FROM SAMPLING TO SELECTION: GROUNDING LLM REASONING WITH PREDICTIVE MODELS OF PHYSICAL SYSTEMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time sampling can place a correct answer in the candidate pool without providing a reliable way to identify it. Scientific workflows offer an additional source of evidence: predictive models of the physical system. We compare three placements of the same predictive-state evidence during LLM inference—post-generation candidate selection, one-step evidence-conditioned revision, and direct state-to-decision mapping when a task rule is specified—and ask whether such evidence helps and when it makes the language model unnecessary. In a controlled simulation-grounded study and an independently implemented diagnostic study on public combined-cycle power-plant data, candidate coverage far outruns majority voting: with eight matched samples a correct candidate is present for 91.3% of questions, while self-consistency returns 63.8% and predictive-state selection 81.3%, recovering 63.6% of the selection gap at no additional sampling cost. Two controls delimit the mechanism. Selection fails when the score evaluates the wrong task quantity and, more stringently, when the required state change is smaller than the predictor's own error. And when the task's answer is a deterministic function of the predicted state, applying the task rule directly to that state (95.0%) matches the best LLM interface (95.6%) with no language-model calls at all. The results give a conditional account of predictive-state grounding and a concrete criterion for when language inference adds value over the predictor alone.

1 INTRODUCTION

Test-time computation improves language-model reasoning by generating, revising, or selecting additional candidate solutions (Wei et al., 2022; Wang et al., 2023; Snell et al., 2025). These operations address different parts of inference. Sampling can increase the chance that a useful answer appears; a selection rule must still recognize it. Recent analyses formalize differences between self-consistency and best-of-N selection (Huang et al., 2026), while learned verifiers and tool-integrated verification provide increasingly capable selection signals (Lightman et al., 2024; Kang et al., 2026). Our focus is a setting common in scientific and engineering workflows: an independently trained model already predicts task-relevant physical state.

Such a predictor supplies evidence with a different provenance from the candidate-generating LLM. A fluent chain may misestimate temperature, reverse an ordering, or cross a threshold incorrectly even when its language is plausible. Numerical simulators, learned surrogates, neural operators, and conventional regressors can provide an external state estimate for these decisions (Karniadakis et al., 2021; Li et al., 2021; Kovachki et al., 2023). This motivates the question studied here: *how should predictive-state evidence be used once the LLM has produced one or more candidate answers?*

We compare three placements while keeping the evidence source explicit (Figure 1). *Grounded selection* samples multiple candidates and ranks them against predicted state without exposing that state to the LLM. *Evidence-conditioned revision* shows the predicted quantities to the model for one revision. *Direct model decision* applies a specified task rule to predicted state and serves as a no-LLM control for the final state-to-decision mapping. The experiments use simple inference rules so that differences can be attributed to the source and placement of predictive-state evidence.

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.

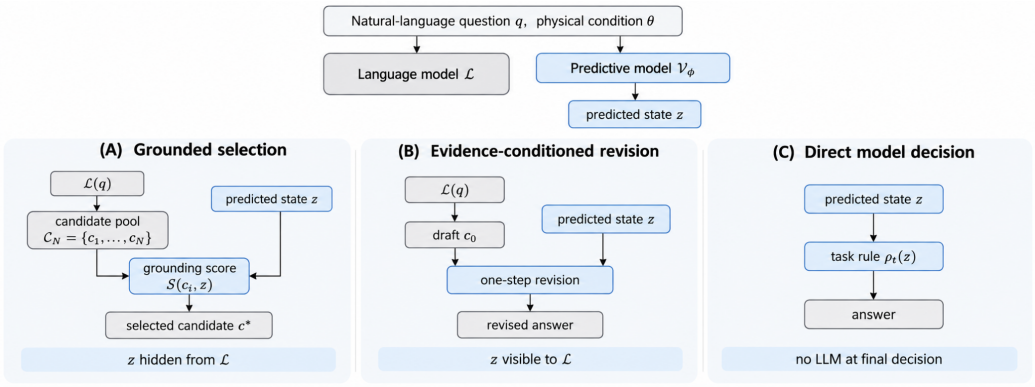


Figure 1: Three placements of the same predictive-state evidence. Grounded selection keeps z hidden from the LLM and uses it only to rank sampled candidates. Evidence-conditioned revision exposes z for one revision. Direct model decision applies a specified task rule to z without an LLM at the final decision.

We evaluate these interfaces in two complementary settings. A controlled laser–arc hybrid welding study uses simulation-derived labels and a separately trained surrogate across 139 questions, five task families, four base LLMs, and four levels of predictive-signal fidelity. Self-consistency reaches 36.7% and grounded selection 75.5%. An independent study uses public Combined Cycle Power Plant (CCPP) measurements, a separately implemented random-forest predictor, and a frozen 160-question diagnostic set. At $N = 8$, self-consistency reaches 63.8%, grounded selection reaches 81.3%, and 91.3% of candidate pools already contain a correct answer. Because self-consistency and grounded selection reuse exactly the same candidates, this comparison isolates the selection signal. Additional controls test two sources of error: inaccurate state estimates and scores that do not match the task decision.

Our empirical contributions are:

- We compare three placements of predictive-state evidence during LLM inference: candidate selection, evidence-conditioned revision, and direct state-to-decision mapping.
- Using exactly matched candidate pools in the independent study, we separate candidate availability from selection: at $N = 8$, candidate coverage is 91.3%, self-consistency accuracy 63.8%, and grounded-selection accuracy 81.3%.
- Controlled state-fidelity variation, a score–task mismatch case, and a direct state-to-decision control delimit when predictive-state evidence helps, when the selector fails, and when a final language-model transformation adds no measurable value.

2 Related Work

Test-time scaling and verification. Chain-of-thought and self-consistency allocate additional inference-time computation to reasoning and repeated sampling (Wei et al., 2022; Wang et al., 2023). Best-of-N, process-reward models, and related verifier-based methods use an explicit score to choose among candidates (Lightman et al., 2024; Snell et al., 2025). Recent theory establishes different sample-complexity behavior for self-consistency and best-of-N under the same proposal distribution (Huang et al., 2026), while empirical studies examine how verification depends on generator capability, verifier capability, and task difficulty (Zhou et al., 2026). Numerical-claim verification likewise uses trained verifiers to select among test-time samples (Chungkham et al., 2025). Our study focuses on a different source of selection evidence: state variables predicted independently by a model of the physical system, with the evidence source held fixed while its placement in inference changes.

External tools and scientific reasoning. Tool-augmented methods expose external computation to language models, from program execution and general actions (Yao et al., 2023; Gao et al., 2023;

Table 1: Relation to common test-time inference patterns. The comparison isolates the placement of independently predicted physical state.

Setting	Selection / feedback signal	Hidden	Visible	Direct
Self-consistency (Wang et al., 2023)	vote over sampled answers	No	No	No
Learned verifier / best-of-N (Lightman et al., 2024; Chungkham et al., 2025)	learned correctness score	Yes	No	No
Tool-integrated verification (Kang et al., 2026)	external tool output + verifier	Yes	Varies	No
This study	predicted physical state	Yes	Yes	Yes

Schick et al., 2023) to scientific toolsets and data-analysis environments (Ma et al., 2024; Chen et al., 2025). Tool-integrated verification also uses external execution to filter or validate candidates during test-time scaling (Kang et al., 2026). Scientific-verifier work studies answer verification across scientific disciplines (Zheng et al., 2026). Evidence-conditioned revision in our study belongs to the broader tool-use family. Our comparison holds the source of predictive-state evidence fixed and changes its placement: hidden post-generation selection, visible revision, or direct state-to-decision mapping.

