TextMineX: Data, Evaluation Framework and Ontology-guided LLM Pipeline for Humanitarian Mine Action

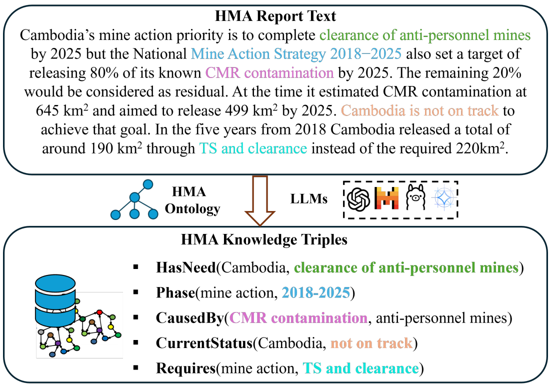
Chenyue Zhou[1,2,4] **Gürkan Solmaz**[1] **Flavio Cirillo**[1]
Kiril Gashteovski[1,3] **Jonathan Fürst**[5]
[1]NEC Laboratories Europe, Germany, [2]University of Stuttgart, Germany
[3]CAIR, Ss. Cyril and Methodius University of Skopje, North Macedonia
[4]VAGO Solutions, Germany, [5]Zurich University of Applied Sciences, Switzerland
zhou@vago-solutions.ai, jonathan.fuerst@zhaw.ch
guerkan.solmaz,flavio.cirillo,kiril.gashteovski@neclab.eu

Abstract

Humanitarian Mine Action (HMA) addresses the challenge of detecting and removing landmines from conflict regions. Much of the life-saving operational knowledge produced by HMA agencies is buried in unstructured reports, limiting the transferability of information between agencies. To address this issue, we propose TextMineX: the first dataset, evaluation framework and ontology-guided large language model (LLM) pipeline for knowledge extraction from text in the HMA domain. TextMineX structures HMA reports into (subject, relation, object)-triples, thus creating domain-specific knowledge. To ensure real-world relevance, we utilized the dataset from our collaborator Cambodian Mine Action Centre (CMAC). We further introduce a bias-aware evaluation framework that combines human-annotated triples with an LLM-as-Judge protocol to mitigate position bias in reference-free scoring. Our experiments show that ontology-aligned prompts improve extraction accuracy by up to 44.2%, reduce hallucinations by 22.5%, and enhance format adherence by 20.9% compared to baseline models. We publicly release the dataset and code[1].

1 Introduction

Humanitarian Mine Action (HMA)—detecting and removing landmines from past (and ongoing) conflicts in order to return land to civilian use—remains a critical humanitarian challenge: in 2022 alone, there were 4,710 casualties globally, 85% of which were civilians (United Nations, 2025; Inclusion, 2023). Over the decades, HMA authorities have published large amount of life-saving knowledge, yet much of it remains locked away in unstructured, free-form text reports, thus making this knowledge largely inaccessible. The automatic extraction and organization of this important information is, therefore, not only a technical advancement, but also a humanitarian imperative that can make life-saving insights more accessible, actionable and transferable across many humanitarian agencies.



Cambodia's mine action priority is to complete clearance of anti-personnel mines by 2025 but the National Mine Action Strategy 2018−2025 also set a target of releasing 80% of its known CMR contamination by 2025. The remaining 20% would be considered as residual. At the time it estimated CMR contamination at 645 km^2 and aimed to release 499 km^2 by 2025. Cambodia is not on track to achieve that goal. In the five years from 2018 Cambodia released a total of around 190 km^2 through TS and clearance instead of the required 220km^2.

- **HasNeed**(Cambodia, **clearance of anti-personnel mines**)
- **Phase**(mine action, **2018-2025**)
- **CausedBy**(**CMR contamination**, anti-personnel mines)
- **CurrentStatus**(Cambodia, **not on track**)
- **Requires**(mine action, **TS and clearance**)

Figure 1: Example knowledge graph triple extraction from texts, guided by the HMA ontology. The task is to extract as many triples as possible while ensuring ontology conformance and source-text faithfulness.

To make HMA knowledge accessible, we turn to Information Extraction (IE). Prior research has shown that structuring information from natural language text data is useful for many downstream tasks, such as question answering (Xiong et al., 2024), fact retrieval (Han et al., 2023) or predicting new facts that were not present originally in the data (Broscheit et al., 2020). More importantly, IE pipelines are useful frameworks for quick information exchange between organizations (Poletto et al., 2021), which is of vital importance for HMA, because the work is typically done in various agencies from different countries.

To effectively structure such domain-specific information, researchers have proposed data, benchmarks and methods for IE in various domains, including finance (Hamad et al., 2024), material science (Cheung et al., 2024), medicine (Romero et al., 2025) and humanitarian crisis response (Fekih et al., 2022). Although LLMs have become ubiquitous, it has been shown that using off-the-shelf LLMs for domain-specific IE is not an optimal ap-

[1] https://github.com/nec-research/TextMineX

Findings of the Association for Computational Linguistics: EACL 2026, pages 2505–2523
March 24-29, 2026 ©2026 Association for Computational Linguistics

proach (Pang et al., 2024; Farzi et al., 2024; Dagdelen et al., 2024). Therefore, there is a need for domain-specific IE applications.

In this work, we propose TextMineX: the first dataset, evaluation framework and pipeline for IE for the HMA domain. The task is to extract information from HMA reports in the form of (subject, relation, object)-triples, thus creating structured knowledge for the HMA domain (Fig. 1). The application provides demining technical information in the form of knowledge triples that can be stored, shared, and queried across demining agencies. Prior data and benchmarks (Mihindukulasooriya et al., 2023, *inter alia*) rely on toy data schemas (i.e., ontologies), which limits their applicability to the complex and specialized HMA documents. By contrast, we utilized the publicly available technical reports from our collaborator Cambodian Mine Action Centre (CMAC). CMAC has been initiated by the UN more than 30 years ago, is one of the world's most experienced and effective demining organizations. Its collaborations within the UN and the Geneva International Centre for Humanitarian Demining (GICHD), ensure that the data is useful in real-world HMA applications beyond Cambodia (Landmine and Cluster Munition Monitor, 2023; United Nations Development Programme in Cambodia, 2021). While there is evidence that ontology-guided methods are effective (Cauter and Yakovets, 2024), they are limited to single-sentence inputs and overlook context-level reasoning. TextMineX addresses this gap by enabling context-level reasoning, while aggregating a set of HMA-related ontologies. Finally, no framework is bias-aware, which might distort the evaluation results, a gap that TextMineX also addresses.

To sum up, our contributions are: ① **Data:** we introduce a curated dataset, curated ontology and annotated ground-truth data for humanitarian demining operations, systematically categorizing operational entities and relationships. ② **Evaluation:** We evaluate extracted triples against our annotated dataset and introduce a bias-aware LLM-as-Judge framework for reference-free scoring. Experiments on closed and open LLMs show that position bias skews rankings.

③ **Knowledge Extraction:** We propose a prompt-based pipeline that combines layout-aware document chunking, ontology-guided extraction, and multi-perspective evaluation. To our knowledge, this is the first LLM application of knowledge extraction for HMA. ④ **In-Context Learning Optimization:** We found that prompts enriched with ontology-aligned examples improve triple extraction accuracy by up to 44.2%, reduce hallucinations by 22.5%, and enhance format conformance by up to 20.9% compared to baseline LLMs. These findings provide practical insights for prompt construction.

2 Related Work

LLMs, such as those from the GPT family (Hurst et al., 2024), Llama (Touvron et al., 2023), BLOOM (Workshop et al., 2022), and PaLM (Chowdhery et al., 2023), have transformed the field of knowledge extraction from text. They possess advanced language understanding and reasoning capabilities, making them well-suited for extracting knowledge from unstructured text or well-structured documents (Colakoglu et al., 2025), especially when paired with prompting techniques like in-context learning (ICL) (Brown et al., 2020).

ICL enables LLMs to learn new tasks by providing input-output demonstrations during inference. Depending on the number of examples provided, this can range from zero-shot (no demonstrations) to one-shot or few-shot learning (multiple demonstrations) (Min et al., 2022; Liu et al., 2023b). This method enhances the models' ability to generalize from minimal data. Zhu et al. (2023) showed the effectiveness of ICL for knowledge extraction. Mihindukulasooriya et al. (2023) introduced an approach that utilizes ontology guidance to extract knowledge from text. Their work highlights the potential of LLMs in extracting domain-specific knowledge constrained by ontological rules. In our study, we adapt this approach to the domain of humanitarian demining, employing a set of specialized ontologies to guide the extraction process.

