MUC-4 Revisited: Document-level Event Analysis Beyond Span-based Arguments

Helene Bøsei Olsen, Erik Velldal, Lilja Øvrelid
University of Oslo
{helenbol, erikve, liljao}@uio.no

Abstract

Automatically predicting structured representations of events has long been a central goal in information extraction, yet most contemporary work remains limited to identifying contiguous text spans as event arguments. This span-centric formulation fails to capture higher-level aspects of real-world events, such as actor identities, temporal scope, and aggregated outcomes, that many event-centred applications depend on. While commonly treated as a standard extractive benchmark, MUC-4 originally combined span-based arguments with normalised, inferred, and categorical fields, reflecting a richer, application-driven design. In this paper, we revisit MUC-4 in its full original formulation, casting it as an abstractive event analysis task that connects traditional event extraction goals with modern generative and document-level paradigms. We provide the first systematic evaluation of fine-tuned generative models in this extended formulation on MUC-4, examining how post-training stages and model size affect performance across both span-based and higher-level, semantically grounded event information. An extensive error analysis highlights practical challenges and directions for future work.

Keywords: Event Extraction, Abstractive Event Analysis, MUC-4, Information Extraction

1. Introduction

Identifying and interpreting events described in unstructured text is a central objective in Information Extraction (IE), bridging natural language and structured representations of real-world events. Within this broad goal, event extraction (EE) represents the dominant formulation, where events, their participants, and attributes are defined based on their linguistic properties, and are extracted as text spans directly from the input text. Prominent datasets such as ACE (Doddington et al., 2004), ERE (Song et al., 2015), and WikiEvents (Li et al., 2021), as well as recent generative modelling and evaluation approaches (Simon et al., 2024), uphold this span-centric assumption that relies on lexical anchors such as trigger words and explicit mentions. As a consequence, important event information expressed as implicit references (Sharif et al., 2024), relative temporal information, or other contextual aspects of real-world events (Olsen et al., 2024), is often left unaddressed. The resulting event representation may therefore be underspecified or ambiguous and less useful in many downstream tasks and practical applications.

Recent work has begun to explore abstractive approaches to the task, motivated by practical applications in domains such as medical discourse and peace research on armed conflicts. These domains often require event representations that capture information not explicitly stated in the text, such as inferred causes, aggregated outcomes, or normalised attributes to support reliable analysis or decision-making (Olsen et al., 2024). Moving beyond the view of event information as merely explicit text spans, recent efforts have developed higher-level annotations, integrating implicit, inferred, or aggregated information from across a document. Datasets such as DiscourseEE (Sharif et al., 2024), Lemonade (Semnani et al., 2025), and UCDP-AEC (Simon et al., 2025) exemplify this paradigm shift by including annotations expressed as categorical, normalised, or numerical values that cannot be captured with extractive approaches alone.

This shift resonates with early foundational work in information extraction, particularly the English MUC-4 dataset (Sundheim, 1992b), which was designed with an application-driven perspective to support information-seeking needs in the military domain. While most modern modelling approaches on the dataset treat it mainly as a span-based extraction benchmark (Du et al., 2021b; Das et al., 2022; Chen et al., 2023; Gantt et al., 2024), the original annotations also cover higher-level and contextual information.

Examples of annotations beyond text spans are illustrated in the MUC-4 sample in Figure 1, which shows a short document alongside two annotated events. Firstly, temporal information may be conveyed with relative expressions, such as the phrase "*yesterday*" in the document, which is annotated as a normalised date (24 Sep 88) in event 1, resolved using the document publication date (25 Sep 1988). Similarly, partially specified timestamps, such as "-23 Sep 88" in event 2, capture temporal uncertainty, reflecting that the document reports the discovery

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 7766–7780
11–16 May 2026.

25 Sep 1988, Ayacucho: According to military sources and people who arrived today in Ayacucho from the La Mar area, 44 peasants have been killed in two townships of La Mar Province, Ayacucho Department. Yesterday, a column of Shining Path terrorists arrived in the village of Chinchipe, in the jungle province of La Mar, and shot 16 peasants who were members of the peasant patrols that oppose the terrorists. The Shining Path guerrillas, who burned the murdered peasants' houses, charged them with collaborating with the army. The town of Chinchipe is 220 km north of Ayacucho. Travellers arriving from La Mar Province said that on 23 September, 28 bodies of alleged peasants were found near the town of Chullas. The bodies showed signs of torture and bullet wounds, but the identity of the murderers apparently could not be determined.

E1	**Date:** 24 Sep 88 **Country:** Peru **City:** *n/a* **Event type:** Attack **Event stage:** Accomplished **Weapon:** *n/a* **Weapon type:** Gun	**E2**	**Date:** -23 Sep 88 **Country:** Peru **City:** Chullas **Event type:** Attack **Event stage:** Accomplished **Weapon:** *"Torture"* **Weapon type:** Torture, Gun
Perp.:	**Category:** Terrorist act **Org.:** *"Shining Path"* **Individual:** *"Shining Path guerrillas"* **Confidence:** Reported as fact	**Perp.:**	**Category:** *n/a* **Org.:** *n/a* **Individual:** *"Murderers"* **Confidence:** *n/a*
Victim:	**Type:** Civilian **Description:** *"Peasants"* **Number:** 16 **Effect:** Death	**Victim:**	**Type:** Civilian **Description:** *"Peasants"* **Number:** 28 **Effect:** Death

Figure 1: Sample of MUC-4 document with two annotated events (E1 and E2). The example shows how the original annotations include both span-based information (field values enclosed in quotation marks) and non-span information.

of victims, rather than the attack itself. Spatial information follows a similar pattern, where places such as *La mar*, *Chinchipe*, and *Chullas* appear in the text but are abstracted to a higher-level geographic-level (Peru).

The sample also shows how categorical participant roles, such as weapon, victim, or perpetrator, are mapped into conceptual categories in addition to their text span representations. The Weapon Type field illustrates how this can require contextual inference: In Event 1, the type *Gun* is not mentioned explicitly in the text but rather inferred from the description "*shot 16 peasants*". In event 2, although "*torture*" can be extracted directly, the additional *Gun* must be inferred from "*bullet wounds*". Accordingly, MUC-4 can be viewed as an early instance of an abstractive event modelling task, positioning it as a foundational benchmark within this shift towards higher-level and contextually grounded event representations.

Following the terminology of Simon et al. (2025), we use the broader notion of *event analysis*, which we take to comprise both extractive and abstractive lines of work on representing events. We reformulate MUC-4 as an abstractive event analysis task, extending traditional event extraction approaches by incorporating both span-based and non-span-based event annotations. Moreover, where traditional event extraction uses the terms *argument role* and *argument value* to distinguish between the type of participants or attribute (e.g., Victim or Perpetrator), and its value, we adopt the terminology of *event field* and *field value* to capture both traditional text span values and higher-level event information.

