ADAPTIVE IE: Investigating the Complementarity of Human-AI Collaboration to Adaptively Extract Information on-the-fly

Ishani Mondal,[1] **Michelle Yuan**,[6] **Anandhavelu Natarajan,**[3] **Aparna Garimella,**[2]
Francis Ferraro,[4] **Andrew Blair-Stanek**,[4] **Benjamin Van Durme**,[5] **Jordan Boyd-Graber**[1]

[1] University of Maryland, College Park, [2] Adobe Research , [3] Aqxle.ai
[4] University of Maryland, Baltimore County, [5] John Hopkins University, [6] Amazon, NY,
imondal@umd.edu

Abstract

Learning template-based information extraction (IE) from documents is a crucial yet difficult task. Prior template-based IE approaches assume foreknowledge of the domain's templates. However, many real-world IE scenarios do not have pre-defined schemas. Despite this, existing IE systems are either fully supervised, requiring expensive human annotations, or fully unsupervised, extracting information that often do not cater to user's needs. To address these issues, we formally introduce the task of "IE on-the-fly", and address the problem using our proposed ADAPTIVE IE framework that uses human-in-the-loop refinement to adapt to changing user questions. Through human experiments on three diverse datasets, we show that ADAPTIVE IE is a *domain-agnostic*, *responsive*, *efficient* framework for helping users *access customized and tailored information* while quickly reorganizing information in response to evolving information needs.

1 Introduction

The goal of IE is to extract structured insights from unstructured data based on a fixed schema. Existing tools help analyze patterns (Li et al., 2022; Móra et al., 2009; Chinchor and Marsh, 1998; Pavlick et al., 2016), but in a dynamic real-world situation, information needs often shift and are subjective, making predefined schemas impractical for use. In Figure 1, after the 2014 Chile Earthquake, the Disaster Emergency Response Team might seek information on safe zones and transportation, while geological teams might look for "number of people trapped", "number of buildings damaged" from the corpus. In such cases, unsupervised IE is ideal for extracting relevant information on-the-fly, catering to evolving user needs.

Recent unsupervised IE systems (Aharoni and Goldberg, 2020; Yu et al., 2022a) often fail to discern specific user needs without clear guidance, potentially overgeneralizing and including non-essential information. For instance, emergency teams may receive broad details on slightly damaged areas (Figure 1) instead of critical information on safe routes for immediate response. While unsupervised approaches (Chambers, 2013; Cheung et al., 2013; Bamman and Smith, 2014; Ferraro and Van Durme, 2016) and template-driven QA methods (Li et al., 2022; Móra et al., 2009) are prevalent, their extraction accuracy (by mapping to desired slots) is quite low. On the other hand, supervised IE systems which require pre-defined schema template annotations to train a model (such as whether or not we need to extract "safety routes" or "casualties" need to be predefined) (Chinchor and Marsh, 1998; Pavlick et al., 2016), are impractical for real-world applications. A minimally supervised system would be preferable in this scenario, which can offer *enhanced accuracy over unsupervised methods* and the ability to *quickly adapt to varying user needs* (map to user-desired slots like "Casualties" or "Damaged Properties" in Fig. 1).

To address these gaps, we make the following contributions: **[1]** First, we introduce the concept of "*IE-on-the-fly*", a dynamic approach that adapts to user-specific information needs (Overview in Figure 1), formally defined in Section 2. **[2]** Second, we define it through on-the-fly schema induction, which involves generating question-answer pairs from a corpus, as questions effectively encapsulate information needs. These pairs are then clustered to identify unique information demands. However, these unsupervised clusters may not fully meet user needs (Step A in Figure 1). **[3]** Third, we propose the idea of understanding user requirements through these cluster modifications, hypothesizing the fact that user eventually groups or wants to group information which they are interested in (Step B). We introduce an interactive "human-in-the-loop" system, ADAPTIVE IE, that takes the initial clusters from Step A (at the first stage else

Proceedings of the 31st International Conference on Computational Linguistics, pages 5870–5889
January 19–24, 2025.

