Bridging Internal Consistency and External Alignment: A Causal and Dynamic Interpretability Framework for LLM Generation

Shuyao Xiao, Shengling Wang*, Ke Chao
College of Artificial Intelligence, Beijing Normal University
xiaoshuyao@mail.bnu.edu.cn, wangshengling@bnu.edu.cn, chaoke@mail.bnu.edu.cn

Abstract

Large Language Models (LLMs) are widely used in high-stakes applications, making their interpretability increasingly important. Existing interpretability methods are typically categorized into internal and external perspectives, which are often studied in isolation and tend to overlook two key aspects: causality and temporal dynamics. Explanations are often limited to surface correlations or static dependencies, failing to capture how influences evolve during autoregressive generation. To address these limitations, we propose a causal and dynamic interpretability framework for LLM generation. We first characterize the backdoor-adjusted causal effects of both the generated prefix and the prompt on the current token using the Structural Causal Model. Next, we introduce two metrics to quantify contextual causal influence and question–answer causal influence. Overall, our work provides a unified causal view of internal consistency and external alignment in LLM generation dynamics.[1]

1 Introduction

Large Language Models (LLMs) (Brown et al., 2020) achieve impressive performance in dialogue, question answering, and content generation. They are widely deployed in high-stakes real-world scenarios such as search, education, healthcare, and decision support. Therefore, clarifying the rationale behind their specific outputs and the information they rely on during generation is critical to ensuring their reliability, safety, and controllability.

Existing research on LLM interpretability can be categorized into two classes based on their objectives: internal and external interpretability. The former focuses on mechanistic explanations of the model's internal generation process (Conmy et al., 2023; Ortu et al., 2024; Marks et al., 2024), while the latter emphasizes explanations of the model's behavior in meeting human requirements, such as evaluating generation quality or diagnosing instruction-following performance (Calderon and Reichart, 2025; Wang et al., 2025; Deutsch et al., 2022; Qin et al., 2024; Madhavan et al., 2023).

Despite notable advances in prior work, a prominent gap in current research is the lack of integration between internal and external interpretability. These two perspectives are often studied in isolation, leading to explanations that either accurately describe internal mechanisms without aiding external evaluation or align with external criteria while detaching from the model's true generative process.

Beyond the lack of integration between internal and external interpretability, existing work also falls short in jointly considering two fundamental analytical perspectives: causality and dynamics. From a causal perspective, interpretability should identify which factors genuinely constrain model generation. Without causal reasoning, explanations are limited to surface correlations and cannot differentiate co-occurrence from true causal influence (Zhang and Nanda, 2023). From a dynamic perspective, interpretability should consider how the generation process unfolds over time. In the absence of such a view, explanations focus on static outputs or isolated decoding steps, overlooking how tiny deviations can accumulate during autoregressive generation (Anagnostidis et al., 2023). When causality and dynamics are not considered together, explanations may seem reasonable at individual steps but fail to reflect the model's decision logic throughout the generation trajectory, limiting their reliability and diagnostic value.

To address these limitations, we propose a causal and dynamic interpretability framework for LLM generation. Our framework provides a unified causal perspective for analyzing how contextual causal constraints from the generated prefix to the current token, together with question–answer

*Corresponding author.

[1]The code and datasets are publicly available at https://github.com/WinnieShaw/Causal-Dynamic.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 20378–20392

causal alignment induced by the input question, evolve along the generation trajectory. Under this perspective, internal contextual dependency within generation and external question–answer alignment can be characterized within the same causal-dynamic framework.

Based on the Structural Causal Model (SCM) (Pearl, 2010a), we further introduce two computable metrics: Contextual Causal Influence (CCI) and Question–Answer Causal Influence (QACI). CCI quantifies the causal influence of the prefix on the current generated token, characterizing contextual constraints during generation. QACI measures the degree of causal alignment between the input question and the generated answer, reflecting how strongly generation depends on the question. Importantly, neither metric requires reference answers, making them especially suitable for open-ended generation tasks without ground truth. Our contributions are as follows:

(1) We propose a causal-dynamic interpretability framework for LLM generation, together with two metrics, CCI and QACI, that provide a unified way to characterize contextual causal constraints, question–answer causal alignment, and their trajectory-level evolution during generation.

(2) Experiments show that CCI establishes relatively stable prefix-level causal constraints early in generation, while QACI reveals systematic relationships between question–answer causal alignment and factors such as question difficulty, length, and semantic complexity. These findings suggest that causal constraints in generation follow observable trajectory-level patterns, rather than being reflected only through static outcome-level correlations.

2 Structural Causal Modeling

We model sequence generation as an autoregressive time-step process (Vaswani et al., 2017). The original question is first transformed into a token sequence Q. At decoding step t, the model computes the conditional distribution $P(\varepsilon_t \mid Q, A_{t-1})$ based on the input question Q and the previously generated prefix A_{t-1}, then generates the next token ε_t and updates the current sequence as $A_t = \mathrm{concat}(A_{t-1}, \varepsilon_t)$. Starting from $A_1 = \varepsilon_1$, this process continues autoregressively until an end-of-sequence token (<EOS>) is produced or a maximum length is reached. Therefore, each token ε_t, and thus the sequence A_t, is jointly determined by the question Q and the preceding sequence A_{t-1}.

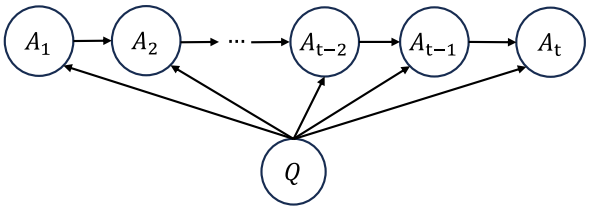


Figure 1: Causal structure of sequence generation in LLMs. Here, A_t denotes the generated sequence up to step t.

Based on the above analysis, we create a structural causal graph (Pearl, 2010b) for sequence generation in LLMs. As shown in Figure 1, for the sequence A_t, there are two influential paths presented on the graph. The first is $A_1 \rightarrow A_2 \rightarrow \cdots \rightarrow A_{t-2} \rightarrow A_{t-1} \rightarrow A_t$, indicating that each time step's generated sequence is influenced by the previous one. The second is $Q \rightarrow \{A_1, A_2, \cdots, A_{t-2}, A_{t-1}, A_t\}$, showing that Q affects generation at each time step.