Contributions This work revisits the MUC-4 dataset in its full complexity, including all span-based and inferred fields, to realign the benchmark with its relevant real-world use cases. We approach MUC-4 as an abstractive event analysis task, bridging traditional event extraction with the modern generative and document-level paradigm. To better understand the challenges posed by the dataset, we provide a quantitative framework for analysing each field's properties, combining established measures of domain openness with a novel span-match ratio that captures how each field relates to the source text. We further present the first systematic evaluation of generative models fine-tuned on the original MUC-4 schema, comparing how post-training steps and model size influence performance. Our experiments are complemented by an error analysis, which motivates future work. The code for preprocessing and training is available at https://github.com/helenebol/MUC-4-Revisited

2. Background and Related Work

The Message Understanding Conferences Towards the end of the 1980s and the beginning of the 1990s, the Message Understanding Conferences (MUC; Sundheim, 1996) were organised to advance research in information extraction from natural language text. The fourth iteration, MUC-4 (Sundheim, 1992b), defined a template-filling task covering terrorism events in Latin America, with predefined fields such as date, location, perpetrator, weapon, victim, and effect. Designed with an application-driven perspective, the annotation guidelines described the field values to be nor-

malised, inferred, mapped to a category, or directly extracted from the text. Reflecting the information-seeking needs of the intended users, MUC-4 was designed for a high-level document focus requiring more than just span identification.

From extractive to abstractive The seminal standard introduced by ACE (Doddington et al., 2004), and subsequent datasets such as ERE (Song et al., 2015), WikiEvents (Li et al., 2021) and Doc-EE (Tong et al., 2022), moved away from the high-level MUC-4 design by requiring all argument values to be contiguous text spans. While these resources have contributed significant progress for the event extraction task, they exclude information that does not exist verbatim in the text.

The growing demand for high-quality event data to support socio-political research has drawn attention to this limitation. Olsen et al. (2024) argue that traditional NLP event extraction datasets are poorly suited for socio-political applications, as many domain databases rely on normalised abstractions designed to capture what happened in the world rather than merely what is stated in the text. More broadly, researchers have advocated for shifting the event extraction task from an "extractive" paradigm focused on identifying spans within text to an "abstractive" paradigm that leverages free-form inference to capture higher-level event information (Simon et al., 2024).

In response, recent work has introduced datasets that cast conflict event analysis as an abstractive task, moving beyond text spans toward structured representations grounded in human expert annotations. Among these, the UCDP-AEC dataset (Simon et al., 2025) and the Lemonade dataset (Semnani et al., 2025) represent complementary efforts built on the established armed conflict databases UCDP GED (Sundberg and Melander, 2013) and ACLED (Raleigh et al., 2010). The DiscourseEE dataset (Sharif et al., 2024) illustrates a similar direction in health discourse by introducing implicit arguments (inferred from context) and scattered arguments (information distributed across multiple sections of a document) along with a semantic evaluation metric to capture predictions that are semantically accurate but differ in surface form.

Generative approaches to event extraction Sequence-to-sequence approaches such as Text2Event (Lu et al., 2021) and DEGREE (Hsu et al., 2022) reformulate event extraction as structured text generation. More recent work has applied instruction-tuned decoder-only models in both zero-shot and supervised settings (Huang et al., 2024; Gao et al., 2023; Zhu et al., 2024). The shift towards generative modelling for information extraction has enabled more flexible output formats that are better suited to non-span fields. Still, most previous work on generative approaches remains span-centric in practice, either by constraining the token generation to only include tokens present in the input, or by evaluating predictions against span-based annotations (Simon et al., 2024).

Recent work on MUC-4 In recent years the MUC-4 dataset has resurfaced in research on template filling (Du et al., 2021b; Das et al., 2022; Chen et al., 2023; Gantt et al., 2024), role-filler extraction (Du and Cardie, 2020; Huang et al., 2021; Du et al., 2021a), document level event extraction (Wang et al., 2023; Huang and Riloff, 2011; Fang et al., 2024; Huang et al., 2024) and event summarisation (Gantt et al., 2024).

Consistent with the extractive nature of the traditional event extraction task formulation, most previous work on MUC-4 focuses on a simplified set of event fields, limited to those with values that appear as text spans. Beyond event type, this simplified schema covers the perpetrator, victim, weapon, and physical target, with assumed mentions in the source documents, but omits other fields with categorical, numerical, or spatio-temporal information.[1] This paper addresses that gap, evaluating generative models on the full MUC-4 dataset.

3. Dataset Reintroduction

In this section, we provide an overview of the MUC-4 dataset, describe our preprocessing steps, and report statistics for the modified version of the dataset. Next, we analyse the different types of event information annotated for MUC-4, which we refer to as fields. To better characterise the structural properties of the fields, we perform a field-level analysis using three metrics. We use two established metrics for domain openness introduced by Simon et al. (2025), and we introduce a new metric, the Span match ratio (SMR), which quantifies the extent to which the field values exist as strings in the source document, i.e. are span-based. Finally, we analyse these metrics jointly to document the variation in how each event field is realised as text spans in the source text, lexical openness, and discuss how these differences can inform evaluation.

3.1. The Original MUC-4 Schema

The original MUC-4 dataset consists of 1700 documents covering terrorism events in Latin America, annotated at the document-level with zero or more events per document. The schema records categorical, numerical, and descriptive information about

[1] A notable exception is Gantt et al. (2024), which in their work on event summarisation also include location and date, when they occur as a text spans in the text.

perpetrators, physical targets, human victims and general information about the event, such as spatio-temporal information. Figure 1 shows a sample of a document and two corresponding annotated events. Unlike most event extraction datasets, MUC-4 is not annotated with trigger words. Instead, the event definitions are rooted in political science theory, delineating what constitutes a relevant incident.

3.2. Dataset Filtering and Preprocessing

Based on label frequencies and field distribution over the full corpus, we introduce a modified version of the MUC-4 dataset. The modified version is a result of the following filtering to the original annotations.

First, we restrict Event Type to the two most frequently labelled categories: Attack and Bombing. Together they account for over 83% of all annotated events, while remaining types such as Kidnapping (10.5%), Arson (3.6%), and Robbery (1.4%) are significantly less frequent and highly imbalanced across splits.

We further remove fields that only occur in a small fraction of the data or have low occurrence globally. Specifically, we exclude the fields Human Target Foreign Nation, Physical Target Foreign Nation, Human Target Total Number, and Physical Target Total Number, each of which appears in fewer than 5% of all events.

Finally, we split the location field into two separate fields, Country and City, where the City field covers previously annotated information labelled as either city or town.

These modifications result in a more balanced and interpretable dataset with improved coverage of field types. After filtering, 1263 of the original 1514 annotated events (83.4%) remain in the modified dataset.

3.3. Statistics

The dataset contains 1700 documents, but only about 55% of them contain event annotations, indicating a significant portion (45%) of documents without any recorded events, highlighting the importance of event detection for this task. Among documents with events, there is a moderate distribution between single-event documents and multi-event documents, with an average of 1.6 events per document. As a result of removing the less frequent event types, the number of documents with multiple events has decreased in the modified dataset. Since rare event types frequently co-occur with high-frequency ones, filtering them out leads to a notable reduction in multi-event documents.

