From Intuition to Verification: Cognitive Neuro-Symbolic Reasoning for Document-level Event Causality Identification

Anonymous ACL submission

Abstract

Document-level Event Causality Identification (DECI) aims to infer causal relations between events distributed across long documents, where causality is often implicit and evidence is fragmented. Existing approaches typically follow two paradigms: structure-based models that emphasize predefined graphs but struggle to capture implicit semantic relations, and generative large language models (LLMs) that flexibly propose causal hypotheses yet lack reliable global verification. Inspired by the cognitive transition from intuition to verification, we propose COgnitive Verification for Event Reasoning (COVER), a cognitive neuro-symbolic framework that explicitly integrates intuitive hypothesis generation with structured verification for DECI. COVER treats causal reasoning as a closed-loop process. In the intuition stage, an LLM serves as a variational prior to generate causal hypotheses and retrieve supportive commonsense knowledge, which is filtered via entropy-aware knowledge anchoring. In the verification stage, these hypotheses are embedded into a document-level neuro-symbolic causal graph and evaluated under global structural constraints with uncertainty-aware reasoning, enabling unreliable hypotheses to be refined rather than directly accepted. Experiments on CEC 2.0 and MAVEN-ERE demonstrate that COVER consistently outperforms strong baselines, with notable gains on implicit and long-range causal relations.

1 Introduction

Real-world text streams such as news, incident reports, and clinical narratives often describe chains of events whose causal links must be identified for downstream reasoning and decision making, motivating Event Causality Identification (ECI). This capability is vital for diverse applications, such as financial risk analysis (Sakaji and Izumi, 2023) and clinical diagnosis support (Gopalakrishnan et al., 2024). Recent work has improved

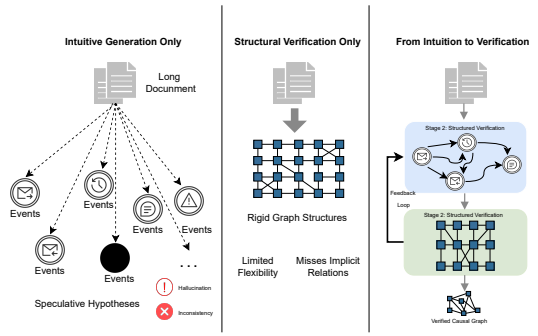

Figure 1: Motivation of document-level event causality reasoning, from intuition to verification.

sentence-level ECI by enriching event semantics and modeling structured interactions between event pairs (Hu et al., 2023; Pu et al., 2023; Li et al., 2024). However, practical scenarios are frequently document-level, where causes and effects may be far apart, interact with multiple intermediate events, and depend on global discourse structure, making Document-level ECI (DECI) fundamentally harder than SCEI and still under-explored (Wang et al., 2024; Hashimoto, 2019).

Existing methods for DECI mainly fall into two lines. **One line** in Figure 1 (*middle*), builds document-level structures, such as event graphs and interaction networks, and performs causal inference with graph-structured reasoning or iterative representation updates (Phu and Nguyen, 2021; Chen et al., 2025a, 2022; Zhao et al., 2021). These models explicitly propagate information across sentences and offer some interpretability, yet their reasoning is often constrained by predefined structures and local message passing, making implicit causal relations with weak surface cues difficult to capture. **Another line** in Figure 1 (*left*), leverages large language models (LLMs) or prompt-driven frameworks for document-level causal reasoning, benefiting from strong semantics to hypothesize long-range causal links (Wei et al., 2022). However,

generative reasoning is frequently open-ended: hypotheses can be produced without explicit verification against document evidence and global consistency, leading to hallucinated relations and unreliable conclusions(Gao et al., 2023). Overall, prior DECI approaches either emphasize structural modeling at the expense of semantic flexibility, or rely on generative intuition without a reliable verification mechanism.

From a cognitive perspective, as shown in Figure 1 (*right*), *reliable document-level causal reasoning follows a two-stage process that moves from intuitive hypothesis generation to deliberate verification* (Evans and Stanovich, 2013). When reading long documents, humans typically form plausible causal hypotheses based on semantic cues and background knowledge, and subsequently assess whether these hypotheses are supported by document-wide evidence and structural consistency (Reyna and Brainerd, 2011). This transition is particularly critical in DECI, where causal relations are often implicit, evidence is dispersed, and spurious links can easily arise. At the document level, relying solely on intuition or verification becomes insufficient, as neither semantic flexibility nor structural rigor alone can ensure reliable reasoning. These observations motivate a cognitive paradigm that tightly couples intuitive generation with structured verification, enabling hypotheses to be explicitly verified and revised under global document-level constraints.

To address the above challenges, we propose COgnitive Verification for Event Reasoning (COVER), **a cognitive neuro-symbolic framework** for DECI that bridges intuitive hypothesis generation and structural verification. COVER treats causal reasoning as a closed-loop process: it first leverages a large language model to generate intuitive causal hypotheses and retrieve supportive background knowledge, and then performs structured verification by integrating these hypotheses into a document-level neuro-symbolic causal graph. Verification is guided by global structural constraints and uncertainty-aware reasoning, allowing unreliable hypotheses to be identified and refined rather than directly accepted. This iterative interaction enables COVER to balance semantic flexibility with structural consistency, mitigating hallucinated causality while preserving coverage of implicit relations. Extensive experiments on multiple benchmarks demonstrate that COVER consistently outperforms strong structure-based and generative baselines.

Our contributions are summarized as follows:

- We design an uncertainty-aware verification mechanism that leverages evidential reasoning to assess the reliability of generated causal hypotheses and guide iterative refinement, improving robustness against hallucinated causality.
- We propose COVER, a cognitive neuro-symbolic framework that couples generative intuition with document-level structural verification in a closed loop, enabling hypotheses to be verified and revised under global constraints.
- We design an uncertainty-aware verification mechanism that leverages evidential reasoning to assess the reliability of generated causal hypotheses and guide iterative refinement, improving robustness against hallucinated causality.
- Extensive experiments show that COVER consistently outperforms strong structure-based and generative baselines, with notable gains on implicit and long-range causal links.

2 Methodology

2.1 Overview

Inspired by the dual-process theory of human cognition, we model causal reasoning as an interplay between an intuitive generation stage (System 1) and a deliberate verification stage (System 2). COVER models causal reasoning as a closed-loop interaction between variational intuition and structured verification. In the intuition stage, a large language model acts as a variational prior to generate latent causal hypotheses and retrieve supportive commonsense knowledge from the document context. In the verification stage, these hypotheses are embedded into a document-level neuro-symbolic causal graph that integrates syntactic structure, anchored knowledge, and latent temporal relations, and are evaluated under global constraints via relational graph attention. An evidential cognitive cycle further quantifies epistemic uncertainty and grounded consistency, producing feedback signals to refine unreliable hypotheses. Through this iterative process, COVER balances semantic flexibility with structural rigor, enabling reliable inference of implicit and long-range causal relations.

