

# SC-RED: A Knowledge-Graph-Grounded Supply Chain Relation Extraction Dataset

Siva Indraneel Dhulipala<sup>1</sup>, Debayan Banerjee<sup>1</sup> and Ricardo Usbeck<sup>1</sup>

<sup>1</sup>Artificial Intelligence & Explainability (AIX), Leuphana University Lüneburg, Germany

## Abstract

Automated relation extraction from supply chain text offers a principled path toward the structured network knowledge that due diligence legislation, such as the German Supply Chain Due Diligence Act, now requires organisations to maintain. The absence of standard supply chain annotation schemas and labelled benchmarks makes existing relation extraction systems difficult to apply and evaluate in this domain. We address this gap by introducing **SC-RED**, a silver-standard dataset of 1,671 supply chain relation triples across 2,328 entities and 1,150 source documents, derived from Wikipedia and grounded in Wikidata and a formal supply chain process reference standard. Its deliberately small scale mirrors a common low-resource scenario. Agentic relation extraction is one candidate approach for such settings, so we explore how AgentRE, a recent agent-based extraction framework, performs under these conditions. Auditing its released implementation against its published description, we identify and correct several undocumented gaps before evaluation. Our ablations demonstrate that simple few-shot retrieval outperforms all memory-augmented variants even on the original baseline dataset. Evaluating this optimal few-shot setup on SC-RED, our error analysis traces its extraction failures directly to unconstrained generation, pointing to explicit document chunking and constrained decoding as clear next steps.

## Keywords

SC-RED, Supply Chain, Dataset, Relation Extraction, Large Language Models, Agentic AI

## 1. Introduction

Supply chains connect raw material sources, manufacturers, logistics operators, and end markets across hundreds of countries. Regulations such as the German Supply Chain Due Diligence Act (LkSG) [1] and the EU's Corporate Sustainability Due Diligence Directive (CSDDD) [2] now require companies to identify and manage risks within their supplier relationships, presupposing a structured understanding backed by something like a queryable database, that most organisations lack. The information needed to build this understanding largely exists as unstructured text, corporate filings, press releases, and trade publications, that is dispersed, heterogeneous, and produced at a rate that makes manual extraction infeasible at scale.

Relation Extraction (RE) offers a principled solution: identifying typed semantic relationships between named entities and producing structured (*subject, predicate, object*) triples that can populate a supply chain knowledge graph [3]. Traditional RE systems, however, require large annotated corpora and generalise poorly outside their training domain. This is a particular limitation for supply chains, where relation annotation is typically conducted internally by individual organisations against their own supplier records rather than released as public benchmarks, making low resource conditions the common rather than the exceptional setting for this domain.

Large Language Models (LLMs) bypass this data bottleneck, enabling few-shot RE through in-context learning [4]. Going a step further, they can act as reasoning agents that interleave logic with tool calls and retrieved evidence [5]. AgentRE [6] applies this paradigm to RE, combining dense retrieval, session-level memory, and iterative reflection. It reports state-of-the-art results on sentence-level

---

Posters, Demos, Blue Sky, and Tutorials at SEMANTiCS 2026, Sep 2026, Ghent, Belgium

✉ [indraneeldhulipala@gmail.com](mailto:indraneeldhulipala@gmail.com) (S. I. Dhulipala); [debayan.banerjee@leuphana.de](mailto:debayan.banerjee@leuphana.de) (D. Banerjee);

[ricardo.usbeck@leuphana.de](mailto:ricardo.usbeck@leuphana.de) (R. Usbeck)

🌐 <https://github.com/sivadhulipala1999> (S. I. Dhulipala);

<https://www.leuphana.de/institute/iis/personen/debayan-banerjee.html> (D. Banerjee);

<https://www.leuphana.de/institute/iis/personen/ricardo-usbeck.html> (R. Usbeck)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

benchmarks (DuIE2.0 and SciERC) in low-resource settings. While its reliance on retrieval makes it a natural candidate for supply chain RE, applying it directly reveals two major obstacles: the supply chain domain lacks a public benchmark to evaluate it on, and the model’s released codebase diverges significantly from its published description.

