FROM INFORMATION VALUE TO INTERVENTION VALUE IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Additional information can improve language model performance, but measuring that potential does not tell us what system change will realize it. We study the gap between the value of information and the value of interventions. We separate four questions: whether oracle information improves task utility (*opportunity*), whether the channel used to supply it preserves that gain (*instrument validity*), whether the remaining shortfall can be reliably assigned to information acquisition or use (*attribution*), and whether acting on that assignment improves utility on held out data (*decision validity*). Across eight benchmarks, oracle information value is the most robust quantity in our audits, while attribution can change under subsampling, reference construction, and reader family. On BigCodeBench (Zhuo et al., 2025), the injection channel itself destroys the oracle gain, making recovery measurements uninterpretable. Our prospective experiment is a sufficiency test, not a prevalence study. It asks whether stable shortfall attribution is sufficient evidence for selecting the corresponding intervention. On FEVER and QASPER, with 56 diagnostic items per system, attributions are stable under subsampling and threshold perturbations, yet the mapped interventions reduce held out utility relative to doing nothing ($\tau = -0.078$ and -0.048). FEVER provides a registration consistent counterexample. QASPER shows the same pattern under the documented menu deviation. These cases do not estimate how often attribution and intervention value diverge.

1 INTRODUCTION

Supplying language models with additional information is now a common way to improve their performance. Retrieval systems provide documents, tool using models obtain external results, grounding methods expose structured evidence, and hindsight methods turn later outcomes into supervision (Schick et al., 2023; Qin et al., 2024; Andrychowicz et al., 2017; Madaan et al., 2023). When these systems fall short, however, the final task score says little about why. In retrieval augmented question answering, poor accuracy may mean that useful evidence would not help the model, that the retriever failed to find it, or that the model failed to use the evidence once it was available (Liu et al., 2024). The same ambiguity appears whenever additional information is expected to improve an outcome.

A natural way to separate these possibilities is to compare the system with an oracle. If supplying oracle information improves task utility, then useful information exists for the model. If retrieved information captures only part of that gain, acquisition appears to leave value unrealized. If useful information is available but the model still fails to realize its benefit, the shortfall instead appears to lie in how that information is used. Such decompositions turn a single task failure into more specific hypotheses about where information value is lost.

But identifying where value is lost is not the same as knowing what to change. A diagnostic may attribute the largest shortfall to retrieval, yet a realistic retrieval improvement may add little utility or even make the system worse. Another intervention, such as changing the reader, may have a larger effect. The diagnostic describes the gap between the observed system and an oracle. An engineering decision depends on the effect of a concrete action. These are different quantities. Our central question is whether the first tells us enough about the second. We ask whether a stable shortfall attribution, by itself, is sufficient evidence for selecting the corresponding intervention.

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.

We organize this question into four validity claims. *Opportunity* asks whether a specified oracle source improves utility for a fixed model and channel. *Instrument validity* asks whether the channel used to supply that information preserves the oracle benefit. *Attribution* asks whether the remaining shortfall can be assigned reliably to information acquisition or use. Finally, *decision validity* asks whether acting on that attribution improves utility on untouched data. These claims form a procedural sequence rather than a chain of logical implications. Establishing opportunity does not establish attribution, and a stable attribution does not establish that the corresponding intervention has value. Section 2 defines an estimand for each claim.

We first examine the first three claims across eight benchmarks spanning six task settings. Opportunity is the most robust quantity in our audits. Oracle information reveals substantial headroom where additional information is useful and near zero headroom on the calibration cases. Attribution is less robust: the inferred shortfall can change under subsampling, reference construction, or reader family. BigCodeBench exposes a different boundary. There, the channel used to measure recovery destroys the oracle benefit itself, so the resulting recovery measurements are not interpretable.

We test this prospectively on FEVER and QASPER. We run the diagnostic on 56 items, freeze the mapped action, and only then evaluate candidate actions on untouched test data. The mapped interventions reduce utility relative to doing nothing ($\tau = -0.078$ and -0.048). FEVER follows the registered protocol throughout. QASPER shows the same pattern under the documented menu deviation. These two systems provide counterexamples to sufficiency, not an estimate of how often the failure occurs.

Our contributions are:

1. **A framework from information value to intervention value.** We separate opportunity, instrument validity, attribution, and decision validity, and define measurable quantities for each stage. The framework distinguishes whether useful information exists, whether it can be measured through the chosen channel, where its value appears to be lost, and whether acting on the attribution improves utility.
2. **An audit across eight benchmarks.** We test information opportunity, realization, and attribution stability, including a case where the recovery channel itself invalidates recovery analysis.
3. **A preregistered prospective decision test.** On FEVER and QASPER, stable shortfall attributions are insufficient to license the corresponding targeted interventions: both mapped actions reduce utility relative to `NOOP` on held out data.

Experiments use Qwen2.5-3B-Instruct (Yang et al., 2024) as the anchor model, with Qwen3 (Yang et al., 2025) scale sweeps through 14B where reported. The first three claims are studied across eight benchmarks, while decision validity is evaluated prospectively on FEVER and QASPER and supported by held out analyses on four additional systems.

2 From information value to intervention value

The introduction separates two questions: where information has value, and which intervention has value. We now make both questions measurable. Fix a language model, a task, a candidate information source, and a channel through which that information is supplied. For item i, let $m_i(Y) \in [0, 1]$ denote the task utility of model output Y. When information g is supplied, we define

$$u_i(g) = \mathbb{E}[m_i(Y) \mid i, g].$$

Binary correctness is a common special case, but the definition also covers execution accuracy, judge accuracy, verdict accuracy, and other task specific utilities used in our experiments. On a diagnostic sample of n items, we construct a specified oracle source g^* and compare it with the same system without that information. All quantities below are conditional on the model, the choice of g^*, and the channel through which information is supplied.

Table 1 summarizes four validity claims. The first three describe how the system responds to information. The fourth changes the object of measurement from information to actions. The claims are ordered as a measurement procedure rather than as logical implications: evidence for one claim does not by itself establish the next. Appendix A gives the full definitions, edge cases for ratios, and the complete diagnostic screen.

Table 1: Four validity claims from information value to intervention value. Each claim has its own estimand. The final column states what a failed check means for the measurement or decision that follows. Passing one claim does not establish the next.

Claim	Question	Estimand	If the claim fails
Opportunity	Does oracle information improve utility?	Δ, Scope	No detectable headroom for this source
Instrument validity	Does the delivery channel preserve oracle benefit?	Δ_{instr}	Recovery cannot be interpreted through this channel
Attribution	Where is oracle value left unrealized, and is that reading reliable?	R_{ret}, R_{self}, alignment, stability	The shortfall cannot be assigned reliably
Decision validity	Does acting on the attribution improve held out utility?	τ_a, R_{raw}, R_{norm}, coverage	The mapped action is not supported by held out utility

Opportunity measures the value of available information. For each item, the oracle lift is

$$L_i = u_i(g^*) - u_i(\emptyset),$$

and the *Information Gap* is

$$\Delta = \mathbb{E}_i[L_i].$$

Δ measures the average utility available from the specified oracle source, while *Scope* describes how that opportunity is distributed across items. A broad Scope means many items benefit. A concentrated Scope means that most of the gain comes from a smaller subset.

Opportunity gate. The registered protocol uses the oracle probe to decide whether recovery analysis should proceed. If the specified oracle source shows no detectable headroom, analysis stops at near zero opportunity. Otherwise the protocol reports Scope as broad or concentrated and proceeds to instrument validation and shortfall attribution. Registered labels `GO` and `STOP` denote these branches in the supplement. The main estimands are Δ and Scope.

Instrument validity asks whether recovery can be measured. A positive Information Gap does not guarantee that every way of supplying information preserves that gain. Before interpreting retrieval or self recovery, we therefore test the channel used for those measurements. Let $\bar{u}_\emptyset$ be mean utility without supplied information and $\bar{u}_{\text{gold}-\text{forced}}$ mean utility when oracle information is inserted through the recovery channel. We define

$$\Delta_{\text{instr}} = \bar{u}_{\text{gold}-\text{forced}} - \bar{u}_\emptyset.$$

If $\bar{u}_{\text{gold}-\text{forced}} < \bar{u}_\emptyset$, the channel itself reduces utility. We mark the instrument invalid and do not interpret recovery measurements through that channel. The same channel destruction check is applied before interpreting self recovery on every benchmark. BigCodeBench is the setting in which gold forced utility falls below the no information arm.

Attribution asks where oracle value remains unrealized. Once the channel is valid, we compare realized information with the oracle. For retrieval,

$$R_{\text{ret}} = \frac{\bar{u}_{\text{retrieved}} - \bar{u}_\emptyset}{\Delta},$$

which measures the fraction of oracle opportunity recovered by retrieved information. For self recovery,

$$R_{\text{self}} = \frac{\bar{u}_{\text{self}-\text{forced}} - \bar{u}_\emptyset}{\bar{u}_{\text{gold}-\text{forced}} - \bar{u}_\emptyset},$$

which measures how much of the utility available from gold information is recovered by information produced by the model itself. We also measure *quality–utility alignment*, the association between a recovery quality proxy and task utility, and test whether the resulting shortfall assignment survives changes in the diagnostic sample, reference construction, and reader family.

These quantities support a descriptive attribution. They can indicate that oracle value appears to be lost in acquiring useful information, in producing or using it, or that no dominant shortfall can be assigned. They do not measure the effect of changing any component. This distinction is the point at which information value ends and intervention value begins.

Decision validity measures the value of concrete actions. For each system s, we declare an action menu A_s before evaluating decision outcomes. The menu contains `NOOP`, the intervention corresponding to the diagnosed shortfall, a reader upgrade, and additional sampling. The menu defines the decision problem: "best action" always means best among the actions declared for that system.

Let U_a denote mean utility on held out items when action a is applied. The intervention value of a is

$$\tau_a = U_a - U_{\mathrm{NOOP}},$$

with $\tau_{\mathrm{NOOP}} = 0$. Thus Δ measures the value available from information, whereas τ_a measures the value produced by a concrete intervention.

Let

$$a^* = \arg\max_{a \in A_s} \tau_a$$

be the best declared action, and let a_{diag} be the action selected by the frozen diagnostic mapping. Our primary decision quantity is *operational regret*,

$$R_{\mathrm{raw}} = \tau_{a^*} - \tau_{a_{\mathrm{diag}}}, \tag{1}$$

the utility forgone by following the diagnostic rather than the best declared action. We interpret R_{raw} within a system only, since task utilities differ across domains and their magnitudes are not directly comparable.

We also report normalized regret,

$$R_{\mathrm{norm}} = R_{\mathrm{raw}} / \tau_{a^*},$$

when the best declared action has a detectable positive effect. It is undefined otherwise. On this scale, $R_{\mathrm{norm}} = 0$ means that the mapping selected the best declared action, $R_{\mathrm{norm}} = 1$ means that it did no better than `NOOP`, and $R_{\mathrm{norm}} > 1$ means that the selected action reduced utility while a better declared action had positive value. STOP, ABSTAIN, and WARN fall back to the registered baseline when a menu is open, while `INVALID` prevents the decision estimand from being defined. We compare the frozen mapping with fixed policies on the same items and report decision stability separately from regret. Full definitions, coverage measures, equivalence rules, and stability axes are in Appendix A.

3 Information value is more robust than shortfall attribution

Section 2 separates the value available from information from the value produced by an intervention. Before asking which intervention has value, we first examine the measurements on the information side. Across eight benchmarks, three patterns emerge. First, oracle opportunity is the most robust quantity in our audits. Second, retrieved and self generated information often realizes only a fraction of oracle value, and the channel used to measure recovery can itself be invalid. Third, assigning the remaining shortfall to acquisition or use is less robust than measuring whether opportunity exists. Detailed benchmark results, ablations, scale sweeps, and the BIRD analysis are in Appendix B.

Table 2: Information opportunity and realized gain at the Qwen2.5-3B anchor (n=300 per task, BIRD SQL n=1534). Δ measures oracle opportunity. R_{ret} and R_{self} measure how much of that opportunity is realized by retrieved and self generated information. Dashes indicate that an arm was not evaluated: the calibration tasks do not open recovery arms, and BIRD has no retrieval arm. "n/i" marks a recovery reading invalidated by the channel check. Task utilities differ across benchmarks, so magnitudes are not comparable across rows.