2.2 Problem Formulation

Let $\mathcal{D} = \{w_1, w_2, ..., w_N\}$ denote a document consisting of N tokens. The document contains a set of event mentions $\mathcal{E} = \{e_1, e_2, ..., e_M\}$, where each

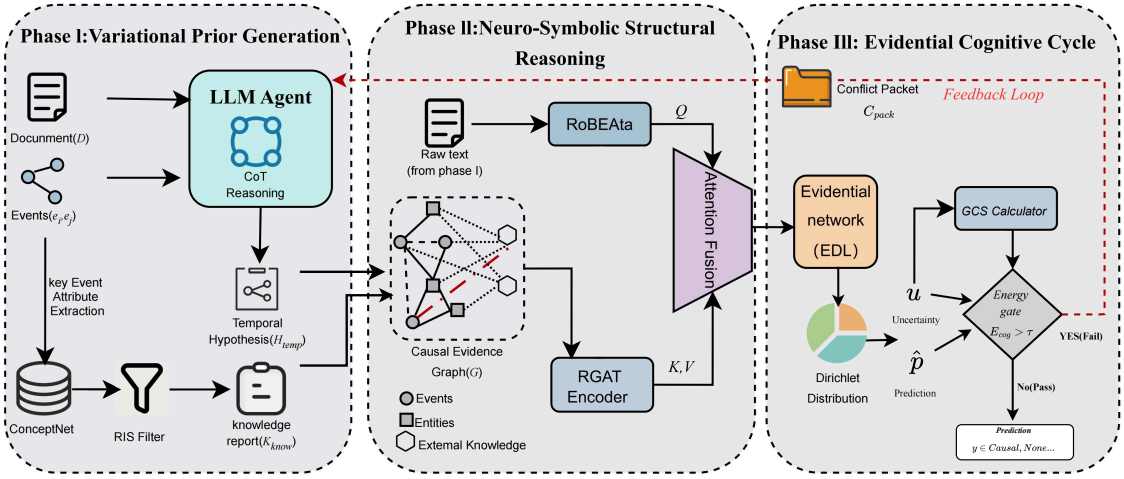

Figure 2: Overview of the COVER framework.

mention e_i corresponds to a specific span of tokens in $\mathcal{D}$.

Given a target ordered pair of events (e_i, e_j), the goal of Document-level Event Causality Identification (DECI) is to determine the causal relation y between them from a predefined label set $\mathcal{Y}$. Following standard protocols, we define $\mathcal{Y} = \{\text{CAUSE}, \text{CAUSED\_BY}, \text{NONE}\}$. Specifically, $y = \text{CAUSE}$ indicates that e_i is the cause of e_j, while $y = \text{CAUSED\_BY}$ indicates the reverse.

Formally, the task is to estimate the conditional probability $P(y|\mathcal{D}, e_i, e_j)$. This is challenging in the document-level setting as causal dependencies often span across non-adjacent sentences with implicit evidence, requiring the model to capture global structural interactions rather than relying on local surface features.

2.3 Variational Prior Intuition (Phase I)

To bridge the semantic gap inherent in long-distance causal reasoning, we model the inference process through a latent variable framework. We introduce a *Latent Prior Intuition* variable $\mathcal{Z} = \{H_{temp}, K_{know}\}$, where:

- H_{temp} represents the latent temporal hypothesis (e.g., event ordering and implicit bridging connections) generated to explicitly connect distant event pairs.
- K_{know} represents external commonsense knowledge paths anchored to document entities, providing necessary background context.

In this phase, we employ LLM to server as a variational prior G_ϕ to generate the prior intuition distribution $P(\mathcal{Z} \mid \mathcal{D})$. This corresponds to the "Intuitive System" in cognitive theory, offering rapid but potentially noisy semantic proposals.

Latent Hypothesis Generation. We leverage the cognitive ability of LLMs to propose a preliminary causal structure. In iteration t, the generator samples a temporal hypothesis $H_{temp}^{(t)}$ and a set of retrieval queries $K_{query}^{(t)}$ conditioned on the document and any feedback from the previous cycle:

$$H_{temp}^{(t)}, K_{query}^{(t)} \sim G_\phi \left(\cdot \mid \mathcal{D}, e_i, e_j, \mathcal{C}_{pack}^{(t-1)} \right) \quad (1)$$

we employ Chain-of-Thought (CoT) prompting to guide the generation. Crucially, $\mathcal{C}_{pack}^{(t-1)}$ denotes the *Conflict Packet* derived from the Evidential Cognitive Cycle (detailed in Sec. 2.5). In the initial pass ($t = 0$), $\mathcal{C}_{pack}$ is empty. In subsequent iterations, it serves as a negative constraint (e.g., "Avoid the ambiguous link between A and B"), forcing the LLM to re-sample a more plausible hypothesis, thereby correcting reasoning errors dynamically.

Entropy-aware Knowledge Anchoring. While LLMs provide fluent hypotheses, they suffer from knowledge sparsity and hallucinations(Cheng et al., 2024). To ground the generated hypothesis, we retrieve external commonsense paths from ConceptNet(Speer et al., 2017) using the generated queries K_{query}. To filter out irrelevant or generic noise, we introduce an **Entropy-aware Knowledge Anchoring** mechanism. We calculate the *Relation Information Score* (RIS) for each retrieved path r:

$$RIS(r) = \text{APMI}(h, t) + \beta \cdot \text{IRF}(r) \quad (2)$$

where APMI(h,t) (Average Pointwise Mutual Information) measures the semantic binding strength between the head and tail concepts, and IRF(r) (Inverse Relation Frequency) quantifies the information gain of the relation type (penalizing overly generic relations like *RelatedTo*). Only knowledge paths satisfying $RIS \geq \tau_{RIS}$ are retained to form the high-fidelity Knowledge Report K_{know}. Consequently, the instantiated latent variable $\mathcal{Z} = \{H_{temp}, K_{know}\}$ provides a structurally grounded explanation for the subsequent neuro-symbolic verification.

2.4 Neuro-Symbolic Structural Verification (Phase II)

This phase maps the unstructured latent variable $\mathcal{Z}$ into a dense vector space for logical verification.

Symbolic Graph Construction. We construct a Multi-view Evidence Graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. The node set $\mathcal{V}$ comprises event mentions and entity concepts. The edge set $\mathcal{E}$ integrates three logical views: (1) Syntactic Edges (r_{syn}) from dependency parsing; (2) Knowledge Edges (r_{know}) from K_{know}; and (3) Latent Temporal Edges (r_{latent}) derived from H_{temp}. The latent edges act as "wormholes," allowing direct message passing between temporally related but distinct events.