Predictive models of physical systems. Scientific machine learning provides physics-informed models, neural operators, reduced-order surrogates, and conventional regressors for approximating physical systems (Karniadakis et al., 2021; Li et al., 2021; Kovachki et al., 2023). These models are usually evaluated by predictive fidelity. We use their state estimates as inference-time evidence and study how predictive fidelity and score-task match affect downstream LLM decisions. The independent study intentionally uses a conventional random forest so that the result does not depend on a particular scientific-ML architecture.

3 Predictive-State Evidence in LLM Inference

3.1 Problem formulation

Let q be a natural-language question about physical parameters θ, and let $c \sim p_L(c \mid q)$ denote a reasoning candidate sampled from language model L. An independently trained predictive model V_ϕ maps the physical condition to task-relevant state variables,

$$z = V_\phi(\theta). \tag{1}$$

The evaluation target is derived from observed or simulated ground truth and is kept separate from V_ϕ. The predictive model can be inaccurate; its role is to provide an independent estimate of task-relevant state during inference.

For a pool $\mathcal{C}_N = \{c_1, \ldots, c_N\}$, define oracle candidate coverage as

$$A_{\text{pool}}(N) = \frac{1}{M} \sum_{m=1}^{M} \mathbb{I}\left[\exists c_i \in \mathcal{C}_N^{(m)} : \ c_i \text{ is correct}\right]. \tag{2}$$

This quantity separates candidate availability from the selection rule used to produce the final answer.

3.2 Grounded selection

Grounded selection samples N candidates using the same candidate format as the matched self-consistency baseline. A parser extracts the final answer and, when present, explicit claims about task-relevant physical quantities. The external model remains hidden from the LLM. Each candidate receives a grounding score

$$s_i = S(c_i, z), \qquad c^* = \arg\max_{c_i \in \mathcal{C}_N} s_i. \tag{3}$$

Table 2: Controlled study on 139 simulation-grounded welding questions. The table compares LLM-only inference, the two predictive-state interfaces, and the direct model control.

Method	Overall	Num.	Risk	CFact.	Comp.	Select.
Direct	.324	.000	.455	.611	.640	.111
CoT	.360	.111	.545	.556	.560	.148
Self-Consistency ($N = 8$)	.367	.028	.700	.444	.600	.148
Grounded selection	.755	.528	**1.000**	**1.000**	.800	**.556**
Evidence-conditioned revision	.849	**1.000**	.939	.944	**.920**	.407
Direct model decision	.935	1.000	1.000	1.000	.960	.700

S combines agreement between stated physical quantities and z with consistency between the candidate answer and the task rule. Detailed parsing and scoring rules are provided in Appendix A.1.3. No scoring parameter is tuned on test accuracy.

3.3 Evidence-conditioned revision

Evidence-conditioned revision treats V_ϕ as readable evidence. The LLM first produces a draft c_0. The relevant components of z are then added to a revision prompt,

$$\hat{y}_{\text{vis}} = L(q, c_0, z). \tag{4}$$

We use a single revision step. This interface tests the value of giving the model direct access to the same physical evidence used by grounded selection.

3.4 Direct model decision

For tasks with a deterministic decision rule ρ_t over the predicted state, we also evaluate

$$\hat{y}_{\text{direct}} = \rho_t(V_\phi(\theta)). \tag{5}$$

This serves as a systems control, quantifying the value of language inference after the relevant state and decision rule are already available.

4 Controlled Simulation-Grounded Study

4.1 Setting

The controlled study uses aluminum laser–arc hybrid welding. High-fidelity CFD simulations provide evaluation labels derived from peak temperature, melt-pool depth, and maximum local composition deviation (Chen et al., 2018; 2020). A fixed offline-trained surrogate predicts the same task-relevant quantities for inference-time verification. Ground-truth labels never come from the surrogate.

The evaluation contains 139 questions: 36 numerical predictions, 33 risk diagnoses, 18 counterfactual judgments, 25 binary comparisons, and 27 constrained selections. This setting provides controlled variation of task type and predictive signal.

4.2 External grounding improves reliability

Table 2 reports the main results. Additional LLM-only inference produces little improvement: CoT reaches 36.0% and self-consistency 36.7%. Grounded selection reaches 75.5%, and evidence-conditioned revision reaches 84.9%. Direct model decision with the learned surrogate reaches 93.5%, showing that several tasks become largely determined by the predicted state once the relevant state is available.

The task breakdown matters. Evidence-conditioned revision is strongest for numerical prediction and comparison, where the predicted quantity can be used directly. Grounded selection is stronger on risk, counterfactual, and constrained-selection tasks in this controlled setting. The preferred interface varies by task.

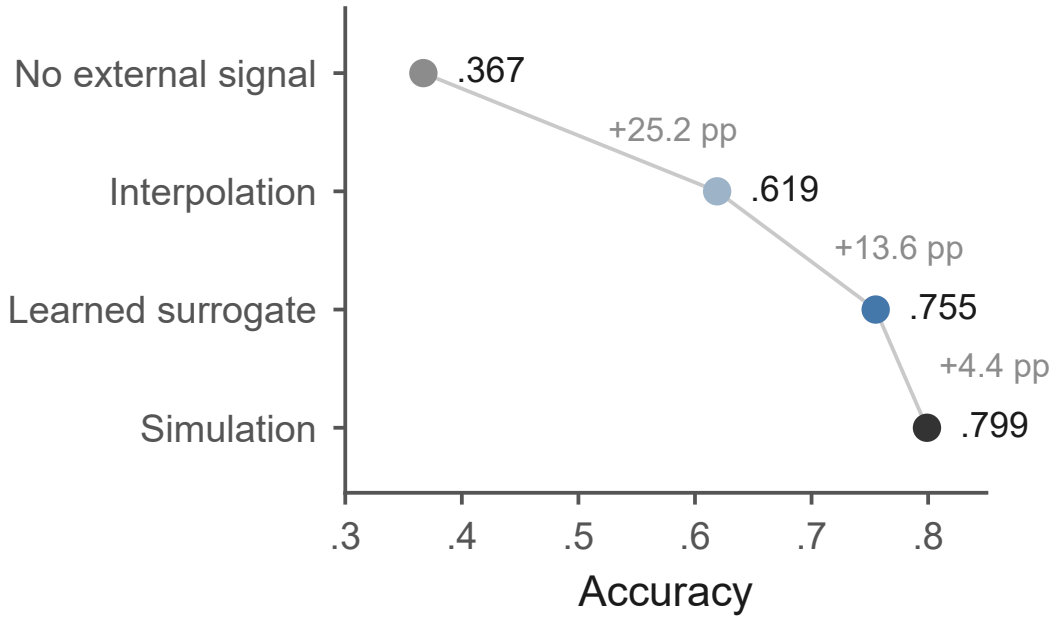


Figure 2: Controlled variation of the external predictive signal. Grounded-selection accuracy increases as the state estimate becomes more faithful, with smaller increments near the simulation reference.

Table 3: Cross-model accuracy in the controlled study.

Method	GLM-4-flash	GPT-4o-mini	GPT-4.1-mini	GPT-4o
CoT	.360	.353	.410	.374
Self-Consistency	.367	.360	.388	.338
Grounded selection	.755	.655	.683	.683
Evidence-conditioned revision	**.849**	**.842**	**.899**	**.892**

4.3 Predictive-model fidelity and test-time sampling

Replacing the external signal while holding the hidden-selection mechanism fixed produces a clear ordering: no external signal 36.7%, trilinear interpolation 61.9%, learned surrogate 75.5%, and simulation 79.9%. The final 4.4-point gap is much smaller than the earlier gains, indicating diminishing empirical returns to predictive-model fidelity in this setting.

Across four base LLMs, grounded selection achieves 65.5–75.5% accuracy and evidence-conditioned revision 84.2–89.9%, while CoT and self-consistency remain between 33.8% and 41.0% (Table 3). The controlled effect is therefore not tied to the primary GLM model.