Task	Headroom Δ	R_{ret}	R_{self}	Recovery instrument
GSM8K	-0.022	—	—	—
MBPP	$+0.010$	—	—	—
BIRD SQL	$+0.118$	—	≈ 0	valid
HotpotQA	$+0.43$	0.48	≈ 0	valid
2WikiMultiHopQA	$+0.24$	—	≈ 0	valid
MuSiQue	$+0.44$	—	≈ 0	valid
BrowseComp-Plus	$+0.54$	0.09	≈ 0	valid
BigCodeBench	$+0.070$	—	n/i	**rejected**

Finding 1: oracle information separates substantial opportunity from calibration near zero. Supplying g^* produces near zero gain on the two calibration cases and clear positive gains elsewhere. Oracle intent information does not improve GSM8K over the baseline, and the small positive lift on MBPP is not statistically detectable. These cases therefore register little opportunity rather than a failure of the measurement pipeline. On BIRD, the multi hop QA tasks, BrowseComp-Plus, and BigCodeBench, supplying oracle information instead produces positive task utility gains (see Table 2).

The opportunity reading also survives resampling. On HotpotQA, BrowseComp-Plus, and BIRD, positive opportunity together with the broad versus concentrated Scope label reproduce with probability 1.00 at n=100, above their registered thresholds (see Appendix B.5). BrowseComp-Plus utility is defined by the registered semantic judge, under which the no context arm scores zero. We interpret Δ only within that judge based utility and do not compare its magnitude with other tasks. Within the domains we audited, whether the specified oracle source carries value is therefore more reproducible than where that value is lost.

Finding 2: available information value can be much larger than realized value. A positive Δ says that useful information exists for the model. It does not say that the deployed system can obtain or reproduce that information. HotpotQA illustrates the acquisition gap: accuracy rises from 0.18 without evidence to 0.38 with retrieved evidence, but reaches 0.61 with gold evidence. BrowseComp-Plus shows the same pattern more sharply, with retrieval realizing only a small fraction of the available oracle gain (see Table 2).

Self-generated information realizes even less. On BIRD, utility with self-recovered grounding is 0.162, close to free generation at 0.163, while gold grounding reaches 0.271. This gives $R_{\text{self}} \approx 0$ even though the same model benefits when gold grounding is supplied. The association between grounding quality and execution utility is also weak, with $\text{corr}(\text{F1}, \text{EX}) \approx 0.03$. The same near zero self recovery pattern appears on all three multi hop QA benchmarks and on BrowseComp-Plus. Increasing model size does not remove it through 14B on BIRD and HotpotQA or through 8B on the remaining settings we tested. The scale experiments test whether the observed realization gap disappears within the 3B to 14B range. They are not evidence about frontier scale models. The conceptual distinction between information value and intervention value does not depend on this scale result.

Positive opportunity is not sufficient even to make recovery measurable. BigCodeBench (Zhuo et al., 2025) has positive oracle headroom, but inserting gold information through the recovery channel reduces accuracy below the baseline without supplied information: 0.103 versus 0.273 for Qwen2.5-3B. The same failure appears on the second model, and three channel redesigns at n=75 do not restore validity. We therefore mark the instrument `INVALID`, do not interpret recovery through this channel, and do not open a decision menu. Opportunity can therefore exist even when the channel needed to locate the shortfall cannot support that measurement.

Finding 3: locating the shortfall is less robust than detecting opportunity. Once opportunity and instrument validity are established, the diagnostic asks where the remaining value is lost. That attribution is more sensitive to the choices used to measure it. The clearest case is BIRD: at n=100, the opportunity gate and Scope reproduce with probability 1.00, while the dominant shortfall reproduces with probability only 0.52. Attribution is also sensitive to oracle construction: on BrowseComp-Plus, changing the reference moves the Information Gap from 0.53 to 0.05 and changes the dominant shortfall from acquisition to self recovery. Reader family and controlled calibration audits show the same pattern: opportunity is more robust than the resulting shortfall attribution. Detailed subsample, reference, and reader family stability results are reported in Appendix B.5.

What the information side establishes. These measurements can establish that useful information exists, measure how much of its value is realized, and form a hypothesis about where the remaining shortfall lies. They do not measure intervention value. Supporting analyses show that an analyst can use these readings to design useful interventions, even when attribution is unstable. For example, on held out BIRD items, deterministic schema and value linking realizes $H_a = 0.785$ $[0.620, 0.986]$ even though the final shortfall attribution is unstable (see Appendix B.6). Section 4 removes that analyst choice by freezing the mapping from attribution to action before outcomes are observed.

4 From shortfall attribution to intervention value

The information measurements in Section 3 do not identify the effect of changing the system. We therefore freeze a mapping from shortfall attribution to a registered intervention and evaluate that intervention on untouched data. The prospective test asks a narrow question: whether the attribution, without observing intervention outcomes, is sufficient to license the mapped action. It is not intended as a production policy, and failure does not imply that the attributed component is irrelevant.

4.1 A prospective decision test

For each system, we declare the action menu and mapping before observing decision outcomes, run the diagnostic on n=56 items, freeze the selected action, and then evaluate every declared action on an untouched `final_test` split. The diagnostic budget n=56 was registered in advance rather than chosen by retrospective power analysis. Sanity gates could inspect data loading, verification, and oracle correctness, but not intervention effects. Estimands, menus, and failure conditions were preregistered (Preregistration v2, anonymous supplement). FEVER (Thorne et al., 2018) follows the registered protocol. QASPER (Dasigi et al., 2021) has the documented menu deviation described in Section 4.2. BrowseComp-Plus, BIRD, HotpotQA, and SciFact provide supporting analyses only.

4.2 Stable shortfall attributions selected interventions that reduced utility

Both prospective diagnostics passed their sanity gates and produced shortfall attributions that mapped to targeted interventions. The selected action was unchanged under every subsample and threshold perturbation we tested. On untouched held out data, the attribution selected intervention reduced utility relative to `NOOP` in both systems: $\tau = -0.078$ verdict accuracy on FEVER and $\tau = -0.048$ QASPER score on QASPER (see Table 3). FEVER provides a registration consistent counterexample. QASPER shows the same pattern under the documented menu deviation. These cases do not estimate how often attribution and intervention value diverge.

On FEVER, the diagnostic attributes the shortfall to acquisition and freezes the acquisition intervention. That intervention performs full corpus retrieval and supplies the retrieved evidence to the reader. Its negative treatment effect is consistent with retrieved context degrading verdict prediction, but the experiment does not identify whether the loss comes from irrelevant, conflicting, or otherwise poorly matched evidence. On QASPER, the diagnostic attributes the shortfall to self recovery and freezes that intervention. Thus, in both systems, $\tau_{a_{\text{diag}}} < 0$: the mapped targeted action reduces utility relative to doing nothing.

Table 3: Prospective intervention value at the Qwen2.5-3B-Instruct anchor. Shortfall attribution is frozen on n=56 items. τ_a is the mean utility gain of action a over `NOOP` on the untouched `final_test` split. Results are reported separately because the study contains only two prospective task clusters. Registered gate labels appear in the supplement. Lower operational regret is better when comparing policies. The reader excluded comparison is a descriptive sensitivity analysis. Operational regret under the full declared menu remains the registered estimand.

	FEVER	QASPER
Frozen attribution (n=56)		
Opportunity reading	pos., concentrated	pos., broad
Information Gap Δ	0.459 [0.323, 0.600]	0.255 [0.154, 0.356]
Dominant shortfall $\rightarrow$ frozen action	acquisition	self recovery
Stability (subsample / threshold)	1.00 / 1.00	1.00 / 1.00
Untouched `final_test`		
τ (frozen targeted action)	$\mathbf{-0.078}$	$\mathbf{-0.048}$
τ (reader intervention)	$+0.324$	$+0.036$
τ (more samples)	-0.004	-0.020
Best menu action a^*	reader	reader
Operational regret R_{raw}	0.403	0.084
Sensitivity: regret with reader excluded	0.078	0.048
Fixed policies, same items (R_{raw})		
$\pi_{\text{always NOOP}}$ / π_{random}	0.324 / 0.264	0.036 / 0.044

The sign of the result does not depend on the cross family reader control. Excluding the reader from the menu leaves `NOOP` as the best remaining action, giving regret 0.078 on FEVER and 0.048 on QASPER. We report this only as a descriptive sensitivity analysis. Full menu regret remains the registered estimand.

The frozen mapping does not outperform policies that ignore it. On both systems, the frozen mapping has greater regret than always choosing `NOOP` and than uniform random choice from the declared menu. Full fixed-policy and outside-menu comparisons are in Appendix C.

Stability does not establish intervention value. Both prospective decisions have perfect subsample and threshold stability at n=56. Resampling the diagnostic data and perturbing the registered thresholds never changes the selected action. The attribution is therefore reproducible along both axes measured prospectively, yet the corresponding intervention reduces held out utility in both systems. Stability establishes reproducibility of the attribution, but intervention value still requires evidence of attribution to treatment alignment (see Appendix D).

For QASPER, the freeze commit table lists acquisition, while the executable menu, held out drivers, and powered runs use `self_recovery` as the targeted action. The executable path predates the table mismatch, and the intervention was not changed after observing held out outcomes. We therefore treat QASPER as a prospective diagnostic with a downstream menu deviation and FEVER as the cleanest prospective decision cell.

Supporting systems exhibit other outcomes. BrowseComp-Plus aligns with the best declared action, HotpotQA has little intervention opportunity, and BIRD and SciFact abstain despite positive value from another declared action. These cases show that harmful targeted intervention is not the only possible outcome of mapping attribution to action. Full results and the six system taxonomy are in Appendix C.

5 Discussion

Information value and intervention value answer different questions. An oracle probe measures whether a specified source can improve utility, while attribution asks where that value appears to

remain unrealized. Neither quantity, by itself, identifies the effect of changing the mechanism that acquires or uses information.

A related distinction appears in causal inference, where effects are defined relative to particular interventions (Pearl, 2009; Hernán & Robins, 2020). Our oracle probe already intervenes on information availability. What it does not identify is the effect of intervening on the mechanism that acquires or uses that information. Likewise, feature attribution need not coincide with interventional relevance (Lundberg & Lee, 2017; Janzing et al., 2020). Our contribution is not this general distinction, but its operationalization for information diagnostics in language models and a prospective test of the resulting attribution to action step.

Our audits show that opportunity is more reproducible than shortfall attribution. The prospective tests show that stable attribution is not sufficient for intervention value in the systems tested. With only two prospective systems, they do not estimate how often the two diverge.

The distinction follows from the estimands. The Information Gap measures the utility available from ideal information, whereas τ_a measures what a concrete action actually produces. A large shortfall therefore need not imply that a feasible intervention can recover it, and the best action necessarily depends on the available menu. One empirical condition that would support actionability is attribution to treatment alignment: across systems, the identity or severity of an attributed shortfall should predict the relative effect of interventions targeting that shortfall. Stability establishes reproducibility of the attribution. Such alignment would provide evidence that it is also informative for action selection.

We therefore use information diagnostics to generate intervention hypotheses and evaluate candidate actions directly on held out outcomes. Adaptive policies that incorporate outcome feedback answer a different question because they use information beyond the attribution itself.

6 Related work

Measuring and adapting information use. Retrieval augmented language modeling has been studied through learned retrieval during pretraining and open domain question answering (Guu et al., 2020; Karpukhin et al., 2020; Izacard & Grave, 2021; Lewis et al., 2020). Later systems make retrieval modular or optimize it toward downstream model utility (Shi et al., 2024). Evaluation frameworks separately measure retrieval, context, and generation quality (Es et al., 2024; Saad-Falcon et al., 2024; Ru et al., 2024; Chen et al., 2024), while long context studies show that making relevant information available does not ensure that models use it effectively (Liu et al., 2024). Adaptive methods instead change the information mechanism itself, deciding when or how to retrieve, reflect, or refine (Jiang et al., 2023; Asai et al., 2024; Yan et al., 2024; Jeong et al., 2024; Yao et al., 2023; Schick et al., 2023; Qin et al., 2024; Madaan et al., 2023; Shinn et al., 2023). Our work asks a complementary question: how much utility a specified oracle source makes available, how much is realized, and whether the resulting attribution is reliable enough to motivate an action. Oracle ablations themselves are common. Our contribution is to separate oracle opportunity, channel validity, and attribution robustness, then evaluate intervention selection as a separate prospective question. Appendix E gives a more detailed comparison.