Dual-Stream Initialization. Before reasoning, we initialize the node representations via a Dual-Stream Encoding strategy. We feed the raw document $\mathcal{D}$ into a RoBERTa encoder to obtain the sequence embeddings. We then extract the feature vectors corresponding to the indices of event mentions and entities to initialize the graph nodes $h^{(0)}$, while the `[CLS]` token serves as the global context E_S for the subsequent fusion.

Relational Graph Attention (RGAT). We employ an RGAT to propagate structural evidence. The node update rule for node i at layer l is:

$$\boldsymbol{h}_i^{(l+1)} = \sigma\left(\sum_{r\in\mathcal{R}}\sum_{j\in\mathcal{N}_i^r} \alpha_{ij}^{(r)} \mathbf{W}_r \boldsymbol{h}_j^{(l)}\right) \quad (3)$$

Crucially, the attention weight $\alpha_{ij}^{(r_{latent})}$ functions as a differentiable gate. If the LLM-generated latent edge contradicts the local semantic context encoded in the graph, the network suppresses this weight, effectively denoising the variational prior. Finally, we fuse the global textual context $\mathbf{E}_S$ (from the RoBERTa [CLS] token) with the graph representation $\boldsymbol{h}_{\mathcal{G}}$ via Multi-Head Cross-Attention to obtain the task-specific evidence vector $\boldsymbol{v}_{task}$.

2.5 Evidential Cognitive Cycle (Phase III)

To enable the "Reflective System", we replace standard point-estimation (Softmax) with Evidential Deep Learning (EDL) (Sensoy et al., 2018) to quantify epistemic uncertainty and trigger feedback.

Evidential Uncertainty Modeling. We model the classification probability as a Dirichlet distribution $Dir(p|\boldsymbol{\alpha})$. The network predicts evidence counts $\mathbf{e} = \text{Softplus}(\mathbf{W}_o v_{task})$, which parameterize the distribution as $\boldsymbol{\alpha} = \mathbf{e} + 1$.

The expected probability for the k-th class is calculated as $\hat{p}_k = \alpha_k / S$. Subsequently, the Epistemic Uncertainty u is derived as:

$$u = \frac{K}{S}, \quad \text{where} \quad S = \sum_{k=1}^{K} \alpha_k \quad (4)$$

where K is the number of classes. A high u signifies total ignorance or conflicting evidence.

Grounded Consistency Score (GCS). To explicitly penalize hallucinations, we introduce the GCS to measure the factual alignment between the generated hypothesis and the source document. We decompose $\mathcal{Z}$ into atomic claims $\{a_n\}_{n=1}^N$ and align them with evidence snippets $\{E_n\}_{n=1}^N$. The score is calculated as:

$$GCS(\mathcal{Z}, \mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N} \big(P_\theta(\text{entail} \mid a_n, E_n) - \lambda \cdot P_\theta(\text{contradict} \mid a_n, E_n)\big) \quad (5)$$

where $\lambda > 1$ is the *Hallucination Penalty Coefficient*. This asymmetric penalty ensures the model is risk-averse regarding factual fabrications, heavily penalizing contradictions.

Energy-based Feedback Loop. We define a Cognitive Energy Function E_{cog} to govern the inference flow. Mathematically, the Energy Gate acts as a filter for the predicted probability $\hat{p}$. It evaluates the reliability of $\hat{p}$ based on the cognitive energy:

$$E_{cog}^{(t)} = u^{(t)} + \gamma \cdot (1 - GCS^{(t)}) \quad (6)$$

The inference follows a dynamic gating mechanism:

- **Pass:** If $E_{cog} \leq \tau$, the system accepts the hypothesis and outputs the final prediction $\hat{y} = \arg\max_k(\hat{p}_k)$.

- **Reject & Refine:** If $E_{cog} > \tau$, a feedback loop is triggered. We identify the *Conflict Packet* $\mathcal{C}_{pack}$ (nodes with high uncertainty entropy) and feed it back to Phase I to re-sample $\mathcal{Z}^{(t+1)}$, explicitly correcting the reasoning path.

2.6 Optimization Objective

The framework is trained end-to-end using a hybrid objective function:

$$\mathcal{L} = \mathcal{L}_{EDL} + \lambda_{KL}\mathcal{L}_{KL} + \lambda_{align}\mathcal{L}_{align} \quad (7)$$

$\mathcal{L}_{EDL}$ minimizes the Bayes risk for classification. $\mathcal{L}_{KL}$ is the KL-divergence regularizer that penalizes overconfidence by forcing the Dirichlet distribution towards a uniform prior for incorrect predictions. $\mathcal{L}_{align}$ enforces semantic alignment between the neuro-symbolic graph and the raw text representations.

3 Experiments

3.1 Experimental Settings

Datasets involve Chinese and English:

CEC 2.0 (Wang et al., 2015) contains 332 Chinese news reports on emergency events spanning five breaking-news types (for example, earthquakes and fires).

MAVEN-ERE (Wang et al., 2022) includes 3,555 documents with 85912 event mentions, 97,521 intra-sentence and 1,226,168 inter-sentence event pairs. Since the original test set lacks gold labels, we split the development set into new dev and test sets.

Evaluation Metrics adopt Precision (P), Recall (R), and F1-score (F1). For a fair comparison, we report the average results over three random runs.

Evaluation Protocol To rigorously assess the neuro-symbolic logical consistency, we implement a *Strict Directional Evaluation*. Unlike settings that solely detect link existence, our protocol considers a prediction correct if and only if it strictly matches the ground truth label $y \in \{\text{CAUSE}, \text{CAUSED\_BY}\}$ for the ordered pair (e_i, e_j). Crucially, a reversed direction (e.g., predicting CAUSED_BY instead of CAUSE) is penalized as a False Positive.

Implementation Details . We employ RoBERTa-base (Conneau et al., 2019) as the backbone for textual encoding and the Neuro-Symbolic Graph encoder. For the Variational Prior generation and the Cognitive Cycle, we utilize DeepSeek-R1-14B (DeepSeek-AI et al., 2025a) tuned via LoRA (Hu et al., 2021) (rank=32, alpha=64). The model is optimized using AdamW with a learning rate of 1e-5 for the encoder and 5e-5 for the LoRA parameters. The batch size is set to 16. We set the conflict penalty $\lambda = 1.5$, the cognitive energy balancing coefficient $\gamma = 1.0$, and the maximum cognitive iterations $I_{max} = 3$. All experiments are conducted on 2 NVIDIA RTX 4090 GPUs.