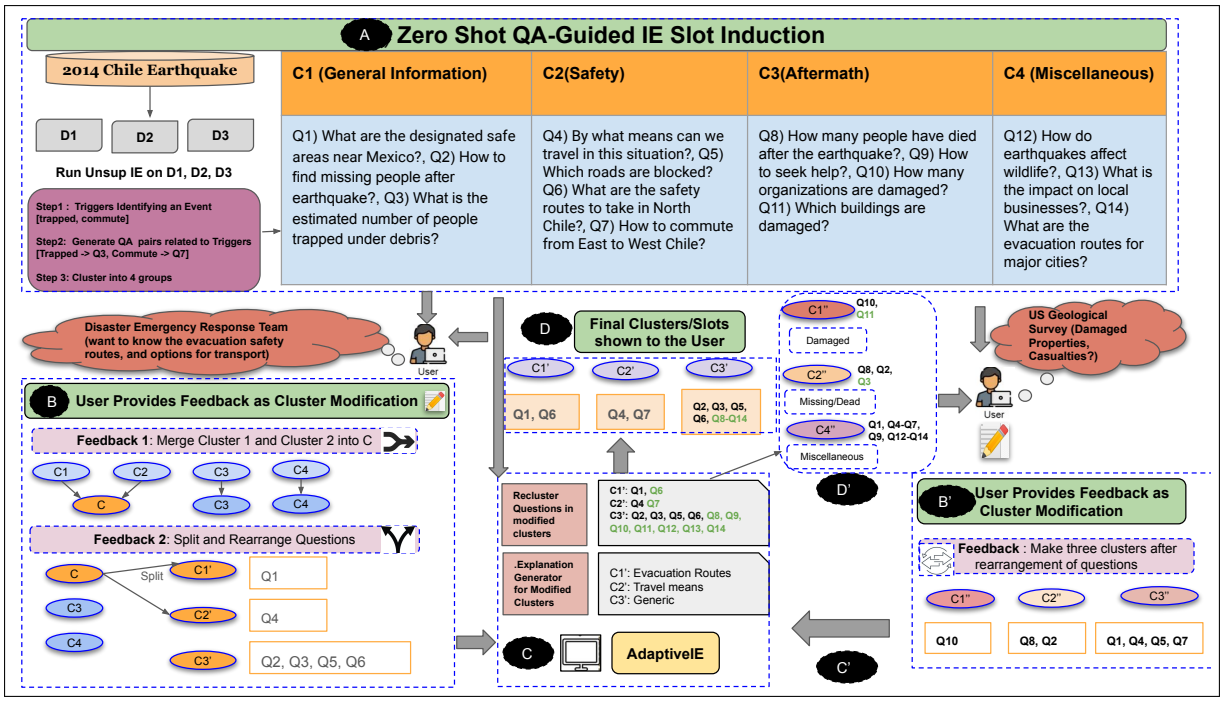

Figure 1: ADAPTIVE IE allows users to refine clusters based on their specific needs, as demonstrated by User A and User B's feedback on the initial clusters produced by UnsupervisedIE. Based on user feedback, our system dynamically updates and reclusters the information, ensuring that it is tailored to each user.

output from previous iteration), and user feedback from Step B as input, reclusters all other questions from the corpus, then explains the intuitions behind each cluster by providing the updated set of clusters in (Step C). The user can again glean on the next set of clusters (Step D) and further tweak or modify if they are not satisfied fully in the next iteration. This iterative loop (steps B, C, D) continues until the extracted information meets the user's expectations, illustrating a practical application of adaptive information extraction in emergency response scenarios. **[4]** Finally, we conducted human experiments on three datasets and demonstrated that ADAPTIVE IE significantly improved F1 score for extracted information compared to unsupervised methods within 30 minutes, highlighting its adaptability. This is practically usable when there are insufficient human annotations to train a supervised IE system. Any user can quickly obtain desired information with minimal interactions with the system.

2 Task Description and Formulation

The task "IE on-the-fly" involves dynamically adjusting information extraction processes based on real-time user feedback to meet evolving information needs. For instance, during the 2014 Chile Earthquake, this involved reorganizing queries, such as merging "Damaged Properties" and "Casualties" for the USGS, or combining "Evacuation Routes" and "Travel means" for emergency teams (Figure 1). This approach ensures that information is more accurately aligned with the immediate requirements of users during critical events.

We formulate this task as a slot filling (Louvan and Magnini, 2020) task that involves extracting and assigning specific values (slots) from unstructured input data, where each slot type represents a unique information need. We explain the key terminologies that will be frequently referred to in the paper using Figure 1 as an example:

[1] IE Template Schema: A predefined structure for extracting information. For example, in the figure, templates include information that may be of interest to the Disaster Emergency Response Team (User A) or the US Geological Survey (User B).

[2] Questions: The goal of asking a question nicely intersects with the definition of an information need (Srihari and Li, 2000). For example, in Figure 1, the goal is to extract as much information as possible about the 2014 Chile earthquake event, such as Q1, Q2. So we represent the user needs as questions / queries.

[3] Clusters/IE Slot Type: Groups of related questions form a cluster or slot type, e.g., *Cluster 1* focuses on *General Information* (Q1: Safe areas,

Q2: Missing people), while *Cluster 3* covers the *Aftermath* (Q8: Death toll, Q10: Damaged buildings). The answers to the questions fill the values for the slots. The concept of an IE template aligns closely with the slot-filling task in information extraction. In this context, each slot type corresponds to a specific question, while the values filled in these slots represent the answers. For example, a slot might be defined to extract information about "safety routes", and the corresponding slot filler would be the actual route details provided in response to an event. We treat the slots as questions/queries, and the slot fillers as answers to those questions (slots).

3 QA-Guided UnsupervisedIE

We propose a QA-guided UnsupervisedIE approach to dynamically create and populate IE slots, adapting to evolving informational needs without relying on pre-labeled training data. This process is demonstrated in Step A of Figure 1 (2014 Chile earthquake). The goal of this approach is to efficiently organize large volumes of unstructured data into coherent clusters, enabling rapid and accurate information retrieval and question answering. This method uses the QA format for schema and slot induction, aligning with human cognition to make complex IE actionable and enhance decision-making in crises. The steps are:

Event Trigger Identification. From all the documents, we extract the trigger words that describe the occurrence of events (Prompt LLM_t to extract the most important triggers $T = t_1, ...t_n$) from each document (Prompt in Appendix B)). In Figure 1, "trapped", "commute" are examples of event triggers generated from input documents D_1, D_2, D_3.

Question-Answer (QA) Pair Generation. Given a document d and set of triggers $T = tr_1, ...tr_n$, we generate "WH"-type questions by prompting LLM_{QA} such that they contain one of the triggers tr_i whose answer is a continuous span in d. Our questions answer about Who, Whom, What, When, Where, Why, How of an event (Prompt in Appendix B). For example, Q3, Q7 are the corresponding questions generated for the triggers "trapped" and "commute".[1]

[1]Note: To mitigate hallucination concerns, we optimized prompts using 100 examples for trigger identification and QA pair generation. After finalizing, 200 additional samples were tested, confirming the outputs were accurate and free of hallucinations, ensuring factual consistency.