From measured utility to action selection. The closest connection to our decision question comes from work that evaluates actions by their utility rather than inferring an action from a component diagnosis. Classical value of information compares the expected benefit of acquiring information with its cost (Howard, 1966). Recent work applies related ideas to language models. Uplift RAG defines document utility by the marginal improvement a document provides over the model's internal knowledge and trains retrieval toward that quantity (Qu et al., 2025). Value of Information for language model agents compares the expected utility of requesting additional information with the cost of asking the user (Dong et al., 2026). Both make utility consequences part of the decision itself.

D2R-RAG is particularly close to the diagnosis to action setting: it derives failure signals from retrieval and generation, then uses a contextual bandit to select corrective actions under resource constraints (Hashemifar et al., 2026). Our experiment is complementary: we withhold intervention

feedback and freeze the attribution to action mapping before outcomes are observed, isolating what the attribution alone supports.

7 LIMITATIONS

The prospective evidence comprises two independent task clusters and does not estimate how often attribution and intervention value diverge. QASPER has the documented menu deviation, leaving FEVER as the cleanest registration consistent cell. Both diagnostics use the registered $n{=}56$ budget. Perfect subsample and threshold stability establishes local reproducibility, not representativeness.

Operational regret is conditional on the declared menu. In particular, the reader intervention changes model family and capacity, so its status as the best full-menu action is not a recommendation to upgrade readers. Reader-excluded sensitivity is reported in Section 4, and raw regret is not comparable across task utilities.

Finally, the framework assumes access to credible oracle information on a diagnostic sample and evaluates a frozen batch decision rather than an adaptive online policy. We test one mapping from diagnostic labels to actions and do not establish how alternative mappings, richer action menus, intervention costs, or online feedback would change the result. Appendix C gives further protocol details, and Appendix D describes evidence that would revise our interpretation.

8 CONCLUSION

Additional information can have substantial value for a language model without telling us which change to the system will realize that value. Across our audits, oracle opportunity is the most robust information side measurement, while recovery can be limited by the information mechanism or even invalidated by the channel used to measure it, and shortfall attribution is less robust.

On FEVER and QASPER, stable shortfall attributions selected interventions that reduced held out utility relative to doing nothing ($\tau = -0.078$ and $\tau = -0.048$). FEVER provides the registration consistent counterexample, while QASPER shows the same pattern under the documented menu deviation. These cases establish that attribution stability alone does not guarantee intervention value. They do not estimate how frequently the two diverge.

The central lesson is that information value, shortfall attribution, and intervention value are different quantities. A diagnostic can establish that useful information exists and suggest where its value appears to be lost, but choosing what to change requires evidence about the actions themselves. Our results therefore support using these measurements to form intervention hypotheses, followed by direct evaluation of candidate actions, `NOOP`, and plausible alternatives on held out outcomes. Moving from information value to intervention value is an additional empirical step, not a consequence of the diagnosis.

AI USE STATEMENT

In this work, we used generative AI tools to provide feedback on research methodology and experimental design, assist with method implementation, clean and reformat data, and help interpret experimental results. We did not use generative AI to generate synthetic datasets, develop theoretical models or conceptual frameworks, propose or refine research hypotheses, assist with translation, or conduct qualitative or thematic data analysis. Formulating mathematical claims, providing ingredients for mathematical proofs, and assisting with proof writing are not applicable to this work.

We also used generative AI tools to suggest experimental parameters and assist with software development. All AI assisted work was reviewed by the authors. The core implementation was written and tested by the authors, and code changes produced with AI assistance were checked against the original implementation using unit tests. Most of the literature review was conducted manually by the authors before being extended and refined with AI assistance. Code changes were managed through pull requests and reviewed by other authors before approval, and all parts of the manuscript were reviewed by the authors. We take responsibility for the final content of this work, including text, claims, code, and artifacts produced with the aid of generative AI.

Reproducibility Statement

We provide detailed definitions of the estimands, experimental protocols, action menus, bootstrap procedures, robustness checks, and decision analyses in the main paper and appendices. The supplement includes the preregistration used for the prospective FEVER and QASPER experiments, and we report model variants, sample sizes, random seeds where applicable, protocol deviations, and analysis corrections needed to interpret the results. The code used to run the experiments and analyses will be released upon acceptance.

References

Marcin Andrychowicz, Filip Wolski, Alex Ray, Jonas Schneider, Rachel Fong, Peter Welinder, Bob McGrew, Josh Tobin, Pieter Abbeel, and Wojciech Zaremba. Hindsight experience replay. *Advances in Neural Information Processing Systems*, 30, 2017.

Akari Asai, Zeqiu Wu, Yizhong Wang, Alexander Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, 2024.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*, 2021.

Jiawei Chen, Hongyu Lin, Xianpei Han, Le Sun, et al. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762, 2024.

Zijian Chen, Xueguang Ma, Shengyao Zhuang, Ping Nie, Kai Zou, Sahel Sharifymoghaddam, Andrew Liu, Joshua Green, Kshama Patel, Ruoxi Meng, Mingyi Su, Yanxi Li, Haoran Hong, Xinyu Shi, Xuye Liu, Hosna Oyarhoseini, Nandan Thakur, Crystina Zhang, Luyu Gao, Wenhu Chen, and Jimmy Lin. BrowseComp-Plus: A fair and disentangled evaluation benchmark for deep search agents. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 22349–22370. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1023.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Rei Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

Pradeep Dasigi, Matt Gardner, Tushar Khot, Christopher Clark, Sean Welleck, Ashish Sabharwal, and Hannaneh Hajishirzi. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4599–4610, 2021.

Yijiang River Dong, Tiancheng Hu, Zheng Hui, Caiqi Zhang, Ivan Vulić, Andreea Bobu, and Nigel Collier. Value of information: A framework for human–agent communication. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 42879–42896. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1987.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. RAGAS: Automated evaluation of retrieval augmented generation. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pp. 150–158, 2024.

Gemma Team et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

Aaron Grattafiori, Abhimanyu Dubey, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Kelvin Guu, Kenton Lee, Zora Tung, Ming-Wei Pasupat, and Ming-Wei Chang. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3929–3938, 2020.

Soroush Hashemifar, Havva Alizadeh Noughabi, Fattane Zarrinkalam, and Ali Dehghantanha. Diagnosing and repairing factual errors in RAG under budget constraints. In *Proceedings of the 39th Canadian Conference on Artificial Intelligence*, volume 318 of *Proceedings of Machine Learning Research*. PMLR, 2026.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC, 2020.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6609–6625, 2020.

Ronald A Howard. Information value theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1):22–26, 1966.

Gautier Izacard and Edouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 874–880, 2021.

Dominik Janzing, László Minorics, and Patrick Blöbaum. Feature relevance quantification in explainable AI: A causal problem. In *International Conference on Artificial Intelligence and Statistics*, pp. 2907–2916. PMLR, 2020.

Soyeong Jeong, Jinheon Baek, Sukmin Cho, Sung Ju Hwang, and Jong Park. Adaptive-RAG: Learning to adapt retrieval-augmented large language models through question complexity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7036–7050. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.389.

Zhengbao Jiang, Frank F Xu, Jun Araki, and Graham Neubig. FLARE: Active retrieval augmented generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13153–13167, 2023.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6769–6781, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474, 2020.

Jinyang Li, Binyuan Hui, Ge Qu, Jiaxi Yang, Binhua Li, Bowen Li, Bailin Bai, Lei Yuan, Jian Li, et al. BIRD: A big bench for large-scale database grounded text-to-SQL evaluation. *Advances in Neural Information Processing Systems*, 36, 2023.

Nelson F Liu, Kevin Khandelwal, et al. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Aman Madaan et al. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2 edition, 2009.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. ToolLLM: Facilitating large language models to master 16000+ real-world apis. In *International Conference on Learning Representations*, 2024.

Changle Qu, Sunhao Dai, Hengyi Cai, Yiyang Cheng, Jun Xu, Shuaiqiang Wang, and Dawei Yin. Uplift-RAG: Uplift-driven knowledge preference alignment for retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9632–9644. Association for Computational Linguistics, 2025.

Stephen Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

Dongyu Ru, Lin Qiu, Xiangkun Hu, Tianhang Zhang, Peng Shi, Shuaichen Chang, Jiayang Cheng, Cunxiang Wang, Shichao Sun, Huanyu Li, Zizhao Zhang, Binjie Wang, Jiarong Jiang, Tong He, Zhiguo Wang, Pengfei Liu, Yue Zhang, and Zheng Zhang. RAGChecker: A fine-grained framework for diagnosing retrieval-augmented generation. In *Advances in Neural Information Processing Systems*, 2024. Datasets and Benchmarks Track.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. ARES: An automated evaluation framework for retrieval-augmented generation systems. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 338–354. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.20.

Timo Schick, Jane Dwivedi-Yu, et al. Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 2023.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Richard James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. REPLUG: Retrieval-augmented black-box language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 8371–8384, 2024.

Noah Shinn et al. Reflexion: Language agents with verbal reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 809–819, 2018.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022.

David Wadden, Jenny Lin, Kyle Lo, Lucy Lu Wang, Emma van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 7534–7550, 2020.

Shi-Qi Yan, Jia-Chen Gu, Yun Zhu, and Zhen-Hua Ling. Corrective retrieval augmented generation. *arXiv preprint arXiv:2401.15884*, 2024.

An Yang, Baosong Yang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

An Yang, Baosong Yang, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

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

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, Simon Brunner, Chen Gong, James Hoang, Armel Randy Zebaze, Xiaoheng Hong, Wen-Ding Li, Jean Kaddour, Ming Xu, Zhihan Zhang, Prateek Yadav, Naman Jain, Alex Gu, Zhoujun Cheng, Jiawei Liu, Qian Liu, Zijian Wang, Binyuan Hui, Niklas Muennighoff, David Lo, Daniel Fried, Xiaoning Du, Harm de Vries, and Leandro Von Werra. BigCodeBench: Benchmarking code generation with diverse function calls and complex instructions. In *International Conference on Learning Representations*, 2025.

APPENDIX

A ESTIMANDS FROM INFORMATION VALUE TO INTERVENTION VALUE

This appendix gives the full definitions summarized in Section 2. We first define the measurements of information value, realized value, and shortfall attribution, then define intervention effects and operational regret. Table A.1 gives the registered diagnostic outcomes and clarifies how they relate to the frozen action mapping evaluated in Section 4. The information side measurements end with a descriptive shortfall attribution. Intervention value is then evaluated separately over concrete actions.

A.1 INFORMATION VALUE AND SHORTFALL ATTRIBUTION

Fix a language model, a task metric, a candidate information source, and a channel through which candidate information is supplied at inference. On a diagnostic sample of n items, we construct a specified oracle source g^*. For item i, utility under supplied information g is

$$u_i(g) = \mathbb{E}[m_i(Y) \mid i, g],$$

where the registered task utility satisfies $m_i \in [0, 1]$. Binary correctness is a common special case. Our experiments also use execution accuracy, exact match, unit test pass rate, judge accuracy, verdict accuracy, and task specific continuous scores.

All quantities below are conditional on the model, the representation of g^*, and the channel through which information is supplied. Because task utilities differ across domains, their absolute magnitudes are not comparable across benchmarks.

Opportunity. For item i, oracle lift is

$$L_i = u_i(g^*) - u_i(\emptyset).$$

The *Information Gap* is

$$\Delta = \mathbb{E}_i[u_i(g^*)] - \mathbb{E}_i[u_i(\emptyset)] = \mathbb{E}_i[L_i].$$

Δ measures the average utility made available by the specified oracle source. *Scope* describes how the item level lifts L_i are distributed, distinguishing broad opportunity from opportunity concentrated on a smaller set of items. Because Δ already averages L_i, Scope describes where opportunity occurs rather than adding another source of headroom.

If the registered opportunity screen returns $\Delta \approx 0$, the diagnostic does not proceed to recovery for that source and channel. This is a statement about detectable oracle opportunity under the specified setup, not a claim that no other source or intervention could improve the system.

