BEYONDATOMS: Benchmarking and Verifying Relational Factuality Beyond Atomic Facts

Zeyang Liu, Jingfeng Xue, Jiarun Fu, Wenbiao Du,
Xiuqi Yang, Yifeng Fu, Junbao Chen, Yong Wang
Beijing Institute of Technology
liuzeyang@bit.edu.cn

Abstract

As Large Language Models (LLMs) increasingly generate reasoning-driven outputs that weave isolated events into intricate logical topologies, factuality depends not only on isolated atomic facts but critically on the faithful relations connecting them. However, current verification paradigms prioritize individual facts and remain fundamentally blind to relational distortions. This oversight creates a dangerous false sense of security, where verifiers inadvertently endorse relationally corrupted claims simply because their constituent events are factual. Furthermore, the community lacks a systematic benchmark to evaluate this crucial relational layer. To address these critical vulnerabilities, we introduce BEYONDATOMS, a novel benchmark for relational fact verification. Our evaluations reveal that generic verification pipelines struggle significantly by misclassifying relational manipulations as supported. While explicitly guided reasoning models can resolve these structural errors, they incur prohibitive computational costs. To counter this pervasive relation blindness, we propose CARVE, a completely training-free verification framework. By decoupling structural dependencies through deterministic extraction and bi-directional relational alignment, CARVE achieves a favorable accuracy–efficiency trade-off: it remains highly competitive on relational and general fact verification while avoiding the heavy inference cost of reasoning-based LLMs. All datasets and code are publicly available.[1]

1 Introduction

While Large Language Models (LLMs) excel at generation, their susceptibility to hallucinations makes evidence-grounded fact verification an imperative line of research (Ji et al., 2023; Li et al., 2023; Servedio et al., 2025). As LLMs increasingly generate reasoning-driven outputs like Chain-of-Thought explanations, they naturally weave isolated events into intricate logical topologies (Wei et al., 2023; Chen et al., 2026). Consequently, factuality is no longer determined solely by the correctness of isolated facts; it critically depends on the faithful logical relations connecting them.

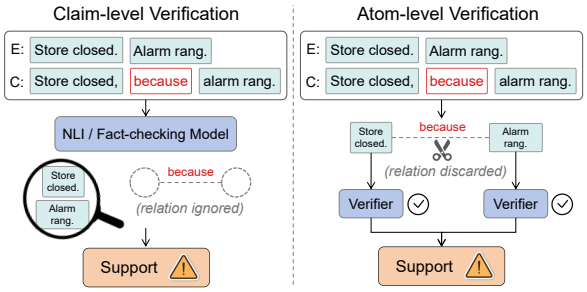


Figure 1: Limitation of current paradigms on relationally corrupted claims. Claim-level verifiers misclassify due to entity-overlap bias. Atom-level decomposition systematically discards logical connections.

Current verification pipelines operate at two primary granularities: holistic sentence-level entailment and sub-sentence atomic verification (Min et al., 2023; Wei et al., 2024; Tang et al., 2024). However, both paradigms suffer from structural blind spots that are visually illustrated in Figure 1. Claim-level verifiers suffer from severe entity-overlap bias (McCoy et al., 2019), while atomic methods systematically sever relational edges during proposition decomposition. This oversight creates a dangerous false sense of security, inadvertently endorsing relationally corrupted claims merely because their constituent events are factual. Furthermore, mitigating this vulnerability is bottlenecked by the lack of a comprehensive benchmark. Existing relation-aware datasets, such as DiscInfer (Zhang et al., 2025) and LoReFact (Xie et al., 2026), primarily adopt a semantic-centric approach. They categorize errors by linguistic relation types and predominantly focus on relation *substitution*, capturing only a fraction of real-world distortions. Therefore, an evaluation framework covering the

[1] https://github.com/lzy248/beyondatoms

full topological spectrum of structural manipulations is urgently required.

To bridge this gap, we introduce **BEYONDATOMS**, a novel benchmark that isolates relational fact verification by manipulating only relational edges while holding constituent events constant. It adopts an operation-centric taxonomy comprising four structural perturbations: *Injection*, *Omission*, *Re-anchoring*, and *Substitution*. Evaluations reveal a pervasive vulnerability across widely adopted pipelines. Traditional NLI and atomic methods (Min et al., 2023; Tang et al., 2024; Wei et al., 2024) exhibit exceptionally low negative recall due to event-overlap bias and structural fragmentation. Similarly, when evaluated with standard verification prompts, reasoning-enhanced LLMs remain structurally blind. Although these models can detect such structural distortions if the prompt explicitly reveals the underlying error taxonomy, this heavy reasoning process incurs prohibitive computational costs.

To address these challenges, we propose **CARVE** (**C**laim-**A**nchored **R**elational **V**erification over **E**vidence). To overcome the pervasive structural blindness of existing paradigms, CARVE explicitly decouples structural logic from constituent events and offloads the verification of brittle logical links to a deterministic, bi-directional rule alignment pipeline. This design creates a completely training-free framework requiring no costly parameter tuning. Even powered by a lightweight open-source backbone, CARVE remains competitive on relational and general fact verification while avoiding the heavy inference cost of reasoning-based LLMs.

Our contributions are as follows:

- We identify a critical structural blind spot in current fact verification paradigms and introduce BEYONDATOMS, a meticulously constructed, operation-centric benchmark dedicated to isolating and diagnosing relation-level factual errors.
- We demonstrate that generic verification methods struggle at this task. Our evaluations show that accurately diagnosing these errors benefits substantially from explicit relational modeling. Although LLMs can perform this modeling via taxonomy-revealed reasoning prompts, this process incurs prohibitive computational costs.
- We propose CARVE, a training-free verification framework that achieves a strong accuracy–efficiency trade-off. CARVE outperforms standard verification baselines and generic-prompt LLMs on relational fact verification, while avoiding the heavy inference-time reasoning required by prompt-based LLM verifiers.

2 Related Work

2.1 Fact Verification and Atomic Factuality Evaluation

Fact verification and natural language inference (NLI) traditionally assess claim-evidence entailment via sentence-pair classification (Bowman et al., 2015; Williams et al., 2018; Thorne et al., 2018). Subsequent works extend this paradigm to document-grounded consistency evaluation for generated text, computing entailment scores to detect unsupported hallucinations (Laban et al., 2022; Zha et al., 2023; Tang et al., 2024). To provide precise error localization, recent frameworks have transitioned toward fine-grained atomic factuality. Methods like FActScore (Min et al., 2023) and WiCE (Kamoi et al., 2023) decompose long-form generations into isolated subclaims and measure the proportion supported by external evidence. While effective at grounding individual entities, this decomposition paradigm systematically discards the structural dependencies among units. Consequently, existing pipelines remain structurally blind to relation-level errors, where individual events are factual but their causal, temporal, or conditional connections are distorted.

2.2 Discourse-aware and Relation-level Factuality

Linguistic resources for discourse parsing and event relation extraction highlight the distinct challenge of relational understanding (Prasad et al., 2017; Caselli and Vossen, 2017; Ning et al., 2018). Recent temporal relation extraction models further show that relation direction and event ordering require dedicated structural modeling (Zhang et al., 2022). However, their primary objective is relation identification within source texts rather than evidence-grounded claim verification. The most relevant prior work introduces structured discourse representations to evaluate factual consistency, demonstrating that adversarial discourse relations severely challenge existing models (Zhang et al., 2025; Zhong and Litman, 2025). Existing

benchmarks such as DiscInfer and LoReFact primarily adopt a semantic-centric approach. They categorize instances based on linguistic discourse relations (such as causal, conditional, temporal, and concessive relations). In contrast, BEYONDATOMS introduces an operation-centric taxonomy based on structural perturbations. It systematically covers fundamental topological operations on a relational edge between preserved constituent events. This structural perspective allows for a more exhaustive diagnostic evaluation of how models fail on distinct relational manipulations.