Statistic	Original	Modified
Total documents	1700	1700
Total events annotated	1514	1263
Avg. events per doc	1.60 ($\pm$1.18)	1.48 ($\pm$0.97)
Docs with events	942 (55.4%)	852 (50.1%)
w/ exactly 1 event	620 (35.4%)	606 (35.6%)
w/ >1 event	322 (18.9%)	246 (14.4%)
Docs without events	758 (44.6%)	848 (49.9%)

Table 1: Comparison of dataset statistics before and after filtering for event types (*Attack*, *Bombing*)

3.4. Field Overview

The original MUC-4 dataset covers event information in four main categories: *Incident*-related information, such as where and when the event occurred, type of event and weapon used, *perpetrator* information, covering both organisational and individual actors, as well as the degree of certainty of their involvement and information on both *physical targets*, and *human victims*, specifying what and who, their conceptual categories, the effect of the incident, and the number of those affected.

According to the annotation guidelines, these fields are classified into three types on the basis of their values. *Text string values* are directly derived from the source document and correspond to explicit text spans representing the perpetrator, weapon, victim, and target as they appear in the text. These fields align with the traditional extractive event extraction task formulation, and have been the main focus of prior work on MUC-4. In contrast, *Canonical values* reflect values that must be normalised or standardised from the text, and include temporal and numerical values, such as the number of victims or targets and the event date. Finally, *text conversion values* are taken from a predefined finite set of possible values, capturing higher-level classifications and judgments about the status, nature, or impact of the event and its participants, such as event type, effect on victim or target, perpetrator category, and location. A complete list of the fields included after the preprocessing step is provided in Table 2.

Domain openness We define domain openness as the extent to which admissible values in a field are diverse and extend beyond those observed during training. We report on the two complementary measures Value density and Unique value overlap, in order to better understand how the distribution of admissible values may affect model performance. Value density measures the average number of instances per unique value. Higher ratios indicate a closed domain with a small set of recurring val-

ues, while lower ones indicate a more open domain with a more diverse value set that are more rarely reused. The unique value overlap captures the percentage of test values that were present during training, reflecting the proportion of novel values in the test data. For further details on these measures, see Simon et al. (2025).

Span Match Ratio Some of the annotated fields in MUC-4 are, by definition, text spans that explicitly occur in the source document (e.g., Perpetrator Individual, Physical Target, and Victim Name). Others may be expressed in a higher-level form, such as Perpetrator Confidence or Event Stage, where the annotated value represents an inferred or normalised concept rather than a surface string. Moreover, some fields *may* occur as a span in the document, such as Date or Location, but are not guaranteed. To capture the degree of explicit textual anchoring for each field, we introduce the *Span match ratio (SMR)*, as the proportion of annotated values that can be recovered as exact contiguous spans within their corresponding source document. Higher SMR values suggest that the annotations are directly grounded in the text, typical of traditional extractive EE, whereas low SMR values indicate annotated values that are normalised or inferred from context.

Formally, let $V(f)$ denote the set of all annotated values for a field f across the dataset, and let $\text{doc}(v)$ be the source document for field value v. The Span match ratio of field f is defined as the fraction of values that appear as a contiguous substring in their corresponding documents $\text{doc}(v)$:

$$\text{SMR}(f) = \frac{|\{v \in V(f) : v \in \text{doc}(v)\}|}{|V(f)|} \quad (1)$$

We report SMR as a percentage for each field in Table 2 for readability and direct comparison.

Event field variation By analysing the field-level statistics on unique value overlap, value density, and span match ratio as represented in Table 2, clear differences appear in how each event field relates to the source text and the degree of novelty they introduce at test time. Sorted based on unique value overlap, the event fields positioned at the top of the table have no unseen values in the test set (100% overlap), have a limited set of frequently occurring values (high value density), which rarely appear as explicit text spans in the source document (low span match ratio). For instance, the Event Stage field, with only three unique annotated values across the dataset – *Accomplished, Threatened, and Attempted* – occurs as a text span in only 1.77% of their corresponding documents.

In contrast, many event fields with low overlap between test and train values exhibit high annotation consistency as exact text spans, as evidenced by SMRs close to 100% and low value density. The fields aligned with the traditional extractive paradigm, such as Weapon, Victim Name, Physical Target, share these properties. Notably, the City field has similar properties, despite originating from a broader location field composed of concatenated categorical values.

Some event fields, such as Physical Target Number, Victim Number, and Date, exhibit high novelty in test-set values (exceeding 50%), whereas their presence as explicit text spans in source texts is limited. For example, only 46.15% of the Victim Number values are observed during training, and the annotated values appear as text spans in about 20% of the associated documents. These characteristics indicate that such fields cannot be captured using extractive methods alone, and some degree of abstractive approaches is needed.

Field	Unique Value Overlap	Value Density	Span Match Ratio
Perp. Confidence	100.00%	135.80	1.77%
Event Stage	100.00%	421.00	2.45%
Perp. Category	100.00%	551.00	7.89%
Victim Type	100.00%	149.40	19.08%
Effect on Physical Tgt.	100.00%	182.33	21.02%
Event Type	100.00%	631.50	37.29%
Physical Tgt. Type	92.86%	56.13	12.71%
Country	91.67%	74.35	59.97%
Effect on Victim	87.50%	157.11	20.16%
Weapon Type	87.50%	40.28	61.24%
Physical Tgt. Number	47.62%	23.91	18.28%
Victim Number	46.15%	19.85	20.55%
Date	40.37%	2.27	22.13%
Perp. Org.	40.35%	3.73	99.56%
City	36.59%	3.82	98.46%
Victim Name	35.48%	1.53	99.69%
Victim Description	35.22%	2.09	99.15%
Weapon	33.33%	2.96	99.57%
Physical Target.	23.16%	1.38	98.81%
Perp. Ind.	21.54%	1.86	99.52%

Table 2: Overview of field-level statistics, including value density, unique value overlap and instance-level overlap between the test and train split, as well as the Span match ratio, which measures the proportion of values for each field that occur as a text string in the source document.

3.5. Dataset Balance and Composition

Overall, distributions across dataset splits remain stable following modifications, with relative frequencies differing by less than 5% for most fields. Despite this split-level consistency, the dataset as a whole is notably imbalanced. Most incidents ($\approx$ 68%) are annotated as attacks, while bombing ac-

counts for the remainder. The majority of events occur in El Salvador (≈ 45%) and Colombia (≈ 30%), and nearly all are labelled as Accomplished (≈ 92%), as opposed to Threatened or Attempted. Bombs and Guns are the most common weapons, together representing almost 60% of all cases, and terrorist acts strongly outweigh state-sponsored violence. Civilians represent the most frequent victims (≈ 68%), and the event outcomes are skewed toward severe effects, with death reported in roughly 70% of incidents involving human victims and some damage in about 65% of those involving physical targets.

Although these patterns likely reflect real-world distributions captured in the source data, the overall imbalance may influence model behaviour. Limited exposure to less frequent event structures can lead to overfitting to dominant values and reduced generalisation for rarer types, roles, effects, or outcomes. Moreover, as noted in the previous subsection, because many fields are not strictly categorical, this imbalance may bias the model towards treating frequently observed values as discrete classes rather than contextual information inferred from the input.

4. Evaluation

We break down the evaluation of event analysis on MUC-4 into three views: event alignment, field value matching, and event detection evaluation.

