Large Language Models for Knowledge Graph Extraction: A Schema-Constrained Evaluation Framework

Markus Ilves, Eduard Barbu, Jaan Übi
Institute of Computer Science, University of Tartu
Tartu, Estonia
markus.ilves@ut.ee, eduard.barbu@ut.ee, jaan.ubi@ut.ee

Abstract

Large language models enable zero-shot knowledge graph extraction from text, yet evaluation at the level of complete typed graphs remains an open challenge. We present a schema-constrained evaluation framework that combines an explicit ontology of six entity types and 96 relation types with structured generation guided by schema-injected prompts. Supporting both single-step and two-step extraction modes, controlled inference settings, and repeated-run stability analysis, the framework enables systematic benchmarking of LLM-based graph construction under closed ontology constraints. Four large language models Gemini 3 Pro, GPT-5.1, Claude Opus 4.5, and Mistral 7B are evaluated on DocRED using entity and triple F1, schema adherence, and run consistency. Manual review reveals that automatic triple F1 systematically underestimates extraction quality, as a substantial portion of model-predicted triples are textually valid but absent from the incomplete gold annotations. The framework, prompts, and experimental outputs are publicly available for download and experimentation.

Keywords: knowledge graph extraction, large language models, schema constrained generation, evaluation framework, DocRED

1. Introduction

Knowledge graphs (KGs) represent information as structured entity–relation–entity triples and support applications such as information retrieval, question answering, and explainable AI. The automatic construction of KGs from unstructured text has traditionally relied on supervised relation extraction pipelines trained on annotated corpora such as DocRED (Yao et al., 2019). The emergence of large language models (LLMs) capable of zero-shot and few-shot structured generation has introduced a different paradigm for KG construction. Instead of training task-specific extraction models, LLMs can directly generate typed entities and relations from raw text. While this capability substantially simplifies pipeline design, it raises a fundamental methodological question: how should we evaluate complete typed graphs produced by LLMs under explicit schema constraints? Most prior work evaluates LLM-based extraction at the level of isolated subtasks, such as named entity recognition or pairwise relation classification, often restricted to sentence-level inputs. Evaluation at the level of complete KGs is substantially more demanding: entities, relations, and schema constraints must jointly hold, introducing additional challenges related to structured generation, ontology enforcement, and reproducibility. These challenges also expose limitations in widely used benchmarks. DocRED, for instance, is constructed by projecting Wikidata annotations onto text, a process that leaves many text-grounded relations unannotated and thereby complicates automatic scoring. In this work, we introduce a schema-constrained framework for LLM-based KG extraction and evaluation. The system enforces structured generation through explicit entity and relation ontologies, Pydantic-validated outputs, and schema-injected prompts. Single-step and two-step extraction modes are supported alongside controlled inference parameters, persistent caching, and repeated-run stability analysis. Together, these components enable systematic benchmarking of LLM-based graph construction under closed ontology constraints across models and configurations. The framework is dataset-agnostic and can be applied to any corpus providing annotated entities and relations under a defined schema. To provide a concrete benchmark, we evaluate the system on DocRED using four LLMs: Gemini 3 Pro, GPT-5.1, Claude Opus 4.5, and Mistral 7B. Experiments assess entity and triple F1, schema adherence, and run-to-run consistency. A manual analysis further examines the reliability of DocRED as a gold standard, identifying annotation incompleteness, schema rigidity, and temporal instability as systematic confounds that cause automatic triple F1 to underestimate true extraction quality. The main contributions of this work are threefold:

1. A reproducible, schema-constrained extraction system for LLM-based KG construction;
2. A dataset-agnostic benchmarking protocol measuring entity and triple accuracy, schema adherence, and output stability across repeated runs;
3. An empirical demonstration that strict triple-level evaluation against incomplete gold standards systematically underestimates LLM ex-

Proceedings of the Knowledge Graphs and Large Language Models Workshop 2026 (KG-LLM@LREC26), pages 221–228
May 16, 2026.

traction quality, with implications for benchmark design.

Section 2 reviews document-level relation extraction, LLM-based structured generation, and benchmark reliability. Section 3 presents the framework architecture, extraction modes, and schema enforcement mechanisms. Section 4 describes the evaluation dataset and annotation schema. Section 5 reports automatic and manual evaluation results. Section 6 concludes and outlines directions for future work. The framework, prompts, and experimental outputs are publicly available for download and experimentation (Section 7).