3 BEYONDATOMS Benchmark

While existing benchmarks evaluate whether factual units are supported, they leave a critical blind spot: whether the relations connecting those units are faithfully preserved. We introduce BEYONDATOMS to address this gap.

3.1 Task Definition

We formulate relational fact verification as determining whether a claim c is supported by evidence passages E. Formally, we represent the structural logic of a claim as a set of relational triples: $T_c = \{(e_i, r, e_j)\}$ where e_i and e_j are constituent events, and r denotes the explicit logical relation connecting them (e.g., causal, temporal, or conditional). Crucially, r is inherently directional: for asymmetric relations such as causality, (e_i, r, e_j) is strictly distinct from (e_j, r, e_i).

Unlike conventional atomic verification, which merely checks the veracity of isolated factual units, relational verification strictly requires that both the constituent events *and* the semantic edge r are faithfully grounded in E. Consequently, even if all base events in c are fully entailed by E, any distorted or unsupported relation r renders the claim as *not supported*. This formulation explicitly targets structural errors that routinely bypass methods relying on entity overlap or atomic decomposition.

3.2 Relational Error Taxonomy

Unlike existing datasets that classify instances by semantic relation types (e.g., causal vs. temporal), BEYONDATOMS adopts a novel **operation-centric taxonomy**. From an anchor-centric graph perspective, we categorize errors based strictly on how the topological edges between events are manipulated. As shown in Table 1, all four operations preserve the base constituent events and perturb only their logical connections. Specifically, a relational edge can only be **artificially introduced** (Injection), **unconditionally dropped** (Omission), **structurally shifted in its attachment** (Re-anchoring), or **semantically replaced** (Substitution). This ensures that model failures reflect relation-level blindness rather than entity-level hallucination.

3.3 Comparison with Existing Benchmarks

Table 2 compares BEYONDATOMS with representative fact-checking, factuality evaluation, and relation-aware benchmarks. Traditional benchmarks (e.g., FEVER, FActScore) evaluate factuality either at the holistic sentence level or the atomic-fact level, leaving structural relations largely unexamined. Recent relation-aware benchmarks, specifically DiscInfer and LoReFact, recognize this gap but predominantly adopt a *semantic-centric* approach. They categorize instances based on linguistic logic types (such as causal, temporal, or conditional relations) and their negative instances primarily map to relation *substitution*. In contrast, BEYONDATOMS utilizes an *operation-centric* taxonomy. By systematically applying four fundamental topological perturbations (Injection, Omission, Reanchoring, Substitution) to the relational edges, our benchmark explicitly isolates relation-level factuality and guarantees comprehensive coverage across the full spectrum of structural manipulations.

3.4 Dataset Construction

Each instance in BEYONDATOMS is formulated as a tuple: $\{\texttt{id}, \texttt{claim}, \texttt{label}, \texttt{evidences}\}$, where the label is either `support` or `not_support`. To ensure a diverse and linguistically rich set of logical relations, we source our foundational texts primarily from Wikipedia and popular science articles. Our construction pipeline is fundamentally driven by our operation-centric taxonomy, and strictly adheres to two core principles to isolate structural errors:

- **Event Preservation:** The constituent events in all negative claims must be fully grounded in the evidence, guaranteeing `not_support` labels stem exclusively from relational distortions rather than event-level mismatches.
- **Single-Operation Perturbation:** Each negative instance is manipulated using exactly one of the four defined operations, preventing confounding factors and enabling fine-grained diagnosis across specific error types.

Operation	Description	Evidence	Claim
Injection	Introduces a relation between events that are individually supported but remain disconnected in the evidence.	A happened. B happened.	B happened [so/if/...] A happened.
Omission	Asserts an event unconditionally by dropping a required condition or premise present in the evidence.	B happens [if/when/...] A holds.	B happens.
Re-anchoring	Preserves the connective but shifts its attachment, altering the relational scope between events.	A happened [so/if/...] B happened.	[so/if/...] A happened, B happened.
Substitution	Replaces the existing relation in the evidence with a semantically different connective.	A happened [if/so/...] B happened.	A happened [but/when/...] B happened.

Table 1: Taxonomy of structural perturbations in BEYONDATOMS. Each operation explicitly isolates relation-level errors by manipulating only the structural edges while holding base constituent events (A and B) constant.

Benchmark	Granularity	Scope	Injection	Omission	Re-anchoring	Substitution
FEVER (Thorne et al., 2018)	Sentence	Fact	–	–	–	–
WiCE (Kamoi et al., 2023)	Atom	Fact	–	△	–	–
FActScore (Min et al., 2023)	Atom	Fact	–	–	–	–
LLM-AggreFact (Tang et al., 2024)	Sentence	Fact	–	–	–	–
DiscInfer (Zhang et al., 2025)	Sentence	Relation	–	–	–	✓
LoReFact (Xie et al., 2026)	Sentence	Relation	△	–	–	✓
BEYONDATOMS	Sentence	Relation	✓	✓	✓	✓

Table 2: Comparison with representative factuality benchmarks. "Granularity" denotes the verification unit, and "Scope" indicates whether the benchmark targets factual units or relations. △ indicates partial coverage.

3.5 Annotation and Statistics

Positive instances are directly sourced from original texts containing clear relational expressions. Each is then perturbed according to one of the four operation types to produce a corresponding negative counterpart, ensuring that base events are preserved while only the relation is distorted. To guarantee the diagnostic purity of the benchmark, we implemented an exceptionally strict pre-filtering protocol. Out of 4,000 initial candidate text segments sourced from Wikipedia and science articles, only 1,550 unambiguous instances (spanning 647 unique text segments) were retained for final annotation. The annotation was conducted by three graduate-level NLP researchers. Cases involving unclear event references, insufficient evidence, or multiple plausible interpretations were discarded prior to the structural annotation phase. This rigorous upstream selection eliminated ambiguity at the source. Consequently, it resulted in a 100% verification pass rate during the final cross-verification stage with no candidates discarded.

To maintain strict quality control, all selected instances must be unambiguous and straightforwardly verifiable. We rigorously excluded cases involving unclear event references, insufficient evidence, multiple plausible interpretations, or overlapping error types. This conservative filtering ensures that the benchmark remains exclusively diagnostic of relation-level factuality, avoiding confounders from lexical or event-level ambiguity.

Following this rigorous pipeline, the final BEYONDATOMS benchmark comprises 1,550 instances. As summarized in Table 3, the dataset is strictly balanced between positive and negative labels across all four operation subsets. To illustrate the precise nature of these structural manipulations, we provide detailed qualitative examples of each relational error type in Appendix A.

Subset	#Sup.	#Not-sup.	Total
Injection	186	186	372
Omission	187	187	374
Re-anchoring	241	241	482
Substitution	161	161	322
Total	775	775	1,550

Table 3: Statistics of BEYONDATOMS.

