ReBeCA: Unveiling Interpretable Behavior Hierarchy behind the Iterative Self-Reflection of Language Models with Causal Analysis

Anonymous ACL submission

Abstract

While self-reflection can enhance language model reliability, its underlying mechanisms remain opaque, with existing analyses often yielding correlation-based insights that fail to generalize. To address this, we introduce **ReBeCA** (self-**Re**flection **Be**havior explained through **C**ausal **A**nalysis), a framework that unveils the interpretable behavioral hierarchy governing the self-reflection outcome. By modeling self-reflection trajectories as causal graphs, ReBeCA isolates genuine determinants of performance through a three-stage Invariant Causal Prediction (ICP) pipeline. We establish three critical findings: (1) **Behavioral hierarchy:** Semantic behaviors of the model influence final self-reflection results hierarchically: directly or indirectly; (2) **Causation matters:** Generalizability in self-reflection effects is limited to just a few semantic behaviors; (3) **More $\neq$ better:** The confluence of seemingly positive semantic behaviors, even among direct causal factors, can impair the efficacy of self-reflection. ICP-based verification identifies sparse causal parents achieving up to 49.6% structural likelihood gains, stable across tasks where correlation-based patterns fail. Intervention studies on novel datasets confirm these causal relationships hold out-of-distribution ($p = .013, \eta_p^2 = .071$).

1 Introduction

1.1 Self-Reflection of Language Models

With the rapid development of large language models (LMs), prompt-driven pipelines have profoundly reshaped many aspects of modern society (OpenAI, 2023; Bai et al., 2023). Despite their strong generative capabilities, LMs still suffer from reliability challenges, including factual inaccuracies and limitations in faithful reasoning (Alkaissi and McFarlane, 2023; Zheng et al., 2023). To address these issues, self-reflection has been proposed as a potential mechanism that enables LMs to iteratively critique and revise their own outputs over multiple rounds, where earlier decisions may influence subsequent reflections and final outcomes, without external supervision. (Madaan et al., 2023; Yan et al., 2025). However, growing empirical evidence indicates that self-reflection is far from universally effective and may even degrade performance in certain settings (Huang et al., 2023).

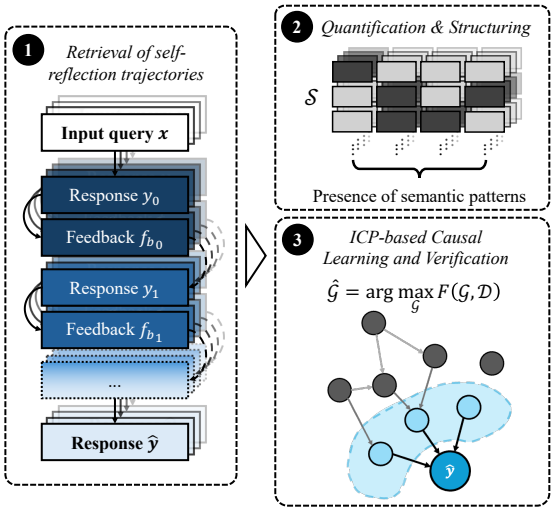

Figure 1: ReBeCA is a novel causal-driven framework for analyzing the self-reflection of language models. Drawing on vast collections of self-reflection trajectories, it reveals the underlying dynamics of the model's semantic behaviors and pinpoints the direct causal drivers of the self-reflection outcome.

Previous studies have reported a range of failure modes, including performance degradation on specific datasets (Li et al., 2024), oracle label leakage (Kim et al., 2023), and overly engineered prompts (Paul et al., 2023; Wang et al., 2023a). In response, several studies have proposed practical guidelines to improve self-reflection performance by adjusting design factors such as model size, the number of refinement iterations (Li et al., 2024), or by augmenting self-reflection with external tools or auxiliary feedback (Kamoi et al., 2024; Gou et al., 2023; Chen et al., 2023). While these heuristics yield sporadic improvements, they rely on trial-and-error and fail to generalize across different

tasks or models, leaving the underlying success or failure mechanisms opaque. Beyond performance-oriented studies, several works have begun to analyze self-reflection more directly. Some probe whether LMs genuinely exhibit self-correction abilities by modifying the design of self-reflection (Li et al., 2024; Xu et al., 2024), while others attempt to interpret reflective behavior through white-box analyses of internal representations or reasoning traces (Stechly et al., 2024). Although these analyses offer valuable perspectives, they tend to produce fragmented or task-specific observations. Taken together, most existing studies focus on isolated design factors or individual reflection steps, without modeling how reflective behavior evolves across each round. This gap limits the interpretability and reliability of existing findings.

Together, these challenges motivate the need for an analytical framework that treats self-reflection as a structured, multi-round trajectory and enables systematic analysis of how different reflective behaviors interact and contribute to final outcomes.

1.2 ReBeCA: The Causal-Driven Framework

Redesigning the analytical framework for self-reflection requires addressing two key challenges: **the reliability of collected trajectories** and **the validity of derived conclusions**. ReBeCA is designed to tackle both challenges within a unified framework.

A major obstacle in collecting data from LMs is their inherent *output stochasticity* (Wang et al., 2023b; Kuhn et al., 2023). Specific hyperparameters control the stochasticity, but manipulations are not always feasible (Laban et al., 2025). To mitigate the issue under a default setting, ReBeCA incorporates **Consistency-Enhanced Self-Refine (CESR)**, where every intermediate output in the self-reflection iteration undergoes multi-sampling followed by semantic equivalence clustering and centroid-based selection (Cheng et al., 2024). CESR improves the statistical significance of each self-reflection trajectory.

Beyond data reliability, ReBeCA focuses on deriving interpretable and ubiquitous conclusions about which factors genuinely influence the self-reflection outcome. In this context, "factors" correspond to observable self-reflection behaviors, operationalized via **semantic patterns** (see Appendix B.5 for detailed explanations). Consider "…the third part of the response should be more logical …" which instantiates the "Clarity and Organization" semantic pattern, denoting targeted refinement intent. Semantic patterns enable quantization of any text segment into a binary vector reflecting the presence/absence of each pattern, which supports computational analysis. "Genuinely influence" is delineated using *causal parents* from causality theory: with A $\rightarrow$ B $\rightarrow$ C, B genuinely influences C, whereas A–C forms a spurious correlation. Sec. 2.2 elaborates why causation beats correlation in analyzing self-reflection.

Accounting for spurious correlations proves essential in self-reflection analysis for one primary reason: Semantic patterns exhibit spatiotemporal structuring in self-reflection trajectories ("temporal" capturing round-to-round variance; "spatial" representing positional variance). Far from independent, semantic patterns interconnect via propagation chains, shared causes, and coupling effects, etc. Observational data alone breed abundant spurious correlations (Zhao et al., 2023). Semantic patterns linked merely by upstream ancestors or intermediary paths thus masquerade as pivotal drivers of the self-reflection outcome. This obfuscates true causal parents and warps interpretations of the genuine self-reflection dynamics. ReBeCA leverages causal discovery (CD), which provides the most straightforward solution for handling such data. It uniquely enables the detection of spurious correlations (Cai et al., 2023). The causal graph (CG) obtained via CD offers an intuitive depiction of relationships among all factors in LMs across the self-reflection trajectories, alongside the causal parents influencing the self-reflection outcome. Integrating causal associations with definitions of semantic patterns yields highly interpretable insights (Moraffah et al., 2020).

The reliability of CD outcomes hinges on validating the causal relationships. To the best of our knowledge, no prior research has established a ground-truth CG for self-reflection dynamics. This absence poses a substantial challenge. Guided by *causal invariance theory*: valid causal structures remain stable across distributions (Peters et al., 2016) we devised a three-phase causal parent selection and verification strategy based on *Invariant Causal Prediction (ICP)* (Peters et al., 2016; Gamella and Heinze-Deml, 2020), independent of any reference causal relationships. Applying ReBeCA to the Qwen3 family on mathematical reasoning and translation tasks, we uncover three key findings that

challenge prevailing assumptions. First, we reveal a **behavioral hierarchy**: semantic behaviors of LMs exhibit time-dependent self-reflection efficacy, with *Clarity and Organization* causally decisive only at $t = 2$ and *Specificity* only at $t = 4$, contradicting the notion that uniform criteria apply across all rounds. Second, we demonstrate that **correlations can be spurious**: only a sparse set of semantic behaviors genuinely affects the self-reflection outcome, grounding why heuristic prompt optimization often fails and shifting the paradigm from empirical "prompt engineering" to rigorous "causal behavior verification." Third, we discover that **stacking positives $\neq$ better outcomes**: while individual semantic behaviors among direct causal factors significantly improve outcomes, simultaneously activating multiple ones yields no cumulative gains and may even degrade performance, suggesting the non-additive character of the self-reflection dynamics. These insights position ReBeCA not just as a robust analytical tool, but as a foundational framework for distinguishing generalizable causal drivers from transient correlations.