2. Related Work

Recent work on LLM-based knowledge graph construction spans three interconnected areas: document-level relation extraction, prompt-driven structured generation, and the reliability of benchmarks used to evaluate these systems.

Document-level relation extraction. Knowledge graph extraction from text has been studied extensively in the context of document-level relation extraction (DocRE), where the goal is to predict all relations between entity pairs mentioned across an entire document rather than within a single sentence. Supervised approaches based on pre-trained language models have established strong baselines on DocRED, with ATLOP (Zhou et al., 2021) introducing adaptive thresholding and localised context pooling to address the multi-label and multi-entity challenges of document-level inference. Subsequent work such as DREEAM (Ma et al., 2023) improved performance further by using evidence sentences as supervisory signals to guide model attention. These systems are trained and evaluated under a closed, schema-defined relation set, a design assumption shared by the present framework, but applied to prompt-driven LLM generation rather than fine-tuned classifiers.

The reliability of DocRED as a benchmark has itself been questioned. Tan et al. (2022) systematically re-annotated 4,053 documents and showed that the original dataset contains widespread false negatives arising from its recommend-revise annotation scheme, with models trained and evaluated on the corrected Re-DocRED achieving gains of around 13 F1 points. This finding directly motivates the manual evaluation in Section 5.2, which identifies the same incompleteness problem in the context of LLM-generated outputs.

LLM-based structured generation and KG construction. The use of LLMs for relation extraction has shifted the paradigm from fine-tuning dedicated models towards prompt-based and zero-shot generation. Wadhwa et al. (2023) showed that few-shot prompting with GPT-3 achieves near state-of-the-art performance on standard relation extraction benchmarks, while also demonstrating that exact-match evaluation underestimates LLM performance due to paraphrase and label mismatch: a concern directly relevant to the triple-level evaluation reported here. Papaluca et al. (2024) evaluated zero- and few-shot triplet extraction across LLMs of different scales, finding that contextual knowledge from a knowledge base strongly correlates with extraction quality and that model size yields only logarithmic improvements.

Schema-guided approaches have emerged as a means of improving output structure and controllability. GoLLIE (Sainz et al., 2024) demonstrated that fine-tuning LLMs on annotation guidelines substantially improves zero-shot generalisation to unseen information extraction tasks, confirming that explicit structural specifications benefit extraction quality. The present system takes a complementary approach: rather than fine-tuning on guidelines, schema constraints are injected at inference time and enforced through Pydantic-validated generation, enabling direct comparison across off-the-shelf LLMs without additional training.

Practical tools for LLM-based KG construction. Beyond research prototypes, practical open-source tools have emerged that perform LLM-based KG construction from unstructured text. McDermott's AI Knowledge Graph Generator (McDermott, 2024) implements a three-phase pipeline: SPO triple extraction, entity standardisation, and relationship inference operating over chunked documents with any OpenAI-compatible LLM endpoint. The Neo4j LLM Knowledge Graph Builder (Neo4j Labs, 2024) offers a production-oriented application supporting schema-configurable entity and relation extraction from diverse input formats, integrated directly with a graph database backend. Both systems demonstrate the practical viability of prompt-driven KG construction, but neither provides a formal evaluation framework, schema adherence metrics, or systematic benchmarking against annotated gold standards. The present framework is designed to address this gap.

3. System Overview

Benchmarking LLM-based KG extraction requires a pipeline that enforces schema constraints consistently across models and configurations, while producing evaluation outputs that are comparable and reproducible. To this end, we present a modular, end-to-end framework that takes document texts with stable identifiers as input and produces

typed entity nodes and relation triples aligned with a fixed ontology. In addition to per-document graphs, the pipeline generates run-level summaries aggregating extraction and evaluation statistics across the corpus.

The overall architecture is shown in Figure 1. The pipeline is configuration-driven and modular: data loading, schema management, extraction, graph normalisation, evaluation, and caching operate as independent components that exchange structured records. This design enables controlled benchmarking across extraction modes and LLMs while maintaining a uniform representation of predicted and gold standard graphs.

