REGen: A Reliable Evaluation Framework for Generative Event Argument Extraction

Omar Sharif, Joseph Gatto, Madhusudan Basak, Sarah M. Preum
Department of Computer Science, Dartmouth College
{omar.sharif.gr, sarah.masud.preum}@dartmouth.edu

Abstract

Event argument extraction identifies arguments for predefined event roles in text. Existing work evaluates this task with exact match (EM), where predicted arguments must align exactly with annotated spans. While suitable for span-based models, this approach falls short for large language models (LLMs), which often generate diverse yet semantically accurate arguments. EM severely underestimates performance by disregarding valid variations. Furthermore, EM evaluation fails to capture implicit arguments (unstated but inferable) and scattered arguments (distributed across a document). These limitations underscore the need for an evaluation framework that better captures models' actual performance.

To bridge this gap, we introduce **REGen**, a **R**eliable **E**valuation framework for **Gen**erative event argument extraction. REGen combines the strengths of exact, relaxed, and LLM-based matching to better align with human judgment. Experiments on six datasets show that REGen reveals an average performance gain of +23.93 F1 over EM, reflecting capabilities overlooked by prior evaluation. Human validation further confirms REGen's effectiveness, achieving 87.67% alignment with human assessments of argument correctness.

1 Introduction

Information extraction is a key area in natural language processing (Gaizauskas and Wilks, 1998). Event argument extraction (EAE) is a core information extraction task that transforms text into structured information. As EAE identifies and extracts event-specific arguments from texts, it is essential for a wide range of applications such as document understanding (Tong et al., 2022), misinformation detection (Wu et al., 2022), discourse understanding (Sharif et al., 2024), pharmacovigilance (Sun et al., 2022). With the emergence of generative models (e.g., LLMs), EAE has gained significant attention in recent years (Zhang et al., 2024, 2025). However, previous studies (Gao et al., 2023; Sun et al., 2024) indicate that LLMs perform poorly on EAE tasks. This is largely due to the disconnect between the nature of generative predictions and the exact span-based evaluation method commonly used for EAE (Huang et al., 2024).

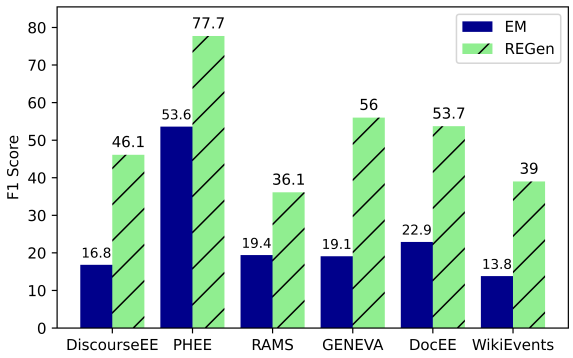

Figure 1: Performance comparison of the best-performing EAE model across six datasets under Exact Match (EM) and the REGen framework. The results highlight that, on average, EM underestimates model performance by 54.8%, which is captured by REGen.

Span-based exact matching (EM) significantly underestimates the performance of LLMs as they often predict accurate arguments in surface forms that differ from the ground truth. For example, if the ground truth annotation for a role is `'pain relief'`, the model might output terms like `[alleviates pain, reducing discomfort, analgesia]`. Depending on the context, all or multiple of these outputs are correct, but none would be accepted by EM. Even minor variations would result in no match. Authors in (Sharif et al., 2024) highlighted that this problem is even more pronounced when evaluating the arguments composed of information from different parts of the text (scattered arguments) or the arguments that are not directly mentioned (implicit arguments).

Previous works have attempted to address these

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 12146–12168
November 4-9, 2025

issues using embedding-based relaxed matching, which considers two arguments similar if they have high embedding similarity (Han et al., 2024). However, this approach fails to capture semantically similar arguments with different lexical forms and wrongly classifies arguments with high token overlap as similar (Sharif et al., 2024). For example, in Figure 2 for the role *'patient concerns'*, the ground-truth argument `'limited insurance coverage'` and the predicted argument `'coverage limitations for FLA and cryotherapy'` refer to the same issue. Due to lexical variation, relaxed matching fails to capture this. In contrast, consider a role *'date'* for which ground-truth and predicted arguments are `'18 April 2024'` and `'20 April 2024'`, respectively. These two arguments are different, but relaxed matching considers them the same due to high token overlap. Context is needed when evaluating these arguments. Recent work by Lu et al. (2024) used LLMs as judges to identify similar arguments. This approach requires a large number of inferences, adding significant computational costs. Additionally, without human validation, LLM-based judgments can produce unreliable results. Relying solely on relaxed match or judge-based approaches can overestimate performance by incorrectly classifying non-match arguments as matches, leading to inflated and unreliable model assessments. A detailed analysis of argument correctness by each method is shown in Table 6.

To address these limitations, we introduce **REGen**, a reliable evaluation framework for event argument extraction. REGen systematically combines the strengths of exact, relaxed, and LLM-based matching by **maximizing the evaluation reliability while minimizing the computation costs**. Figure 2 illustrates the framework, and it is structured into four sequential phases: *Exact Match (EM)*, *Relaxed Match (RM)*, *Complex Match (CM)*, and *Alignment with Human Judgments*.

The EM level filters arguments that match exactly, reducing computational costs for subsequent stages by eliminating obvious matches. This level does not require human evaluation as exact matches indicate perfect agreement with humans. The RM stage identifies arguments that are semantically similar, making evaluation robust to minor syntactic variations. This matching is performed based on the contextual embedding of the arguments. Setting up a high embedding similarity threshold ensures higher reliability and minimizes human evaluation.

After filtering out exact and relaxed matches, unmatched arguments are carried forward for complex matching. The CM stage captures semantically similar arguments based on context despite lexical and/or syntactic differences. We leverage LLM as a judge (Zheng et al., 2023) to determine argument similarity. Finally, in the judgment alignment stage, we propose a novel **Judgment Aligned Match (JAM)** score to factor in the scores from each level to account for misjudgments based on human validation. This framework ensures evaluation accuracy, cost-effectiveness, and better alignment with human judgments.

To the best of our knowledge, this is the first systematic evaluation of LLMs on popular EAE datasets. Unlike prior studies (Lu et al., 2024; Huang et al., 2024) that experimented on small test subsets sampled and merged from multiple datasets, we evaluate the complete test sets of the original datasets. This provides a more reliable assessment of LLMs' performances on these benchmarks and highlights their potential in solving the EAE task, which has been previously underestimated. Our key contributions are as follows.

- We present **REGen**, a **R**eliable **E**valuation framework for **Gen**erative event argument extraction, minimizing inference costs and the need for human validation. REGen yields 87.67% alignment with humans thus ensuring higher reliability. We also introduce a scoring mechanism to systematically measure how well REGen's evaluation aligns with human judgments. Finally, we curate a novel, human-annotated dataset with 900 samples to select LLM models as judges for EAE evaluation.

- We demonstrate the generalizability of REGen through extensive evaluation using multiple LLMs on six widely-used EAE datasets, including DiscourseEE (Sharif et al., 2024), PHEE (Sun et al., 2022), RAMS (Ebner et al., 2020), GENEVA (Parekh et al., 2023), DocEE (Tong et al., 2022), and WikiEvents (Li et al., 2021). The results show an average improvement of 23.93 F1 points across all datasets while reducing inference costs by 41.2% than the LLM-as-judge-only approach (Lu et al., 2024).

Reproducibility: Our code, evaluation framework, the judge and alignment datasets, and other relevant resources are available at https://github.com/Omar-Sharif/REGen.

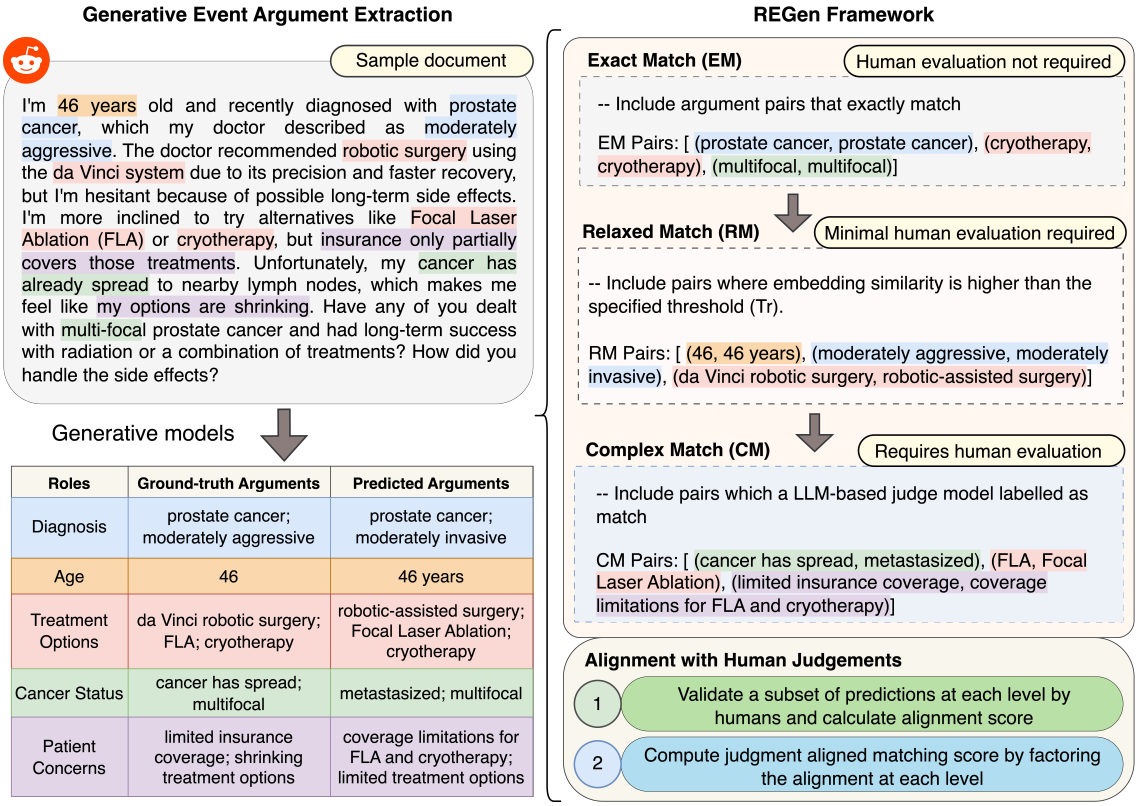

Figure 2: Proposed REGen evaluation framework for event argument extraction. ***Left:*** An example of getting role-specific arguments from documents using generative models. Different colors indicate arguments for different roles. Semicolons separate multiple arguments for a role. ***Right:*** Illustration of the REGen's sequential evaluation process: Exact Match (2.2), Relaxed Match (2.3), and Complex Match (2.4) and Alignment with Human Judgments (2.5). Only the arguments that do not match at the previous level are carried forward to the next level. Due to space constraints, the mathematical illustration of the framework is provided in the Appendix Figure 4.