To solve the first obstacle, we introduce **SC-RED**, a silver-standard supply chain dataset akin to TACRED [7] (sentence-level) or DocRED [8] (document-level). Grounded in SCOR-DS [9] and the Enterprise Ontology [10], SC-RED comprises 1,671 triples across 1,150 documents. We intentionally restrict its scale to mirror authentic low-resource conditions and openly release the pipeline to enable future domain adaptations.

To solve the second obstacle, we audit AgentRE and correct several undisclosed implementation gaps (releasing the fixed architecture as **AgentRE<sub>redux</sub>**). Using this reconstructed model, we conduct controlled ablations on DuIE2.0, revealing that a simple few-shot retrieval configuration actually outperforms the full memory-augmented pipeline. Applying this winning few-shot setup to our SC-RED benchmark, our error analysis identifies recursive self-verification loops and hallucinated pseudo-chunking as critical bottlenecks. By mapping these specific failures, we establish explicit algorithmic chunking and constrained decoding as necessary next steps for supply chain relation extraction.<sup>1</sup>

## 2. SC-RED Construction

Constructing a supply chain relation extraction dataset at scale rules out manual sentence-level annotation as a starting point: the space of supplier, ownership, and distribution relationships is too large, and the domain expertise required to label them correctly is too specialised, for manual annotation to be tractable within a limited annotation budget. The only practical path forward is distant supervision [11], under which a knowledge base relation between two entities is treated as weak, potentially noisy evidence that any co-mention of those entities in text expresses that relation. Distant supervision lets us derive relation labels from an existing structured knowledge base and verify them against naturally occurring text, without requiring a human annotator to label each sentence from scratch.

We instantiate this paradigm using two complementary data sources: Wikidata, which provides structured entity properties and relation statements, and Wikipedia, which provides the natural language text in which these relations are expressed. Wikidata alone yields relation labels without textual grounding, while Wikipedia alone yields text without structured supervision. Combining the two allows us to derive relation labels from Wikidata and verify them against Wikipedia mentions. Crucially, Wikidata and Wikipedia are independently maintained: a relation recorded in Wikidata is not guaranteed to be explicitly stated in the corresponding Wikipedia article, and vice versa. This independence is precisely why distant supervision, rather than a direct mapping between the two sources, is the appropriate approach here, we treat each Wikidata relation only as weak evidence of a potential relation in text, and retain it only once it is verified against an actual textual mention.

To scope this process to the supply chain domain, we ground our ontology in two established schemas. The Supply Chain Operations Reference model’s data standard (SCOR-DS) defines standard supply chain process relationships across five phases: Plan, Source, Transform, Order/Return, and Orchestrate. The Enterprise Ontology defines organisational and product-level entity classes, ensuring that each category maps onto a role with operational relevance to at least one SCOR-DS phase. To extract the specific Wikidata property IDs (PIDs) and entity IDs (QIDs) underlying these frameworks, we initially prompted an LLM to generate candidates (prompts are detailed in Appendix A). We then performed extensive manual corrections and added supplementary PIDs and QIDs to guarantee strict alignment with the schemas. To facilitate efficient in-context learning, we explicitly group these curated QIDs into five overarching entity types, allowing us to easily inject them into the LLM’s context window. The resulting fifteen relation types and five entity categories are summarised in Table 1 and Table 2 respectively, while the full exhaustive list of QIDs can be found in Appendix B.

---

<sup>1</sup>Code and appendix: [https://github.com/semantic-systems/AgentRE\\_redux](https://github.com/semantic-systems/AgentRE_redux)

Using this seed list of PIDs and QIDs, we execute a SPARQL query (Appendix C) against Wikidata to fetch all triples whose subject and object entity types and whose relation type satisfy our ontology. We group the resulting triples by source entity and, for each unique source entity, fetch the corresponding Wikipedia article text. A triple is discarded if its target entity does not appear in the fetched text for its source entity, which anchors each Wikidata-derived label to a verifiable textual span rather than an unsupported knowledge base fact and constitutes our co-mention filtering step. The retained triples, together with their source text, form the SC-RED silver-standard dataset.