For the current sequence A_t, the first type of causal path is the direct path: $A_{t-1} \rightarrow A_t$ and $Q \rightarrow A_t$. The second type of causal path includes confounding factors that affect both the prefix A_{t-1} and the current sequence A_t, leading to the observed influence of A_{t-1} on A_t being partially attributable to the influence of Q through A_{t-1}. Thus, the observed influence of A_{t-1} on A_t is biased and does not represent the true causal effect.

To eliminate amplified effects, exclusion experiments can be considered; however, such experiments are often time-consuming, costly, and difficult to scale in practice (Feder et al., 2021). Therefore, we adopt the SCM to analyze LLM generation. By analytically blocking the backdoor path induced by the input question, this approach enables direct estimation of causal effects from standard decoding probabilities. It removes the confounding influence of Q without altering the generation trajectory or requiring counterfactual decoding, thereby isolating the causal effect under the assumed SCM in an efficient and scalable manner.

To quantify the causal effects and make them computable without the do-operator, we derive the following theorem on the unbiased causal effect. In the following derivations, we use uppercase letters to denote random variables and lowercase letters to denote their observed values.

Theorem 1. *The backdoor-adjusted causal effect of A_{t-1} on A_t, obtained by intervening while marginalizing over Q drawn from the same batch*

of questions, is given by:

$$P(A_t = a_t \mid do(A_{t-1} = a_{t-1})) = \sum_{Q=q} P(\varepsilon_t \mid Q, A_{t-1} = a_{t-1}) \cdot P(Q = q).$$

Proof. Here, $P(A_t = a_{t-1} \mid \mathrm{do}(A_{t-1} = a_{t-1}))$ denotes the interventional distribution that blocks the backdoor path (Pearl, 2009) via Q. According to Figure 1, Q is a common cause of both A_{t-1} and A_t, which induces the backdoor path $A_{t-1} \leftarrow Q \rightarrow A_t$. Therefore, without controlling for Q, the observed influence of A_{t-1} on A_t contains not only the direct effect of A_{t-1} on A_t, but also spurious dependence induced by the question Q.

Thus, Q is the confounding variable and forms a valid adjustment set. By the backdoor criterion (Pearl, 2009), we have

$$P(A_t = a_t \mid \mathrm{do}(A_{t-1} = a_{t-1})) = \sum_{Q=q} P(A_t = a_{t-1} \mid Q = q, A_{t-1} = a_{t-1}) \cdot P(Q = q).$$

Furthermore, as discussed earlier, in autoregressive generation the difference between A_t and A_{t-1} lies only in the newly generated token ε_t. Therefore, the conditional term above can be equivalently written using $P(\varepsilon_t \mid Q = q, A_{t-1} = a_{t-1})$, which yields Theorem 1. This procedure distinguishes the "spurious association" from the "true causal effect" in LLM generation, providing a foundation for unbiased effect estimation.

To verify whether the numerical estimation of the do-operator in Theorem 1 is affected by the size of the set $|Q|$, we conducted experiments under different scales of $|Q|$. The results show that it remains generally stable; detailed findings are provided in Appendix A. □

Theorem 2. *When there is no confounding factor between Q and A_t, the causal effect of Q on A_t is identifiable without adjustment, and the interventional distribution is:*

$$P(A_t = a_t \mid do(Q = q)) = P(A_t = a_t \mid Q = q).$$

We thus derive all unbiased causal effects on A_t.

3 Causal Influence Metrics

This section proposes two causal influence metrics based on the aforementioned causal effects.

3.1 Contextual Causal Influence

We define *Contextual Causal Influence* (CCI) to quantify the internal consistency between the current generated sequence A_t and its prefix A_{t-1}.

Definition 1. *The Contextual Causal Influence (CCI) at token t is defined as*

$$CCI(A_t = a_t \mid A_{t-1} = a_{t-1}) = \log_2 \frac{P(A_t = a_t \mid do(A_{t-1} = a_{t-1}))}{P(A_t = a_t)}.$$

We propose the following theory on the computability of CCI:

Theorem 3. *CCI can be computed as:*

$$CCI(A_t = a_t \mid A_{t-1}) = \log_2 \frac{\sum_{Q=q} P\left(\varepsilon_t \,\middle|\, {}^{Q=q}_{A_{t-1}=a_{t-1}}\right) P(Q = q)}{\sum_{Q=q} P(A_t = a_t \mid Q = q)\, P(Q = q)}. \quad (1)$$

Proof. Specifically, $\mathrm{CCI}(A_t = a_t \mid A_{t-1} = a_{t-1})$ measures the log-ratio of the probabilities of generation with and without prefix A_{t-1} guidance. For the numerator, according to Pearl (2010a), we intervene on $A_{t-1} = a_{t-1}$ to block the backdoor path from the question variable Q, thereby eliminating confounding effects:

$$\mathrm{CCI}(A_t = a_t \mid A_{t-1} = a_{t-1}) = \log_2 \frac{\sum_{Q=q} P\left(A_t = a_t \,\middle|\, {}^{Q=q}_{A_{t-1}=a_{t-1}}\right) P(Q = q)}{P(A_t = a_t)}. \quad (2)$$

For the denominator, by applying the law of total probability, we have:

$$\mathrm{CCI}(A_t = a_t \mid A_{t-1} = a_{t-1}) = \log_2 \frac{\sum_{Q=q} P\left(A_t = a_t \,\middle|\, {}^{Q=q}_{A_{t-1}=a_{t-1}}\right) P(Q = q)}{\sum_{Q=q} P(A_t = a_t \,|\, Q = q)\, P(Q = q)}. \quad (3)$$

When the current sequence A_t differs from the prefix A_{t-1} only in the new token ε_t, the difference in generation probability arises solely from ε_t. Thus, given $Q = q$ and $A_{t-1} = a_{t-1}$, the probability of $A_t = a_t$ can be expressed as the probability of ε_t, leading to (1). □

3.2 Question-Answer Causal Influence

To capture causal dependency between a model's generated answer and the input question, we define the *Question–Answer Causal Influence* (QACI), which quantifies the causal influence of the question on the generated answer.