Instrument validity. A positive Information Gap does not guarantee that the channel used to measure recovery preserves that value. Let $\bar{u}_\emptyset$, $\bar{u}_{\text{retrieved}}$, $\bar{u}_{\text{self}-\text{forced}}$, and $\bar{u}_{\text{gold}-\text{forced}}$ denote mean utility under the corresponding evaluation arms. Instrument gain is

$$\Delta_{\text{instr}} = \bar{u}_{\text{gold}-\text{forced}} - \bar{u}_\emptyset.$$

The channel destruction check marks the instrument `INVALID` when

$$\bar{u}_{\text{gold}-\text{forced}} < \bar{u}_\emptyset.$$

In this case, supplying gold information through the recovery channel reduces utility relative to supplying none. Recovery measurements through that channel are therefore not interpreted. Instrument validity is a precondition for attribution rather than a measure of recovery itself.

Realization and attribution. Given a valid recovery channel, retrieval realization is

$$R_{\text{ret}} = \frac{\bar{u}_{\text{retrieved}} - \bar{u}_\emptyset}{\Delta}.$$

It measures how much of the oracle opportunity is realized by retrieved information.

Self recovery realization is

$$R_{\text{self}} = \frac{\bar{u}_{\text{self}-\text{forced}} - \bar{u}_{\emptyset}}{\bar{u}_{\text{gold}-\text{forced}} - \bar{u}_{\emptyset}}.$$

It measures how much of the utility available from gold information in the same forcing channel is realized by information generated by the model.

We also report *quality utility alignment*, defined as the Pearson correlation between a registered recovery quality proxy and teacher forced task utility on the diagnostic sample. This statistic asks whether examples judged better by the proxy also produce greater downstream utility. It is an association measure and does not enter the Information Gap.

These readings support a descriptive shortfall attribution. They can indicate that value appears to remain unrealized in information acquisition or self recovery, or that no registered dominant shortfall can be assigned. They do not measure the effect of changing either component.

Undefined and unstable ratios. Ratio based readings require a positive denominator. Retrieval realization is undefined when $\Delta \leq 0$. Self recovery realization is undefined when

$$\bar{u}_{\text{gold}-\text{forced}} - \bar{u}_{\emptyset} \leq 0.$$

For preregistered ratio inference on held out analyses, if more than 5% of paired bootstrap replicates have a nonpositive oracle denominator, the ratio is declared unstable and we report raw paired utility differences descriptively instead.

Supporting headroom realization. For analyst designed interventions reported as supporting evidence in Section 3, we additionally report

$$H_a = \frac{\bar{u}_{\text{intervention}} - \bar{u}_{\emptyset}}{\bar{u}_{\text{gold}} - \bar{u}_{\emptyset}}$$

on the registered held out items. H_a measures how much oracle headroom a particular intervention realizes. Arm level H_a is computed from aggregate utilities, whereas ΔH_a in paired arm contrasts is the paired mean of per item headroom fraction differences. The two estimators therefore need not coincide. It is not part of the shortfall attribution and is not the decision estimand used in the prospective test.

A.2 Intervention value and operational regret

The decision analysis changes the object of measurement. The quantities above describe responses to information. The quantities below describe the effects of concrete actions.

Declared action menu. Each system s has a declared action menu A_s. The menu contains `NOOP`, the targeted intervention associated with the diagnostic attribution, a reader intervention, and additional sampling, with the exact implementation specified for each system. For the prospective study, the menu and the mapping from diagnostic labels to actions are frozen before decision outcomes are observed.

All decision quantities are conditional on A_s. Thus, "best action" means best among the actions in the declared and executed menu, not best among all interventions that could have been constructed.

Intervention effects. On held out items, let U_a denote mean utility under action a. The treatment effect of action a relative to `NOOP` is

$$\tau_a = U_a - U_{\text{NOOP}},$$

with

$$\tau_{\text{NOOP}} = 0.$$

Let

$$a^* = \arg\max_{a \in A_s} \tau_a$$

be the best declared action, and let a_{diag} be the action selected by the frozen diagnostic mapping. Operational regret is

$$R_{\text{raw}} = \tau_{a^*} - \tau_{a_{\text{diag}}},$$

as defined in Equation 1. It measures the utility forgone by following the diagnostic instead of the best declared action.

Raw regret is reported in the registered task utility and is therefore interpreted within a system only. Its magnitude is not comparable across tasks with different utility definitions.

Normalized regret. When the registered paired bootstrap detects a positive effect for a^* at $\alpha = 0.05$, we additionally report

$$R_{\text{norm}} = \frac{R_{\text{raw}}}{\tau_{a^*}}.$$

If a positive effect for a^* is not detectable, normalized regret is undefined and we report that there is no detectable intervention opportunity at the available sample size.

On this scale, $R_{\text{norm}} = 0$ means that the diagnostic selected the best declared action, $R_{\text{norm}} = 1$ means that its selected action has the same utility as `NOOP`, and $R_{\text{norm}} > 1$ means that the selected action reduces utility relative to `NOOP` while a better declared action has positive value.

Bootstrap inference and equivalent best actions. The registered equivalence tolerance for system s is

$$\epsilon_s = \frac{1}{n_s},$$

the largest contribution one held out item can make to mean utility. Within bootstrap replicate r, let

$$U^{(r)}_{\max} = \max_{a \in A_s} U^{(r)}_a.$$

Action a is counted as best within tolerance when

$$U^{(r)}_{\max} - U^{(r)}_a \leq \epsilon_s.$$

We report

$$P(a \text{ is best})$$

as the fraction of bootstrap replicates satisfying this condition. The tolerance affects this descriptive probability only. It does not modify a^* or either regret estimand.

Bootstrap inference for operational regret uses the system level a^* defined by the full held out evaluation rather than an itemwise maximum.

Abstention and invalid measurements. `STOP`, `ABSTAIN`, and `WARN` map to the registered baseline when a decision menu is open. Their realized cost therefore enters operational regret rather than being excluded from evaluation.

`INVALID` has a different meaning. It indicates that the measurement channel itself failed the instrument check. No decision menu is opened, and operational regret is not defined. We report C_{decision} as the fraction of systems that reach the decision stage and $C_{\text{actionable}}$ as the fraction for which the diagnostic selects a targeted intervention.

Fixed policy baselines. We evaluate fixed policies on the same held out items as the diagnostic mapping: always `NOOP`, always use the reader intervention, always use the targeted intervention, always use additional sampling, and a uniform random policy over the declared menu.

For the random policy, we use its expectation rather than drawing one random action:

$$\tau_{\pi_{\text{random}}} = \frac{1}{|A_s|} \sum_{a \in A_s} \tau_a.$$

Its operational regret is

$$R_{\text{raw}}(\pi_{\text{random}}) = \tau_{a^*} - \tau_{\pi_{\text{random}}}.$$

Decision stability. Separately from intervention value, we ask whether the diagnostic reproduces the same action when its inputs or decision rule are perturbed. We evaluate four axes where available: diagnostic subsample, decision threshold, reference construction, and reader family. Each axis is reported separately rather than combined into a single stability score.

Stability measures reproducibility of the selected action. Operational regret measures its consequence. A stable action can therefore have either positive or negative intervention value.

A.3 Diagnostic outcomes and the frozen action mapping

The information side measurements produce the six registered outcomes in Table A.1. These outcomes summarize what the diagnostic measurements support. They should not be read as estimates of intervention effects.

The labels `improve acquisition` and `improve self recovery` are retained here because they are the registered protocol outputs. Their names identify the component associated with the measured shortfall. They do not establish that changing that component will improve utility. The prospective experiment in Section 4 tests that additional claim by freezing a mapping from these labels to concrete actions and measuring the resulting intervention effects.

`WARN: attribution unstable` is an audit flag triggered when shortfall attribution fails the registered stability checks. It is not a third value of `dominant_shortfall` and was not part of the original intervention selection mapping.

Table A.1: Six registered diagnostic outcomes. The screen maps information side measurements to shortfall hypotheses. When a decision menu is open, the frozen action mapping is evaluated separately using held out intervention value in Section 4.

Outcome	Diagnostic condition	Interpretation
`STOP: no headroom`	$\Delta \approx 0$	No detectable opportunity for this oracle source and channel
`improve acquisition`	Sharp acquisition shortfall with a valid instrument	Hypothesis: acquisition is the dominant shortfall
`improve self recovery`	Sharp self recovery shortfall with a valid instrument	Hypothesis: self recovery is the dominant shortfall
`INVALID: channel invalid`	gold forced $<$ none	Recovery is not interpreted and no decision menu is opened
`WARN: attribution unstable`	Shortfall disagreement under a registered stability check	Opportunity may remain valid but the attribution is withheld
`ABSTAIN: no actionable shortfall`	Opportunity gate `GO` with no registered dominant shortfall	No targeted shortfall is assigned

When a decision menu is open, `STOP`, `WARN`, and `ABSTAIN` fall back to the registered baseline and are priced by operational regret. Only `INVALID` prevents the decision estimand from being defined.

B Information value and shortfall attribution in detail

This appendix provides the detailed evidence behind the three findings in Section 3. We first measure information opportunity across tasks in Section B.1, then examine how much of that value is realized and whether the recovery channel supports interpretation in Section B.2 and Section B.3. We next test whether the realization gap persists across model scale in Section B.4 and whether shortfall attribution survives changes in diagnostic sample, oracle construction, controlled perturbation, and reader family in Section B.5. Finally, we report supporting interventions designed after observing

the diagnostic and the earlier registered SciFact abstention, which provide context for the transition from attribution to intervention value in Section 4.

B.1 Information opportunity across tasks

We first expand the opportunity results summarized in Section 3. For each benchmark, we fix the model and information delivery channel and compare task utility with the specified oracle source g^* withheld and supplied. The resulting Information Gap Δ measures the utility made available by that source under the fixed setup. The probe returns near zero opportunity on the calibration tasks, substantial positive opportunity on BIRD, multi hop QA, and BrowseComp-Plus, and positive opportunity on BigCodeBench before the recovery channel is tested (see Table 2).

Calibration cases: GSM8K and MBPP. GSM8K (Cobbe et al., 2021) and MBPP (Austin et al., 2021) test whether the oracle probe can register little additional opportunity when the supplied information adds little beyond what the model already has. On GSM8K (n=300, seeds 42 and 123), with conditioning verified, oracle intent injection does not improve over the baseline: the Oracle B2 lift is -0.022. On MBPP (n=300), baseline accuracy is 0.39, Oracle A gives a lift of $+0.010$ that is not statistically detectable, and Oracle B2 is negative. We therefore treat both as calibration cases with near zero headroom rather than as failures of the measurement pipeline.

BIRD SQL. BIRD (Li et al., 2023) provides our most instrumented positive opportunity case. Using gold SQL grounding as g^*, execution accuracy increases from 0.179 to 0.297, giving

$$\Delta = 0.118$$

with a 95% confidence interval of $[0.097, 0.139]$ under the paired bootstrap at n=1534 (see Figure B.1a). Ablations in Figure B.1b show that the gain is associated with value and schema grounding rather than empty evidence or solution hints.

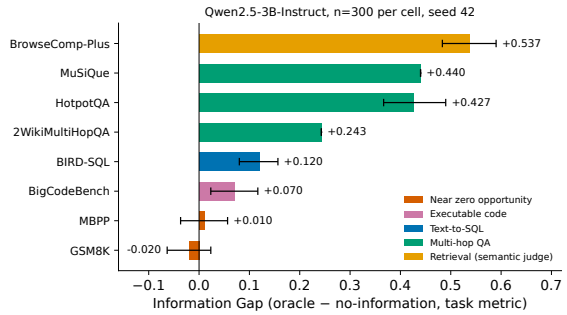
Under the corrected templated oracle protocol used for the scale sweep (n=300), oracle lift increases from 0.120 at 3B to 0.173, 0.180, and 0.203 through 14B. An earlier version of the sweep used an untemplated prompt and produced a different scale pattern. The reported sweep therefore uses the templated protocol throughout.