Clustering with Explanations. The generated questions and their corresponding answers are grouped into K clusters, where each cluster represents a distinct information need. For each cluster, the questions corresponding to its centroid are selected to prompt $LLM_{Cluster}$ to generate an explanation for why the questions in that cluster (names of clusters viz. "General information", "Safety", "Aftermath", "Miscellaneous" are generated as explanations) (Prompt in Appendix B). This step helps the users in assessing the coherence information within the generated clusters.

4 ADAPTIVE IE Methodology

QA-Guided UnsupervisedIE often produces generalized outputs that do not address the user's actual information needs (Output of Step A in Figure 1). For this, we introduce an interactive human-in-the-loop system, ADAPTIVE IE (explained in Algorithm 1), that takes two inputs: 1) initial clusters from Step A (Section 3)) at the first stage or output from previous iteration, and 2) user feedback from Step B (Section 4.1), then adjusts other information in the schema and finally show the updated set of clusters (Step C) to users in the next iteration (Section 4.2). The user can then glean on these clusters (Step D) and further modify if they are not satisfied fully with their need. This iterative loop (steps B, C, D) continues until the extracted information meets the user's needs.

4.1 User Feedback as input (Step B):

Let $\mathbf{Q} = \{q_1, q_2, \ldots, q_n\}$ be the set of questions to be clustered, $\mathbf{C}(t) = \{C_1(t), C_2(t), \ldots, C_k(t)\}$ be the set of clusters at iteration t, where each $C_j(t)$ contains questions grouped by semantic roles (C1-C4 in **Step A** are the initial slots generated by the Zero-Shot QA-Guided approach with the slots being mapped by answers of questions from Q1-Q14 in Figure 1). We define $\mathbf{F}(t)$ as the user feedback at iteration t. A user, driven by specific information needs, initially attempts to categorize the available information by leveraging the predefined clusters to group questions that are likely to fulfill their requirements (as depicted in Figure. 1, where the Disaster Emergency Team adjusts clusters C1, C2, C3, and C4 in **Step B**). $\mathbf{F}(t)$ can be implemented in three distinct manners:

A) Merge Clusters (≻): This feedback is used when two or more clusters are semantically related and can be combined to form a single, cohesive

cluster. In Figure 8, the user merges Cluster 1 (General Information) and Cluster 2 (Safety) into a single cluster C because both the clusters deal with the broader context of safety and general information relevant to an emergency situation. If C_1 and C_2 are two clusters, the merge operation can be represented as $C = C_1 \cup C_2$.

B) Split Clusters (⛙): This feedback is used when a cluster is too broad and needs to be split into more focused and fine-grained subtopics. In Figure 8, Cluster C is split into C1', C2' where each new cluster contains questions more tightly grouped by subtopic, such as evacuation routes and travel means. If C is an original cluster, splitting it into two can be denoted by $C \rightarrow \{C1', C2'\}$, where each C' represents a subset of C such that $C1' \cup C2' = C$ and $C1' \cap C2' = \emptyset$.

C) Rearrange Questions (⇵): The questions need to be reassigned to ensure they fit well into the correct clusters, maintaining the contextual alignment. Let Q be a set of all questions, and $Q_i \subset Q$ a subset of questions originally in a cluster C_i. If questions are reallocated such that Q_i now belongs to a new cluster C_j, this can be denoted as $Q_i \rightarrow C_j$.

4.2 Reclustering in Adaptive IE (Step C):

Once feedback $\mathbf{F}(t)$ is provided in Step B, the system utilizes an update function u to refine the clusters. The update function is defined as $u : (\mathbf{C}(t), \mathbf{F}(t)) \rightarrow \mathbf{C}(t+1)$, where feedback is incorporated to update the clusters for the next iteration, yielding $\mathbf{C}(t+1)$. The clustering process adheres to the following principles:

1) Recluster-Rename (Rec-Ren) In this approach, questions are initially reorganized based on user feedback, followed by renaming clusters to reflect their updated content. User feedback incorporates two specific types of constraints to guide the reclustering process: a) **Must-have constraints ($M \subseteq Q \times Q$)**: It specifies pairs of questions that must be included in the same cluster. **b) Cannot-have constraints ($N \subseteq Q \times Q$)**: Specifies pairs of questions that must not be included in the same cluster. Once the questions are reorganized, we use LLM to generate meaningful names for each cluster by analyzing the questions closest to the cluster centroid. This naming step offers users the flexibility to review, edit, and further refine the clusters.

2) Rename-Recluster (Ren-Rec) We initially use LLM to generate meaningful names for each cluster by analyzing the representative questions located nearest to the cluster centroids. Next, questions are reassigned to clusters based on the semantic alignment of their content with the cluster names. Let N_i represent the name of cluster C_i, and $e(N_i)$ its embedding. For any question $q_j \in Q$, its embedding is denoted as $e(q_j)$. The assignment of q_j to a cluster C_i is determined by maximizing the cosine similarity between $e(q_j)$ and $e(N_i)$:

$$\text{assign}(q_j) = \arg\max_i \cos(e(q_j), e(N_i))$$

where cos denotes the cosine similarity.