2 Related Work

2.1 Unveiling the Mechanism of Self-Reflection

Inspired by self-reflection as a cognitive strategy for consolidating human experience, prior work has incorporated self-reflection into language models as a mechanism for reducing hallucinations and enabling agentic behaviors (Madaan et al., 2023; Shinn et al., 2023). However, subsequent empirical studies have shown that self-reflection is not universally reliable and may even degrade performance under certain conditions (Huang et al., 2023).

In response, recent work has begun to examine the behaviors that models exhibit during self-reflection, showing that certain semantic cues and revision patterns recur across reflection rounds and are correlated with the self-reflection outcome (Kamoi et al., 2024; Yan et al., 2025). Other studies investigate factors such as over-confidence, prompt sensitivity, or refinement depth, linking them to self-reflection effectiveness (Xu et al., 2024; Kirchhof et al., 2025; Huang et al., 2023). Although these analyzes provide useful empirical observations, they often focus on individual factors or isolated steps without modeling how behaviors evolve and interact across multiple rounds.

Several recent works attempt to address these limitations by examining self-reflection from a temporal or mechanistic perspective. Some modify the design or constraints of self-reflection procedures to probe whether language models genuinely exhibit self-correction capabilities (Li et al., 2024), while others adopt white-box analytical approaches, inspecting internal activations or parameter-level behaviors to explain self-reflection (Stechly et al., 2024). While informative, these approaches typically emphasize local mechanisms or latent features, which are difficult to interpret at the behavioral level and may not generalize across tasks or model settings. Overall, prior research provides valuable insights into self-reflection but largely lacks trajectory-level analyses that model self-reflection as a structured, multi-round process.

2.2 Causation vs. Correlation

Our finding of behavioral hierarchy in model self-reflection rooted from the hierarchy assumption, a prior drawn from relevant theoretical frameworks: understanding the outcomes of temporal trajectories requires accounting for the hierarchical structure of their underlying determinants (Pearl, 2009; Schölkopf et al., 2021). Correlation modeling falls short in representing this hierarchy, whereas causal inference excels. First, hierarchical variable relationships are inherently directed, e.g., A influences B, unidirectionally ($A \rightarrow B$ (Pearl, 2009). Causation represents this asymmetry, while correlation describes only a bidirectional linkage. Second, spurious correlations are prevalent in deep hierarchical representations and can only be mitigated through causal modeling (Ye et al., 2024). In a chain $A \rightarrow B \rightarrow C$, purely correlational methods conflate the indirect predictor A with the direct cause B. Causal modeling resolves this ambiguity by leveraging conditional independence to prune the spurious connection between A and C.

The sequential dynamics of self-reflection necessitate a hierarchical approach to factor analysis. Such kind of approaches enhances generalizability by transcending the limitations of correlation-based studies (Xu et al., 2024; Kamoi et al., 2024; Kirchhof et al., 2025), which often erroneously treat factors as isolated or parallel. A case in point is overconfidence: it may not directly trigger hallucinations but instead induces specific reasoning patterns that lead to errors (Kalai et al., 2025), a distinction missed by flat, parallel analyses. ReBeCA distinguishes itself by discovering and verifying

the analytical results without relying on ground truth. Instead, it employs an ICP-based framework to validate relationships through distributional invariance. A detailed justification for emphasizing generalizability over ground-truth causal graphs is provided in this paragraph.

3 Problem Formulation

Iterative Self-Reflection. The *Self-Refine* framework represents a typical self-reflection approach (Madaan et al., 2023) (see Appendix B.2 for more information). We quantify the outcome of self-reflection using a standard and intuitive formulation. Specifically, for a given input query x, we define the outcome as the degree of alignment between the final output $\hat{y}$, generated after the last iteration of self-reflection and refinement, and the ground truth y.

Semantic Patterns. Semantic patterns capture key semantic features in an LM's output, comparable to distinctive linguistic behaviors in human speech. This paper employs semantic patterns as a fixed-length binary (0-1) encoding for text, where 1 signals presence and 0 signals absence. Suppose N patterns are determined, a self-reflection trajectory with T rounds yields a binary matrix $\mathcal{S}^{T\times N} \sim \{0,\, 1\}^{T\times N}$.

Hypothesis $\mathrm{H_a}$. For model $\mathcal{M}_\theta$ and task set $\mathcal{C}$ (with $\mathcal{C}_1, \mathcal{C}_2, \ldots$ as its mutually exclusive subsets), there exists a collection of semantic patterns as causal parents to $\mathcal{M}_\theta$'s self-reflection outcome. These causal relationships hold stably across the task subsets $\mathcal{C}_1, \mathcal{C}_2, \ldots$.

The Core Objective. Identify a collection of semantic patterns capable of:

1. Maximizing the average structural log-likelihood (LL) across all task subsets: $\mathcal{C}_1, \mathcal{C}_2, \ldots$.

2. Ensuring the uniqueness of this collection: the LL maximization is not stochastic. In particular, across every task subset, the LL for the empty collection and the full collection must be strictly inferior to the LL of the selected one.

3. Satisfying the linear stability of causal effects: for every semantic pattern within the selected collection, the coefficient and residual capturing the linear link between it and the self-reflection outcome display no substantial variation over task subsets.

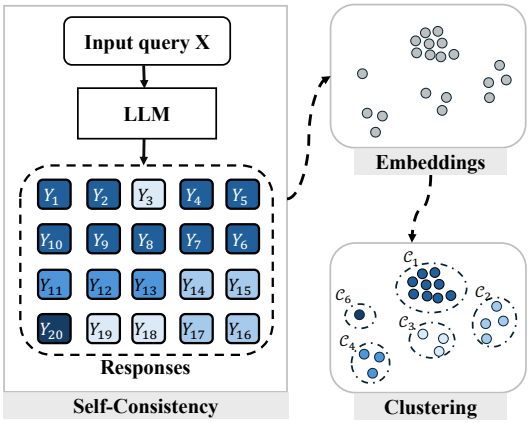

Figure 2: The flowchart of the Consistency-Enhanced Self-Refine (CESR).

These three points ensure causal invariance within a bounded scope (Peters et al., 2016), namely that these semantic patterns exert theoretically invariant influences on the model's self-reflection outcome across analogous tasks.

4 The ReBeCA Pipeline

4.1 Self-Reflection Trajectory Generation

To capture the temporal dynamics of self-reflection, we adopt an iterative prompting framework and choose Self-Refine as a representative instantiation (Madaan et al., 2023). While Reflexion is designed for memory-centric agents (Shinn et al., 2023), in conversational question answering without long-horizon memory requirements it follows the same *generate* → *feedback* → *refine* schema as Self-Refine. We therefore focus on Self-Refine for concreteness, without loss of generality.

A key challenge in analyzing self-reflection trajectories is the inherent stochasticity of language model outputs (Shorinwa et al., 2025). Evidence derived from a single trajectory is often unrepresentative, limiting the robustness of downstream analysis. Moreover, stochastic variation accumulates across multi-turn refinement rounds, leading to a rapidly expanding space of possible trajectories. Exhaustively enumerating all trajectories is infeasible in practice. Our remedy is a linear-time approximation, resampling-and-clustering over *feedback* and *refine* steps that yields a representative branch, maintaining statistical coverage without exponential overhead.

Building on this idea, we introduce Consistency-Enhanced Self-Refine pipeline as shown in fig 2.

Specifically, it yields a representative trajectory while maintaining computational efficiency. At each feedback and refine step, multiple samplings are performed and semantic clustering is applied to the resulting outputs in a shared semantic space, following prior work on semantic equivalence (Kuhn et al., 2023). Rather than selecting the most frequent output, we identify the largest cluster and select its centroid as the representative outcome. This procedure approximates the most typical semantic response under the model's output distribution, reducing sensitivity to outliers and low-probability variations (Yan et al., 2022).