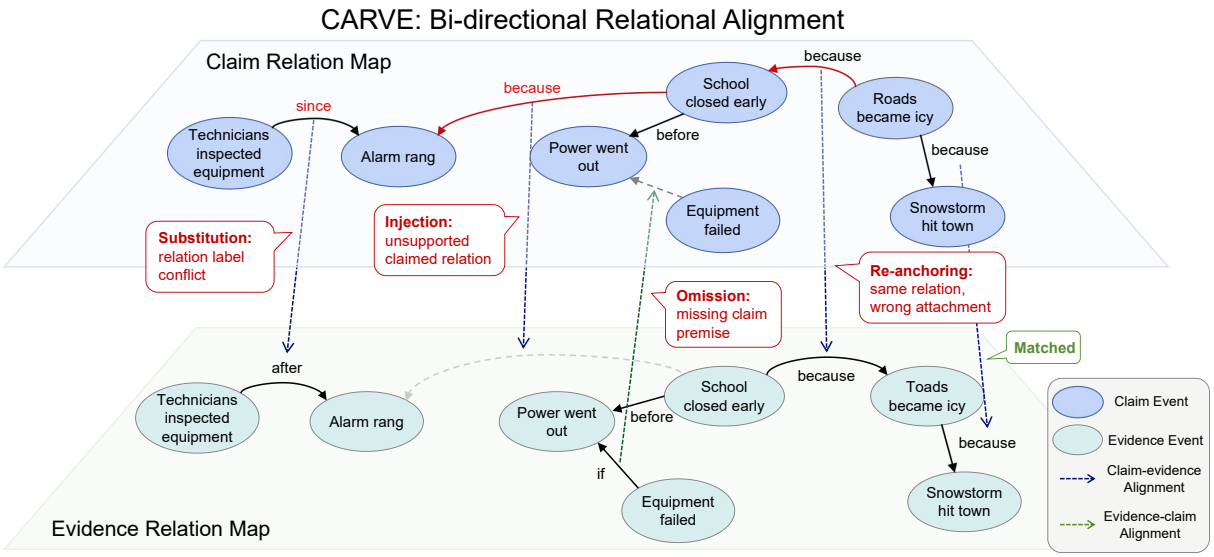


Figure 2: Overview of CARVE's bidirectional relational alignment. CARVE aligns claim relations with a claim-anchored evidence structure: claim-to-evidence alignment detects Injection, Re-anchoring, and Substitution through missing edges, endpoint conflicts, and label conflicts, while evidence-to-claim alignment detects Omission by finding required evidence relations absent from the claim. This explicit alignment improves transparency and reduces reliance on costly end-to-end LLM reasoning.

4 CARVE: Claim-Anchored Relational Fact Verification

BEYONDATOMS exposes relational factual errors that easily bypass current verification systems. Fundamentally, claim-level verifiers overlook relations due to an inherent event-overlap bias, while atomic verifiers completely dismantle relational structures. To effectively address this, relations must be explicitly surfaced and structurally aligned.

We propose CARVE, a training-free **C**laim-**A**nchored **R**elational **V**erification framework over **E**vidence. As illustrated in Figure 2, CARVE represents the claim and the claim-anchored evidence as aligned event–relation structures. Rather than relying on holistic black-box judgments that may blur rigid structural nuances, CARVE decomposes verification into an interpretable pipeline: relation-preserving canonicalization, claim and evidence extraction, and bi-directional relation alignment.

4.1 Relation-Preserving Canonicalization

The same logical relation can be expressed through highly diverse surface structures (e.g., fronted subordinate clauses, cross-sentence discourse markers, or pronominal references). This syntactic variance severely hinders reliable deterministic extraction. To neutralize this noise, CARVE first canonicalizes both the claim and the evidence. This module performs transformations such as repairing segmentation artifacts caused by fronted connectives (e.g., "Because A, B" $\rightarrow$ "B because A") and resolving local pronominal references near connectives. Crucially, this canonicalization is strictly relation-preserving: it standardizes existing structures without introducing any new semantic relations.

4.2 Relation Extraction

Constructing a complete relation graph over the entire evidence is not only computationally prohibitive but also introduces massive irrelevant noise. Verification must therefore be strictly claim-driven. To achieve this, CARVE adopts an asymmetric, two-stage extraction pipeline: it first parses the claim to establish exact structural anchors, and then uses these anchors to strictly constrain the extraction scope on the evidence side.

Claim-Side Extraction. Following canonicalization, CARVE parses the claim into a set of explicit relation triples $T_c = \{(u, r_c, v)\}$, where u and v are the constituent events and r_c is the connective. It additionally extracts isolated events I_c that are not bound by any connective. Together, these elements form the claim anchor set $A(c)$, which explicitly defines the verification targets and dictates the boundaries for subsequent evidence processing.

Evidence-Side Extraction. Instead of processing the evidence blindly, CARVE performs a targeted extraction strictly conditioned on $A(c)$. It scans the canonicalized evidence, retaining only

Evidence	Semantics vs. r_c	Diagnosis
Claim-to-Evidence Alignment (Claim target: (u, r_c, v))		
(u, r_e, v)	Same family	`support`
(v, r_e, u)	Semantic inverse	`support`
(v, r_e, u)	Same family	`Re-anchoring`
(u, r_e, v)	Different family	`Substitution`
u, v found	No connective	`Injection`
Evidence-to-Claim Alignment (Claim target: y)		
(x, r_e, y)*	Premise x missing	`Omission`

*r_e is a conditional connective.

Table 4: Bi-directional alignment logic in CARVE.

sentences that exhibit sufficient content-word overlap with at least one claim anchor. Triples extracted from these filtered sentences form a compact, focused evidence relation pool $P_E(c)$, ensuring that the verification process remains completely immune to out-of-scope relational noise.

4.3 Bi-Directional Alignment and Error Diagnosis

The core of CARVE is aligning claim relations T_c with the evidence pool $P_E(c)$. This alignment process corresponds to the central comparison space in Figure 2: claim-to-evidence alignment verifies whether claimed relations are supported, while evidence-to-claim alignment checks whether required evidence relations have been omitted. The deterministic diagnosis logic is detailed in Table 4.

Claim-to-Evidence Alignment. For each $(u, r_c, v) \in T_c$, CARVE searches $P_E(c)$ for a match in both argument directions. Connective compatibility is evaluated using predefined semantic families and inverse pairs (e.g., *before*/*after*). A consistent semantic match yields `support`. Otherwise, specific distortions are flagged: a direction mismatch within the same connective family triggers Re-anchoring; a semantic family mismatch triggers Substitution; and constituent events lacking a connective edge trigger Injection.

Evidence-to-Claim Alignment. Omission involves a relation entirely absent from the claim. CARVE performs a reverse scan over $P_E(c)$ for unmatched conditional relations (e.g., *if*). If the result-side event of an evidence condition is asserted unconditionally in the claim, an Omission error is diagnosed.

4.4 Auxiliary Verification and LLM Fallback

After relational alignment, CARVE performs event-level verification to ensure that the constituent events are also grounded in the evidence. It checks independent events in I_c and the arguments of verified relations using an NLI model or a constrained LLM prompt; unsupported events trigger a `fact_unsupported` label.

CARVE also uses a constrained LLM fallback when deterministic parsing or connective interpretation fails. The LLM is not used for end-to-end claim verification; instead, it serves as a localized semantic helper for resolving ambiguous structures or unrecognized connectives. This design preserves interpretability while adding limited semantic flexibility.

5 Experiments

5.1 Settings

Datasets. We evaluate our proposed framework and baseline models on three datasets to comprehensively assess both relation-level and general fact verification capabilities. **BEYONDATOMS**: Our proposed benchmark, evaluated across its four operation-centric subsets: *Substitution*, *Re-anchoring*, *Injection*, and *Omission*. **DiscInfer** (Zhang et al., 2025): A related discourse-aware textual entailment benchmark. We evaluate on its *Comparison*, *Contingency*, and *Temporal* subsets. Since DiscInfer predominantly focuses on relation substitution, it serves as a complementary relation-aware testbed. **FEVER** (Thorne et al., 2018): A standard fact verification dataset. We include FEVER to assess whether relation-focused verification methods maintain robust performance on traditional entity- and event-level factuality.