4.3 Users decide the next steps (Step D)

In this step, the user concludes if the clustering configuration aligns with his objectives. For example, as illustrated in Figure 6, after each iteration, humans evaluate whether the question-answer pairs correctly segregate concepts like "Increased Effect" from "Decrease" based on semantic content. Otherwise, he proceeds to next iteration with $t = t + 1$ and the steps B, C and D get repeated.

Evaluation. In evaluating slot mapping performance, each iteration assesses the alignment of slots with user-defined needs. We define the user need using a subset of slots from the gold standard dataset, serving as a proxy for the specific information the user seeks.

In Figure 6, the user need is represented by two slots: downregulation and upregulation. At each iteration, we assume that fulfilling these two slots from the gold standard will meet the user's requirements. Initially, the clusters generated by UnsupervisedIE do not coherently reflect the user's desired information. A fuzzy mapping is first applied to map clusters to the required slots (upregulation and downregulation), but none could be correctly mapped in the initial iteration. After receiving feedback from the user, the system splits and regroups questions, allowing ADAPTIVE IE to better organize the information. Now, Cluster 1 represents upregulation (e.g., cholesterol upregulating SREBP-1c), while Cluster 3 represents downregulation (e.g., ATM decreasing the effect of caffeine). At each iteration, the goal is to match the information in the gold standard (user's requirement) with the corresponding answers in the predicted slots. For instance, "Cholesterol [Upregulates] SREBP-1c" can be accurately mapped to Q7 and its answer, confirming that the user's needs are being met through the iterative feedback process.

5 Datasets and Baselines

Datasets. We conduct experiments on the three following datasets from diverse domains to test the generalizabilty of our approach:
(1) GENEVA (Parekh et al., 2023) is a generic-domain Event Extraction dataset comprising of 179 event types and 362 argument roles.
(2) Biomedical Slot Filling (Papanikolaou et al., 2022) comprises of different relation types betwee the biomedical entities, out of which we evaluate on 1000 passages containing the most-occuring relations (interacts with, downregulation, upregulation, cause and regulation) between biomedical entities.
(3) CrisisNLP (Imran et al., 2016) is a classification dataset comprising of crisis-related tweets between 2013 and 2015. We repurpose this dataset to create a slot filling dataset for emergency domain. Using GPT-4, we initially identified precise information from each tweet, ensuring it matched predetermined categories. For instance, in "Emergency Aids" category, we focused on extracting specific details like locations of emergency and availability of emergency supplies, organizing this information into slot-value pairs. Manual examination was conducted to guarantee the accuracy of slots, which involved removing entries that were not relevant, finally creating a dataset comprising 3,000 tweets from Chile Earthquake, Ebola Outbreak, Typhoon and 6,940 slot-value pairs (Appendix D).

Baselines. We compare **UnsupervisedIE** with the following baselines:
(1) BERTQA by Du and Cardie (2020): Based on BERT, it enhances label semantics through a QA objective. It scales to a broad range of argument roles by posing questions in the format "What is arg-name?" for each specific role,
(2) TE (Transfer Entailment) by Lyu et al. (2021): A zero-shot transfer model that leverages a pre-trained entailment model to autonomously extract events. Similar to BERTQA, it crafts hypothesis questions like "What is arg-name?" for every role, facilitating direct comparison.
(3) OpenIE by Angeli et al. (2015): A triple-extraction baseline that extracts open-domain relation triples, representing a subject, a relation, and object of the relation,
(4) PromptORE by Genest et al. (2022): It extracts trigger words surrounding the context, followed by clustering and slot mapping. However, our methods do not rely on heuristics to find trigger words between two or more entities in sentences, instead consider the overall context to ask questions conditioned on the tagged entities.
(5) Span-Extraction Method by Yu et al. (2022a): It comprises of bottomup span extraction method regularized by unsupervised probabilistic context-free grammar (PCFG), followed by clustering. Furthermore, we experiment IE-on-the-fly using zero-shot and few-shot prompting of GPT-3 (*text-davinci-003*), ChatGPT (*gpt-3.5-turbo*), GPT-4 to extract information in an unsupervised way. Our implementation and hyperparameter details are in Appendix A and prompts are in Appendix B.

6 Human Experimental Setup

Our primary objective is to evaluate whether our proposed system, ADAPTIVE IE, can effectively assist the users in extracting essential information from large datasets during emergency situations. So we explore three main research questions:

- **RQ1)** How does ADAPTIVE IE compare to both manual methods and automatic unsupervised approaches in effectiveness in extracting the desired information? (Section 6.1)
- **RQ2)** How easily can multiple individuals with different information needs engage with ADAPTIVE IE to extract desired information? (Section 6.2)
- **RQ3)** How quickly can an individual engage with ADAPTIVE IE to extract desired information as their information needs change over time? (Section 6.3)

6.1 RQ1: Testing the Effectiveness

We aim to assess how well ADAPTIVE IE assists users in identifying and refining relevant question clusters to suit their specific needs (Appendix E).