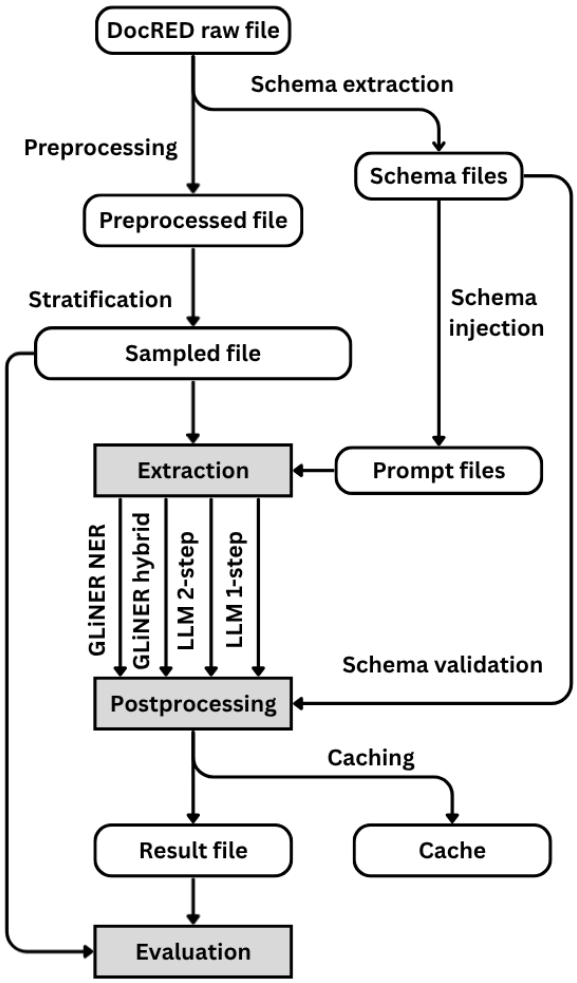

Figure 1: Pipeline architecture for schema-constrained KG extraction and evaluation.

Extraction layer. The extraction layer operates on preprocessed documents and supports either single-step or two-step extraction. In the single-step setting, an LLM jointly predicts entities and relation triples in a structured response. In the two-step setting, entity extraction precedes relation extraction, which is conditioned on the predicted entities. This staged design reduces combinatorial ambiguity and improves consistency. Optionally, GLiNER (Zaratiana et al., 2024; Stepanov et al., 2026) can replace the LLM entity extraction step, either as a standalone NER component or in a hybrid configuration with LLM-based relation extraction. Regardless of configuration, the extraction component returns a standardised record containing entities, triples, and metadata such as latency and truncation flags, ensuring that downstream evaluation is independent of the extraction configuration used.

Schema constraints. Schema constraints operate at three stages: schema extraction, schema injection, and schema validation. The ontology is derived from the dataset and serialised into machine-readable files defining admissible entity types and relation predicates, which serve as the authoritative specification for all downstream components. During schema injection, allowed entity types and relation labels are explicitly enumerated in the prompts, creating a closed schema. A representative fragment is shown in Listing 1. Relation predicates can optionally be enforced at decode time by dynamically constructing a Pydantic enumeration from the ontology labels; when enabled, the structured generation backend restricts the token space so that only schema-compliant predicates can be produced. When this constraint is disabled, model outputs are parsed into typed response models and revalidated against the ontology post-hoc, with out-of-schema predictions explicitly flagged. An excerpt of the relation ontology is shown in Listing 2. Together, these mechanisms ensure structural alignment between predicted graphs and the target schema.

Listing 1: Representative fragment of the schema-injected extraction prompt.

```
You are a knowledge graph extraction
    system.
ALLOWED ENTITY TYPES:
{{allowed_types}}
ALLOWED RELATIONS (closed schema):
{{relation_mappings}}
Return a SINGLE valid JSON object.
Only use allowed types and relations.
```

Listing 2: Excerpt of the relation ontology schema used for validation.

```
schema_version: '3.0'
relations:
  P1001:
    label: applies to jurisdiction
  P102:
    label: member of political party
```

Postprocessing. Extracted outputs are converted into a canonical graph representation that is independent of the specific LLM or extraction mode used. Entity names are normalised, types mapped to schema codes, and triples represented as subject–predicate–object tuples with optional identifiers. The same normalisation is applied to both predicted and gold standard graphs, reducing formatting mismatches and ensuring structurally meaningful comparison.