Baselines. We compare CARVE against three distinct categories of verification systems. **LLM-based Verifiers**: We evaluate LLMs under two settings. The *generic* setting prompts for standard binary entailment (`support` or `not_support`). Additionally, we introduce a *taxonomy-revealed* (TR) setting. The TR prompt explicitly reveals the four relational error operations evaluated in BEYONDATOMS. We separate TR prompting from the main comparison because it reveals the target error taxonomy at inference time and serves as a prompt-engineered upper reference rather than a standard verifier. The exact generic and taxonomy-revealed prompts are provided in Ap-

pendix D. Evaluations are conducted primarily on the open-source `Qwen3.5` series (Yang et al., 2025), alongside `gpt-5-mini` (Singh et al., 2025) as a closed-source reference. **NLI and Fact-Checking Models**: We select commonly used verification baselines, including a strong NLI model `nli-deberta-v3-large` (Reimers, 2019), a representative document-grounded fact-checking model `MiniCheck-DeBERTa` (Tang et al., 2024), and `FDSpotter` (Zhang et al., 2025), a recent method specifically fine-tuned for discourse-level factual consistency. **Fine-grained fact verification Pipeline**: We implement the FActScore pipeline (Min et al., 2023) as a representative fine-grained verification approach. This serves to explicitly demonstrate the structural limitations and potential information loss caused by decomposing claims prior to verification. For **CARVE**, we use Qwen3.5-4B as the default lightweight auxiliary backbone for event-level verification and constrained LLM fallback.

Evaluation Metrics. For BEYONDATOMS and FEVER, we report Macro-F1 to account for class balance across support and not-support labels. For DiscInfer, we report Accuracy following the convention of prior work, as the dataset is heavily skewed toward negative instances.

5.2 Results and Analysis

Tables 5 and 6, alongside Figure 3, present the performance and inference efficiency of all evaluated systems. The results substantiate our core hypothesis. Relational factual errors expose a critical structural blindness in current verification paradigms under generic prompts. While taxonomy-revealed reasoning prompts can substantially improve performance, they incur prohibitive computational costs. In contrast, CARVE natively operationalizes this explicit structural modeling to successfully mitigate the vulnerability while maintaining extreme inference efficiency.

BEYONDATOMS Exposes Broad Relational Blindness. The evaluation reveals a pervasive relation-blind status among current verification systems. While a powerful proprietary model like GPT-5-mini saturates prior benchmarks (99.01% on DiscInfer), it degrades significantly on BEYONDATOMS. Crucially, traditional NLI models and small-scale open LLMs show severe degradation on BEYONDATOMS, struggling to differentiate structural manipulations from event overlaps. By exposing these vulnerabilities, BEYONDATOMS establishes a highly diagnostic challenge for structural factuality.

Holistic and Atomic Paradigms Struggle Significantly. Traditional NLI pipelines (e.g., MiniCheck) perform dismally on BEYONDATOMS, yielding a Macro-F1 of 48.61%. Because they rely heavily on event overlap, they uniformly default to `support` when the constituent events are present, misclassifying nearly all negative instances. Similarly, the fine-grained atomic verification method FActScore exhibits severe degradation (46.09% Macro-F1). This confirms that atomic decomposition structurally dismantles the relational edges binding events together, rendering both paradigms fundamentally incapable of diagnosing relation-level manipulations.

Specialized Models Lack Structural Generalization. FDSpotter is the strongest specialized relation-aware baseline on BEYONDATOMS, especially on Substitution, the category most aligned with discourse relation alteration. However, its performance plummets to 55.21% on Omission, demonstrating exceptionally low negative recall. This indicates that while specialized training captures semantic connective replacements within its distribution, it fails to diagnose the covert dropping of logical premises, highlighting the necessity for a unified diagnostic framework.

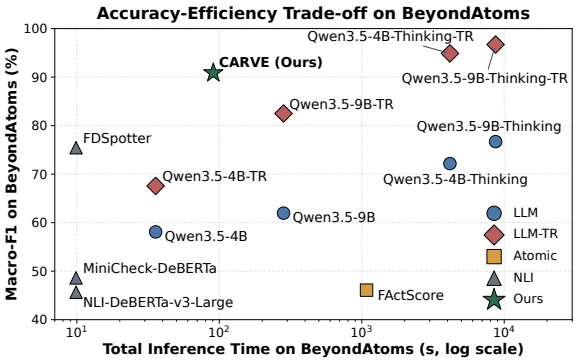


Figure 3: Inference efficiency comparison on BEYONDATOMS. The x-axis uses a log scale due to the large runtime gap across models.

Reasoning Requires Explicit Structural Guidance. Under generic prompts, reasoning models still struggle with existential errors like Omission (56.88% Macro-F1 for Qwen3.5-9B-Thinking). However, Table 6 shows that explicitly exposing the error taxonomy (the TR setting) enables remarkable recovery, with Qwen3.5-9B-Thinking-

Model	BeyondAtoms					DiscInfer*				FEVER
	Sub.	Re-anch.	Inj.	Omi.	Avg.	Comp.	Cont.	Temp.	Avg.	
Generic prompt LLMs										
GPT-5-mini	85.06	83.68	82.25	72.41	80.85	97.92	99.11	100.00	99.01	94.53
Qwen3.5-9B-Thinking	84.75	90.79	74.37	56.88	76.70	66.67	84.82	92.77	81.42	94.12
Qwen3.5-9B	71.63	78.56	55.11	42.55	61.96	47.92	77.23	84.34	69.83	94.58
Qwen3.5-4B-Thinking	81.73	90.37	64.86	51.76	72.18	68.75	77.23	90.36	78.78	94.32
Qwen3.5-4B	66.04	75.24	49.39	41.67	58.09	35.42	62.05	73.49	56.99	94.40
Fine-grained fact verification										
FActScore	55.04	49.44	36.97	42.92	46.09	66.67	80.36	86.75	77.93	91.70
NLI and Fact-Checking Models										
NLI-DeBERTa-v3-Large	49.91	46.12	40.30	46.27	45.65	37.50	70.09	73.49	60.36	91.74
MiniCheck-DeBERTa	57.31	49.35	47.99	39.79	48.61	77.08	91.52	96.39	88.33	89.85
FDSpotter	89.03	75.49	82.01	55.21	75.44	68.75	95.98	89.16	84.63	94.22
Our Method										
CARVE	**91.60**	**82.88**	**96.25**	**92.77**	**90.87**	**93.75**	**98.66**	**98.80**	**97.07**	**92.63**

Table 5: Overall evaluation results. Metrics are reported in percentages (%). We report Macro-F1 for BeyondAtoms and FEVER, and Accuracy for DiscInfer (marked with *). Sub., Re-anch., Inj., Omi., Comp., Cont., and Temp. correspond to Substitution, Re-anchoring, Injection, Omission, Comparison, Contingency, and Temporal.

Model	Sub.	Re-anch.	Inj.	Omi.	Avg.
Taxonomy-revealed prompt LLMs					
Qwen3.5-9B-Thinking-TR	98.76	97.51	98.66	91.96	96.72
Qwen3.5-9B-TR	85.96	87.15	83.96	71.93	82.25
Qwen3.5-4B-Thinking-TR	96.58	97.51	96.77	88.68	94.89
Qwen3.5-4B-TR	75.81	79.83	65.59	49.04	67.57

Table 6: Performance of taxonomy-revealed (TR) prompting on BeyondAtoms. TR prompts explicitly describe the four relational error operations.

TR achieving a 96.72% average. This massive gap confirms that standard prompts leave models structurally blind. Furthermore, removing thinking capabilities causes a severe performance drop. This proves that merely knowing the error types is insufficient. To actually resolve relational distortions, LLMs must actively model the logical connections via a heavy reasoning process.

CARVE Operationalizes Explicit Modeling Efficiently. Taxonomy-revealed thinking achieves the highest accuracy, but it depends on costly model-side reasoning after being explicitly informed of the relational error taxonomy. CARVE implements the same principle of explicit structural awareness through deterministic extraction and bi-directional alignment. With only a lightweight 4B backbone, CARVE reaches 90.87% on BeyondAtoms, remains competitive on FEVER with 92.63%, and runs over 40× faster than reasoning baseline. These results show that explicit structural alignment provides a more practical accuracy-efficiency trade-off than relying on inference-time generative reasoning.

Generalization to Naturally Occurring Errors.