We recruited ten participants via Upwork to evaluate the effectiveness of our ADAPTIVE IE system (Figure 9). All the participants were not previously exposed to this task and interface. To help them become familiar, they were first asked to read 50 questions, answers and mapped slots (Appendix C). Initially, the participants were first tasked with examining initial clusters of questions generated from 400 documents of biomedical dataset. Next, they interacted with the ADAPTIVE IE interface to organize information regarding potential side effects of biomedical entities, which were categorized under the "Cause-Effect","upregulation" and "downregulation" slot types. In the second part of the study, participants used 1000 documents from CrisisNLP

Dataset to identify 'affected areas", "casualties" and "recovery rates" from documents concerning the Ebola outbreak, and "Damaged Properties" and statistics on "missing or dead people" from the 2014 Chile Earthquake corpus. We initially calculated the F1-scores for these slots before any user interaction (only with unpervised approach). The study then measured how user feedback improved F1-scores within a 30-minute window. We compare different experimental configurations in time taken to improve slot mapping, as measured by F1-score progression. The experimental mechanism includes the following approaches: (1) the Control Group, which involves manual grouping of information into predefined clusters based on a set of questions generated from documents; (2) an Experimental Group consisting of three sub-configurations that start from unsupervised clustering methods, specifically, the approaches proposed by Yu et al. (2022a) and Genest et al. (2022), both of which uses Rec-Ren approach. Finally, additional configurations apply Rec-Ren or Ren-Rec strategy using K-Means and HDBScan clustering techniques. Genest et al. (2022) and Yu et al. (2022a)'s approach use Rec-Ren process for clustering and updating slot names to enhance information grouping, while K-Means and HDBScan variants use both Rec-Ren and Ren-Rec methods to optimize slot mappings iteratively.

6.2 RQ2: Adaptability for Multiple Individuals

Three students participated in a role-playing exercise, each focusing on distinct information needs: one concentrated on medical supplies and aid, another on affected individuals and regions, and the third on Ebola symptoms and preventive measures. These roles were designed to align with the gold-standard slot-value pairs annotated in the dataset. We compared the performance (F1-score in slot mapping), time taken, API cost, compute power of the model of four zero-shot LLMs in extracting the user-specified information dynamically compared to that of ADAPTIVE IE. We wanted to compute the trade-off of all the models in catering to *adapting to* dynamic information needs.

6.3 RQ3: Adaptability over time

Our goal of this experiment was to assess the time-efficiency of ADAPTIVE IE in responding to evolving information needs, and to compare this with the best-performing LLM (GPT-4 in this case, since it provides the best F1 in Table 2). Therefore, we simulated dynamic shifts in information needs using the Ebola Outbreak from the CrisisNLP Dataset as a real-world case study. We segmented the Ebola Outbreak timeline into three distinct phases: **T1** focused on transmission and symptoms, **T2** on affected areas and casualties, **T3** when information needs are related to vaccines and treatments. Two graduate students participated in a role-playing exercise starting at T1, where they gathered information about the transmission and symptoms of the disease. In T2, they shifted their focus to the areas affected by the outbreak, and by T3, their inquiries centered around vaccines and treatments. After completing each phase, the participants preserved their findings and continued searching while maintaining the same state of clusters. We measured the average time taken by the participants to find answers to their evolving slot mapping requirement. The performance was evaluated using a sample of 300 tweets, with time efficiency being a key metric.

	Biomed	Crisis	GENEVA
	F1	F1	F1
Random	0.09	0.07	0.05
Angeli et al. (2015)	0.15	0.14	0.11
Genest et al. (2022)	0.23	0.24	0.13
Du and Cardie (2020)	0.13	0.17	0.08
Lyu et al. (2021)	0.18	0.22	0.13
Yu et al. (2022a)	**0.23**	**0.26**	0.13
UnsupIE (Ours)	0.20	0.24	**0.15**

Table 1: Compares Macro-F1 of unsupervised baselines on Biomedical Slot Filling, CrisisNLP (Crisis), and GENEVA. Note that, our goal is not to use the best-performing model, but to use a system that can serve as a *good foundation model for user experience*.

7 Main Results and Discussion

Before comparing the performance of ADAPTIVE IE, we explored how well does the "QA-guided UnsupervisedIE" perform compared to the existing unsupervised baselines.

Table 1 compares the performance of various unsupervised information extraction models across three datasets—Biomedical Slot Filling (Biomed), CrisisNLP (Crisis), and GENEVA—using the Macro-F1 metric. UnsupervisedIE, while not necessarily the best in every domain, provides a strong foundation for user-oriented tasks. Notably, it obtains the highest F1 score on the GENEVA dataset (0.15), outperforming other models that struggled to generalize in this domain. In the Crisis dataset,

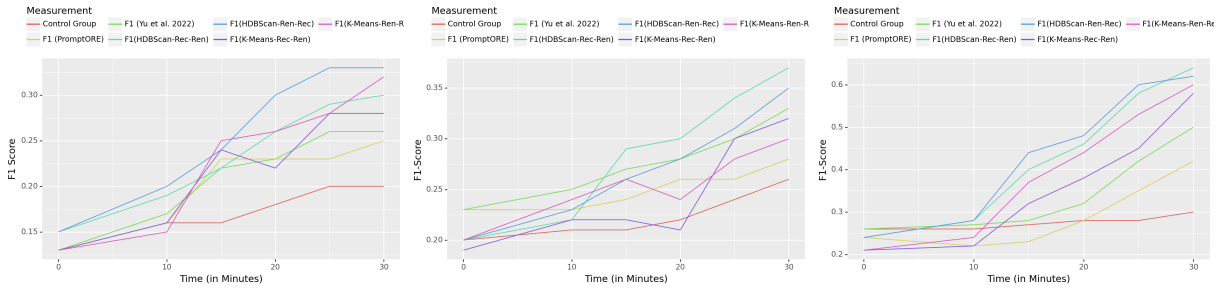

Figure 2: Average Macro F1-scores obtained by ten users at an interval of 10 minutes on the GENEVA (left), Biomedical (middle) and CrisisNLP Datasets (right). At time 0, *UnsupervisedIE* clusters are shown initially and participants kept interacting with ADAPTIVE IE for 30 minutes. *Across all datasets,* ADAPTIVE IE *gets a consistent and significant improvement in Macro F1-scores over time, outperforming unsupervised baselines as participants interact with the system, highlighting its adaptability and effectiveness.*