4.2 Uncovering Determinants of the Self-Reflection Outcome

Semantic outputs from LMs serve as a key characterization of their behavior, while causality provides a robust foundational logic for interpreting behavioral patterns (Moriwaki et al., 2013; Zhu et al., 2024). Thus, it is intuitive to explain self-reflection mechanisms by modeling causal relationships over trajectories of these semantic outputs. Grounded in causality and its core methodology, causal discovery (CD), ReBeCA recovers a causal graph (CG) given the semantic outputs of self-reflection trajectories. Crucially, the determinants, or causal parents, of the self-reflection outcome are isolated from this graph, a step that is central to reaching our analytical conclusions. Realizing these objectives entails addressing two fundamental challenges: (i) quantifying and formatting unstructured semantic outputs into structured representations compatible with existing CD frameworks, and (ii) devising a method to verify the validity of the inferred causal parents without relying on ground-truth CGs. The solutions are unfolded as follows.

4.2.1 Quantification and Structuring

We leverage **semantic patterns** as an intuitive method to quantify and structure multi-turn dialogue trajectories. Consider a set of five semantic patterns, $\{\mathrm{SP}_0, \mathrm{SP}_1, \ldots, \mathrm{SP}_4\}$. When an output instance demonstrates patterns SP_0 and SP_2 but lacks SP_1, SP_3, and SP_4, it is encoded via the multi-hot encoding vector $[1, 0, 1, 0, 0]$. The choice of semantic patterns over embedding models for text encoding is motivated by two factors: interpretability and computational efficiency. First, unlike embeddings, semantic patterns provide transparent representations. Second, the dimensionality of these patterns is typically much lower than that of embeddings. Since CD algorithms frequently scale exponentially with dimensionality (Zanga et al., 2022), this reduction is critical for scalable and efficient computation.

The definition of semantic patterns demands a balance in occurrence frequency, as patterns that are either overly rare or excessively common lack the distinctiveness required for effective CD (Chiu et al., 2023). We employ a hybrid strategy, integrating GPT-5 generation with manual refinement to identify representative patterns. First, we leverage GPT-5 to mine a comprehensive list of candidate patterns from a random subset of our self-reflection trajectories. Subsequently, we manually refine the candidates by removing instances with frequencies under 5%. We also consolidate overlapping semantic patterns to ensure clarity. Appendix B.5 outlines the definitions of all semantic patterns considered for our case study. Semantic patterns structure the collection of self-reflection dialogues into a binary matrix $\mathcal{S}$, providing a rigorous basis for subsequent analysis.

4.2.2 Invariant Causal Prediction

ICP is a blanket term for causal validation predicated on the invariance assumption, ensuring the cross-domain generalizability of causal associations (Peters et al., 2016). ReBeCA uncovers and verifies the semantic patterns driving the self-reflection outcome, denoted hereafter as the *causal parents* or *causal parent sets*, via a three-stage ICP-based pipeline. We explicitly distinguish this approach from standalone CD because ground-truth labels for verification are unavailable in this domain. To address this challenge, the ICP-based pipeline not only yields candidate causal parent sets but also performs selection and self-verification without reliance on ground truth.

Stage 1: Selecting the Optimal Set of Causal Parents. ReBeCA implements **Ensemble CD** using *Greedy Equivalence Search* (GES, see Appendix B.3 for more information) (Chickering, 2002), a well-established family of score-based graph optimization algorithms with multiple choices for the score function.

The ensemble strategy counteracts the sensitivity of results to the specific score function, as different functions can produce divergent causal structures from identical data (Montagna et al., 2023). Its procedure involves three main steps. First, the self-reflection dataset is partitioned into K folds. Sec-

ond, GES is applied to every combination of the K training folds and Q score functions, generating up to $K \times Q$ unique candidate causal parent sets. Finally, we evaluate these candidates via Cross-Validation Log-Likelihood (CVLL, higher values indicate better performance) (Zhang et al., 2025). We compute the average structural LL for each candidate on the validation sets, typically utilizing the Bayesian Dirichlet sparse (BDs, see Appendix B.4 for more information) score for sparse data[1] (Scutari, 2016). The candidate set with the highest CVLL is selected, identifying the causal structure (the candidate causal parents $\rightarrow$ the self-reflection outcome $\hat{y}$) that exhibits the best generalizability across diverse data distributions.

Stage 2: Uniqueness Verification. The causal parent set derived in Stage 1 should be the unique configuration that maximizes the CVLL, not merely an arbitrary candidate. Uniqueness is essential, as the number of independent candidates returned by standard score functions and CD algorithms is typically far smaller than the power set of semantic patterns. The empty set $\text{Pa}_{\hat{y}} = \emptyset$ and the full set $\text{Pa}_{\hat{y}} = \Omega$ warrant specific attention as the sparsest and densest possible configurations. To mitigate computational costs, we verify uniqueness approximately by demonstrating that the identified causal parents generalize better across distributions than these two cases (Zanga et al., 2022). In particular, we compute the CVLL for mappings from the empty and full sets to $\hat{y}$ and compare the results against the optimal CVLL from Stage 1, serving as a proxy validation for structural uniqueness.

Stage 3: Linear Stability Verification. Linear stability is grounded in the assumption of a linear mapping from causal parents to $\hat{y}$. Independent of the strict validity of the linearity assumption, a key advantage of this formulation is the ability to intuitively quantify generalizability through the variance of regression coefficients and residuals across distribution shifts (Peters et al., 2016; Arjovsky et al., 2019). This offers a more granular verification of generalizability than Stages 1 and 2. While maximizing CVLL optimizes average performance, it does not ensure that the influence of each specific causal parent is consistent across data. Stage 3 addresses this limitation by enforcing effect stability at the level of individual parents, thereby establishing a robust final barrier against spurious correlations.

We underscore the importance of causal generalizability in ReBeCA, driven less by the lack of ground truth and more by the fact that generalizability supersedes accuracy in our framework. The inherent heterogeneity and uncertainty of self-reflection trajectories render high-accuracy claims susceptible to spurious correlations (Lu et al., 2025). Therefore, we anchor our analysis in causal invariance to derive universally applicable insights, distinguishing our findings from purely empirical results. This critical rationale explicitly underpins our adoption of the ICP-based workflow.

5 Case Study on the Qwen3 Model Family

5.1 Analytical Outcomes

Using the ~8B parameter Qwen3 models (Bai et al., 2023), we curated a structured dataset of self-reflection trajectories based on two low-leakage benchmarks: MATH (logical reasoning) and BOUQuET (multilingual capacity) (Hendrycks et al., 2021; Andrews et al., 2025). Table 3 entails the configurations of the case study. The three-stage ICP-based framework was utilized to derive the optimal causal parent set of the self-reflection outcome, conducted on the dataset, with configurations of the study outlined in Appendix A.

5.1.1 Stage 1: Selecting the Optimal Set of Causal Parents

The structured data were split into five cross-validation folds, each comprising training and testing subsets. On each training fold, Ensemble CD applied four score functions to generate diverse candidate parent sets. Testing folds then assessed these sets using BDs-based CVLL. Temporal constraints ensured acyclicity across time steps (Ling, 1982): SPs at round $t+1$ cannot causally precede those at round t. These are supplied to CD as prior constraints through adjacency matrix masking, a feature natively supported by Causal-Learn (Zheng et al., 2024). This is the open-source CD toolkit employed in our work.

Across all training folds in the MATH dataset, Ensemble CD detected 12 independent candidate causal parent sets converging on the outcome. For BOUQuET training folds, the method yielded six candidate parent sets each for the COMET and MetricX outcomes. Table 4 reports the CVLL for each candidate-to-$\hat{y}$ prediction over the five testing folds. This metric is the arithmetic mean of the

[1]The term "sparse" is grounded in empirical observations of $\mathcal{S}$ derived from our case study.

structural LL computed across all testing folds. In self-reflection trajectories derived from the MATH benchmark, emphasis on "Clarity and Organization (CO)" during round $t = 2$ and on "Specificity (Sp)" during round $t = 4$ was identified as the strongest causal influencers ($\mathcal{L}^*_{\text{BDs}} = -0.434 \pm 0.563$) shaping the self-reflection outcome. On BOUQuET trajectories, "Clarity and Organization (CO)" from round $t = 3$ qualified as the optimal causal parent for both $\hat{y}$ indicators ($\mathcal{L}^*_{\text{BDs}|\text{COMET}} = -0.978 \pm 0.495$, $\mathcal{L}^*_{\text{BDs}|\text{MetricX}} = -1.003 \pm 0.536$). "Clarity and Organization (CO)" emerged most often as the top causal parent among these factors, indicating that the causal edge to the outcome exhibits robust structural generalizability over all testing folds.