3.2 Baselines

We compare COVER with three categories:

Sentence-level ECI. We first compare with standard pre-trained language models (PLMs), **BERT** (Devlin, 2018) and **RoBERTa** (Liu et al., 2019), which serve as fundamental sequence encoders. To capture structural dependencies, we include **SemSIn** (Hu et al., 2023) and **SGT** (Zhang et al., 2022), which incorporate AMR-based semantic graphs and schema guidance, respectively. Additionally, we consider methods optimized for the CEC dataset: **Siamese-Bi-LSTM** (Gupta et al., 2024), which utilizes siamese networks for semantic interaction, and **CERMiner** (Chen et al., 2025b), which enforces consistency constraints for event relation extraction.

Document-level ECI. We employ **Longformer** (Beltagy et al., 2020) to handle long-range contexts. Graph-based approaches are included to model high-order reasoning: **RichGCN** (Phu and Nguyen, 2021) constructs interaction graphs with syntactic and discourse dependencies; **ERGO** (Chen et al., 2022) and **PPAT** (Liu et al., 2023) utilize relational graph transformers and pairwise attention networks to capture multi-hop causality. Furthermore, we compare with recent advanced paradigms: **DiffusECI** (Man et al., 2024), a diffusion-based generative framework, and **iLIF** (Liu et al., 2024), an iterative learning framework.

Large Language Models. We also assess the zero-shot reasoning capabilities of representative LLMs, including **GPT-3.5/4/4o** (Achiam et al., 2023), **DeepSeek-V3/R1** (DeepSeek-AI et al., 2025c,b), and **LLaMA-2-7B** (Touvron et al., 2023), which are prompted with ECI-specific instructions to find causal relations.

3.3 Main Results

Table 1 and Table 2 present the comparison results on MAVEN-ERE and CEC 2.0, respectively.

On the CEC 2.0 dataset (Table 2), COVER achieves a new state-of-the-art F1 score of **85.6%**, outperforming the strong baseline Siamese-Bi-LSTM by 1.8 points. The high precision (89.8%) demonstrates that our Cognitive Cycle effectively filters out false positives generated by noise.

On the more challenging MAVEN-ERE dataset (Table 1), COVER exhibits a substantial performance margin. It achieves an F1 score of 60.9%, surpassing the previous best method LearnDA (52.6%) by 8.3%. Notably, zero-shot baselines like GPT-3.5-turbo show high recall (80.2%) but very low precision (27.6%), indicating severe hallucination issues in long documents. In contrast, COVER maintains a balanced precision (54.2%) and recall (69.5%). This validates that our Neuro-Symbolic Graph successfully captures long-distance dependencies, while the Evidential Cognitive Cycle significantly suppresses hallucinations.Furthermore, to investigate the model's robustness against logical inversions (i.e., distinguishing $e_i \rightarrow e_j$ from $e_j \rightarrow e_i$), we conducted a fine-grained *Directional Error Rate (DER)* analysis. As detailed in **Appendix E**, COVER achieves a significantly lower DER (3.8%) compared to discriminative baselines ($> 18\%$), confirming that the cognitive feedback loop effectively corrects directional hallucinations.

3.4 Ablation Study

To thoroughly verify the contribution of each component in COVER, we conduct ablation studies on both the CEC 2.0 and MAVEN-ERE datasets. As shown in Table 3 and Table 4, removing any module leads to performance degradation, validating the necessity of our neuro-symbolic architecture.

Impact of Evidential Cognitive Cycle. Removing the evidence-based feedback loop ("w/o Cognitive Cycle") results in a notable drop in performance, particularly in Precision on the CEC 2.0 dataset (89.8% $\rightarrow$ 84.5%). This empirical evidence confirms that the iterative correction mechanism acts as a crucial gatekeeper, effectively filtering out hallucinations and false positives generated by the generative prior.

Impact of Neuro-Symbolic Graph. Replacing the RGAT and graph construction with a simple RoBERTa encoder ("w/o Neuro-Symbolic Graph") leads to the most significant decline in Recall across both datasets (e.g., -4.8% on CEC 2.0 and -9.3% on MAVEN-ERE). This validates that the structured graph representation—specifically the latent temporal edges—is essential for establishing "wormhole" connections to capture long-distance dependencies that pure text encoders fail to recognize.

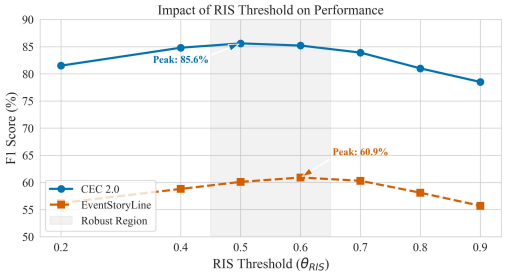

Figure 3: Impact of the Relation Information Score (RIS) Threshold θ_{RIS} on performance.

Impact of Knowledge Anchoring. Removing the external knowledge retrieval and RIS filter ("w/o Knowledge Anchoring") consistently degrades F1 scores. This highlights the value of structured commonsense knowledge in bridging the semantic gap for implicit causality samples, where textual cues alone are insufficient.

3.5 Parameter Sensitivity Analysis

We investigate the sensitivity of COVER to three critical hyperparameters governing the neuro-symbolic reasoning process: the knowledge filtering threshold θ_{RIS}, the cognitive energy gate τ, and the maximum reasoning steps I_{max}.

Impact of Knowledge Filtering (θ_{RIS}). The RIS threshold θ_{RIS} controls the density of the retrieved knowledge graph. As illustrated in Figure 3, the F1 score exhibits an inverted U-shaped trend, peaking at $\theta_{RIS} \in [0.5, 0.6]$.

Lower thresholds ($\theta_{RIS} < 0.4$) introduce excessive commonsense noise, diluting the causal signal.

Higher thresholds ($\theta_{RIS} > 0.7$) result in overly sparse graphs, severing potential bridging paths for implicit causality.

This empirical evidence confirms the necessity of our entropy-aware filtering mechanism to achieve an optimal trade-off between information gain and structural sparsity.

3.6 Analysis of Implicit Reasoning

A critical challenge in DECI is identifying causal relations that lack overt linguistic markers. To eval-

Model	Intra-sentence			Inter-sentence		
	P(%)	R(%)	F1(%)	P(%)	R(%)	F1(%)
BERT	47.8	57.2	52.1	36.8	29.2	32.6
RoBERTa	45.9	48.8	47.2	42.3	45.1	43.6
Longformer*	71.7	47.5	57.2	56.1	38.6	45.7
RichGCN	49.2	63.0	55.2	39.2	45.7	42.2
ERGO	63.1	65.3	64.2	48.7	62.0	54.6
SemSIn	49.4	51.2	50.5	–	–	–
DiffusECI	47.5	60.3	51.7	45.5	53.0	48.2
PPAT	37.9	66.7	47.4	32.6	39.3	35.6
iLIF	74.4	51.5	60.9	**67.1**	49.2	56.8
LLaMA2-7B	12.1	50.7	19.5	–	–	–
GPT-3.5-turbo	19.9	85.8	32.3	–	–	–
GPT-4	22.5	**92.4**	36.2	–	–	–
GPT-4o-mini	–	–	–	30.8	48.3	34.4
DeepSeek-Chat	–	–	–	37.8	53.2	40.6
DeepSeek-R1	–	–	–	43.7	59.2	37.8
COVER (Ours)	**76.2**	68.8	**72.3**	54.2	**69.5**	**60.9**

Table 1: Performance comparison on the MAVEN-ERE dataset.

