SCHEMA-GROUNDED TWO-STAGE 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 important for building knowledge graphs from long documents, but in cost-constrained deployments, low-cost large language models (LLMs) prompted in a one-shot manner often over-generate relations, producing many false positives and unreliable graphs. We ask whether precision can be recovered without training by separating candidate generation from verification while enforcing machine-checkable outputs. We propose a two-stage, schema-grounded pipeline: a recall-oriented generator produces candidate relation triples and a verifier filters them, and we additionally apply simple domain–range (entity-type pair) constraints derived from the training data to eliminate type-incompatible predictions. On 10 stratified-sampled JacRED development documents, this workflow consistently reduces false positives and improves micro-precision and usually F1 across multiple Gemini Flash configurations (e.g., precision improves by 10–15 points and F1 by up to 8 points), while recall remains the main bottleneck. The results indicate that structured, verifiable decomposition plus lightweight schema grounding can make low-cost LLM DocRE more dependable for practical Japanese knowledge graph extraction.

1 INTRODUCTION

Document-level relation extraction (DocRE) identifies semantic relations between entities mentioned anywhere in a document, including cases where evidence spans multiple sentences or requires discourse-level reasoning **??**. As a result, DocRE is both scientifically interesting—because it stresses cross-sentence inference, coreference, and implicit reasoning—and practically important for building knowledge graphs from long-form text used in downstream applications such as search, question answering, and knowledge base population.

Despite rapid progress in supervised DocRE, deploying DocRE systems remains challenging when training data, compute budgets, or engineering time are limited. In such settings, practitioners increasingly rely on LLM prompting because it is training-free and can be adapted quickly **?**. However, recent evidence from JacRED—the first public Japanese Wikipedia-based DocRE benchmark with evidence sentences—shows that LLM in-context learning underperforms supervised Japanese DocRE models and struggles with Japanese-specific discourse and surface-form variation **?**. In our experience, the most acute failure mode in low-cost LLM deployments is not merely low recall, but a flood of false positives: one-shot prompts that ask the model to output all relations at once tend to hallucinate unsupported links, flip relation directionality, or connect entities that co-occur but are unrelated. For knowledge graph construction, this precision failure can render the extracted graph unusable even when some correct relations are present.

This paper addresses a concrete gap: how to make low-cost, training-free Japanese DocRE more reliable under realistic API constraints. Stronger models, multi-sample self-consistency, and iterative agentic refinement can improve accuracy, but they also raise cost and latency and are often infeasible in production. Our direction is to recover precision through workflow design rather than model scaling: we separate "generate candidates" from "verify support," enforce structured outputs to reduce brittleness, and add lightweight constraints grounded in labeled data.

We propose a schema-grounded two-stage pipeline. First, a recall-oriented generator produces candidate relation triples in schema-valid JSON. Second, a verifier re-judges each candidate against the document, also returning schema-valid decisions, enabling deterministic filtering and batching for cost control. Finally, we apply relation-specific domain and range constraints represented as allowed (head_type, tail_type) pairs empirically observed in the JacRED training set; candidates violating these constraints are removed.

We evaluate on 10 JacRED development documents selected by character-length–based stratified sampling and compare against a one-shot baseline across multiple Gemini Flash configurations (2.0/2.5/3 preview, with and without thinking). The proposed workflow consistently improves precision by reducing false positives and often improves micro-F1, while recall remains low, indicating that candidate coverage and entity alignment dominate remaining errors.

Contributions: - We identify and quantify false-positive overgeneration as a dominant practical failure mode of low-cost, one-shot LLM prompting for Japanese DocRE. - We demonstrate that a verification-centered, schema-grounded decomposition improves reliability: across model configurations, it consistently reduces false positives and increases micro-precision under the same model family. - We show that simple, data-derived relation type-pair constraints provide an additional, training-free guardrail that further suppresses incompatible predictions. - We provide a controlled, cost-aware evaluation on stratified-sampled JacRED documents across multiple low-cost LLM configurations, highlighting robustness patterns and recall bottlenecks.

Improving recall without sacrificing the achieved precision—via better candidate generation, entity normalization, and directionality handling—remains the key open problem for making low-cost LLM DocRE broadly practical **?**.

2 Related Work

Our work connects three themes: (i) document-level relation extraction as a benchmarked reasoning problem, (ii) LLM-based information extraction under deployment constraints, and (iii) structured output enforcement and verification.

DocRE as cross-sentence reasoning. DocRED established DocRE as a task requiring document-level evidence aggregation and introduced evidence annotation to facilitate analysis **?**. Subsequent surveys emphasize that DocRE is difficult due to the quadratic space of entity pairs, reliance on coreference and discourse cues, and the prevalence of implicit and multi-hop relations **?**. What is missing for our setting is not another supervised architecture optimized for peak F1, but guidance on how to obtain reliable predictions when training is unavailable and model calls must be cheap.