Evaluating generated texts is a challenging task, especially when limited ground-truth data are available. To address this problem, recent approaches include multi-faceted fact-based evaluation (Gashteovski et al., 2022), generating synthetic data to train an evaluator model (Saad-Falcon et al., 2023; Kim et al., 2025), annotating datasets using a human-in-the-loop methodology (Dagdelen et al., 2024), or leveraging strong LLMs as judges (Zheng et al., 2024; Bavaresco et al., 2024). Our work involves annotating extracted triples to create an evaluation dataset, applying LLMs as judges and analyzing the alignment between these two evaluation methods. For more detailed discussion

Table 1: Dataset Statistics for TextMineX Corpus

Document	Pages	Chars	Words	Sent.	Nums
Annual progress report	170	347,249	49,164	1,887	3,672
Mine clearance report	14	57,537	8,908	452	1,222
Integrated work plan	21	44,952	6,153	323	611
Cluster munition remnant report	9	37,846	5,865	327	893
Article 7 report	19	35,139	5,003	164	920
Total	233	522,723	75,093	3,153	7,318

on related work, see Appendix A.

3 Benchmark Dataset Creation

We illustrate the process of creating the annotated humanitarian dataset in Fig. 2. We curated 120 online available technical reports: 60 reports from the Cambodian Mine Action Center (CMAC)[2] and 60 from Geneva International Centre for Humanitarian Demining (GICHD) websites[3] ①. The reports are all in the PDF format. The curation is based on relevance to the humanitarian demining, language (English), and recency. From the initial humanitarian dataset of 120 PDF documents ①, we filtered a dataset of five m most recent mine action reports from CMAC of overall 233 pages (see Table 1) ②.

Simultaneously, we selected domain-specific ontologies through working with domain experts, by first incorporating data models from the Information Management System for Mine Action (IMSMA)③. We performed a survey of demining related ontologies together with domain experts and we incorporate six ontologies from Information Management System for Mine Action (IMSMA) Core[4]. These ontologies are used for demining information system but not strictly for HMA. Furthermore, we add a more general humanitarian domain ontology from *Empathi* (Gaur et al., 2019) to make the overall HMA ontology more comprehensive. We filter out the concepts that are not relevant to HMA from Empathi by utilizing knowledge of subject-matter expert. For instance, certain concepts related to natural disasters (e.g., flood response) are not directly related to demining. As a result, HMA ontology integrates seven ontologies (160 entity types, 86 relation types) covering diverse aspects of HMA (see details in Sec. 5.1).

The annotation prompt generation ④ first parses reports into text chunks, then systematically combines each chunk with ontology templates. Each template specifies the relation types for a specific domain (e.g., mine action events, land contamination, mine clearance). This generates 2,520 prompts (360 chunks × 7 templates), where each prompt contains both report context and ontology specifications for that domain.

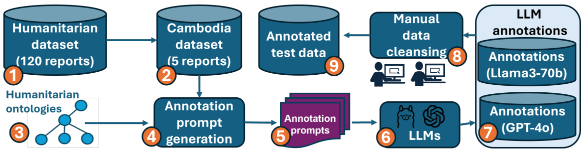


Figure 2: Semi-automatic creation of the humanitarian mine action dataset. The dataset contains a large and diverse set of technical reports from the global mine action and a smaller LLM- and human-annotated portion for Cambodian mine action.

We randomly select 100 prompts ⑤ out of the 2,520, and apply them with the same prompts using GPT-4o and Llama3-70B ⑥ to initially annotate the data as the "LLM annotations⑦.

For each prompt and data chunk pair, each LLM generates a set of knowledge triples as outputs. Then, a human annotator who is an expert in natural language processing working with knowledge graphs reviewed these outputs, filtering out incorrect knowledge triples and aggregating the valid ones in a new set. Incorrect knowledge triples can include triples following wrong format, mixed order of entities, hallucinated entities/relations (e.g., triples that are not included in the data chunk) and so on. In addition, there exist duplicate triples which are removed to create the final reference set. Those included 1,095 unique triples across the 100 prompts⑧. The end result of the human annotation is a set of knowledge triples which are relevant and correct without hallucination or formatting problems, faithful to the given data chunk. This clean set comprises the "annotated test data" ⑨. The annotated test data can serve as the ground-truth for evaluations of performance of different techniques and LLMs.

Finally, to assess the reliability of the annotations, a subject-matter expert—working in the demining innovation area—annotated 20 sets of data (corresponding to 286 knowledge triples) following the same methodology. Because we are interested in the overlap between the triples extracted by two independent annotators, we quantify the inter-annotator agreement with Jaccard Index, Dice Coefficient and Overlap Coefficient for these 20 sets. The two annotators have on average 0.89, 0.94, and 0.97 agreements respectively. These con-

[2] https://cmac.gov.kh/publications/

[3] https://www.gichd.org/publications-resources/publications/

[4] https://www.gichd.org/our-response/information-management/imsma-core/

sistently high agreement scores indicate that the annotations in the full dataset are of similarly high quality.

4 TextMineX Overview

4.1 Humanitarian Mine Action Task

Knowledge triple extraction from HMA reports can be formally defined as follows: given an ontology $\mathcal{O} = (\mathcal{E}, \mathcal{R})$, where $\mathcal{E}$ is a set of entities and $\mathcal{R}$ is a set of relations, and a textual context C, the objective is to design a prompt $P(C, \mathcal{O})$ that guides an extractor model M to extract triples $T = \{(s_1, r_1, o_1), \cdots, (s_n, r_n, o_n) \mid s, o \in \mathcal{E},\ r \in \mathcal{R}\}$, where n is the number of extracted triples, and subjects/objects (s_i/o_i), and relations (r_i) are extracted from the source text and mapped to $\mathcal{E}$ and $\mathcal{R}$ respectively. The extracted triples must remain consistent with both the ontology and the source text; i.e. $T = M\big(P(C, \mathcal{O})\big)$.

4.2 Overall Pipeline

Figure 3 shows our triple extraction method. In the **Layout-Aware Document Chunking** phase, PDF reports are split into paragraph chunks. These are used with a newly constructed HMA ontology in the **Ontology-Guided Knowledge Extraction** phase. For evaluation, we apply a **Multi-Perspective Evaluation** combining reference-based metrics on our annotated dataset and a reference-free LLM-as-a-Judge approach. All LLM calls use greedy decoding (temperature = 0, top_p = 1.0) to ensure deterministic outputs.

Layout-Aware Document Chunking The input to our pipeline is PDF-formatted demining reports, which contain rich human-readable structures (chapters, sections, tables, lists, and figures). To prepare them for LLM consumption, we segment each document into semantically coherent chunks that preserve context while fitting within typical model context windows (Liu et al., 2024). We leverage Open-Parse's document understanding capabilities (Smock and Pesala, 2021) to identify layout elements and extract paragraph-level segments. On our reports, this yields chunks averaging 127 words (std. 6), which aligns well with both small and large LLM context limits.

Ontology-Guided Knowledge Extraction Given the text chunks as input, our goal is to extract (s, r, o)-triples. First, the extracted triples must be accurate w.r.t. the source text. An ideal extraction system would extract all triples (recall=1) precisely (precision=1). Second, extracted triples must conform to a specified ontology so that extracted demining knowledge can be stored and shared between organizations in a compatible way. We address this by combining a domain-specific HMA ontology with LLM in-context learning to extract triples from paragraphs.

Prompt Templates We design five prompting strategies for knowledge triple extraction: (1) *Zero-shot* (instruction and context only), (2) *One-shot with Random Sentence (RS)*, (3) *One-shot with Random Paragraph (RP)*, (4) *One-shot with Ontology-Aligned Sentence (OS)*, and (5) *One-shot with Ontology-Aligned Paragraph (OP)*. One example prompt is provided in Appendix.

Ontology-aligned demonstrations share the same ontology (entity types and relations) as the target context, while *random* demonstrations use unrelated ontology. Sentences are extracted from paragraphs using NLTK sentence tokenizer[5]. For OS and OP prompts, we design a retrieval algorithm that selects demonstrations from our annotated dataset (§ 3) by identifying the shortest context-answer pair matching the target ontology, for minimizing token costs while retaining high semantic similarity. To prevent data leakage and ensure a fair evaluation, we implement a second retrieval step if the initially-retrieved demonstration contains the same or part of target context as the test instance. In such cases, we select the next shortest matching example instead. This ensures that the retrieved demonstrations do not overlap with the evaluation context, preserving the integrity of the inference.

We design these prompt templates to test two hypotheses: (a) *Ontology alignment enhances accuracy* by priming the model with ontology-specific reasoning. Semantically aligned demonstrations help constrain the label space and improve precision (Min et al., 2022; Long et al., 2024). This is evident in our results, as the contrast between RS/RP and OS/OP confirms the benefit of ontology alignment. (b) *Paragraph-level context improves extraction performance* by providing richer demonstrations that reflect how entities and relations are introduced across sentences in real-world reports. However, our results do not support this hypothesis as comparisons between RS vs. RP and OS vs. OP show no consistent improvement from paragraph-level context.

