger, E., Bastani, H., Yueh-Han, C., Jacobs, Z., Halawi, D., Zhang, F., and Tetlock, P. E. ForecastBench: A Dynamic Benchmark of AI Forecasting Capabilities, 2024. URL https://arxiv.org/abs/2409.19839.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering, 2022. URL https://arxiv.org/abs/2209.09513.

Luo, X., Rechardt, A., Sun, G., Nejad, K. K., Yáñez, F., Yilmaz, B., Lee, K., Cohen, A. O., Borghesani, V., Pashkov, A., et al. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2):305–315, 2025.

Lyu, B., Huang, S., Liang, Z., Sun, Q.-A., and Zhang, J. SURGE: On the Potential of Large Language Models as General-Purpose Surrogate Code Executors, 2025. URL https://arxiv.org/abs/2502.11167.

Park, J., Mendes, E., Stanovsky, G., and Ritter, A. Anticipatory evaluation of language models, 2026. URL https://arxiv.org/abs/2509.20645.

Parv Mahajan, Yilin, y. Tastybench: Toward measuring research taste in llm, 2025. URL https://www.lesswrong.com/posts/Mxsy7wYvsCRv5dGrw/tastybench-toward-measuring-research-taste-in-llm.

Rein, D., Hou, B. L., Stickland, A. C., Petty, J., Pang, R. Y., Dirani, J., Michael, J., and Bowman, S. R. GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023. URL https://arxiv.org/abs/2311.12022.

Schoenegger, P., Tuminauskaite, I., Park, P. S., and Tetlock, P. E. Wisdom of the silicon crowd: LLM ensemble prediction capabilities rival human crowd accuracy, 2024. URL https://arxiv.org/abs/2402.19379.

Sun, L., Han, Y., Zhao, Z., Ma, D., Shen, Z., Chen, B., Chen, L., and Yu, K. Scieval: A multi-level large language model evaluation benchmark for scientific research, 2024. URL https://arxiv.org/abs/2308.13149.

Wang, M., Lin, R., Hu, K., Jiao, J., Chowdhury, N., Chang, E., and Patwardhan, T. FrontierScience: Evaluating AI's Ability to Perform Expert-Level Scientific Tasks, 2026. URL https://arxiv.org/abs/2601.21165.

Zhu, M., Tian, M., Yang, X., Zhou, T., Yuan, L., Zhu, P., Chertkov, E., Liu, S., Du, Y., Ji, Z., Das, I., Chen, Q., Cao, J., Du, Y., Yu, J., Wu, P., He, J., Su, Y., Jiang, Y., Zhang, Y., Liu, C., Huang, Z.-M., Jia, W., Wang, Y., Jafarpour, F., Zhao, Y., Chen, X., Shelton, J., Young, A. W., Bartolotta, J., Xu, W., Sun, Y., Chu, A., Colussi, V., Akers, C., Brooks, N., Fu, W., Zhao, J., Qi, M., Mu, A., Yang, Y., Zang, A., Lyu, Y., Mai, P., Wilson, C., Guo, X., Zhou, J., Inafuku, D., Xue, C., Gao, L., Yang, Z., Hein, Y., Kahn, Y., Zhou, K., Luo, D., Wilson, J. D., Reilly, J. T., Bandak, D., Press, O., Yang, L., Wang, X., Tong, H., Chia, N., Huerta, E., and Peng, H. Probing the Critical Point (CritPt) of AI Reasoning: a Frontier Physics Research Benchmark, 2026. URL https://arxiv.org/abs/2509.26574.

A. Scoring rules and metric definitions

We detail here the per-type proper scoring rules, their constants, and the random baselines summarised in Section 3.