Japanese DocRE resources and LLM gaps. JacRED highlights the scarcity and distinctiveness of Japanese DocRE data and reports that LLM in-context learning performs poorly on Japanese documents **?**. Prior work primarily uses this observation to motivate supervised modeling for Japanese. Our perspective differs: we treat JacRED as a realistic stress test for training-free extraction and focus on the failure mode that matters operationally—false positives that corrupt downstream knowledge graphs.

LLMs for generative information extraction and decomposition. Surveys of LLM-based IE catalog prompt-based extraction, decomposition, and self-correction strategies **?**. Many approaches assume either stronger models, multiple sampling, or iterative refinement, which can be effective but costly. What remains insufficiently explored is a minimal, cost-aware decomposition that explicitly separates recall-oriented generation from precision-oriented verification, aiming for predictable improvements in precision without training.

Structured outputs and constrained decoding. Recent work on constrained decoding and schema-driven generation shows that enforcing JSON Schema compliance improves output validity and can improve downstream task reliability **?**. Most applications use structured outputs primarily to prevent parsing errors. Our use is more task-directed: we employ schema-constrained outputs to enable batch verification and deterministic post-processing, turning structure into an operational tool for reducing false positives.

In short, supervised DocRE work optimizes for overall accuracy with training; Japanese DocRE work motivates the need for better methods in Japanese; and LLM IE work offers many techniques but often at higher cost. Our contribution is a training-free, low-cost DocRE workflow that emphasizes precision improvements through structured verification and lightweight schema grounding.

3 Background

Task definition. We consider the standard document-level relation extraction (DocRE) problem: given a document d and a set of entities E detected/defined for that document, predict a set of directed relation instances $T \subseteq E \times R \times E$, where R is the relation label set. A relation instance (e_h, r, e_t) is correct if the document entails that relation from head entity e_h to tail entity e_t.

Dataset characteristics. Our experiments use JacRED, a Japanese Wikipedia-based DocRE dataset with evidence sentences, 2,000 documents, 42,241 relation triples, 35 relations (including inverse-augmented relations), and 8 entity types following an IREX-style taxonomy **?**. We evaluate on a small subset of the development split sampled to cover document length variation.

Evaluation protocol. We report micro-averaged precision, recall, and F1 over predicted relation instances aggregated across documents, using standard definitions: precision = TP / (TP+FP), recall = TP / (TP+FN), and F1 = 2PR / (P+R). This aligns with the evaluation practice popularized by DocRED **?** and summarized in DocRE surveys **?**.

Cost-constrained, training-free setting. We study a deployment-motivated regime in which predictions are produced by low-cost LLM API calls without task-specific training. Access to labeled data is limited to computing simple corpus statistics (e.g., relation-wise distributions) that can be used as deterministic constraints, rather than optimizing model parameters.

4 Method

Conceptual overview. The method targets a specific failure mode of low-cost, one-shot LLM extraction: overgeneration of unsupported relations (false positives). Instead of asking the model to perform extraction and implicit verification in a single generation, we decompose the task into (i) producing a broad set of candidates and (ii) explicitly verifying whether each candidate is supported by the document. Throughout, we enforce schema-valid structured outputs so that the pipeline is robust to formatting drift and amenable to deterministic filtering.

Stage 1: candidate generation (maximize coverage). Given a document, the system prompts an LLM to output a set of candidate directed relation triples. This stage is intentionally permissive: it is acceptable to include uncertain candidates, because later steps are designed to remove unsupported ones. The output is generated under JSON Schema–constrained decoding so each candidate triple has consistent fields (e.g., head, relation label, tail).

Stage 2: relation verification (maximize precision). The verifier takes the document and a batch of candidates and returns a binary supported/not-supported decision for each candidate, again under a JSON schema. Batching amortizes overhead and encourages consistent judgments across candidates. This step directly addresses the dominant practical failure mode: hallucinated or weakly implied relations that a one-shot extractor tends to include.

Deterministic schema grounding via type-pair constraints. Even after verification, errors can persist due to relation directionality confusions or semantically plausible but schema-incompatible predictions. We therefore apply relation-specific domain–range constraints represented as allowed (head_type, tail_type) pairs. For each relation r, the allowed set is computed from the JacRED training split as the type pairs observed with r. A verified triple is kept only if its entity-type pair is in the allowed set for its relation.

Terminology and outputs. We refer to the Stage 1 output as candidates, the Stage 2 accepted subset as verified candidates, and the final set after type-pair filtering as predictions. This consistent naming clarifies where false positives are removed and where recall can be lost.