5.1.2 Stage 2: Uniqueness Verification

To approximate a uniqueness check on the optimal causal parents $\text{Pa}^*_{\hat{y}}$, two baseline conditions were verified. Specifically, the CVLL of $\emptyset \to \hat{y}$ should fall below that of $\text{Pa}^*_{\hat{y}} \to \hat{y}$. The same holds for the full-set predictor $\Omega \to \hat{y}$.

Table 1 and Figure 3 substantiate the approximate uniqueness of the selected $\text{Pa}^*_{\hat{y}}$s:

1. **MATH:** Relative to $\emptyset \to \hat{y}$, $\text{Pa}^*_{\hat{y}} \to \hat{y}$ yields a CVLL gain of 22.36%; relative to $\Omega \to \hat{y}$, it yields a gain of 49.59%.

2. **BOUQuET (COMET):** Relative to $\emptyset \to \hat{y}$, $\text{Pa}^*_{\hat{y}} \to \hat{y}$ yields a CVLL gain of 6.32%; relative to $\Omega \to \hat{y}$, it yields a gain of 15.10%.

3. **BOUQuET (MetricX):** Relative to $\emptyset \to \hat{y}$, $\text{Pa}^*_{\hat{y}} \to \hat{y}$ yields a CVLL gain of 3.93%; relative to $\Omega \to \hat{y}$, it yields a gain of 12.09%.

The results consistently reveal uniqueness in the Stage 1-selected $\text{Pa}^*_{\hat{y}}$ groups. Such uniqueness is markedly reliable ($< 5\%$) in most instances.

5.1.3 Stage 3: Linear Stability Verification

A causal relationship under the linear assumption is formalized in the following form:

$$y \sim \mathbf{k} \cdot \text{Pa}_y^\top + b \tag{1}$$

$\mathbf{k}$ denotes the coefficient vector, and b the residual term. For causal relationships to generalize across data distributions, the linear coefficients and residuals must exhibit distributional stability. This stability ensures that every variable in a causal parent set has a consistently stable influence on y,

Table 1: Stage 2 outputs: uniqueness validation results comparing CVLLs computed using the empty set, the full set, and optimal sets identified in Stage 1.

Configuration	MATH	BOUQuET (COMET)	BOUQuET (MetricX)
Empty set	-0.559 ± 0.486	-1.044 ± 0.337	-1.044 ± 0.342
Optimal set	$\mathbf{-0.434 \pm 0.563}$	$\mathbf{-0.978 \pm 0.495}$	$\mathbf{-1.003 \pm 0.536}$
Full set	-0.861 ± 1.227	-1.152 ± 0.784	-1.141 ± 0.763

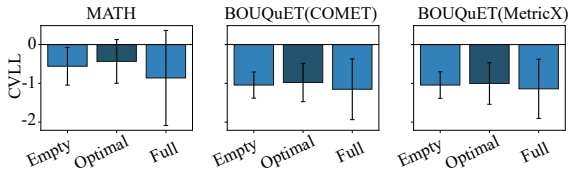

Figure 3: Table 1 illustrated as a bar chart.

offering a more robust evaluation of causal generalizability compared to Stage 2.

Levene's Test assesses homogeneity of variance (Levene, 1960), while One-Way ANOVA examines mean differences (Fisher and Fisher, 1971). Together, they evaluate the linear stability of $\text{Pa}^*_{\hat{y}} \to \hat{y}$. The data undergo re-stratification into 20 mutually exclusive folds for these analyses, organized into 4 balanced groups (each containing 5 folds) to support group-wise significance inference. We specify the common null hypothesis employed in both tests,

Hypothesis $\mathbf{H_0}$. Linear coefficients and residuals exhibit no significant differences (in mean or variance) across groups.

$\mathbf{H_0}$ is supported for both tests only if the following hold together: $p > .05$ and $\eta^2_{\text{p}} < .06$ (a small effect size), validating the linear stability. Table 5 presents the outcomes. Except for a single moderate effect, the results confirm that all causal relationships exhibit linear stability across data groups.

5.1.4 Debriefing

Thus far, we arrive at the following insights across observable samples: semantic patterns from a LM do not invariably determine the self-reflection outcome at every iteration. Direct-influencing patterns, in the meantime, reliably predict outcomes consistently for the same task across input variations. Importantly, the structured dataset incorporates and randomizes the complete set of self-reflection trajectories from the four Qwen3 models. Although sourced from shared datasets, these trajectories reflect queries spanning diverse difficulties and domains, without differentiation during validation.

Intervention	0.6B	1.7B	4B	8B
None	1	12	19	21
$\mathrm{CO}_{t=2}$ *excl.*	1	14	26	23
$\mathrm{Sp}_{t=4}$ *excl.*	1	10	32	21
$\mathrm{CO}_{t=2}$ & $\mathrm{Sp}_{t=4}$ *excl.*	0	9	19	23

Table 2: Intervention results on 50 AIME problems across four LMs, reported as the count of correct final answers. **None**: the unmodified CESR pipeline with five rounds of self-reflection and refinement. $\mathbf{SP_t}$ ***excl.***: A controlled variant where the t-th round is guided to emphasize the semantic pattern SP during self-reflection and refinement.

5.2 Intervention Study

While the ICP-based validations provide essential observational evidence, its assessments of generalizability are bound by the distribution of observed, non-manipulable data (Li et al., 2022). Intervenability allows for stronger validation of factor effects, underpinned by Randomized Controlled Trial (RCT) theory (Pearl, 2010). This constitutes the primary objective of our intervention study. Specifically, it produces RCTs through prompt-based constraints on selected semantic patterns in designated Self-Refine iterations. These trials draw from the AIME dataset (AI-MO, 2024), a higher-difficulty counterpart to the MATH dataset in structure but with distinct problems[2]. Central to this approach is the intervention hypothesis: *should a semantic pattern prove a legitimate factor of the self-reflection outcome, directing the model to emphasize it during reflection and refinement will demonstrably alter the outcome.* For MATH trajectories, the optimal causal parents comprise two semantic patterns: $\mathrm{CO}_{t=2}$ and $\mathrm{Sp}_{t=4}$. The intervention study will examine whether a 2×2 factorial intervention[3] on these patterns, applied to the AIME dataset, produces four distinct outcome trials with statistically significant mutual differences.

Hypothesis $\mathbf{H_0^i}$. All intervention groups demonstrate no significant divergence in the self-reflection outcome.

Hypothesis $\mathbf{H_a^i}$. All intervention groups demonstrate significant divergence in the self-reflection outcome.

Cochran's Q test was employed to evaluate $\mathbf{H_a^i}$ (Cochran, 1950). It assesses consistency for binary variables across matched samples. We consider $\mathbf{H_a^i}$ supported if the results show statistical significance ($p < .05$) and a large effect size ($\eta_p^2 \geq .06$). Based on 50 randomly selected AIME problems (–2024), the outcomes are summarized in Table 2. The corresponding Cochran's Q test produced $p = .013$ and $\eta_p^2 = .071$, supporting $\mathbf{H_a^i}$. Additionally, results indicate that targeted prompting for reflection and optimization focusing on Clarity and Organization at round $t = 2$ or Specificity at round $t = 4$ achieves the highest average self-reflection performance on the selected dataset, with 16 correct responses per model on average. Joint optimization of these two semantic patterns, however, offers no benefit and marginally underperforms the non-intervened CESR (None) across models.

6 Conclusion

In this work, we introduce ReBeCA, a framework that leverages causal analysis to explain the mechanisms behind self-reflection. Our investigation yields three critical insights into how language models refine their outputs: **1) A dynamic behavioral hierarchy:** We find that self-reflection is not a uniform process; instead, the impact of semantic behaviors is strictly time-dependent. For instance, mathematical reasoning dominates in early stages ($t = 2$), while translation nuances become crucial only in later stages ($t = 3$ on BOUQuET); **2) Correlations are often spurious:** Many behaviors that seemingly correlate with success are merely unstable associations. By filtering out these spurious patterns, ReBeCA identifies a sparse set of genuine causal drivers, achieving significantly higher likelihood ($+49.6\%$ on MATH and $+15.1\%$ on BOUQuET) compared to dense baselines; **3) Stacking positives $\neq$ better outcomes:** Our intervention study reveals a critical non-additivity in reflection dynamics. While individual causal factors successfully generalized to out-of-distribution datasets ($p = .013$), activating multiple positive factors simultaneously failed to yield cumulative gains due to complex interference effects.

ReBeCA moves beyond trial-and-error heuristics, offering a rigorous blueprint for interpreting and optimizing language agent behaviors.

