# TWO-STAGE VERIFIED AND TYPE-CONSTRAINED LLM PIPELINES FOR LOW-COST JAPANESE DOCUMENT-LEVEL RELATION EXTRACTION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Japanese document-level relation extraction (DocRE) is a practical bottleneck for building knowledge graphs from long-form text, yet low-cost large language models (LLMs) often yield low precision when asked to extract all relation triples in a single pass. This is challenging because documents contain many entity pairs, evidence can be distributed across sentences, and generative models tend to over-propose plausible but unsupported relations, producing many false positives that are expensive to manually review. We propose a two-stage pipeline using JSON Schema–constrained structured outputs that explicitly separates recall-oriented candidate generation from precision-oriented verification, and then applies deterministic relation-specific domain/range filtering based on empirically observed (head_type, tail_type) pairs from training data. We evaluate on 10 documents from the JacRED development set selected via character-length–based stratified sampling, comparing a one-shot baseline with the proposed pipeline across Gemini Flash family variants (2.0/2.5/3 preview, with and without "thinking"). Across all configurations, the proposed method reduces false positives and improves precision; it improves micro-F1 in three of five matched settings, demonstrating that verification and simple data-driven constraints can substantially improve low-cost DocRE reliability.

## 1 INTRODUCTION

Document-level relation extraction (DocRE) aims to extract directed semantic relations between entities when supporting evidence may be distributed across multiple sentences. Unlike sentence-level relation extraction, DocRE must cope with long-range dependencies, discourse structure, and phenomena such as coreference and ellipsis, while also handling a combinatorial explosion in candidate entity pairs. DocRED established DocRE as a benchmarked task in English and highlighted that a substantial fraction of relations require multi-sentence reasoning, motivating models that integrate evidence globally rather than locally Yao et al. (2019). For Japanese, progress has historically been constrained by the lack of public resources; JacRED addresses this gap with a Japanese Wikipedia-based dataset that provides entity types, relation labels, and evidence sentences, enabling systematic evaluation and analysis Ma et al. (2024).

At the same time, large language models (LLMs) have become an attractive option for information extraction due to their flexibility and low setup cost. Generative information extraction can be framed as prompting an LLM to output structured predictions without task-specific training, but surveys emphasize persistent reliability issues such as hallucination, sensitivity to prompts, and output-format instability Xu et al. (2023). In DocRE, these issues can be exacerbated: asking a model to "extract all triples in the document" invites over-generation, where the model proposes relations that are plausible given world knowledge but not supported by the document itself.

This paper studies a concrete and practically consequential failure mode in low-cost LLM-based Japanese DocRE: one-shot extraction yields low precision because false positives dominate. This is problematic for knowledge graph construction and curation workflows, where manual verification cost is often the dominant expense and where erroneous triples can propagate into downstream analytics. While one can often improve quality with larger models, extensive prompt engineering, or supervised fine-tuning, such solutions may be incompatible with settings that prioritize low latency and low cost and therefore use smaller, cheaper model variants.

We propose a simple decomposition that targets the dominant error source directly. Our approach uses a two-stage LLM pipeline with schema-constrained structured outputs. Stage 1 is recall-oriented candidate generation: the model outputs a potentially over-complete set of candidate triples in JSON under an explicit schema. Stage 2 is precision-oriented verification: the model is asked, in batches, to judge whether each candidate triple is supported by the document, again producing JSON under a verification schema to ensure deterministic parsing and downstream processing. Finally, we apply a deterministic post-processing step that filters verified triples using relation-specific domain/range constraints computed from the JacRED training data: for each relation, we retain only type pairs (head_type, tail_type) that were empirically observed in training. This yields a lightweight, transparent consistency check that further suppresses spurious extractions.

We validate the approach empirically on 10 documents from the JacRED development set selected via character-length–based stratified sampling. We compare a one-shot baseline to our proposed pipeline across multiple Gemini Flash family configurations (Gemini 2.0 Flash, Gemini 2.5 Flash, Gemini 3 Flash Preview), including configurations with "thinking" disabled and with a thinking budget of 2048 tokens where available. We report micro-averaged precision, recall, and F1 aggregated over the sampled documents, along with TP/FP/FN counts to expose the false-positive dynamics explicitly.