2 REGen Framework

2.1 Preliminaries

Document: A document D is a piece of text, which can be a sentence, a paragraph, or a full document.
Events, Roles and Arguments: Events (E) refer to occurrences or actions described in document D. A document can have multiple events. Each event is characterized by its roles (R), which define the participants or entities involved. Arguments (A) are specific details or attributes associated with these roles, providing context such as specific time, location, and other information. For example, consider the sentence: *'Alice sent a package to Bob on Monday'*. The event here is '`Send`', with potential roles such as sender, recipient, and time. The corresponding arguments for these roles are Alice, Bob, and Monday, respectively.
Generative Event Argument Extraction: This approach leverages generative models such as LLMs to extract arguments from the source document D. Given the source document, along with information about events and roles, the model generates a structured list of associated arguments.

2.2 Level-1: Exact Match

Let's assume we have a list of predicted and ground-truth argument strings for each role R_i in a D.

$$P = [p_1, \ldots, p_{x1}], \quad G = [g_1, \ldots, g_{y1}]$$

An exact match (EM) pair is defined when $p_i = g_j$, forming a list of EM pairs such as $[(p_1, g_2), (p_3, g_5), \ldots, (p_x, g_y)]$. Precision and recall for the EM level are computed as:

$$\text{EM}_p = \frac{\text{NP}_e}{|P|}, \quad \text{EM}_r = \frac{\text{NG}_e}{|G|}$$

Here, NP_e and NG_e represent the number of correctly predicted arguments from the predicted (P) and ground-truth (G) argument lists, respectively. Note that $\text{NP}_e = \text{NG}_e$ under exact match.

2.3 Level 2: Relaxed Match

Predicted and ground-truth argument lists are updated by removing arguments matched in Level-1.

$$P_{rm} = [p_1, \ldots, p_{x2}], \quad G_{rm} = [g_1, \ldots, g_{y2}]$$

We compute the embedding-based similarity for all possible argument pairs of P_{rm} and G_{rm}. A pair is considered a relaxed match (RM) if its similarity score exceeds the predefined threshold T_r. The threshold selection method is described in section 3.1. The resulting list of RM pairs is a subset of all possible pairs. Precision and recall for relaxed matching are computed as:

$$\text{RM}_p = \frac{\text{NP}_e + \text{NP}_r}{|P|}, \quad \text{RM}_r = \frac{\text{NG}_e + \text{NG}_r}{|G|}$$

Here, NP_r and NG_r represent the arguments matched under relaxed conditions form P_{rm} and G_{rm}, while NP_e and NA_e are taken from Level 1. Relaxed matching allows an argument (p_x or g_y) to appear in multiple pairs where the similarity exceeds T_r. To avoid overcounting, separate counts (NP_r, NG_r) are maintained for the number of arguments correctly matched from the prediction list and the ground-truth list.

2.4 Level 3: Complex Match

After exact and relaxed matching, unmatched arguments are carried forward for complex matching.

$$P_{cm} = [p_1, \ldots, p_{x3}], \quad G_{cm} = [g_1, \ldots, g_{y3}]$$

For the possible pairs from these lists, a preselected judge model determines similarity based on context. Details on how a judge model is selected for complex matching are discussed in section 3.2. If a pair is predicted as similar, it is added to the complex match (CM) pair list. Precision and recall for complex matching are computed as:

$$\text{CM}_p = \frac{\text{NP}_e + \text{NP}_r + \text{NP}_c}{|P|}, \quad \text{CM}_r = \frac{\text{NG}_e + \text{NG}_r + \text{NG}_c}{|G|}$$

Here, NP_c and NG_c represent arguments correctly matched by the judge model. Similar to relaxed match, separate counts ensure that arguments are not overcounted when they appear in multiple matches. NP_e, NG_e, NP_r, and NG_r are precomputed values from previous levels.

Generic Equation: We define three match levels $L = [\text{EM}, \text{RM}, \text{CM}]$. NP_e, NP_r, and NP_c denote the number of correctly predicted arguments from the prediction list (P), while NG_e, NG_r, and NG_c represent the number of correctly matched arguments from the ground-truth list (G) at each level. Finally, precision, recall, and F1-score for a given level are calculated using the Equations 1-2.

$$\text{P}_l = \frac{\sum_{i=1}^{l} \text{NP}_i}{|P|}, \quad \text{R}_l = \frac{\sum_{i=1}^{l} \text{NG}_i}{|G|} \tag{1}$$

$$\text{F1}_l = \frac{2 * \text{P}_l * \text{R}_l}{\text{P}_l + \text{R}_l} \tag{2}$$

2.5 Alignment with Human Judgments

In Levels 2 (Relaxed Match) and 3 (Complex Match), the performance can be overestimated if the relaxed matching model or the complex match judge incorrectly classifies a non-match pair as a match. To account for this overestimation, we introduced a novel **Judgment Aligned Match (JAM)** Score, which penalizes the counts on each level based on the deviation from human judgment.

We first calculate the *deviation rate* of a matching model (M) on a dataset (DT) by measuring the number of disagreements between the model and the human evaluator. The deviation rate is computed using equation 3. Addition details on the deviation rate or the alignment calculation are provided in Appendix B.

$$\text{E}_{(M,DT)} = \frac{\text{N}_d}{\text{N}_o} \tag{3}$$

Here, N_d and N_o denote the number of disagreements and the total number of observations, respectively. We calculate the **JAM Score** for a dataset factoring the model's score at each matching level (EM, RM, CM) by the deviation rate of that level following equations 4-6.

$$\text{JAM}_p = \frac{\sum_{i=1}^{L}((1 - \text{E}_i) * \text{NP}_i)}{|P|} \tag{4}$$

$$\text{JAM}_r = \frac{\sum_{i=1}^{L}((1 - \text{E}_i) * \text{NG}_i)}{|G|} \tag{5}$$

$$\text{JAM}_{f1} = \frac{2 * \text{JAM}_p * \text{JAM}_r}{\text{JAM}_p + \text{JAM}_r} \tag{6}$$

The JAM Score improves the alignment with human judgment, providing a more reliable reflection of the model's true performance.

3 REGen Implementation Details

3.1 Threshold Selection for Relaxed Match

Usually, the model-predicted arguments contain the core words from the corresponding ground-truth arguments (Sharif et al., 2024; Lu et al., 2024).

While these predictions may have redundant words or miss some surrounding words, such discrepancies do not alter the overall semantics. We can identify these variations by using a high threshold relaxed match for accurate evaluation. We consider two arguments similar if their semantic similarity score exceeds 0.85, calculated using SBERT embeddings (Reimers and Gurevych, 2019).

This threshold is determined as follows. We tested three thresholds: 0.95, 0.85, and 0.75, across 500 argument pairs sourced from the six EAE datasets evaluated. The disagreement (error) rates were 0.0%, 1.78%, and 8.33% for these thresholds, respectively. Although the 0.95 threshold yielded perfect agreement with human assessments, it allowed us to filter only a limited number of arguments. Conversely, the 0.75 threshold led to many incorrect matches. Therefore, we selected 0.85 as the optimal threshold. Our judgment alignment step ensures that our results are reliable and not inflated due to misjudgments.

3.2 Judge Selection for Complex Match

Studies show that LLMs achieve a strong correlation with human judgment across various tasks (Fu et al., 2024; Liu et al., 2023). We also used LLMs to determine whether the ground truth and predicted arguments match. This approach makes the evaluation scalable across datasets and models.

Judge data annotation: We construct a *judge dataset comprising 900 argument pairs* (150 pairs per dataset) to select the best judge model. Specifically, we randomly select pairs not matched under exact or relaxed criteria, meaning they inherently represent challenging or ambiguous cases. Each pair is annotated as *'match'* or *'non-match'* by a human annotator. A second human verifies the labels, and disagreements are resolved through discussion to finalize the annotations.

Judge LLM selection: We evaluate both open-source (Llama3.1-70B) and closed-source (GPT-4o, GPT-4o-mini, GPT-3.5) models as potential judges, assessing their performance in *zero-shot* and *chain-of-thought* settings. GPT-4o, with a zero-shot prompt, achieves the highest agreement with human judgments, scoring 86.17. Therefore, we selected GPT-4o as the judge model for complex match evaluation. Note that the choice of judge is orthogonal to our proposed framework. The selected judge model can easily be swapped with newer or better alternatives without further modifications. Appendix E provides additional details on judge selection and relevant prompts.

3.3 Judgment Alignment

Datasets	Deviation Rate (%) EM	RM	CM	Alignment (%) (1-$\sum$ deviation)
DiscourseEE	0.0	2.67	13.33	84.0
PHEE	0.0	0.0	7.33	92.67
RAMS	0.0	1.33	8.66	90.0
GENEVA	0.0	2.0	8.0	90.0
DocEE	0.0	3.33	16.0	80.67
WikiEvents	0.0	0.0	11.33	88.67
	Avg. Alignment (%)			87.67

Table 1: Alignment and judgment deviation rate from humans at different matching levels on the evaluated EAE datasets.

We manually evaluated a subset of predictions from each matching level to determine the alignment with human judgments. In total, we analyzed 2,700 arguments (900 for each level) to quantify the frequency of disagreements with humans. To ensure unbiased judgments, we randomly selected 150 outputs from each level for each evaluated dataset. Table 1 presents the alignment and deviation rates. The EM consistently showed perfect alignment with human judgments, while RM exhibited minimal disagreement. However, CM demonstrated the highest deviation rates.