UnsupervisedIE is competitive with existing models, achieving F1 score of 0.24, slightly behind the leading model by Yu et al. (2022b). In the Biomed dataset, although UnsupIE does not surpass the best performing models, it remains in a competitive range with F1 score of 0.20. Overall, the results highlight UnsupIE's versatility across diverse datasets making it a **promising baseline model for user experience despite not being the highest performer in every category**.[2]

Next, through a series of human experiments described in Section 6, we answer the research questions as follows:

Answer to RQ1): ADAPTIVE IE observes the best trend in helping the users obtain higher F1-gain compared to other baselines. Figure 2 reports the average Macro F1-scores achieved by ten users at an interval of 10 minutes on the GENEVA, Biomedical and CrisisNLP Datasets. The control group consistently shows the lowest F1-score improvements across all intervals. In the GENEVA dataset, the highest performing method, HDBScan-Ren-Rec, shows an improvement from about 0.2 to 0.6 over 30 minutes. In the Biomedical dataset, a similar trend is observed, with the leading approach (K-Means-Ren-Rec) jumping from approximately 0.15 to 0.35. For the CrisisNLP dataset, the leading method is HDBScan-Ren-Rec, which escalates from 0.2 to 0.55.

[2]Note on QA Coverage: The question-answer pairs covered 92%, 89.23%, and 94.56% of the event-related information from the Geneva, Biomedical, and CrisisNLP corpora. Coverage was assessed by fuzzy matching, comparing the recall of generated slots against gold-standard slots. This process measured the proportion of information slots produced by the UnsupervisedIE pipeline, with the final evaluation focusing on the recall of extracted information to ensure comprehensive coverage.

	F1 (↑)	Time (↓)	Comp (↓)	API (↓)
GPT-3	0.82	90 m	Low	High
ChatGPT	0.84	88 m	Low	High
GPT-4	0.84	92 m	Low	High
LLama-13b	0.77	67 m	High	Low
Adaptive IE	0.75	50 m	Low	Low

Table 2: Shows the trade-off between models compared to our approach (300 emergency tweets), where we show our model's efficacy in emergency situations. Notably, ADAPTIVE IE *obtains competitive runtime (Time) and compute efficiency (Comp) with comparatively low API costs, making it a practical solution in emergency scenarios despite a slight trade-off in F1 performance.*

In comparing the KMeans and HDBScan methods to other baseline methods like Genest et al. (2022) and Yu et al. (2022b)'s approach, distinct trends emerge. For the GENEVA dataset, HDBScan-Ren-Rec outperforms all others, achieving F1-score increase from 0.2 to 0.6 over 30 minutes, whereas KMeans-Ren-Rec also performs robustly, though slightly lower. Yu's method and PromptORE are consistently outperformed by these clustering techniques across all datasets. In the Biomedical and CrisisNLP datasets, both KMeans and HDBScan show superior improvement rates, with KMeans-Ren-Rec reaching F1 of 0.35 in Biomedical from a lower baseline, and HDBScan-Ren-Rec escalating to 0.55 in CrisisNLP, both showing more substantial gains than the baselines.

Answer to RQ2): ADAPTIVE IE is the winner in cost-performance trade-off. Table 2 presents a comparative analysis of zero-shot LLM prompting effectiveness during IE tasks from a corpus of 300 documents related to the Ebola Outbreak in the CrisisNLP dataset. ADAPTIVE IE emerged as the most efficient, taking the least overall time (50

Stage	ADAPTIVE IE (↓)	GPT4 (↓)
T1	72.23	91.0
T2	20.22	78.00
T3	28.23	84.00

Table 3: Comparison of Average Time Taken by ADAPTIVE IE and LLM During Different Phases (T1: Transmission & Symptoms, T2: Affected Areas & Casualties, T3: Vaccines & Treatments) of the Ebola Outbreak Crisis Scenario. ADAPTIVE IE *takes significantly lower average time across all phases, making it a more time-efficient solution for crisis response tasks.*

minutes) to obtain a reasonable Macro F1-score (0.75) across all tasks. In contrast, GPT-based models, despite higher F1-scores (up to 0.84), required longer runtime and were less cost-effective due to higher API usage. LLAMA-13b, while showing good performance, demanded high computational resources. ADAPTIVE IE's efficiency is attributed to its one-time use of LLMs for question generation, after which users could refine information clusters without additional LLM overhead, making it faster and more economical (Details in Appendix F).

Answer to RQ3): ADAPTIVE IE adapts quickly over time. Table 3 reveals that ADAPTIVE IE consistently outperforms the best performing LLM GPT-4 in time efficiency across all three stages of the Ebola outbreak crisis scenario, taking significantly less time to extract user tailored information. Specifically, ADAPTIVE IE took 72.23, 20.22, and 28.23 minutes for stages T1, T2, and T3, compared to GPT-4's 91.0, 78.0, and 84.0 minutes. Initially, ADAPTIVE IE takes longer due to one-time overhead of question generation by LLMs in the zero-shot UnsupervisedIE pipeline. Nevertheless, our system shows higher performance, ensuring quicker response time to information that aligns with dynamic requirements during Ebola Outbreak.