Our contributions are: - We isolate and empirically characterize a low-cost Japanese DocRE setting where the primary failure mode of one-shot LLM extraction is false-positive over-generation, leading to low precision. - We introduce a two-stage, schema-constrained LLM pipeline that separates candidate generation from verification, enabling targeted reduction of false positives. - We incorporate deterministic, relation-specific domain/range filtering based on empirically observed type pairs from training data, providing a transparent and costless inference-time constraint. - We provide an evaluation across Gemini Flash family variants showing consistent precision gains and frequent F1 improvements from the proposed pipeline on JacRED development documents.

The results support the central hypothesis that verification and type constraints reduce false positives, but they also reveal that recall remains uniformly low, making recall improvement the primary bottleneck. Future work should therefore prioritize methods that recover additional gold relations without sacrificing the precision benefits of verification, including improving entity alignment robustness, correcting relation directionality, and better capturing implicit or multi-hop document-level relations.

## 2 Related Work

Our work is situated at the intersection of document-level relation extraction, LLM-based generative information extraction, and structured output generation.

DocRE benchmarks and supervised modeling. DocRED introduced a large-scale Wikipedia-based DocRE benchmark with sentence-level evidence annotations and established micro-averaged metrics for relation prediction, making explicit the prevalence of cross-sentence reasoning in relation extraction Yao et al. (2019). The subsequent literature, surveyed comprehensively by Delaunay et al., includes graph-based approaches, transformer-based architectures, evidence selection modules, and global reasoning mechanisms that address long-context integration and multi-hop inference Delaunay et al. (2023). These methods typically treat DocRE as multi-label classification over entity pairs, often with learned thresholds or adaptive decision rules; their core difficulty is modeling and inference over many candidate pairs, rather than producing outputs in a strict structured format.

In contrast, our work focuses on a prompting-based, training-free pipeline using low-cost LLMs, where the dominant error is not primarily miscalibration of classification thresholds but uncontrolled generative over-proposal. This shifts emphasis from representation learning to output reliability and post-hoc validation.

Japanese DocRE resources and LLM limitations. JacRED provides the first public Japanese Wikipedia-based DocRE dataset with evidence sentences and a tailored relation schema, enabling reproducible studies in Japanese Ma et al. (2024). Importantly for our motivation, the JacRED study reports that in-context LLM baselines perform poorly compared to supervised Japanese models, underscoring that naive prompting is insufficient for high-quality Japanese DocRE. Our work uses JacRED as an evaluation setting but targets a different objective: improving precision and practical usability of low-cost LLM extraction through decomposition and deterministic constraints, rather than maximizing absolute performance via supervised training.

Generative IE and constrained decoding. Generative IE surveys emphasize that LLMs can serve as general-purpose extractors via prompting and in-context learning, but they also highlight well-known reliability issues such as hallucination and sensitivity to output formats Xu et al. (2023). Constrained decoding and schema-based structured outputs are widely used to improve parseability and compliance. Recent benchmarking work systematizes this area by evaluating JSON Schema–constrained generation engines and showing that structured output constraints can improve compliance and sometimes even downstream task accuracy Geng et al. (2025). Our use of JSON Schema constraints is aligned with these findings, but our central aim is not schema validity alone; instead, schema constraints serve as an enabling mechanism for a two-stage decision process (generate, then verify) and for deterministic downstream filtering.

Post-hoc verification and consistency constraints. While many DocRE models incorporate global consistency implicitly through architecture or joint objectives (for example, entity-centric graphs or evidence-aware modules summarized in Delaunay et al. (2023)), we employ a simpler and fully transparent alternative: a verifier step that explicitly checks support for each candidate triple, followed by deterministic type-pair constraints learned from the training data. This differs from learned global inference in assumptions and failure modes: our constraints are lightweight and costless at inference time but may be brittle to domain shift or to valid but rare type combinations that were unseen in training.

Overall, prior work either (i) advances supervised DocRE modeling and benchmarking Yao et al. (2019); Delaunay et al. (2023); Ma et al. (2024), or (ii) studies LLM-based IE and the role of constrained decoding for reliable structured outputs Xu et al. (2023); Geng et al. (2025). Our contribution is to connect these strands in a low-cost Japanese DocRE pipeline that uses structured outputs to operationalize verification and uses empirical type constraints to further suppress false positives.

