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

Anonymous ACL submission

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 a 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. The code and datasets are available at: https://anonymous.4open.science/r/Causal-Dynamic-9ECA.

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 (Turpin et al., 2023; 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 that jointly bridges internal and external perspectives to analyze LLMs' generation behavior. We first characterize the unbiased causal effects of both the generated

prefix and the prompt on the current token using a Structural Causal Model (SCM) (Pearl, 2010a). Based on it, we introduce Contextual Causal Influence (CCI) and Question Answer Causal Influence (QACI). CCI characterizes the causal effect of the prefix on the current generated token, revealing internal information dependencies and dynamical evolution. QACI quantifies the causal alignment between generated content and input questions, enabling an externally consistent evaluation of whether generation meets users' requirements. Critically, our metrics require no golden reference answers. This strength makes them uniquely suited for interpreting open-ended generation tasks with no ground truth. Our contributions are as follows:

(1) We propose a causal-dynamic interpretability framework. We introduce two metrics, namely CCI and QACI, to characterize the internal consistency of prefix-level causal constraints and the external consistency of question-answer causal alignment. These metrics capture both the strength of causal influence and how it evolves and changes over time, without requiring reference answers.

(2) Extensive experiments show that CCI establishes stable prefix-level causal constraints early in generation, with its cumulative speed and acceleration rapidly converging. Meanwhile, QACI reveals systematic dependencies between question–answer causal influence and factors such as question difficulty, length, and semantic complexity.

2 Structural Causal Modeling

We model sequence generation as an autoregressive time-step process (Vaswani et al., 2017). First, the original question is transformed into a sequence of basic units through tokenization, making it compatible with the model's input format. We denote this question token sequence as Q. At time $t = 1$, the decoder generates the first token ε_1 conditioned on Q, yielding the initial answer $A_1 = \varepsilon_1$. At $t = 2$, using the generated prefix $A_1 = \varepsilon_1$ and decoding under Q, the model produces ε_2. Then, the model concatenates ε_2 with ε_1 to form $A_2 = \varepsilon_1\varepsilon_2$. Similarly, at $t = 3$, the model uses A_2 and Q to generate ε_3, yielding $A_3 = \varepsilon_1\varepsilon_2\varepsilon_3$. This process continues until an end-of-sequence token (<EOS>) is produced or a maximum length is reached.

Specifically, when generating ε_t, the model computes the conditional distribution $P(\varepsilon_t \mid Q, A_{t-1})$ based on the already generated prefix A_{t-1} and the question Q, and then selects the most likely token for coherent sequence generation. In summary, the probability of each token depends on the previously generated prefix and the original question, namely ε_t is influenced by Q and $A_{t-1} = \varepsilon_1\varepsilon_2 \ldots \varepsilon_{t-1}$. Thus, with the deterministic update $A_t = \mathrm{concat}(A_{t-1}, \varepsilon_t)$, the sequence A_t is influenced by Q and the preceding sequence A_{t-1}.

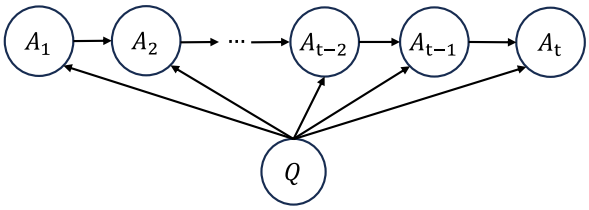


Figure 1: Causal structure of sequence generation in LLMs.

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 express the causal relationships in Figure 1,

we present the SCM equations:

$$Q = f(u_Q), A_1 = f(Q, u_1),$$
$$A_2 = f(Q, A_1, u_2), ..., A_t = f(Q, A_{t-1}, u_t).$$

u_t denotes the exogenous random noise at time step t. We assume the input question Q causally influences both the generated prefix A_{t-1} and the current token A_t. Exogenous noise variables u_t capture stochasticity in decoding but do not act as confounders, as they affect only A_t and are independent of both Q and A_{t-1}. Under this model, Q is the sole confounding variable for estimating the causal effect of A_{t-1} on A_t.

To quantify the causal effects captured by structural equations and make them computable without the do-operator, we derive the following theorem on the unbiased causal effect. Our analysis focuses on causal identification under the SCM, rather than performing physical interventions on the model.

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:*

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

Proof. Here, $P(A_t \mid \mathrm{do}(A_{t-1} = a_{t-1}))$ denotes the interventional distribution that blocks the backdoor path (Pearl, 2009) via Q. To implement the intervention $do(A_{t-1} = a_{t-1})$, we remove the structural equation defining A_{t-1} and replace it with the fixed assignment $A_{t-1} = a_{t-1}$. In the original structural equation, A_{t-1} is determined by the previous sequence A_{t-2} and the encoded question Q, namely $A_{t-1} = f\big(Q, A_{t-2},\ u_{t-1}\big)$.