[2]We focus exclusively on reporting and dissecting findings from math-reasoning intervention experiments here. Machine translation is set aside. Key rationales include overlapping validation procedures in both setups and MATH trajectories' association with 2 direct causes, offering broader applicability to intervention analyses versus BOUQuET's lone cause.

[3]With two possible intervention states for each semantic pattern (activation or deactivation), two patterns jointly define a 2×2 matrix of intervention combinations.

Limitations

The present paradigm of ReBeCA has certain limitations. Firstly, the consistency enhancement strategy detailed in Sec. 4.1 applies a fixed resampling rate (20 iterations) to each intermediate output. This static threshold may fail to adequately capture the semantic uncertainty and output distribution across all LMs, hyperparameters, and inputs. As a result, certain model outputs may retain bias, potentially exerting a marginal effect on the generalizability of our findings. To the best of our knowledge, no existing research offers a robust method for determining this optimal resampling count under arbitrary conditions, and establishing such a method lies outside the scope of the present study.

Secondly, the semantic cues used for causal graph construction in this study are manually defined. While this ensures interpretability, it may not be the most efficient strategy and risks overlooking critical yet underexplored candidate variables. One alternative approach combines Bag-of-Words representation with unsupervised word embedding clustering followed by class definition. However, this method presents notable limitations: it may fail to capture latent semantic structures, and clustering often yields overly numerous or fragmented categories. Our literature review has not identified superior methods for automating semantic cue discovery while maintaining interpretability and coverage. Addressing this limitation constitutes a valuable avenue for future research.

Finally, although the direct causal parent set identified by ReBeCA has been verified through both ICP-based multi-step procedures and interventional experiments, achieving robust validation ultimately requires reliable ground truths, which are currently non-existent within the scope of self-reflection. Addressing these gaps represents another potential direction for our future endeavors.

Ethical Considerations

This study was designed and conducted in full accordance with the ACL Code of Ethics.

This study involves a limited analysis and interpretation of machine behavior, and we are fully aware of the associated ethical considerations. We cannot guarantee that the conclusions presented herein, such as the results of causal discovery, generalize to all scenarios, whether within or beyond the settings of our case studies. For the scenarios covered by our case studies, we have striven to minimize the risk of spurious conclusions through multi-step, comprehensive invariance verification and intervention experiments. However, we acknowledge the inherent uncertainty of LMs, as well as the intrinsic limitations of the causal discovery algorithms and statistical tests employed. We urge readers to interpret and adopt the findings of this paper with caution.

Additionally, this study incorporates limited human participation in the form of annotation and quality checking of model outputs. The annotation process does not constitute any human subjects study, and annotators were not exposed to personal or sensitive information. All data used in our experiments are either publicly available benchmarks or model-generated outputs and do not contain personal information. We provide detailed descriptions of the pipeline and experimental settings to support reproducibility, while results may depend on the stochastic behavior of proprietary language models.

References

AI-MO. 2024. aimo-validation-aime. https://huggingface.co/datasets/AI-MO/aimo-validation-aime.

Hussam Alkaissi and Samy I McFarlane. 2023. Artificial hallucinations in chatgpt: implications in scientific writing. *Cureus*, 15(2).

Pierre Andrews, Mikel Artetxe, Mariano Coria Meglioli, Marta R Costa-jussà, Joe Chuang, David Dale, Mark Duppenthaler, Nathanial Paul Ekberg, Cynthia Gao, Daniel Edward Licht, et al. 2025. Bouquet: dataset, benchmark and open initiative for universal quality evaluation in translation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27503–27523.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Levin Brinkmann, Fabian Baumann, Jean-François Bonnefon, Maxime Derex, Thomas F Müller, Anne-Marie Nussberger, Agnieszka Czaplicka, Alberto Acerbi, Thomas L Griffiths, Joseph Henrich, et al. 2023. Machine culture. *Nature Human Behaviour*, 7(11):1855–1868.

Hengrui Cai, Yixin Wang, Michael Jordan, and Rui Song. 2023. On learning necessary and sufficient causal graphs. *Advances in Neural Information Processing Systems*, 36:42148–42160.

Xinyun Chen, Maxwell Lin, Nathanael Schärli, and Denny Zhou. 2023. Teaching large language models to self-debug. *arXiv preprint arXiv:2304.05128*.

Furui Cheng, Vilém Zouhar, Simran Arora, Mrinmaya Sachan, Hendrik Strobelt, and Mennatallah El-Assady. 2024. Relic: Investigating large language model responses using self-consistency. In *Proceedings of the 2024 CHI conference on human factors in computing systems*, pages 1–18.

David Maxwell Chickering. 2002. Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Max Chickering. 2020. Statistically efficient greedy equivalence search. In *Conference on Uncertainty in Artificial Intelligence*, pages 241–249. Pmlr.

Chih-Yuan Chiu, Kshitij Kulkarni, and Shankar Sastry. 2023. Towards dynamic causal discovery with rare events: A nonparametric conditional independence test. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pages 7610–7616. IEEE.

William G Cochran. 1950. The comparison of percentages in matched samples. *Biometrika*, 37(3/4):256–266.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. 2022. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158.

Ronald Aylmer Fisher and Ronald A Fisher. 1971. *The design of experiments*. Springer.

Juan L Gamella and Christina Heinze-Deml. 2020. Active invariant causal prediction: Experiment selection through stability. *Advances in Neural Information Processing Systems*, 33:15464–15475.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. 2023. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*.

Qiyu Hu, Guangyi Zhang, Zhijin Qin, Yunlong Cai, Guanding Yu, and Geoffrey Ye Li. 2023. Robust semantic communications with masked vq-vae enabled codebook. *IEEE Transactions on Wireless Communications*, 22(12):8707–8722.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Adam Tauman Kalai, Ofir Nachum, Santosh S Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Geunwoo Kim, Pierre Baldi, and Stephen McAleer. 2023. Language models can solve computer tasks. *Advances in Neural Information Processing Systems*, 36:39648–39677.

Michael Kirchhof, Luca Füger, Adam Goliński, Eeshan Gunesh Dhekane, Arno Blaas, and Sinead Williamson. 2025. Self-reflective uncertainties: Do llms know their internal answer distribution? *arXiv preprint arXiv:2505.20295*.

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

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. Llms get lost in multi-turn conversation. *arXiv preprint arXiv:2505.06120*.

Howard Levene. 1960. Robust tests for equality of variances. *Contributions to probability and statistics*, pages 278–292.

Bo Li, Yifei Shen, Yezhen Wang, Wenzhen Zhu, Dongsheng Li, Kurt Keutzer, and Han Zhao. 2022. Invariant information bottleneck for domain generalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 7399–7407.

Yanhong Li, Chenghao Yang, and Allyson Ettinger. 2024. When hindsight is not 20/20: Testing limits on reflective thinking in large language models. *arXiv preprint arXiv:2404.09129*.

Robert F Ling. 1982. Correlation and causation.

Haolang Lu, Yilian Liu, Jingxin Xu, Guoshun Nan, Yuanlong Yu, Zhican Chen, and Kun Wang. 2025. Auditing meta-cognitive hallucinations in reasoning large language models. *arXiv preprint arXiv:2505.13143*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Francesco Montagna, Atalanti Mastakouri, Elias Eulig, Nicoletta Noceti, Lorenzo Rosasco, Dominik Janzing, Bryon Aragam, and Francesco Locatello. 2023. Assumption violations in causal discovery and the robustness of score matching. *Advances in Neural Information Processing Systems*, 36:47339–47378.

Raha Moraffah, Mansooreh Karami, Ruocheng Guo, Adrienne Raglin, and Huan Liu. 2020. Causal interpretability for machine learning-problems, methods and evaluation. *ACM SIGKDD Explorations Newsletter*, 22(1):18–33.

Norihiko Moriwaki, Ken-ichiro Nomura, and Dai Senoo. 2013. A behavior-based approach for analyzing knowledge-process dynamics. *Journal of service science and management*, 6(2):160–169.

R OpenAI. 2023. Gpt-4 technical report. arxiv 2303.08774. *View in Article*, 2(5):1.

Debjit Paul, Mete Ismayilzada, Maxime Peyrard, Beatriz Borges, Antoine Bosselut, Robert West, and Boi Faltings. 2023. Refiner: Reasoning feedback on intermediate representations. *arXiv preprint arXiv:2304.01904*.

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

Judea Pearl. 2010. An introduction to causal inference. *The international journal of biostatistics*, 6(2):7.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012.

Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W Crandall, Nicholas A Christakis, Iain D Couzin, Matthew O Jackson, et al. 2019. Machine behaviour. *Nature*, 568(7753):477–486.

Ricardo Rei, Ana C Farinha, José G. C. de Souza, Craig Stewart, Taisiya Glushkova, Alisa Tsvigun, Chris Hokamp, and Alon Lavie. 2022. Metricx-22: A new generation of quality metrics for machine translation. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 706–715.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7665–7678.

Raanan Y Rohekar, Yaniv Gurwicz, and Shami Nisimov. 2023. Causal interpretation of self-attention in pre-trained transformers. *Advances in Neural Information Processing Systems*, 36:31450–31465.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. 2021. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634.

Marco Scutari. 2016. An empirical-bayes score for discrete bayesian networks. In *Conference on probabilistic graphical models*, pages 438–448. PMLR.

Marco Scutari. 2018. Dirichlet bayesian network scores and the maximum relative entropy principle. *Behaviormetrika*, 45(2):337–362.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. 2025. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *ACM Computing Surveys*.

Peter Spirtes, Clark Glymour, and Richard Scheines. 2001. Causation, prediction, and search. *MIT Press Books*, 1.

Kaya Stechly, Karthik Valmeekam, and Subbarao Kambhampati. 2024. Chain of thoughtlessness? an analysis of cot in planning. *Advances in Neural Information Processing Systems*, 37:29106–29141.

Tianlu Wang, Ping Yu, Xiaoqing Ellen Tan, Sean O'Brien, Ramakanth Pasunuru, Jane Dwivedi-Yu, Olga Golovneva, Luke Zettlemoyer, Maryam Fazel-Zarandi, and Asli Celikyilmaz. 2023a. Shepherd: A critic for language model generation. *arXiv preprint arXiv:2308.04592*.

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

Huiqiang Xie, Zhijin Qin, Geoffrey Ye Li, and Biing-Hwang Juang. 2021. Deep learning enabled semantic communication systems. *IEEE transactions on signal processing*, 69:2663–2675.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. 2024. Pride and prejudice: Llm amplifies self-bias in self-refinement. *arXiv preprint arXiv:2402.11436*.

Tianqiang Yan, Ziqiao Lin, Lin Zhang, Zhenglong Sun, and Yuan Gao. 2025. Entrospect: Information-theoretic self-reflection elicits better response refinement of small language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 24563–24577.

Xuyang Yan, Shabnam Nazmi, Biniam Gebru, Mohd Anwar, Abdollah Homaifar, Mrinmoy Sarkar, and Kishor Datta Gupta. 2022. A clustering-based active learning method to query informative and representative samples. *Applied Intelligence*, 52(11):13250–13267.

Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Alessio Zanga, Elif Ozkirimli, and Fabio Stella. 2022. A survey on causal discovery: theory and practice. *International Journal of Approximate Reasoning*, 151:101–129.

Yuelei Zhang, Qingcui Li, Jiachen Wang, Xiao Chang, Luonan Chen, and Xiaoping Liu. 2025. Causal network inference based on cross-validation predictability. *Communications Physics*, 8(1):173.

Yu Zhao, Pan Deng, Junting Liu, Xiaofeng Jia, and Mulan Wang. 2023. Causal conditional hidden markov model for multimodal traffic prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 4929–4936.

Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023. Why does chatgpt fall short in providing truthful answers? *arXiv preprint arXiv:2304.10513*.

Yujia Zheng, Biwei Huang, Wei Chen, Joseph Ramsey, Mingming Gong, Ruichu Cai, Shohei Shimizu, Peter Spirtes, and Kun Zhang. 2024. Causal-learn: Causal discovery in python. *Journal of Machine Learning Research*, 25(60):1–8.

Sujia Zhu, Guodao Sun, Yue Shen, Zihao Zhu, Wang Xia, Baofeng Chang, Jingwei Tang, and Ronghua Liang. 2024. Vac 2: visual analysis of combined causality in event sequences. *IEEE Transactions on Visualization and Computer Graphics*.

A Configurations of the Case Study

This section details the supplementary information about the configurations of the case study (Sec. 5).

Model family. The 0.6B, 1.7B, 4B, and 8B Qwen3 LMs were designated as the target models for the case study. We prioritize this selection for several reasons: 1) a lack of rigorous analysis for LMs under 10B parameters in existing literature (Yan et al., 2025), 2) a unified model family facilitates strict variable control, and 3) the Qwen3 family offers the most extensive range of variants within this parameter spectrum.

Table 3: Overview of base datasets and LM configurations for the present case study.

Component	Item	Configuration Details & Stratification
Base Datasets	MATH	• **Difficulty Levels:** Level 1 to 5 • **Topics:** 7 distinct mathematical fields
	BOUQUET	• **Granularity:** Paragraphs and Sentences • **Content Type:** Conversation and Web Misc. • **Languages:** 5 distinct languages
Models	Qwen Model Family (4-bit Quantized)	• Qwen3-0.6B-bnb-4bit • Qwen3-1.7B-bnb-4bit • Qwen3-4B-bnb-4bit • Qwen3-8B-bnb-4bit

Base datasets. To generate self-reflection trajectories, we employ two base datasets: MATH and BOUQuET (Hendrycks et al., 2021; Andrews et al., 2025). MATH is a mathematical reasoning benchmark characterized by moderate difficulty and minimal data leakage or contamination. For this dataset, self-reflection outcomes are binary, where 0 and 1 indicate incorrect and correct answers, respectively. BOUQuET is a recent multilingual-to-English translation dataset that similarly offers the advantage of low leakage and contamination. The outcomes for BOUQuET are derived from two specific metric models, namely MetricX (the lower $\downarrow$ the better) and COMET (the higher $\uparrow$ the better) (Rei et al., 2020, 2022), benchmarked against the initial translation. Should the final Self-Refine output elevate any given metric, the corresponding $\hat{y}$ is labeled as 1. Otherwise, the label is 0. We generate trajectories for each model using a stratified subset of 100 samples from each dataset.

B Additional Related Work & Explanations

B.1 Related Work: Machine Behavior

As intelligent systems are increasingly deployed in society, it has become insufficient to understand them solely through implementation details or performance metrics. Prior work on Machine Behavior (Rahwan et al., 2019) argues that intelligent systems exhibit observable behavioral patterns when operating in socio-technical environments, arising from both algorithmic mechanisms and interactions with humans and other systems. However, characterizing such behaviors alone is often not enough to explain their broader downstream effects. Relatedly, Machine Culture (Brinkmann et al., 2023) points out that when machines generate content

and influence information exposure, they participate in processes of cultural variation, transmission, and selection, leading to cultural outcomes shaped jointly by humans and machines.

An intermediate aspect between machine behavior and cultural outcomes is human self-reflection. Humans do not simply accept machine behavior or machine-generated outputs; instead, they interpret and evaluate them in terms of reliability, fairness, and usefulness, and adjust their own behavior accordingly (e.g., whether to trust or adopt the output). Through this process of self-reflection, machine behavior does not translate directly into cultural outcomes, but is mediated by reflective judgments and subsequent behavioral choices. While this perspective suggests a connection between machine behavior, self-reflection, and downstream outcomes, the concrete behavioral mechanisms involved remain unclear.

Recent language models have begun to incorporate self-reflection mechanisms inspired by this human process, typically by prompting models to examine, revise, or critique their own outputs. Although such model-level self-reflection does not replicate human reflection, it provides a practical setting in which reflective processes can be operationalized and studied as part of machine behavior. In this context, ReBeCA focuses on identifying which behaviors within self-reflection causally drive performance changes in language models. By modeling self-reflection as a set of behaviors and analyzing them using causal methods, ReBeCA provides a concrete and testable approach for studying self-reflection in LLMs as a form of machine behavior. While this work is limited to the language model setting and does not attempt to characterize broader cultural feedback loops, it takes a first step toward grounding self-reflection in a causal behavioral framework.

B.2 The Formulation of Self-Refine

Self-Refine formalizes the procedure for an LM, $\mathcal{M}_\theta$, as a trajectory of iterative reflection and refinement (referred in Sec. 3):

$$y_{t+1} = \mathcal{M}_\theta \left(p_{\text{refine}} \| x \| y_0 \| f_{b_0} \| \ldots \| y_t \| f_{b_t}\right) \quad (2)$$

In (2), $f_{b_t} = \mathcal{M}_\theta \left(p_{\text{fb}} \| x \| y_t\right)$. p_{fb} denotes the prompt used to elicit the model's self-reflection. The term f_{b_t} represents the resulting feedback on the model's response at round t. This feedback might include, for instance, a list of logical fallacies or suggestions for revision. y_{t+1} is the refined response based on y_t, incorporating both the feedback f_{b_t} and the historical context.