Among the datasets, the PHEE dataset showed the highest alignment (92.67%) with human judgments, while DocEE had the lowest (80.67%). On average, the **REGen framework achieved 87.67% alignment with human evaluators**. Our analysis reveals primary reasons for judgment disagreements are (1) *Numerical nuances:* the model often failed to distinguish numerical differences. Such as for a role '`drug-dosage`,' it incorrectly treated '`14 mg`' and '`6 mg`' as equivalent. (2) *Temporal variations:* dates such as '`18 April`' versus '`20 April`' or days like '`Thursday`' versus '`Friday`' were incorrectly judged as similar. (3) *Coreference handling:* datasets like RAMS and WikiEvents frequently used pronouns (e.g., 'he', 'they') in the ground truth, while models predicted specific names (e.g., 'John'). This mismatch led to judgment errors, especially when documents contained multiple names, confusing the model. We identified a total of 111 disagreements. Of these, 15 cases were due to numerical nuances, 10 cases involved temporal variations, 57 cases were related to coreference handling, and 29 cases were due to other issues, such as the model incorrectly matching unrelated arguments. A detailed breakdown of

Datasets	Disagreement category			
	Numerical	Temporal	Coreference	Other
DiscourseEE	7	1	15	1
PHEE	1	1	4	5
RAMS	0	0	10	5
GENEVA	0	0	8	7
DocEE	6	8	12	3
WikiEvents	1	0	8	8
Total	15	10	57	29

Table 2: Detailed breakdown of disagreement cases between human and judge model in evaluated EAE datasets.

these disagreement categories for each dataset is provided in Table 2. Additional error analysis is provided in Appendix A.

The proposed JAM score accounts for these judgment errors. The score for each dataset is calculated based on alignment, providing a more reliable estimate of a model's true performance when using relaxed matching and LLM as judge models instead of human evaluators.

4 Experiments

4.1 Datasets and Experimental Setup

We used six standard EAE datasets from diverse domains to evaluate REGen. These datasets include: **RAMS** (Ebner et al., 2020) (news), **GENEVA** (Parekh et al., 2023) (book, news, journal articles), **DocEE** (Tong et al., 2022) (long news documents), **WikiEvents** (Li et al., 2021) (Wikipedia texts), **DiscourseEE** (Sharif et al., 2024) (online health discourse), and **PHEE** (Sun et al., 2022) (pharmacovigilance texts). Prior works, such as Huang et al. (2024) and Lu et al. (2024) have evaluated LLMs using small test subsets sampled and merged from multiple datasets. Thus not reflecting actual performance of LLMs on these datasets. We conduct evaluations using the complete official test sets of the selected datasets to provide a more reliable assessment of LLMs' performance on these benchmarks. Detailed statistics for these test datasets are presented in Table 4. Appendix C contains additional details on the data preparation steps.

Performance Metrics: We report the precision, recall, and F1-score at each evaluation phase: exact match, relaxed match, complex match, and post-judgment alignment. Scores are computed following prior works (Peng et al., 2023) and calculation details are discussed in Section 2.

4.2 EAE Models

Baselines: Following prior works (Sharif et al., 2024; Lu et al., 2023), we implement question-answering-based baselines. We use two models: BERT and FLAN-T5. Both models are fine-tuned on SQuAD (Rajpurkar et al., 2016) data to extract `arguments` from `context` based on the `question`.

LLM Based Models: We perform comprehensive experiments using open-source and closed-source LLMs from different model families of various parameter sizes, including Phi-3.5 (3.8B), Gemma-1.1 (7B), Mixtral (8x7B), Llama-3.1 (70B), and GPT-4o. We evaluate all the models in two prompt settings: *zero-shot* and *chain-of-thought*. We employed question-guided prompting as previous works achieved SOTA performance using this approach (Lu et al., 2023; Hsu et al., 2022; Du and Cardie, 2020). Specifically, models are prompted with `(Instruction, Document, Question)` to generate → `(Arguments)`, where each question is tailored to extract specific role[1]. Sample questions for the datasets are presented in Table 18.

Different LLMs require prompts and in-context samples tailored to each model and dataset. In practice, users select the optimal prompt using a trial-and-error approach (Ziems et al., 2024; Zamfirescu-Pereira et al., 2023). However, in our experiments, iterating over various prompts to find the optimal prompt for each model and dataset is impractical. Instead, we opted to use a consistent prompt across all models and datasets to (i) ensure a fair comparison among the models and (ii) eliminate the confounding factors related to prompt optimization. Generic templates for zero-shot and chain-of-thought prompts for argument extraction are illustrated in Figures 8 and 9, respectively. Additional descriptions of the models are provided in Appendix D.

5 Results

Significant improvement in F1-score across all datasets: Table 5 illustrates performance of various models using REGen framework. We observed a notable performance boost when models transitioned from the EM to the JAM score. For instance, the F1-score for the top-performing GPT-4o model increased from 16.82 with EM to 46.16 with JAM in the DiscourseEE dataset. Additionally, the average F1-scores of all the LLMs shown in Table 3

[1]Role-specific questions for each dataset can be found here: https://tinyurl.com/38en8e94

Datasets	Avg. EM-F1	Avg. REGen-F1	Δ F1	Gain (%)
DiscourseEE	10.74	37.45	+26.71	+248.69
PHEE	39.26	62.34	+23.08	+58.79
RAMS	13.38	28.27	+14.89	+111.25
GENEVA	13.62	46.12	+32.49	+238.52
DocEE	17.33	41.65	+24.32	+140.36
WikiEvents	8.93	31.02	+22.08	+247.18
	Avg. Δ F1		+23.93	

Table 3: Average F1-scores of LLMs in zero-shot and chain-of-thought settings, comparing Exact Match (EM) and REGen evaluation frameworks. Additional comparisons with Relaxed Match (RM) and Complex Match (CM) are reported in Table 9.

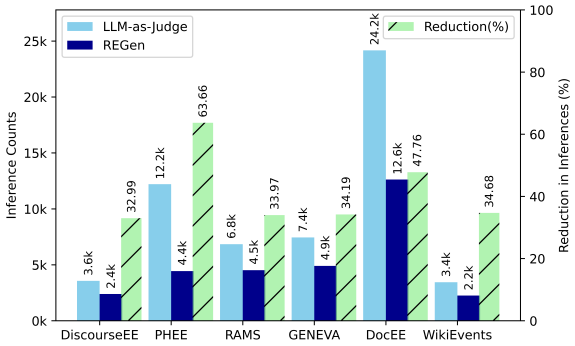

Figure 3: Comparison of required inference counts and reduction in inferences when using LLM-as-Judge versus the REGen framework for the GPT-4o prediction model. Additional statistics are presented in Table 11.

exhibit that all evaluated datasets achieved considerable performance gains, averaging 23.93 points. The increase in F1 score for the GENEVA dataset was 32.49, representing a 238.52% improvement over the standard EM evaluation. Similar substantial gains were noted in other datasets, such as 26.71 for DiscourseEE and 24.32 for DocEE.

On average, 41.20% of inferences are reduced under the REGen framework: Our results in Figure 3 and Table 11 demonstrate that the REGen framework significantly lowers the number of inferences needed for evaluation compared to solely using the LLMs-as-judge approach (Lu et al., 2024). For example, in the PHEE dataset, the inference count drops dramatically from 12,206 to 4,436, resulting in a reduction of 63.6%. Similarly, the DocEE dataset sees a decrease from 24,166 to 12,624, corresponding to a 47.7% reduction. These results highlight the efficiency of the REGen framework in streamlining the inference process. It enables effective evaluation by significantly decreasing the computational burden. Moreover, the systematic reduction in judgment errors through the REGen framework lessens the need for human validation without compromising reliability.

REGen framework is more reliable (87.67% alignment): REGen shows no/minimal errors in the performance under exact match and relaxed match scoring. While there is some overestimation due to misjudgments in the complex match step, our extensive validation indicates an 87.67% alignment with human judgments (see Table 1). The JAM score incorporates this human alignment, ensuring the overall reliability of the framework. Additionally, the reported scores are more explainable, as they include a clear breakdown of performance gains at each matching level (EM, RM, CM, and JAM).

Recall is on average higher than precision in all settings: Our fine-grained analysis (see Tables 12 to 17) reveals LLMs achieve higher recall than precision. Such as the GPT-4o model in the DocEE dataset achieved a JAM recall of 68.41 compared to a precision of only 42.12. This indicates while the models are effective in identifying ground-truth arguments, they tend to over-predict, impacting the overall F1-score. In this work, we used a single prompt for all the models and datasets, which might have contributed to this overprediction. Future research should focus on pushing the performance through dataset- and model-specific prompting to enhance precision without sacrificing recall.

6 Related Work

Generative Event Argument Extraction: Early studies on event argument extraction (EAE) treated it as an extractive or token-level classification task (Doddington et al., 2004; Du and Cardie, 2020). These efforts primarily focused on identifying argument spans directly found in the text (Sun et al., 2022). Recently, EAE has been formulated as a generative task where pre-trained language models are guided with natural language to fill templates or generate arguments (Hsu et al., 2022). Sharif et al. (2024) argue that this generative formulation better suits real-world applications as it can capture implicit and scattered arguments better. With the emergence of LLMs, generative model-based argument extraction gained more traction (Sun et al., 2024; He et al., 2024; Gatto et al., 2025). So, we focus on generative extraction covering diverse models and datasets.

Evaluations for Generative EAE: Existing works for generative EAE primarily rely on *exact match-*

Datasets	#Events	#Roles	#Docs	#Arguments	Doc-length (words)	Argument Density	Domain
DiscourseEE	3	34	98	997	121.21	10.17	Online health discourse
PHEE	2	14	968	4952	20.12	5.11	Pharmacovigilance
RAMS	129	63	754	2023	133.70	2.68	News
GENEVA	115	196	899	3078	29.74	3.42	General (book, news, journal)
DocEE	57	266	500	3453	635.60	6.90	News
WikiEvents	33	44	19	473	653.87	24.89	Wikipedia

Table 4: Test set statistics of the six datasets used for evaluation show broad variability among these datasets. The columns #Events, #Roles, #Docs, and #Args represent the number of unique event types, unique role types, unique documents, and number of arguments, respectively. The average document length is measured in words, and argument density reflects the average number of arguments per document.