To temper extreme outliers in the underlying relation distribution, we apply post-hoc stratified sampling per (`org_type`, `relation`) bucket prior to finalising the dataset. This pipeline yields 1,150 source documents and 1,671 triples. Because we query a scoped subset of Wikidata rather than the full knowledge base, only 11 of our 15 defined relation types are represented in the final dataset. We treat this coverage gap as a deliberate feature rather than a limitation: it allows us to evaluate agentic relation extraction under a low resource setting in the supply chain domain, which we do in Section 4. Figure 1 shows the resulting relation distribution as a percentage breakdown, revealing a pronounced class imbalance. This imbalance likely reflects two compounding factors: Wikidata’s own labelling is unevenly distributed across relation types, and public knowledge of supply chain networks is inherently skewed. For example, companies are far more likely to be publicly known for having subsidiaries than for processing raw materials for one another.

**Table 1**  
SC-RED relation types, Wikidata PIDs, and SCOR-DS phase.

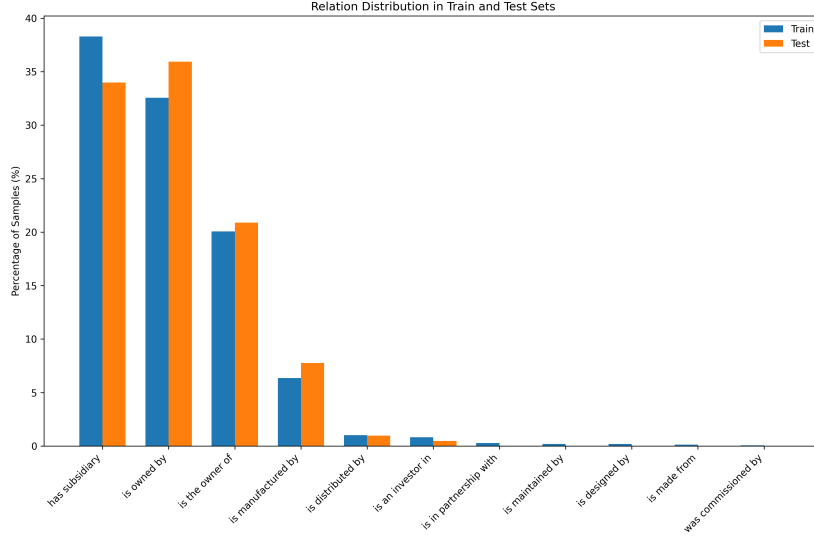
| Relation               | PID   | Phase       | Relation               | PID    | Phase      |
|------------------------|-------|-------------|------------------------|--------|------------|
| is owned by            | P127  | Orchestrate | typically sells        | P7163  | Plan/Order |
| has subsidiary         | P355  | Orchestrate | sources energy from    | P618   | Plan/Order |
| is in partnership with | P2652 | Orchestrate | is manufactured by     | P176   | Source     |
| is regulated by        | P3719 | Orchestrate | is made from           | P186   | Source     |
| is an investor in      | P1951 | Orchestrate | is designed by         | P287   | Transform  |
| is the owner of        | P1830 | Orchestrate | processes raw material | P10672 | Transform  |
| is distributed by      | P750  | Plan/Order  | is maintained by       | P126   | Return     |
| was commissioned by    | P88   | Plan/Order  |                        |        |            |

**Table 2**  
SC-RED entity categories and semantic roles.

| Category           | Semantic role & examples                                                     |
|--------------------|------------------------------------------------------------------------------|
| CoreEnterprise     | Core supply chain company: logistics co., public company, defense contractor |
| SupplyChainPartner | External supplier/service provider: vendor, OEM, contract manufacturer       |
| OrganizationalUnit | Internal/regulatory body: agency, regulator, joint venture                   |
| MarketActor        | Demand-side endpoint: customer, retailer, competitor                         |
| ProductMaterial    | Goods flowing through the chain: product, raw material, software             |
| SupplyChainEntity  | Abstract supertype used only in relation constraints                         |

### 3. Reconstructing AgentRE

Having constructed SC-RED, we turn to agentic relation extraction. AgentRE [6] remains the primary work in this space: a ReAct-style agent [5] utilizing three modules. The **retrieval module** fetches few-shot samples (*RetrieveExamples*) alongside external context for entities identified in the input text (*RetrieveRelevantInfo*). Reusing this retrieval mechanism, the **memory module** builds two session-level stores frozen for inference: shallow memory (storing correct and incorrect triple predictions in separate memories) and deep memory (holding self-critiquing reflexion, and a historical update revising entity