To implement the intervention, remove this equation so that A_{t-1} is no longer affected by Q and A_{t-2}, and is directly assigned the fixed value a_{t-1}. Next, substitute $A_{t-1} = a_{t-1}$ into all subsequent equations. For example, consider the generation equation of A_t, we have $A_t = f\big(Q, A_{t-1},\ u_t\big)$. After the intervention $\mathrm{do}(A_{t-1} = a_{t-1})$, it becomes $A_t = f\big(Q, a_{t-1},\ u_t\big)$. The generation of A_t depends only on Q, the fixed value a_{t-1}, and random noise u_t, eliminating the confounding influence of Q via A_{t-1}. This operation cuts off the correlation between A_{t-1} and Q, blocking the backdoor path $A_{t-1} \leftarrow Q \rightarrow A_t$. Thus, Theorem 1's expression removes the confounding effect of Q, allowing $P\big(A_t \mid \mathrm{do}(A_{t-1} = a_{t-1})\big)$ to reflect the true causal effect of A_{t-1} on A_t, rather than a mere statistical correlation. This procedure distinguishes the "spurious association" from the "true causal effect" in LLM generation, providing a foundation for unbiased effect estimation. □

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:*

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

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:

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

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

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

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

$$\textit{QACI}(A_t = a_n \mid Q = q) = \log_2 \frac{P(A_n = a_n \mid do(Q = q))}{P(A_n = a_n)}. \quad (4)$$

Furthermore, we have the following theory on QACI computability:

Theorem 4. *The calculation for QACI is:*

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

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

$$\mathrm{QACI}(A_t = a_n \mid Q = q) = \log_2 \frac{P(A_n = a_n \mid Q = q)}{P(A_n = a_n)}. \quad (6)$$

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:

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

Similarly, we have:

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

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

Specifically, $\mathrm{QACI}(A_t = a_n \mid Q = q)$ represents the log-ratio of the probability of generating the answer with question guidance versus without question guidance. The Appendix B 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, widely regarded as 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 necessary for causal interpretation. In autoregressive generation, temporality is inherently satisfied: the input question precedes the answer, 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 CNN/DailyMail dataset (See et al., 2017). We evaluated LLMs including T5 family (Raffel et al., 2020), Qwen2.5-Instruct-7B (Team, 2024), Mistral-7B-v0.1(Jiang et al., 2023) and Llama3-8B base model (Meta AI, 2024). Experiments are conducted on an NVIDIA A100 GPU.

Model	CCI↑	Cosine↑	Jaccard↑	Topic-JS↓	QACI↑	ROUGE-L↑	CHRF↑	TER↓
T5-small	13.83	0.74	0.23	0.73	502.13	27.72	33.53	90.03
T5-base	13.21	0.57	0.05	0.92	594.38	27.21	35.93	90.36
T5-large	12.68	0.62	0.05	0.93	1024.95	31.27	41.28	87.21
Mistral-7B-v0.1	27.06	0.70	0.10	0.57	164.49	17.34	29.50	170.15
Qwen2.5-Instruct-7B	19.66	0.78	0.06	0.57	342.36	24.81	40.68	120.57
LLaMA 3-8B	18.49	0.67	0.04	0.64	168.27	20.36	35.87	157.85
Pearson r		0.44	-0.08	-0.83		0.91	0.60	-0.83
Spearman ρ		0.66	0.37	-1.00		0.94	0.71	-0.94

Table 1: Model-level comparison of CCI/QACI and conventional metrics, ↑ indicates that higher values are better, while ↓ indicates that lower values are better.

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 create a validation subset of 50 samples for CCI and 100 for QACI by shuffling the data with a random seed of 42 and repeating the evaluations 10 times. After averaging model performance, we compute Pearson and Spearman correlation coefficients between the metrics (Benesty et al., 2009; Spearman, 1961).

Table 1 shows the covariations between CCI/QACI and other correlation-based metrics. CCI shows significant correlations with Cosine and Topic-JS. The stable negative correlation with Topic-JS indicates that strong contextual causal dependence leads to reduced topic drift during generation, while the weak correlation with Jaccard suggests that surface-level lexical overlap fails to capture contextual causal dependencies.

QACI is positively correlated with ROUGE-L/CHRF and negatively correlated with TER, indicating a stronger question–answer causal influence relates to better content coverage and lower edit distance compared to references. Overall, these results show that although our metrics do not rely on reference answers, they maintain strong covariation with reference-based metrics. Moreover, CCI/QACI can capture causal influence characteristics not directly reflected by traditional measures, demonstrating complementarity and validation.