Multi hop question answering. We apply the same opportunity probe to HotpotQA (Yang et al., 2018), 2WikiMultiHopQA (Ho et al., 2020), and MuSiQue (Trivedi et al., 2022) as one task family. Across these experiments, the measurement code, BM25 (Robertson & Zaragoza, 2009) retrieval procedure, and exact match verifier are fixed, while the datasets differ in candidate pool size and compositional structure. Here, gold passages are evidence that the retriever is expected to recover and can contain text directly supporting the answer. This differs from BIRD and GSM8K, where the oracle source is constructed to exclude the final answer.

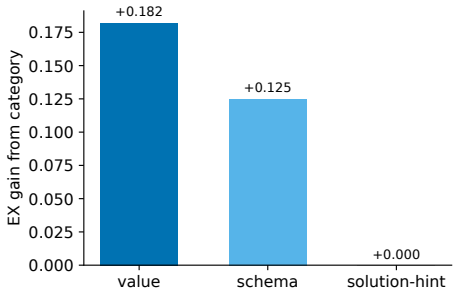
At the Qwen2.5-3B-Instruct anchor with n=300, oracle headroom is $+0.43$ on HotpotQA, $+0.24$ on 2WikiMultiHopQA, and $+0.44$ on MuSiQue. HotpotQA also illustrates the separation between opportunity and realized retrieval value: accuracy is 0.18 without evidence, 0.38 with retrieved evidence, and 0.61 with gold evidence, with recall@$k = 0.61$.

BrowseComp-Plus. BrowseComp-Plus (Chen et al., 2026) is the strongest retrieval opportunity case in our audit. Each item contains roughly 84 candidate documents and 6 gold documents. Answers are evaluated with a Qwen3-32B semantic judge, and the protocol includes explicit `retrieved` and `hard_negative` arms. The gold documents represent evidence that a retrieval system could in principle acquire. They are not answer withholding controls.

At n=300, oracle headroom is $+0.54$ for Qwen2.5-3B and $+0.68$ for Qwen3-8B. Utility is defined by the registered semantic judge, under which the no context arm scores zero. We therefore interpret Δ only within that judge based utility and do not compare its magnitude with other tasks. For Qwen2.5-3B, utility is 0.00, 0.54, 0.05, and 0.00 under none, gold, retrieved, and hard negative evidence. For Qwen3-8B, the corresponding values are 0.00, 0.68, 0.09, and 0.00. The no information arm scores zero under this judge, while supplying oracle evidence produces substantial utility. These values are specific to the BrowseComp-Plus judge and are not directly comparable to execution accuracy or exact match on other tasks.



(a) Information Gap across eight benchmarks at the Qwen2.5-3B anchor: GSM8K, MBPP, BigCodeBench, BIRD SQL, 2WikiMultiHopQA, HotpotQA, MuSiQue, and BrowseComp-Plus.



(b) BIRD execution accuracy lift under value, schema, and solution hint oracle ablations.

Figure B.1: Information opportunity and oracle ablations. **Left:** the opportunity probe distinguishes the two calibration cases from benchmarks with positive oracle headroom. **Right:** BIRD ablations identify which parts of the oracle grounding account for its measured opportunity.

BigCodeBench: positive opportunity before instrument validation. BigCodeBench (Zhuo et al., 2025) also shows positive oracle opportunity. The Information Gap is $+0.070$ with a 95% confidence interval of $[0.023, 0.117]$ for Qwen2.5-3B and $+0.073$ for Qwen3-8B (n=300, unit test pass rate, leakage 0.017). This contrasts with the near zero MBPP calibration result.

The category arms order as

$$\texttt{gold} > \texttt{examples} > \texttt{contract} > \texttt{libs} > \texttt{none} > \texttt{shuffled},$$

which is consistent with the oracle source providing useful API grounding rather than simply exposing a solution. This opportunity result is retained even though the later recovery measurement is invalidated. Section B.3 shows that the recovery channel itself reduces utility when gold information is forced through it, so BigCodeBench contributes an opportunity reading but no interpretable recovery reading.

B.2 Realized information value across tasks

Having established positive oracle opportunity, we next ask how much of that value is realized by information that the system can actually retrieve or generate. We interpret these measurements only when the recovery channel passes the instrument check defined in Appendix A. Across BIRD, multi hop QA, and BrowseComp-Plus, retrieved or self generated information realizes substantially less utility than the corresponding oracle source. BIRD provides the most detailed analysis of self recovery, while HotpotQA and BrowseComp-Plus make the acquisition gap especially clear.

BIRD SQL: self recovery realizes little oracle value. BIRD first lets us separate the ability to use grounding from the ability to produce grounding that carries the same utility. A grounding then SQL student produces grounding with token F1 around 0.40. SQL generation is sensitive to the supplied grounding: permuting grounding across items reduces execution accuracy to 0.133. Yet self recovered grounding realizes little of the utility available from gold grounding.

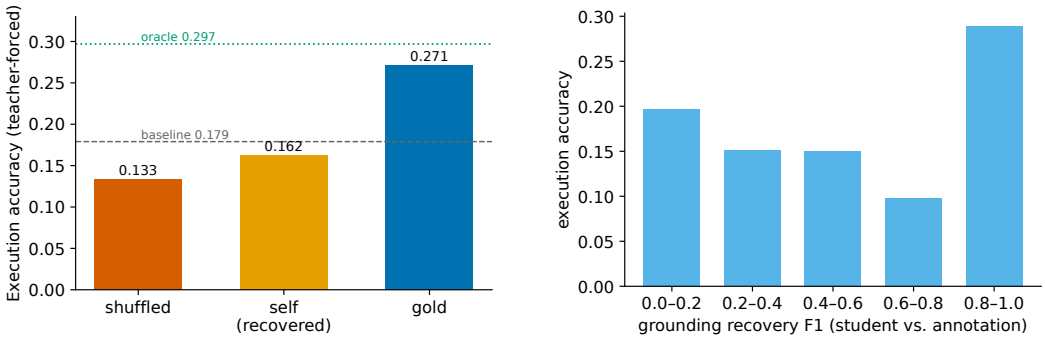
Under the teacher forcing channel, execution accuracy is

$$0.133 \text{ (shuffled)} < 0.162 \text{ (self)} \approx 0.163 \text{ (free generation)} < 0.271 \text{ (gold)}.$$

This gives

$$R_{\text{self}} \approx 0$$

even though the same model benefits substantially when gold grounding is supplied, as Figure B.2a shows. With gold grounding forced through the same channel, execution accuracy reaches 0.271, compared with the baseline value of 0.179. The remaining shortfall therefore cannot be explained simply by an inability to use grounding once useful grounding is available.



(a) Execution accuracy under shuffled, self recovered, and gold grounding in the teacher forcing channel.

(b) Execution accuracy stratified by token F1 of the recovered grounding.

Figure B.2: Realization of grounding value on BIRD. Gold grounding produces a clear utility gain, while self recovered grounding remains close to the baseline and surface recovery quality is weakly associated with execution utility.

Surface grounding quality is also weakly associated with downstream utility. On the diagnostic sample,

$$\mathrm{corr}(\mathrm{F1}, \mathrm{EX}) \approx 0.03,$$

and items with the highest grounding F1 do not outperform the baseline systematically in Figure B.2b. Thus, token overlap is a poor proxy for whether recovered grounding carries execution value in this setting.

BIRD SQL: training on recovered grounding does not close the gap. We additionally test whether selecting or training on different forms of grounding improves transfer to SQL execution. On a matched item set, students trained with gold, random, token similarity, execution verified, or no grounding score between 0.158 and 0.179 execution accuracy. Recovered grounding F1 remains near 0.39 across these variants. Execution aware selection does not improve over random selection (-0.006) or token similarity selection (-0.011).

These results provide a second view of the same shortfall. The model can exploit gold grounding when it is supplied, but neither the recovered grounding nor the selection rules we tested recover comparable downstream utility. They do not identify a unique cause for the failure, but they show that better surface matching alone is not sufficient in this setting.

BIRD SQL: oracle opportunity is concentrated. The aggregate Information Gap also hides substantial variation across items. On $n{=}500$ items with $M{=}5$ to 10 decodes, approximately 11 to 15% of items show large positive lift from gold grounding, while roughly 58 to 60% show little lift even when gold grounding is supplied. The distribution is stable across the tested decode counts. Most of the aggregate $+0.118$ gain therefore comes from a minority of items (see Figure B.3).

This concentration is important for interpretation. A large average opportunity does not imply that missing information is the main limitation on every item. For many examples, supplying the specified oracle grounding changes utility little, while a smaller subset accounts for much of the measured gain.

Multi hop question answering. The same separation between available and realized value appears across HotpotQA, 2WikiMultiHopQA, and MuSiQue. Retrieved passages improve over `none` but remain below gold evidence on all three datasets. Recall@k ranges from 0.41 to 0.61, with MuSiQue lowest under its larger candidate pool.

HotpotQA makes the acquisition gap concrete. Accuracy is 0.18 without evidence, 0.38 with retrieved evidence, and 0.61 with gold evidence at recall@$k = 0.61$. Retrieval therefore realizes part of the available oracle value but leaves a substantial gap.

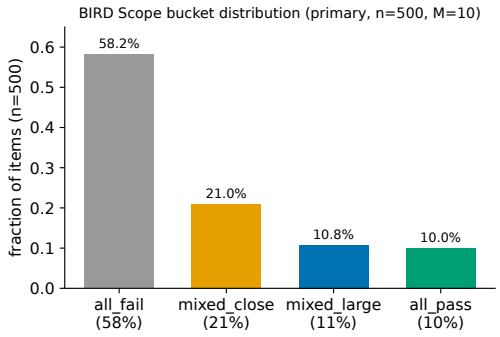


Figure B.3: Distribution of per item oracle lift on BIRD (n=500, M=10). Opportunity is concentrated: much of the aggregate Information Gap comes from a minority of items. The earlier n=150 analysis used a different item pool and is reported in Appendix G.

Table B.1: Realization on multi hop question answering (Qwen2.5-3B-Instruct, n=300). All three datasets show positive oracle opportunity and near zero self recovery.

Dataset	Headroom	Recall@k	Self recovery
HotpotQA	$+0.43$	0.61	≈ 0
2WikiMultiHopQA	$+0.24$	0.57	≈ 0
MuSiQue	$+0.44$	0.41	≈ 0

Self generated evidence realizes still less. Self evidence is nonempty on at least 99% of examples across the family, yet R_{self} remains near zero on all three benchmarks and quality utility alignment is weak. On 2WikiMultiHopQA, shuffled forced evidence can even outperform self forced evidence. The common pattern is therefore not failure to produce any evidence, but failure of that evidence to carry utility comparable to the oracle passages.

BrowseComp-Plus. BrowseComp-Plus shows the largest separation between oracle opportunity and retrieved utility in the retrieval settings we study. At the operational retrieval depth, BM25 recall is approximately 8.1% with $k \approx 6$. Retrieved evidence improves utility by $+0.05$ for Qwen2.5-3B and $+0.09$ for Qwen3-8B, while the remaining gap between retrieved and gold evidence is $+0.49$ and $+0.59$, respectively. Hard negative evidence produces no gain.

Self generated evidence again realizes little oracle value. Self evidence is nonempty on every item, yet forcing it into the model produces essentially no utility gain and gives

$$R_{\text{self}} \approx 0.$$

For Qwen3-8B, quality utility alignment is undefined because utility under self forced evidence has zero variance.

BrowseComp-Plus therefore separates three quantities cleanly. Oracle evidence has substantial value, BM25 retrieves only a small part of that value, and self generated evidence realizes almost none of it. The recovery channel itself remains valid, so these shortfalls are interpretable as realization failures rather than artifacts of the information delivery mechanism.

B.3 Instrument validity boundary on BigCodeBench

BigCodeBench (Zhuo et al., 2025) separates information opportunity from the validity of the channel used to measure recovery. As shown in Section B.1, oracle information has positive value: the Information Gap is $+0.070$ $[0.023, 0.117]$ for Qwen2.5-3B and $+0.073$ for Qwen3-8B. The opportunity measurement therefore indicates that useful information exists for both models.

The recovery channel does not preserve that value. When gold information is inserted through the teacher forcing channel used for recovery measurement, utility falls below the no information

baseline for both models. For Qwen2.5-3B, accuracy decreases from 0.273 with no supplied information to 0.103 with gold information forced through the channel. This violates the instrument validity condition in Appendix A.