Numeric (`float`, `integer`). Numeric forecasts are elicited as distributions, not point estimates. For every numeric field the model must output `distribution` $\in \{$`normal`, `log_normal`$\}$ and quantiles (p_{10}, p_{50}, p_{90}), with p_{50} used as the point prediction. For `normal` fields, we fit a Gaussian in the target's original units with $\mu = p_{50}$ and $\sigma = (p_{90} - p_{10})/(2\Phi^{-1}(0.9)) = (p_{90} - p_{10})/2.5631$, then set $z = (y - p_{50})/\sigma$. For `log_normal` fields, all quantiles and the reference value must be positive; the same fit is performed after applying $\log_{10}$, so σ and the resulting CRPS are measured in dex and $z = (\log_{10} y - \log_{10} p_{50})/\sigma$. In both cases the numeric score is the Gaussian CRPS,

$$\mathrm{CRPS} = \sigma \left[z(2\Phi(z) - 1) + 2\phi(z) - \frac{1}{\sqrt{\pi}} \right].$$

Raw CRPS is capped at 30 to keep malformed or catastrophically overconfident forecasts finite. Because raw CRPS carries physical units, the cross-paper metric uses a relative CRPS: $\widetilde{\mathrm{CRPS}} = \mathrm{CRPS}/|y|$ for `normal` targets with $y \neq 0$, and the dex CRPS directly for `log_normal` targets. Relative CRPS is capped at 3 per field. A normal target with $y = 0$ has no relative CRPS and is omitted from relative-quality averages, although its raw CRPS is still recorded. Calibration is reported through the empirical fractions with $|z| < 1$, $|z| < 2$, and $|z| < 3$. Unlike the discrete types, numeric fields admit no well-defined uniform-guess baseline—relative CRPS depends on the scale the forecaster chooses—so we instead benchmark calibration against the ideal Gaussian observer, whose $1\sigma/2\sigma/3\sigma$ coverage is $68.3\%/95.4\%/99.7\%$.

Boolean. Boolean forecasts return both an argmax, `result` $\in \{$`true`, `false`$\}$, and a calibrated probability `prob_true`. The exact-match statistic uses the predicted boolean (falling back to `prob_true` ≥ 0.5 if needed). The proper score is the binary Brier score, $\mathrm{Brier} = (p - y)^2$, where $p = \Pr($`true`$)$ and $y \in \{0, 1\}$ is the reference. Boolean quality is $1 - \mathrm{Brier}$; an uninformative $p = 0.5$ forecast has Brier 0.25 and quality 0.75, while the exact-match random-guess baseline is 0.5.

Categorical. Each categorical field carries an explicit allowed-value list of size $k \in \{2, 3, 4, 5\}$ (e.g. *s-/p-/d-/f-*wave; below/at/above the Fermi level). The model returns a best-guess `result` and probabilities over the allowed values. The scorer restricts probabilities to the reference allowed list, clips negative entries to zero, and renormalizes; if the explicit `result` is valid it is used as the argmax, otherwise the highest-probability class is used. Exact-match accuracy is recorded. The proper score is the multiclass Brier score $\mathrm{Brier} = \sum_i (p_i - \mathbb{1}_{i=y})^2$, and categorical quality is $\max(0, 1 - \frac{1}{2}\mathrm{Brier})$. Because k varies across questions, the per-question uniform-guess exact accuracy is $1/k_q$ and its quality is $1 - \frac{1}{2}\frac{k_q - 1}{k_q}$; averaging over our k distribution gives a categorical random baseline of 0.293 exact accuracy and 0.646 quality.