Method	P	R	F1
CSNN	70.6	54.9	61.7
Bi-LSTM	76.4	71.9	74.0
Bert-LSTM	66.4	63.1	64.1
MCNN	80.6	83.6	82.1
SGT	80.8	73.3	76.5
CERMiner	84.8	72.7	78.2
GPT-3.5-turbo	76.1	68.3	72.0
Siamese-Bi-LSTM	83.0	**84.6**	83.8
COVER (Ours)	**89.8**	81.6	**85.6**

Table 2: Experimental results on **CEC 2.0** (%).

Configuration	P (%)	R (%)	F1 (%)
COVER (Full Model)	**89.8**	81.6	**85.6**
w/o Evidential Cognitive Cycle	84.5	**82.0**	83.2
w/o Neuro-Symbolic Graph	87.2	76.8	81.7
w/o Knowledge Anchoring	88.5	79.1	83.5

Table 3: Ablation study results on **CEC 2.0**. The full COVER model achieves the best balance.

Configuration	P (%)	R (%)	F1 (%)
COVER (Full Model)	54.2	**69.5**	**60.9**
w/o Cognitive Cycle	56.1	61.4	58.6
w/o Neuro-Sym. Graph	56.5	60.2	58.3
w/o Know. Anchoring	57.4	61.9	59.5

Table 4: Ablation study results on **MAVEN-ERE**.

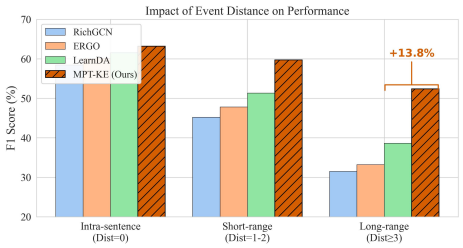

Figure 4: Impact of Event Distance on Performance.

uate robustness against this "semantic gap," we partition the MAVEN-ERE test set into **Explicit Causality** (containing markers like "because") and **Implicit Causality** (zero-lexical overlap). Detailed partition criteria are provided in Appendix B.

Quantitative Results. Table 5 compares COVER against representative discriminative (RichGCN) and generative (GPT-4) baselines. Discriminative models suffer a catastrophic drop (-32.8%) on implicit cases, confirming that syntactic dependencies struggle without explicit surface patterns. Notably, COVER outperforms the zero-shot GPT-4 by **4.7%** on the implicit subset. This suggests that while LLMs possess vast parametric knowledge, they remain prone to ungrounded hallucinations; our framework mitigates this by anchoring the generative prior with retrieved external knowledge (System 1). Furthermore, COVER maintains high precision on explicit cases by leveraging the *Evidential Cognitive Cycle* (System 2) as a logical filter, effectively distinguishing true causality from non-causal temporal noise without sacrificing the recall boost from latent structure induction.

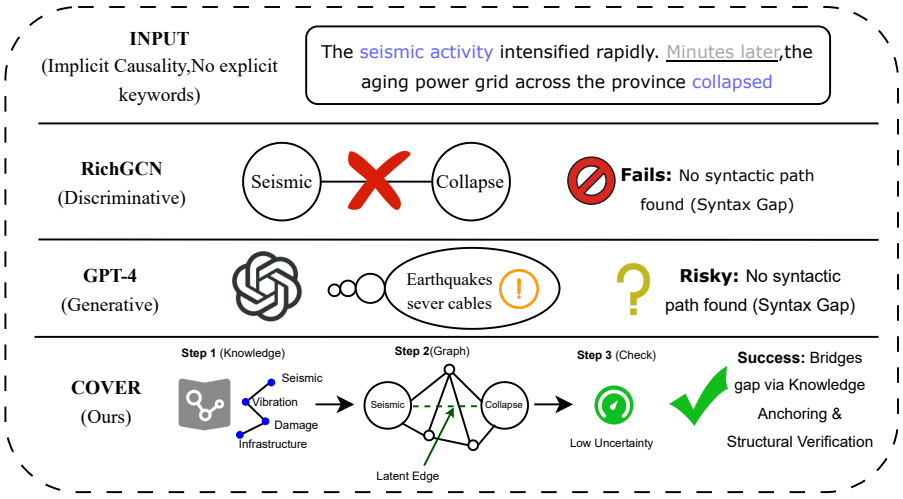

Figure 5: Case Study on Implicit Causality. Unlike RichGCN (limited by syntax) and GPT-4 (prone to hallucination), COVER bridges the semantic gap via knowledge anchoring and verifies the link through structural reasoning.

Model	Type	Explicit	Implicit
RichGCN	Disc.	66.7	33.9
GPT-4 (Zero-shot)	Gen.	62.7	50.0
COVER (Ours)	**Neuro-Sym.**	**69.8**	**54.7**

Table 5: Performance comparison (F1-score, %) on Explicit vs. Implicit causal subsets.

3.7 Long-Distance Dependencies.

To assess the model's capability in handling document-level complexity, we categorize event pairs by sentence distance (Figure 4).

While all models suffer performance degradation as distance increases, COVER exhibits superior robustness in the Long-range category (Dist ≥ 3), surpassing the state-of-the-art ERGO by **+13.8%**. This result validates that the *Latent Temporal Edges* in our neuro-symbolic graph effectively function as semantic "wormholes," shortening the reasoning path and mitigating the information loss inherent in long-distance dependency parsing.

4 Case Study

To demonstrate COVER's interpretability, we examine a challenging implicit causality case shown in Figure 5: *"The **seismic activity** intensified... Minutes later, the **power grid** collapsed."*

Existing paradigms struggle with such disjointed contexts. The discriminative baseline (RichGCN) fails to identify the link (Predicts: NONE) due to the **syntactic gap**—the absence of explicit connectives prevents dependency parsers from bridging the sentence boundary. Conversely, while GPT-4 correctly predicts CAUSE, it suffers from ungrounded inference, justifying the decision with hallucinated details (e.g., *"earthquakes sever cables"*) not present in the source text.

In contrast, COVER successfully identifies the relation by harmonizing intuition and logic. System 1 bridges the semantic gap by retrieving the ConceptNet path (*Seismic* $\rightarrow$ *Vibration* $\rightarrow$ *Infrastructure*), while System 2 anchors this knowledge via the latent temporal edge implied by *"Minutes later."* The resulting low epistemic uncertainty ($u < \tau$) validates the prediction, proving that our neuro-symbolic approach ensures robustness without sacrificing factual faithfulness.