	DiscourseEE				**PHEE**				**RAMS**			
Model	EM	RM	CM	JAM	EM	RM	CM	JAM	EM	RM	CM	JAM
Baselines												
BERT	5.88	8.66	33.56	30.18	27.78	34.98	52.61	51.33	14.63	18.14	33.61	32.24
Flan-T5	6.74	10.16	36.46	32.87	42.34	50.44	66.98	65.77	12.61	15.13	28.62	27.43
LLMs with Zero-Shot Prompt												
Phi-3.5	3.40	5.00	14.73	13.39	43.03	50.46	67.67	66.42	15.34	17.92	34.19	32.76
Gemma-1.1	11.87	15.86	50.14	45.48	45.00	54.34	76.93	75.28	14.87	17.50	32.43	31.11
Mixtral	13.10	17.74	48.59	44.38	36.58	42.55	59.19	57.98	12.97	15.46	29.93	28.65
Llama-3.1	13.38	18.73	43.57	40.13	39.17	46.95	63.96	62.72	11.95	14.56	25.44	24.47
GPT-4o	16.82	23.08	49.87	46.16	53.67	61.92	78.96	77.72	19.44	23.15	37.42	36.15
LLMs with Chain-of-thought Prompt												
Phi-3.5	7.08	12.42	41.41	37.43	32.09	38.01	54.03	52.86	15.53	18.65	34.54	33.13
Gemma-1.1	9.35	13.06	43.27	39.16	34.14	42.28	61.46	60.06	10.71	13.57	26.28	25.15
Mixtral	4.99	7.00	26.39	23.76	29.46	37.28	50.75	49.77	6.85	8.28	17.00	16.23
Llama-3.1	12.65	17.31	44.75	40.98	31.29	39.93	52.51	51.59	10.66	12.76	23.72	22.75
GPT-4o	14.77	20.86	47.33	43.66	48.14	55.66	70.01	68.96	15.50	19.58	33.56	32.30

	GENEVA				**DocEE**				**WikiEvents**			
Baselines												
BERT	15.24	26.58	53.09	50.74	18.66	25.74	47.81	44.05	6.46	9.55	29.44	27.2
Flan-T5	18.34	30.85	57.76	55.36	18.55	24.97	45.4	41.92	9.27	11.8	29.4	27.41
LLMs with Zero-Shot Prompt												
Phi-3.5	13.20	25.46	49.80	47.61	14.26	19.95	38.39	35.25	9.08	10.90	34.53	31.86
Gemma-1.1	11.69	24.40	50.73	48.37	17.99	26.78	46.77	43.28	6.22	7.31	34.37	31.32
Mixtral	13.31	24.86	48.44	46.32	22.91	32.55	58.16	53.74	9.89	12.36	38.24	35.31
Llama-3.1	16.36	29.62	55.09	52.79	17.56	25.14	46.44	42.80	12.81	15.48	38.94	36.29
GPT-4o	19.16	33.35	58.30	56.02	21.91	31.65	56.40	52.14	13.80	17.00	41.85	39.04
LLMs with Chain-of-thought Prompt												
Phi-3.5	10.76	21.76	46.31	44.12	19.84	27.79	48.51	44.93	6.87	8.53	31.97	29.32
Gemma-1.1	9.38	21.21	45.51	43.33	9.87	14.71	26.02	24.05	3.25	4.61	17.63	16.16
Mixtral	16.37	26.77	47.07	45.24	6.90	9.63	19.28	17.65	4.49	6.07	19.58	18.06
Llama-3.1	9.46	17.29	32.05	30.72	19.79	29.68	53.99	49.78	10.78	13.11	36.56	33.91
GPT-4o	16.54	28.50	48.48	46.65	22.26	31.92	57.27	52.90	12.15	15.10	41.93	38.90

Table 5: Evaluation results using the REGen framework for event argument extraction across the six datasets. The table reports F1-scores for models assessed at different evaluation levels: Exact Match (EM), Relaxed Match (RM), Complex Match (CM), and Judgment-Aligned Match (JAM). Due to space constraints, detailed precision, recall, F1-scores, and additional results are provided in Appendix Tables 10-17. The highest and the second-highest values in a column are highlighted using a dark shade and light shade, respectively.

ing for evaluation (Huang et al., 2024). This strict approach unfairly penalizes models, even when the generated output is correct (Fane et al., 2025). To address this, Han et al. (2024) adopts a relaxed matching approach, considering arguments similar if their embedding-based similarly exceeds a threshold of 0.5. Similarly, Sharif et al. (2024) used a threshold of 0.75. However, this approach has limitations. It fails to capture semantically similar arguments with different lexical or syntactic forms and wrongly classifies arguments with high token overlap as similar. Thus, performance reported solely on relaxed matching is unreliable. More recently, Lu et al. (2024) employed LLMs to determine argument similarity. Nonetheless, this approach incurs significant computational overhead and demands extensive human validation. Our REGen framework combines the strengths of exact, relaxed, and LLM-based matching. It systematically reduces misjudgments, computational costs, and the need for human validation.

7 Conclusion

This paper presents REGen, a novel evaluation framework for EAE. Our extensive experiments and human validation demonstrate its effectiveness, with a 23.93-point gain in average F1 score across six EAE datasets, and reliability, with 87.67% alignment with human judgments. We highlight the limitations of current evaluation approaches and illustrate how REGen addresses these issues. Furthermore, our analysis reveals that previous studies have underestimated the true performance of LLMs. We believe that REGen fills a critical gap in EAE research and motivates future work to explore the generative model's capability in solving other information extraction tasks, e.g., relation extraction, entity extraction, and beyond.

8 Limitations

One limitation of this work is that we did not conduct statistical significance testing on the reported results. We chose not to conduct statistical testing for two reasons. First, our goal is not to conclude which model is best but to highlight performance gaps and show how existing evaluation approaches underestimate model performance. The results clearly demonstrate a significant performance gap with exact match evaluation, which is not diminished by the lack of statistical testing. Second, performing statistical tests across all datasets and models with multiple runs is time-consuming and prohibitively expensive. For example, averaging over 3 runs would require an additional 320k inferences.

Another limitation is that we did not optimize prompts for each model. Performance could be improved with dataset- and model-specific prompting. However, we chose to focus on benchmarking a wide range of datasets using model-agnostic prompting. Conducting a thorough, prompt engineering for every model and dataset in a single study is not feasible. Our results show significant performance gains and future work can explore dataset- and model-specific prompts to further enhance performance. Additionally, future work can explore few-shot experiments and find optimal prompting strategies for different datasets, such as self-consistency (Wang et al., 2023b) or plan-and-solve (Wang et al., 2023a). We exclude few-shot experiments as they require selecting demonstration examples through trial and error, finding the optimal order of demonstration, and running multiple iterations, which significantly increases the experimental cost and complexity.

Ethical Considerations

Intended Use: We released our judge and alignment datasets to facilitate future research on generative argument extraction evaluation.
Annotation: Judge and alignment data annotation were conducted by trained NLP researchers. All annotators were compensated as per the standard paying rate of the author's institution. Key characteristics of our annotators include: (a) graduate students, (b) 3-6 years of research experience, and (c) a mix of native and non-native English speakers. We provided annotators detailed annotation guidelines, including argument extraction, semantical similarity, and the type of information we wanted to compare to mitigate potential biases.
Reproducibility Details on models and dataset processing are provided in Appendices C and D. The evaluation framework, code, and processed datasets are available at https://github.com/Omar-Sharif/REGen..

Acknowledgments

This work was supported in part by National Institutes of Health (NIH) grant 1R21DA059665-01A1. We also thank the reviewers for their valuable feedback.

References

George Doddington, Alexis Mitchell, Mark Przybocki, Lance Ramshaw, Stephanie Strassel, and Ralph Weischedel. 2004. The automatic content extraction (ACE) program – tasks, data, and evaluation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Xinya Du and Claire Cardie. 2020. Event extraction by answering (almost) natural questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online. Association for Computational Linguistics.

Seth Ebner, Patrick Xia, Ryan Culkin, Kyle Rawlins, and Benjamin Van Durme. 2020. Multi-sentence argument linking. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8057–8077, Online. Association for Computational Linguistics.

Enfa Fane, Md Nayem Uddin, Oghenevovwe Ikumariegbe, Daniyal Kashif, Eduardo Blanco, and Steven Corman. 2025. BEMEAE: Moving beyond exact span match for event argument extraction. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5734–5749, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576, Mexico City, Mexico. Association for Computational Linguistics.

Robert Gaizauskas and Yorick Wilks. 1998. Information extraction: Beyond document retrieval. In *International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 2, August 1998*, pages 17–60.

Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. 2023. Exploring the feasibility of chatgpt for event extraction. *Preprint*, arXiv:2303.03836.

Joseph Gatto, Omar Sharif, Parker Seegmiller, and Sarah M. Preum. 2025. Document-level event-argument data augmentation for challenging role types. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25109–25131, Vienna, Austria. Association for Computational Linguistics.

Gemma-Team. 2024. Gemma: Open models based on gemini research and technology. *Preprint*, arXiv:2403.08295.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2025. A survey on llm-as-a-judge. *Preprint*, arXiv:2411.15594.

Ridong Han, Chaohao Yang, Tao Peng, Prayag Tiwari, Xiang Wan, Lu Liu, and Benyou Wang. 2024. An empirical study on information extraction using large language models. *Preprint*, arXiv:2305.14450.

Shiming He, Yu Hong, Shuai Yang, Jianmin Yao, and Guodong Zhou. 2024. Demonstration retrieval-augmented generative event argument extraction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4617–4625, Torino, Italia. ELRA and ICCL.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, and et al. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

AI @ Meta1 Llama Team. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Di Lu, Shihao Ran, Joel Tetreault, and Alejandro Jaimes. 2023. Event extraction as question generation and

answering. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1666–1688, Toronto, Canada. Association for Computational Linguistics.

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Chen Xu, and Heyan Huang. 2024. Beyond exact match: Semantically reassessing event extraction by large language models. *Preprint*, arXiv:2410.09418.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2024. Gpteval: A survey on assessments of chatgpt and gpt-4. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7844–7866, Torino, Italy. ELRA and ICCL.

Microsoft. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *Preprint*, arXiv:2404.14219.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Tanmay Parekh, I-Hung Hsu, Kuan-Hao Huang, Kai-Wei Chang, and Nanyun Peng. 2023. GENEVA: Benchmarking generalizability for event argument extraction with hundreds of event types and argument roles. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3664–3686, Toronto, Canada. Association for Computational Linguistics.