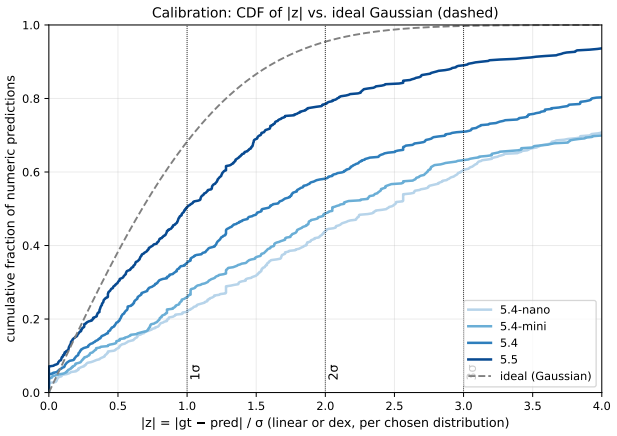

Figure 2. Distribution of prediction errors and coverage on numeric fields, per model. GPT-5.5 captures over half of its numerical predictions within 1σ, though this remains significantly below the 68.3% expected for a calibrated Gaussian distribution.

Formula. For symbolic fields the prediction is a closed-form expression plus a confidence in $[0, 1]$. Exact equality after whitespace removal is accepted immediately; otherwise the expression is scored by an LLM-as-judge (GPT-5.4-nano) against the reference. The judge receives the experiment description, result description, reference formula, and predicted formula, and is asked to apply a strict equivalence rubric: 1. the prediction must contain every term that appears in the reference; 2. it must contain no spurious extra terms; 3. signs and dimensional prefactors must match; 4. the functional dependence on each input variable must match. Equivalence up to harmless formatting differences, equivalent notation, algebraic rearrangement, and renaming of bound or dummy variables is accepted. Each $(\text{experiment}, \text{reference}, \text{prediction})$ triple is cached so that re-scoring is deterministic across reruns. Formula accuracy is the fraction judged equivalent. Formula quality uses the same Brier form as booleans, $1 - (c - y)^2$, where c is the model's confidence and $y = 1$ iff the formula is judged equivalent. A guesser that cannot reconstruct the expression scores near-zero accuracy, and an uninformative confidence-0.5 forecast scores 0.75 quality. The judge is intentionally the weakest model in our lineup, so any judge-induced bias penalises the strongest predictors at least as much as the weakest—a conservative choice for relative comparisons.

Importance strata, baselines, and decisions. The annotation turns one aggregate into several: we report quality restricted to *headline* fields and stratified by centrality and surprise. To separate physical foresight from observable-type priors we add a *name-only* baseline that re-runs prediction with the experiment and result descriptions withheld, leaving only the typed key (which encodes the observable and its units) and any allowed values; the *foresight lift* is full-context quality minus this baseline. For numeric fields we also record order-of-magnitude accuracy—the fraction whose median lands within a factor of three (and within a decade) of the truth—which, unlike relative CRPS, has a meaningful prior-only floor. Finally, each comparison set is scored as a decision: top-1 *selection accuracy* and a normalised *regret* against the oracle, with a no-model random pick as the reference whose gap is the *decision value*; a *directional accuracy* and a proper sign Brier read off the two predicted Gaussians; and order-of-magnitude coverage over the set's members.

B. Extended limitations and future work

Model diversity. We restrict our present evaluation to the GPT-5.{4,5} model family. Our primary objective is to introduce and validate the Foresight-Phys benchmarking methodology, rather than to provide an exhaustive survey of all available models. By standardizing on this frontier model family, we demonstrate the benchmark's ability to differentiate between model tiers and calibrate performance against state-of-the-art reasoning capabilities. The modular architecture of our evaluation harness allows for straightforward expansion to a broader range of open and closed-source models in future work.

Harness. The models in this study rely exclusively on chain-of-thought to make their predictions. In reality, scientists have access to computational tools and external resources. The latter is the harder of the two to grant fairly: unrestricted web access would trivially leak the target. A useful compromise is carefully sandboxed, proxied access to a corpus frozen at the dataset collection start.

Repairing the corpus. Reducing the positivity bias remains future work along three complementary lines: mining Supporting Materials, Appendices, and PhD theses, where failed experiments are more commonly recorded; harvesting refutation literature ("Comment on…" and "Reply to…" pieces, failed replications); and synthetically injecting a predictable flaw into otherwise successful experiments so that the correct forecast is the negative one.