Event alignment Because a single document in MUC-4 may describe multiple events, an important step in evaluation is to align the predicted events with the correct gold annotations. Building on prior work on template extraction (Du et al., 2021a; Chen et al., 2023), we frame event alignment as a one-to-one assignment problem, solved with the Kuhn-Munkres algorithm (Kuhn, 1955; Munkres, 1962), to maximise the similarity between predicted and gold events. Specifically, we use the Constrained Entity-Alignment F-Measure-Role-filler Mention Extraction (CEAF-RME; Chen et al., 2023), which measures the correspondence between the sets of predicted and gold values for a field, awarding partial credit when some values overlap. This overlap is captured by the similarity function $\phi_3(\mathrm{R}, \mathrm{S}) = |\mathrm{R} \cap \mathrm{S}|$, where R and S denote the sets of predicted and gold values for a given field, respectively, and $|\mathrm{R} \cap \mathrm{S}|$ measures the cardinality of their intersection, i.e., the number of overlapping string values shared by both sets.

Field-level performance We evaluate field-level performance on the matched event pairs using exact string matching, where a prediction is scored as correct only if it exactly equals the gold value. Because MUC-4 annotations for the span-based fields often list multiple correct mentions for the same entity occurring in the input document, we use a *prediction-to-any* policy, where a predicted value is correct if it matches any of the listed mentions in gold. This ensures that equivalent or synonymous surface forms occurring in the input document, such as "*Local government officials*" and "*Local authorities*", are treated as correct matches. We apply this policy during both alignment and final field level scoring. We report field-level exact-match precision, recall, and micro-averaged F_1 on the aligned event pairs.

Event detection Finally, we evaluate the models' ability to correctly identify the presence and number of events within a document. After the alignment with CEAF-RME, each predicted event is paired with at most one gold event. A predicted event that does not align with any gold event is counted as spurious, and a gold event without any aligned prediction is counted as missing. Event detection precision is the proportion of predicted events correctly aligned with a gold event, and recall is the proportion of gold events correctly aligned with a predicted event. We also report the event detection F_1 as the harmonic mean of the two.

5. Experimental Setup

We frame abstractive event analysis as a supervised text generation problem. Each training example consists of a document, its publication date and location, and the corresponding structured event representation. The objective is to identify relevant events within the document and generate a structured, schema-conformant representation containing up to 20 annotated fields per event. Full hyperparameter details are provided in Appendix B.3

All experiments use the original MUC-4 splits, with 1300 documents for training, 200 for validation, and 200 for testing. While certain fields include multiple valid surface forms, we use only the first listed value as the supervision signal.

We evaluate two families of openly available decoder-only models: Qwen 3 (Yang et al., 2025) and Olmo 3 (Olmo et al., 2025). To examine the effect of post-training alignment, we fine-tune the Olmo 3 base model as well as well as several instruction-tuned variants (SFT, DPO, RLVR), all with 7B parameters. For Qwen 3, we fine-tune four model sizes (0.6B, 4B, 8B, and 14B parameters) to analyse the scaling effect.[2] All models are fine-tuned to generate complete event structures directly in JSON format. We use a unified prompt

[2]To ensure comparability across model families, we disable Qwen's thinking mode.

Model	Size	Exact Match (%)			Event Detection (%)		
		P	R	F_1	P	R	F_1
Qwen3-0.6B	0.6B	17.7	24.8	20.6	32.1	43.4	36.9
Qwen3-4B	4B	21.4	**61.1**	31.7	29.5	**77.5**	42.7
Qwen3-8B	8B	53.4	46.8	49.9	75.2	53.3	**62.4**
Qwen3-14B	14B	**67.2**	39.8	**50.0**	**86.7**	46.7	60.7
OLMo-3-7B-Base	7B	62.5	**31.2**	**41.6**	85.3	**35.2**	**49.8**
OLMo-3-7B (SFT)	7B	**74.5**	11.9	20.5	**100.0**	14.3	25.0
OLMo-3-7B (DPO)	7B	60.2	22.6	32.8	79.7	25.8	39.0
OLMo-3-7B (RLVR)	7B	63.1	27.3	38.1	84.9	34.1	48.6

Table 3: Precision, recall, and F_1 for field-level exact match and event detection across Qwen 3 model sizes and Olmo-3-7B post-training configurations. The highest score for each metric in each model family is marked in bold.

template, described in Appendix B.2, across model families to ensure a fair and consistent comparison.

6. Results

In this section, we evaluate the fine-tuned models on the MUC-4 dataset, reporting exact match and event detection scores with a focus on scaling-behaviour of the Qwen 3 models, and the effect of post-training stages for Olmo 3. We then report per-field exact match performance to identify consistent difficulties and where model performance diverges. Next, we examine whether field-level properties described in Section 3.4 can explain variation in model performance. Finally, we focus on the best performing model and introduce semantic evaluation metrics to distinguish between surface-form errors and reasoning errors.

Table 3 reports precision, recall, and F_1 for both field-level exact match and event detection, across Qwen 3 model sizes and Olmo-3-7B post-training configurations. For Qwen 3, exact match F_1 shows a clear positive scaling trend, increasing from 20.6% at 0.6B to 31.7% at 4B and 49.9% at 8B. The 14B model yields a comparable 50.0%, suggesting that scaling beyond 8B provides minimal improvements. Precision generally improves with larger models, while recall is less consistent. Notably, the 4B model achieves the highest recall but at the cost of lower precision, indicating a tendency to over-generate. In contrast, larger models appear more selective, yielding higher precision, but lower recall. A similar trend holds for event detection among the Qwen 3 models, where the 8B model achieves the best precision–recall balance, resulting in the highest F_1 of 62.4%.

For Olmo 3, the base model achieves the highest exact match F_1 of 41.6%, outperforming all post-trained variants. The instruction-tuned models show consistently high precision, but low recall. This is most pronounced for the SFT model, with precision of 74.5%, but only 11.9% recall. This behaviour might reflect instruction tuning objectives that reward concise and safe responses, which can discourage generating uncertain events. Among the post-trained models, RLVR achieves performance closest to the base model, with an F_1 of 38.1%. A similar precision–recall pattern is evident for event detection, where the base model achieves the highest F_1 score, and post-trained models yield high precision but reduced recall.

Error analysis The per-field exact match scores reported in Table 4 show wide variation across event fields, with different models performing well on different subsets of fields. Notably, the larger Qwen 3 variants (8B and 14B) generally yield the highest scores, with Olmo-3-Base performing competitively.

Fields such as Weapon and Country consistently receive high scores across models, likely due to their clearly identifiable lexical forms. In contrast, most fields fall into the low-to-mid range, suggesting that field difficulty is not just determined by model scale or capacity, but that field-specific complexity also plays a role.

Section 3.4 introduced field-level intrinsic statistics (Table 2), including value space, unique value overlap, and span-match ratio, as potential difficulty indicators. However, when comparing these to the exact-match F_1 scores, no single property consistently explains the model's performance.