Hao Peng, Xiaozhi Wang, Feng Yao, Kaisheng Zeng, Lei Hou, Juanzi Li, Zhiyuan Liu, and Weixing Shen. 2023. The devil is in the details: On the pitfalls of event extraction evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9206–9227, Toronto, Canada. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. 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.

Zhaoyue Sun, Jiazheng Li, Gabriele Pergola, Byron Wallace, Bino John, Nigel Greene, Joseph Kim, and Yulan He. 2022. PHEE: A dataset for pharmacovigilance event extraction from text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5571–5587, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zhaoyue Sun, Gabriele Pergola, Byron Wallace, and Yulan He. 2024. Leveraging ChatGPT in pharmacovigilance event extraction: An empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 344–357, St. Julian's, Malta. Association for Computational Linguistics.

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.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023a. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2609–2634, Toronto, Canada. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Xueqing Wu, Kung-Hsiang Huang, Yi Fung, and Heng Ji. 2022. Cross-document misinformation detection based on event graph reasoning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 543–558, Seattle, United States. Association for Computational Linguistics.

J.D. Zamfirescu-Pereira, Richmond Y. Wong, Bjoern Hartmann, and Qian Yang. 2023. Why johnny can't prompt: How non-ai experts try (and fail) to design llm prompts. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23, New York, NY, USA. Association for Computing Machinery.

Xinliang Frederick Zhang, Carter Blum, Temma Choji, Shalin Shah, and Alakananda Vempala. 2024. ULTRA: Unleash LLMs' potential for event argument extraction through hierarchical modeling and pairwise self-refinement. In *Findings of the Association for Computational Linguistics: ACL 2024*,

pages 8172–8185, Bangkok, Thailand. Association for Computational Linguistics.

Zikang Zhang, Wangjie You, Tianci Wu, Xinrui Wang, Juntao Li, and Min Zhang. 2025. A survey of generative information extraction. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4840–4870, Abu Dhabi, UAE. Association for Computational Linguistics.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623. Curran Associates, Inc.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

Appendix

A Error Analysis

Table 6 presents examples of generative models' prediction variations from the ground-truth arguments for a specific role. It also demonstrates how these argument pairs are evaluated under exact-match, relaxed-match, and complex-match evaluation schemes. For instance, consider the role `Cancer status` for which the ground-truth and predicted arguments are `'cancer has spread'` and `'metastasized'`. The SBERT similarity score of these two arguments is 0.1597. But based on the context (document, event, role), these two arguments should be considered a match, and both exact and relaxed-match fail to recognize this. This causes severe underestimation of model performance and unfair evaluation. As shown in the table, complex matching performs well in such cases. However, the complex match can sometimes fail to distinguish differences. Such as for the role `Place`, it incorrectly treated `'Chelsea, New York'` and `'Manhattan'` as similar. The complex match model might consider them as similar because Manhattan is part of New York, and Chelsea is part of Manhattan. We further discuss these errors in Section 3.3. The REGen framework hierarchically combines complex matching with exact, relaxed matching, and judgment alignment, helping capture true model performance. We illustrate the need for a context-grounded evaluation approach and the effectiveness of the REGen framework through examples in Table 7.

B Alignment Calculation Details

In the judgment alignment calculation (section 2.5) we count disagreement (N_d) only when the judge model marks a prediction as a match, but the human annotator marks it as a non-match. This choice helps mitigate issues of over-penalization and overestimation.

- **Over-penalization issue:** The precision and recall for argument evaluation are calculated differently than in standard classification tasks. Specifically, there are no true negative cases because the total number of negative arguments is unknown (Sharif et al., 2024; Sun et al., 2022). Instead, precision measures how many predicted arguments are correct, and recall measures how many ground-truth arguments were correctly identified by the model

Role	Ground-truth	Prediction [variations]	Exact Match	Relaxed Match	Complex Match
Date	20 April 2024	20 April 2024	M | ✓	M | ✓	M | ✓
		18 April 2024	NM | ✓	M | ✗	NM | ✓
		17 December 2022	NM | ✓	NM | ✓	NM | ✓
Cancer status	Cancer has spread	Patient cancer has spread	NM | ✗	M | ✓	M | ✓
		Metastasized	NM | ✗	NM | ✗	M | ✓
		Cancer spread from lung to liver	NM | ✗	NM | ✗	M | ✓
Occasion	July 4th	US Independence Day	NM | ✗	NM | ✗	M | ✓
		Fourth of July	NM | ✗	M | ✓	M | ✓
Participant	John Kerry	Senator John Kerry	NM | ✗	M | ✓	M | ✓
		US secretary of state in 2014	NM | ✗	NM | ✗	M | ✓
Drug dosage	6 mg	6 mg	M | ✓	M | ✓	M | ✓
		14 mg	NM ✓	M | ✗	M | ✗
		6 milligram	NM | ✗	NM | ✗	M | ✓
		take 6 mg drug	NM | ✗	NM | ✗	M | ✓
Duration	1 month	from July 30 to August 30	NM | ✗	NM | ✗	M | ✓
		30 days	NM | ✗	NM | ✗	M | ✓
		One month	NM | ✗	M | ✓	M | ✓
Medical condition	Chronic kidney disease	Long term kidney disease	NM | ✗	NM | ✗	M | ✓
		long term heart disease	NM | ✓	NM | ✓	NM | ✓
		CKD	NM | ✗	NM | ✗	NM | ✗
Causalities and losses	7 dead in central provinces flooding	18 died in central provinces due to flood	NM | ✓	M | ✗	NM | ✓
		Flooding in the central provinces killed seven people	NM | ✗	M | ✓	M | ✓
Place	Chelsea, New York	Chelsea neighborhood in NYC	NM | ✗	NM | ✗	M | ✓
		Chelsea, NYC	NM | ✗	M | ✓	M | ✓
		Manhattan	NM | ✓	NM | ✓	M | ✗

Table 6: Illustration of how predicted arguments may differ from ground-truth arguments for a specific role and how they are evaluated under different approaches—exact match, relaxed match, and complex match. Each cell shows whether the argument pair is classified as a match (M) or not a match (NM) under the respective method. Green checkmarks (✓) indicate correct evaluations, while red crosses (X) indicate errors. Our analysis shows that relying on a single evaluation method results in inaccurate assessments. REGen addresses this by systematically combining the strengths of each approach, enhancing evaluation accuracy while reducing computational cost.

(Sections 2.2–2.4). These two numbers can vary because multiple predicted arguments can be matched with a single ground-truth argument and vice versa. Because the score is computed only based on matches, any argument pairs predicted as non-matches do not contribute to the final score. Therefore, cases where the judge model predicts a non-match but a human annotator marks it as a match do not affect the evaluation score. Including such cases in the disagreement count will increase the deviation rate and would unfairly penalize the model's performance during JAM score calculation. To prevent over-penalization, we exclude these cases from our disagreement rate.

- **Ensuring no overestimation:** To obtain the most accurate evaluation score, all evaluations were made by the relaxed match model, and the complex match judge needs to be evaluated by humans. It is not feasible to conduct all the reevaluations manually. Deviation rate helps us obtain a reliable estimate of model performance. There is a trade-off – if relaxed

Role	Query	Ground-truth	Prediction	Observations
		Example 1		
I'm 46 years old and recently diagnosed with prostate cancer, which my doctor described as moderately aggressive. The doctor recommended robotic surgery using the da Vinci system due to its precision and faster recovery, but I'm hesitant because of possible long-term side effects. [...]. My cancer has already spread to nearby lymph nodes, which makes me feel like my options are shrinking. [..]				
`Treatment options`	What treatment options patient have?	da Vinci robotic surgery	robotic assisted surgery	Scattered arguments; Arguments refer to the same treatment options.
`Age`	What is the age of the patient?	46 years	46	Here, 46 refers to patient age. Exact and relaxed match approach fails to correlate.
`Cancer status`	What is patient cancer status?	Cancer has spread	Metastasized	Semantically similar arguments based on context despite lexical and syntactic differences.
		Example 2		
On March 15, a group of militants launched a coordinated assault on a military outpost in northern Mali. The surprise attack, believed to be conducted by a terrorist group linked to al-Qaeda (AQ) , involved several trucks and heavy gunfire . The attack resulted in the deaths of 17 military personnel . Government officials reported that the AQ seized control of the outpost before reinforcements could arrive.				
`Attack type`	What type of attack occurred?	coordinated assault	surprise attack	Example of subjective annotation. Both arguments can be accurate based on the role.
`Attacker`	Who carried out the attack?	al-Qaeda	AQ	Coreference: referring to same entity
`Attack Weapon`	What weapons were used in the attack?	several trucks and guns	trucks, guns	Difference in span boundary
`Causalities`	How many were killed?	17 military personnel	17 militants	Lexical variation

Table 7: Examples illustrating the effectiveness of REGen framework. In these cases, existing evaluation approaches (Exact match, Relaxed match, Head noun phrase match) fail to capture the semantic similarity between ground-truth and predicted arguments. For each example, representative argument roles are shown. Highlighted colors indicate the source segment of the ground-truth annotation.

and complex match criteria are too strict, we risk underestimation; if they are too lenient, we risk overestimation. To address this, we prioritize precision: when our framework says there is a match, we want to be confident it is indeed a true match. By factoring the score with the deviation rate, the final score we report is intentionally conservative — it serves as a lower bound on the model's true performance, avoiding inflated scores due to unverified matches.

C Dataset Details

We transform all datasets into a unified format as explained in Section 2.1. Each document includes a set of predefined roles based on the event, with each role having a list of ground-truth argument strings. In this work, we use a trigger-free approach for argument extraction (Tong et al., 2022). We adopt this formulation because many datasets lack trigger annotations or include implicit or scattered arguments that can not be tied to trigger phrases (Sharif et al., 2024). We use the official test split of all the datasets. Table 4 exhibits the detailed statistics of the datasets. We will release our processed datasets and associated scripts upon acceptance of the paper. Detailed descriptions of each dataset are provided in the following.