Definition 2. *Suppose that $A_n = a_n$ is the entire generated answer to the question $Q = q$. We define the Question–Answer Causal Influence (QACI) as:*

$$\begin{aligned} &QACI(A_n = a_n \mid Q = q) \\ &= \log_2 \frac{P(A_n = a_n \mid do(Q = q))}{P(A_n = a_n)}. \end{aligned} \tag{4}$$

Furthermore, we have the following theory on QACI computability:

Theorem 4. *QACI can be computed as follows:*

$$\begin{aligned} &\mathrm{QACI}(A_n = a_n \mid Q = q) = \\ &\log_2 \frac{P(\varepsilon_1 = a_1 \mid Q = q) \prod_{t=2}^{n} P\left(\varepsilon_t \,\middle|\, {}^{Q=q}_{A_{t-1}=a_{t-1}}\right)}{P(\varepsilon_1 = a_1) \prod_{t=2}^{n} P(\varepsilon_t \mid A_{t-1} = a_{t-1})}. \end{aligned} \tag{5}$$

Proof. According to the backdoor criterion (Pearl, 2010a), there exists no backdoor path from the question Q to the generated token A_t in Figure 1. Therefore, the causal effect of Q on A_t is identifiable without adjustment, and the interventional distribution is

$$\begin{aligned} &\mathrm{QACI}(A_n = a_n \mid Q = q) = \\ &\log_2 \frac{P(A_n = a_n \mid Q = q)}{P(A_n = a_n)}. \end{aligned} \tag{6}$$

We further decompose $P(A_n = a_n \mid Q = q)$ based on the chain rule (Murphy, 2012). As the model generates answers autoregressively, the generation of each token ε_t depends on the question $Q = q$ and the prefix A_{t-1}. We fix the prefix A_{t-1} to the partial answer a_{t-1}. Thus, the full answer generation probability can be expressed as:

$$\begin{aligned} P(A_n = a_n \mid Q = q) = P(\varepsilon_1 = a_1 \mid Q = q) \\ \cdot \prod_{t=1}^{n} P(\varepsilon_t \mid Q = q, A_{t-1} = a_{t-1}). \end{aligned} \tag{7}$$

Similarly, we have:

$$\begin{aligned} P(A_n = a_n) = P(\varepsilon_1 = a_1) \\ \cdot \prod_{t=1}^{n} P(\varepsilon_t \mid A_{t-1} = a_{t-1}). \end{aligned} \tag{8}$$

Substituting (7) and (8) into (6), we can derive (5). □

Specifically, $\mathrm{QACI}(A_n = a_n \mid Q = q)$ represents the log-ratio of the probability of generating the answer with versus without question guidance. Appendix C provides the feasibility and complexity analysis of CCI and QACI.

4 Experimental evaluation

4.1 Causal Validation

To justify the causal validity of CCI and QACI, we analyze three classical criteria—temporality, covariation, and exclusivity, which are fundamental in causal inference (Reichardt, 2002).

4.1.1 Temporality

Temporality requires that a cause precede its effect in time, ensuring the correct temporal ordering for causal interpretation. In autoregressive generation, temporality is inherently satisfied: the input question precedes the answer, and the prefix is formed before the model produces the next token. Therefore, our experimental analysis focuses on covariation and exclusivity.

4.1.2 Covariation

Covariation implies that variations in a causal measure are systematically linked to changes in other metrics, indicating non-independent variations among them. We focus on summarization using the CNN/DailyMail dataset (See et al., 2017). We evaluated LLMs including the T5 family (Raffel et al., 2020), Qwen2.5-Instruct-7B (Team, 2024), Mistral-7B-v0.1 (Jiang et al., 2023) and the Llama3-8B base model (Meta AI, 2024). Experiments are conducted on a single NVIDIA A100 GPU.

CCI is compared to similarity-based metrics like Adjacent-Sentence Embedding Cosine (Cosine) (Gao et al., 2021), Jaccard Overlap (Jaccard) (Jaccard, 1901; Shao et al., 2024), and topic Jensen–Shannon Divergence (Topic-JS) (Lin, 2002; Wang et al., 2024). QACI is benchmarked against classical summarization accuracy metrics, including Recall-Oriented Understudy for Gisting Evaluation-Longest Common Subsequence (ROUGE-L) (Lin, 2004; Saha and Zhang, 2023), character n-gram F-score (chrF) (Popović, 2015; Winata et al., 2024), and Translation Edit Rate (TER) (Snover et al., 2006; Deguchi et al., 2024). We use 250 samples for CCI and 500 for QACI,

Model	CCI↑	Cosine↑	Jaccard↑	Topic-JS↓	QACI↑	ROUGE-L↑	CHRF↑	TER↓
T5-small	14.66	0.72	0.16	0.80	322.19	26.43	34.60	90.41
T5-base	15.35	0.56	0.03	0.93	471.43	24.14	34.86	95.64
T5-large	14.45	0.50	0.06	0.70	877.39	30.63	41.83	88.91
Mistral-7B-v0.1	22.25	0.62	0.06	0.65	168.46	17.21	29.62	162.01
Qwen2.5-Instruct-7B	17.57	0.86	0.05	0.62	334.27	25.24	41.12	122.18
LLaMA 3-8B	19.37	0.66	0.05	0.63	166.35	19.63	35.10	168.16
Pearson r		0.16	-0.29	-0.59		0.80	0.66	-0.77
95% CI		[0.04, 0.28]	[-0.40, -0.17]	[-0.67, -0.51]		[0.77, 0.82]	[0.61, 0.71]	[-0.80, -0.73]
Spearman ρ		0.31	-0.09	-0.54		0.90	0.57	-0.91

Table 1: Model-level comparison of CCI/QACI with conventional metrics. An upward arrow ↑ indicates better performance with higher values, while a downward arrow ↓ indicates better performance with lower values.

shuffling with fixed random seeds and repeating evaluations 10 times. After averaging model performance, we compute Pearson correlation coefficients, 95% confidence intervals, and Spearman correlation coefficients between the metrics (Benesty et al., 2009; Fisher, 1921; Spearman, 1961).

Table 1 presents the model-level comparison between CCI/QACI and other correlation-based metrics. The results show that Mistral-7B-v0.1 achieves the highest CCI value, indicating optimal contextual causal consistency. Overall, CCI exhibits a relatively clear negative correlation with Topic-JS, suggesting that stronger contextual causal dependence is generally associated with lower topic drift, and the confidence-interval analysis further supports the robustness of this relationship. Meanwhile, CCI shows a weak positive correlation with Cosine, indicating a certain degree of consistency with semantic similarity. In contrast, the correlation between CCI and Jaccard is relatively weak, suggesting that surface-level lexical overlap is insufficient to fully reflect the contextual causal information captured by CCI. Overall, CCI appears to be more closely related to topical consistency and semantic coherence than to superficial lexical matching.