For example, Perpetrator Confidence, Event Stage, and Perpetrator Category have closed and frequently reused value sets, which suggest that, in principle, they should be easy to learn once the label set is known. Yet, their varied and sometimes low scores across models, indicate that other factors, such as annotation ambiguity or conceptual complexity, may contribute more to model error. Conversely, semi-open fields, such as Weapon Type and Country, contain several novel values

Models	Qwen3				Olmo-3 7B			
Field	0.6B	4B	8B	14B	Base	SFT	DPO	RLVR
Perp. Confidence	00.0	10.3	**35.2**	30.6	18.3	03.9	09.4	07.4
Event Stage	30.7	35.9	**61.9**	59.2	48.8	25.0	39.0	46.2
Perp. Category	01.3	29.9	**48.0**	46.5	33.2	07.7	24.3	33.5
Victim Type	17.1	29.7	39.0	**43.3**	38.4	20.1	26.5	40.4
Effect on Physical Tgt.	16.8	19.6	52.6	**53.1**	46.9	24.7	46.2	40.0
Event Type	35.5	25.8	**61.9**	58.2	48.6	24.2	38.3	47.4
Physical Tgt. Type	04.1	19.2	37.0	**48.1**	37.2	15.7	35.0	25.9
Country	30.2	41.6	**61.0**	59.7	49.8	25.0	38.3	47.8
Effect on Victim	24.7	41.5	44.3	**46.9**	43.4	25.0	35.9	44.1
Weapon Type	35.0	53.6	**63.5**	62.4	53.7	26.3	41.7	48.5
Physical Tgt. Number	24.7	36.6	45.1	**47.0**	38.5	19.2	37.7	30.3
Victim Number	31.6	**45.9**	40.6	43.3	38.4	21.3	27.5	37.7
Date	23.5	25.6	50.9	**53.4**	43.6	23.6	32.5	41.7
Perp. Org.*	15.7	23.7	**36.1**	33.3	28.8	05.8	16.7	14.3
City	14.2	31.7	**55.9**	55.4	49.0	26.1	38.6	37.8
Victim Name*	32.3	47.6	**55.7**	49.5	47.7	32.9	32.9	46.3
Victim Description*	18.6	32.6	36.1	**40.5**	35.5	16.8	23.7	32.7
Weapon*	36.0	56.3	**65.3**	65.1	55.9	34.9	55.0	49.4
Physical Tgt.*	15.4	33.7	**45.1**	33.3	33.3	17.5	29.3	25.9
Perp. Ind.*	14.8	24.1	**46.5**	42.2	34.7	16.3	34.7	34.8

Table 4: Overview of per field exact match F_1 score across model families and sizes. Darker cells imply higher F_1 scores, and the best score per field is marked in bold

at test time and have varying span presence, yet achieve comparatively high scores. This might suggest that these arguments tend to occur in predicable syntactic patterns, e.g., "*attacked with a [Weapon]*", or "*attack in [Country]*". Finally, fields with near-perfect span-match ratio, such as Perpetrator Individual and Victim Description, show large performance variation, suggesting that span-based annotations alone do not necessarily make them easier.

6.1. Semantic Evaluation

The previous section analyses the overall and per-field performance across all models using strict exact match. Below, we evaluate the strongest model, Qwen3-14B, with semantic metrics in order to distinguish reasoning errors from lexical mismatches. For example, predicting "*politician*" for a gold annotation of "*political figure*" reflects semantic understanding, yet receives the same exact match score as a conceptual error such as "*civilian*".

We use two complementary approaches based on BERT embeddings (Devlin et al., 2019). BERTScore (Zhang et al., 2019) measures semantic similarity using token-level cosine similarity aggregated via greedy matching. Precision@1(P@1), introduced by Simon et al. (2025), is inspired by information retrieval. For each field, all unique values are embedded using BERT, and the prediction is ranked against this candidate set by cosine similarity, and P@1 equals 1 only if the gold value ranks first. Unlike BERTScore, P@1 requires the prediction to be nearest the correct value, making it suitable for quantifying reasoning success on fields with a closed value space, such as Event Type and Event Stage.

We report exact match precision, P@1, and BERTScore F_1 in Figure 2, excluding Victim Number, Physical Tgt. Number, and Date, as semantic similarity measures are not meaningful for numeric values. Across all fields, BERTScore F_1 exceeds exact match precision, indicating that a large proportion of errors are semantically close to the gold value. We observe the largest gap for Physical Target (+18.4), Victim Description (+18.4), and Victim Type (+14.7), while Event Type, Event Stage, and Country have minimal difference across metrics.

By including P@1, we can determine whether errors reflect just semantic proximity or correct relative alignment within a fields' value space. For several fields, P@1 exceeds exact match, indicating that many surface form errors are semantically closest to the gold value. However, divergence between BERTScore and P@1 indicates that semantic similarity does not always imply conceptual correctness, as predictions may be closely related to the gold value while still collapsing distinctions that are meaningful within that field.

For example, predicting *Bombing* instead of *Attack* yields a BERTScore F_1 of 80.5%, but the task requires that these values are treated as distinct, rather than near paraphrases. P@1 captures this distinction by penalising semantically similar but conceptually incorrect predictions.

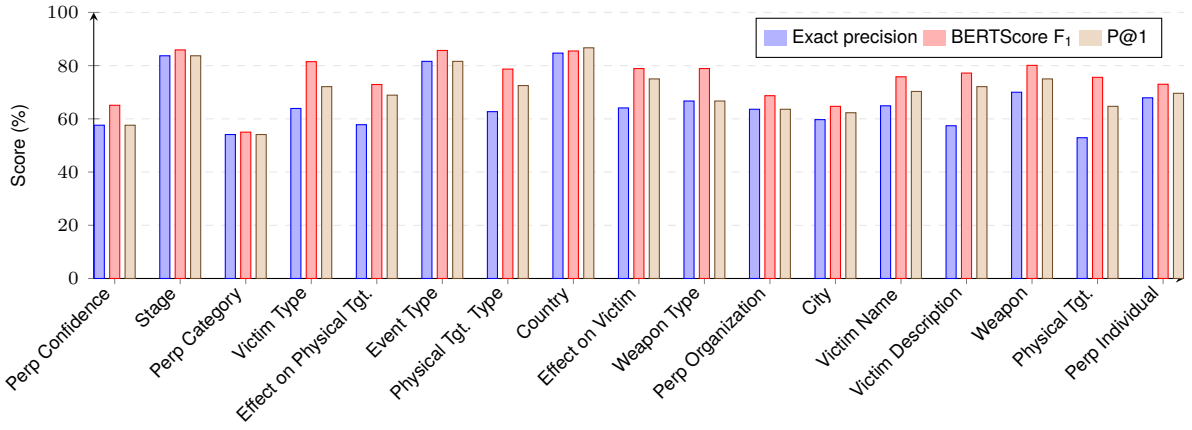

Figure 2: Per field evaluation for Qwen3-14B comparing exact match precision to BERTScore F_1 and Precision@1.

On the other hand, P@1 reflects relative alignment within the candidate set rather than absolute semantic equivalence. Consequently, semantically equivalent predictions may still receive a score of 0 if another candidate value ranks higher. For example, predicting "*10 bombs*" for a gold value of "*ten bombs*" receives a high BERTScore, whereas P@1 may assign a score of 0 if another candidate value is identical to the prediction.