[5] https://www.nltk.org/howto/stem.html

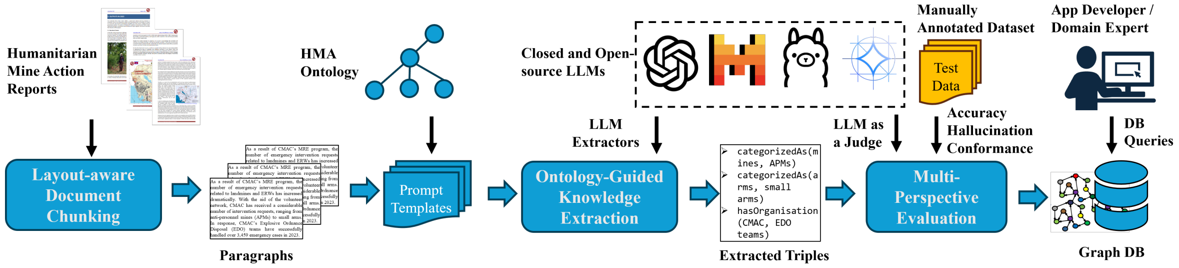

Humanitarian Mine Action Reports
HMA Ontology
Closed and Open-source LLMs
Manually Annotated Dataset
Test Data
App Developer / Domain Expert
Layout-aware Document Chunking
Paragraphs
Prompt Templates
LLM Extractors
Ontology-Guided Knowledge Extraction
categorizedAs(mines, APMs)
categorizedAs(arms, small arms)
hasOrganisation(CMAC, EDO teams)
Extracted Triples
LLM as a Judge
Accuracy Hallucination Conformance
Multi-Perspective Evaluation
DB Queries
Graph DB

Figure 3: **TextMineX Overview.** Reports are preprocessed into paragraph chunks, used as test inputs during inference. Each chunk is combined with an instruction template and ontology, then passed through LLMs for triple extraction. We apply a multi-perspective evaluation using both reference-based and reference-free methods. Extracted triples are stored in a database, queried by developers and HMA domain experts.

4.3 Multi-Perspective Evaluation

Reference-based Evaluation HMA is a high-stakes decision-making domain, where incorrect triples—especially hallucinated ones—can misinform demining operations, leading to inefficiency. A comprehensive evaluation of triple extraction requires assessing accuracy, reliability, and structural validity. We evaluate models across three dimensions: (1) Triple Extraction Accuracy, (2) Hallucination Rate, and (3) Format Conformance.

Triple Extraction Accuracy Triple extraction accuracy serves as the primary metric, as it directly measures how well models extract knowledge triples from text. However, accuracy alone does not fully capture model reliability. The hallucination rate evaluates faithfulness by detecting extraneous or fabricated information, while format conformance ensures that outputs adhere to a syntactically valid structure, enabling seamless integration into downstream applications. We employ N-gram matching-based metrics such as BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and METEOR (Banerjee and Lavie, 2005) to assess the extracted triples. Additionally, we incorporate BERTScore (Zhang et al., 2019), which leverages word embeddings to capture semantic similarity beyond lexical overlap. To enhance the accuracy of our metrics, we first apply stemming and lemmatization by NLTK to normalize morphological variations. We then compute the accuracy metrics against the manually annotated test set.

Hallucination Rate Accuracy measures how well the extracted triples match reference triples, but it does not fully capture whether the generated content is grounded in the input. A model could produce plausible triples that are semantically similar to the original text—thus achieving a high accuracy score—yet still be incorrect; i.e. they do not appear in the input but seem reasonable. Hallucination is a prevalent issue (Ji et al., 2023; Huang et al., 2023; Xu et al., 2024b) and a critical aspect of our evaluation. To quantify this, for each extracted triple $t = (s, r, o)$ we first normalize s, r, and o, as well as the entire input context (tokenization, lemmatization, lowercasing, and punctuation removal). We flag t as a hallucination if the normalized subject s and object o are not found as contiguous substrings in the normalized report text, or if the normalized relation r is absent from the ontology set. This procedure ensures that even "plausible" but unsupported triples are detected and penalized, thereby maintaining faithfulness and trustworthiness.

Format Conformance Format conformance metric assesses whether the generated triples adhere to the correct syntactic format of $r(s, o)$, where r is the relation and s and o denote the subject and object, respectively. We consider a triple well-formatted if it follows this structure. We accommodate edge cases where the subject or object contains numerical values with commas, such as *hasReliabilityInfo(2,500,011 square meters, landmine/ERW affected areas)*, or phrases in parentheses, such as *hasAccidentOrganisationInfo(Quality of Life Survey (QLS), Department of Victim Assistance of CMAA)*. Format conformance ensures that extracted triples follow a structured format necessary for practical use. A model with high accuracy but poor format conformance may fail to produce usable outputs, limiting its applicability in real-world.

Combined Score To unify evaluation metrics into a single representative score, we apply min-max normalization and compute the overall *Com-*

bined Score as:

$$S_{\text{combined}} = \frac{1}{k}\big(S'_{\text{BLEU}} + S'_{\text{ROUGE}} + S'_{\text{METEOR}} + S'_{\text{BERTScore}} + (1 - S'_{\text{Hallucination}})\big) \tag{1}$$

where S' represents the normalized metric values within $[0, 1]$, and the hallucination rate is inverted to penalize more hallucinations. $k = 5$ is the number of metrics included in the score. Format conformance is excluded in the Combined Score, as our experimental results show consistently high format conformance across all extraction models and prompts, making it non-differentiating. *Combined Score* provides a holistic measure of extraction quality while mitigating scale differences among individual metrics.

Reference-Free Evaluation Evaluating generated texts is particularly challenging in domains like HMA, where annotated datasets are scarce. Demining reports are highly technical and domain-specific, requiring extracted triples to align with predefined ontologies of landmine types, clearance operations, and affected areas. Constructing a manually labeled test set is time-consuming and resource-intensive, limiting large-scale reference-based evaluation. Given these constraints, we explore an LLM-as-a-Judge approach as a potential reference-free evaluation framework for evaluating extracted triples. LLM-as-a-Judge offers a potential alternative to evaluation when ground-truth data is limited (Friel and Sanyal, 2023; Saad-Falcon et al., 2023; Es et al., 2023). The ultimate objective of our approach is to find an optimal judge LLM setting where the LLM consistently identifies the best candidate answer and provides a reasoned justification for its decision.

We try to find the optimal LLM Judge setting by conducting systematic ranking experiments and analyzing correlations between the LLMs judged ranking and reference-based rankings. For these ranking experiments, we design Judge Prompts that instruct LLMs on evaluation criteria. We use five models: Mistral-7B (Jiang et al., 2023), Llama3-8B (Grattafiori et al., 2024), Gemma2-9B (Team et al., 2024), LLaMA3-70B and GPT-4o (Hurst et al., 2024), as extraction models. We rank five responses from five models using GPT-4o, Llama3.1-70B, and Llama3.3-70B as our judge models. The the correlation between different judge models and methods are included in Table 2. The judge prompts follow a fixed template with seven placeholders, where the ontology placeholder represents entity and relation types. Formally, let the input set for the LLM judge prompts be $\{O, C, R_{m_1}, R_{m_2}, R_{m_3}, R_{m_4}, R_{m_5}\}$, where O is the ontology set, C is the set of test contexts, and $R_{m_1}, \cdots, R_{m_5}$ are the five sets of candidate answers from the five extractor models. The judge LLM produces a verdict (output as a ranking):

$$V = \text{LLM}\big(\{O, C, R_{m_1}, R_{m_2}, R_{m_3}, R_{m_4}, R_{m_5}\}\big), \tag{2}$$

assigning a rank from best (1) to worst (5) based on predefined instructions and ranking criteria.

To mitigate evaluation biases, we design three judge prompt templates: (1) Basic Judge Prompt, (2) Fair Judge Prompt, and (3) Randomized Fair Judge Prompt. These templates differ in their instructions and ranking methodologies. Fair Judge Prompt enforces explicit reasoning criteria to mitigate position bias, a known issue when LLMs evaluate multiple candidate answers simultaneously (Li et al., 2024; Shi et al., 2024). Randomized Fair Judge Prompt further reduces this bias by randomizing the position of candidate answers, ensuring that response order does not influence rankings.

Once the optimal judge LLM setting is determined, we adopt it as the reference-free evaluator to identify the best answer from each extraction, leveraging its reasoning process. Detailed prompt templates and an example of the reasoning process used for evaluation are provided in Appendix.

5 Experimental Results

We assess the effectiveness of our knowledge triple extraction method through *reference-based* and *reference-free* evaluations. Reference-based evaluation compares extracted triples against our curated dataset, while reference-free evaluation relies on LLM judges to assess generated triples without reference data. A key aspect of our analysis is to examine the correlation between these two evaluation paradigms to evaluate the reliability of LLM-based judgments. In addition, we investigate how different prompt strategies influence extraction performance between models.

Reference-Based Evaluation We employ five LLMs as extractor models: Llama3-70B, GPT-4o, Gemma-9B, Llama-8B and Mistral-7B. Figure 4 illustrates the impact of model selection and prompt strategy on extraction performance. The box plot (top) shows the distribution of *Combined Scores* across models. Llama3-70B achieves the highest overall performance score, closely followed by GPT-4o. Gemma2-9B demonstrates moderate performance, while Llama3-8B and Mistral-7B receive lowest overall scores. The line plot (bottom) highlights prompt effects, with OS and OP yielding the highest scores across most models, supporting the effectiveness of ontology-aligned prompting. These findings reinforce our hypothesis that ontology-aligned prompts enhance extraction accuracy.