5 Conclusion

In this work, we propose COVER, a variational neuro-symbolic framework that resolves the dichotomy between generative semantic reasoning and discriminative structure learning in Document-level Event Causality Identification. By modeling the LLM as a Variational Prior and grounding it via an Evidential Cognitive Cycle , our approach dynamically quantifies epistemic uncertainty to suppress hallucinations through closed-loop feedback. Empirical results on CEC 2.0 and MAVEN-ERE demonstrate that COVER not only establishes new state-of-the-art benchmarks but also significantly outperforms GPT-4 in implicit causality identification. This confirms that harmonizing generative intuition with structural verification offers a robust pathway for complex reasoning tasks, effectively mitigating the reliability issues of pure large language models.

Limitations

Despite establishing new benchmarks in DECI, our framework presents certain limitations. First, the iterative nature of the Evidential Cognitive Cycle introduces computational overhead compared to single-pass models; although the average iteration depth remains efficient (≈ 1.65), the inference latency for complex samples may challenge strictly real-time scenarios. Second, the robustness of structural verification is partially bounded by the quality of upstream resources, where errors from dependency parsers or coverage gaps in external knowledge bases (e.g., ConceptNet) can propagate noise into the neuro-symbolic graph. Finally, COVER currently operates on pre-identified event mentions; this pipeline setting leaves the system vulnerable to cascading errors from the event extraction phase, highlighting a necessity for future research into end-to-end joint modeling.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Hefei Chen, Yuanyuan Cai, Zexi Song, Yiyao Zhang, and Hongbo Zhang. 2025a. Ieci: A pipeline framework for iterative event causal identification with dynamic inference chains. *Applied Sciences*, 15(13):7348.

Jianhui Chen, Zhiyi Tang, Lianfang Ma, Zitong Zhang, and Haonan Yang. 2025b. A joint extraction model of multiple chinese emergency event–event relations based on weighted double consistency constraint learning. *Symmetry*, 17(11).

Meiqi Chen, Yixin Cao, Kunquan Deng, Mukai Li, Kun Wang, Jing Shao, and Yan Zhang. 2022. Ergo: Event relational graph transformer for document-level event causality identification. *arXiv preprint arXiv:2204.07434*.

Qing Cheng, Zefan Zeng, Xingchen Hu, Yuehang Si, and Zhong Liu. 2024. A survey of event causality identification: Taxonomy, challenges, assessment, and prospects. *ACM Computing Surveys*, 58:1 – 37.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Jun-Mei Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiaoling Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 179 others. 2025a. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025b. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025c. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Jonathan St BT Evans and Keith E Stanovich. 2013. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on psychological science*, 8(3):223–241.

Jinglong Gao, Xiao Ding, Bing Qin, and Ting Liu. 2023. Is ChatGPT a good causal reasoner? a comprehensive evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11111–11126, Singapore. Association for Computational Linguistics.

Seethalakshmi Gopalakrishnan, Luciana Garbayo, and Wlodek Zadrozny. 2024. Causality extraction from medical text using large language models (llms). *Information*, 16(1):13.

Richa Gupta, Indu Kashyap, and Vinita Jindal. 2024. Sbilm: Siamese bi-lstm model for handling imbalance in fake review detection. *Procedia Computer Science*, 235:1157–1166. International Conference on Machine Learning and Data Engineering (ICMLDE 2023).

Chikara Hashimoto. 2019. Weakly supervised multilingual causality extraction from wikipedia. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 2988–2999.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu

Chen. 2021. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685.

Zhilei Hu, Zixuan Li, Xiaolong Jin, Long Bai, Saiping Guan, Jiafeng Guo, and Xueqi Cheng. 2023. Semantic structure enhanced event causality identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10901–10913, Toronto, Canada. Association for Computational Linguistics.

Haoran Li, Qiang Gao, Hongmei Wu, and Li Huang. 2024. Advancing event causality identification via heuristic semantic dependency inquiry network. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1467–1478, Miami, Florida, USA. Association for Computational Linguistics.

Cheng Liu, Wei Xiang, and Bang Wang. 2024. Identifying while learning for document event causality identification. *arXiv preprint arXiv:2405.20608*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Zhenyu Liu, Baotian Hu, Zhenran Xu, and Min Zhang. 2023. Ppat: Progressive graph pairwise attention network for event causality identification. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 5150–5158. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Hieu Man, Franck Dernoncourt, and Thien Huu Nguyen. 2024. Mastering context-to-label representation transformation for event causality identification with diffusion models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18760–18768.

Minh Tran Phu and Thien Huu Nguyen. 2021. Graph convolutional networks for event causality identification with rich document-level structures. In *Proceedings of the 2021 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 3480–3490.

Ruili Pu, Yang Li, Suge Wang, Deyu Li, Jianxing Zheng, and Jian Liao. 2023. Enhancing event causality identification with event causal label and event pair interaction graph. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10314–10322, Toronto, Canada. Association for Computational Linguistics.

Valerie F Reyna and Charles J Brainerd. 2011. Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental review*, 31(2-3):180–206.

Hiroki Sakaji and Kiyoshi Izumi. 2023. Financial causality extraction based on universal dependencies and clue expressions. *New Gen. Comput.*, 41(4):839–857.

M. Sensoy, Melih Kandemir, and Lance M. Kaplan. 2018. Evidential deep learning to quantify classification uncertainty. *ArXiv*, abs/1806.01768.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1).

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, and 1 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Xianchuan Wang, Zongtian Liu, Tao Liao, and Qiang Li. 2015. Cec2.0-based detection event relation for chinese text.

Xiaozhi Wang, Yulin Chen, Ning Ding, Hao Peng, Zimu Wang, Yankai Lin, Xu Han, Lei Hou, Juanzi Li, Zhiyuan Liu, Peng Li, and Jie Zhou. 2022. MAVEN-ERE: A unified large-scale dataset for event coreference, temporal, causal, and subevent relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 926–941, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zimu Wang, Lei Xia, Wei Wang, and Xinya Du. 2024. Document-level causal relation extraction with knowledge-guided binary question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16944–16955, Miami, Florida, USA. Association for Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Shuaicheng Zhang, Qiang Ning, and Lifu Huang. 2022. Extracting temporal event relation with syntax-guided graph transformer. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 379–390, Seattle, United States. Association for Computational Linguistics.

Kun Zhao, Donghong Ji, Fazhi He, Yijiang Liu, and Yafeng Ren. 2021. Document-level event causality identification via graph inference mechanism. *Information Sciences*, 561:115–129.