8 Further Analysis

In this section, we aim to answer two questions:
a) *Which user feedback results in increased F1-score compared to the previous iterations?* We measure this using hit rate of each type of user feedback (defined as the proportion of instances where the application of a specific feedback type leads to an increase in the Macro F1-score between consecutive iterations). Let $F1_i$ denote the Macro F1-score at iteration i, and let $\Delta F1_i = F1_{i+1} - F1_i$ represent the change in F1-score following the user feedback from iteration i to $i+1$. The user feedback at iteration i is represented by U_i, with specific actions categorized into splitting (U_i^{split}), merging (U_i^{merge}), and rearranging ($U_i^{\text{rearrange}}$). The hit rate for each feedback type k is calculated as Hit Rate$_k = \frac{\text{hits}_k}{n_k}$, where hits$_k$ is the count of instances $\sum_i \mathbf{1}(U_i^k$ and $\Delta F1_i > 0)$, and n_k is the total number of times feedback type k was applied. Here, $\mathbf{1}(\text{condition})$ is an indicator function that returns 1 if the condition is true and 0 otherwise. (Experimental phase of RQ1).
b) *To what extent do the cluster-specific content explanations help users obtain an improved F1-score?* We tested the success of the cluster content explanations on users' ability to make improvements to the slot mapping performance by creating two configurations. In a study with 6 participants, 3 users were provided with cluster content explanations before reclustering (Example in 10), while the other 3 using ADAPTIVE IE system saw no explanations. We compared mean F1 after 20 minutes of their interactions to evaluate how much the explanations contributed to improved slot mapping.

Rearrangement of questions and splitting of clusters are the most prominent feedback for improved F1. Figure 3 illustrates the comparative effect of user feedback operations on the performance improvement on all the datasets. In the Biomedical dataset, splitting clusters shows the highest hit rate at 75%, indicating its strong effectiveness in improving data organization and retrieval compared to rearranging questions (65%) and merging clusters (50%). The Disaster dataset reveals even stronger performance enhancements with rearranging questions leading at an 80% hit rate, followed by splitting clusters at 70% and merging clusters at 55%. The GENEVA dataset demonstrates a more balanced effect with rearranging questions and splitting clusters yielding hit rates of 60% and 68%, while merging clusters shows the least impact at 48%.

Explanations after Reclustering improve the F1 scores most of the times. Figure 4 illustrates the usefulness of providing explanations on slot mapping performance across three datasets. In all cases, participants with explanations achieved higher mean F1 scores compared to those without, with the highest difference observed in CrisisNLP dataset where the group with explanations reached F1 score of 0.75, while the group without explanations achieved only about 0.35. It was because

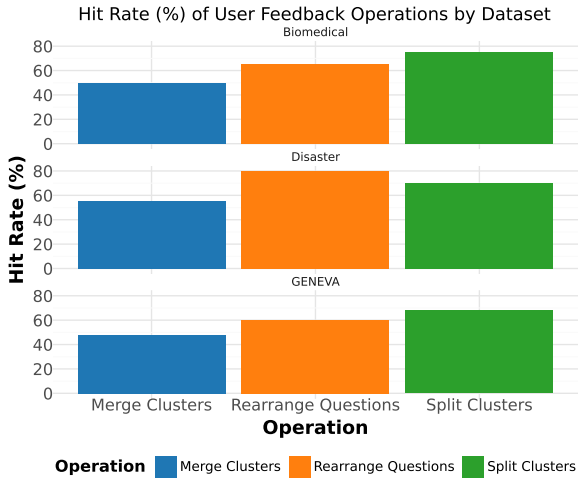

Figure 3: Hit Rate (%) of User Feedback Operations across all the datasets, where we observe, *split clusters and rearrangement of questions stand out.*

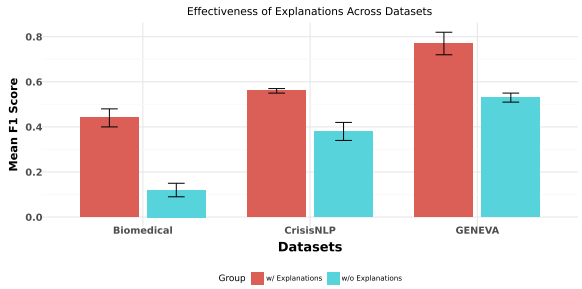

Figure 4: Effect of Explanations indicating that *participants who were provided with explanations consistently achieved higher mean F1 scores in slot mapping performance across all datasets.*

users could infer the pattern of why a certain cluster was formed before providing any further feedback.

9 Background and Related Work

Early work (Chambers and Jurafsky, 2008, 2009) automatically learned a schema from newswire text based on coreference and statistical probability models. Later, Peng et al. (2016) generated an event schema based RNN (Schmidt, 2019). Other studies by Zhang et al. (2022) had focused on modeling event-type semantics by aligning the definition of events with the sentences in a zero-shot manner. However, these methods considered prior annotations of templates or event definitions to extract information from documents. Unsupervised IE by Yates et al. (2007) aimed to extract intents without having access to a labeled dataset during training. Roy et al. (2019) proposed an ensemble method to aggregate results of multiple OpenRE models but relies on surface forms which makes it difficult to group instances with same relations expressed us ing different syntax. Li et al. (2020b) and Li et al. (2021) used transformers to handle schema generation and viewed a schema as a graph instead of a linear sequence. However, they could not make it domain-agnostic.