B.3 Greedy Equivalence Search

The optimization procedure of GES is formulated in (3) (referred in Sec. 4.2.2):

$$\begin{aligned} \text{Forward (FES):} \quad & \mathcal{G}_{k+1} = \underset{\mathcal{G}' \in \text{Insert}(\mathcal{G}_k)}{\arg\max} \ \mathcal{F}\left(D, \mathcal{G}'\right) \\ & \text{s.t.} \quad \mathcal{F}\left(D, \mathcal{G}'\right) > \mathcal{F}\left(D, \mathcal{G}_k\right) \\ \text{Backward (BES):} \quad & \mathcal{G}_{k+1} = \underset{\mathcal{G}' \in \text{Delete}(\mathcal{G}_k)}{\arg\max} \ \mathcal{F}\left(D, \mathcal{G}'\right) \\ & \text{s.t.} \quad \mathcal{F}\left(D, \mathcal{G}'\right) > \mathcal{F}\left(D, \mathcal{G}_k\right) \end{aligned} \quad (3)$$

$\mathcal{G}_k$ denotes the essential graph (CPDAG[4]) at step k, and $\mathcal{F}\left(D, \mathcal{G}\right)$ represents the score of the equivalence class $\mathcal{G}$ given dataset D. The operators Insert$(\cdot)$ and Delete$(\cdot)$ generate the set of valid neighboring equivalence classes obtainable by adding or removing a single edge respectively, guiding the greedy maximization of the score until convergence (Chickering, 2020).

The application of GES to model self-reflection trajectories is theoretically grounded in the inherent properties of the language modeling process, satisfying the core assumptions of score-based CD (Spirtes et al., 2001; Pearl, 2009). First, the *Acyclicity/Markov Assumption* is naturally enforced by the sequential structure of multi-turn reflection. Since the model's state at iteration t is strictly conditioned on the history $H_{<t}$, the resulting causal graph is inherently a DAG, eliminating the possibility of feedback loops violating the temporal order. Second, regarding *Sufficiency Assumption*, while LLMs operate on high-dimensional latent states, recent work on causal abstractions suggests that high-level model behaviors can be faithfully represented by discrete causal variables (Geiger et al., 2021; Rohekar et al., 2023). By mapping continuous trajectories into specific semantic patterns (as detailed in Sec. 4.2.1), ReBeCA constructs a closed-system approximation where the distinct "behavioral drivers" are observed. Finally, the *Faithfulness Assumption* is supported by the structural sensitivity of LMs to prompt variations, where specific semantic components (e.g., error detection) exert non-degenerate, recoverable probabilistic influences on the output (Feder et al., 2022), allowing GES to reliably identify the equivalence class of the underlying mechanism.

[4] A CPDAG (Completed Partially Directed Acyclic Graph), also known as an essential graph, is a DAG that incorporates both directed edges and undirected edges.

B.4 Bayesian-Dirichlet sparse Score

The BDs formula (referred in Sec. 4.2.2) for computing the LL: $\mathcal{L}_{\text{BDs}}(\hat{y} \mid \Pi_{\hat{y}})$, is given as follows (Scutari, 2016):

$$\sum_{j:N_j>0} \left[\log \frac{\Gamma\left(\frac{\alpha}{\tilde{q}}\right)}{\Gamma\left(\frac{\alpha}{\tilde{q}} + N_j\right)} + \sum_{k=1}^{r_{\hat{y}}} \log \frac{\Gamma\left(\frac{\alpha}{\tilde{q} r_{\hat{y}}} + N_{jk}\right)}{\Gamma\left(\frac{\alpha}{\tilde{q} r_{\hat{y}}}\right)} \right] \tag{4}$$

In (4), α is the equivalent sample size[5], $r_{\hat{y}}$ the cardinality of $\hat{y}$, and N_{jk} the count of samples where $\hat{y}$ is in state k given the j-th parent configuration (with marginals $N_j = \sum_k N_{jk}$). Crucially, BDs sets the prior scaling factor $\tilde{q} = |\mathcal{J}^+|$ to the count of observed parent configurations (where $N_j > 0$), rather than all theoretically possible ones. This formulation prevents prior dilution in sparse regimes, ensuring the score remains a robust metric for structural verification despite the exponential growth of the state space (Scutari, 2018).

B.5 Semantic Patterns: The Atomic Units of Interpretable Behavior

This additional section offers essential elaborations on key elements recurring across the paper: semantic patterns (referred in Sec. 1.2).

Consider *semantic primitives* from the framework of semantic communication systems: these are atomic carriers of meaning, replacing *bits* as the basic quanta of traditional channels (Xie et al., 2021; Hu et al., 2023). Each primitive is a feature vector in semantic space, where nearby vectors encode similar meanings, such as "supper" and "dinner". This representation enhances transmission reliability: despite channel noise or loss, receivers can reconstruct messages from proximate primitives, enabling effective semantic exchange. In comparison, *semantic patterns* offer a higher-order abstraction by capturing **semantic behaviors**. While semantic primitives correspond to specific expressions, semantic patterns encapsulate the underlying intent. This allows semantic patterns to describe general behaviors across diverse expressions. For an intuitive illustration of the difference, see the example provided below:

Consider two expressions:

1) I apologize for the mistake I had last night.

2) Thank you for being so accommodating of my actions.

[5]The equivalent sample size α is a hyperparameter representing the strength of the prior belief. Its typical values include: 1, 5, 10, etc.

Composed of virtually disjoint semantic primitives, these two sentences nonetheless uniformly exhibit the speaker's polite mode of expression. Where behavioral features warrant emphasis, unification via a semantic pattern like "*the expression of politeness*" proves effective.

Specific semantic content in self-reflection outputs holds limited interest for this type of research. With scaling data volumes, individual semantics becomes impractically rare, rendering it unsuitable for generalizability discussions. Higher-order abstractions are thus required to organize text data. Such abstractions facilitate interpretable, quantitative analyses in frameworks like ReBeCA, leading to broadly applicable insights. Defining semantic patterns serves precisely this purpose: leveraging combinations of abstract semantic behaviors to quantify model outputs across diverse text corpora. The definitions (identification prompts) of all semantic patterns included in this paper are listed here (referred in Sec. 4.2.1):

MATH (logic reasoning):

1. **Completeness:** `Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the completeness of the previous response, i.e., feedback evaluates if all parts of the math problem were solved and if there is any information missing or omitted, or refined response makes the corresponding refinement from the perspective of completeness.`

2. **Specificity:** `Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the specificity of the previous response, i.e,. feedback evaluates if all parts of the math problem were solved and if there is any information missing or omitted, or refined response makes the corresponding refinement from the perspective of specificity.`

3. **Clarity and Organization:** `Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the clarity and organization of the`

previous response, i.e., feedback evaluates if all parts of the math problem were solved and if there is any information missing or omitted, or refined response makes the corresponding refinement from the perspective of clarity and organization.

4. **Depth of Analysis:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the depth of analysis of the previous response, i.e., feedback evaluates if all parts of the math problem were solved and if there is any information missing or omitted, or refined response makes the corresponding refinement from the perspective of depth of analysis.

BOUQuET (machine translation):

1. **Completeness:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the completeness of the previous translation, i.e., feedback evaluates if all parts of the source text were translated and if there is any information missing or omitted, or refined response makes the corresponding refinement from the perspective of completeness.

2. **Fluency:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the fluency of the previous translation, i.e., feedback evaluates if the translation reads naturally in the target language and if there are any grammatical errors or awkward expression, or refined response makes the corresponding refinement from the perspective of fluency.

3. **Cultural Appropriateness:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the cultural appropriateness of the previous translation, i.e., feedback evaluates if cultural references and context are appropriately adapted for the target language, or refined response makes the corresponding refinement from the perspective of cultural appropriateness.

4. **Consistency:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the consistency of the previous translation, i.e., feedback evaluates if terminology and style are consistent throughout the translation, or refined response makes the corresponding refinement from the perspective of consistency.

5. **Clarity and Organization:** Evaluate and judge whether the model's self-reflect-and-refine record below demonstrates an attempt to optimize the clarity and organization of the previous translation, i.e., feedback evaluates if the translation is clear and easy to understand for native speakers of the target language, or refined response makes the corresponding refinement from the perspective of clarity.