Generalization to Naturally Occurring Errors. To validate whether our findings transfer beyond controlled perturbations, we evaluate on a balanced subset of 80 real-world news summaries from the FRANK dataset (Pagnoni et al., 2021) containing naturally occurring relation errors. As shown in Table 7, CARVE maintains highly competitive performance on this out-of-domain set. It successfully outperforms implicit learning models and generic-prompt LLMs, approaching the accuracy of the explicitly guided, compute-heavy taxonomy-revealed reasoning baselines. Error analysis reveals that 76.47% (13/17) of CARVE's misclassifications stem purely from *implicit* relational distortions (e.g., causality implied by semantic verbs rather than overt connectives). This clearly defines the boundary of explicit syntactic alignment (further discussed in Limitations) while proving CARVE's robustness on unstructured, real-world texts.

5.3 Ablation Study

Table 8 confirms that the gains of CARVE come from complementary components. Removing canonicalization leads to a consistent but moderate drop across all error types, showing that normalizing fronted clauses, discourse markers, and local references stabilizes deterministic relation extraction. Removing bi-directional alignment causes the largest degradation, with Omission falling from

Method	FRANK-Relation (Macro-F1)
FDSpotter	65.33
Qwen3.5-9B	73.33
Qwen3.5-9B-Thinking	79.68
Qwen3.5-9B-TR	82.46
Qwen3.5-9B-Thinking-TR	**84.96**
CARVE (Ours)	78.48

Table 7: Macro-F1 performance on the FRANK-Relation subset containing naturally occurring errors. TR denotes Taxonomy-Revealed prompting.

Variant	Inj.	Omi.	Re-anch.	Sub.	Avg.
Full CARVE	96.23	92.77	82.88	91.60	**90.87**
w/o Norm.	93.23	91.58	80.59	89.23	88.66
w/o Bi-dir.	96.23	35.70	82.34	91.23	76.38
w/o LLM FB	93.61	92.77	80.88	87.26	88.63

Table 8: Ablation study on the core components of CARVE across BEYONDATOMS. Metrics are reported in Macro-F1 (%).

92.77% to 35.70%, because dropped premises are absent from the claim and cannot be recovered by claim-to-evidence matching alone. Removing the LLM fallback also reduces performance, especially on Substitution, indicating that bounded semantic assistance helps when connective meanings fall outside predefined rule families. Overall, CARVE's performance is primarily driven by explicit structural extraction and alignment, with the LLM acting as a localized parser rather than an end-to-end reasoner.

5.4 Error Analysis

We further analyze failures in Appendix F. The key finding is that LLM-thinking and CARVE fail for different reasons. LLM-thinking mainly suffers from *relation under-specification*: it often verifies whether the events are present, but does not consistently verify whether the relation between them is explicitly supported. This leads to three typical errors: weakening conditional statements into unconditional facts, hallucinating relations from event co-occurrence, and importing pragmatic or world-knowledge assumptions.

By contrast, CARVE mainly suffers from *relation representation errors*. Since CARVE explicitly verifies event–relation structures, its mistakes usually arise when the extracted structure is wrong, such as spurious connective detection, incorrect cross-sentence anchoring, over-aggressive normalization, or unresolved discourse references. Thus, LLM-thinking fails because relations are not treated as independent verification targets, whereas CARVE fails when the relation target is explicit but imperfectly represented.

6 Conclusion

We introduced BEYONDATOMS, an operation-centric benchmark for relational fact verification, showing that generic fact verification methods often fail to detect distorted relations when constituent events are preserved. We further proposed CARVE, a training-free framework that verifies relational fidelity through claim-anchored extraction and bi-directional alignment. Experiments show that CARVE provides a strong accuracy–efficiency trade-off, outperforming standard baselines on relational verification while avoiding costly inference-time reasoning.

Limitations

While CARVE demonstrates strong performance and efficiency, we acknowledge several limitations in the current framework.

First, CARVE mainly targets explicit relational markers. In natural language, logical relations such as causality, temporality, and conditions are often implied across adjacent sentences without overt discourse connectives (e.g., "The road was icy. The car skidded."). Error analysis on our FRANK-Relation evaluation reveals that 13 out of 17 (76.47%) of CARVE's misclassifications stem purely from these implicit relational distortions. Since CARVE relies on explicit lexical or syntactic triggers to surface event relations, it cannot process logical links encoded purely through semantic verbs. Resolving these requires heavy semantic reasoning beyond the scope of a syntactic pipeline, which we position as future work.

Second, regarding the evaluation framework, CARVE encodes a task-specific structural prior that mirrors the benchmark's taxonomy. This might be perceived as hard-coding the answer key. However, any specialized verifier must possess such a structural prior to effectively address relational vulnerabilities. CARVE does not receive the gold category of an instance at inference time. It deterministically verifies edge existence, endpoint alignment, and label compatibility, from which the distinct diagnoses naturally follow. Our goal is to demonstrate that computationally expensive relation-checking processes can be successfully compiled into highly efficient and reusable structural operators, rather

than relying exclusively on heavy generative reasoning.

Finally, CARVE inherits the limitations of rule-based NLP extraction. Its deterministic extraction is efficient but may become less robust when facing highly diverse sentence structures, noisy generations, or long-range dependencies. Errors from sentence segmentation, dependency parsing, or event matching can propagate to later alignment decisions. While the constrained LLM fallback mitigates some of these failures, an increase in parsing ambiguities triggers more fallback calls, which inherently weakens the extreme efficiency advantage observed in our current experiments.

Acknowledgments

This work was supported by the National Science and Technology Major Project under Grant 2022ZD0120204.

References

Samuel Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 632–642.

Tommaso Caselli and Piek Vossen. 2017. The event storyline corpus: A new benchmark for causal and temporal relation extraction. In *Proceedings of the Events and Stories in the News Workshop*, pages 77–86.

Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2026. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *Science China Information Sciences*, 69(6):161101.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. Wice: Real-world entailment for claims in wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583.

Philippe Laban, Tobias Schnabel, Paul N Bennett, and Marti A Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Junyi Li, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3428–3448.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv:2305.14251*.

Qiang Ning, Hao Wu, and Dan Roth. 2018. A multi-axis annotation scheme for event temporal relations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1318–1328.

Artidoro Pagnoni, Vidhisha Balachandran, and Yulia Tsvetkov. 2021. Understanding factuality in abstractive summarization with frank: A benchmark for factuality metrics. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4812–4829.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv:1908.10084*.

Giovanni Servedio, Alessandro De Bellis, Dario Di Palma, Vito Walter Anelli, and Tommaso Di Noia. 2025. Are the hidden states hiding something? testing the limits of factuality-encoding capabilities in llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6089–6104.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, and 1 others. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*.

Liyan Tang, Philippe Laban, and Greg Durrett. 2024. Minicheck: Efficient fact-checking of llms on grounding documents. In *EMNLP*.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. Fever: a large-scale dataset for fact extraction and verification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language*

Technologies, Volume 1 (Long Papers), pages 809–819.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Jerry Wei, Chengrun Yang, Xinying Song, Yifeng Lu, Nathan Hu, Jie Huang, Dustin Tran, Daiyi Peng, Ruibo Liu, Da Huang, and 1 others. 2024. Long-form factuality in large language models. *Advances in Neural Information Processing Systems*, 37:80756–80827.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122.

Qiming Xie, Wenjie Zheng, Xiangqing Shen, and Rui Xia. 2026. Lorefact: Bridging the logic gap in fact-checking. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 6974–6996.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. Alignscore: Evaluating factual consistency with a unified alignment function. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348.

Kun Zhang, Oana Balalau, and Ioana Manolescu. 2025. Structured discourse representation for factual consistency verification. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 820–838.

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.

Yang Zhong and Diane Litman. 2025. Discourse-driven evaluation: Unveiling factual inconsistency in long document summarization. *Preprint*, arXiv:2502.06185.