## 3 Background

DocRE task definition and evaluation. We consider a document D consisting of text segmented into sentences, along with a set of entities (each potentially having multiple mentions). DocRE seeks a set of directed relation triples of the form (h, r, t), where h is a head entity, t is a tail entity, and r is a relation label from a fixed schema. Following standard DocRE practice, evaluation compares a predicted triple set P against a gold triple set G and reports micro-averaged precision, recall, and F1 aggregated across documents Yao et al. (2019). Let TP = $|P \cap G|$, FP = $|P \setminus G|$, and FN = $|G \setminus P|$. Then precision is TP / (TP + FP), recall is TP / (TP + FN), and F1 is 2PR / (P + R) when denominators are nonzero. Micro-averaging is appropriate for DocRE because each document contains many candidate pairs and because the output is naturally a set of factual assertions.

JacRED as a Japanese DocRE setting. JacRED is a Japanese Wikipedia-based DocRE dataset that provides entity types, relation labels, and evidence sentences supporting each annotated relation Ma et al. (2024). These characteristics are central to our design. First, evidence may be cross-sentence, which increases ambiguity and raises the risk that a generative model proposes relations that are plausible but unsupported. Second, entities are typed, enabling type-based constraints on candidate triples.

Generative IE with LLMs and the over-generation problem. In generative information extraction, a model produces structured outputs Y conditioned on input text X and an instruction or prompt P, conceptually following an autoregressive conditional generation objective $p(Y|X, P)$ Xu et al. (2023). For DocRE, a naive "extract all relations" prompt induces a mismatch between the open-ended nature of generation and the strict requirement of text-grounded factuality: the model can easily propose unsupported relations. In our experiments, this manifests as many false positives under one-shot extraction, depressing precision.

Structured outputs and constrained decoding. Constrained decoding restricts model outputs to a formal language, such as JSON that conforms to a schema, and is commonly used to increase reliability by ensuring outputs can be parsed deterministically Geng et al. (2025). In our context, structured outputs serve two roles. First, they enforce that both extraction and verification stages emit machine-readable JSON, avoiding brittle heuristics for parsing free-form text. Second, they help stabilize the interface between stages, enabling batch verification and deterministic post-processing.

Problem emphasis and assumption. The central hypothesis tested in this paper is that in low-cost LLM DocRE settings, false positives (FP) dominate the error profile under one-shot extraction, and thus precision is the limiting factor for practical knowledge graph construction. Our method is designed to reduce FP through verification and type consistency constraints, accepting that recall may change (and may decrease) as a trade-off. We do not introduce new datasets or new relation schemas; we operate within the JacRED relation and type inventory.

## 4 Method

We propose a pipeline that turns one-shot, open-ended LLM DocRE into a controlled process consisting of recall-oriented generation, precision-oriented verification, and deterministic type-based filtering.

Pipeline overview. Given a document D, the pipeline outputs a predicted triple set P. The baseline approach produces P in a single LLM call. Our proposed approach produces P through three steps: Stage 1 candidate generation, Stage 2 verification, and Stage 3 type-constraint filtering.

Stage 1: Candidate generation under EXTRACTION_SCHEMA. Stage 1 prompts the LLM to extract entities and relation candidates from D and to return them in JSON that conforms to a predefined EXTRACTION_SCHEMA. Constrained decoding ensures the output is schema-compliant, which makes parsing deterministic and avoids downstream failures due to malformed JSON. The Stage 1 policy is intentionally recall-oriented: it may include candidates that are weakly supported or uncertain, with the goal that later stages will remove false positives.

Stage 2: Batch verification under VERIFICATION_SCHEMA. Stage 2 takes each candidate triple c = (h, r, t) from Stage 1 and asks the LLM to judge whether the relation is supported by D. To reduce overhead and encourage consistent decisions, candidates are verified in batches of size 10. The verifier returns decisions in JSON that conforms to a predefined VERIFICATION_SCHEMA, again via constrained decoding. Operationally, this turns the extraction problem into many smaller support-checking decisions, which is aligned with our objective of reducing hallucinated relations.

Stage 3: Deterministic domain and range filtering from training data. After verification, we apply a deterministic filter based on entity types. For each relation r, we compute a set of observed type pairs Tr = { (type (h), type (t))} from the JacRED training data (as exposed by our preprocessing pipeline). A verified candidate (h, r, t) is retained only if (type (h), type (t)) is in Tr. This approximates relation-specific domain/range constraints using empirical evidence from training annotations and provides a transparent, cost-free consistency check at inference time.

Rationale and expected effects. Stage 2 is designed to reduce FP by requiring explicit support judgments rather than relying on the model's tendency to propose plausible relations. Stage 3 further reduces FP by removing type-inconsistent relations that can arise from entity confusion or directionality mistakes. The main trade-off is the possibility of recall loss: verification can be conservative, and type constraints can filter out correct relations whose type pairing is rare or absent in training.