5 Independent External Validation

5.1 Independent diagnostic study on CCPP

We use the Combined Cycle Power Plant dataset from the UCI Machine Learning Repository (Tufekci & Kaya, 2014). It contains 9,568 full-load operating points collected from 2006–2011 with four hourly averaged inputs—ambient temperature, exhaust vacuum, ambient pressure, and relative humidity—and net electrical energy output (PE) as the target. After removing 41 exact duplicate rows, we split the remaining 9,527 observations 80/20 into 7,622 training and 1,905 test examples (seed 42), ensuring that no identical input vector crosses the split.

The independently implemented predictor is a 300-tree random-forest regressor trained only on the training split. On held-out data it obtains $R^2 = 0.960$, MAE $= 2.41$ MW, RMSE $= 3.43$ MW,

Table 4: CCPP predictor robustness. Pairwise accuracy uses the same 2 MW margin as the primary analysis.

Split	R^2	MAE	Spearman ρ	Pairwise
Primary random	.9599	2.412	.9774	.9643
Cluster-separated	.9514	3.122	.9302	.9217

Table 5: Independent CCPP diagnostic study on 160 held-out questions. Confidence intervals are question-level bootstrap intervals.

Method	Accuracy	95% CI
Direct	.544	[.469, .619]
CoT	.594	[.519, .669]
Self-Consistency ($N = 8$)	.638	[.563, .706]
Grounded selection ($N = 8$)	.813	[.750, .875]
Evidence-conditioned revision	**.956**	[.925, .988]
Direct model decision	.950	[.913, .981]

Spearman $\rho = 0.977$, and 96.4% pairwise ordering accuracy with a 2 MW margin. The predictor is a conventional random forest, so the source and use of external evidence is easy to audit.

From held-out observations we deterministically generate 160 questions, 32 each for numerical prediction, threshold judgment, pairwise comparison, matched-change directional comparison, and constrained selection. Ground truth always comes from measured PE. Predictor estimates never define labels. Prompts, margins, aggregation rules, and the grounding score were frozen before the first model call. Nine automated integrity checks test train/test overlap, duplicate questions, label recomputation, leakage, units, and parser validity. We evaluate Kimi K3 with thinking disabled under Direct, CoT, Self-Consistency ($N = 8$), Grounded Selection ($N = 8$ using the same candidate pool and format), Evidence-Conditioned Revision, and Direct Model Decision. All API responses are cached. Confidence intervals use 10,000 question-level bootstrap resamples; method differences use paired bootstrap.

Predictor robustness. The primary study uses the frozen random split above. We separately prespecified a predictor-only check that standardizes the four inputs, groups observations with K-means ($k = 20$, seed 42), assigns whole clusters to train or test, and retrains the same 300-tree random forest without tuning. Predictive fidelity remains high under this cluster-separated split (Table 4). This check does not change the benchmark or any LLM result.

5.2 Replication across an independent physical system

Figure 3 and Table 5 show a large separation between ungrounded sampling and external grounding. Direct prompting reaches 54.4%, CoT 59.4%, and self-consistency 63.8%. Grounded selection reaches 81.3%, a 17.5-point gain over self-consistency (paired 95% CI $[10.6, 24.4]$). Evidence-conditioned revision reaches 95.6%, a 31.9-point gain (CI $[24.4, 40.0]$). The no-LLM direct model decision reaches 95.0%, statistically indistinguishable from evidence-conditioned revision under the paired bootstrap. Grounded selection reuses exactly the same eight sampled candidates as self-consistency, so this 17.5-point gain requires no additional LLM generation; only the lightweight predictive-model inference is added for selection. Evidence-conditioned revision uses two LLM calls per question, compared with eight sampled calls for self-consistency. Table 7 reports the measured call and token counts.

Task-level behavior. Performance varies across task types (Table 6). Grounded selection is strongest on numerical prediction, threshold judgments, and constrained selection, but it does not improve the matched-change directional task over self-consistency. Evidence-conditioned revision performs strongly across all five task families. The matched-change result is examined in Section 5.5.

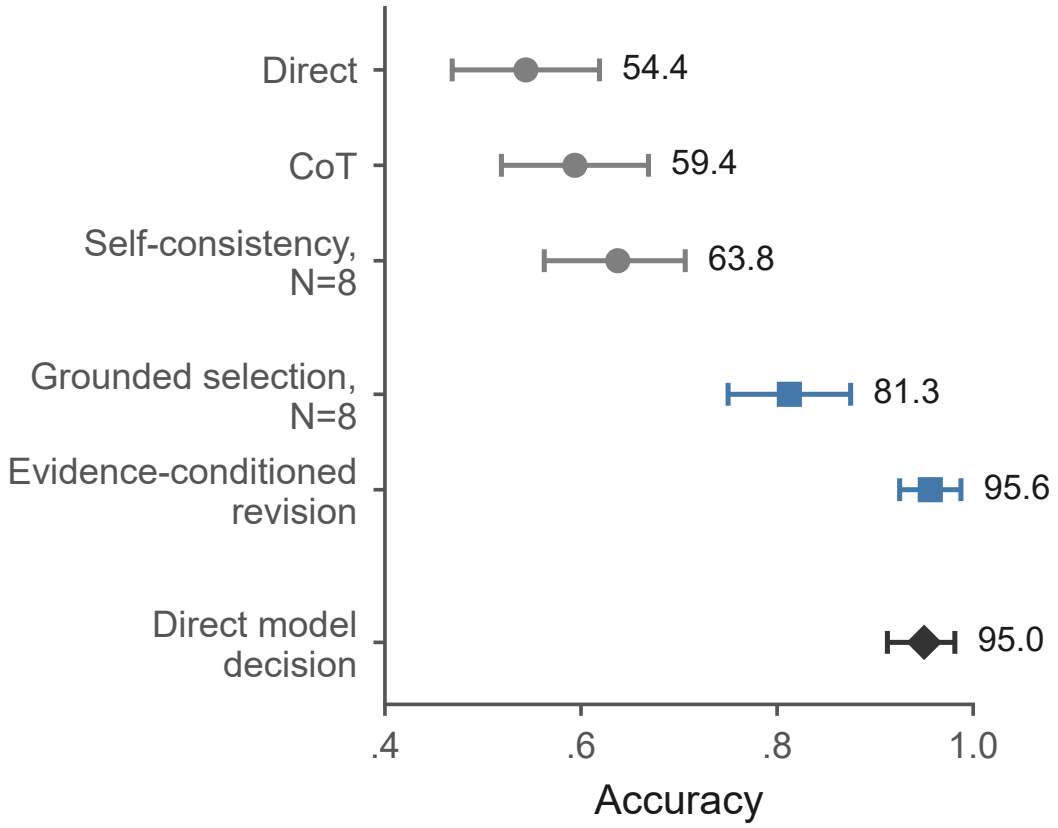


Figure 3: Independent CCPP diagnostic study on 160 held-out questions. Points show accuracy and whiskers show question-level 95% bootstrap intervals. Grounded selection and evidence-conditioned revision use the same independently trained state predictor; direct model decision is a no-LLM control.

Table 6: Task-wise accuracy in the independent CCPP study (32 questions per task).

Task	Direct	CoT	SC	Grounded	Revision	Direct model
Numerical	.313	.344	.500	.844	1.000	1.000
Threshold	.656	.813	.906	1.000	1.000	1.000
Pairwise	.875	.875	.906	.906	.938	.938
Matched change	.656	.594	.500	.500	.875	.844
Selection	.219	.344	.375	.813	.969	.969

5.3 Inference budget

The matched-pool design also controls LLM generation cost. Self-consistency and grounded selection both use eight sampled completions per question and have identical measured token counts; grounded selection adds one random-forest inference for selection. Evidence-conditioned revision uses one draft and one revision. Table 7 reports the cached API usage. The accuracy ordering therefore cannot be attributed to a larger language-model sampling budget for grounded selection.