For QACI, T5-large model attains the highest value, indicating that it possesses the strongest causal association strength between questions and answers. The results show that QACI is strongly positively correlated with ROUGE-L/CHRF and strongly negatively correlated with TER, indicating that a stronger question–answer causal influence is generally associated with better reference alignment and lower edit distance. Confidence-interval analysis further suggests that these correlations are overall robust, with the relationships to ROUGE-L and TER appearing particularly stable. Overall, although QACI does not rely on reference answers, it still exhibits clear covariation with multiple reference-based metrics, while also capturing

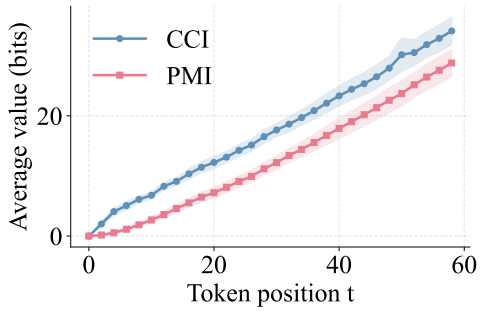


Figure 2: Average values of causal and correlational metrics.

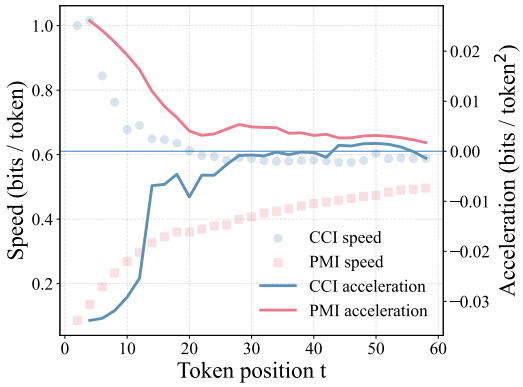
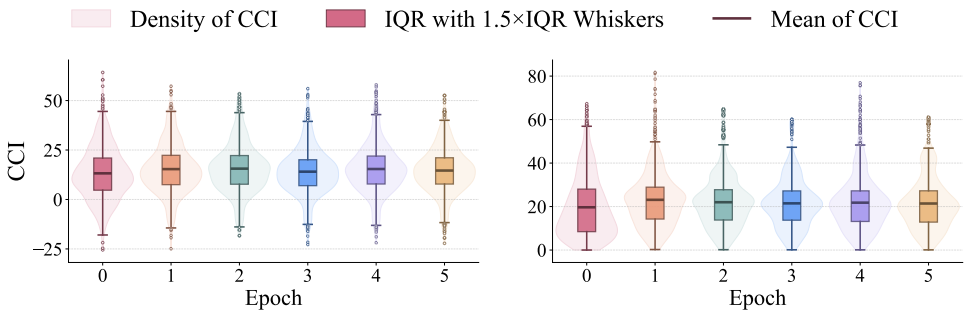


Figure 3: Speed and acceleration dynamics of causal (CCI) and correlational (PMI) metrics, averaged over 10 repeated runs.

causal influence information that traditional metrics cannot directly reflect.

4.1.3 Exclusivity

Exclusivity characterizes whether the influence attributed to a specific factor cannot be explained away by other correlated variables, thereby reflecting the uniqueness of its contribution to the outcome. A natural baseline for this purpose is Pointwise Mutual Information (PMI) (Xu et al.), which quantifies statistical dependence but may conflate unique influence with dependence induced by other correlated variables. It measures statistical depen-



(a) T5-base (b) Qwen2.5-Instruct-7B

Figure 4: CCI varying trajectories over the fine-tuning process.

dence between two random variables by comparing their joint probability to the product of their marginals. In the generation setting, we use PMI to quantify the observational dependence between the generated prefix A_{t-1} and the current sequence A_t, conditioned on the question Q:

$$\mathrm{PMI}(A_{t-1}; A_t \mid Q) = \log_2 \frac{P(A_t \mid A_{t-1}, Q)}{P(A_t \mid Q)}.$$

Due to a lack of high-quality datasets for open-ended questions, we randomly selected 50 questions from our original dataset (see Appendix D for details) and repeated the experiments 10 times.

As shown in Figure 2, both CCI and PMI demonstrate increasing dependence between the prefix A_{t-1} and the current sequence A_t as the token position advances, indicating that the model's generation becomes more conditioned on the preceding context. CCI consistently achieves higher values than PMI across all positions, indicating a systematic offset between the two measures.

In Figure 3, as generation proceeds, the speed of PMI increases with token position, and its acceleration gradually decreases but remains above 0. This suggests that the observed dependence in PMI mainly arises from the accumulation of longer contexts, particularly surface-level co-occurrence, rather than a strong causal constraint. In contrast, the speed of CCI increases rapidly in the early stages, with negative acceleration approaching 0, indicating that the causal constraint imposed by the prefix is established early and maintained with stable strength throughout generation.

These observations indicate that when PMI is computed without accounting for the confounding influence of the question Q, it confounds the estimated effect of the prefix with that of Q, thereby consistently undervaluing the prefix's actual contribution. By contrast, CCI rapidly establishes causal constraints between the prefix and the current sequence early in generation, maintaining a stable increase in the middle and later stages (with acceleration rising from negative values to stabilizing around zero), thus providing a more stable and intuitive measure of contextual influence.

4.2 Multi-Task Evaluation

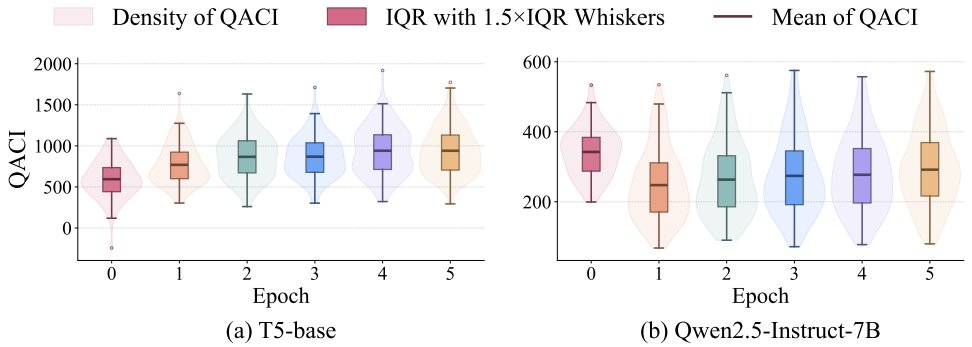
This section analyzes the trajectories of CCI and QACI across fine-tuning stages, summarization, extractive QA, and open-ended tasks with varying difficulty levels.