Implementation. The pipeline is implemented with the Google GenAI SDK and the Gemini API using structured outputs. It uses two JSON schemas (EXTRACTION_SCHEMA and VERIFICATION_SCHEMA), performs batch verification with batch size 10, and applies relation-specific (head_type, tail_type) filtering derived empirically from training data. Our use of schema-constrained decoding is motivated by evidence that structured-output engines improve the reliability of JSON generation Geng et al. (2025), but our method applies this capability specifically to enabling a verifiable, multi-step DocRE pipeline rather than solely to ensure syntactic validity.

## 5 Experimental Setup

Dataset and sampling procedure. Experiments are conducted on the JacRED development set Ma et al. (2024). To keep evaluation costs low while controlling for document-length variation, we select 10 development documents using character-length–based stratified sampling. Evaluation is performed at the document level: for each sampled document, the system predicts a set of relation triples, which are then compared against gold triples for that document.

Compared methods. We evaluate two pipelines: - Baseline (One-shot extraction): a single LLM call extracts entities and relations simultaneously, followed by simple filtering. - Proposed (Two-Stage KG Extraction): Stage 1 candidate generation, Stage 2 batch verification, and Stage 3 deterministic type-constraint filtering.

Models and configurations. We test multiple Gemini Flash family models and settings: - Gemini 2.0 Flash with thinking set to none. - Gemini 2.5 Flash with thinking off. - Gemini 2.5 Flash with thinking budget 2048. - Gemini 3 Flash Preview with thinking off. - Gemini 3 Flash Preview with thinking budget 2048.

Evaluation metrics and reporting. We compute micro-averaged precision, recall, and F1 over extracted triples aggregated across the 10 documents, consistent with standard DocRE evaluation Yao et al. (2019). To diagnose the hypothesized false-positive problem, we additionally report TP, FP, and FN counts for each configuration.

Implementation details. The experimental pipeline uses the Gemini API through the Google GenAI SDK, with structured outputs (JSON Schema–constrained decoding) for both extraction and verification. Verification is executed in batches of size 10. Final prediction sets are filtered using empirically observed relation-specific (head_type, tail_type) constraints derived from the JacRED training data. Because inference primarily consists of API calls, we do not report or assume specific local hardware characteristics.

Fairness considerations. Baseline and Proposed methods are evaluated on the same 10 sampled documents and under the same model configurations. The key difference is the computational structure: the Proposed method uses more API calls due to verification and includes deterministic filtering, which can change the precision–recall trade-off. We therefore interpret improvements with attention to both TP and FP movements, not only to F1.

## 6 Results

All results reported below come directly from the executed runs over the 10 sampled JacRED development documents. No figures were generated or saved in the experiment logs (figures list is empty), so results are presented numerically.

Experiment 1: Gemini 2.0 Flash (thinking: none). The one-shot Baseline achieves precision 0.20, recall 0.15, and F1 0.17, with TP = 22, FP = 86, and FN = 126. The Proposed pipeline achieves precision 0.35, recall 0.19, and F1 0.25, with TP = 28, FP = 51, and FN = 121. This setting illustrates the intended behavior of the pipeline: FP decreases substantially (86 to 51) while TP increases (22 to 28), yielding an F1 improvement.

Experiment 2: Gemini 2.5 Flash (thinking: off). The Baseline achieves precision 0.17, recall 0.16, and F1 0.17 (TP = 24, FP = 115, FN = 124). The Proposed pipeline achieves precision 0.30, recall 0.12, and F1 0.17 (TP = 18, FP = 42, FN = 130). FP decreases sharply (115 to 42), confirming the precision effect, but TP also decreases (24 to 18), offsetting the precision gains in terms of F1.

Experiment 3: Gemini 2.5 Flash (thinking: 2048). The Baseline achieves precision 0.18, recall 0.17, and F1 0.17 (TP = 25, FP = 115, FN = 123). The Proposed pipeline achieves precision 0.36, recall 0.21, and F1 0.27 (TP = 31, FP = 54, FN = 117). Relative to the Baseline, FP decreases (115 to 54) and TP increases (25 to 31), producing the largest F1 improvement observed for Gemini 2.5 Flash.