5.4 Sampling expands coverage; selection determines final accuracy

The complete candidate logs let us measure A_{pool}. At $N = 8$, at least one correct candidate is present for 91.3% of questions. Self-consistency reaches only 63.8%, while grounded selection reaches 81.3%. Relative to the gap between self-consistency and candidate coverage, grounded selection recovers 63.6% of the available improvement.

Table 7: Inference usage in the independent CCPP study. Token counts are means per question from cached API usage fields.

Method	Acc.	Calls	Input tok.	Output tok.
Direct	.544	1	254.2	26.8
CoT	.594	1	269.6	202.4
Self-Consistency ($N = 8$)	.638	8	2162.2	1752.1
Grounded selection ($N = 8$)	.813	8	2162.2	1752.1
Evidence-conditioned revision	.956	2	897.9	346.2
Direct model decision	.950	0	0.0	0.0

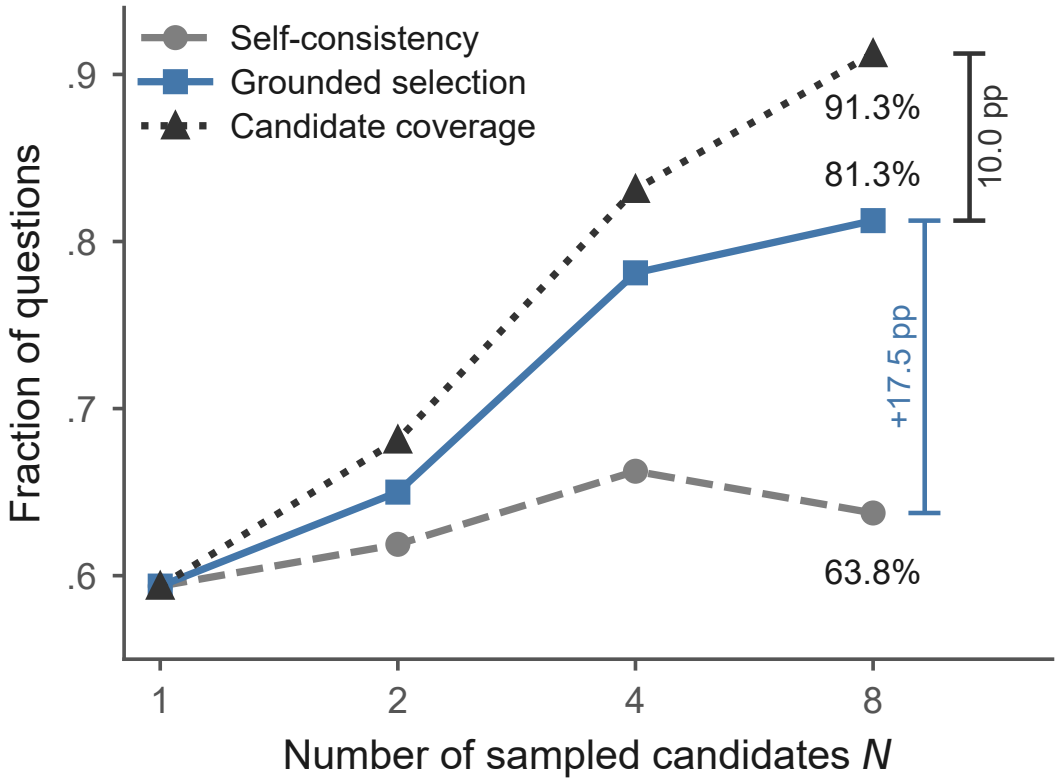


Figure 4: Sampling expands candidate availability faster than majority voting converts it into final accuracy. Self-consistency and grounded selection use the same sampled candidate pool at each N. At $N = 8$, a correct candidate appears for 91.3% of questions, self-consistency reaches 63.8%, and grounded selection reaches 81.3%, recovering 63.6% of the available selection gap.

The trajectories across N show where the gap emerges. As N increases from 1 to 2, 4, and 8, candidate coverage rises from 59.4% to 68.1%, 83.1%, and 91.3%. Grounded selection follows the expanding opportunity set, rising from 59.4% to 65.0%, 78.1%, and 81.3%. Self-consistency changes from 59.4% to 61.9%, 66.3%, and 63.8%. Across this range of N, candidate coverage grows much faster than majority-vote accuracy.

5.5 Score–task mismatch in matched-change questions

Matched-change direction judgments provide a useful negative case: self-consistency and grounded selection both achieve 50.0%. Prediction quality is high in this setting: the random-forest predictor has $R^2 = 0.960$ on held-out PE. Instead, the frozen hidden score measures magnitude agreement while the task asks for the sign of a change. Across matched-change candidates, the frozen score has correlation $r = -0.05$ with correctness, while a direction-agreement score has correlation $r = 0.72$.

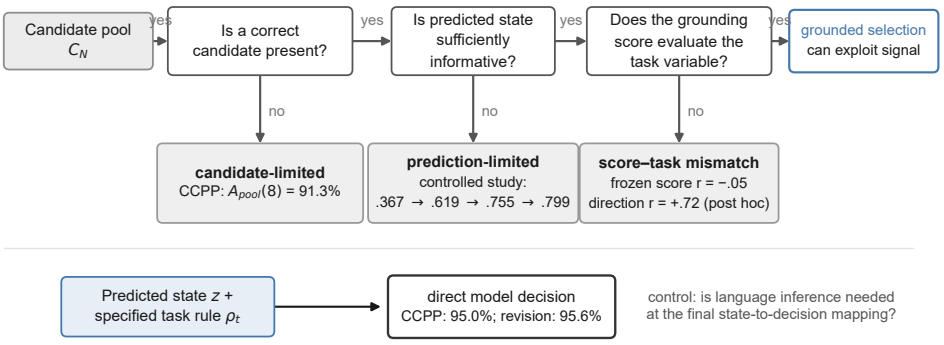


Figure 5: Failure sources observed in the experiments. The lower branch reports the separate direct-decision control.

The frozen score is therefore poorly matched to the directional decision. Appendix D turns this negative case into a constructive one: re-scoring the *same* cached pools with a task-aligned score (agreement on the decision variable, not magnitude) raises matched-change selection from 50.0% to 71.9% and overall grounded selection from 81.3% to 88.1%, with the residual gap set by the predictor's own direction accuracy (84.4%) when the required state change is near its noise floor.

6 Failure Analysis

Figure 5 summarizes the failure sources directly supported by the experiments: an absent correct candidate, an insufficiently informative state estimate, or a ranking score that does not track the task decision. These cases arise at different stages of the inference pipeline and need not occur together.

The lower branch in Figure 5 is a control, not an additional failure category. Direct model decision reaches 93.5% in the controlled study and 95.0% in CCPP; the latter is statistically indistinguishable from evidence-conditioned revision at 95.6%. This control concerns only the final state-to-decision mapping; more open-ended roles for language models are outside the experiments here.

7 Limitations

The evidence covers two physical systems, not scientific reasoning broadly. CCPP is low-dimensional (four inputs, one target), lacks timestamps, and its matched-change questions use observational pairs rather than causal interventions. A cluster-separated predictor check retains high predictive fidelity, although evaluation on higher-dimensional and multi-regime systems remains outside the present study. CCPP uses one base LLM, while cross-model robustness is measured in the controlled study. The approach also presumes access to a predictive model of task-relevant state.

8 Conclusion

We studied how predicted physical state can be used during inference after an LLM generates candidate answers. In the independent study, candidate coverage rises substantially with additional samples while majority-vote accuracy changes much less; grounded selection recovers part of this gap using the same candidate pools. The controlled fidelity sweep and the matched-change negative case show that selection also depends on the quality of the predicted state and on whether the score evaluates the quantity required by the task. For tasks whose answer is already determined by predicted state and a specified rule, the direct-model control provides a useful reference for the final state-to-decision mapping.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Appendix A provides the task construction, predictive-signal conditions, candidate scoring, and cross-model details for the controlled welding study. Appendix B specifies the complete independent CCPP evaluation, including data preprocessing, predictor training, the frozen 160-question set, task-specific correctness rules, and inference conditions. Self-consistency and grounded selection reuse exactly the same sampled candidate pools.