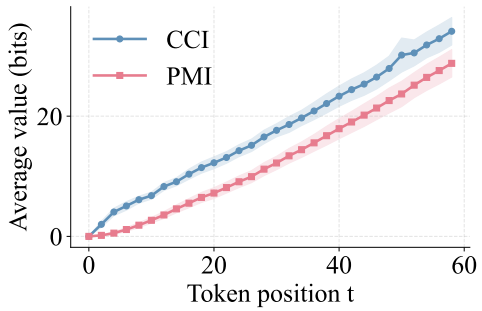


Figure 2: Average values of causal and correlational metrics.

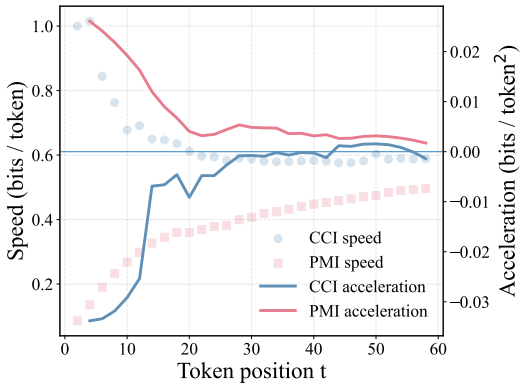
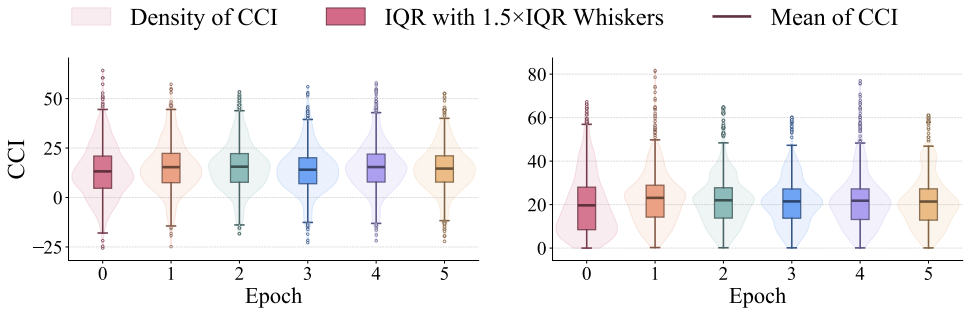


Figure 3: Speed and acceleration dynamics of causal and correlational metrics.

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 dependence between two random variables by comparing their joint probability to the product of their marginals. In the generation setting,we use PMI to



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

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

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 C 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 consistently higher than 0. It suggests that the observed dependence in PMI mainly comes from the accumulation of longer contexts, specifically 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.

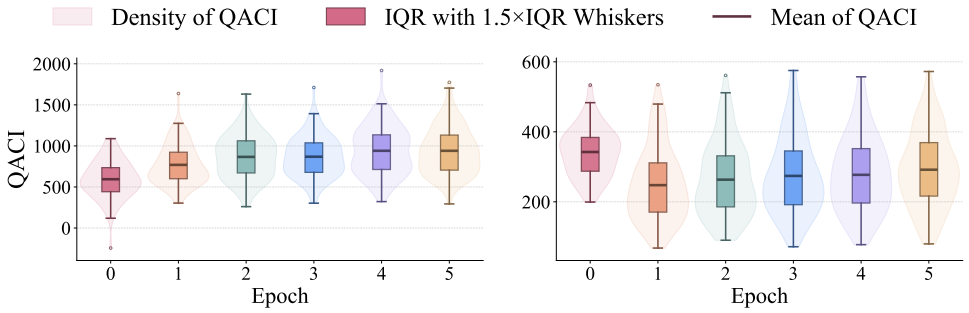
4.2 Multi-Task Evaluation

4.2.1 Fine-Tuning

We finetune 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.

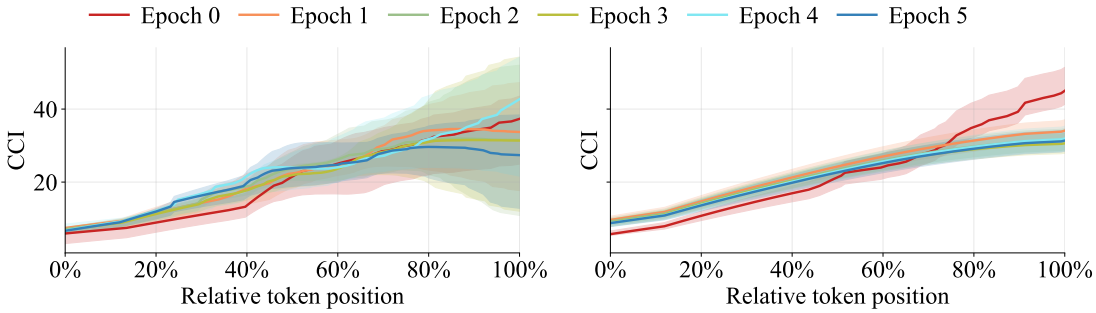
(1) Varying trajectories of CCI. Figure 4 illustrates that during training, both models' CCI remain above zero, indicating that generations are strongly constrained by prior context. From epoch 0 to epoch 1, CCI significantly increases, followed by minor fluctuations from epoch 2 to epoch 5. Both models 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.