7. Conclusion and Future Work

This work provides the first systematic evaluation of modern generative models for predicting the rich document-level event representations annotated in the original MUC-4 dataset. While recent previous work focused on a small subset of fields whose values can be identified as substrings in the input text, we include fields that may require reasoning over implicit information. As such, we position MUC-4 within the recently proposed paradigm of abstractive event analysis.

To better characterise the dataset and its modelling challenges, we propose a quantitative framework combining established metrics for domain openness with a novel metric, span-match ratio, which captures the degree to which each field's values are explicitly stated in the text. Our analysis suggests that textual grounding, as measured by span-match ratio, provides valuable insights on the data, but does not fully explain variation in model performance across fields.

We report results for fine-tuning a selection of decoder models across two model families, Qwen 3 (0.6B–14B) to examine scaling effects, and Olmo 3 (7B) across post-training configurations (Base, SFT, DPO and RLVR). Our results show a clear positive scaling trend for Qwen 3, with performance plateauing between 8B and 14B parameters. For Olmo 3, the base model outperforms all post-trained variants, with an imbalance of precision-recall for the instruction-tuned models.

To move beyond strict surface form evaluation, we introduce a semantic evaluation protocol combining BERTScore and P@1. Our results indicate that exact match scoring underestimates model performance, as many predicted values are semantically close to the gold value despite surface differences. At the same time, divergence between BERTScore and P@1 reveals cases where models collapse meaningful conceptual distinctions within a field's value space. This suggests that semantic similarity alone is insufficient for event extraction and should be complemented with ranking based similarity metrics such as P@1.

Several directions remain open for future work. First, while scaling beyond 8B provides minimal gains in our experiments, it is not clear whether this is due to limitations of the dataset, the fine-tuning setup, or model architecture. Future work could investigate whether larger models benefit from alternative training regimes, variation in instructions, or increased training data. Second, examining ways to improve over the Olmo 3 base model warrants more attention. Rather than general instruction following, exploring strategies designed for structured extraction could be an interesting path for future work. Moreover, enabling chain-of-though reasoning, including Olmo 3's think variants and Qwen 3 reasoning mode, may benefit fields requiring inference over subtle contextual cues. Finally, our semantic evaluation highlights the need for more nuanced evaluation for event extraction. Future research should focus on developing evaluation metrics better aligned with MUC-4's hybrid nature, rewarding semantic correctness while penalising conceptual errors, and explore dedicated handling of fields with special structure, such as dates, potentially drawing on external resources such as calendar information.

8. Limitations

Part of this study evaluates the effects of model size on abstractive event analysis. Our largest model, Qwen3-14B, is not considered a "*large*" model, which now often exceeds 70B parameters. As a result, the observed scaling behaviour may not scale to larger models.

However, to our knowledge, this is the first study to evaluate LLMs on the original MUC-4 dataset. We therefore prioritise smaller models (0.6B–14B) to ensure that the experiments are computationally accessible and reproducible for the research community. Future work should investigate whether scaling beyond 14B confirms the trends observed in this paper.

Another limitation concerns the small size of the MUC-4 dataset. With only 1700 documents, the dataset provides limited supervision for fine-tuning large generative models. This may also influence the models' abilities to learn the less frequent fields. Furthermore, the MUC-4 dataset contains only English documents and annotations, and the findings in this work may not generalise to other languages. Although important efforts have been made to translate the dataset into other languages (Gantt et al., 2024), these translations only include the text span fields. An important path for future work is to extend this translation effort to include the full event schema for MUC-4.

The temporal scope of MUC-4 is a potential limitation of this work. The documents in MUC-4 report on events that occurred more than 20 years ago. Models trained on MUC-4 may therefore capture geopolitical context and linguistic patterns specific to conflict reporting during that time period. As a result, the models might not generalise well to more recent descriptions of terrorism events. We plan to examine the robustness of abstractive event analysis models when applied to more recent news corpora, such as UCDP-AEC, in future work.

Acknowledgements

This work was supported by the Research Council of Norway with funding to the Peace Science Infrastructure project (grant number 322425). The computations were performed on resources provided through Sigma2 – the national research infrastructure provider for High-Performance Computing and large-scale data storage in Norway.

9. Bibliographical References

Yunmo Chen, William Gantt, Weiwei Gu, Tongfei Chen, Aaron White, and Benjamin Van Durme. 2023. Iterative document-level information extraction via imitation learning. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1858–1874, Dubrovnik, Croatia. Association for Computational Linguistics.

Aliva Das, Xinya Du, Barry Wang, Kejian Shi, Jiayuan Gu, Thomas Porter, and Claire Cardie. 2022. Automatic error analysis for document-level information extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3960–3975, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Xinya Du and Claire Cardie. 2020. Document-level event role filler extraction using multi-granularity contextualized encoding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8010–8020, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021a. Grit: Generative role-filler transformers for document-level event entity extraction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics.

Xinya Du, Alexander Rush, and Claire Cardie. 2021b. Template filling with generative transformers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 909–914, Online. Association for Computational Linguistics.

Quntian Fang, Feng Liu, Zhen Huang, Zhenliang Guo, Changjian Wang, Dongsheng Li, and Minghao Hu. 2024. Controllable template generation for document-level event extraction. In *2024 4th International Conference on Neural Networks, Information and Communication Engineering (NNICE)*, pages 325–329. IEEE.

William Gantt, Alexander Martin, Pavlo Kuchmiichuk, and Aaron Steven White. 2024. Event-keyed summarization. *arXiv preprint arXiv:2402.06973*.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction.

I-Hung Hsu, Kuan-Hao Huang, Elizabeth Boschee, Scott Miller, Prem Natarajan, Kai-Wei Chang, and Nanyun Peng. 2022. DEGREE: A data-efficient generation-based event extraction model. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1890–1908, Seattle, United States. Association for Computational Linguistics.

Kuan-Hao Huang, I-Hung Hsu, Tanmay Parekh, Zhiyu Xie, Zixuan Zhang, Prem Natarajan, Kai-Wei Chang, Nanyun Peng, and Heng Ji. 2024. TextEE: Benchmark, reevaluation, reflections, and future challenges in event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12804–12825, Bangkok, Thailand. Association for Computational Linguistics.

Kung-Hsiang Huang, Sam Tang, and Nanyun Peng. 2021. Document-level entity-based extraction as template generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5257–5269, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ruihong Huang and Ellen Riloff. 2011. Peeling back the layers: Detecting event role fillers in secondary contexts. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1137–1147, Portland, Oregon, USA. Association for Computational Linguistics.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Sha Li, Heng Ji, and Jiawei Han. 2021. Document-level event argument extraction by conditional generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 894–908, Online. Association for Computational Linguistics.

Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. 2021. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online. Association for Computational Linguistics.

James Munkres. 1962. Algorithms for the assignment and transportation problems. *Society for Industrial and Applied Mathematics*, 15:196–210.

Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, Pete Walsh, Pradeep Dasigi, Robert Berry, Saumya Malik, Saurabh Shah, Scott Geng, Shane Arora, Shashank Gupta, Taira Anderson, Teng Xiao, Tyler Murray, Tyler Romero, Victoria Graf, Akari Asai, Akshita Bhagia, Alexander Wettig, Alisa Liu, Aman Rangapur, Chloe Anastasiades, Costa Huang, Dustin Schwenk, Harsh Trivedi, Ian Magnusson, Jaron Lochner, Jiacheng Liu, Lester James V. Miranda, Maarten Sap, Malia Morgan, Michael Schmitz, Michal Guerquin, Michael Wilson, Regan Huff, Ronan Le Bras, Rui Xin, Rulin Shao, Sam Skjonsberg, Shannon Zejiang Shen, Shuyue Stella Li, Tucker Wilde, Valentina Pyatkin, Will Merrill, Yapei Chang, Yuling Gu, Zhiyuan Zeng, Ashish Sabharwal, Luke Zettlemoyer, Pang Wei Koh, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. Olmo 3.

Helene Olsen, Étienne Simon, Erik Velldal, and Lilja Øvrelid. 2024. Socio-political events of conflict and unrest: A survey of available datasets. In *Proceedings of the 7th Workshop on Challenges and Applications of Automated Extraction of Socio-political Events from Text (CASE 2024)*, pages 40–53, St. Julians, Malta. Association for Computational Linguistics.

Omar Sharif, Joseph Gatto, Madhusudan Basak, and Sarah Masud Preum. 2024. Explicit, implicit, and scattered: Revisiting event extraction to capture complex arguments. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12061–12081, Miami, Florida, USA. Association for Computational Linguistics.

Étienne Simon, Helene Olsen, Huiling You, Samia Touileb, Lilja Øvrelid, and Erik Velldal. 2024. Generative approaches to event extraction: Survey and outlook. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pages 73–86, Miami, Florida, USA. Association for Computational Linguistics.

Barry Wang, Xinya Du, and Claire Cardie. 2023. Probing representations for document-level

event extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12675–12683, Singapore. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

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

Mengna Zhu, Kaisheng Zeng, JibingWu JibingWu, Lihua Liu, Hongbin Huang, Lei Hou, and Juanzi Li. 2024. LC4EE: LLMs as good corrector for event extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12028–12038, Bangkok, Thailand. Association for Computational Linguistics.

10. Language Resource References

George R Doddington, Alexis Mitchell, Mark A Przybocki, Lance A Ramshaw, Stephanie M Strassel, and Ralph M Weischedel. 2004. The automatic content extraction (ACE) program-tasks, data, and evaluation. In *International Conference on Language Resources and Evaluation*, volume 2, pages 837–840. Lisbon.

William Gantt, Shabnam Behzad, Hannah An, Yunmo Chen, Aaron White, Benjamin Van Durme, and Mahsa Yarmohammadi. 2024. MultiMUC: Multilingual template filling on MUC-4. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 349–368, St. Julian's, Malta. Association for Computational Linguistics.

Clionadh Raleigh, rew Linke, Håvard Hegre, and Joakim Karlsen. 2010. Introducing ACLED: An armed conflict location and event dataset. *Journal of Peace Research*, 47(5):651–660.

Sina Semnani, Pingyue Zhang, Wanyue Zhai, Haozhuo Li, Ryan Beauchamp, Trey Billing, Katayoun Kishi, Manling Li, and Monica Lam. 2025. LEMONADE: A large multilingual expert-annotated abstractive event dataset for the real world. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 25813–25852, Vienna, Austria. Association for Computational Linguistics.

Étienne Simon, Helene Bøsei Olsen, Ramón Carreño, Rahul Mishra, Nikolay Arefyev, Mert Can Yilmaz, Lilja Øvrelid, and Erik Velldal. 2025. Abstractive event analysis of armed conflicts: Introducing the UCDP-AE dataset. In *Proceedings of the 21st Conference on Natural Language Processing (KONVENS 2025): Workshops*, pages 104–119, Hannover, Germany. HsH Applied Academics.

Zhiyi Song, Ann Bies, Stephanie Strassel, Tom Riese, Justin Mott, Joe Ellis, Jonathan Wright, Seth Kulick, Neville Ryant, and Xiaoyi Ma. 2015. From light to rich ERE: annotation of entities, relations, and events. In *Proceedings of the the 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 89–98.

Ralph Sundberg and Erik Melander. 2013. Introducing the UCDP Georeferenced Event Dataset. *Journal of Peace Research*, 50(4):523–532.

Beth M. Sundheim. 1992a. Overview of the fourth Message Understanding evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Beth M. Sundheim. 1992b. Overview of the fourth Message Understanding Evaluation and Conference. In *Fourth Message Understanding Conference (MUC-4): Proceedings of a Conference Held in McLean, Virginia, June 16-18, 1992*.

Beth M Sundheim. 1996. The message understanding conferences. In *TIPSTER TEXT PROGRAM PHASE II: Proceedings of a Workshop held at Vienna, Virginia, May 6-8, 1996*, pages 35–37.

MeiHan Tong, Bin Xu, Shuai Wang, Meihuan Han, Yixin Cao, Jiangqi Zhu, Siyu Chen, Lei Hou, and Juanzi Li. 2022. DocEE: A large-scale and fine-grained benchmark for document-level event extraction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3970–3982, Seattle, United States. Association for Computational Linguistics.

A. Dataset Details

Table 6 shows an overview of the full set of fields in MUC-4, a short description of the type of information they represent, and examples of such values. The fields are grouped by the field value category as defined by the annotation guidelines. The original MUC-4 dataset is available here: https://www-nlpir.nist.gov/related_projects/muc/muc_data/muc_data_index.html

A.1. Preprocessing Details

Table 5 shows the relative prevalence of event types annotated in the original MUC-4 dataset. This distribution serves as the motivation for our focus on *Attack* and *Bombing* events, as they make up together more than 83% of the annotated events across splits. Moreover, the two event types are compatible with a unified template which supports all event fields listed in MUC-4, whereas the less prevalent event types impose field constraints. For example, *Arson* and *Kidnapping* never include a Weapon or Weapon type field, while *Kidnapping* additionally excludes physical targets. To maintain schema consistency and simplify model training, we filter out the entire event annotation for types other than *Attack* or *Bombing* during pre-processing.

A.2. Field Distribution and Annotation Rate

Table 7 summarises per-field statistics after filtering, covering the number of unique values, values in the test set that are not seen during training, annotation rates across Train/Val/Test sets, and shows the absolute changes from the original dataset. For the text string fields, we report the type-token ratio instead of unique values, as a more informative measure for span-based fields.

Not all fields are required to be annotated for every event. There is variation in annotation rates across fields, where fields such as Date, Event Type, Country, and Event Stage are nearly always annotated. Other fields are annotated in fewer than a third of all events, such as Effect on Physical target (31%) and Weapon (32%). While a low annotation rate means that there are few training instances available for a field, this does not seem to directly affect model performance in the fine-tuning experiments. As reported in Table 4, the best performing model yields an exact match F_1 score of over 50 % for Effect on Physical target, and models perform consistently well on the Weapon field.

#Unseen measures the number of field values appearing in the test set that were never seen during training. While most of the fields have no new values during training, fields such as City, Victim Number, and Physical Tgt. Number require models to generalise to previously unseen values.

Most fields show stable annotation rates across splits. Some exceptions include Perp. Category with 93% coverage in train, but only 73% in test, and Perp. Ind, with 50% in train vs. 67% in test.