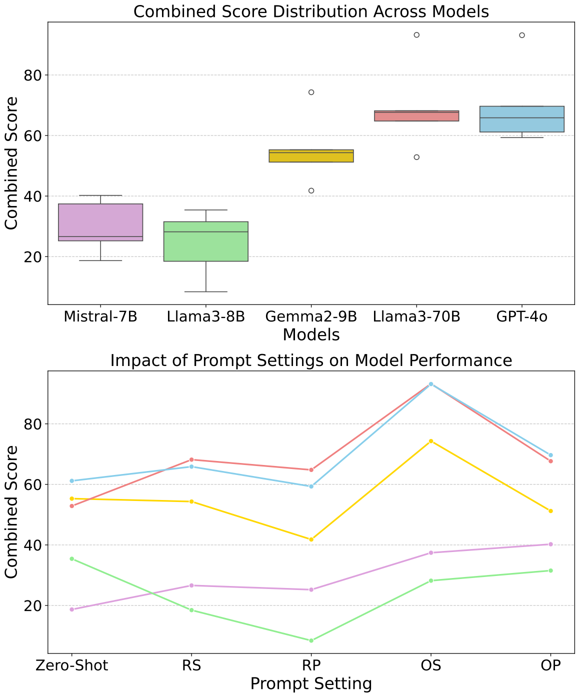


Figure 4: The combined visualization illustrates the impact of model selection and prompt strategy on extraction performance. Abbreviations on x-axis are about One-shot with: RS = Random Sentences; RP = Random Paragraphs; OS = Ontology-Aligned Sentence; OP = Ontology-Aligned Paragraph. The top Combined Score is achieved by Llama3-70B (93.24) closely followed by GPT-4o (93.13), both with OS prompt setting.

Figure 5 further breaks down the performance of five models across four accuracy evaluation metrics: BLEU, ROUGE, METEOR, and BERTScore. OS demonstration prompts consistently result in the best accuracy across all four metrics and all five models, highlighting their effectiveness for the triple extraction task. BLEU scores peak with OS prompts, with GPT-4o achieving the best performance. ROUGE results show a similar trend, with GPT-4o and Llama3-70B excelling in the OS prompt setting. METEOR follows the same pattern as BLEU and ROUGE, reinforcing the advantages of OS prompts. BERTScore, which measures semantic similarity, shows high clustering across models, suggesting minimal differentiation in performance. Overall, OS demonstration prompts consistently enhance extraction accuracy across all models and metrics.

Figure 6 presents the hallucination rates for subjects, relations, and objects across different models and prompt settings. GPT-4o, Llama3-70B, and Gemma2-9B exhibit lower hallucination rates for subjects and objects, while Llama3-8B and Mistral-7B tend to have higher hallucination rates. For high-performing models, the OS prompt type generally helps reduce hallucination, whereas RP prompts tend to increase it. Zero-Shot prompts often lead to increased hallucination for subjects and objects across models. Interestingly, however, Zero-Shot prompts show lower hallucination rates for relations, which may be due to the additional demonstrations in other prompts introducing noise that negatively impacts relation extraction. Most models exhibit high format conformance across all prompt types, with only Gemma-9B under the Zero-Shot prompt scoring below 80%. GPT-4o and Llama3-70B consistently achieve FC above 95% across all prompt types, demonstrating superior adherence to the expected format.

Reference-Free Evaluation Results show that the OS prompt consistently achieves the highest performance across most models, so for the reference-free ranking experiments we only consider five models under OS prompt setting. To assess the alignment between reference-based and reference-free rankings, we compute correlations between the rankings derived from the *Combined Score* calculated based on references, and the rankings derived from the *Expectation Score* based on LLM judges.

Expectation Score As each extractor model received multiple rankings from different judge models, we compute a single *Expectation Score* per extractor model m. This score is defined as:

$$\textit{Expectation Score } E(m) = \frac{\sum_{i=1}^{\mu}\left(i \times P_m(i)\right)}{\sum_{i=1}^{\mu} P_m(i)} \tag{3}$$

where i represents a specific rank, $P_m(i)$ denotes the number of times the model m was assigned rank i. In our case from $i = 1, \cdots, \mu$ and $\mu = 5$ for the five extractor models. E provides a weighted average that reflects the overall tendency of judge models to place an extractor model at a particular rank. The extractor models are ranked such that the one with lowest *Expectation Score* is ranked the

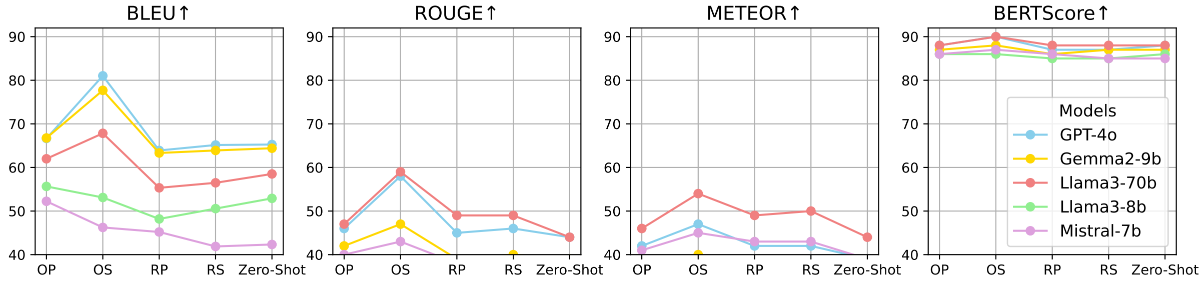


Figure 5: Accuracy metrics scores across prompt types for each model. OS demonstration prompts consistently result in the best accuracy across all four metrics and all five models. *Note: ROUGE, METEOR are scaled by 150, BERTScore by 100 for better visibility.*

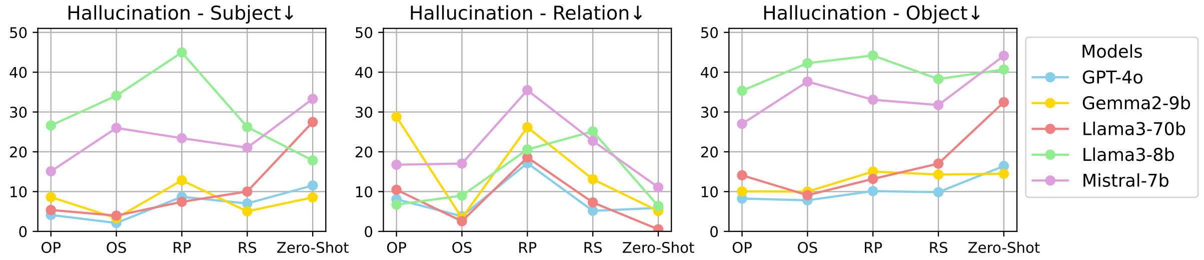


Figure 6: Hallucination rate of subject, relation, and object across prompt types for each model.

highest and vice versa. To systematically analyze the consistency and reliability of the judge methods, we compute the *Expectation Scores* of extractor models separately for Basic Judge, Fair Judge, and Randomized Fair Judge.

Correlation Between LLM Judged and Reference-Based Rankings To assess the reliability of LLM judges, we compute Spearman's correlation (ρ) and Kendall's Tau (τ) to quantify the alignment between LLMs judged rankings and reference-based rankings. ρ measures the monotonic relationship between rankings, where values close to 1 indicate strong agreement. τ evaluates ranking concordance by analyzing the number of concordant and discordant rank pairs, making it particularly useful for detecting minor positional changes. The results of our iterative ranking experiments, shown in Table 2, demonstrate how different judge methods impact ranking alignment.

Our findings reveal that introducing randomization significantly enhances ranking consistency for GPT-4o, improving from $\rho = 0.4$ (Basic) to $\rho = 1.0$ (Randomized). In contrast, Llama3.1-70B shows no improvement across judge methods, indicating persistent positional bias. Llama3.3-70B exhibits weaker alignment overall, with minor improvements under Fair and Randomized judging. These results suggest that while randomization effectively mitigates positional bias for GPT-4o, its impact varies across models.

Table 2: The correlation results for different judge models, judge methods.

Judge Model	Judge Method	Spearman's Correlation	Kendall's Tau
GPT-4o	Basic	0.4	0.4
	Fair	0.9	0.8
	Randomized	1.0	1.0
Llama3.1-70B	Basic	0.4	0.4
	Fair	0.4	0.4
	Randomized	0.4	0.4
Llama3.3-70B	Basic	0.0	-0.2
	Fair	0.3	0.2
	Randomized	0.2	0.2

The ranking experiments identify GPT-4o with Randomized Fair Judge as the optimal judge LLM setting for our evaluation task. We further apply this setting to identify the optimal triples from each extraction output using GPT-4o with the Randomized Fair Judge method, aggregating the top-ranked triples across all extractions to generate the final output. We compare these aggregated results against the test dataset, yielding a *Combined Score* of 83.93. This achieves 90% of the current best score 93.24 (see figure 4), reinforcing the effectiveness of our reference-free LLM-as-a-Judge paradigm as a viable alternative to conventional reference-based evaluation methods. Future research could refine this approach by exploring additional LLMs and judge strategies to approximate a even better judge LLM setting for knowledge triple extraction tasks.