4.2.1 Fine-Tuning

This section aims to track CCI and QACI throughout training, characterizing how causal constraints and question–answer alignment evolve during generation rather than demonstrating performance improvements during fine-tuning. This allows us to observe how the model gradually develops stable contextual dependencies and stronger alignment with the input, providing a quantitative view of generation behavior evolution during training. Compared to the pre-training stage, which lacks an explicit question–answer alignment structure, the fine-tuning stage offers a clearer question–answer generation format, making it better suited for analyzing causal relations among the question, the prefix, and the current token.

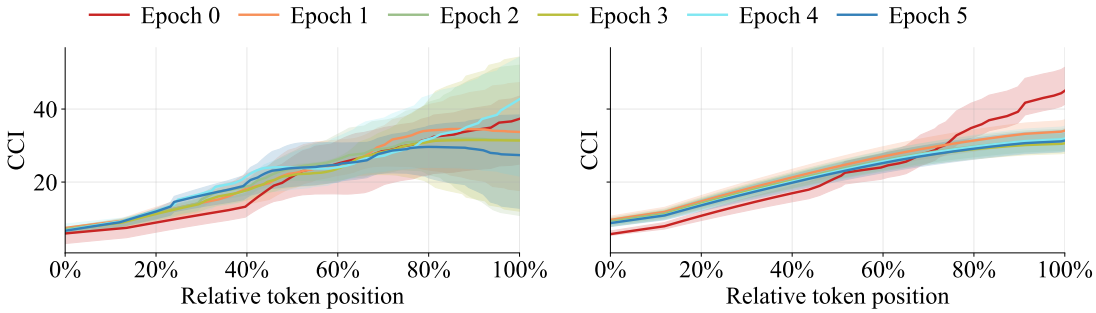
We fine-tune T5-base and Qwen2.5-Instruct-7B on the CNN/DailyMail dataset for 5 epochs, tracking causality dynamics during learning in a standard supervised fine-tuning setup (Raffel et al., 2020; Team, 2024; Yu et al., 2025). After each epoch, we compute QACI and CCI on the validation split, plotting their empirical distributions and epoch-wise trajectories.

(1) Varying trajectories of CCI. Figure 4 illustrates that during training, both models' CCI



(a) T5-base (b) Qwen2.5-Instruct-7B

Figure 5: QACI varying trajectories over the fine-tuning process.



(a) T5-base (b) Qwen2.5-Instruct-7B

Figure 6: Average CCI (±95% CI) vs. relative token position.

remain above zero, indicating that generations are strongly constrained by prior context. From epoch 0 to epoch 1, CCI significantly increases, followed by minor fluctuations from epoch 2 to epoch 5. Both models exhibit distributional contraction: the Inter Quartile Range (IQR) and its whiskers shrink modestly, suggesting reduced variability and more consistent contextual dependence.

(2) Varying trajectories of QACI. As shown in Figure 5, QACI increases starting from epoch 1. It indicates that as training progresses, the model's generation becomes increasingly dependent on the questions, reflecting a growing degree of causal alignment between the question and the answer. However, the temporary decrease in QACI for Qwen2.5 from epoch 0 to epoch 1 arises from its rapid adaptation to generic summarization templates, resulting in a transient causal misalignment between the question and the answer. In contrast, T5-base establishes task-specific templates earlier, leading to a steady increase in QACI from the start.

4.2.2 Generation

This section further investigates the proposed causal dynamics during generation. We first analyze how CCI varies with relative generation position in the summarization task (CNN/DailyMail) and the QA task (CMRC 2018). Then, we examine the effects of question difficulty on CCI and question length on QACI in open-ended question-answering tasks.

(1) CNN/DailyMail. Figure 6 shows the per-epoch CCI averaged by relative token position, with the solid curve denoting the mean and the shaded region representing the 95% confidence interval. Relative token position normalizes each token's index within a sequence, expressing its location from 0% (beginning) to 100% (end). As the sequence progresses, CCI consistently increases in both models. In T5-base, CCI slightly decreases for later tokens in epochs 2 and 5. In Qwen2.5-Instruct-7B, the upward trend in epoch 0 is more pronounced than in subsequent epochs, reflecting the original model's significant potential for further optimization. Overall, these results highlight that each newly generated token increasingly reflects the preceding content.

(2) CMRC 2018. To further enrich the empirical results and strengthen the persuasiveness of our findings, we conduct experiments on a passage-understanding and extractive QA task using the CMRC 2018 dataset. In this setting, given a passage and a question, the model is required to locate and output a continuous answer span directly from the passage. The experiments reveal two clear and

Interval%	0–20	20–40	40–60	60–80	80–100
Mean CCI	2.26	3.45	4.54	5.32	6.34

Table 2: Mean CCI across relative position intervals.

mutually reinforcing phenomena: (i) Contextual Causal Influence (CCI) dynamics along the generation trajectory exhibit a smooth and monotonically increasing trend with respect to the relative token position. The value ranges and mean statistics are summarized in Table 2. Approximately 92% of the samples exhibit a positive-slope increase in CCI as the relative token position grows, indicating that this pattern appears consistently across the vast majority of cases. This suggests that the model gradually forms and strengthens its causal constraint structure during generation. (ii) The correlation analysis further shows that QACI is strongly associated with answer quality in the passage-understanding and extractive QA setting. Specifically, the Pearson correlation coefficient is $r \approx 0.72$ ($p \approx 3.1 \times 10^{-9}$), and the Spearman correlation coefficient is $\rho \approx 0.74$ ($p \approx 8.0 \times 10^{-10}$). These results indicate that QACI reliably captures the semantic correctness and alignment quality between questions and answers, and can serve as a causal alignment metric that is reusable across datasets and question types. The extremely small p-values suggest that the observed positive correlations are highly unlikely to arise from random variation, demonstrating strong statistical significance and robustness. Together, these findings support our central claim that CCI and QACI provide a unified and stable cross-task causal interpretability perspective, respectively characterizing causal dynamics during generation and answer-level causal alignment quality.

(3) Open-ended tasks. We evaluate some reference-free question-answering tasks using Qwen2.5-Instruct-7B. The dataset sizes and hierarchical categorizations of questions by abstraction and length are shown in Appendix D.