Prompt template for LLM-based matching of semantic patterns:

What to Know before Your Task

Semantic Pattern

A **semantic pattern** refers to a high-level characterization that summarizes a kind of conversational behaviors recognizable across multiple single-turn dialogues generated by a language model (LM). For example, the semantic pattern "repeating" could represent repetitive wording within a single response from a LM. It is a straightforward yet sensible way of quantifying and structuring multi-turn dialogues.

For instance, five responses produced by a LM from a multi-turn dialogue could be quantified into a $5 \times N$ binary matrix by defining N semantic patterns,

where the existence of a semantic pattern S_i ($i = 1, 2, \ldots, N$) regarding response k ($k = 1, 2, \ldots, 5$) is denoted as 0 (S_i appears somewhere within response k) or 1 (S_i does not appear anywhere within response k).

We conduct quantitative analyses on a batch of LM-generated self-reflection data, represented in pure text, on the basis of machine translation requests. Self-reflection denotes a specific kind of iterative self-prompting framework, which, for any input query, generates a multi-turn self-reflect-and-refine dialogue to reach an optimal response. Analyzing this type of data quantitatively is non-trivial; therefore, we introduce **semantic patterns**. By treating each sample, namely a multi-turn self-reflect-and-refine dialogue comprising revision feedback and refined responses, as a binary semantic-pattern matrix, we are able to quantify the textual data and apply statistical analyses in downstream tasks.

Notations

- Text content wrapped in $\langle q \rangle$ represents a specific translation request.

- Text content wrapped in $\langle y_0 \rangle$ represents the language model's previous response to the request.

- Text content wrapped in $\langle y^* \rangle$ represents the ground-truth or reference translation.

- Text content wrapped in $\langle fb \rangle$ represents the revision feedback generated by self-reflecting on the previous response $\langle y_0 \rangle$.

- Text content wrapped in $\langle y_1 \rangle$ represents the refined translation produced by modifying $\langle y_0 \rangle$ based on the feedback $\langle fb \rangle$.

Task Definition

Please carefully read and understand the preceding instructions and the meanings of the text variables corresponding to each tag, and then complete the following identification task: evaluate and judge whether the model's self-reflect-and-refine record demonstrates an attempt to optimize the $\langle SP \rangle$ of the previous translation. Specifically, $\langle SP \rangle$ optimization refers to $\langle \mathrm{Explain}(SP) \rangle$.

- $\langle q \rangle$ {query}

- $\langle y^* \rangle$ {ground_truth}

- $\langle y_0 \rangle$ {prev_resp}

- $\langle fb \rangle$ {feedback}

- $\langle y_1 \rangle$ {refined_resp}

Please make your judgment by replying **True** (optimization of $\langle SP \rangle$ exists in the provided record) or **False** (optimization of $\langle SP \rangle$ does not exist in the provided record), and state the reason for your judgment.

C Presentation of Additional Results

Table 4: (Referred in Sec. 5.1.1) Stage 1 outputs: CVLLs of candidate causal parent sets $\mathrm{Pa}_{\hat{y}}$s across all testing folds. Abbreviations for all semantic patterns and the two jointly considered uncertainty metrics: CO = Clarity and Organization, DA = Depth Analysis, Sp = Specificity, MI = Mutual Information, PPL = Perplexity[6]. Candidates obtaining highest CVLLs are shown in **bold**.

MATH		**BOUQuET**		
Causal Parent Set $\mathrm{Pa}_{\hat{y}}$	**CVLL $\mathcal{L}_{\mathrm{BDs}}$**	**$\hat{y}$ indicator**	**Causal Parent Set $\mathrm{Pa}_{\hat{y}}$**	**CVLL $\mathcal{L}_{\mathrm{BDs}}$**
$\{\mathrm{CO}_{t=2}, \mathrm{CO}_{t=5}, \mathrm{MI}_{t=5}\}$	-0.532 ± 0.849	COMET	$\{\mathrm{CO}_{t=2}, \mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=3}\}$	-1.231 ± 1.374
$\{\mathrm{CO}_{t=2}, \mathrm{DA}_{t=5}, \mathrm{PPL}_{t=4}\}$	-0.481 ± 0.667	COMET	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=4}\}$	-1.088 ± 0.976
$\{\mathrm{CO}_{t=2}, \mathrm{MI}_{t=2}\}$	-0.447 ± 0.561	COMET	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=5}\}$	-1.009 ± 0.880
$\{\mathrm{CO}_{t=2}, \mathrm{MI}_{t=5}\}$	-0.482 ± 0.680	COMET	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}\}$	-0.981 ± 0.736
$\{\mathrm{CO}_{t=2}, \mathrm{PPL}_{t=2}, \mathrm{MI}_{t=2}\}$	-0.466 ± 0.616	COMET	$\{\mathrm{Sp}_{t=4}, \mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}\}$	-1.036 ± 0.945
$\{\mathrm{CO}_{t=2}, \mathrm{PPL}_{t=4}, \mathrm{PPL}_{t=5}, \mathrm{MI}_{t=4}\}$	-0.602 ± 1.105	**COMET**	$\{\mathbf{CO_{t=3}}\}$	$\mathbf{-0.978 \pm 0.495}$
$\{\mathrm{CO}_{t=2}\}$	-0.445 ± 0.548	MetricX	$\{\mathrm{CO}_{t=2}, \mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=3}\}$	-1.216 ± 1.283
$\{\mathrm{Sp}_{t=2}, \mathrm{CO}_{t=2}, \mathrm{CO}_{t=3}\}$	-0.457 ± 0.688	MetricX	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=4}\}$	-1.075 ± 0.960
$\{\mathrm{Sp}_{t=3}, \mathrm{CO}_{t=2}, \mathrm{CO}_{t=4}\}$	-0.464 ± 0.726	MetricX	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}, \mathrm{PPL}_{t=5}\}$	-1.085 ± 0.935
$\{\mathrm{Sp}_{t=3}, \mathrm{CO}_{t=2}, \mathrm{MI}_{t=2}\}$	-0.449 ± 0.608	MetricX	$\{\mathrm{Sp}_{t=4}, \mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}\}$	-1.054 ± 0.907
$\{\mathrm{Sp}_{t=4}, \mathrm{CO}_{t=2}, \mathrm{MI}_{t=4}\}$	-0.471 ± 0.747	**MetricX**	$\{\mathbf{CO_{t=3}}\}$	$\mathbf{-1.003 \pm 0.536}$
$\{\mathbf{Sp_{t=4}}, \mathbf{CO_{t=2}}\}$	$\mathbf{-0.434 \pm 0.563}$	MetricX	$\{\mathrm{CO}_{t=3}, \mathrm{CO}_{t=4}\}$	-1.007 ± 0.733

Table 5: (Referred in Sec. 5.1.3) Stage 3 outputs: results of linear stability validation, evaluated via Levene's test and one-way ANOVA on 20 folds of regrouped data.

Dataset	$\hat{\mathbf{y}}$	$\mathbf{p_a} \in \mathbf{Pa}^*_{\hat{\mathbf{y}}}$	**M (95% CI)**	**Levene's Test**	**ANOVA**	η^2_{p}	**Status**
MATH	–	$\mathrm{Sp}_{t=4}$ (k_1)	0.160 [0.122, 0.198]	$F(3,16) = 1.27, p = .317$	$F(3,16) = 0.25, p = .860$	.045	Stable
		$\mathrm{CO}_{t=2}$ (k_2)	0.298 [0.259, 0.338]	$F(3,16) = 0.11, p = .950$	$F(3,16) = 0.10, p = .956$	.019	Stable
		Intercept (b)	0.514 [0.488, 0.540]	$F(3,16) = 0.69, p = .571$	$F(3,16) = 0.13, p = .942$	.024	Stable
BOUQuET	COMET	$\mathrm{CO}_{t=3}$ (k_1)	−0.526 [−0.588, −0.465]	$F(3,16) = 0.53, p = .671$	$F(3,16) = 0.34, p = .798$	.060	Stable
		Intercept (b)	1.059 [1.018, 1.101]	$F(3,16) = 0.71, p = .558$	$F(3,16) = 0.09, p = .966$	.016	Stable
	MetricX	$\mathrm{CO}_{t=3}$ (k_1)	−0.553 [−0.613, −0.494]	$F(3,16) = 0.61, p = .618$	$F(3,16) = 0.43, p = .732$	.075	Moderate
		Intercept (b)	1.077 [1.028, 1.126]	$F(3,16) = 0.84, p = .491$	$F(3,16) = 0.21, p = .888$	.038	Stable