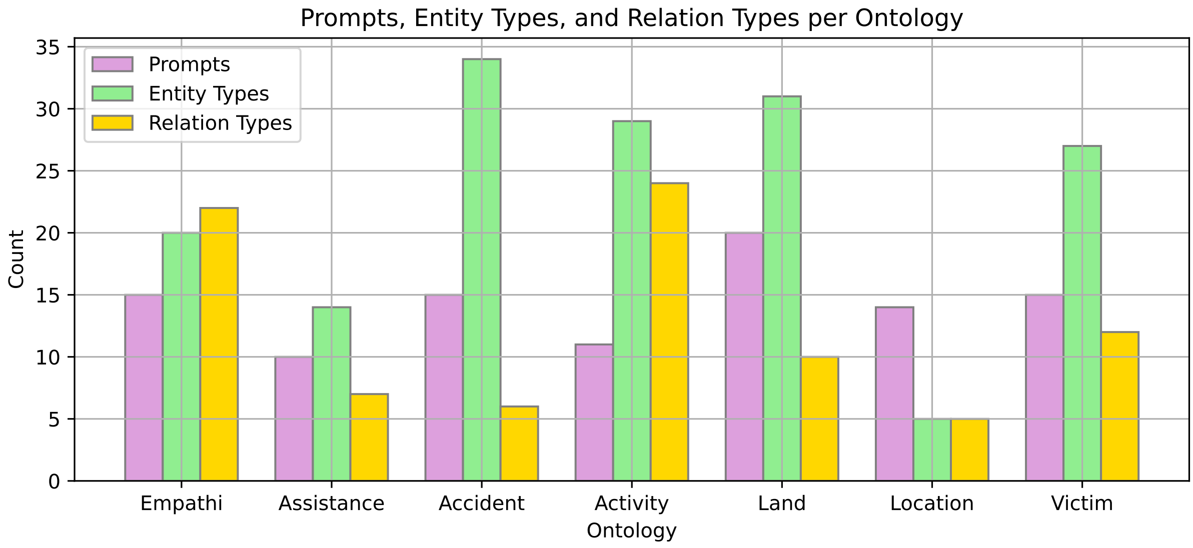


Figure 7: Number of prompts, entity types, and relation types per the utiized ontologies from IMSMA Core and Empathi.

5.1 Additional Information and Insights for the HMA Ontology

We visualize the statistics regarding the combined ontologies. Fig. 7 includes the number of prompts, entity types, and relation types per ontology.

The Empathi (Gaur et al., 2019) ontology is designed for "Emergency Managing and Planning about Hazard Crises". It is the most advanced ontology that is openly-available in the research domain for hazard crises, as an outcome of a research study. However, this ontology is designed for the broader concept of disasters as opposed to being specialized in the HMA domain. For instance, it has many concepts such as hurricanes or firefighters that are outside of the scope of the demining-relevant disasters such as mine explosions. The concepts (entity types and relation types) that are relevant to the demining are manually filtered and used as the ontology. Thus, the number of entities and relations are after the filtering process. GICHD's new data schema IMSMA Core (previous standard was called IMSMA-NG (Desantis and Eriksson, 2013), where NG stands for "Next Generation") is the agreed standard between mine agencies of governments and non-governmental organizations (NGOs). IMSMA Core serves the data model for databases of mine agencies globally. On the other hand, the number of entities and relation types are limited.

We consider the separate portions of the GICHD data model, which are all related to the mine action but logically separated as *"Assistance, Accident, Activity, Land, Location, and Victim"* ontologies. We utilize both as the best state-of-the-art models and convert them into relatively small ontologies. On the other hand, we utilize the concepts from all of the 7 ontologies in our prompt generation process for LLMs, resulting in different number of prompts per ontology. The proposed "HMA Ontology" can be extended with more special entity and relation types based on special needs of the organizations.

6 Conclusion

TextMineX addresses the need for knowledge extraction from HMA reports by LLMs and domain ontologies to transform unstructured technical reports into structured knowledge triples. This paper introduces the curated data of HMA reports, ontology, and LLM pipeline, aiming to enable innovations for mine action and natural language processing.

Limitations

While TextMineX provides the first dataset, evaluation framework and pipeline that demonstrates the feasibility of an LLM-driven extraction pipeline in the specialized domain of humanitarian demining, several limitations are considered to be addressed in the future, including further extension of the dataset. Our evaluation relies on 100 prompt–response examples (yielding 1,095 unique triples).

Assembling and annotating demining data demand extensive domain expertise, so even this modest set offers valuable proof-of-concept insights, but we plan to extend to larger, multilingual collections in future work. Second, this work focuses on reports that are written only in English. In the future, we plan to extend the work to include other languages, because mine-action data are often produced in multiple languages, such as Arabic, French or Ukrainian, depending on the conflict region. Making TextMineX multilingual is, therefore, needed to ensure more inclusiveness.

In this paper, we focus on humanitarian demining and actively collaborate with demining agencies. Thus, the scope of the evaluation framework and pipeline is limited to the demining domain, whereas the applicability of the pipeline in similar domains (e.g, natural disaster management) could be tested as a potential future work.

Ethical Considerations

Humanitarian demining is a high importance decision making domain where incorrect or hallucinated triples can misinform planning and lead to wasted time and resources. To address this, TextMineX combines reference based validation with a bias aware LLM as Judge framework and publishes all prompts and decoding settings for full transparency. We also engage landmine clearance experts throughout development to review and validate outputs. TextMineX is not used to locate mines, as safety remains governed by established GICHD standards (GICHD Mine Action Standards, https://www.gichd.org/our-response/mine-action-standards/), and instead supports expert analysis and planning. By documenting our methods and keeping an expert in the loop, we aim to minimize misinformation and ensure responsible AI deployment in demining operations.

The dataset used in this study consists of publicly available or institutionally provided humanitarian demining reports. These reports were reviewed to ensure they do not contain personally identifiable information (PII) or offensive content. Our usage of the data adheres to privacy standards and is strictly confined to research contexts.

All datasets used in this study were accessed under conditions permitting research use. The curated demining report dataset we constructed is intended solely for academic and research purposes and complies with the original access and licensing conditions. The ontology and pipeline components developed in TextMineX are likewise designed for research and evaluation within humanitarian domains. We do not support or promote deployment of these artifacts in operational or commercial contexts without further validation and ethical review.

AI usage: AI assistance tool is utilized for grammar/spelling check of the paper.

References

Vahan Arsenyan, Spartak Bughdaryan, Fadi Shaya, Kent Small, and Davit Shahnazaryan. 2023. Large language models for biomedical knowledge graph construction: information extraction from emr notes. *arXiv preprint arXiv:2301.12473*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, and 1 others. 2024. Llms instead of human judges? a large scale empirical study across 20 nlp evaluation tasks. *arXiv preprint arXiv:2406.18403*.

Sangnie Bhardwaj, Samarth Aggarwal, and Mausam Mausam. 2019. Carb: A crowdsourced benchmark for open ie. In *Conference on Empirical Methods in Natural Language Processing*.

Iva Bojic, Jessica Chen, Si Yuan Chang, Qi Chwen Ong, Shafiq Joty, and Josip Car. 2023. Hierarchical evaluation framework: Best practices for human evaluation. *arXiv preprint arXiv:2310.01917*.

Samuel Broscheit, Kiril Gashteovski, and Martin Achenbach. 2017. Openie for slot filling at tac kbp 2017-system description. In *TAC*.

Samuel Broscheit, Kiril Gashteovski, Yanjie Wang, and Rainer Gemulla. 2020. Can we predict new facts with open knowledge graph embeddings? a benchmark for open link prediction. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Valentina Anita Carriero, Antonia Azzini, Ilaria Baroni, Mario Scrocca, and Irene Celino. 2024. Human evaluation of procedural knowledge graph extraction from text with large language models. *ArXiv*, abs/2412.03589.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *arXiv preprint arXiv:2307.01128*.

Zeno Cauter and Nikolay Yakovets. 2024. Ontology-guided knowledge graph construction from maintenance short texts. In *Proceedings of the 1st Workshop on Knowledge Graphs and Large Language Models (KaLLM 2024)*, pages 75–84.

Hanzhu Chen, Xu Shen, Qitan Lv, Jie Wang, Xiaoqi Ni, and Jieping Ye. 2024. Sac-kg: Exploiting large language models as skilled automatic constructors for domain knowledge graphs. *arXiv preprint arXiv:2410.02811*.