The remaining appendices report the candidate-coverage and sampling analysis, cached inference accounting, the pre-specified cluster-separated predictor check, the matched-change negative result, statistical evaluation, prompting and parsing protocol, and benchmark-integrity checks. The supplementary material provides the frozen CCPP benchmark, data-processing and predictor-training code, parsing and scoring tools, evaluation outputs, statistical analyses, and representative model outputs used for the reported results. The candidate-selection, candidate-coverage, error, and inference-accounting analyses can be recomputed from these released artifacts.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for designing and providing feedback on research methodology and experiments, providing feedback on mathematical arguments, and interpreting results. We have not used generative AI tools for generating synthetic data sets, formulating mathematical claims, assisting in the writing of proofs, or implementing methods, and the remaining required disclosure tasks are not applicable to this work. Additionally, we used generative AI tools for literature discovery, brainstorming, drafting and editing parts of the manuscript, creating and editing software code and figures, and simulated peer-review feedback. We have reviewed all AI-assisted work: references identified with AI assistance were verified against the original publications, and AI-assisted code was re-run with its outputs checked against the experimental artifacts reported in this paper. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Xuyang Chen, Gang Yu, Xiuli He, Shaoxia Li, and Haibin Miao. Effect of droplet impact on molten pool dynamics in hybrid laser-mig welding of aluminum alloy. *The International Journal of Advanced Manufacturing Technology*, 96:209–222, 2018.

Xuyang Chen, Gang Yu, Xiuli He, Shaoxia Li, and Zhiyong Li. Numerical study of heat transfer and solute distribution in hybrid laser-mig welding. *International Journal of Thermal Sciences*, 149:106182, 2020.

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*, 2025.

Primakov Chungkham, Venktesh V, Vinay Setty, and Avishek Anand. Think right, not more: Test-time scaling for numerical claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 24345–24363, 2025. doi: 10.18653/v1/2025.findings-emnlp.1322.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *International Conference on Machine Learning*, 2023.

Baihe Huang, Shanda Li, Tianhao Wu, Yiming Yang, Ameet Talwalkar, Kannan Ramchandran, Michael Jordan, and Jiantao Jiao. Sample complexity and representation ability of test-time scaling paradigms. In *International Conference on Learning Representations*, 2026.

Minki Kang, Jongwon Jeong, and Jaewoong Cho. T1: Tool-integrated verification for test-time compute scaling in small language models. In *International Conference on Learning Representations*, 2026.

George Em Karniadakis, Ioannis G. Kevrekidis, Lu Lu, Paris Perdikaris, Sifan Wang, and Liu Yang. Physics-informed machine learning. *Nature Reviews Physics*, 3:422–440, 2021.

Nikola Kovachki, Zongyi Li, Burigede Liu, Kamyar Azizzadenesheli, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Neural operator: Learning maps between function spaces with applications to pdes. *Journal of Machine Learning Research*, 24(89):1–97, 2023.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations*, 2021.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let's verify step by step. In *International Conference on Learning Representations*, 2024.

Yubo Ma, Zhibin Gou, Junheng Hao, Ruochen Xu, Shuohang Wang, Liangming Pan, Yujiu Yang, Yixin Cao, and Aixin Sun. Sciagent: Tool-augmented language models for scientific reasoning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 15701–15736, 2024. doi: 10.18653/v1/2024.emnlp-main.880.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2023.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling llm test-time compute optimally can be more effective than scaling parameters for reasoning. In *International Conference on Learning Representations*, 2025.

Pinar Tufekci and Heysem Kaya. Combined cycle power plant. UCI Machine Learning Repository, 2014.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837, 2022.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R. Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023.

Shenghe Zheng, Chenyu Huang, Fangchen Yu, Junchi Yao, Jingqi Ye, Tao Chen, Yun Luo, Ning Ding, Lei Bai, Ganqu Cui, and Peng Ye. Sci-verifier: Scientific verifier with thinking. In *International Conference on Learning Representations*, 2026.

Yefan Zhou, Austin Xu, Yilun Zhou, Janvijay Singh, Jiang Gui, and Shafiq Joty. Variation in verification: Understanding verification dynamics in large language models. In *International Conference on Learning Representations*, 2026.

A Additional Experimental Details

This appendix provides additional experimental and implementation details for the two studies in the main paper. The controlled welding study and the independent Combined Cycle Power Plant (CCPP) study serve complementary roles. The controlled study evaluates several task types, multiple language models, and predictive signals of different quality. The CCPP study uses public measured data, a frozen question set, matched candidate pools, cached language-model outputs, and question-level uncertainty estimates to provide a more directly auditable analysis of candidate generation and selection.

A.1 Controlled Welding Study

A.1.1 Task setup

The controlled study contains 139 questions derived from a laser–arc hybrid welding setting. Questions span five task families that require different mappings from physical state to an answer. The evaluation includes numerical prediction, threshold or risk judgments, comparisons between physical conditions, directional changes under controlled parameter variation, and constrained decisions.

The language model receives the natural-language question and the information specified by the corresponding inference condition. Predictive-state quantities used for grounding are produced independently of language-model candidate generation. This separation allows the study to distinguish the model's proposal distribution from the external signal used to select or revise a response.

The aggregate accuracies reported in the main paper are reproduced in Table 8.

Table 8: Aggregate accuracy in the controlled welding study.

Method	Accuracy
Direct	0.324
Chain-of-thought	0.360
Self-consistency ($N = 8$)	0.367
Grounded selection	0.755
Evidence-conditioned revision	0.849
Direct model decision	0.935

These aggregate values summarize the controlled experiment. The independent CCPP study provides the exact matched-candidate comparison used in the main paper to separate candidate generation from candidate selection.

A.1.2 Predictive-signal conditions

The controlled study varies the source of the external state estimate while keeping the downstream grounding procedure fixed. The four conditions and their aggregate grounded-selection accuracies are shown in Table 9.

Table 9: Grounded-selection accuracy under different external predictive signals in the controlled study.

External signal	Accuracy
No external signal	0.367
Interpolation	0.619
Learned surrogate	0.755
Simulation	0.799

The learned condition uses a previously trained neural surrogate of the physical system. The simulation condition uses the corresponding simulation state directly. The comparison tests whether the informativeness of the external state estimate affects candidate selection.

The four conditions are categorical sources of predictive information. Their ordering is not a quantitatively linear fidelity scale, and the observed increments do not constitute a general scaling law.

A.1.3 CANDIDATE SCORING

The controlled implementation parses a typed final answer $\hat{y}_i$ and, when present, explicit scalar claims about task-relevant physical quantities. For scalar values u and v, relative error is defined as

$$\mathrm{re}(u, v) = \frac{|u - v|}{\max(|u|, |v|, 1)}. \tag{6}$$

Scalar agreement is then

$$\kappa(u, v) = \exp[-2\,\mathrm{re}(u, v)]\,. \tag{7}$$

For a candidate containing one or more explicit scalar claims, physical-state agreement is computed from κ between the parsed claims and the corresponding quantities predicted by the external model. Candidates containing no explicit scalar claim receive the neutral physical-agreement value

$$s_p(c_i, z) = 0.5. \tag{8}$$

A separate task-level term evaluates whether the candidate's final answer is consistent with the answer induced by the predicted state. For discrete tasks, this term uses exact agreement. For numerical tasks, it uses the same agreement kernel.