We tested three alternative channel designs on a controlled sample of n=75. None restored

$$\mathrm{acc}_{\text{gold-forced}} \geq \mathrm{acc}_{\text{none}}.$$

We therefore mark the recovery instrument `INVALID`. The result does not imply that information recovery is impossible on BigCodeBench. It means that recovery cannot be interpreted through the full function body teacher forcing channel we tested. This contrasts with the shorter completion channels used for BIRD and question answering, where supplying gold information through the recovery channel preserves positive utility.

The same limitation applies to quality utility alignment. We report

$$\mathrm{corr}(\mathrm{F1}, \mathrm{utility}) = -0.04 \quad \text{and} \quad +0.04$$

for the two models descriptively, but do not use these values in cross task conclusions because the utility measurement itself is taken through an invalid channel.

BigCodeBench therefore contributes an opportunity result but no recovery attribution and no decision analysis. Because the instrument is invalid, no action menu is opened and operational regret is not defined. This is the intended role of the instrument check: it prevents a recovery number from being interpreted when the channel used to produce that number has already destroyed the oracle benefit.

B.4 Realization under model scaling

The anchor results in Section B.2 show substantial oracle opportunity but little self recovery. We next test whether this separation disappears as model size increases. Starting from the Qwen2.5-3B anchor, we evaluate Qwen3-4B and Qwen3-8B (Yang et al., 2025) on BIRD and all three multi hop QA datasets, Qwen3-14B on BIRD and HotpotQA, and Qwen3-8B on BrowseComp-Plus.

Across benchmarks with a valid recovery instrument, increasing model size does not produce a consistent increase in self recovery realization. R_{self} remains near zero through 14B on BIRD and HotpotQA and through 8B on 2WikiMultiHopQA, MuSiQue, and BrowseComp-Plus. Quality utility alignment likewise remains weak without a consistent upward trend. Oracle utility can increase with model size, including on BIRD and MuSiQue under the templated oracle protocol, but realized self recovery does not increase proportionally. Thus, larger models can make better use of oracle information without becoming correspondingly better at producing information that carries the same utility.

The preregistered 4B versus 8B scaling gate did not reveal a recovery trend that warranted further scaling under the protocol. Additional 14B runs on BIRD and HotpotQA left the qualitative reading unchanged. These experiments do not establish that scale can never close the gap. They show only that the self recovery shortfall persists across the model sizes tested here. The scale experiments test whether the observed realization gap disappears within the 3B to 14B range. They are not evidence about frontier scale models. The conceptual distinction between information value and intervention value does not depend on this scale result.

Table B.2: Summary of information side measurements at the Qwen2.5-3B anchor. The eight benchmarks span six task settings and are not a random sample. Task utilities differ across settings, so raw magnitudes are not comparable across rows.

Task	Opportunity	Realization	Scope	Instrument	Q–U alignment
GSM8K / MBPP	near zero	not evaluated	—	—	—
BIRD	moderate	self ≈ 0	concentrated	valid	weak
Multi hop QA	large	retrieval partial, self ≈ 0	moderate	valid	weak
BigCodeBench	small positive	not interpreted	mostly all fail	invalid	not interpreted
BrowseComp-Plus	large	retrieval partial, self ≈ 0	broad	valid	weak

Retrieved information behaves differently. When retrieval recall is nontrivial, as on HotpotQA, retrieved evidence realizes part of the available oracle value. When recall is low, as on BrowseComp-Plus, realized retrieval remains far below oracle utility. The distinction is therefore not simply between small and large models: acquisition depends on what evidence the retrieval mechanism obtains, while self generated evidence remains low utility across the tested scales.

Table B.2 summarizes the information side readings before attribution robustness is considered. The calibration tasks show little opportunity for the specified source and channel. BIRD, multi hop QA, and BrowseComp-Plus show positive opportunity but incomplete realization. BigCodeBench shows that positive opportunity can coexist with an invalid recovery channel. Across the settings where recovery is interpretable, available oracle value can therefore remain substantially larger than the value realized by retrieved or self generated information.

These measurements establish that a shortfall exists, but locating that shortfall is a stronger claim. The next subsection tests whether the attribution survives changes in diagnostic sample, oracle construction, and reader family.

B.5 Robustness of shortfall attribution

The previous analyses establish that oracle opportunity can remain much larger than realized information value. We now test the stronger claim required to locate that shortfall: whether the resulting attribution survives reasonable changes to the diagnostic procedure. We keep the registered decision rules fixed and vary the diagnostic sample, oracle construction, controlled perturbation, or reader family. Across these audits, opportunity is more reproducible than shortfall attribution. Acquisition attribution calibrates cleanly on HotpotQA, while BIRD exposes several cases in which the opportunity reading remains stable but the shortfall label does not.

Subsample stability. We repeatedly subsample HotpotQA, BrowseComp-Plus, and BIRD at $n \in \{20, 50, 100, 200, 300\}$. The opportunity gate and the broad versus concentrated Scope label reproduce above their registered thresholds on all three domains. The registered gate stability criterion is at least 90% agreement with median stopping size at most 100 items. The observed median stopping sizes are 30, 30, and 50 for HotpotQA, BrowseComp-Plus, and BIRD, respectively.

Table B.3: Stability of opportunity and shortfall attribution. The first two columns report the probability that a random n=100 subsample reproduces the full sample decision over 1000 trials. Median stop n is the sample size at which the opportunity gate stabilizes. Degraded g^* counts how many of three alternate reference constructions preserve both labels. Reader family records whether Llama-3.1-8B and Gemma-2-9b reproduce the dominant shortfall.

Domain	Opportunity gate / Scope	Dominant shortfall	Median stop n	Degraded g^*	Reader family
HotpotQA	1.00 / 1.00	1.00	30	$3/3$	agrees
BrowseComp-Plus	1.00 / 1.00	1.00	30	$2/3$	agrees
BIRD	1.00 / 1.00	**0.52**	50	$3/3$	**disagrees**
Registered threshold	≥ 0.90	≥ 0.80	—	—	—

Shortfall attribution is less stable. Dominant shortfall agreement exceeds the registered 80% threshold on HotpotQA and BrowseComp-Plus but not on BIRD. At n=100, BIRD reproduces the full sample shortfall label with probability 0.52 in Table B.3, and agreement remains close to this level across subsample sizes below the full sample. The full sample label is `none`, but that attribution is therefore not reliably recovered under resampling. We report this as a registered failure rather than changing the threshold after observing the result.

Reference construction. The attribution can also depend on how the oracle source g^* is represented. Alternative oracle constructions leave the main opportunity reading stable on HotpotQA and BIRD. BrowseComp-Plus is more sensitive. The canonical construction gives an Information Gap

Table B.4: Sensitivity to the construction of g^* on BrowseComp-Plus (n=300). ΔGap is the change in Information Gap relative to the canonical construction.

Construction	Gap	ΔGap	Dominant shortfall
Canonical	0.53	0.00	acquisition
Minimal evidence	0.11	−0.42	acquisition
Utility selected	0.05	−0.48	`self_recovery`

of 0.53 and an acquisition shortfall, while a utility selected construction reduces the gap to 0.05 and changes the dominant shortfall to `self_recovery` (see Table B.4).

The shortfall attribution therefore inherits choices made when the oracle reference is constructed. This does not invalidate the opportunity measurement, but it limits how strongly the resulting attribution can be interpreted independently of that construction.

Controlled calibration. We next ask whether the attribution responds as expected when a shortfall is deliberately introduced. On HotpotQA, degrading retrieval quality moves retrieval realization monotonically while oracle opportunity remains stable. The acquisition measurement therefore tracks the planted acquisition failure in the expected direction.

The corresponding recovery test is less successful on BIRD. Perturbing grounding quality does not move the borderline full sample attribution from `none` into a recovery shortfall at an Information Gap of approximately 0.09. The registered recovery calibration criterion is therefore not met in this setting.

Recovery quality proxies also show weak association with task utility. On the Qwen anchor, the strongest reported associations are 0.271 for token multiset F1 on HotpotQA, 0.089 on BrowseComp-Plus, and −0.090 for entity title recall on BIRD. Cross domain Kendall τ for the proxy rankings is near zero between BIRD and the question answering domains. These comparisons provide no evidence that one proxy ranking transfers consistently across the domains tested.

Reader family robustness. We repeat the diagnostic readings with Llama-3.1-8B (Grattafiori et al., 2024) and Gemma-2-9b (Gemma Team et al., 2024) on a shared set of 300 items per domain. On BrowseComp-Plus and HotpotQA, per item oracle lift correlates with the Qwen2.5-3B reference at approximately $r = 0.49$ to 0.65, and on scope overlap ranges from 79% to 95%. The opportunity gate, Scope, and dominant shortfall labels agree across reader families.

BIRD again differs. Per item lift correlation remains moderate, approximately $r = 0.51$ to 0.55, but on scope overlap falls to 44% to 61%. More importantly, the dominant shortfall changes from `none` under Qwen to `self_recovery` under the alternate readers in Figure B.4. Thus, the presence of oracle opportunity transfers more reliably across readers than the attribution of where that opportunity remains unrealized.

B.6 Supporting interventions designed after diagnosis

The attribution audits above ask whether the diagnostic reading itself is reproducible. A separate, weaker question is whether an analyst can use those readings to design an intervention that realizes oracle headroom. These experiments allow the analyst to inspect the diagnostic before choosing the intervention and therefore do not test decision validity. They provide supporting evidence for the hypothesis generation role described in Section 3.

BrowseComp-Plus. On 530 held out BrowseComp-Plus items, dense retrieval followed by reranking yields $H_a = 0.180$ $[0.135, 0.229]$, compared with 0.134 for the BM25 baseline. Arm level H_a is computed from aggregate utilities, whereas ΔH_a is the paired mean of per item headroom fraction differences. The two estimators therefore need not coincide. The targeted intervention also increases raw judge accuracy over BM25 by 0.025 under a one sided paired bootstrap ($p = 0.033$).

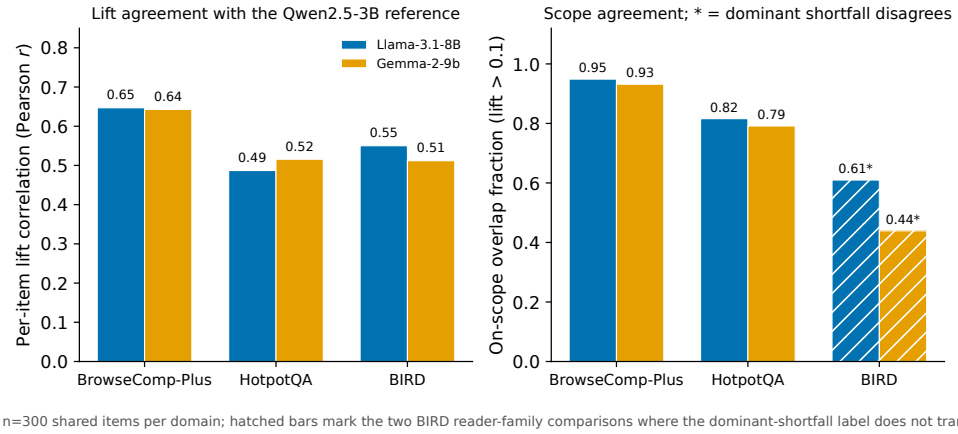


Figure B.4: Reader family robustness on 300 shared items per domain. Left: correlation of per item oracle lift with the Qwen2.5-3B reference. Right: overlap in items above the registered lift threshold of 0.1. Hatched bars mark the two BIRD comparisons for which the dominant shortfall label changes across reader families.

Table B.5: BrowseComp-Plus held out contrasts (n=530, 10,000 paired bootstrap replicates, seed 42). Table ΔH_a uses the paired per item estimator defined above. Ratio based inference is unstable under the registered denominator check, so raw accuracy contrasts are descriptive.

Comparison	ΔH_a	Holm p (ratio)	Δacc	p (raw acc)
Targeted vs baseline	$+0.028$	0.11	$+0.025$	0.033
Targeted vs `CTRL_OTHER`	$+0.166$	< 0.001	$+0.096$	< 0.001
Targeted vs `CTRL_READER`	$+0.011$	0.32	$+0.006$	0.36
`CTRL_SAMPLE`	—	not run	—	—

The registered ratio analysis is not stable, however. More than 5% of bootstrap replicates have a nonpositive oracle denominator. The observed fraction is 46.6%. Under the registered rule, ratio based inference is therefore declared unstable and the raw paired utility differences are reported descriptively rather than as confirmatory evidence.