- **DiscourseEE** (Sharif et al., 2024) dataset is annotated from online health discussions and includes explicit, implicit, and scattered arguments. This dataset is hi-

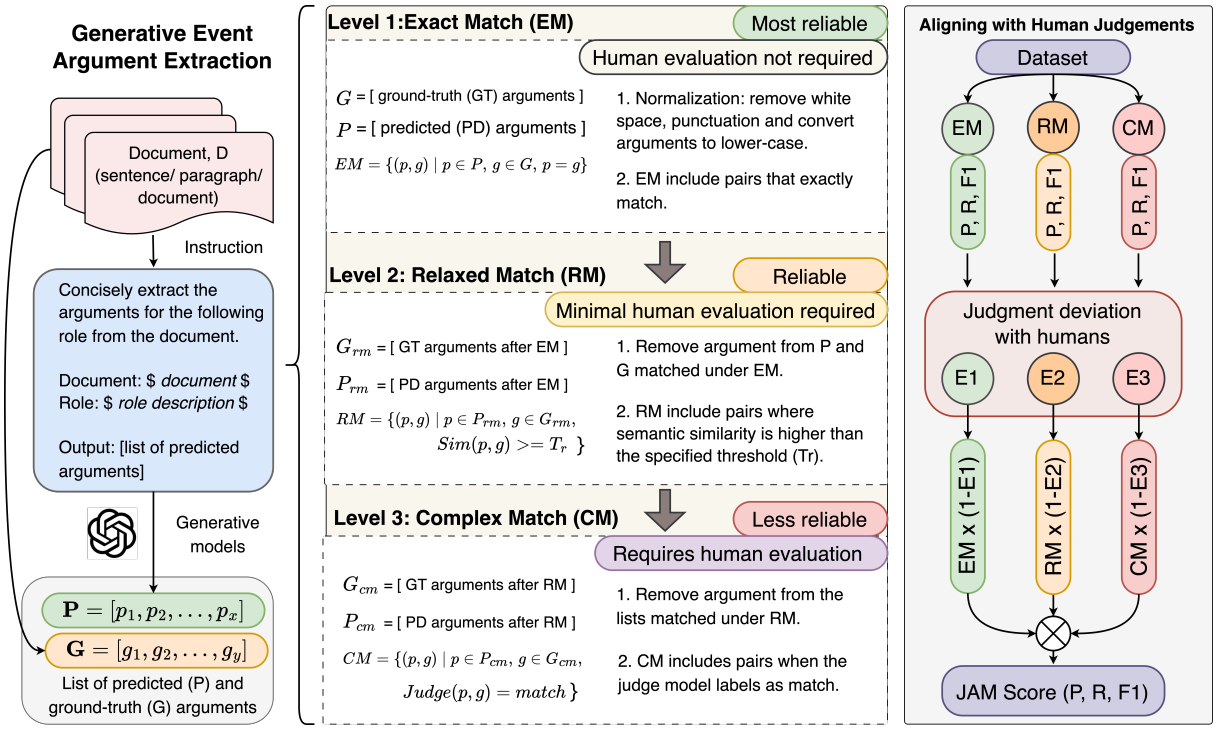

Figure 4: Illustration of REGen evaluation framework. ***Left:*** Example showing getting role-specific arguments from documents using generative models. ***Middle:*** Evaluation process at different levels: Exact Match, Relaxed Match, and Complex Match. ***Right:*** Judgment Aligned Match (JAM) score calculation process on a specific dataset, where E1, E2, and E3 represent the deviation rates from human judgments at different matching levels. Typically, E1 equals zero for a dataset, as an exact match indicates a perfect agreement with humans. *P, R,* and *F1* denote Precision, Recall, and F1-score, respectively. JAM score ensures that reported scores are reliable and not inflated due to misjudgments.

erarchical, with each role further classified into four types: core, type-specific, subject-specific, and effect-specific arguments. We sourced the test set from the official repository https://github.com/omar-sharif03/DiscourseEE. It features 34 unique roles across 3 event types, with all arguments annotated as strings. In this work, we do not use the argument types or hierarchical structure, as they are not essential. We process the dataset using the author's provided code.

- **PHEE** (Sun et al., 2022) is an event extraction dataset sourced from the pharmacovigilance domain. It contains 14 unique roles across 2 event types. We obtain the dataset from https://github.com/ ZhaoyueSun/PHEE. The dataset includes annotations for both trigger and argument spans. Following our formulation, we discard the trigger and only take the argument strings. We combine multiple arguments under the same role into a single argument list, separating them with semicolons.

- **RAMS** (Ebner et al., 2020) is an event extraction dataset from the news domain. We downloaded the dataset from https://nlp.jhu.edu/rams/ and processed it leveraging the script provided by TextEE (Huang et al., 2024). We ignored the trigger annotation and used the argument string to map the dataset into our formulation. The test set contains 129 unique events and 63 roles.

- **GENEVA** (Parekh et al., 2023) is a general-domain event extraction dataset developed using FrameNet. This dataset includes samples from books, articles, journals, and Wikipedia. We used the provided test set from https://github.com/PlusLabNLP/GENEVA. The test set includes 115 events and 196 unique roles. We applied the preprocessing script from TextEE (Huang et al., 2024) to convert the dataset to our format.

- **DocEE** (Tong et al., 2022) is a trigger-free document-level event extraction dataset with very long documents. We obtained the

test set from the official GitHub repository https://github.com/tongmeihan1995/DocEE. The official test set contains 2,771 documents. Due to high inference time, we selected 500 random samples to reduce complexity. Our test set includes 57 unique events and 266 unique roles.

- **WikiEvents**: (Li et al., 2021) is a document-level event extraction dataset based on Wikipedia texts. We sourced the dataset from https://github.com/raspberryice/gen-arg and processed it using the TextEE (Huang et al., 2024) preprocessing script. We retained only the argument annotations and discarded the rest. The test set includes 33 event types and 44 roles. WikiEvents is highly argument-dense compared to other datasets, with a density of 24.89. It also has longer documents, averaging 654 words per document.

D Model Details

Baselines: As baselines, we used transformer-based BERT (110M parameters) and instruction-fine-tuned FLAN-T5 (250M parameters) models. Both models were implemented using the HuggingFace pipeline and fine-tuned on the SQuAD (Rajpurkar et al., 2016) dataset. BERT was fine-tuned with a learning rate of 2×10^{-5}, a batch size of 8, and trained for 3 epochs. FLAN-T5 was fine-tuned for 4 epochs with a batch size of 16 and the same learning rate. During prediction, we provide a role-specific question and the associated document as context. The input is formatted as `[CLS] Question [SEP] Document [SEP]`. The output span is then decoded as the argument for the specific role. Arguments for each role are extracted independently.

LLMs: To investigate the feasibility of our proposed evaluation framework, we experimented with various LLMs used in previous studies on event argument extraction (Sharif et al., 2024; Lu et al., 2024). We evaluated open-source models ranging from 4B to 70B parameters and the closed-source GPT-4o model. This diverse selection allowed us to assess performance across different scales. We used five LLMs for the experimentation.

- **Phi-3.5** : We used the Phi-3.5-mini, a 3.8 billion parameter model trained on 3.3 trillion tokens (Microsoft, 2024). It achieves comparable performance to Mixtral 8x7B and GPT-3.5 models on academic benchmarks despite being a very small model.
- **Gemma-1.1** model trained on 6T tokens with novel RLHF method, based on the architecture and training recipe of Gemini models (Gemma-Team, 2024). It performs better than similar open-source models in 11 out of 18 text tasks. Gemma is available in two versions, with 2 billion and 7 billion parameters. We use the 7 billion parameter version for our experiments.
- **Mixtral (8x7B)** is a sparse mixture of expert language designed with an architecture similar to Mistral 7B (Jiang et al., 2024). It has a total of 47 billion parameters, with only 13 billion being active at a time. These architectural changes allow Mixtral to outperform models with more parameters (e.g., Llama-2, GPT-3.5) across several benchmarks.
- **Llama-3.1** is a state-of-the-art open-source language model pretrained and instruction-fined with 8B, 70B, and 405B parameters. It builds upon the Llama-3 model (Llama Team, 2024), incorporating grouped query attention (GQA) and RLHF. We use the 70B version of the model.
- **GPT-4o** (OpenAI, 2024) is one of the best-performing models that can reason across audio, vision, and text. It achieved state-of-the-art performance across most benchmarks[2].

We utilized the instruction-tuned versions of all the models. The HuggingFace inference strings for the open-source LLMs are Phi-3.5 (`microsoft/ Phi-3.5-mini-instruct`), Gemma-1.1 (`google/ gemma-1.1-7b-it`), Mixtral (`mistralai/ Mixtral-8x7B-Instruct-v0.1`), and Llama-3.1 (`meta-llama/ Llama-3.1-70B-Instruct`). We assess the performance of the GPT-4o model through API calls, using version (`gpt-4o-2024-11-20`).

E Judge Selection Process

There is a growing trend to leverage LLM as a judge to reduce the high cost of human evaluation (Zheng et al., 2023; Mao et al., 2024; Gu et al., 2025). Following this approach, we employed LLMs to mimic human evaluation and automatically determine whether the ground truth and

[2]https://lmarena.ai/

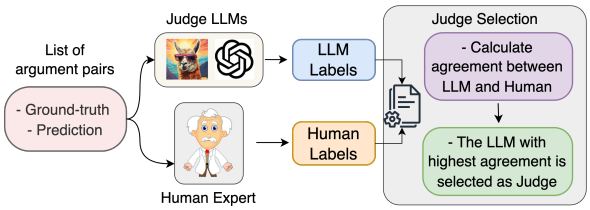

Figure 5: Schematic diagram of the judge selection process.

predicted arguments match. Figure 5 shows the schematic diagram of the judge selection process.

First, we create a judge dataset through manual annotation. We then use this dataset to experiment with multiple models and select the most suitable judge. Specifically, we evaluate four models: Llama3.1-70B, GPT-3.5, GPT-4o-mini, and GPT-4o. Each model is tested using both *zero-shot* and *chain-of-thought* prompts, with a single prompt uniformly applied across all the evaluated datasets and models. Figures 6 and 7 show the prompts for zero-shot and chain-of-thought, respectively. Table 8 presents the agreement rate between the judge models and human evaluations.

	GPT-3.5	GPT-4o-mini	Llama3-70B	GPT-4o
ZS	73.05	84.27	51.52	86.17
COT	79.91	68.11	73.95	78.39

Table 8: Agreement percentage of different LLMs with human judgments. ZS and COT indicate zero-shot and chain-of-thought prompting approaches, respectively.