Jerry Junyang Cheung, Yuchen Zhuang, Yinghao Li, Pranav Shetty, Wantian Zhao, Sanjeev Grampurohit, Rampi Ramprasad, and Chao Zhang. 2024. Polyie: A dataset of information extraction from polymer material scientific literature. *arXiv preprint arXiv:2311.07715*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and 1 others. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Gaye Colakoglu, Gürkan Solmaz, and Jonathan Fürst. 2025. Problem solved? information extraction design space for layout-rich documents using llms. In *Findings of the EMNLP'25*.

John Dagdelen, Alexander Dunn, Sanghoon Lee, Nicholas Walker, Andrew S Rosen, Gerbrand Ceder, Kristin A Persson, and Anubhav Jain. 2024. Structured information extraction from scientific text with large language models. *Nature communications*, 15(1):1418.

Angela Desantis and Daniel Eriksson. 2013. The new imsma and victim assistance. *The Journal of ERW and Mine Action*, (8).

Daniel Deutsch, Rotem Dror, and Dan Roth. 2022. On the limitations of reference-free evaluations of generated text. *arXiv preprint arXiv:2210.12563*.

Jacob Eisenstein. 2019. *Introduction to natural language processing*. MIT press.

Shahul Es, Jithin James, Luis Espinosa-Anke, and Steven Schockaert. 2023. Ragas: Automated evaluation of retrieval augmented generation. *arXiv preprint arXiv:2309.15217*.

Saeed Farzi, Soumitra Ghosh, Alberto Lavelli, and Bernardo Magnini. 2024. Get the best out of 1b llms: insights from information extraction on clinical documents. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 266–276.

Selim Fekih, Nicolò Tamagnone, Benjamin Minixhofer, Ranjan Shrestha, Ximena Contla, Ewan Oglethorpe, and Navid Rekabsaz. 2022. Humset: Dataset of multilingual information extraction and classification for humanitarian crisis response. *arXiv preprint arXiv:2210.04573*.

Niklas Friedrich, Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2022. AnnIE: An annotation platform for constructing complete open information extraction benchmark. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 44–60, Dublin, Ireland. Association for Computational Linguistics.

Robert Friel and Atindriyo Sanyal. 2023. Chainpoll: A high efficacy method for llm hallucination detection. *arXiv preprint arXiv:2310.18344*.

Jinlan Fu, Pengfei Liu, and Graham Neubig. 2020. Interpretable multi-dataset evaluation for named entity recognition. *arXiv preprint arXiv:2011.06854*.

Kiril Gashteovski, Rainer Gemulla, Bhushan Kotnis, Sven Hertling, and Christian Meilicke. 2020. On aligning openie extractions with knowledge bases: A case study. In *EVAL4NLP*.

Kiril Gashteovski, Sebastian Wanner, Sven Hertling, Samuel Broscheit, and Rainer Gemulla. 2019. Opiec: an open information extraction corpus. *arXiv preprint arXiv:1904.12324*.

Kiril Gashteovski, Mingying Yu, Bhushan Kotnis, Carolin Lawrence, Mathias Niepert, and Goran Glavaš. 2022. BenchIE: A framework for multi-faceted fact-based open information extraction evaluation. In *Proc. of the ACL'22*, pages 4472–4490. Association for Computational Linguistics.

Julia Gastinger, Timo Sztyler, Lokesh Sharma, Anett Schuelke, and Heiner Stuckenschmidt. 2023. Comparing apples and oranges? on the evaluation of methods for temporal knowledge graph forecasting. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 533–549. Springer.

Manas Gaur, Saeedeh Shekarpour, Amelie Gyrard, and Amit Sheth. 2019. Empathi: An ontology for emergency managing and planning about hazard crisis. In *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, pages 396–403.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Hassan Hamad, Abhinav Kumar Thakur, Nijil Kolleri, Sujith Pulikodan, and Keith Chugg. 2024. Fire: A dataset for financial relation extraction. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3628–3642.

Xiaoqi Han, Ru Li, Hongye Tan, Wang Yuanlong, Qinghua Chai, and Jeff Pan. 2023. Improving sequential model editing with fact retrieval. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 11209–11224.

Shibo Hao, Bowen Tan, Kaiwen Tang, Hengzhe Zhang, Eric P Xing, and Zhiting Hu. 2022. Bertnet: Harvesting knowledge graphs from pretrained language models. *arXiv preprint arXiv:2206.14268*.

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2025. Enabling LLM knowledge analysis via extensive materialization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16189–16202.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Liangyi Huang and Xusheng Xiao. 2024. Ctikg: Llm-powered knowledge graph construction from cyber threat intelligence. In *First Conference on Language Modeling*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Humanity & Inclusion. 2023. Landmine monitor 2023: Current conflicts & long-lasting contamination cause high number of mine casualties. Accessed: 29-Jan-2025.

Sameer Jain, Vaishakh Keshava, Swarnashree Mysore Sathyendra, Patrick Fernandes, Pengfei Liu, Graham Neubig, and Chunting Zhou. 2023. Multi-dimensional evaluation of text summarization with in-context learning. *arXiv preprint arXiv:2306.01200*.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *arXiv*, arXiv:2310.06825. [cs.CL].

Martin Josifoski, Nicola De Cao, Maxime Peyrard, Fabio Petroni, and Robert West. 2022. GenIE: Generative information extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4626–4643, Seattle, United States. Association for Computational Linguistics.

Seungone Kim, Juyoung Suk, Xiang Yue, Vijay Viswanathan, Seongyun Lee, Yizhong Wang, Kiril Gashteovski, Carolin Lawrence, Sean Welleck, and Graham Neubig. 2025. Evaluating language models as synthetic data generators. In *Association for Computational Linguistics (ACL)*.

Bhushan Kotnis, Kiril Gashteovski, Julia Gastinger, Giuseppe Serra, Francesco Alesiani, Timo Sztyler, Ammar Shaker, Na Gong, Carolin Lawrence, and Zhao Xu. 2022a. Human-centric research for nlp: Towards a definition and guiding questions. *arXiv preprint arXiv:2207.04447*.

Bhushan Kotnis, Kiril Gashteovski, Daniel Rubio, Ammar Shaker, Vanesa Rodriguez-Tembras, Makoto Takamoto, Mathias Niepert, and Carolin Lawrence. 2022b. MILIE: Modular & iterative multilingual open information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 6939–6950.

Landmine and Cluster Munition Monitor. 2023. Cambodia: Mine action. In *Landmine Monitor 2023*. International Campaign to Ban Landmines - Cluster Munition Coalition (ICBL-CMC).

Anne Lauscher, Yide Song, and Kiril Gashteovski. 2019. Minscie: citation-centered open information extraction. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 386–387. IEEE.

Xiang Li, Penglei Sun, Wanyun Zhou, Zikai Wei, Yongqi Zhang, and Xiaowen Chu. 2025. Finkario: Event-enhanced automated construction of financial knowledge graph. *arXiv preprint arXiv:2508.00961*.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024. Split and merge: Aligning position biases in llm-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Jixiong Liu, Yoan Chabot, Raphaël Troncy, Viet-Phi Huynh, Thomas Labbé, and Pierre Monnin. 2023a. From tabular data to knowledge graphs: A survey of semantic table interpretation tasks and methods. *Journal of Web Semantics*, 76:100761.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Pengfei Liu, Jinlan Fu, Yang Xiao, Weizhe Yuan, Shuaicheng Chang, Junqi Dai, Yixin Liu, Zihuiwen Ye, Zi-Yi Dou, and Graham Neubig. 2021. Explainaboard: An explainable leaderboard for nlp. *arXiv preprint arXiv:2104.06387*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023b. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9):1–35.

Runxuan Liu, Bei Luo, Jiaqi Li, Baoxin Wang, Ming Liu, Dayong Wu, Shijin Wang, and Bing Qin. 2025. Ontology-guided reverse thinking makes large language models stronger on knowledge graph question answering. *arXiv preprint arXiv:2502.11491*.

Quanyu Long, Yin Wu, Wenya Wang, and Sinno Jialin Pan. 2024. Does in-context learning really learn? rethinking how large language models respond and solve tasks via in-context learning. In *First Conference on Language Modeling*.

Christian Meilicke, Manuel Fink, Yanjie Wang, Daniel Ruffinelli, Rainer Gemulla, and Heiner Stuckenschmidt. 2018. Fine-grained evaluation of rule-and embedding-based systems for knowledge graph completion. In *International semantic web conference*, pages 3–20. Springer.

Nandana Mihindukulasooriya, Sanju Tiwari, Carlos F Enguix, and Kusum Lata. 2023. Text2kgbench: A benchmark for ontology-driven knowledge graph generation from text. In *International Semantic Web Conference*, pages 247–265. Springer.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2018. Rankme: Reliable human ratings for natural language generation. *arXiv preprint arXiv:1803.05928*.

Juri Opitz. 2024. A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *transactions of the association for computational linguistics*, 12:820–836.