5 Experimental Setup

Data selection and scope. To keep evaluation cost low while covering document-length variability, we evaluate on 10 documents sampled from the JacRED development set using character-length–based stratified sampling **?**. This design improves internal validity relative to ad hoc selection but limits statistical power and generality.

Compared systems (controlled vs varied). We compare (i) a one-shot baseline that extracts all relation triples in a single call and (ii) the proposed two-stage pipeline with verification and type-pair constraints. Across comparisons, we hold constant the underlying LLM family for a given condition and the evaluation procedure; we vary only the pipeline structure (one-shot vs two-stage) and the model configuration.

Model configurations. We test multiple Gemini Flash configurations: Gemini 2.0 Flash (thinking none), Gemini 2.5 Flash (thinking off and thinking 2048), and Gemini 3 Flash Preview (thinking off and thinking 2048). This probes whether the observed effects are robust across low-cost model variants and reasoning budgets.

Evaluation. For each configuration and method, we compute micro precision/recall/F1 over relation instances across the 10 documents and also report TP, FP, and FN counts. These metrics directly quantify whether the proposed workflow reduces false positives (precision gains) and whether any gains come at the cost of missing true relations (recall losses).

Design limitations. The small evaluation set and single sampling draw constrain external validity; results should be interpreted as an indicative, cost-aware comparison rather than a definitive benchmark. Additionally, using training-derived type-pair constraints may filter correct but rare type combinations, potentially underestimating recall.

6 Results

Overall pattern. Across all tested Gemini Flash configurations, the two-stage pipeline reduces false positives relative to one-shot extraction, yielding consistent precision improvements. F1 improves in a majority of configurations but not all, because recall remains low and sometimes decreases when verification and constraints become conservative.

Precision robustness (observation). Precision increases in every configuration, from 0.17–0.31 (baseline) to 0.30–0.37 (two-stage). In terms of error counts, false positives drop substantially in all cases (e.g., 86→51 for Gemini 2.0 Flash; 74→52 for Gemini 3 Flash Preview with thinking 2048).

F1 and recall trade-offs (observation). F1 improves in three of five configurations: Gemini 2.0 Flash (0.17→0.25), Gemini 2.5 Flash with thinking 2048 (0.17→0.27), and Gemini 3 Flash Preview with thinking off (0.20→0.27). In the remaining two settings, F1 is unchanged because true positives also drop (e.g., Gemini 2.5 Flash thinking off: TP 24→18; Gemini 3 Flash Preview thinking 2048: TP 33→30). Recall is consistently the bottleneck for both methods (0.12–0.22), with false negatives dominating the error budget (FN 115–130).

Effect of thinking (observation). Adding thinking improves the one-shot baseline for Gemini 3 Flash Preview (F1 0.20→0.26). For the proposed pipeline, thinking can help (Gemini 2.5 Flash reaches its best F1 under thinking 2048) but can also reduce retained true positives (Gemini 3 Flash Preview: TP 32 with thinking off vs 30 with thinking 2048).

Interpretation (hypotheses). The consistent precision gains support the hypothesis that explicitly separating verification from generation counteracts one-shot overgeneration. The persistent recall ceiling likely reflects candidate coverage and entity alignment limitations, as well as document-level phenomena such as implicit relations and directionality sensitivity; the deterministic type-pair constraints may further reduce recall when gold relations involve rare or unseen type combinations.

Takeaway. The method's primary, robust benefit is precision improvement via false-positive suppression, while improving recall remains the central open challenge in this low-cost regime.

7 Conclusion

We investigated how to make low-cost, training-free Japanese document-level relation extraction more reliable, focusing on the practical failure mode of one-shot LLM prompting: excessive false positives that undermine knowledge graph usability **?**. We introduced a schema-grounded two-stage workflow that separates candidate generation from explicit verification and adds lightweight domain–range constraints based on relation-wise entity-type pairs observed in training data.

Across multiple Gemini Flash configurations on stratified-sampled JacRED development documents, the approach consistently reduced false positives and increased micro-precision, and it improved micro-F1 in a majority of settings. What remains unresolved is recall: both one-shot and two-stage pipelines miss many gold relations, indicating that candidate generation coverage, entity alignment, and implicit cross-sentence reasoning are the limiting factors in this cost-constrained regime.

More broadly, the results suggest that structured, verifiable decomposition and simple data-derived constraints can shift low-cost LLM DocRE from brittle extraction toward more dependable pipeline behavior. Future work should explore recall-oriented improvements—such as better entity normalization, directionality handling, and evidence-focused prompting—while preserving the precision gains demonstrated here **?**.

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