**Figure 1: SC-RED Relation Distribution**

**Table 3**

AgentRE<sub>redux</sub> ablation on DuIE2.0 (g1m-4.7,  $n = 50$ , F1 over 3 runs)

| Configuration                                                          | Run 1        | Run 2        | Run 3        | Macro F1     |
|------------------------------------------------------------------------|--------------|--------------|--------------|--------------|
| AgentRE <sub>orig</sub>                                                | 48.86        | 55.14        | 39.76        | 47.92        |
| AgentRE <sub>redux-1</sub> (shallow corr., shallow incorr., reflexion) | 33.33        | 30.60        | 29.41        | 31.11        |
| AgentRE <sub>redux-2</sub> (shallow corr., reflexion)                  | 59.00        | 41.00        | 54.74        | 51.58        |
| AgentRE <sub>redux-3</sub> (shallow corr. only)                        | 49.02        | 36.08        | 34.34        | 39.81        |
| ReAct_FSL                                                              | <b>49.16</b> | <b>53.71</b> | <b>51.72</b> | <b>51.53</b> |

data based on successful predictions). Finally, the **extraction module** draws on both to produce relation triples.

Auditing the released codebase revealed four implementation gaps. First, *RetrieveRelevantInfo* is absent. Second, reflexion memory is inaccessible during inference. Third, shallow memory is incorrectly implemented: it unconditionally stores the golden label in correct memory regardless of the model’s actual prediction. This leaves incorrect memory empty and collapses the correct/incorrect distinction described in the paper. Finally, historical updates are missing entirely. We release **AgentRE<sub>redux</sub>**, which reconstructs the first three gaps—including fixing shallow memory to properly route actual predictions into their respective stores—and documents historical updates as an open limitation.

Beyond these codebase gaps, the paper’s reported metrics proved difficult to reproduce. The original work claimed peak performance using a 1000-sample retrieval pool driven by bge-large-zh-v1.5. In preliminary tests on 50-sample subsets, gpt-3.5-turbo failed entirely ( $F1=0$ ) across both DuIE2.0 (a Chinese sentence-level RE dataset) and SciERC (an English sentence-level RE dataset with coreference). While gpt-4-turbo-preview matched the paper’s claims on DuIE2.0, its SciERC performance remained irreproducible ( $F1 \approx 3\%$ ). Consequently, we drop SciERC entirely. We use DuIE2.0 solely for ablation experiments to identify the optimal configuration for SC-RED. Because g1m-4.7 yields comparable performance to gpt-4, we standardise all subsequent runs on this model alongside the 1000-sample pool (see Appendix E for detailed gpt-4 comparisons and our model selection rationale).

Ablating AgentRE<sub>redux</sub> yields five configurations: ReAct\_FSL (retrieval only), AgentRE<sub>orig</sub> (the original flawed implementation), and three variants incrementally adding the corrected memory components. Averaging performance across three runs on a 50-sample DuIE2.0 subset, the retrieval-only ReAct\_FSL matched or exceeded all memory-augmented pipelines (Table 3). Consequently, we carry only ReAct\_FSL forward to evaluate SC-RED.

**Table 4**SC-RED full test set results ( $n = 150$ , F1 over 3 runs).

| Setup              | Run 1       | Run 2       | Run 3       | Macro F1    |
|--------------------|-------------|-------------|-------------|-------------|
| Zero-shot baseline | 3.15        | 3.46        | 3.08        | 3.23        |
| ReAct_FSL          | <b>7.37</b> | <b>6.44</b> | <b>6.85</b> | <b>6.88</b> |

## 4. SC-RED Evaluation and Error Analysis

We evaluate ReAct\_FSL on the 150-sample SC-RED test set, restricting the retriever pool to 1,000 samples to reflect low-resource conditions. We compare against a zero-retrieval g1m-4.7 baseline prompted only with the task description and ontology. Both systems are scored via exact-match macro F1. As Table 4 shows, ReAct\_FSL outperforms the baseline, confirming few-shot retrieval benefits even under severe domain shift. However, this gain is smaller than on DuIE2.0, indicating SC-RED’s longer, heterogeneous documents pose a significantly harder extraction challenge.