A) Question difficulty on CCI. We first conduct a question difficulty analysis by grouping questions into multiple levels. We examine how CCI evolves over relative token positions during the generation process. Figure 7 shows a clear and monotonic increase of CCI with question difficulty. CCI increases as question difficulty rises. For level 1 questions, CCI grows slowly and displays larger

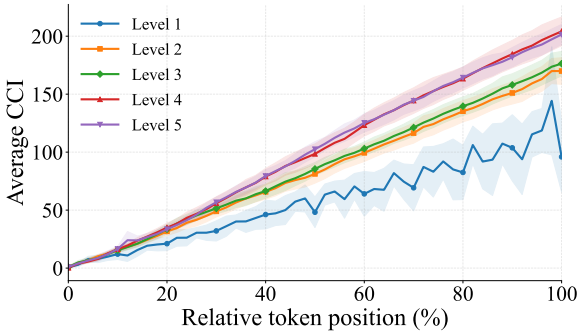


Figure 7: CCI across question difficulty levels.

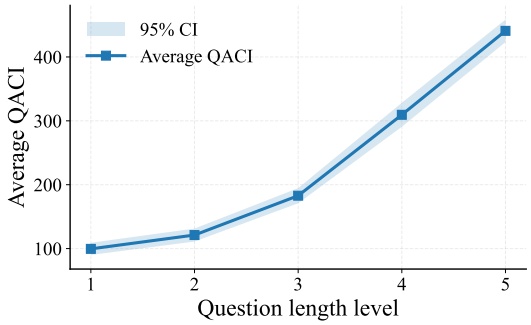


Figure 8: QACI dynamics across question length levels.

fluctuations, indicating a weaker causal constraint of the prefix on the current sequence. As question difficulty increases, CCI rises rapidly in early decoding and continues to strengthen throughout the generation process, suggesting that complex questions rely more on previously generated context for sustained reasoning and structured generation.

B) Question length on QACI. We analyze the effect of question length by grouping questions into five levels based on length and semantic complexity, ranging from short, keyword-like queries to long, multi-sentence questions requiring integrative reasoning. As shown in Figure 8, the average QACI increases monotonically with question length, with a notably steeper rise at higher levels, indicating that longer questions induce substantially stronger causal dependence between the question and the generated answer. This trend suggests that as questions become longer and more informative, the model's generation is increasingly constrained by the global question context, highlighting QACI's sensitivity to question-level causal influence beyond local token-wise dynamics.

5 Related Work

Research on the interpretability of large language models (LLMs) can generally be divided into internal and external interpretability. The distinction is as follows: internal interpretability focuses on

whether the model's mechanisms are accurately represented, while external interpretability emphasizes the role of explanations in understanding, evaluation, and governance.

Internal interpretability reveals the generative foundations and causal mechanisms of the model's behavior. Previous studies have noted that attention weights or model-generated explanations are often merely post-hoc signals, which may not accurately reflect the true decision-making process (Jain and Wallace, 2019). Subsequent work has focused more on verifiable mechanism analysis. For instance, methods like activation patching (Zhang and Nanda, 2023) and circuit discovery (Elhage et al., 2021; Conmy et al., 2023) identify the neurons, components, or circuits responsible for specific behaviors, with activation patching also testing local causal effects through counterfactuals (Ortu et al., 2024); causal mediation analysis (CMA) (Rocchetti et al., 2024) decomposes the causal effects from input to output, measuring the intermediary role of specific internal representations in the causal chain. Some research has analyzed how knowledge and behavior evolve in the model based on representations and training dynamics (Belrose et al., 2023; Marks et al., 2024). Overall, these methods address which internal structures realize a certain behavior.

External interpretability emphasizes the role of explanations in real-world settings, assessing their support for user understanding, model assessment, and system governance rather than reconstructing internal generation mechanisms. The effectiveness of interpretability depends on the stakeholders it serves, as different audiences have distinct preferences for explanations and content (Calderon and Reichart, 2025). Empirical studies show that structured or verifiable information typically enhances user trust and satisfaction more than lengthy reasoning traces, especially in high-risk applications (Wang et al., 2025). Related work in model evaluation and alignment has revealed systematic biases in reference-free evaluation methods, highlighting that external assessments require interpretability and calibration (Deutsch et al., 2022). Other studies break down complex instructions into explicit, checkable requirements, making successes or failures in instruction following more transparent and diagnosable (Qin et al., 2024). Research on fairness and safety uses causal analysis to identify factors contributing to harmful attributes, providing interpretable signals for external governance and risk control (Madhavan et al., 2023).

Although the aforementioned studies have enhanced the interpretability of LLMs from different perspectives, most methods either focus on identifying internal mechanisms or mediation decomposition, or they emphasize external utility and evaluation support. There is still a lack of a unified framework that can simultaneously connect internal consistency with external alignment while capturing the dynamic evolution of causal effects during the generation process. As shown in Appendix E, we compare different explanation paradigms across key capability dimensions. In particular, attention or saliency methods mainly provide static correlation signals, which cannot guarantee intervenable causal meanings; while activation patching, circuit discovery, and CMA can analyze internal causal structures, they cannot describe the dynamic process of when and how constraints form, shrink, and stabilize during the generation trajectory. This paper offers a macroscopic view of causal dynamics: while mechanism methods address which internal structures realize **what or where** constraints, it explains **how and when** these constraints form and evolve during the generation process. Specifically, this paper characterizes causal dynamics at the generation trajectory level, focusing on how constraints gradually form, shrink, and stabilize, and how these processes change with task conditions or training stages, thus providing a complementary unified perspective for existing behavior-level and mechanism-level research.

6 Conclusion

In this work, we propose a causal and dynamic interpretability framework for LLM generation and introduce two reference-free metrics, CCI and QACI, to analyze causal dependencies during autoregressive decoding. Extensive experiments show that CCI establishes stable prefix-level causal constraints early in generation, with its speed and acceleration converging rapidly. QACI reveals systematic dependencies between question–answer causal influence and factors such as question difficulty, length, and semantic complexity. These results suggest that causal-dynamic analysis can be a valuable trajectory-level tool for studying the evolution of contextual constraints and question-answer alignment during generation.

Limitation

This framework focuses on autoregressive text generation and does not cover multimodal or cross-modal generation scenarios. Extending the proposed causal–dynamic interpretability framework to visual or other modalities is reserved for future work. The proposed metrics also assume access to token-level logits, which may limit direct applicability to fully closed-source models.

Acknowledgments

This work has been supported by National Key R&D Program of China (2024YFC3308200), National Natural Science Foundation of China (No. 62293555, 62402047), the Major Program of Science and Technology Innovation 2030 of China (No. 2022ZD0117105), and the Fundamental Research Funds for the Central Universities (No. 2233100006).