The final candidate score is

$$S(c_i, z) = s_p(c_i, z) + 1.5\, s_t(c_i, z), \tag{9}$$

where s_p denotes agreement between explicit physical claims and predicted state, and s_t denotes agreement between the final task answer and the decision induced by that state.

Grounded selection returns

$$c^* = \arg\max_{c_i \in \mathcal{C}_N} S(c_i, z). \tag{10}$$

Parse failures receive score -1. These rules correspond to the reported controlled-study results and were not modified based on the later CCPP matched-change diagnostic.

A.1.4 LANGUAGE-MODEL ROBUSTNESS

The controlled study includes GLM-4-flash, GPT-4o-mini, GPT-4.1-mini, and GPT-4o. These experiments examine whether the qualitative benefit of predictive-state evidence is tied to one language-model generator.

The cross-model experiment and the CCPP experiment address different sources of uncertainty. The controlled study supplies generator diversity, task heterogeneity, and variation in predictive-state quality. The CCPP study provides a frozen public-data setting in which candidate pools and selection outcomes can be inspected directly.

B INDEPENDENT CCPP DIAGNOSTIC STUDY

B.1 DATASET AND PREPROCESSING

The independent study uses the Combined Cycle Power Plant dataset. Each observation contains four operating variables:

- ambient temperature (AT),

- exhaust vacuum (V),
- ambient pressure (AP), and
- relative humidity (RH),

with net electrical energy output (PE) as the prediction target.

After removal of exact duplicate observations, the cleaned dataset contains

$$9{,}527 \tag{11}$$

observations. The primary split contains

$$7{,}622 \tag{12}$$

training observations and

$$1{,}905 \tag{13}$$

held-out observations.

The primary predictor split is an 80/20 random split with seed 42. PE values from held-out observations supply the ground truth for downstream question evaluation. Predictions from the external model are not used to construct correctness labels.

B.2 Predictive model

The CCPP predictive model is a Random Forest regressor with 300 trees. It is trained on the training portion of the primary split and evaluated on the held-out observations.

Table 10 reports its held-out performance.

Table 10: Held-out predictive performance of the CCPP Random Forest under the primary split.

Metric	Value
R^2	0.9599
MAE	2.412 MW
RMSE	3.434 MW
Spearman ρ	0.9774
Pairwise ordering accuracy	0.9643

The predictor is used as an independent source of estimated physical state. Its predictive accuracy is therefore an experimental property of the grounding signal, not a contribution by itself.

B.3 Frozen question construction

The CCPP evaluation contains 160 questions, with 32 questions in each of five task families. The question set, inference protocol, parsing rules, and evaluation rules were frozen before benchmark language-model calls.

The five task families probe different mappings from operating state to an answer:

1. numerical prediction of PE;
2. threshold-based decisions;
3. pairwise comparisons;
4. matched-change directional comparisons; and
5. constrained selection among alternatives.

The task set is used as a diagnostic evaluation. Its low-dimensional structure makes candidate generation, candidate availability, predicted state, and final selection directly inspectable.

B.4 Correctness and construction rules

The task-specific correctness and construction rules were fixed before language-model inference.

Numerical prediction. A numerical answer is counted as correct when its predicted PE lies within 2% of the held-out observed PE.

Threshold decisions. Threshold questions use thresholds from

$$\{440, 450, 460, 470\}\ \mathrm{MW}. \tag{14}$$

Examples lying within a 2 MW exclusion region around the selected threshold are not used for question construction. This reduces ambiguity near the decision boundary.

Pairwise comparisons. Eligible pairs require an observed PE difference of at least 5 MW. The correct answer is determined from the held-out observed measurements.

Matched-change directional comparisons. Matched pairs vary one input by at least 15% of its observed range while the remaining inputs stay within 2% of their respective ranges. Eligible pairs also satisfy

$$|\Delta PE| \geq 3\ \mathrm{MW}. \tag{15}$$

We refer to these questions as *matched-change* comparisons. The pairs are constructed from observational measurements and are not interpreted as causal interventions.

Constrained selection. Constrained-selection questions use alternatives whose relevant PE values satisfy a 2 MW runner-up margin. The margin reduces ambiguity between the best and second-best admissible choices.

B.5 Inference conditions

Let q denote a question and

$$\mathcal{C}_N = \{c_1, \ldots, c_N\} \tag{16}$$

the set of N sampled candidate responses.

Direct. The language model answers the question once without external predictive-state evidence.

Chain-of-thought. The language model produces one response using the corresponding reasoning instruction.

Self-consistency. The model generates N sampled responses. The frozen self-consistency aggregation rule selects the final answer without access to predicted physical state.

Grounded selection. Grounded selection uses exactly the same sampled candidate pool as self-consistency. No additional language-model response is generated. The external predictor supplies the state estimate used by the frozen grounding score to rank the existing candidates.

For every question and every evaluated N,

$$\mathcal{C}_N^{\mathrm{SC}} = \mathcal{C}_N^{\mathrm{grounded}}. \tag{17}$$

The matched-pool design isolates the effect of candidate selection from the effect of candidate generation.

Evidence-conditioned revision. The language model first produces a draft without access to the predicted state. A second call receives the draft together with the predicted state and produces one revised answer. No iterative revision loop is used.

Direct model decision. The language model is omitted from the final state-to-decision mapping. The specified task rule is applied directly to the predicted state. This condition serves as a control for tasks in which the predicted quantity and task rule already determine the answer.

B.6 AGGREGATE RESULTS

Table 11 reproduces the aggregate CCPP results.

Table 11: Aggregate accuracy in the independent CCPP diagnostic study.

Method	Accuracy
Direct	0.544
Chain-of-thought	0.594
Self-consistency ($N = 8$)	0.638
Grounded selection ($N = 8$)	0.813
Evidence-conditioned revision	0.956
Direct model decision	0.950

Grounded selection improves over self-consistency by

$$81.25 - 63.75 = 17.5 \tag{18}$$

percentage points while using exactly the same eight sampled candidates.

Evidence-conditioned revision reaches 95.6%, while direct model decision reaches 95.0%. Their similar performance is specific to these structured tasks. When predicted state and a specified task rule already determine the answer, an additional language-model transformation has little opportunity to improve the final mapping in this evaluation.

B.7 CANDIDATE AVAILABILITY AND SAMPLING

For candidate pool $\mathcal{C}_N$, define candidate coverage as

$$A_{\text{pool}}(N) = \frac{1}{M} \sum_{j=1}^{M} \mathbb{I}\Big[\exists c \in \mathcal{C}_N^{(j)} \text{ that is correct}\Big], \tag{19}$$

where M is the number of questions.

Candidate coverage measures proposal availability. It does not use the grounding score. It is also not an upper bound for methods such as evidence-conditioned revision, which can produce a new answer after receiving predictive-state evidence.

The observed trajectories are reported in Table 12.

At $N = 8$, the sampled pool contains at least one correct response for 91.3% of questions. Self-consistency reaches 63.8%, while grounded selection reaches 81.3%.

The difference between candidate coverage and self-consistency is

$$91.25 - 63.75 = 27.5 \tag{20}$$

Table 12: Sampling trajectories in the CCPP study. Grounded selection and self-consistency use identical candidate pools at each N.

	$N = 1$	$N = 2$	$N = 4$	$N = 8$
Self-consistency	0.594	0.619	0.663	0.638
Grounded selection	0.594	0.650	0.781	0.813
Candidate coverage	0.594	0.681	0.831	0.913

percentage points. Grounded selection recovers

$$81.25 - 63.75 = 17.5 \tag{21}$$

points of that difference. The corresponding fraction is

$$\frac{17.5}{27.5} = 0.636. \tag{22}$$

The 63.6% value is a descriptive decomposition of the observed $N = 8$ result, not a general efficiency coefficient for selection algorithms.

B.8 Inference budget and token accounting

Table 13 reports language-model calls and cached token usage in the CCPP experiment. Token counts come from saved API usage fields.