Error analysis identifies two unconstrained generation failure modes driving this gap. First, **recursive self-verification** (perpetual self-correction loops without finalizing answers) dominates DuIE2.0 and persists on SC-RED. Second, on SC-RED’s longer texts, the model hallucinates **pseudo-chunking**—attempting to segment the text on its own by randomly selecting isolated sentences for extraction, rather than systematically chunking the full document. These findings motivate explicit input chunking and stricter prompt constraints. Furthermore, exact-match scoring on the silver-standard SC-RED unfairly penalizes valid surface form variations (e.g., “Google” vs. “Google LLC”), requiring generation-aware metrics like GenRES [12]. It also penalizes valid but unannotated relations. Reformulating extraction as link prediction could leverage these novel extractions to automatically propose new edges for community-verified Wikidata enrichment.

## 5. Conclusion and Limitations

We introduce SC-RED, a silver-standard supply chain relation extraction benchmark, alongside an open pipeline for future schema adaptations. By reconstructing AgentRE, our evaluations demonstrate that a purely few-shot retrieval configuration outperforms memory-augmented variants. Applying this optimal setup to SC-RED highlighted recursive self-verification loops and hallucinated pseudo-chunking as primary failure modes, establishing explicit document chunking and constrained decoding as critical next steps.

While SC-RED provides a critical foundation, future work must address three specific limitations. First, only 11 of the 15 ontology relations are currently attested in the dataset, requiring a broader article sample to capture the remaining types. Second, to maintain architectural parity with the baseline, we retained the bge-large-zh-v1.5 retriever. Future English-dominant pipelines should optimise semantic matching using bge-large-en-v1.5. Finally, strict exact-match F1 scores should not be taken at face value. Silver-standard evaluation inherently penalises valid but unannotated triples as false positives, alongside semantically correct surface-form variations. Future benchmarking must adopt generation-aware frameworks like GenRES to accurately assess true extraction quality and support the automated enrichment of Wikidata.

## Declaration on Generative AI

During the preparation of this work, the author(s) used Claude Sonnet 5.0 and Grammarly to draft content, paraphrase existing text, and check grammar. After using these tools, the authors extensively reviewed and edited the content, taking full responsibility for the final publication.

## References

- [1] Bundesministerium für Arbeit und Soziales, Lieferkettensorgfaltspflichtengesetz, <https://www.bmas.de/lksg>, n.d. Retrieved 2026.
- [2] European Parliament, Council of the European Union, Directive (EU) 2024/1760 on corporate sustainability due diligence, Official Journal of the European Union (2024).
- [3] N. Bach, S. Badaskar, A review of relation extraction, Technical Report, Carnegie Mellon University, Language Technologies Institute, Pittsburgh, PA, 2007. Literature review for Language and Statistics II. <https://www.cs.cmu.edu/~nbach/papers/A-survey-on-Relation-Extraction.pdf>.
- [4] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al., Language models are few-shot learners, in: *Advances in Neural Information Processing Systems*, volume 33, 2020, pp. 1877–1901.
- [5] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, et al., ReAct: Synergizing reasoning and acting in language models, in: *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.
- [6] Y. Shi, G. Jiang, T. Qiu, D. Yang, AgentRE: An agent-based framework for navigating complex information landscapes in relation extraction, in: *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, ACM, 2024.
- [7] Y. Zhang, V. Zhong, D. Chen, G. Angeli, C. D. Manning, Position-aware attention and supervised data improve slot filling, in: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017, pp. 35–45.
- [8] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, et al., DocRED: A large-scale document-level relation extraction dataset, in: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019, pp. 764–777.
- [9] Association for Supply Chain Management, SCOR digital standard (SCOR DS), <https://www.ascm.org/corporate-solutions/standards-tools/scor-ds/>, 2022. Retrieved May 7, 2026.
- [10] M. Uschold, M. King, S. Moralee, Y. Zorgios, The enterprise ontology, *The Knowledge Engineering Review* 13 (1998) 31–89.
- [11] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, in: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, Association for Computational Linguistics, 2009, pp. 1003–1011.
- [12] P. Jiang, J. Lin, Z. Wang, J. Sun, J. Han, GenRES: Rethinking evaluation for generative relation extraction in the era of large language models, in: *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, Volume 1: Long Papers, Association for Computational Linguistics, 2024, pp. 2820–2837.