Huitong Pan, Qi Zhang, Mustapha Adamu, Eduard Dragut, and Longin Jan Latecki. 2025. Taxonomy-driven knowledge graph construction for domain-specific scientific applications. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4295–4320.

Jeff Z Pan, Simon Razniewski, Jan-Christoph Kalo, Sneha Singhania, Jiaoyan Chen, Stefan Dietze, Hajira Jabeen, Janna Omeliyanenko, Wen Zhang, Matteo Lissandrini, and 1 others. 2023. Large language models and knowledge graphs: Opportunities and challenges. *arXiv preprint arXiv:2308.06374*.

Chaoxu Pang, Yixuan Cao, Chunhao Yang, and Ping Luo. 2024. Uncovering limitations of large language models in information seeking from tables. *arXiv preprint arXiv:2406.04113*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Kevin Pei, Ishan Jindal, and Kevin Chang. 2023. Abstractive open information extraction. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6146–6158.

Maxime Peyrard, Wei Zhao, Steffen Eger, and Robert West. 2021. Better than average: Paired evaluation of nlp systems. *arXiv preprint arXiv:2110.10746*.

Fabio Poletto, Yunbai Zhang, André Panisson, Yelena Mejova, Daniela Paolotti, and Sylvain Ponserre. 2021. Developing annotated resources for internal displacement monitoring. In *Companion Proceedings of the Web Conference 2021*, pages 136–144.

Gorjan Radevski, Kiril Gashteovski, Chia-Chien Hung, Carolin Lawrence, and Goran Glavaš. 2023. Linking surface facts to large-scale knowledge graphs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7189–7207.

Gorjan Radevski, Kiril Gashteovski, Shahbaz Syed, Christopher Malon, Sebastien Nicolas, Chia-Chien Hung, Timo Sztyler, Verena Heußer, Wiem Ben Rim, Masafumi Enomoto, and 1 others. 2025. On synthesizing data for context attribution in question answering. *arXiv preprint arXiv:2504.05317*.

André Gomes Regino and Julio Cesar Dos Reis. 2025. Can llms be knowledge graph curators for validating triple insertions? In *Proceedings of the workshop on generative AI and knowledge graphs (GenAIK)*, pages 87–99.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. *arXiv preprint arXiv:2005.04118*.

Wiem Ben Rim, Carolin Lawrence, Kiril Gashteovski, Mathias Niepert, and Naoaki Okazaki. 2021. Behavioral testing of knowledge graph embedding models for link prediction. In *3rd Conference on Automated Knowledge Base Construction*.

Pablo Romero, Lifeng Han, and Goran Nenadic. 2025. Insightbuddy-ai: Medication extraction and entity linking using pre-trained language models and ensemble learning. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 18–27.

Jon Saad-Falcon, Omar Khattab, Christopher Potts, and Matei Zaharia. 2023. Ares: An automated evaluation framework for retrieval-augmented generation systems. *arXiv preprint arXiv:2311.09476*.

Oscar Sainz, Iker García-Ferrero, Rodrigo Agerri, Oier López de Lacalle, Germán Rigau, and Eneko Agirre. 2023. Gollie: Annotation guidelines improve zero-shot information-extraction. *ArXiv*, abs/2310.03668.

Hendrik Schuff, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. 2023. How to do human evaluation: A brief introduction to user studies in nlp. *Natural Language Engineering*, 29(5):1199–1222.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.

Brandon Smock and Rohith Pesala. 2021. Table Transformer.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Craig Thomson, Ehud Reiter, and Anya Belz. 2024. Common flaws in running human evaluation experiments in nlp. *Computational Linguistics*, 50(2):795–805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

United Nations. 2025. International mine awareness day.

United Nations Development Programme in Cambodia. 2021. Clearing for results - mine action for human development: Annual report 2020. Technical report, UNDP Cambodia.

Chris van der Lee, Albert Gatt, Emiel van Miltenburg, and Emiel J. Krahmer. 2021. Human evaluation of automatically generated text: Current trends and best practice guidelines. *Comput. Speech Lang.*, 67:101151.

Peter Vickers, Loïc Barrault, Emilio Monti, and Nikolaos Aletras. 2024. We need to talk about classification evaluation metrics in nlp. *arXiv preprint arXiv:2401.03831*.

Gerhard Weikum, Xin Luna Dong, Simon Razniewski, Fabian Suchanek, and 1 others. 2021. Machine knowledge: Creation and curation of comprehensive knowledge bases. *Foundations and Trends® in Databases*, 10(2-4):108–490.

Haris Widjaja, Kiril Gashteovski, Wiem Ben Rim, Pengfei Liu, Christopher Malon, Daniel Ruffinelli, Carolin Lawrence, and Graham Neubig. 2022. Kgxboard: Explainable and interactive leaderboard for evaluation of knowledge graph completion models. *arXiv preprint arXiv:2208.11024*.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, and 1 others. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Guanming Xiong, Junwei Bao, and Wen Zhao. 2024. Interactive-kbqa: Multi-turn interactions for knowledge base question answering with large language models. *arXiv preprint arXiv:2402.15131*.

Zhao Xu, Wiem Ben Rim, Kiril Gashteovski, Timo Sztyler, and Carolin Lawrence. 2024a. A human-centric evaluation platform for explainable knowledge graph completion. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024b. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Bowen Zhang and Harold Soh. 2024. Extract, define, canonicalize: An llm-based framework for knowledge graph construction. *arXiv preprint arXiv:2404.03868*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Ying Zhou, Xuanang Chen, Ben He, Zheng Ye, and Le Sun. 2022. Re-thinking knowledge graph completion evaluation from an information retrieval perspective. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*.

Yuqi Zhu, Xiaohan Wang, Jing Chen, Shuofei Qiao, Yixin Ou, Yunzhi Yao, Shumin Deng, Huajun Chen, and Ningyu Zhang. 2023. Llms for knowledge graph construction and reasoning: Recent capabilities and future opportunities. *arXiv preprint arXiv:2305.13168*.

A Detailed Discussion of Related Work

A.1 Automated Knowledge Graph Construction

Constructing knowledge graphs (KGs) from text is a well-established problem (Weikum et al., 2021). The methods for constructing KGs are either schema-free (Pei et al., 2023; Kotnis et al., 2022b; Gashteovski et al., 2019) or are fixed either to general-purpose schemas (Zhang and Soh, 2024; Josifoski et al., 2022; Broscheit et al., 2017) or to domain-specific schemas (Dagdelen et al., 2024; Lauscher et al., 2019; Broscheit et al., 2017). Both the schema-free methods and the methods that are tied to general-purpose schemas do not address the complexities of the domain at hand, which is why we worked on a specific data, benchmarks and methods for the HMA domain.

A.2 LLMs for Knowledge Graph Construction

Prior work has explored the use of LLMs for constructing KGs from text (Pan et al., 2023). One line of work focuses on harvesting knowledge from pretrained LLMs without the use of annotated data via zero-shot iterative prompting (Hu et al., 2025; Carta et al., 2023; Hao et al., 2022) or in-context learning (ICL) (Brown et al., 2020; Min et al., 2022; Liu et al., 2023a). However, these methods are not well-suited for the construction of domain-specific knowledge graphs.

For these reasons, there have been methods for constructing domain-specific KGs by the use of LLMs (Chen et al., 2024). For example, there are LLM methods for constructing KGs in various domains, such as medicine (Arsenyan et al., 2023), finance (Li et al., 2025), cyber-thread intelligence (Huang and Xiao, 2024) or climate science (Pan et al., 2025). To the best of our knowledge, however, there is no prior work on KG construction for the humanitarian mine action domain.

A.3 Ontology-Guided for Knowledge Graph Construction

Ontology-guided (a.k.a. schema-driven) methods have been shown to be effective, though they are often limited to single-sentence input and, consequently, overlook context-level reasoning (Cauter and Yakovets, 2024). Text2KGBench (Mihindukulasooriya et al., 2023) is a benchmark and evaluation framework for extracting triples from text, which are bound to a predefined ontology. However, it relies on toy data schemas, which limits their applicability to the complex, domain-specific HMA documents.

Other line of work focuses on reasoning and validation within already structured knowledge graphs, not extraction from free-form text as with TextMineX (Liu et al., 2025; Regino and Dos Reis, 2025). Their ontological constraints apply post-hoc to existing triples, whereas TextMineX uses the ontology during extraction to guide triple formation and reduce hallucinations.

Finally, none of these works address bias-aware evaluation: they rely on deterministic metrics or small-scale manual validation. TextMineX introduces a bias-aware LLM-as-Judge evaluation framework that explicitly mitigates positional bias in reference-free scoring—an issue absent from prior ontology-guided pipelines.

A.4 Multi-Faceted Knowledge Graph Evaluation

They are typically evaluated with a single score metric, such as accuracy (Eisenstein, 2019). In recent years, however, the NLP community has increasingly been pointing out that such evaluation approach can obscure the model performance nuances, thus revealing misleading results (Ribeiro et al., 2020).