The targeted intervention exceeds the best of N sham by 0.096 raw judge accuracy ($p < 0.001$), but does not establish an advantage over the reader control, for which the difference is 0.006 ($p = 0.36$). The registered `CTRL_SAMPLE` arm was not executed, which is a protocol deviation. We therefore do not claim that the registered all controls criterion was satisfied.

BIRD. On 1234 held out BIRD items, deterministic schema and value linking yields

$$H_a = 0.785 \quad [0.620, 0.986],$$

compared with 0.000 for the baseline. The intervention was developed with the diagnostic readings available. The final audited attribution is `none` with the additional flag `WARN: attribution unstable`.

This result therefore does not show that the diagnostic selected the intervention. It shows something weaker but important: a large intervention gain can exist even when the corresponding shortfall is not stably inferable from the diagnostic. A useful intervention and a reliably inferred intervention are different properties.

Why these results are supporting evidence. The BrowseComp-Plus and BIRD diagnostic readings were available while the corresponding interventions were being developed. The intervention therefore reflects both the diagnostic and analyst knowledge about the system. These experiments show that the measurements can inform intervention hypotheses, but they cannot establish that the

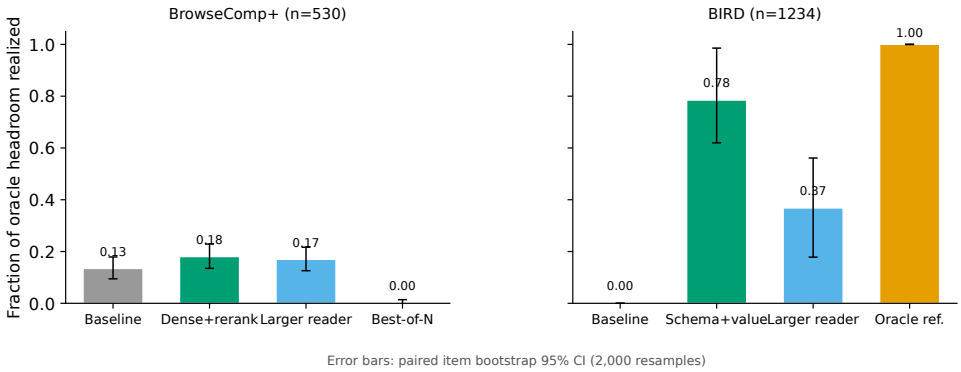


Figure B.5: Supporting interventions designed after inspecting the diagnostic readings. The figure reports the fraction of oracle headroom realized by each arm on BrowseComp-Plus and BIRD. The BIRD intervention was developed under the earlier self recovery interpretation, while the final audited diagnosis is `none` with unstable attribution. Error bars show paired item bootstrap 95% confidence intervals.

Table B.6: Earlier registered SciFact diagnosis on the frozen diagnostic split (n=56). Retrieval and self recovery both realize most of the measured oracle opportunity, so no dominant shortfall is assigned and the frozen mapping selects `NOOP`.

Reading	Value
Opportunity gate / Scope	GO / concentrated
Information Gap Δ	0.634 $[0.511, 0.748]$
Retrieval realization R_{ret}	0.730
Self recovery realization R_{self}	0.969
Dominant shortfall	`none`
Frozen mapping	`NOOP`

diagnostic label itself selects a useful action. The prospective test in Section 4 removes this analyst choice by freezing the mapping before intervention outcomes are observed.

B.7 Earlier registered abstention on SciFact

SciFact (Wadden et al., 2020) provides a separate supporting case in which the diagnostic prospectively selected `NOOP` under an earlier registration. On the frozen diagnostic split of n=56 items, the opportunity gate is `GO`, Scope is concentrated, and the Information Gap is 0.634 $[0.511, 0.748]$. Retrieval realization is 0.730 and self recovery realization is 0.969. No registered dominant shortfall is assigned.

The registered mapping therefore selects `NOOP`, and no intervention split is opened at diagnosis time (see Table B.6). This preserves the prospective abstention from the earlier study rather than retroactively choosing an intervention. Section 4 later opens a menu only to measure the cost of that frozen decision.

C Intervention value and decision outcomes in detail

This appendix expands the decision analysis in Section 4. We first compare the frozen diagnostic mapping with fixed policies on the same prospective test items in Table C.1. We then summarize the outcomes across all systems with declared action menus in Table C.2 and give the detailed interpretation of alignment, immaterial opportunity, abstention, and coverage. FEVER and QASPER provide the prospective evidence. BrowseComp-Plus, BIRD, and HotpotQA provide supporting analyses, while SciFact verifies a decision frozen under an earlier registration.

C.1 FIXED POLICY COMPARISONS

Table C.1: Operational regret R_{raw} for the frozen diagnostic mapping and fixed policies on the same untouched test items. Lower regret is better. π_{random} uses the expected utility of choosing uniformly from the declared menu rather than one sampled action. π_{oracle} selects the best declared action and therefore has zero regret by construction.

Policy	FEVER	QASPER
π_{diag} (frozen diagnostic mapping)	0.403	0.084
$\pi_{\text{always sample}}$	0.328	0.056
$\pi_{\text{always NOOP}}$	0.324	0.036
π_{random} (uniform over menu)	0.264	0.044
$\pi_{\text{always reader}}$	0.000	0.000
π_{oracle}	0.000	0.000

On both prospective systems, the frozen diagnostic mapping has greater operational regret than every executable fixed policy shown in Table C.1 when the reader intervention is in the menu. On FEVER, its regret is 0.403, compared with 0.324 for always choosing `NOOP`, 0.328 for always sampling, and 0.264 for choosing uniformly from the menu. On QASPER, the corresponding values are 0.084, 0.036, 0.056, and 0.044. If the reader action is removed from both menus in a descriptive sensitivity analysis, `NOOP` is the best remaining action and regret of the frozen mapping against that restricted menu is 0.078 on FEVER and 0.048 on QASPER, matching the negative targeted treatment effects relative to `NOOP`. Operational regret under the full declared menu remains the registered estimand.

Neither prospective diagnostic abstains. Within each system, the frozen mapping is therefore behaviorally identical to deploying the intervention associated with its attributed shortfall. The prospective results consequently provide no evidence that conditioning intervention choice on these diagnostic readings improves over committing to the corresponding targeted intervention in advance.

The always reader policy has zero regret on both systems because the reader intervention is the best action in each declared menu. As discussed in Section 4, this should not be interpreted as evidence that reader upgrades are generally preferable. The comparison is menu specific, and the reader intervention also changes model family and capacity.

C.2 OUTSIDE MENU CONTROLS ON PROSPECTIVE SYSTEMS

Plausible actions outside the declared menu can outperform the frozen targeted intervention. On QASPER, BM25 paragraph retrieval gives $\tau \approx 0.139$, above both the reader intervention at 0.036 and the targeted self recovery intervention at -0.048. On FEVER, self evidence generation gives $\tau \approx 0.299$, below the reader intervention at 0.324 but far above the targeted acquisition intervention at -0.078. These controls do not enter operational regret. They illustrate that the ranking of interventions depends on the declared menu.

C.3 DECISION OUTCOMES ACROSS SYSTEMS

Aligned ranking on BrowseComp-Plus. BrowseComp-Plus is the one system in which the frozen attribution selects the best action in the declared menu. Dense retrieval followed by reranking is both the targeted acquisition intervention and the menu maximum, giving

$$R_{\text{raw}} = 0.$$

Its effect is small, however: $\tau = 0.025$ judge accuracy, and a positive effect for the best action is not detectable at n=530. Normalized regret is therefore undefined. The result supports the ranking of the declared actions, not a claim that the selected intervention produces a large or detectable gain.

This system is also analyzed in Section B.6, where the intervention was designed after inspecting the diagnostic readings. That earlier analysis establishes supporting intervention evidence rather than prospective decision validity.

Table C.2: Decision outcomes across the evaluated systems. τ_{diag} is the held out effect of the action selected by the frozen mapping, and R_{raw} is regret relative to the best action in that system's declared menu. FEVER and QASPER are prospective. BrowseComp-Plus, BIRD, and HotpotQA are supporting analyses. SciFact verifies an abstention frozen under an earlier registration. Task utilities differ across systems, so magnitudes are not comparable across rows.

System	Evidence	n_{test}	Frozen decision	Best menu action	τ_{diag}	R_{raw}
Targeted intervention reduces utility						
FEVER	prospective	4178	acquisition	reader (+0.324)	−0.078	**0.403**
QASPER	prospective	285	self recovery	reader (+0.036)	−0.048	**0.084**
Abstention forgoes positive intervention value						
BIRD	supporting	1234	`NOOP`	self recovery (+0.092)	0.000	0.092
SciFact	earlier registration	76	`NOOP`	reader (+0.105)	0.000	0.105
Aligned or immaterial						
BrowseComp-Plus	supporting	530	acquisition	acquisition (+0.025)	+0.025	0.000
HotpotQA	supporting	7105	self recovery	acquisition (+0.0009)	+0.0001	0.0008
No decision analysis						
GSM8K, MBPP	calibration	—	`STOP` ($\Delta\approx 0$)	no menu	—	—
BigCodeBench	instrument boundary	—	`INVALID`	no menu	—	—

Immaterial decision opportunity on HotpotQA. On HotpotQA, the frozen mapping selects self recovery while acquisition has the highest measured utility in the menu. The difference is practically negligible. Every declared intervention changes utility by less than 0.001 on 7105 held out items, and the effect of the best action is not detectable. Operational regret is 0.0008 because little utility separates the available actions.

We therefore treat HotpotQA as an immaterial decision opportunity rather than as evidence that the diagnostic selected the correct intervention. The decision problem is well defined, but there is little measured utility to gain from choosing differently.

Costly abstention on BIRD and SciFact. BIRD and SciFact show that declining to select a targeted intervention can itself have measurable cost. On BIRD, the frozen mapping assigns no dominant shortfall and therefore selects `NOOP`. A declared self recovery intervention produces

$$\tau = 0.092$$

execution accuracy on 1234 held out items, giving

$$R_{\text{raw}} = 0.092.$$

The BIRD attribution is also unstable under the audits in Section B.5. The decision analysis therefore prices the consequence of abstaining without implying that the diagnostic had enough evidence to identify the missed intervention reliably.

SciFact provides a different case. Its `NOOP` decision was frozen under an earlier registration and is verified rather than derived again here. When a decision menu is later opened on 76 items, the reader intervention gives

$$\tau = 0.105$$

verdict accuracy. The frozen `NOOP` decision therefore has

$$R_{\text{raw}} = 0.105.$$

These cases illustrate why abstention is included in the decision estimand. Avoiding a targeted action can prevent exposure to a poor intervention, but it can also forgo positive intervention value.

Systems without a decision estimand. GSM8K and MBPP stop after the opportunity measurement because the specified oracle source provides near zero detectable headroom. BigCodeBench stops for a different reason: positive opportunity is present, but the recovery channel fails the instrument validity check. No action menu is opened in either case.

BigCodeBench is therefore not assigned zero regret. Operational regret is undefined because the system never enters the decision analysis. This distinction preserves the role of the instrument check described in Appendix A.

Coverage. Among the six systems for which a decision menu is opened,

$$C_{\text{actionable}} = \frac{4}{6} \approx 0.667.$$

BIRD and SciFact account for the two nonactionable decisions.

Within this six system set,

$$C_{\text{decision}} = 1.0$$

by construction because every included system reaches the decision analysis. This quantity should not be confused with coverage over the full benchmark portfolio: BigCodeBench is excluded before a menu is opened because its recovery instrument is `INVALID`.

Both prospective systems produce actionable decisions. Thus, the prospective result is not driven by a rule that avoids selecting an intervention. On FEVER and QASPER, the mapping commits to a targeted action, and that action reduces held out utility relative to `NOOP`.

We do not aggregate operational regret across systems because each task defines utility differently.

D Falsifiability and future tests

Our conclusions are conditional on the model, oracle source g^*, information delivery channel, diagnostic procedure, and, for decision analysis, the declared action menu. The experiments therefore do not support a universal claim that information diagnostics cannot guide interventions. This appendix states which parts of our interpretation can be tested directly and what evidence would change how broadly we read the present results.