The Δ column reports the change in annotation rate in percentage points relative to the full unfiltered dataset. The table shows that our filtering has a modest overall effect, with most fields changing by at most $\pm$3pp.

B. Experimental Settings

B.1. Model selection

All models used for fine-tuning are publicly available on Hugging Face. We list the exact checkpoints used in our experiments below.

Qwen3

- `Qwen/Qwen3-0.6B`
- `Qwen/Qwen3-4B`
- `Qwen/Qwen3-8B`
- `Qwen/Qwen3-14B`

OLMo-3

- `allenai/OLMo-3-1025-7B (Base)`
- `allenai/Olmo-3-7B-Instruct-SFT`
- `allenai/Olmo-3-7B-Instruct-DPO`
- `allenai/OLMo-3-7B-Instruct`

B.2. Prompt

We use the instruction shown in Figure 3 for all models. For Qwen3 models, we include "no_think" in the instruction in order to disable the model's reasoning mode.

> Instruction: "Extract all events described in the document as a JSON array where each event is an object with fields:{_LABEL_CSV}. Omit any field not mentioned. Reply with JSON only."

Figure 3: Instruction used to fine-tune models on MUC-4. For Qwen3 models, "no_think" is included in the instruction to disable reasoning mode.

Both Qwen3 and Olmo-3 Instruct models are typically used with chat-style prompt formatting. In preliminary experiments, we also used chat-styled prompt wrapping, but observed that this often caused models to produce invalid json outputs. We therefore use the simpler instruction format shown in Figure 3 for all experiments.

	Train	Val	Test	Total prop
# Annotated templates	1114	191	209	1514
Attack	57.09%	50.79%	60.77%	56.80%
Bombing	26.75%	26.18%	26.32%	26.62%
Kidnapping	10.23%	16.23%	6.70%	10.50%
Arson	3.86%	4.19%	1.44%	3.57%
Robbery	1.53%	1.57%	0.48%	1.39%
Forced Work Stoppage	0.54%	0.52%	1.91%	0.73%
Attack / Bombing	–	–	1.91%	0.26%
Bombing / Attack	–	0.52%	0.48%	0.13%

Table 5: Overview of event type distribution with conditional distribution of values given annotation for each data split and global frequencies relative to all annotated events.

	Field Name	Description	Examples:
Set fill	Incident Type	Event type*	Attack, Bombing, Arson, Kidnapping
	Event Stage	Event status	accomplished, attempted, threatened
	Location	The location the event occurred	Peru: Lima (City)
	Weapon Type	Weapon category	Gun, Explosive, projectile
	Perp. Category	Event subcategory	Terrorist act, State-sponsored violence
	Perp. Confidence	Certainty of the perp org. involvement	Reported as fact, Possible
	Physical Tgt. Type	Category of the physical target	Transport route, Civilian residence
	Physical Tgt. Foreign Nation	Nationality physical target if different from location	United states
	Effect on Physical Tgt.	The impact on a physical target	Destroyed, Damaged, Some damage
	Victim Type	Category of human target	Civilian, Active military, "Political figure"
	Victim Foreign nation	nationality of hum tgt if different from location	Peru
	Effect on Victim	The impact of the incident on a human target(s)	"Death", "Injury", "No injury", "Escaped"
Text conversion	Date	Date(s) the event occurred	25 AUG 88 , 25 AUG 88–27 AUG 88
	Physical Tgt. Number	The number of physical targets	1
	Physical Tgt. Total number	The total number of physical targets	4 , 10
	Victim Total number	The total number of human targets	2
	Victim Number	The number of human targets	3
Text string	Weapon*	Weapon used by perpetrator if mentioned	"bombs", "Czechoslovak bombs"
	Perp. Individual*	Person responsible for the attack	"Four terrorists", "Salvadorian rebels"
	Perp. Organization*	An organization responsible for the incident	"FARC", "ELP", "Army"
	Physical Target *	A physical thing that was attacked if mentioned.	"Bridge", "Houses", "power system"
	Victim Name*	The name of the victim	"Adolfo Spezua"
	Victim Description	Title or description of the victim	"Engineer", "Mayor", "Bodyguard"

Table 6: Overview of fields, their data type as defined in the MUC-4 annotation guidelines, description and examples. * marks the fields that are used as a simplified template in previous work.

B.3. Hyperparameters for Supervised Fine-tuning

All models are fine-tuned using the AdamW optimiser with a cosine learning-rate schedule with 50 warm-up steps and fp16 precision. Training runs for for up to 10 epochs, with validation evaluated at the end of each epoch. Early stopping is applied based on validation loss with a minimum improvement threshold of 0.001. For the smaller models (Qwen3 0.6B and 4B), we use a smaller learning rate of 5×10^{-5} to stabilise training and set patience to 2 for early stopping. For all models of size 7B and larger across both model families, we set the learning rate to 2×10^{-4}, and increase early stopping patience to 4. All experiments use a per-device batch size of 1 with gradient accumulation of 4. We train all models using LoRA with the PEFT library with a single LoRA configuration across all models with r16, α =16, dropout=0.05.

B.4. Evaluation Detail

For the semantic evaluation, we follow Simon et al. (2025) and compute Precision@1 using contextual embeddings from `bert-base-uncased`. To ensure consistency between the two For simplicity, we use the same model for BERTScore. While the official recommendation for BERTScore is `microsoft/deberta-xlarge-mnli`, we prioritise using the same backbone to make sure that differences between the metrics reflects the evaluation formulation, and not differences in embedding models. BERTScore is computed using the official implementation. [3]

[3] https://github.com/Tiiiger/bert_score

Field	Vocabulary		Annotation Rate (%)				Δ
	#Unique	#Unseen	Train	Val	Test	Total	
Date	575	99	98	95	98	98	0
Event Type*	2	0	100	100	100	100	0
Country	17	0	100	100	100	100	0
City	187	37	56	51	53	55	+1
Event Stage	3	0	100	100	100	100	0
Weapon Type	18	1	51	49	47	50	+7
Perp. Category	2	0	93	72	73	87	−1
Perp. Confidence	5	0	49	52	53	50	−3
Physical Tgt. Type	15	1	46	48	52	47	+2
Effect on Physical Tgt.	4	0	29	31	34	30	+1
Physical Tgt. Number	35	9	45	48	52	46	+2
Victim Type	10	0	70	71	74	71	0
Effect on Victim	9	1	71	69	69	70	+6
Victim Number	86	16	70	71	74	71	0
Text string fields							
	TTratio	#Unseen	Train	Val	Test	Total	
Weapon*	0.17	23	31	35	28	31	+5
Perp. Ind.*	0.18	90	50	67	67	55	−2
Perp. Org.*	0.10	12	49	52	53	50	−3
Physical Tgt.*	0.22	124	46	48	52	47	+2
Victim Name*	0.34	63	35	31	36	35	−2
Victim Description	0.20	51	32	30	31	32	+2

Table 7: Overview of fields in MUC-4 with event types **Attack, Bombing**. * marks fields used in the simple template in previous work. #Unique = number of unique values; for text string fields this is the type-token ratio instead. #Unseen = unique values in test not seen during training. Δ = absolute difference in percentage points after filtering.