To mitigate these effects, researchers have turned to multiple metric evaluations, which reveal more complete picture about the performance of the models (Vickers et al., 2024; Opitz, 2024; Jain et al., 2023; Deutsch et al., 2022; Peyrard et al., 2021; Liu et al., 2021). Particularly within the field of information extraction, such multi-faceted evaluations are already well established (Radevski et al., 2023; Friedrich et al., 2022; Gashteovski et al., 2022; Fu et al., 2020). Likewise, KG-related work also employs such multi-faceted evaluations, such as for the task of link prediction (Gastinger et al., 2023; Widjaja et al., 2022; Rim et al., 2021; Meilicke et al., 2018).

In order to evaluate aspects of the models that are important to users or that need immediate human verifications, another line of work have used manual evaluations (Bojic et al., 2023; Novikova et al., 2018). This could be done with user studies that target the specific aspects that one tries to measure

(Radevski et al., 2025; Schuff et al., 2023; Kotnis et al., 2022a) or by simply manually validating the final results (Thomson et al., 2024; van der Lee et al., 2021). Such manual evaluations have been practiced for both information extraction research works (Sainz et al., 2023; Gashteovski et al., 2020; Bhardwaj et al., 2019), as well as for KG prediction tasks, such as link prediction (Carriero et al., 2024; Xu et al., 2024a; Zhou et al., 2022). Manual evaluations, however, are both expensive and not scalable, thus making them practically infeasible.

To avoid the pitfalls of the narrow single-score metrics, as well as the scalability issues of manual evaluations, we propose a multi-faceted evaluation framework that incorporates several aspects of the information extraction problem, fused into the final score.

B In-Context Learning Prompt Example

The below is an example of one-shot prompt with the RS prompt setting. All in-context learning prompts are stored in CSV files as part of the supplementary material for easier reproducibility.

One-shot with Random Sentence Demonstration (RS)

Instruction:
Extract and list only the triples from the following sentence based on the specified entity types and relation types. Do not include any explanatory or intermediate text in your output. In the output, only include the triples in the given output format: relation(subject, object). Attempt to extract as many entities and relations as you can.
Entity Types:
AdministrativeArea, Association, Location, Organisation, MedicalFacility
Relation Types:
hasAdministrativeArea, hasAssociation, hasLocation, hasOrganisation, locatedNear
Example:
Sentence:
The accidental detonation of old wartime munitions causes significant infrastructure damage to the nearby village roads and buildings.
Output:
CausedBy(infrastructure damage, old wartime munitions)
Context: On Thursday, March 16, 2023, at CMAC Headquarters in Phnom Penh, Delegate of the Royal Government in charge as Director General of CMAC, met with a delegation from the Japan International Cooperation Agency (JICA) General Director of Governance and Peacebuilding Department. During the meeting, the JICA side briefed on the results of its cooperation with CMAC, in particular training for Ukraine with good results.

C LLM Judge Prompts

Basic Judge Prompt

Instruction:
You are a judge who ranks five models from 1 to 5 on a triple extraction task. You must assign 1 to the model with the best answer and 5 to the model with the worst answer. Your ranking should be provided directly in this format: [1: model x; 2: model x; 3: model x; 4: model x; 5: model x].
Ranking Criteria:
Correctness:
The triples must conform to the format relation(subject, object) and must accurately reflect relationships stated in the context. Models with significant formatting errors should be penalized.
Coverage:
The number of correct triples extracted. More accurate triples are better, but avoid penalizing slight redundancies unless they detract from the overall relevance.
Relevance:
The triples must be relevant to the specified entity and relation types and should align well with the specific context provided.
Edge Cases:
If a model extracts many triples but includes incorrect or redundant ones, balance accuracy and redundancy in your ranking. Correctness should be prioritized, followed by Relevance, then Coverage.
Entity Types: {entity_types}
Relation Types: {relation_types}
Context: {Context}
Model Outputs: {model 1 output} {model 2 output} {model 3 output} {model 4 output} {model 5 output}
Your ranking:

The below are two example prompts used during the experimental study: 1) Basic judge prompt, and 2) (Randomized) Fair Judge Prompt. These methods are explained in Sec. 4. The latter prompt example below includes both cases of regular and randomized fair judge prompts at once. The only difference is the shuffling of the positions of candidate answers in "Model Outputs" part of the prompt. All LLM judge prompts are stored in CSV files as part of the supplementary material. The prompts can be used for reproducing as well as applying in different datasets (without additional annotation efforts).

(Randomized) Fair Judge Prompt

Instruction:
You are a judge tasked with evaluating and ranking five models based on their performance in a **triple extraction task**. Your role is to ensure **fairness, impartiality, and accuracy** by independently evaluating each model's output without any positional bias. Do not assume that the first model is better or worse simply because of its position—all models must be treated equally.

Evaluation Guidelines:

1. **Independence of Evaluation**:
   Evaluate each model **independently** without comparing it to others until all models are scored. Avoid assumptions based on position or order in the list.
2. **Evaluation Criteria:**

(a) **Correctness of Triples (Highest Priority):**

- Triples must strictly conform to the format `relation(subject, object)`.
- Relationships must match the **Given Relation Types** provided below.
- Triples containing fabricated or hallucinated relationships must result in a significant penalty.

(b) **Relevance:**

- Triples must accurately reflect relationships mentioned in the **Context**.
- Irrelevant triples or hallucinations must receive a lower score.

(c) **Coverage:**

- The number of correct triples extracted. Higher coverage is better **only** if triples meet correctness and relevance criteria.

3. **Ranking Process:**

- **Step 1:** Independently evaluate each model's output and assign scores (from 1 to 10) for each criterion: Correctness, Relevance, and Coverage. Summarize the total score for each model.
- **Step 2:** Rank all five models from 1 (best) to 5 (worst) based solely on their total scores.
- Break ties by prioritizing **Correctness** first, then **Relevance**, and finally **Coverage**.

Important Note:
Treat all models equally regardless of position. Your goal is to be an impartial judge. Avoid positional bias and ensure the evaluation reflects only the quality of the model outputs.

Output Format:
Provide your output strictly in this format:
Evaluation:
Model 1: Correctness = X, Relevance = X, Coverage = X, Total = Y
Model 2: Correctness = X, Relevance = X, Coverage = X, Total = Y
Model 3: Correctness = X, Relevance = X, Coverage = X, Total = Y
Model 4: Correctness = X, Relevance = X, Coverage = X, Total = Y
Model 5: Correctness = X, Relevance = X, Coverage = X, Total = Y
Ranking: [1: Model X; 2: Model X; 3: Model X; 4: Model X; 5: Model X]
Entity Types: {entity_types}
Relation Types: {relation_types}
Context: {Context}
Model Outputs: {model 1 output} {model 2 output} {model 3 output} {model 4 output} {model 5 output}
(For Randomized Fair Judge Prompts, the positions of candidate answers in Model Outputs are randomly shuffled.)
Your evaluation and ranking:

D Evaluation Reasoning Process

Output of GPT-4o with Randomized Judge Prompt for Extraction Prompt 89

Model's Answer for Prompt 89

Evaluation

Model 1

- **Correctness** = 4
  - Many relations are either incorrect or irrelevant to the context (e.g., Monitoring, Advises, Coordinates), leading to significant penalties.
- **Relevance** = 5
  - The model produces correct triples but includes several irrelevant triples not related to the context (e.g., Physical-Health).
- **Coverage** = 7
  - High number of triples extracted, but this value is diminished by the correctness and relevance issues.
- **Total** = 4 + 5 + 7 = 16

Model 2

- **Correctness** = 6
  - Includes both relevant and some fabricated or unclear relations (e.g., Triggers(InfrastructureDamage, old wartime munitions)).
- **Relevance** = 6
  - Many triples are relevant, though some irrelevant or vague triples are present.
- **Coverage** = 7
  - Covers a broad range of the context accurately but is diminished slightly by prioritizing the correctness and relevance.
- **Total** = 6 + 6 + 7 = 19

Model 3

- **Correctness** = 7
  - Most triples are accurate and correctly formed, with fewer fabricated relations.
- **Relevance** = 7
  - Majority of the triples are highly relevant to the context. Some triples are slightly incoherent but mostly pertinent.
- **Coverage** = 6
  - Good coverage of the context, though a few needed extractions were missing.
- **Total** = 7 + 7 + 6 = 20

Model 4

- **Correctness** = 6
  - Contains some inaccuracies but generally provides correctly formulated triples.
- **Relevance** = 6
  - Various triples are relevant to the context, but some are less directly related.
- **Coverage** = 5
  - Moderate coverage but some key elements are missed or less comprehensive.
- **Total** = 6 + 6 + 5 = 17

Model 5

- **Correctness** = 8
  - High accuracy in triple formation with minimal fabrications.
- **Relevance** = 8
  - Highly relevant triples, capturing the context well without introducing unrelated information.
- **Coverage** = 8
  - Extensive and accurate coverage of the context, handling the core relationships well.
- **Total** = 8 + 8 + 8 = 24

Ranking

1. **Model 5**
2. **Model 3**
3. **Model 2**
4. **Model 4**
5. **Model 1**