A Qualitative Examples of BEYONDATOMS

In this section, we provide qualitative examples from the BEYONDATOMS benchmark to illustrate the four operation-centric relational distortions. As shown in Table 9, the generation of negative instances strictly adheres to our core principle of event preservation. While the constituent entities and baseline events remain fully grounded in the original evidence, the logical connectives bridging them, such as causal, conditional, or temporal markers, are systematically manipulated. These examples demonstrate how structural errors can easily bypass traditional event-overlap verifiers, highlighting the necessity of bi-directional relational alignment.

B Annotation Interface and Construction Guidelines

Figure 4 shows the annotation interface used for constructing BEYONDATOMS. Rather than automatically perturbing existing sentences, annotators manually constructed claim-evidence instances according to operation-specific guidelines. For each instance, annotators were shown the evidence passage and were asked to write or revise a claim such that the target label and relational operation were satisfied.

For positive instances, the claim was required to faithfully preserve both the constituent events and their relation as expressed in the evidence. For negative instances, annotators constructed claims by preserving the constituent events while introducing exactly one relation-level distortion: Injection, Omission, Re-anchoring, or Substitution. This design ensures that `not_support` labels arise from relational inconsistency rather than ordinary event-level factual errors.

Construction rules. Annotators followed three main rules: (1) the main events in the claim must be present in or inferable from the evidence; (2) the label should depend primarily on the relation among events, not on whether the events themselves are correct; and (3) each negative instance should instantiate exactly one operation type. Cases that could not satisfy these constraints, such as ambiguous event boundaries, unclear references, insufficient evidence, or multiple overlapping relation errors, were revised or excluded during quality control.

C Computational Setup.

Experiments are conducted on a workstation with an AMD Ryzen 9 9950X CPU, 128GB RAM, and NVIDIA RTX 4090 GPUs. Open-source LLM inference uses vLLM 0.19.1, PyTorch 2.11.0, and

Table 9: Representative examples from the BEYONDATOMS benchmark across four operation-centric relational error types.

ID & Operation	Text Snippets (Evidence vs. Claim)	Relational Distortion Analysis
Injection		
`injection_9`	**E:** ... Ducks are examples of sexually reproducing organisms. Other organisms reproduce asexually. In asexual reproduction, a single parent can produce offspring alone. ... **C: Because** ducks are examples of sexually reproducing organisms, other organisms reproduce asexually.	The original text presents adjacent, independent factual statements. The claim injects an artificial causal relationship (*because*) where none exists.
`injection_21`	**E:** ... In endocytosis, a substance from outside the cell is engulfed by the membrane. The membrane folds over the substance and it becomes completely enclosed. There are two main kinds of endocytosis... **C:** The membrane folds over the substance and it becomes completely enclosed by the membrane; **therefore**, there are two main kinds of endocytosis...	The evidence describes a biological process followed by a classification. The claim erroneously injects a consequential marker (*therefore*) to link them.
Omission		
`omission_9`	**E:** Red litmus paper turns blue **when placed in a basic solution**. **C:** Red litmus paper turns blue.	The claim completely omits the essential conditional premise (*when...*), converting a restricted scientific fact into an unconditional falsehood.
`omission_11`	**E:** ... On average, a person requires 1500 to 2000 calories for normal daily activity... **Conversely, if you ingest less than that**, the energy stores in your body will be depleted. **C:** On average, a person requires 1500 to 2000 calories for normal daily activity. The energy stores in your body will be depleted.	The claim retains the negative biological consequence but deletes the critical situational trigger (*if you ingest less than that*), causing an omission error.
Re-anchoring		
`re-anchoring_7`	**E:** Water infiltrates the ground **because** soil and rock are porous. **C: Because** water infiltrates the ground, soil and rock are porous.	The base events are fully preserved, but the claim flips the direction of the causal edge, reversing the premise and conclusion (Porous $\rightarrow$ Infiltrate).
`re-anchoring_13`	**E:** Uncontrolled diabetes may damage capillaries of nephrons. **As a result**, the kidneys lose much of their ability to filter blood. **C:** The kidneys lose much of their ability to filter blood. **As a result**, uncontrolled diabetes may damage capillaries of nephrons.	The evidence sets diabetes as the cause of filtration loss. The claim re-anchors the connective to flip the directional dependency entirely.
Substitution		
`substitution_11`	**E:** Binary fission can be described as a series of steps, **although** it is actually a continuous process. **C: Because** binary fission can be described as a series of steps, it is actually a continuous process.	The structural connective is substituted. A concessive relation (*although*) is replaced with an invalid causal relation (*because*).
`substitution_23`	**E:** ... Water and carbon dioxide are needed for this process, **and** oxygen is produced as a byproduct. **C:** Water and carbon dioxide are needed for this process, **so** oxygen is produced as a byproduct.	The evidence lists coordinate and concomitant elements using a simple conjunction (*and*), which the claim misrepresents as a strict cause-and-effect link (*so*).

Transformers 5.9.0. The 9B models are evaluated on two RTX 4090 GPUs, while all other models and CARVE variants run on a single RTX 4090 GPU. We use spaCy 3.8.14 for dependency parsing. CARVE is training-free, and all baselines are evaluated with fixed prompts or public checkpoints without further tuning. For LLM-based baselines and LLM calls within CARVE, we use deterministic decoding with temperature set to 0.

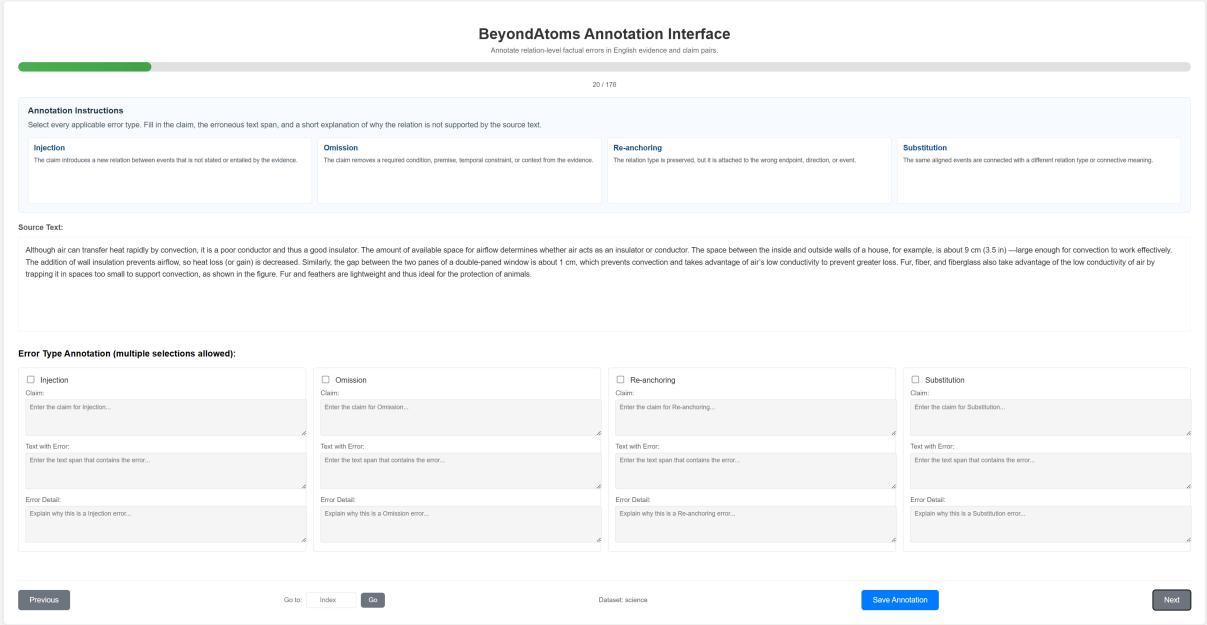


Figure 4: Annotation interface for constructing BEYONDATOMS. Annotators manually write or revise claims based on the evidence passage, target label, and relational operation, while ensuring event preservation and single-operation consistency.