Table 13: Inference accounting for the CCPP experiment. Input and output token counts are means per question. Grounded selection reuses the exact self-consistency candidate pool and therefore has identical LLM token usage.

Method	Acc.	LLM calls/q	Input tok./q	Output tok./q	Predictor calls/q
Direct	0.544	1	254.2	26.8	0
Chain-of-thought	0.594	1	269.6	202.4	0
Self-consistency ($N = 8$)	0.638	8	2162.2	1752.1	0
Grounded selection ($N = 8$)	0.813	8	2162.2	1752.1	1
Evidence-conditioned revision	0.956	2	897.9	346.2	1
Direct model decision	0.950	0	0.0	0.0	1

The self-consistency and grounded-selection conditions are matched in language-model generation. Their 17.5-point accuracy difference is obtained without additional sampled LLM responses; grounded selection adds only the predictive-model inference used for candidate ranking.

Evidence-conditioned revision uses two language-model calls per question, compared with eight sampled calls for self-consistency. We report calls and token usage; monetary API cost is not needed for the methodological comparison.

The accounting is included to rule out additional language-model generation as an explanation for the main CCPP ordering.

B.9 Predictor robustness to split construction

The primary CCPP predictor uses a random train/test split. The dataset does not provide timestamps, and exact duplicate removal alone does not establish separation between nearby operating regimes.

We therefore perform a predictor-only robustness analysis using a pre-specified cluster-separated split, without modifying the frozen primary benchmark.

The procedure is:

1. standardize the four predictor inputs;
2. apply K-means clustering with $k = 20$ and seed 42 in standardized input space;
3. assign entire clusters to training or test to obtain a test set of approximately 20%;
4. retrain the same 300-tree Random Forest without hyperparameter tuning; and
5. evaluate the same predictive metrics.

The resulting split contains 7,545 training observations and 1,982 test observations.

Table 14: Predictor robustness to split construction. The cluster-separated split assigns complete clusters in standardized input space to train or test.

Split	R^2	MAE	RMSE	Spearman ρ	Pairwise ordering
Primary random	0.9599	2.412	3.434	0.9774	0.9643
Cluster-separated	0.9514	3.122	4.068	0.9302	0.9217

Numerical error increases under the cluster-separated split. The predictor nevertheless retains an R^2 of 0.9514, Spearman correlation of 0.9302, and pairwise ordering accuracy of 0.9217.

This analysis evaluates predictor robustness only: the frozen LLM benchmark is not rerun on the alternative split, and the alternative split does not replace the primary evaluation.

C Matched-Change Diagnostic

The matched-change task provides a negative result for the frozen grounding score.

On this task,

$$A_{\mathrm{SC}} = 0.500 \tag{23}$$

and

$$A_{\mathrm{grounded}} = 0.500. \tag{24}$$

The frozen grounding score therefore provides no improvement over self-consistency for these directional change judgments.

Evidence-conditioned revision reaches

$$A_{\mathrm{revision}} = 0.875, \tag{25}$$

while direct model decision reaches

$$A_{\mathrm{direct}} = 0.844. \tag{26}$$

The frozen candidate score emphasizes magnitude agreement, whereas this task asks for the direction of a change. Across candidates, the correlation between the frozen magnitude-based score and correctness is

$$r = -0.048. \tag{27}$$

After observing this failure, we compute a diagnostic that asks whether the direction stated by a candidate agrees with the direction implied by the predicted state. Direction agreement has correlation

$$r = 0.722 \tag{28}$$

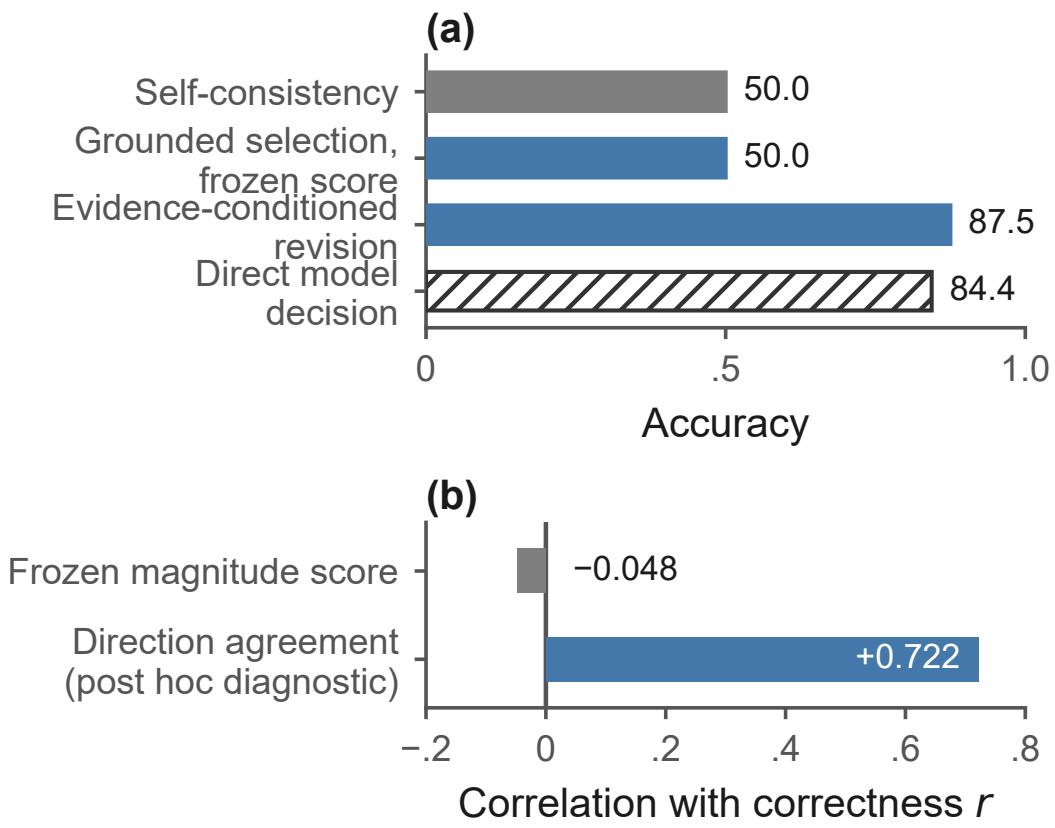


Figure 6: Matched-change diagnostic on CCPP. (a) The frozen magnitude-based grounding score provides no gain over self-consistency on directional comparisons, while evidence-conditioned revision and direct model decision remain substantially more accurate. (b) The frozen score is essentially uncorrelated with candidate correctness, whereas post-hoc direction agreement is strongly associated with correctness.

with candidate correctness.

This direction-agreement quantity is post hoc. It is not used to rescore the primary experiment, modify the reported grounded-selection result, or alter the frozen benchmark. It is reported only to diagnose why an informative predicted state can coexist with an ineffective candidate-scoring rule.

D Task-Aligned Selection Diagnostic

This appendix reports a post-hoc re-scoring of the frozen candidate pools with a task-aligned grounding score. The frozen benchmark and all main-text results are unchanged; this analysis isolates how much of the selection gap is attributable to score–task alignment rather than to the candidate pool or the predictor.

Score. For each task family, the aligned score makes agreement on the task's *decision variable* the primary term and uses magnitude agreement only as a tie-break: value agreement for numerical prediction, threshold-sign agreement for threshold judgments, ordering agreement for pairwise comparisons, direction-of-change agreement for matched change, and choice agreement (the verifier's own arg-min distance to the target) for constrained selection. No parameter is tuned on test accuracy, and the candidate pools are exactly those used by self-consistency and the frozen grounded-selection baseline.