While prompt optimization and alternative techniques (e.g., self-consistency) could further improve agreement, we refrain from such experiments due to the high cost and time requirements. Iterating to find optimal prompts for each model and dataset is impractical. These aspects, along with exploring the applicability of small fine-tuned judge models, are better suited for a separate study.

F Additional Results

F.1 Comparison with Head Noun Phrase Match Approach

Following prior work (Du and Cardie, 2020; Tong et al., 2022), we also evaluate model performance using the Head Noun Phrase Match (HM) approach for comprehensiveness. The results, shown in Table 10, indicate that on average models achieve 19% higher F1 score with REGen than the HM approach across all datasets. This is consistent

Dataset	Avg. EM-F1	Avg. RM-F1	Avg. CM-F1	Avg. REGen-F1 (JAM-F1)
DiscourseEE	10.74	15.11	41.01	37.45
PHEE	39.25	46.94	63.55	62.34
RAMS	13.38	16.14	29.45	28.27
GENEVA	13.62	25.32	48.18	46.12
DocEE	17.33	24.98	45.12	41.65
WikiEvents	8.93	11.05	33.56	31.02

Table 9: Comparison of average F1-scores of LLMs under different evaluation frameworks: Exact Match (EM), Relaxed Match (RM), Complex Match (CM), and REGen.

Datasets	Avg. HM-F1	Avg. REGen-F1	Δ F1	Gain (%)
DiscourseEE	13.09	37.45	+24.36	+186.09
PHEE	38.94	62.34	+23.40	+60.10
RAMS	16.45	28.27	+11.82	+71.85
GENEVA	25.50	46.12	+20.62	+80.86
DocEE	22.20	41.65	+19.45	+87.61
WikiEvents	16.53	31.02	+14.49	+87.66
	Avg. Δ F1		+19.02	

Table 10: Comparison of average F1-scores of the LLMs between Head Noun Phrase Match (HM) and REGen evaluation framework.

with the performance gain compared with the exact match approach (Table 3). We emphasize that, similar to exact and relaxed match strategies, the HM approach can result in inaccurate and misleading evaluations. This is because it only compares the head noun phrases in arguments, ignoring critical contextual information. For examples, for the role `date`, if the ground-truth is `'18 April 2024'` and predicted output is `'20 April 2018'`, the HM approach would consider them a match, as their noun phrase is 'April', despite being semantically and factually different. Moreover, HM fails to assess arguments that do not contain noun phrases at all, resulting in unreliable evaluations.

Datasets	DiscourseEE	PHEE	RAMS	GENEVA	DocEE	WikiEvents	
Inference count and reduction in inference only for zero-shot approach							Avg. Reduction (%)
#Inference (LLM as Judge)	1822	6215	3718	3897	12655	1852	
#Inference (REGen)	1201	2077	2318	2465	6513	1158	
Reduction count	621	4138	1400	1432	6142	694	
Reduction (%)	34.08	66.58	37.65	36.74	48.53	37.47	**43.51**
Inference count and reduction in inference only for only for chain-of-thought approach							
#Inference (LLM as Judge)	1740	5991	3124	3549	11511	1588	
#Inference (REGen)	1186	2359	2200	2435	6111	1089	
Reduction count	554	3632	924	1114	5400	499	
Reduction (%)	31.83	60.62	29.57	31.38	46.91	31.42	**38.62**
Total inference count and reduction in inference (zero-shot + chain-of-thought)							
#Inference (LLM as Judge)	3562	12206	6842	7446	24166	3440	
#Inference (REGen)	2387	4436	4518	4900	12624	2247	
Reduction count	1175	7770	2324	2546	11542	1193	
Reduction (%)	32.98	63.65	33.96	34.19	47.76	34.68	**41.20**

Table 11: Detailed comparison of inference counts and reductions when using LLMs as Judge versus the proposed REGen framework for the GPT-4o prediction model. Results for both zero-shot and chain-of-thought approaches are presented, illustrating total inference counts, achieved reductions, and corresponding percentage reductions across the evaluated datasets.

	Exact-Match			**Relaxed-Match**			**Complex-Match**			**JAM-Score**		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	6.3	5.52	5.88	9.28	8.12	8.66	35.17	32.1	33.56	31.65	28.84	30.18
Flan-T5	7.22	6.32	6.74	10.88	9.53	10.16	37.8	35.21	36.46	34.13	31.71	32.87
LLMs with Zero-Shot Prompt												
Phi-3.5	3.21	3.61	3.40	4.73	5.32	5.00	14.81	14.64	14.73	13.43	13.36	13.39
Gemma-1.1	10.45	13.74	11.87	13.96	18.36	15.86	45.61	55.67	50.14	41.31	50.58	45.48
Mixtral	10.59	17.15	13.10	14.37	23.17	17.74	41.82	57.97	48.59	38.07	53.19	44.38
Llama-3.1	11.05	16.95	13.38	15.49	23.67	18.73	37.52	51.96	43.57	34.47	48.02	40.13
GPT-4o	14.14	20.76	16.82	19.40	28.49	23.08	43.99	57.57	49.87	40.58	53.50	46.16
LLMs with Chain-of-thought Prompt												
Phi-3.5	5.53	9.83	7.08	9.76	17.05	12.42	34.26	52.36	41.41	30.89	47.47	37.43
Gemma-1.1	7.60	12.14	9.35	10.62	16.95	13.06	37.44	51.25	43.27	33.79	46.57	39.16
Mixtral	4.78	5.22	4.99	6.71	7.32	7.00	26.01	26.78	26.39	23.39	24.14	23.76
Llama-3.1	10.81	15.25	12.65	14.79	20.86	17.31	39.83	51.05	44.75	36.40	46.89	40.98
GPT-4o	12.64	17.75	14.77	17.86	25.08	20.86	42.71	53.06	47.33	39.27	49.15	43.66

Table 12: DiscourseEE evaluation results using REGen framework.

	Exact-Match			Relaxed-Match			Complex-Match			JAM-Score		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	29.09	26.6	27.78	36.62	33.48	34.98	54.88	50.53	52.61	53.55	49.28	51.33
Flan-T5	44.32	40.53	42.34	52.8	48.28	50.44	69.96	64.24	66.98	68.71	63.07	65.77
LLMs with Zero-Shot Prompt												
Phi-3.5	39.07	47.88	43.03	45.84	56.12	50.46	62.61	73.63	67.67	61.39	72.35	66.42
Gemma-1.1	44.10	45.94	45.00	53.23	55.49	54.34	75.91	77.99	76.93	74.25	76.35	75.28
Mixtral	33.68	40.02	36.58	39.16	46.59	42.55	55.39	63.55	59.19	54.20	62.31	57.98
Llama-3.1	36.49	42.29	39.17	43.77	50.63	46.95	61.53	66.60	63.96	60.23	65.43	62.72
GPT-4o	51.38	56.18	53.67	59.34	64.74	61.92	77.36	80.63	78.96	76.04	79.47	77.72
LLMs with Chain-of-thought Prompt												
Phi-3.5	30.57	33.76	32.09	36.24	39.96	38.01	52.68	55.45	54.03	51.48	54.32	52.86
Gemma-1.1	33.39	34.92	34.14	41.35	43.26	42.28	60.99	61.93	61.46	59.56	60.57	60.06
Mixtral	29.12	29.81	29.46	36.87	37.70	37.28	50.90	50.61	50.75	49.87	49.66	49.77
Llama-3.1	30.44	32.19	31.29	38.87	41.05	39.93	52.20	52.83	52.51	51.22	51.97	51.59
GPT-4o	46.91	49.43	48.14	54.27	57.13	55.66	69.68	70.34	70.01	68.56	69.37	68.96

Table 13: PHEE evaluation results using REGen framework.

	Exact-Match			Relaxed-Match			Complex-Match			JAM-Score		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	14.63	14.63	14.63	18.14	18.14	18.14	33.61	33.61	33.61	32.24	32.24	32.24
Flan-T5	12.61	12.61	12.61	15.13	15.13	15.13	28.62	28.62	28.62	27.43	27.43	27.43
LLMs with Zero-Shot Prompt												
Phi-3.5	13.75	17.35	15.34	16.07	20.27	17.92	31.23	37.77	34.19	29.90	36.22	32.76
Gemma-1.1	14.40	15.37	14.87	16.95	18.09	17.50	31.45	33.47	32.43	30.17	32.11	31.11
Mixtral	11.30	15.22	12.97	13.47	18.14	15.46	26.42	34.50	29.93	25.28	33.06	28.65
Llama-3.1	9.98	14.88	11.95	12.20	18.04	14.56	21.82	30.50	25.44	20.96	29.39	24.47
GPT-4o	15.01	27.58	19.44	17.89	32.82	23.15	29.91	49.98	37.42	28.84	48.43	36.15
LLMs with Chain-of-thought Prompt												
Phi-3.5	14.15	17.20	15.53	17.00	20.66	18.65	31.96	37.57	34.54	30.64	36.07	33.13
Gemma-1.1	10.17	11.32	10.71	12.88	14.34	13.57	25.13	27.53	26.28	24.04	26.36	25.15
Mixtral	6.08	7.86	6.85	7.34	9.49	8.28	15.33	19.08	17.00	14.63	18.23	16.23
Llama-3.1	9.28	12.51	10.66	11.12	14.98	12.76	21.21	26.89	23.72	20.32	25.83	22.75
GPT-4o	12.77	19.72	15.50	16.13	24.91	19.58	28.55	40.68	33.56	27.44	39.26	32.30

Table 14: RAMS evaluation results using REGen framework.