References

Sotiris Anagnostidis, Dario Pavllo, Luca Biggio, Lorenzo Noci, Aurelien Lucchi, and Thomas Hofmann. 2023. Dynamic context pruning for efficient and interpretable autoregressive transformers. *Advances in Neural Information Processing Systems*, 36:65202–65223.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. In *Noise reduction in speech processing*, pages 1–4. Springer.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nitay Calderon and Roi Reichart. 2025. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 656–693.

Arthur Conmy, Augustine Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. 2023. Towards automated circuit discovery for mechanistic interpretability. *Advances in Neural Information Processing Systems*, 36:16318–16352.

Hiroyuki Deguchi, Masaaki Nagata, and Taro Watanabe. 2024. Detector–corrector: Edit-based automatic post editing for human post editing. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 191–206.

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. In *EMNLP*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Amir Feder, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. Causalm: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386.

Ronald A Fisher. 1921. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Paul Jaccard. 1901. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bull Soc Vaudoise Sci Nat*, 37:547–579.

Sarthak Jain and Byron C Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jianhua Lin. 2002. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151.

Rahul Madhavan, Rishabh Garg, Kahini Wadhawan, and Sameep Mehta. 2023. Cfl: Causally fair language models through token-level attribute controlled generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11344–11358.

Luke Marks, Amir Abdullah, Clement Neo, Rauno Arike, David Krueger, Philip Torr, and Fazl Barez. 2024. Interpreting learned feedback patterns in large language models. *Advances in Neural Information Processing Systems*, 37:36541–36566.

Meta AI. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2404.09323*.

Kevin P Murphy. 2012. *Machine learning: a probabilistic perspective*. MIT press.

Francesco Ortu, Zhijing Jin, Diego Doimo, Mrinmaya Sachan, Alberto Cazzaniga, and Bernhard Schölkopf. 2024. Competition of mechanisms: Tracing how language models handle facts and counterfactuals. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8420–8436.

Judea Pearl. 2009. *Causality*. Cambridge university press.

Judea Pearl. 2010a. Causal inference. *Causality: objectives and assessment*, pages 39–58.

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

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. Infobench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13025–13048.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Charles S Reichardt. 2002. Experimental and quasi-experimental designs for generalized causal inference.

Elisabetta Rocchetti, Alfio Ferrara, and 1 others. 2024. Causal mediation analysis for interpreting large language models. In *CEUR WORKSHOP PROCEEDINGS*, volume 3741, pages 585–594. CEUR-WS.

Swarnadeep Saha and Shiyue Zhang. 2023. Summarization programs: Interpretable abstractive summarization with neural modular trees. In *The International Conference on Learning Representations (ICLR)*.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. *arXiv preprint arXiv:1704.04368*.

Rulin Shao, Jacqueline He, Akari Asai, Weijia Shi, Tim Dettmers, Sewon Min, Luke Zettlemoyer, and Pang Wei W Koh. 2024. Scaling retrieval-based language models with a trillion-token datastore. *Advances in Neural Information Processing Systems*, 37:91260–91299.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Charles Spearman. 1961. The proof and measurement of association between two things.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2024. Adacad: Adaptively decoding to balance conflicts between contextual and parametric knowledge. *arXiv preprint arXiv:2409.07394*.

Yanyun Wang, Xumei Fang, Zan Xu, Jianye Li, and Luping Wang. 2025. Exploring the impact of explainability in large language model (llm) applications on user experience. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Genta Indra Winata, David Anugraha, Lucky Susanto, Garry Kuwanto, and Derry Tanti Wijaya. 2024. Metametrics: Calibrating metrics for generation tasks using human preferences. *arXiv preprint arXiv:2410.02381*.

Shengwei Xu, Yuxuan Lu, Grant Schoenebeck, and Yuqing Kong. Benchmarking llms' judgments with no gold standard. In *The Thirteenth International Conference on Learning Representations*.

	$\lvert Q\rvert$			
	5	10	25	50
Mean	13.30	14.10	15.20	16.05
Std	10.42	10.58	10.82	10.99

Table 3: Sensitivity of do-operator estimation under different values of $|Q|$.

Ziming Yu, Pan Zhou, Sike Wang, Jia Li, Mi Tian, and Hua Huang. 2025. Zeroth-order fine-tuning of llms in random subspaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4475–4485.

Fred Zhang and Neel Nanda. 2023. Towards best practices of activation patching in language models: Metrics and methods. *arXiv preprint arXiv:2309.16042*.

A Stability of Do-Operator Estimation under Varying the Batch Size of Q

To examine the impact of the size of Q (denoted as $|Q|$) on the numerical estimation of the do-operator, we conducted experiments under different settings with $|Q| \in \{5, 10, 25, 50\}$. For each configuration, more than 500 sample points were evaluated, while keeping the sample size consistent with no duplication or omission.

As shown in Table 3, the mean increases smoothly as $|Q|$ grows, while the magnitude of this change remains small relative to the corresponding standard deviations. This indicates that the overall estimation remains stable with respect to variations in $|Q|$.

This trend is consistent with the mathematical form of the estimator. Both terms in CCI involve a log-sum-exp structure, which can be interpreted as a Monte Carlo approximation of marginalization over Q. As $|Q|$ increases, the approximation becomes slightly tighter, leading to a mild numerical increase in the estimated mean. Importantly, this reflects improved approximation quality rather than any change in the underlying causal structure.

B The Use of Large Language Models

In this work, Large Language Models (LLMs) are only employed to assist with language polishing and writing refinement. The LLM did not influence content ideation, data analysis, or experimental design in any way.

C Feasibility and Complexity Analysis.

Feasibility. According to (3), computing CCI at each time step t requires evaluating two weighted sums over the question variable Q. Both the numerator and the denominator can be expressed using token-level log-probabilities from a standard autoregressive language model: the numerator term $P(A_t = a_t \mid Q = q, A_{t-1} = a_{t-1})$ corresponds to the next-token conditional probability given the prefix and Q, while the denominator term $P(A_t = a_t \mid Q = q)$ is obtained via chain-rule accumulation over the prefix up to step t. Here, the summation over the question variable Q is taken over a finite set consisting of all questions within the same evaluation batch, rather than an unbounded question space. These quantities can be computed using standard forward passes and simple log-domain aggregation (e.g., `logsumexp`). They integrate naturally with KV-cache–based incremental decoding, reuse previously computed states, and therefore avoid re-encoding past prefixes, making the computation efficient in practice.