Result. Re-selecting on the frozen pools with the aligned score lifts overall grounded-selection accuracy from 81.3% to 88.1%. The question-level paired bootstrap 95% confidence interval for the improvement over a magnitude-only ablation of the same score is $[+11.3, +23.8]$ points. The gain concentrates where the frozen score was misaligned with the decision variable (counterfactual $+21.9$ points, pairwise comparison $+6.3$, constrained selection $+6.2$), while the numerical and threshold families are unchanged. On matched-change questions the aligned score raises selection from 50.0% to 71.9%; the residual gap is a predictor-fidelity ceiling rather than a scoring failure, because the verifier's own direction accuracy on these pairs is 84.4% and the median matched change in PE (3.65 MW) is comparable to the predictor's mean absolute error (2.4 MW).

Predictor-fidelity sensitivity. Replacing the random-forest verifier ($R^2 = 0.960$) with the gradient-boosting verifier ($R^2 = 0.958$) and the weaker linear verifier ($R^2 = 0.928$) reduces the aligned-selection gain over self-consistency from $+24.4$ to $+22.5$ to $+16.9$ points, respectively. The grounding gain therefore strengthens with predictor fidelity on this public system, consistent with the fidelity ordering observed in the controlled study.

Mechanism. On constrained-selection questions, 27% of incorrect candidates chose the configuration with the highest predicted output, not the one closest to the 440 MW target—a task-misreading error that the verifier's explicit decision variable corrects.

E Error Analysis

Aggregate accuracy does not distinguish failures in candidate generation from failures after a useful candidate has already been generated. We therefore inspect the errors of the frozen CCPP evaluation.

Self-consistency answers 58 questions incorrectly. Of these, 45 are aggregation failures: the sampled candidate pool contains a correct response, but the aggregation rule does not return it. The remaining 13 are proposal failures for which the sampled pool contains no correct candidate.

For grounded selection, 30 questions are incorrect. The observed breakdown is

- 14 proposal failures,
- 13 ranking failures, and
- 3 predictor-associated failures.

For evidence-conditioned revision, seven questions remain incorrect:

- 4 cases in which the revision does not use the supplied evidence correctly, and
- 3 cases associated with predictor error.

These categories are descriptive diagnostics of the CCPP evaluation, not an exhaustive taxonomy of scientific reasoning errors.

The self-consistency breakdown is consistent with the candidate-coverage analysis. Most of its errors occur on questions for which the sampled pool already contains a correct response. Grounded selection removes many of these aggregation errors, while proposal and ranking errors remain.

F Statistical Evaluation

The main CCPP results report question-level bootstrap confidence intervals. The resampling unit is the question, not the individual candidate generation, so that uncertainty reflects variation across evaluation problems.

For question-level correctness indicators

$$y_1, \ldots, y_M, \qquad y_j \in \{0, 1\}, \tag{29}$$

accuracy is

$$\hat{A} = \frac{1}{M} \sum_{j=1}^{M} y_j. \tag{30}$$

Bootstrap samples are formed by resampling questions with replacement and recomputing accuracy from the frozen evaluation outputs.

Self-consistency and grounded selection use matched candidate pools for each question. Their difference is therefore evaluated at the question level as a paired comparison.

G Prompting and Parsing Protocol

The inference conditions differ in the information available to the language model and in the stage at which predicted physical state enters the pipeline.

Direct. The language model receives the task question and produces one answer without external predictive-state evidence.

Chain-of-thought. The model receives the corresponding reasoning instruction and produces one response. Its final answer is evaluated using the same task-specific correctness rule as the direct condition.

Self-consistency. The model produces N stochastic candidate responses. The frozen self-consistency rule aggregates these candidates without access to predicted physical state.

Grounded selection. The same N candidates used by self-consistency are reused without additional language-model generation. Predicted state enters only after candidate generation through the frozen candidate-scoring rule.

Evidence-conditioned revision. The model first produces a draft without seeing predicted state. A second call receives the draft together with the predicted state and produces one revised answer. No iterative revision loop is used.

Direct model decision. No language-model call is used for the final decision. The task-specific rule is applied directly to predicted state.

Parsing. Evaluation uses deterministic task-specific parsing. Parse failures are handled by the frozen evaluation rules and are not manually corrected after inspection.

H Benchmark Integrity

The independent study uses a frozen evaluation pipeline containing data preprocessing, question construction, model prompts, cached model generations, parsing, predictive-model training, scoring, and evaluation.

The benchmark integrity checks verify that:

1. required question fields are present;
2. source observations for evaluation questions come from the held-out split;
3. evaluation rows do not overlap predictor training rows;
4. question identifiers are unique;
5. question texts are unique;
6. ground-truth answers can be recomputed from held-out measurements;
7. PE target values are not exposed in question inputs;

8. units are consistent; and
9. answer spaces satisfy the task-specific construction rules.

All nine checks pass for the frozen 160-question evaluation.

The benchmark contains 32 questions from each of the five task families. The question set and inference protocol were frozen before benchmark language-model calls. Cached model outputs are retained so that parsing, scoring, and analysis can be repeated without regenerating responses.

Grounded selection reuses the self-consistency candidate pool; no second pool is generated. This preserves the exact matched-pool comparison used in the main analysis.

I Evidence Placement

The three predictive-state interfaces differ in where the same class of external evidence enters inference.

For grounded selection,

$$q \longrightarrow \mathcal{C}_N \longrightarrow S(c_i, z) \longrightarrow c^*, \tag{31}$$

and the predicted state z is hidden from the language model during candidate generation.

For evidence-conditioned revision,

$$q \longrightarrow c_0, \qquad (c_0, z) \longrightarrow c_{\text{revised}}, \tag{32}$$

so the second language-model response can condition directly on predicted state.

For direct model decision,

$$z \longrightarrow \rho_t(z) \longrightarrow \hat{y}, \tag{33}$$

where ρ_t is the specified state-to-decision rule for task t.

These are experimental interfaces, not separately branded algorithms. The comparison distinguishes post-generation selection, evidence-conditioned language inference, and deterministic state-to-decision mapping.

J Scope of the Empirical Findings

The two studies provide complementary evidence.

The controlled welding study evaluates multiple task types, several language models, and external predictive signals of different quality. It shows that the effectiveness of grounded selection changes with the information supplied by the external state estimate.

The CCPP study is lower-dimensional, but its structure permits direct inspection of candidate availability, predicted state, selection, inference budget, and failure cases. Its primary role is diagnostic: it separates candidate generation from candidate selection and provides a case in which the frozen grounding score does not evaluate the variable required by the task.

The empirical results do not imply that predictive-state grounding improves every scientific reasoning task. Several conditions affect whether external state estimates are useful:

- post-generation selection cannot recover a correct answer that is absent from the sampled candidate pool;
- errors in the predictive model can propagate into selection or revision;
- an informative predicted state does not guarantee effective selection when the candidate score evaluates a property that does not match the task decision; and

- the CCPP questions are structured state-to-decision tasks, not open-ended scientific discovery problems.

The direct-model condition also has a narrow interpretation. When predicted state and a specified deterministic task rule already determine the answer, the final state-to-decision mapping may not benefit from another language-model transformation.

K Summary of the Independent Diagnostic

The CCPP results can be summarized through three measured quantities at $N = 8$:

$$A_{\mathrm{SC}} = 0.638, \qquad A_{\mathrm{grounded}} = 0.813, \qquad A_{\mathrm{pool}} = 0.913. \tag{34}$$

The self-consistency and grounded-selection values are obtained from the same sampled candidate pools. The difference between them therefore reflects the selection rule alone.

Candidate coverage shows that the sampled pool contains useful responses more often than self-consistency returns them. Grounded selection closes part of this difference using independently predicted physical state.

The controlled fidelity experiment and the matched-change diagnostic qualify this observation. Selection depends on the quality of the predicted state and on whether the grounding score evaluates information relevant to the task. The results therefore support a conditional empirical account of predictive-state grounding rather than a claim that one selection rule is uniformly preferable across scientific reasoning problems.