Experiment 4: Gemini 3 Flash Preview (thinking: off). The Baseline achieves precision 0.26, recall 0.16, and F1 0.20 (TP = 24, FP = 70, FN = 124). The Proposed pipeline achieves precision 0.36, recall 0.22, and F1 0.27 (TP = 32, FP = 56, FN = 116). This configuration shows simultaneous gains in precision and recall, and it was identified in the analysis report as a strong cost–performance option due to improved F1 without enabling thinking.

Experiment 5: Gemini 3 Flash Preview (thinking: 2048). The Baseline achieves precision 0.31, recall 0.22, and F1 0.26 (TP = 33, FP = 74, FN = 115). The Proposed pipeline achieves precision 0.37, recall 0.20, and F1 0.26 (TP = 30, FP = 52, FN = 118). As in other settings, FP decreases (74 to 52), but TP also decreases (33 to 30), leading to no net change in F1.

Cross-configuration summary. The Proposed pipeline reduces FP in every matched comparison, directly supporting the paper's central hypothesis that one-shot extraction over-generates unsupported relations. Precision improves in all five configurations, moving from 0.17–0.31 under the Baseline to 0.30–0.37 under the Proposed pipeline. F1 improves in three settings (Gemini 2.0 Flash none; Gemini 2.5 Flash 2048; Gemini 3 Flash Preview off) and is unchanged in two (Gemini 2.5 Flash off; Gemini 3 Flash Preview 2048). Recall remains low across all settings (0.12–0.22), consistent with the analysis report's conclusion that missing gold relations (FN) is the primary remaining bottleneck.

Ablation evidence through component behavior. Although the experiment logs do not include a separate ablation that isolates Stage 2 versus Stage 3, the consistent FP reductions across all configurations are consistent with the combined effect of verification plus type filtering. The observed cases where TP decreases (Experiments 2 and 5) are also consistent with the pipeline's expected failure mode: conservative verification and type constraints can remove correct but difficult relations.

Limitations. First, evaluation uses only 10 JacRED development documents, and thus results may be sensitive to sampling variance and may not reflect the full dev/test distribution. Second, the pipeline sometimes reduces TP along with FP, indicating that verification can be overly conservative in some model configurations. Third, type-pair constraints are derived from observed training data and may remove correct relations that involve rare or unseen type combinations, especially under distribution shift. Finally, even when precision improves, overall recall remains low, aligning with known DocRE challenges such as entity alignment errors, relation directionality mistakes, and difficulty capturing implicit or multi-hop relations described in the DocRE literature Delaunay et al. (2023).

## 7 Conclusion

This paper investigated low-cost LLM-based Japanese document-level relation extraction on JacRED, focusing on a practical reliability issue: one-shot extraction tends to produce many false positives, resulting in low precision and high downstream verification costs. We proposed a two-stage pipeline that separates recall-oriented candidate generation from precision-oriented verification using JSON Schema–constrained structured outputs, followed by deterministic post-processing via relation-specific domain/range type-pair constraints derived from training data.

Across all tested Gemini Flash family configurations, the proposed pipeline consistently reduced false positives and improved precision. Micro-F1 improved in three of five matched settings, with particularly strong gains for Gemini 2.5 Flash with a thinking budget of 2048 and for Gemini 3 Flash Preview with thinking disabled. However, recall remained low overall, and in two configurations the additional verification and constraint steps reduced true positives enough to offset precision gains.

These findings suggest that practical improvements in knowledge graph extraction can be achieved without fine-tuning by combining decomposition, verification, and simple data-driven constraints. The primary remaining challenge is recall. Future work should therefore focus on recovering additional gold relations while preserving precision gains, for example by improving entity alignment robustness, mitigating relation directionality errors, and refining prompts or decompositions to better capture implicit and multi-hop relations at the document level.

This work was generated by AIRAS (**?**).

## References

Julien Delaunay, Hanh Thi Hong Tran, Carlos-Emiliano González-Gallardo, Georgeta Bordea, Nicolas Sidere, and Antoine Doucet. A comprehensive survey of document-level relation extraction. 2023.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. Generating structured outputs from language models: Benchmark and studies. 2025.

Youmi Ma, An Wang, and Naoaki Okazaki. Building a japanese document-level relation extraction dataset assisted by cross-lingual transfer. 2024.

Derong Xu, Wei Chen, Wenjun Peng, Chao Zhang, Tong Xu, Xiangyu Zhao, Xian Wu, Yefeng Zheng, Yang Wang, and Enhong Chen. Large language models for generative information extraction: A survey. 2023.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. Docred: A large-scale document-level relation extraction dataset. 2019.