Complexity. Let M denote the number of questions enumerated (or sampled) for marginalization, L_{ctx} the prompt length, L the generation length, and let CCI be evaluated every B_{step} tokens, yielding $T \approx \lceil L/B_{\text{step}} \rceil$ evaluation points. With incremental decoding, the time complexity for a single generated sequence is

$$O\Big(M\big(L_{\text{ctx}}^2 + T\,L_{\text{ctx}} + T^2\big)\Big),$$

where L_{ctx}^2 accounts for the one-time prefill per q, and the remaining terms arise from incremental decoding across evaluation points. For N sequences, the total complexity scales linearly as

$$O\big(NM(L_{\text{ctx}}^2 + T\,L_{\text{ctx}} + T^2)\big).$$

The space complexity is dominated by the KV cache and can be expressed as

$$O\big(M\,(L_{\text{ctx}} + T)\,n_{\text{layers}}\,d\big),$$

with batching over q reducing constant factors in practice.

D Original Open-ended Datasets

In Section 4.2.2 on open-ended tasks, we propose two open-domain question answering datasets to evaluate the effects of question difficulty and question length on the trends of CCI and QACI. Each

dataset contains 1,000 samples. The hierarchical categorization of questions by abstraction and reasoning depth is shown in Table 4. The hierarchical categorization of questions by length is illustrated in Table 5. All datasets can be found in https://github.com/WinnieShaw/Causal-Dynamic.

E Comparison of Related Works

Table 6 compares representative interpretability paradigms across key capability dimensions. Existing methods typically focus on structural, interventional, or saliency-based analysis, while our approach complements them by modeling causal dynamics along the generation trajectory under black-box settings.

Our work addresses a different layer of the interpretability problem. Unlike locating internal circuits or decomposing mediating components (Conmy et al., 2023; Ortu et al., 2024; Elhage et al., 2021), we characterize the causal dynamics along the generation trajectory: when causal constraints form, how they gradually contract and stabilize as generation progresses, and how this evolution changes under different tasks or training stages. The outputs of our method are thus temporal structures of causal influence strength (CCI/QACI), rather than structural pathways inside the network. Trajectory-level causal characterization is not an explicit objective of existing mechanistic or mediation-based approaches. Therefore, our work is not intended to replace structural interpretability, but to provide a complementary macro-level causal dynamical perspective: mechanistic approaches answer what and where constraints are implemented, whereas our framework addresses how and when these constraints emerge and evolve during generation.

Furthermore, empirical findings consistent with our observations appear across multiple research directions. Studies on semantic entropy and predictive uncertainty show that early-generation tokens tend to exhibit greater volatility and weaker stability (Farquhar et al., 2024). Meanwhile, hallucination and alignment research has repeatedly shown that task type and difficulty substantially affect generation stability, with complex reasoning and long-form tasks being more prone to mismatch or instability (Huang et al., 2025).

Against this backdrop, our CCI/QACI framework goes beyond merely restating these empirical tendencies. It unifies previously scattered observations into a single trajectory-level causal perspective, allowing these phenomena to be compared along the same generation path through the evolution of causal constraints. In this sense, our method complements, rather than replaces, existing behavioral and mechanistic interpretability approaches.

Level	Question Type and Example
Level 1: Factual	Queries about concrete, observable facts or basic properties. *Example: "What basic function do a cat's whiskers serve in daily activities?"*
Level 2: Relational	Questions that explore relationships or causal links between known concepts. *Example: "How is dogs' high olfactory sensitivity related to their survival needs?"*
Level 3: Hypothetical	Reasoning under plausible assumptions, extending known knowledge to new situations. *Example: "If nocturnal vision in felines were further enhanced, how might their hunting behavior change?"*
Level 4: Analytical	Logical reasoning under extreme or counterfactual assumptions, requiring systematic analysis. *Example: "If humans no longer needed sleep, how would work systems and entertainment patterns fundamentally change?"*
Level 5: Abstract	Highly abstract or philosophical speculation that challenges fundamental concepts. *Example: "If humans could perceive all wavelengths of light, how would our definition of reality change?"*

Table 4: Hierarchical categorization of questions by abstraction and reasoning depth.

Level	Question Type and Example
Level 1: Factual	Queries about concrete, observable facts or basic properties, typically short and direct. *Example: "What is the significance of the Sun's core temperature?"*
Level 2: Contextual Factual	Factual questions with an expanded scope that introduce explicit context or target systems. *Example: "What is the specific significance of the Sun's core temperature for the Solar System?"*
Level 3: Relational Reasoning	Questions that require explaining mechanisms or relationships between concepts and their implications. *Example: "How does the Sun's core temperature sustain nuclear fusion, and why is this crucial for understanding stellar evolution?"*
Level 4: Explanatory	Long, background-rich prompts that provide explanatory context and impose strong contextual constraints. *Example: "The stability of the Sun's core temperature is a prerequisite for sustained nuclear fusion, ensuring continuous solar radiation. It directly determines the habitable zone of the Solar System, placing Earth in a suitable environment. It also provides an indispensable energy basis for the origin and persistence of life on Earth."*
Level 5: Abstract Reasoning	Long-form questions with multiple premises and system-level impacts, requiring high-level synthesis and abstract reasoning. *Example: "The Sun's core temperature is the key condition for sustained nuclear fusion. Its stability at around 15 million degrees Celsius not only determines solar radiative output but also affects energy supply to planets. What deeper and critical roles does this stability play in forming the habitable zone of the Solar System and in the origin and evolution of life on Earth?"*

Table 5: Hierarchical categorization of questions by length.

Capability Dimension / Method Category	Activation Patching	Circuit Discovery	Causal Mediation Analysis	Attention/ Saliency Methods	Our Work: Causal Dynamics (CCI/QACI)
Structural Localization (circuits/neurons)	✓	✓	×	×	×
Interventional Causal Explanation (do-level)	✓	×	✓	×	✓
Internal Mediation Analysis	×	×	✓	×	×
Saliency-/Attention-Based Relevance Attribution	×	×	×	✓	×
Generative-Trajectory Causal Dynamics (token-wise evolution)	×	×	×	×	✓
Cross-Task / Cross-Training Robustness	✓	×	×	×	✓
Black-Box Friendliness (logit/probability only)	×	×	×	✓	✓

Table 6: Comparison of different explanation paradigms across key capability dimensions, including Activation Patching (Zhang and Nanda, 2023), Circuit Discovery (Elhage et al., 2021; Conmy et al., 2023), Causal Mediation Analysis (Rocchetti et al., 2024), Attention/Saliency Methods (Ortu et al., 2024; Jain and Wallace, 2019), and our work (Causal Dynamics, CCI/QACI).The highlighted column corresponds to our method.