D Prompts for LLM-based Baselines

This section presents the prompts used for LLM-based verification baselines. The generic prompt only provides the standard fact verification instruction, while the taxonomy-revealed prompt explicitly describes the four relational error types considered in BEYONDATOMS.

E Detailed Class-wise Results on BEYONDATOMS

To better understand model behavior on relationally corrupted claims, we report class-wise precision and recall on each BEYONDATOMS subset. The support class corresponds to positive instances, while the not-support class corresponds to relationally corrupted negative instances. Not-support Recall is particularly diagnostic in our setting, as it measures whether a verifier can detect relational errors when the constituent events remain grounded in the evidence. Macro-F1 is included for alignment with Table 5. For several LLM rows, class-wise metrics are computed from the corresponding direct-run records and may differ slightly from the rounded aggregate values in the main table.

F Qualitative Error Analysis

This section provides representative error cases for LLM-thinking and CARVE. The goal is to clarify that the two systems fail for different reasons: LLM-thinking often performs holistic semantic plausibility checking, while CARVE explicitly verifies relations but depends on the quality of the extracted event–relation representation.

F.1 Failure Patterns of LLM-thinking

We use Qwen3.5-9B-Thinking as a representative reasoning-based LLM. As shown in Appendix E, its Support Recall is close to 100% across the four BEYONDATOMS subsets, while its Not-support Recall is substantially lower. Specifically, its Not-support Recall is 70.81% on Substitution, 82.16% on Re-anchoring, 48.92% on Injection, and only 25.13% on Omission. This pattern suggests that thinking helps with explicit replacement or direction reversal, but remains weak when the error involves a fabricated relation or a dropped premise.

Case 1: Condition weakening in Omission. Consider the following Omission example:

> **Evidence:** . . . it oxidizes to B2O3 **when heated with concentrated nitric or sulfuric acid**.
> **Claim:** it oxidizes to B2O3.

The gold label is `not_support`, because the claim removes the condition under which oxidation occurs. However, Qwen3.5-9B-Thinking predicts `support`. The model focuses on the lexical overlap between the claim and evidence, observing that the phrase "it oxidizes to B2O3" is explicitly

Generic Zero-shot Fact Verification Prompt

You are a strict fact verification assistant for the Fact Verification task.
Given a Premise and a Hypothesis, determine whether the Hypothesis is explicitly true according to the Premise.

Premise:
`{premise}`

Hypothesis:
`{hypothesis}`

Rules:
1. Output `true` only if the Premise directly supports the Hypothesis.
2. If the Premise is missing, ambiguous, contradictory, or insufficient, output `false`.
3. Do not use world knowledge beyond the Premise.

Decision:
Answer exactly `true` or `false`.

Figure 5: Generic zero-shot prompt used for LLM-based fact verification.

present. It fails to treat the *when*-clause as a necessary condition for the fact to hold. This illustrates a common failure mode: the model compresses a conditional statement into an unconditional fact.

Case 2: Relation hallucination from co-occurrence. Consider the following Injection example:

> **Evidence:** A cell is the smallest unit of life. Most cells are so tiny that they cannot be seen with the naked eye. Therefore, scientists use microscopes to study cells.
> **Claim:** Because a cell is the smallest unit of life, most cells are so tiny that they cannot be seen with the naked eye.

The gold label is `not_support`. The evidence states two facts: cells are the smallest unit of life, and most cells are tiny. However, it does not support a causal relation between them. The LLM-thinking baseline often treats the two co-occurring facts as sufficient evidence, and may further regard the causal relation as plausible. This demonstrates that model-side reasoning can conflate "A is true and B is true" with "B because A," which is precisely the relational error targeted by BEYONDATOMS.

Case 3: Pragmatic entailment leakage. We also observe failures on DiscInfer, where models introduce pragmatic or world-knowledge assumptions. For example:

> **Evidence:** The cover story details the disturbing behavior of the Littleton killers before last week's massacre.
> **Claim:** The killers were featured on the front cover before the mass killings from last week.

The gold label is `not_support`. The evidence states that a cover story details the killers' behavior before the massacre, but it does not state that the killers were featured on the front cover. The LLM-thinking baseline may infer from "cover story" that the killers appeared on the front cover, relying on pragmatic or publication-domain assumptions not licensed by the evidence. Thus, reasoning can introduce unsupported relations rather than strictly verifying the evidence.

Summary. The main failures of LLM-thinking can be summarized as follows: (1) *condition weakening*, where conditional statements are reduced to unconditional facts; (2) *relation hallucination from co-occurrence*, where two supported events are treated as relationally connected; and (3) *pragmatic entailment leakage*, where external plausibility or world knowledge fills in relations not stated in the evidence. These failures show that generic reasoning does not automatically perform strict relation-level verification.

Taxonomy-revealed Relational Fact Verification Prompt

You are a strict fact verification assistant for the Fact Verification task.
Given a Premise and a Hypothesis, determine whether the Hypothesis is explicitly true according to the Premise.

False hypotheses may contain one of four relational error types:
1. **injection**: the Hypothesis introduces a relation that is not supported by the Premise.
2. **omission**: the Hypothesis removes a required condition or context from the Premise.
3. **re-anchoring**: the same relation type is attached to the wrong endpoint or direction.
4. **substitution**: the Hypothesis changes the relation type between the same aligned events.

Premise:
`{premise}`

Hypothesis:
`{hypothesis}`

Rules:
1. Output `true` only if the Premise directly supports the Hypothesis.
2. If the Premise is missing, ambiguous, contradictory, or insufficient, output `false`.
3. Do not use world knowledge beyond the Premise.
4. If false, choose the most likely error type.

Output Format:
Decision: `true/false`
ErrorType: `none/injection/omission/re-anchoring/substitution`

Figure 6: Taxonomy-revealed prompt used for LLM-based relational fact verification. The prompt explicitly describes the four relational error operations in BEYONDATOMS.

F.2 Failure Patterns of CARVE

CARVE's errors differ from those of LLM-thinking. It usually does not predict support merely because the events overlap; instead, it explicitly extracts relations and aligns them against the evidence pool. Its failures are therefore concentrated in representation errors, including incorrect connective detection, wrong event anchoring, over-aggressive repair, or unresolved discourse references. In our debug results, CARVE errors are distributed as follows: Substitution has 20 false positives and 7 false negatives, Re-anchoring has 60 false positives and 22 false negatives, Injection has 2 false positives and 12 false negatives, and Omission has 7 false positives and 20 false negatives. Re-anchoring is particularly sensitive because it depends on accurate relation direction, endpoint attachment, and discourse anchoring.

Case 1: Missing reverse anchoring in Omission. Consider again the Omission example:

> **Evidence:** ... it oxidizes to B2O3 **when heated with concentrated nitric or sulfuric acid**.
> **Claim:** it oxidizes to B2O3.

The claim contains no explicit connective, so CARVE extracts it as an independent event. The evidence contains a conditional relation:

> **Extracted evidence relation:**
>
> **Result:** it oxidizes to B_2O_3
> **Condition:** heated with concentrated nitric or sulfuric acid
> **Connective:** *when*

If the omission detector fails to anchor the independent claim event to the consequent side of this evidence relation, no error is triggered and the system predicts `support`. This case indicates that omission detection requires robust evidence-to-claim anchoring even when the claim degenerates into a single-event assertion.