Evaluation layer. The evaluation layer compares predicted and gold standard graphs at entity and triple levels. Entities are matched using canonical names and aliases, distinguishing exact matches, hallucinations, and omissions. Triples are aligned via assignment-based matching that separates predicate and argument errors while preventing double counting. Precision, recall, and F1 are computed per document and aggregated across the corpus, with structured error categories recorded for detailed inspection (Section 5).

Reproducibility and caching. All hyperparameters, including sampling settings, seeds, and extraction modes, are specified through configuration files and stored with run summaries. A persistent cache links documents and prompts to prior outputs, enabling controlled reuse, regeneration, or bypass of predictions. Repeated runs under identical configurations support explicit quantification of run-to-run variability in entity and triple performance, forming the basis for the stability analyses reported in Section 5.

4. Dataset

Extraction quality was evaluated on DocRED (Yao et al., 2019), a document-level relation extraction dataset constructed by projecting Wikidata entity and relation annotations onto Wikipedia articles. Each document contains multiple sentences and is annotated with co-reference clustered entity mentions and inter-entity relation triples drawn from a fixed inventory of Wikidata properties. Relations may span multiple sentences and multiple triples can hold between the same entity pair, making the dataset suitable for evaluating document-level KG extraction under a fixed schema.

To integrate DocRED into the proposed framework, its original JSON format is converted into a unified graph representation aligned with the system's ontology and evaluation pipeline.

Preprocessing. As the framework is dataset-agnostic, DocRED is converted into a unified JSONL graph representation. Document text is reconstructed by concatenating tokenised sentences. Entities are derived from the `vertexSet` field: the first mention in each co-reference cluster is used as the canonical name, and the entity type is determined by majority vote across mention-level annotations. Mention spans with sentence indices and token offsets are retained for potential span-level analysis.

Relation triples are constructed from the `labels` field. Each triple specifies head and tail entity indices, a Wikidata property identifier, and supporting sentence indices. To enable schema injection, property identifiers are resolved via the Wikidata API to obtain English labels and descriptions. These are cached locally and serialised into a YAML ontology mapping property identifiers to labels and descriptions. The ontology is used both to normalise gold triples and to define the closed schema injected into extraction prompts.

The resulting schema comprises six entity types and 96 Wikidata relation types. Each processed document record contains reconstructed text, typed entities, gold subject–predicate–object triples with property identifiers and evidence indices, and summary statistics.

5. Evaluation

The framework supports stratified sampling over document length, entity count, and triple count. For this study, a shared sample of 10 documents was used for both automatic and manual evaluation. Because tertile-based stratification is unstable for very small samples, documents were partitioned via median splits along the three dimensions, yielding up to eight binary strata. One document was sampled from each non-empty stratum, with remaining slots filled by uniform random sampling. All sampling was performed with a fixed random seed to ensure reproducibility.

5.1. Automatic Evaluation

Extraction quality is evaluated at two levels: entities and triples. All string comparisons are performed after Unicode normalisation, lowercasing, whitespace normalisation, and removal of punctuation artifacts. Metrics are computed per document and aggregated using micro-averaging.

Models evaluated. We evaluate four large language models (LLMs) representing different architectural families and deployment settings: Gemini 3 Pro (Comanici et al., 2025), GPT-5.1 (OpenAI, 2023), Claude Opus 4.5 (Anthropic, 2024), and Mistral 7B (Jiang et al., 2023).[1] The selection spans

[1] For GPT-5.1 and Claude Opus 4.5, no technical report has been published for the specific model versions used. Citations refer to the most recent publicly avail-

proprietary frontier-scale systems (Gemini, GPT-5.1, Claude) and an open-weight mid-sized model (Mistral 7B), allowing comparison across scale and instruction-following capabilities under identical schema-constrained prompting conditions.

Entity evaluation. Entities are matched hierarchically. Exact matching on canonical names is attempted first, followed by alias matching. Optional substring matching can capture boundary variation but is recorded separately as a boundary mismatch. Entities matched under the first two criteria are counted as true positives; unmatched predictions are false positives and unmatched gold entities are false negatives.

Triple evaluation. Triple alignment is formulated as an optimal bipartite matching problem using the Hungarian algorithm. Each predicted–gold pair receives a score based on entity alignment and relation match, by Wikidata identifier or normalised label. Only fully correct triples are counted as true positives; partial matches contribute to structured error analysis but not to metric computation. Precision, recall, and F1 are computed strictly over exact triple matches.