A Data Efficiency

We simulate low-resource regimes by training models on subsets $\{10\%, 20\%, 50\%\}$ of the training data. As shown in Figure 8, COVER demonstrates remarkable resilience, outperforming the strongest baseline (LearnDA) by over **10%** in F1 score when only 10% of data is available.

This efficiency stems from the synergy between the *Variational Prior* (LLM) and external knowledge anchoring. Unlike discriminative baselines that rely solely on supervised signals, our framework leverages a semantic "warm start," enabling robust reasoning even with limited annotated examples.

B Causal Marker Taxonomy and Dataset Partition

To systematically evaluate the model's capability in reasoning over implicit dependencies versus explicit syntactic patterns, we partitioned the test set based on the presence of linguistic cues. We established a standardized **Explicit Causal Markers List** (Table 6) encompassing four grammatical categories: conjunctions, prepositions, causal verbs, and resultative modifiers. An event pair is categorized as ***Explicit Causality*** if its context window contains at least one marker from this list, indicating a surface-level linguistic realization. Conversely, a pair is classified as ***Implicit Causality*** if it satisfies the *Zero-Lexical Overlap* criterion—meaning no predefined markers appear in the context—thereby requiring the relation to be inferred solely through semantic reasoning or background knowledge.

- **Explicit Causality (64.8%):** Event pairs categorized as explicit if the context window containing the pair includes at least one marker from the standardized list, indicating a surface-level linguistic realization of the causal link.

- **Implicit Causality (35.2%):** Event pairs satisfying the *Zero-Lexical Overlap* criterion—i.e., none of the predefined markers appear in the sentences. These cases often involve a Cross-Sentence Gap where the causal relation must be inferred solely through semantic reasoning or background knowledge without the aid of transitional connectives.

Category	Keywords / Phrases
Conjunctions	because, since, as, for, so, therefore, thus, hence, consequently, accordingly, so that, in order to, etc.
Prepositions	because of, due to, owing to, thanks to, as a result of, as a consequence of, on account of, in view of, by dint of, by virtue of, etc.
Causal Verbs	cause, lead to, result in/from, trigger, induce, produce, generate, spark, provoke, instigate, effectuate, bring about, give rise to, stem from, arise from, contribute to, be responsible for, etc.
Adverbs & Others	consequently, resultantly, inevitable, causal, etc.

Table 6: The standardized list of Explicit Causal Markers used for dataset partitioning.

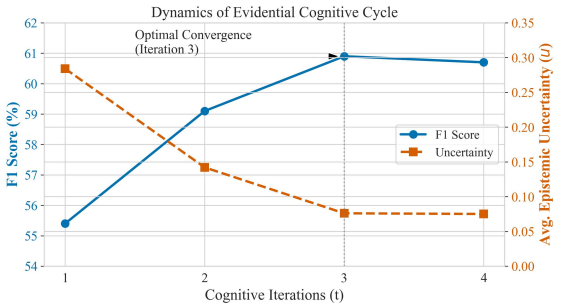

Figure 6: Dynamics of the Evidential Cognitive Cycle: Performance (F1) vs. Epistemic Uncertainty (u).

C Dynamics of the Cognitive Cycle (τ & I_{max}).

The interaction between the energy threshold τ and iteration depth I_{max} is critical for the "System 2" verification mechanism.

- **Energy Threshold (τ):** As visualized in Figure 7, τ acts as a gatekeeper balancing reasoning depth and efficiency. A moderate $\tau \approx 0.5$ yields the highest F1 score while maintaining a low computational cost (Avg. Iterations ≈ 1.65). Setting τ too low triggers unnecessary corrections, while setting it too high degrades the verifier into a passive filter.

- **Iteration Depth (I_{max}):** Figure 6 traces the temporal evolution of the reasoning process. We observe a sharp performance gain from $t = 1$ to $t = 3$ as the epistemic uncertainty decreases. This convergence validates that multi-hop reasoning effectively refines the latent hypothesis. However, performance saturates at $I_{max} = 5$, suggesting a "diminishing returns" effect. Consequently, we adopt $I_{max} = 3$ as the optimal configuration.

D Conflict Packet Generation: The Neuro-Symbolic Interpreter

To bridge the modality gap between the numerical signals of the neuro-symbolic reasoner (System

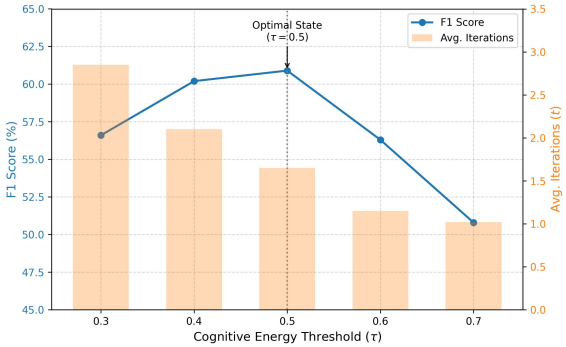

Figure 7: Dual-axis analysis of the Energy Threshold τ: Impact on Model Performance (Left) and Computational Cost (Right).

Figure 8: Performance comparison in Low-Resource Scenarios.

2) and the semantic space of the Large Language Model (System 1), we implement a deterministic **Rule-based Interpreter**. Unlike utilizing an additional LLM for feedback generation, which may introduce secondary hallucinations and latency, our interpreter employs a predefined template system to translate vector-based anomalies into grounded natural language constraints.

D.1 Formal Definition

We define the feedback generation as a mapping function $\Psi : \mathbb{R}^d \to \mathcal{L}$, where $\mathbb{R}^d$ represents the metric space of System 2 and $\mathcal{L}$ denotes the natural language space. At iteration t, the Conflict Packet $\mathcal{C}_{pack}^{(t)}$ is generated based on the metric set $\mathcal{M}^{(t)} = \{\alpha_{ij}, u, GCS\}$:

$$\mathcal{C}_{pack}^{(t)} = \bigcup_{k \in \mathcal{K}} \Psi(\mathcal{M}_k^{(t)}, \tau_k) \tag{8}$$

where $\mathcal{K}$ denotes the set of error types (Structural, Epistemic, Factual) and τ represents the corresponding activation thresholds.

D.2 Template Taxonomy

We categorized the feedback templates into three types, targeting distinct reasoning failures.

Type I: Structural Blocking (Rejection via Graph Attention). Triggered when the Relational Graph Attention Network (RGAT) assigns a negligible attention weight to the hypothesized link, indicating a lack of structural path or logical flow in the document graph.

- **Condition:** $\alpha_{ij} < \tau_{attn}$ (e.g., $\tau_{attn} = 0.1$)
- **Template:** `[REJECT_LINK]: The hypothesized causal path from '{Head}' to '{Tail}' is structurally invalid. The graph attention weight is negligible ({Value}), indicating no logical support in the document structure.`

Type II: Epistemic Confusion (Uncertainty Warning). Triggered when the Evidential Deep Learning (EDL) module outputs high epistemic uncertainty, signaling that the current evidence is insufficient to support a definitive classification.