	Exact-Match			Relaxed-Match			Complex-Match			JAM-Score		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	15.81	14.72	15.24	27.56	25.67	26.58	54.64	51.62	53.09	52.24	49.33	50.74
Flan-T5	19.02	17.71	18.34	32.0	29.79	30.85	59.63	56.01	57.76	57.16	53.67	55.36
LLMs with Zero-Shot Prompt												
Phi-3.5	12.21	14.36	13.20	23.56	27.68	25.46	47.65	52.14	49.80	45.50	49.92	47.61
Gemma-1.1	11.75	11.63	11.69	24.52	24.27	24.40	51.41	50.06	50.73	49.01	47.75	48.37
Mixtral	12.20	14.65	13.31	22.80	27.32	24.86	45.98	51.17	48.44	43.92	49.01	46.32
Llama-3.1	15.04	17.93	16.36	27.24	32.46	29.62	52.74	57.67	55.09	50.45	55.36	52.79
GPT-4o	17.72	20.86	19.16	30.86	36.29	33.35	56.25	60.49	58.30	53.96	58.25	56.02
LLMs with Chain-of-thought Prompt												
Phi-3.5	9.98	11.66	10.76	20.22	23.55	21.76	44.22	48.60	46.31	42.09	46.36	44.12
Gemma-1.1	9.21	9.55	9.38	20.83	21.60	21.21	45.69	45.32	45.51	43.47	43.18	43.33
Mixtral	15.39	17.48	16.37	25.17	28.59	26.77	45.08	49.25	47.07	43.29	47.38	45.24
Llama-3.1	9.29	9.65	9.46	16.98	17.61	17.29	33.15	31.03	32.05	31.70	29.79	30.72
GPT-4o	16.00	17.12	16.54	27.57	29.50	28.50	48.50	48.47	48.48	46.59	46.71	46.65

Table 15: GENEVA evaluation results using REGen framework.

	Exact-Match			Relaxed-Match			Complex-Match			JAM-Score		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	22.93	15.73	18.66	31.59	21.72	25.74	57.09	41.12	47.81	52.73	37.82	44.05
Flan-T5	22.8	15.64	18.55	30.62	21.08	24.97	54.35	38.98	45.4	50.29	35.94	41.92
LLMs with Zero-Shot Prompt												
Phi-3.5	14.16	14.36	14.26	19.86	20.04	19.95	39.18	37.62	38.39	35.90	34.62	35.25
Gemma-1.1	20.47	16.04	17.99	30.38	23.95	26.78	51.92	42.54	46.77	48.15	39.31	43.28
Mixtral	21.84	24.09	22.91	31.06	34.20	32.55	56.50	59.92	58.16	52.12	55.47	53.74
Llama-3.1	15.03	21.11	17.56	21.65	29.97	25.14	42.51	51.17	46.44	38.95	47.49	42.80
GPT-4o	16.82	31.42	21.91	24.45	44.89	31.65	45.79	73.41	56.40	42.12	68.41	52.14
LLMs with Chain-of-thought Prompt												
Phi-3.5	18.95	20.82	19.84	26.57	29.13	27.79	46.94	50.19	48.51	43.43	46.55	44.93
Gemma-1.1	10.90	9.01	9.87	16.20	13.47	14.71	28.44	23.98	26.02	26.30	22.15	24.05
Mixtral	7.84	6.17	6.90	10.93	8.60	9.63	22.38	16.94	19.28	20.44	15.53	17.65
Llama-3.1	16.81	24.07	19.79	25.26	35.97	29.68	48.02	61.66	53.99	44.10	57.15	49.78
GPT-4o	17.78	29.77	22.26	25.56	42.48	31.92	47.90	71.18	57.27	44.07	66.17	52.90

Table 16: DocEE evaluation results using REGen framework.

	Exact-Match			Relaxed-Match			Complex-Match			JAM-Score		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
Baselines												
BERT	9.62	4.86	6.46	14.23	7.19	9.55	38.91	23.68	29.44	36.12	21.82	27.2
Flan-T5	13.81	6.98	9.27	17.57	8.88	11.8	39.33	23.47	29.4	36.87	21.82	27.41
LLMs with Zero-Shot Prompt												
Phi-3.5	9.80	8.46	9.08	11.76	10.15	10.90	36.76	32.56	34.53	33.94	30.03	31.86
Gemma-1.1	8.65	4.86	6.22	10.15	5.71	7.31	40.98	29.60	34.37	37.49	26.90	31.32
Mixtral	10.55	9.30	9.89	13.19	11.63	12.36	39.09	37.42	38.24	36.16	34.51	35.31
Llama-3.1	11.06	15.22	12.81	13.36	18.39	15.48	33.95	45.67	38.94	31.62	42.58	36.29
GPT-4o	11.47	17.34	13.80	14.13	21.35	17.00	35.10	51.80	41.85	32.73	48.36	39.04
LLMs with Chain-of-thought Prompt												
Phi-3.5	7.82	6.13	6.87	9.70	7.61	8.53	35.04	29.39	31.97	32.18	26.93	29.32
Gemma-1.1	4.53	2.54	3.25	6.42	3.59	4.61	22.26	14.59	17.63	20.47	13.35	16.16
Mixtral	5.96	3.59	4.49	8.07	4.86	6.07	24.56	16.28	19.58	22.70	14.99	18.06
Llama-3.1	10.78	10.78	10.78	13.11	13.11	13.11	35.94	37.21	36.56	33.36	34.49	33.91
GPT-4o	10.77	13.95	12.15	13.38	17.34	15.10	37.36	47.78	41.93	34.65	44.34	38.90

Table 17: WikiEvents evaluation results using REGen framework.

Dataset	Event	Role	Question
DiscourseEE	Taking-MOUD	Treatment	What treatments the subject/patient prescribed or undergoing?
	Tapering	Side effects	What are the side effects the subject is experiencing or expects to experience?
	Return to Usage	Intervention	What measures are taken to address or reduce side effects?
	Taking-MOUD	Dosage	What is the current or previous dosage of the Medications?
	Tapering	Age	What is the age of the subject/patient?
	Return to Usage	Conditions	What are the Pre-existing or co-morbid conditions of the subject/patient?
PHEE	Potential therapeutic event	Treatment	What is the therapy administered to the patients?
	Adverse event	Treatment drug	Whare the the drugs used as therapy in the event?
	Potential therapeutic event	Treatment dosage	What is the amount of drug is given?
	Adverse event	Treatment route	What is the route of drug administration?
	Adverse event	Effect	What are the outcomes or side effects of the treatments?
	Potential therapeutic event	Treatment disorder	What is the target disorder of the medicine administration?
RAMS	Artifactexistence	Place	Where does this event occur?
	Transaction	Artifact	What artifact is involved?
	Contact.commandorder	Communicator	Who is the communicator?
	Movement transportartifact	Origin	Where does the movement originate?
	Conflict.yield.retreat	Retreater	Who is the retreater?
	transaction transferownership	Recipient	Who is the recipient?
GENEVA	Statement	Message	What is the message?
	Collaboration	Partners	Who are the partners in this collaboration?
	Supply	Supplier	Who is the supplier?
	Protest	Content	What is the content of the protest?
	Killing	Victim	Who is the victim?
	Research	Topic	What is the research topic?
DocEE	Riot	Location	Where did the riot occur?
	Regime change	Date	When did the change happen?
	Earthquakes	Affected area	Which area was affected by the earthquake?
	Military exercise	Scale	What was the scale of the exercise?
	Diplomatic talks	Participants	Who are the participants?
	Fire	Location	Where did the fire take place?
WikiEvents	Conflict.attack	Instrument	What instrument is used?
	Life.die	Place	Where did the death occur?
	Conflict.detonateexplode	Target	Who or what is the target?
	Movement.transportation	Transporter	Who is the transporter?
	Justice.chargeindict	Defendant	Who is the defendant?
	Transaction	Acquired entity	What entity is being acquired?

Table 18: Details of the argument roles for each event type in the evaluated datasets. Note: for RAMS and WikiEvents datasets some event names are very long. For presentation convenience we use the first part of the event name.

Zero-Shot Judge Selection Prompt

Instruction
Find whether text-1 and text-2 are semantically similar or not based on the context provided.

Context ## \\ Document from where argument extracted
{context}

Texts
text-1: {x} \\ Predicted argument
text-2: {y} \\ Ground-truth argument

Are text-1 and text-2 semantically similar even though they are structurally different? Return "yes" if they are similar and "no" otherwise. Do not provide any extra description.

Figure 6: Zero-shot judge selection prompt

Chain-of-thought Judge Selection Prompt

Determine whether text-1 and text-2 are semantically similar based on the context provided.

Follow these steps to arrive at the answer:
1. Analyze the context and identify the key elements or criteria for semantic similarity.
2. Compare text-1 and text-2 against the key elements from the context.
3. Decide if text-1 and text-2 convey the same meaning even if they are structurally different. We will also consider partial matches with overlapping meanings as similar.

Context ## \\ Document from where argument extracted
{context}

Texts ## \\ Predicted argument
text-1: {x} \\ Ground-truth argument
text-2: {y}

Example
Context
The context is about describing weather conditions.

Texts
text-1: "It's sunny and warm outside."
text-2: "The sun is shining, and it feels warm."

Reasoning
1. Key elements from the context: Describing weather conditions involves mentions of sun, warmth, or similar indicators.
2. Comparing text-1 and text-2: Both mention sunny conditions and warmth, using slightly different phrasing.
3. Conclusion: The texts are semantically similar since they convey the same meaning about the weather.

Answer
yes

Task
Using the above steps, find whether text-1 and text-2 are similar. Return "yes" if they are similar and "no" otherwise.

Your Turn
After reasoning, provides the final output in JSON.
'output': 'yes or no'

Figure 7: Chain-of-thought judge selection prompt

Zero-Shot Argument Extraction Prompt

Instruction
Concisely extract the arguments for the following role from the document. Return 'null' if any argument is not present for a role. Separate multiple arguments of role values by a semicolon (;).

Role Question
{role}: {role_question}

Document
{document}

Extract and return the arguments for the role in the JSON format.

Figure 8: Zero-shot event argument extraction prompt

Chain-of-thought Argument Extraction Prompt

Concisely extract the arguments for the following role from the document.

Follow these steps:

1. Understand the role and the role-specific question.
2. Analyze the document to identify spans that answer the question.
3. Extract relevant arguments, separating multiple arguments with semicolons (;). Return 'null' if any argument is not present for a role. Do not overgenerate arguments; be thoughtful and precise.
4. Return the result in JSON format.

Role Question
{role}: {role_question}

Document
{document}

Example
Responsibility: What are the responsibilities of an Event Planner?
Document: "The Event Planner is responsible for booking venues, coordinating schedules with vendors, and managing budgets to ensure successful events."

JSON
{{"Responsibility": "booking venues; coordinating schedules with vendors; managing budgets"}}

Extract and return the arguments for the role in the JSON format.

Figure 9: Chain-of-thought event argument extraction prompt