Model	Macro-F1	Sup. P	Sup. R	Not-sup. P	Not-sup. R
Generic prompt LLMs					
GPT-5-mini	85.06	88.97	80.12	81.92	90.06
Qwen3.5-9B-Thinking	84.75	77.40	100.00	100.00	70.81
Qwen3.5-9B	71.63	65.18	100.00	100.00	46.58
Qwen3.5-4B-Thinking	81.73	73.85	100.00	100.00	64.60
Qwen3.5-4B	66.04	61.45	100.00	100.00	37.27
Fine-grained fact verification					
FActScore	55.04	56.43	98.14	92.86	24.22
NLI and fact-checking models					
NLI-DeBERTa-v3-Large	49.91	50.72	65.22	51.30	36.65
MiniCheck-DeBERTa	57.31	57.25	95.65	86.79	28.57
FDSpotter	89.03	82.81	98.76	98.46	79.50
Our Method					
CARVE	91.60	95.27	87.58	88.51	95.65

Table 10: Class-wise precision and recall on the BEYONDATOMS Substitution subset. Sup. denotes the support class and Not-sup. denotes the not-support class.

Model	Macro-F1	Sup. P	Sup. R	Not-sup. P	Not-sup. R
Generic prompt LLMs					
GPT-5-mini	83.68	91.37	74.69	78.60	92.95
Qwen3.5-9B-Thinking	90.79	84.86	100.00	100.00	82.16
Qwen3.5-9B	78.56	70.47	100.00	100.00	58.09
Qwen3.5-4B-Thinking	90.37	83.68	100.00	100.00	80.50
Qwen3.5-4B	75.24	68.77	99.59	99.25	54.77
Fine-grained fact verification					
FActScore	49.44	53.97	95.85	81.48	18.26
NLI and fact-checking models					
NLI-DeBERTa-v3-Large	46.12	47.71	60.58	46.02	33.61
MiniCheck-DeBERTa	49.35	54.15	97.51	87.50	17.43
FDSpotter	75.49	69.18	95.02	92.05	57.68
Our Method					
CARVE	82.88	89.16	75.10	78.49	90.87

Table 11: Class-wise precision and recall on the BEYONDATOMS Re-anchoring subset. Sup. denotes the support class and Not-sup. denotes the not-support class.

Case 2: Connective ambiguity in Injection. The following Injection example illustrates an extraction failure:

> **Claim:** The cells of the dorsal root ganglion are unipolar cells, classifying them by shape; therefore, the small round nuclei of satellite cells can be seen surrounding—as if they were orbiting—the neuron cell bodies.

The gold label is `not_support`, because the claim injects a *therefore* relation. However, CARVE may incorrectly detect the *if* inside the idiomatic expression *as if* as a conditional connective, while missing the more important discourse marker *therefore*. If a similar *as if* expression appears in the evidence, the system may incorrectly align the wrong relation and predict `support`. This case shows that connective detection cannot rely only on surface tokens; idiomatic contexts such as *as if*, *even if*, or *what if* require stronger contextual filtering.

Case 3: Over-aggressive claim repair in Re-anchoring. Consider the following Re-anchoring example:

> **Evidence:** Uncontrolled diabetes may damage capillaries of nephrons. As a result, the kidneys lose much of their ability to filter blood.
> **Claim:** The kidneys lose much of their ability to filter blood. As a result, uncontrolled diabetes may damage capillaries of nephrons.

The gold label is `not_support`, because the causal direction is reversed. In some cases, the claim-side repair step may over-normalize the claim into a relation closer to the evidence direction, effectively correcting the erroneous claim instead of preserving its surface structure. If the resulting repaired relation aligns with the evidence, the Re-anchoring error is missed. This case highlights a key requirement for LLM-assisted repair:

Model	Macro-F1	Sup. P	Sup. R	Not-sup. P	Not-sup. R
Generic prompt LLMs					
GPT-5-mini	82.25	83.33	80.65	81.25	83.87
Qwen3.5-9B-Thinking	74.37	66.07	99.46	98.91	48.92
Qwen3.5-9B	55.11	56.40	99.46	97.73	23.12
Qwen3.5-4B-Thinking	64.86	62.00	100.00	100.00	38.71
Qwen3.5-4B	49.39	54.25	99.46	96.77	16.13
Fine-grained fact verification					
FActScore	36.97	50.55	98.39	70.00	3.76
NLI and fact-checking models					
NLI-DeBERTa-v3-Large	40.30	48.23	80.65	40.98	13.44
MiniCheck-DeBERTa	47.99	53.94	99.46	96.55	15.05
FDSpotter	82.01	74.30	99.46	99.19	65.59
Our Method					
CARVE	96.23	93.88	98.92	98.86	93.55

Table 12: Class-wise precision and recall on the BEYONDATOMS Injection subset. Sup. denotes the support class and Not-sup. denotes the not-support class.

Model	Macro-F1	Sup. P	Sup. R	Not-sup. P	Not-sup. R
Generic prompt LLMs					
GPT-5-mini	72.41	68.72	83.42	78.91	62.03
Qwen3.5-9B-Thinking	56.88	57.19	100.00	100.00	25.13
Qwen3.5-9B	42.55	52.25	99.47	94.44	9.09
Qwen3.5-4B-Thinking	51.76	55.49	100.00	100.00	19.79
Qwen3.5-4B	41.67	52.23	100.00	100.00	8.56
Fine-grained fact verification					
FActScore	42.92	49.52	82.35	47.62	16.04
NLI and fact-checking models					
NLI-DeBERTa-v3-Large	46.27	50.17	77.54	50.59	22.99
MiniCheck-DeBERTa	39.79	51.39	98.93	85.71	6.42
FDSpotter	55.22	56.67	100.00	100.00	23.53
Our Method					
CARVE	92.77	90.00	96.26	95.98	89.30

Table 13: Class-wise precision and recall on the BEYONDATOMS Omission subset. Sup. denotes the support class and Not-sup. denotes the not-support class.

it should recover missing structure, but must not rewrite the relational direction or correct the claim using semantic plausibility.

Case 4: Wrong evidence relation anchor. CARVE may also produce false positives when the evidence relation is extracted with the wrong left-side anchor. For example:

> **Claim:** There are many more genes located on the X chromosome than there are on the Y chromosome. As a result, there are many more X-linked traits than there are Y-linked traits.
> **Evidence:** . . . there are many more genes located on the X chromosome than there are on the Y chromosome. As a result, there are many more X-linked traits than there are Y-linked traits.

The claim is supported by the evidence. However, if the evidence pool incorrectly extracts a relation such as:

> **Extracted evidence relation:**
>
> | | |
> |---|---|
> | **Cause:** | Most sex-linked traits are recessive |
> | **Result:** | there are many more X-linked traits |
> | **Connective:** | *as a result* |

then the left argument no longer matches the claim-side cause. The downstream verifier may diagnose a missing relation or Injection error. This is not a failure of the alignment logic itself, but of evidence-side relation anchoring, especially for cross-sentence discourse markers such as *as a result*.

Case 5: Discourse anaphora and bridging. CARVE is also sensitive to discourse anaphora. Consider the following supported claim-evidence pair:

> **Claim:** Winds are stronger than on any other planet in the solar system. This is because Neptune receives little energy from the Sun to power weather systems.

> **Evidence:** Winds are stronger than on any other planet. The rapid winds surprised astronomers. This is because Neptune receives little energy from the Sun to power weather systems.

The evidence supports the claim, but the phrase *This is because* may be incorrectly attached to the nearest sentence, "The rapid winds surprised astronomers," rather than the earlier event about strong winds. The extracted evidence relation then becomes misaligned with the claim relation, leading to a false negative. This illustrates the difficulty of resolving discourse-level references such as *this*, *that*, and *this is because* in multi-sentence evidence.

Summary. CARVE's failures are primarily relation representation errors: the system attempts to verify relations explicitly, but the extracted graph can be wrong. These errors arise from ambiguous connectives, imperfect cross-sentence anchoring, over-aggressive normalization, or unresolved discourse anaphora. This contrasts with LLM-thinking, whose errors mainly stem from relation under-specification: it often verifies whether the events are present, but not whether the relation between them is explicitly supported.