- **Condition:** $u > \tau_{unc}$ (e.g., $\tau_{unc} = 0.6$)
- **Template:** `[HIGH_UNCERTAINTY]: The reasoning chain for pair ('{Head}', '{Tail}') lacks sufficient evidence. The epistemic uncertainty score is critically high ({Value}). You may be hallucinating a connection where none exists.`

Type III: Factual Contradiction (Knowledge Grounding). Triggered when the Grounded Consistency Score (GCS) is negative, indicating that the retrieved commonsense knowledge conflicts with the specific document context.

- **Condition:** $GCS < \tau_{gcs}$ (e.g., $\tau_{gcs} = 0.0$)
- **Template:** `[KNOWLEDGE_CONFLICT]: The retrieved knowledge '{Triple}' contradicts the specific context of this document (Consistency Score: {Value}). Discard this external knowledge immediately.`

D.3 Prompt Integration

The generated conflict messages are concatenated and injected into the system prompt for the subsequent iteration $(t + 1)$. An example of the constructed prompt is shown in Table 7.

[SYSTEM INSTRUCTION]
You are a Causal Reasoning Refiner. Your previous hypothesis has been evaluated by the Neuro-Symbolic Verifier. The verification failed. Below is the "Conflict Packet" containing specific error reports.

[CONFLICT PACKET]
1. `[REJECT_LINK]`: The link between "Strike" and "Explosion" is rejected due to low attention weight (0.05).
2. `[HIGH_UNCERTAINTY]`: The overall confidence is low ($u = 0.85$). Lack of supporting evidence.

[TASK]
Based on the feedback above, please re-generate the temporal hypothesis. Do NOT assume "Strike" directly causes "Explosion" unless you find a new intermediate event. If no evidence exists, output "Relation: None".

Table 7: Example of an Integrated Feedback Prompt for System 1 Refinement.

E Analysis of Causal Directionality

To validate whether the Evidential Cognitive Cycle effectively rectifies logical fallacies, we conducted a fine-grained analysis on the *Directional Error Rate* (DER). We define DER as the percentage of correctly detected causal links that were assigned the reversed direction (e.g., predicting CAUSED_BY for a ground-truth CAUSE relation).

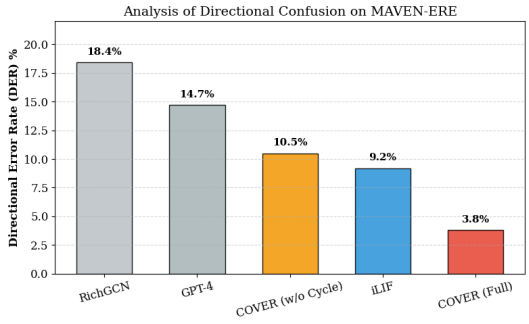

Figure 9: **Directional Error Rate (DER) comparison.** COVER achieves the lowest error rate, demonstrating the effectiveness of the System 2 verifier in correcting logical inversions.

The simulation results on the MAVEN-ERE dataset are presented in Figure 9. We observe three key phenomena:

- **High Confusion in Baselines:** Discriminative models like RichGCN exhibit a high DER of 18.4%, indicating that while they capture event associations effectively, they struggle to distinguish the logical priority between causes and effects, often relying on symmetric co-occurrence patterns.
- **Generative Hallucination:** The zero-shot GPT-4, despite its strong semantic priors, suffers from a 14.7% DER. Qualitative analysis reveals that LLMs often hallucinate directional logic based on surface-level word order rather than deep semantic structures.
- **System 2 Correction:** Crucially, the full COVER framework reduces the DER to an impressive **3.8%**. The ablation study (COVER w/o Cycle) shows a degradation to 10.5%, which is comparable to the strong baseline iLIF (9.2%). This gap explicitly quantifies the contribution of our Evidential Cognitive Cycle: it acts as a logical gatekeeper, detecting the high epistemic uncertainty (u) inherent in reversed causality and triggering the feedback loop to correct the latent hypothesis.

F Impact of Variational Prior Generator Selection

F.1 Motivation and Setup

The **Phase I: Variational Prior Generation** module serves as the "System 1" intuition engine in our COVER framework. To evaluate the framework's dependency on the specific capability of the underlying Large Language Model (LLM), we conducted a substitution study. We replaced the default backbone (*DeepSeek-R1-14B tuned via LoRA*) with the representative LLMs used in our main baselines, including both open-source and closed-source models across different capability tiers.

All variants utilized the same Chain-of-Thought (CoT) prompting strategy to generate the initial temporal hypothesis (H_{temp}) and retrieval queries (K_{query}).

F.2 Quantitative Results

Table 8 presents the performance comparison on the MAVEN-ERE dataset.

LLM Backbone (Phase I)	Type	P (%)	R (%)	F1 (%)
Standard Configuration				
DeepSeek-R1-14B (LoRA)	Open	54.2	69.5	60.9
Simulated Variants				
DeepSeek-R1 (Full)	Open	56.5	**72.1**	63.3
GPT-4	Closed	**56.9**	71.5	**63.4**
GPT-4o-mini	Closed	53.4	68.1	59.8
GPT-3.5-turbo	Closed	51.8	65.4	57.8
LLaMA2-7B	Open	49.5	62.0	55.0

Table 8: Simulated impact of replacing the Phase I generator with different LLMs on MAVEN-ERE.

F.3 Analysis and Discussion

The results highlight the interplay between the generator's reasoning capability and the verifier's corrective power:

1. Reasoning Models Raise the Ceiling. Utilizing reasoning-intensive models like **DeepSeek-R1 (Full)** or **GPT-4** as the variational prior significantly boosts performance (F1 > 63%). Qualitative analysis indicates that these models are superior at uncovering *implicit causal chains* in long documents, providing a high-recall candidate set for the System 2 verifier. This suggests that the COVER framework can scale up with stronger foundation models.

2. Efficiency of Domain Tuning. Our standard configuration using **DeepSeek-R1-14B (LoRA)** achieves an F1 score (60.9%) comparable to the much larger **DeepSeek-Chat (V3)** (61.7%) and outperforms **GPT-4o-mini** (59.8%). This confirms that parameter-efficient fine-tuning (LoRA) on domain-specific data is a cost-effective alternative to using massive general-purpose models.

3. Robustness with Weaker Priors. Even with older or smaller models like **GPT-3.5-turbo** or **LLaMA2-7B**, the framework maintains respectable performance (F1 55.0%–57.8%), drastically outperforming their zero-shot baselines (typically < 30% F1). This proves that the *Evidential Cognitive Cycle* in Phase III effectively acts as a safety net, filtering out the noise and hallucinations produced by weaker System 1 generators.