Schema adherence and stability. In addition to precision, recall, and F1, the evaluator records schema violations, defined as predictions outside the relation ontology. Relation label mismatches are tracked separately to distinguish semantic near-misses from structural errors.

For models evaluated across multiple inference runs, per-document entity and triple F1 mean and standard deviation are computed to quantify run-to-run variability. Stability is assessed only across identical configurations to isolate stochastic effects.

Results. Figure 2 reports entity and triple F1 across models. Entity extraction performance is consistently high across proprietary models, ranging from 74.1% (Mistral 7B) to 84.0% (GPT-5.1). This indicates that schema-injected prompting combined with structured decoding effectively constrains entity predictions in zero-shot settings.

Triple extraction, however, is substantially more challenging. Claude Opus 4.5 (39.9%) and Gemini 3 Pro (39.3%) achieve the strongest triple-level performance. GPT-5.1 shows a pronounced drop from entity F1 (84.0%) to triple F1 (21.7%), while Mistral 7B exhibits severe degradation at the triple level (4.2%) despite reasonable entity extraction performance.

able documentation for each model family at the time of writing.

The consistent gap between entity and triple F1 suggests that relation prediction under a closed ontology remains the primary bottleneck in document-level KG extraction. Because triple correctness requires both accurate entity alignment and predicate selection, small predicate mismatches or ontology ambiguities have amplified effects at the graph level. As discussed in Section 5.2, part of the observed triple-level degradation also reflects annotation incompleteness and schema rigidity in the gold standard rather than purely model failure.

Figure 3 provides a structured breakdown of error categories across models. A consistent pattern emerges: missed and hallucinated triples substantially outnumber entity-level errors for all systems. Even models with strong entity F1 exhibit large numbers of missed triples, indicating that relation grounding rather than entity detection drives overall performance degradation.

Mistral 7B shows the most extreme triple-level failure mode, with a very high number of hallucinated triples (266) and missed triples (140), consistent with its low triple F1. In contrast, Claude Opus 4.5 and Gemini 3 Pro produce fewer hallucinated triples and maintain relatively balanced error profiles, aligning with their stronger triple-level scores.

Schema violations remain comparatively rare for proprietary models but increase for Mistral 7B, suggesting greater difficulty adhering to closed ontology constraints in smaller models. Relation mismatches are present across all systems, reflecting predicate ambiguity and ontology rigidity rather than pure structural invalidity.

Taken together, the error distribution reinforces that document-level KG extraction under schema constraints is primarily limited by relation prediction accuracy and ontology alignment rather than entity detection. These findings are further contextualised by the manual evaluation in Section 5.2, which shows that part of the triple-level error burden stems from annotation incompleteness in the benchmark itself.

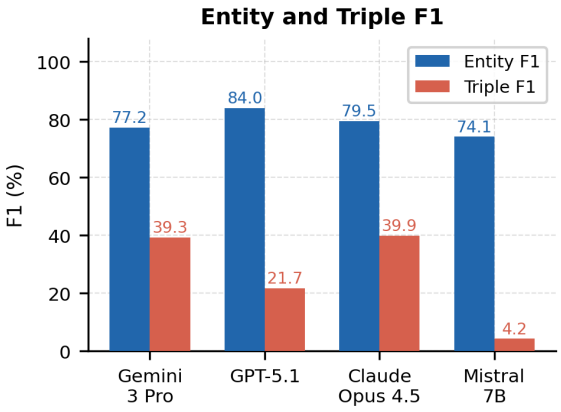

Figure 2: Entity and triple F1 across models.

5.2. Manual Evaluation

To complement automatic scoring, a qualitative analysis of the same 10 documents was conducted. The evaluated system was Gemini 3 Pro Preview in deterministic two-step mode (temperature = 0). Three independent evaluators compared the source documents, the LLM-generated KGs, and the DocRED gold standard annotations.

Aggregate comparison. Across the sample, the gold standard contains 186 entities and 149 triples. The model extracted 205 entities (151 matched, 54 additional) and 156 triples (60 matched, 96 additional). Additional triples account for 61.5% of model predictions, indicating systematic divergence rather than incidental noise.

Systematic discrepancies. Independent review identified three recurring issues.

Annotation incompleteness. Because DocRED annotations are projected from Wikidata, many text-supported relations are absent from the gold standard. In one document describing a journalist and author, the gold annotations contained only two triples, whereas the model extracted fifteen additional relations explicitly supported by the text, including authorship, award received, and publisher relations. Similarly, historical documents containing explicit temporal or geographic information yielded valid `point in time` and `country` relations that were absent from the gold annotations and therefore counted as false positives.