Attribution to treatment alignment. One empirical condition that would support actionability is attribution to treatment alignment: across systems, the identity or severity of an attributed shortfall should predict the relative effect of interventions targeting that shortfall. Stability establishes reproducibility of the attribution. Such alignment would provide evidence that it is also informative for action selection. Evidence that would revise our interpretation includes prospective systems in which stable attributions consistently rank first among targeted interventions on held out data, or controlled perturbations that move both attribution and the relative ordering of τ_a together. Additional registration consistent counterexamples would strengthen the case that attribution alone does not license the registered intervention.

Tests of the information side conclusions. The following implications extend directly from the measurements in this paper.

1. When the specified oracle source provides near zero Information Gap, there is little measured opportunity to recover from that source through the tested setup. A substantial gain from an intervention whose only mechanism is to approximate the same g^* through the same channel would show that the oracle probe was not functioning as the intended upper reference.
2. When oracle opportunity is positive but retrieval obtains little of the relevant evidence, retrieved utility can remain far below oracle utility, as on BrowseComp-Plus. Increasing retrieval quality without increasing task utility would weaken retrieval recall as an explanation of the realization gap. Increasing both together would strengthen it.
3. Weak quality utility alignment provides little evidence that improving a surface recovery proxy alone will improve task utility. An intervention that reliably increases downstream utility by optimizing such a proxy, without changing other parts of the system, would revise this interpretation.
4. The attribution robustness results can be tested directly under new diagnostic samples, oracle constructions, and reader families. More stable BIRD attribution, or systematic instability of the opportunity reading under the same perturbations, would revise the asymmetry observed in our audits.
5. The self recovery gap remains large through the model sizes tested here. Larger models that substantially increase R_{self} while preserving the same oracle and recovery channels would show that the observed gap was specific to the scale range we studied.

Tests that could change the decision interpretation. The prospective evidence establishes two cases in which stable shortfall attributions map to interventions with negative held out value. It does not estimate how frequently this occurs. A larger preregistered collection of independent systems could therefore change the practical interpretation substantially. If a frozen diagnostic mapping consistently selected interventions with positive τ_a and lower operational regret than simple fixed policies, attribution could still be useful as a predictive signal for action even though the FEVER and QASPER counterexamples would remain.

Several experiments would help determine where the present failure originates. A richer mapping could condition on realization fractions, uncertainty, or other diagnostic measurements rather than only the dominant shortfall label. Larger diagnostic samples would test whether $n{=}56$ is representative enough for the target distribution. Perfect subsample and threshold stability rules out some forms of local instability but not distribution mismatch. Menus with more closely matched interventions would test sensitivity to the available action set, including reader and acquisition changes with comparable capacity or engineering effort. An online repair policy over the same actions, such as the contextual bandit approach used by D2R-RAG (Hashemifar et al., 2026), would test how much additional value comes from observing action outcomes rather than relying on the diagnostic attribution alone. These experiments would not erase the intervention effects measured here, but they could change the scope of the conclusion drawn from them.

Extensions not tested here. We do not train directly against downstream information utility. Such training is a natural test of the weak quality utility alignment observed in Section B.2. We also do not study sequential diagnosis, where an intervention is applied, its effect is observed, and the diagnostic is run again. Finally, actions are ranked by registered task utility rather than by compute, latency, or engineering cost. Extending τ_a to include these costs would turn the declared menu into a cost sensitive decision problem rather than the utility only comparison studied here.

E Questions measured by neighboring evaluation styles

Table E.1 summarizes the questions emphasized by evaluation styles discussed in Section 6. The comparison is descriptive rather than a ranking. Different evaluations are designed for different purposes: task scores measure final performance, retrieval and generation evaluations separate parts of a retrieval system, oracle ablations measure available information value, and our decision analysis additionally measures the utility of acting on a frozen attribution.

Table E.1: Questions emphasized by neighboring evaluation styles. Checkmarks indicate that the quantity is measured explicitly. Dashes indicate that it is not the primary target. Individual methods within a category can differ.

Style	Oracle	Retr.	Use	Self	Channel	Interv.
End to end score	—	—	merged	—	—	—
RAG evaluation	—	✓	✓	varies	—	—
Oracle ablation	✓	—	—	—	—	—
This work (information)	✓	✓	✓	✓	✓	—
This work (decision)	✓	✓	✓	✓	✓	✓

F Protocol checklist

For a retrieval system, tool policy, memory mechanism, hindsight procedure, or other information producing system, the framework can be applied as the following sequence. The first three steps measure information value and shortfall attribution. The final step separately measures the value of acting on that attribution.

1. Specify the information opportunity. Choose the candidate information source, model, task utility, and information delivery channel. Construct a credible oracle g^* on a diagnostic sample.

Measure

$$L_i = u_i(g^*) - u_i(\emptyset)$$

and

$$\Delta = \mathbb{E}_i[L_i],$$

together with Scope. If the registered opportunity screen returns near zero headroom, stop the recovery analysis for that source and channel.

2. Verify the recovery instrument. Before interpreting retrieval or self recovery, supply gold information through the same recovery channel. If

$$\bar{u}_{\text{gold-forced}} < \bar{u}_{\emptyset},$$

mark the instrument `INVALID`. Do not interpret recovery measurements or open a decision menu through that channel.

3. Measure realization and shortfall attribution. When the instrument is valid, measure retrieval and self recovery realization fractions together with quality utility alignment. Audit the resulting attribution under diagnostic resampling, decision thresholds, alternate oracle constructions, and reader families where available. The output of this stage is a hypothesis about where information value remains unrealized, not an estimate of intervention value.

4. Measure intervention value separately. Declare the actions under consideration and freeze the mapping from diagnostic output to action before observing decision outcomes. Evaluate every declared action on held out items and estimate

$$\tau_a = U_a - U_{\text{NOOP}}.$$

Compare the selected action with the best declared action using

$$R_{\text{raw}} = \tau_{a^*} - \tau_{a_{\text{diag}}},$$

and report fixed policy baselines such as always `NOOP` and uniform random choice over the menu.

When designing future decision studies, our results also motivate including plausible alternatives that the diagnosis does not nominate when feasible. Such actions should be distinguished from the declared menu if they are introduced only as outside menu controls.

The diagnostic attribution proposes what may be worth changing. Held out intervention effects determine which measured action actually has value.

G Additional validation analyses

Controlled calibration. Figure G.1 gives the controlled perturbation experiments summarized in Section B.5. On HotpotQA, progressively degrading retrieval changes acquisition realization while oracle opportunity remains stable. On BIRD, perturbing grounding quality does not reliably move the borderline attribution into a self recovery shortfall.

Acquisition curve. Figure G.2 shows the relation between retrieval recall and realized oracle headroom on HotpotQA while holding the reader fixed. The curve provides the detailed calibration evidence behind the acquisition result reported in Section B.5.

Recovery proxies and task utility. Figure G.3 compares the recovery proxies evaluated within each domain and their rank agreement across domains. The weak and inconsistent agreement supports the conclusion in Section B.5 that no single proxy ranking transfers consistently across the audited settings.

Historical BIRD Scope analysis. An earlier BIRD analysis used a filtered legacy pool of n=150 items that is not reproduced by the current pipeline. It classified 64% of items as `all_fail`, 14% as `mixed_close`, and 22% as `mixed_large`. The current analysis instead uses the primary n=500 item set reported in Figure B.3. We retain the earlier result only as historical context and do not use its 22% estimate in the paper's quantitative conclusions.

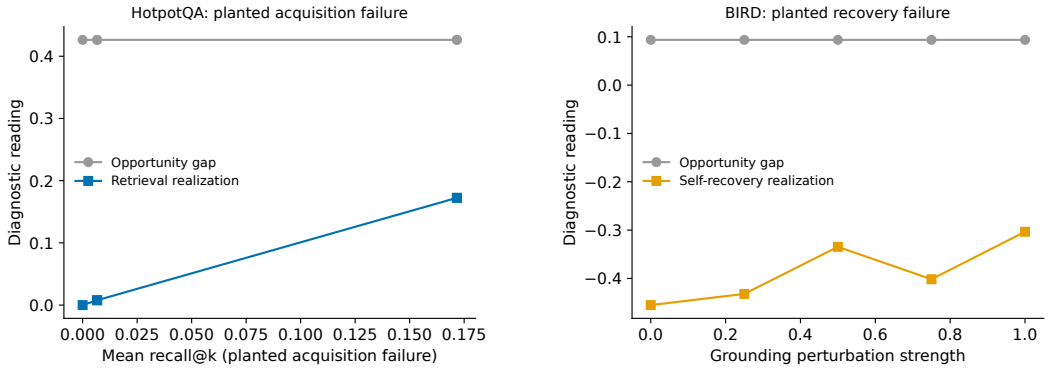


Figure G.1: Controlled calibration of shortfall attribution. Left: HotpotQA acquisition under degraded retrieval. Right: BIRD self recovery under grounding perturbation.

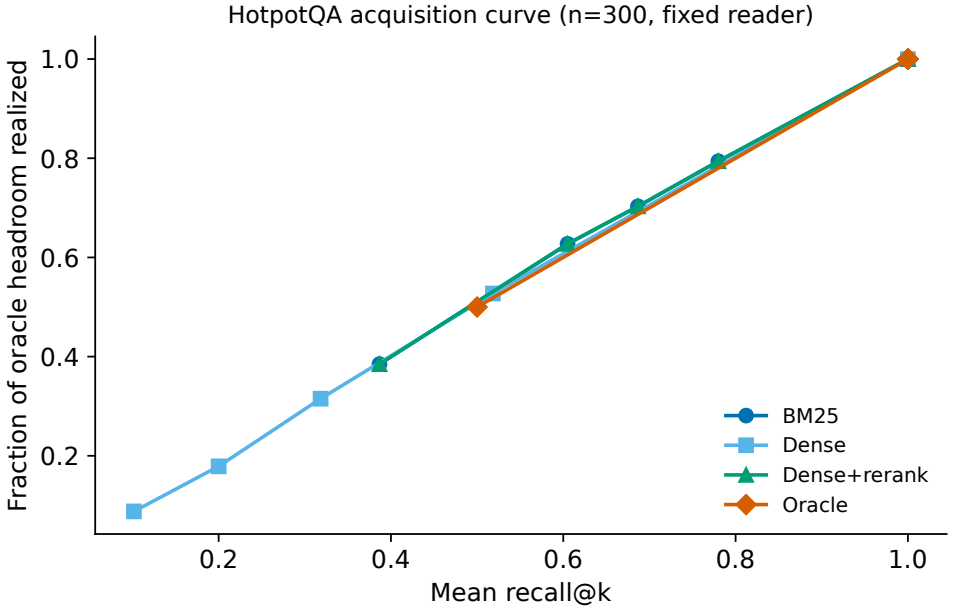
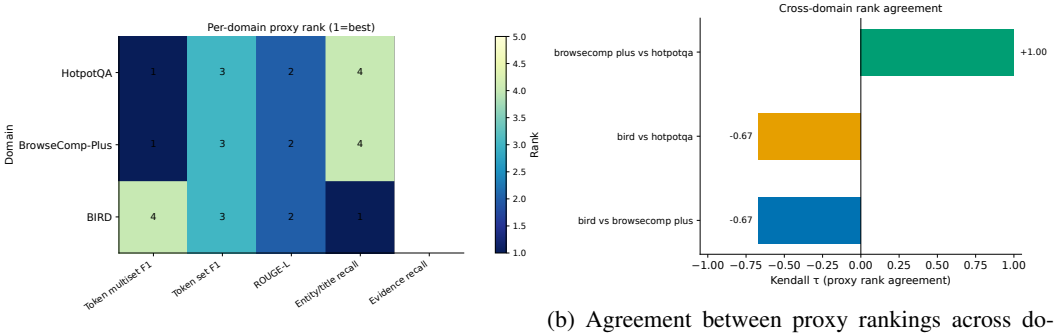


Figure G.2: HotpotQA acquisition curve with a fixed reader. The vertical axis reports the fraction of oracle headroom realized as retrieval recall changes.



(a) Proxy rankings within each audited domain.

(b) Agreement between proxy rankings across domains.

Figure G.3: Recovery proxy analysis.