(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.

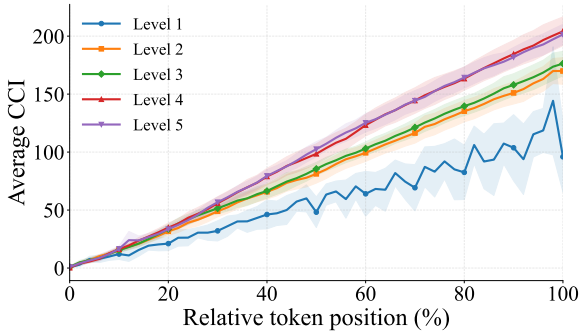


Figure 7: CCI across question difficulty levels.

4.2.2 Generation

(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) Open-ended tasks. We evaluate some reference free Question-Answer tasks using Qwen2.5-Instruct-7B. The dataset sizes and hierarchical categorizations of questions by abstraction and length are shown in Appendix C, respectively.

A) Question difficulty on CCI. We firstly 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 questions in level 1, CCI grows slowly and displays larger fluctuations, indicating a weaker causal constraint of the prefix on current sequence. As question difficulty level 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 aver-

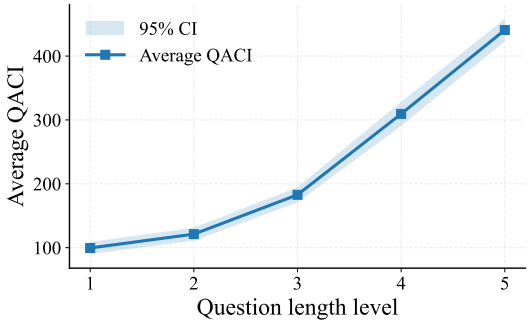


Figure 8: QACI dynamics across question length levels.

age 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

Interpretability research for LLMs can be broadly divided into internal and external perspectives, depending on whether explanation validity is judged by fidelity to internal mechanisms or by usefulness in external applications.

Internal interpretability seeks to reveal the true generative basis and causal mechanisms underlying model behavior. Prior works show that commonly used intermediate signals, such as attention weights or generated rationales, can be misleading or post-hoc and do not necessarily reflect the actual decision process (Jain and Wallace, 2019; Turpin et al., 2023). Consequently, later studies emphasize verifiable faithfulness through interventions like deletion, counterfactual analysis, and manipulation at the parameter-level or activation-level (Zhang and Nanda, 2023), as well as mechanistic interpretability methods that localize model behavior to internal structures through circuit discovery and activation patching (Conmy et al., 2023; Ortu et al., 2024). Marks et al. (2024) analyzes how RLHF training shapes feedback-related activation patterns within the model. Related work also explores representation and training dynamics to explain how internal predictions and knowledge emerge over time (Belrose et al., 2023).

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). From a user-centered perspective, 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). In model evaluation and alignment, related work has revealed systematic biases in reference-free evaluation methods, highlighting that external assessments require interpretability and calibration (Deutsch et al., 2022). Other studies decompose 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, offering interpretable signals for external governance and risk control (Madhavan et al., 2023).

However, existing interpretability approaches face a critical shortcoming. They are unable to unify internal consistency and external alignment within a single framework. What is more, they lack the ability to model causal effects and temporal dynamics simultaneously. Addressing this gap, our work jointly models internal consistency and external alignment from a causal and dynamic perspective within a unified framework.

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 cumulative speed and acceleration rapidly converging. Meanwhile, QACI reveals systematic dependencies between question–answer causal influence and factors such as question difficulty, length, and semantic complexity. These findings indicate that LLM generation follows time-varying causal influences that static or correlation-based measures fail to capture, making causal–dynamic analysis a practical tool for evaluating and diagnosing generative models.

Limitation

This framework focus 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.

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*.

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

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

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.

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.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965.

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*.

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 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.

B 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.

C 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. The size of each dataset is 1000. The hierarchical categorization of questions by abstraction and reasoning depth is shown in Table 2. And the hierarchical categorization of questions by length is illustrated in Table 3. All datasets can be found in https://anonymous.4open.science/r/Causal-Dynamic-9ECA.

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 2: 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 3: Hierarchical categorization of questions by length.