Schema rigidity. Semantically equivalent relations expressed under different but valid property labels are counted as errors. For example, predictions using `parent organization or unit` were penalised when the gold label was `part of`, and `applies to jurisdiction` was penalised when `country` was expected. In addition, several gold triples encoded background Wikidata knowledge, such as nationality of individuals, even when this information was not recoverable from the document itself.

Temporal validity. Some gold triples reflect time-sensitive configurations derived from a live knowledge base, such as office holders at the time of annotation. These relations are independent of the textual content and may become outdated, introducing evaluation noise unrelated to extraction quality.

Implications. The qualitative evidence indicates that a substantial portion of additional model-predicted triples are textually valid but absent from an incomplete gold standard. Automatic triple F1 should therefore be interpreted as a lower bound on actual extraction quality. In documents with rich relational content but sparse annotation coverage, strict matching systematically underestimates model performance.

6. Conclusion

A schema-constrained framework for LLM-based knowledge graph extraction and evaluation has been presented. The system enforces structured generation through explicit ontology injection and output validation, and supports controlled benchmarking across extraction modes, models, and repeated inference settings.

Experiments on DocRED demonstrate that entity extraction under closed-schema prompting is consistently strong across frontier models, while relation prediction remains the primary bottleneck at the graph level. Structured error analysis shows that missed and hallucinated triples dominate performance degradation, and that smaller models struggle to maintain schema adherence under ontology constraints.

Beyond model comparison, the study highlights limitations of benchmark-based evaluation. Manual analysis reveals systematic annotation incompleteness, schema rigidity, and temporal instability in DocRED, indicating that strict triple F1 underestimates true extraction quality. In this setting, triple-level metrics should be interpreted as lower bounds rather than absolute measures of graph correctness.

The proposed framework provides a reproducible, dataset-agnostic protocol for evaluating LLM-based KG extraction under explicit schema constraints. The full system, prompts, evaluation scripts, and experimental outputs are publicly available, and all reported results can be reproduced as described in Section 7. By combining structured generation, stability analysis, and qualitative benchmark assessment, the framework offers a foundation for more reliable graph-level evaluation of LLM-based structured extraction systems.

7. Reproducibility

To ensure transparency and enable reproducibility, the full codebase, evaluation scripts, prompts, and model outputs are shared alongside this paper. The repository includes:

- Benchmarking pipeline code;
- DocRED raw data, preprocessed data, and the evaluated document sample;
- Pipeline results for two-step extraction;
- Prompts and schema used for extraction;
- Evaluation results.

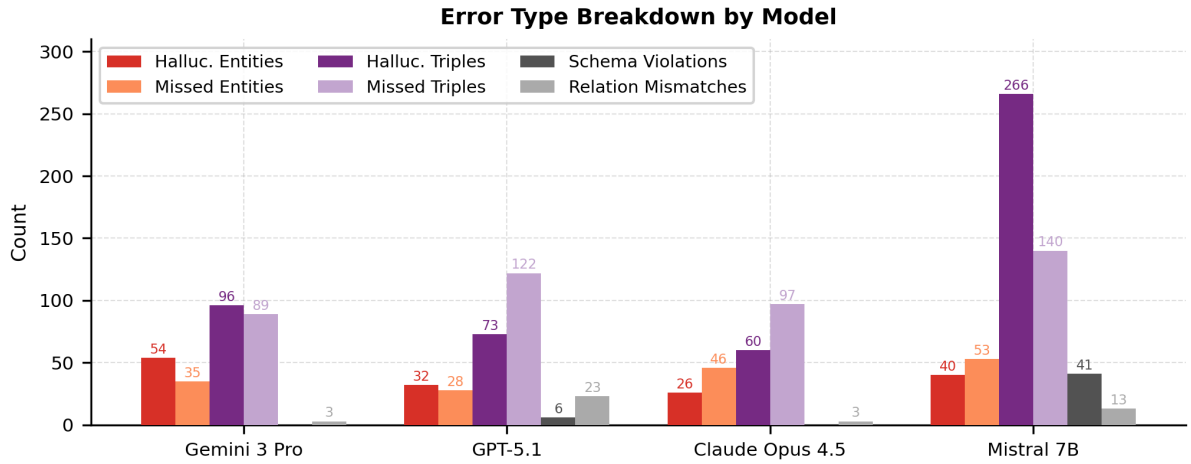

Figure 3: Structured error categories across models.

The repository is publicly accessible via GitHub:

https://github.com/SukSev/kg_benchmarking

All results reported in the paper can be reproduced using the provided scripts. The only external dependency is access to the evaluated LLMs via the OpenRouter API, which requires valid API credentials for each model used in the experiments.

8. Limitations

The evaluation is conducted on a subsample of ten documents from DocRED, which limits the statistical robustness of the reported metrics and may not capture the full variability of model behaviour across different document types and domains. While the subsample was stratified to cover diverse complexity profiles, the results should be interpreted with this scope in mind.

The schema used in the experiments is derived from DocRED and consists of six entity types and 96 Wikidata relation types. This ontology may not generalise to other domains or annotation conventions, and the choice of relation granularity directly affects both extraction difficulty and evaluation outcomes. Systems evaluated against a different schema may exhibit different patterns of schema adherence and relation mismatch.

The framework relies on prompt-based schema injection without fine-tuning, which means that extraction quality is sensitive to prompt formulation and decoding parameters. Small changes in prompt wording or temperature settings may affect both output structure and relational content, particularly for smaller models. The repeated-run consistency analysis captures one dimension of this sensitivity but does not account for variation introduced by prompt design choices.

Finally, the manual evaluation was conducted on the output of a single model and cannot be generalised to characterise the behaviour of all evaluated systems. The qualitative findings regarding DocRED annotation incompleteness are nonetheless expected to apply across models, as they reflect properties of the gold standard rather than of any particular extraction system annotation gaps are model-independent by construction, since any valid extraction absent from the Wikidata-derived annotations will be penalised regardless of which system produced it. Extending manual evaluation to all evaluated models remains a direction for future work.

9. Acknowledgements

This work was supported by the Estonian Research Council grant PRG2006.

10. Bibliographical References

Anthropic. 2024. The Claude 3 model family: Opus, Sonnet, Haiku. Technical report, Anthropic.

Gheorghe Comanici et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. Technical report, Google DeepMind. ArXiv:2507.06261.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Guillaume Lengyel, Guillaume Lample, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Youmi Ma, An Wang, and Naoaki Okazaki. 2023. DREEAM: Guiding attention with evidence for improving document-level relation extraction. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1971–1983, Dubrovnik, Croatia. Association for Computational Linguistics.

Robert McDermott. 2024. AI powered knowledge graph generator. https://github.com/robert-mcdermott/ai-knowledge-graph. Accessed: 2025.

Neo4j Labs. 2024. Neo4j LLM knowledge graph builder. https://github.com/neo4j-labs/llm-graph-builder. Accessed: 2025.

OpenAI. 2023. GPT-4 technical report. Technical report, OpenAI. ArXiv:2303.08774. No technical report has been published for more recent GPT model versions.

Andrea Papaluca, Daniel Krefl, Sergio Rodríguez Méndez, Artem Lensky, and Hanna Suominen. 2024. Zero- and few-shots knowledge graph triplet extraction with large language models. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 12–23, Bangkok, Thailand. Association for Computational Linguistics.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier Lopez de Lacalle, German Rigau, and Eneko Agirre. 2024. GoLLIE: Annotation guidelines improve zero-shot information-extraction. In *Proceedings of the 12th International Conference on Learning Representations*.

Ihor Stepanov, Mykhailo Shtopko, Dmytro Vodianytskyi, and Oleksandr Lukashov. 2026. The million-label ner: Breaking scale barriers with gliner bi-encoder.

Qingyu Tan, Lu Xu, Lidong Bing, Hwee Tou Ng, and Sharifah Mahani Aljunied. 2022. Revisiting DocRED – addressing the false negative problem in relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8472–8487, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Somin Wadhwa, Silvio Amir, and Byron Wallace. 2023. Revisiting relation extraction in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15566–15589, Toronto, Canada. Association for Computational Linguistics.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. GLiNER: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376, Mexico City, Mexico. Association for Computational Linguistics.

Wenxuan Zhou, Kevin Huang, Tengyu Ma, and Jing Huang. 2021. Document-level relation extraction with adaptive thresholding and localized context pooling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14612–14620.