Recently, various methods have been developed to treat Event Extraction (EE) as a form of QA for academic research. This methodology, treating EE as a QA problem, has been explored in works by Du and Cardie (2020), Li et al. (2020a), and Lyu et al. (2021). This process involves generating questions for each argument role, created using pre-defined templates. Pre-defined question templates are effective but lack flexibility and context-specific details (Du and Cardie, 2020). Recent works emphasize Human-in-the-Loop (HITL) approaches to enhance machine learning. Mosqueira-Rey et al. (2024) addressed data bottlenecks with GAN-based augmentation and active learning for iterative expert feedback in medical diagnostics. Similarly, Bobes-Bascarán et al. (2024) integrated HITL to align machine learning outputs with domain-specific standards, enhancing model explainability and reliability. Zeng et al. (2024) leveraged LLMs for summarization and hidden state extraction, enabling scalable, user-guided data analysis across domains. Recently, Dror et al. (2023) took GPT-3 generated documents to build a schema but it suffered from the instability of GPT-3 outputs. Another area related to our work is human-in-the-loop schema generation as done by Ciosici et al. (2021). However, they relied a lot on human input as compared to Zhang et al. (2023)'s work using GPT3 generated candidate steps for schema generation. Due to over-reliance on GPT-3 generations, these models suffered from hallucination in complex domains (Pu and Demberg, 2023; Dror et al., 2023). However, our generated questions are grounded on source documents, ensuring faithfulness. Besides, our method has been benchmarked on multiple domains unlike other works.

10 Conclusion

With the acknowledgements that depending on human annotation is expensive and inefficient, while fully automated generations can be unreliable, we introduce a "human-in-the-loop IE" approach powered by the capabilities of LLMs as the backbone. Our system can be pivotal in analyzing critical information from various sources during emergency.

Acknowledgement

We sincerely thank the anonymous reviewers and the UMD CLIP members—Wichayaporn Wongkamjan, Zongxia Li, Nishant Balepur, Trista Cao, and Calvin Bao—for their valuable feedback and constructive comments on the draft. We also extend our gratitude to Shramay Palta and Yoo Yeon Sung for their support in shaping the interface and assisting with the pilot studies. This work, led by Ishani, was supported by the Adobe Research Gift Fund, the Global Terrorism Database (GTD) research team at the University of Maryland, and the Intelligence Advanced Research Projects Activity (IARPA) through the BETTER (Better Extraction from Text Towards Enhanced Retrieval) program. Previously, Michelle was funded by the COE grant and her work was done before joining Amazon. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the funding agencies.

Limitations and Future Work

While our study demonstrates the effectiveness of ADAPTIVE IE in cost-performance trade-off, we acknowledge the following limitations:

- **User Pool Size:** The experiments were conducted with a relatively small number of users, which might limit the generalizability of the findings. Future work should involve scaling the study to a larger and more diverse participant base.
- **Domain Specificity:** The current study is limited to two domain-specific datasets (e.g., biomedical and crisis datasets). Expanding to additional domains, including low-resource and non-English datasets, would provide a more comprehensive evaluation of the approach.
- **Interface Features:** Some participants expressed interest in interactive visualizations, such as TSNE plots, to better understand the clustering process. The absence of such features in the current system may limit its usability for complex analysis tasks. As a next version of the interface, we hope to include both extrinsic and intrinsic evaluation to provide better guidance to the users.
- **Iterative Refinement Dependency:** While the system demonstrates adaptability, its performance heavily depends on iterative user feedback. This may lead to slower information extraction in scenarios where immediate responses are required.
- **Language Limitations:** The system's evaluation is primarily conducted in English. Testing on non-English datasets, especially low-resource languages, could highlight potential language-specific challenges.
- **Computational Constraints:** The reliance on LLMs for question generation introduces computational overhead, particularly for initial processing stages. Optimizing this step could improve time and resource efficiency.

Ethics Statement

The experiments performed in this study involved human participants. All the experiments involving human evaluation in this paper were exempt under institutional IRB review. We recruited participants for our human study using Upwork and we have fairly compensated all the Upwork freelancers involved in this study, at an average rate of 15.00 USD per hour (respecting their suggested Upwork hourly wage). Prior to the study, the participants provided explicit consent to the participation and to the storage, modification and distribution of the collected data. All the involved participants gave their consent to disclose their interactions with the interface. The documents used in the study are distributed under an open license.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

David Bamman and Noah A. Smith. 2014. Unsupervised discovery of biographical structure from text. *Transactions of the Association for Computational Linguistics*, 2:363–376.

José Bobes-Bascarán, Eduardo Mosqueira-Rey, Ángel Fernández-Leal, Elena Hernández-Pereira,

David Alonso-Ríos, Vicente Moret-Bonillo, Israel Figueirido-Arnoso, and Yolanda Vidal-Ínsua. 2024. Evaluating explanatory capabilities of machine learning models in medical diagnostics: A human-in-the-loop approach.

Nathanael Chambers. 2013. Event schema induction with a probabilistic entity-driven model. In *EMNLP*.

Nathanael Chambers and Dan Jurafsky. 2008. Unsupervised learning of narrative event chains. In *Proceedings of ACL-08: HLT*, pages 789–797, Columbus, Ohio. Association for Computational Linguistics.

Nathanael Chambers and Dan Jurafsky. 2009. Unsupervised learning of narrative schemas and their participants. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 602–610, Suntec, Singapore. Association for Computational Linguistics.

Jackie Chi Kit Cheung, Hoifung Poon, and Lucy Vanderwende. 2013. Probabilistic frame induction. In *NAACL*.

Nancy Chinchor and Elaine Marsh. 1998. Appendix D: MUC-7 information extraction task definition (version 5.1). In *Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29 - May 1, 1998*.

Manuel Ciosici, Joseph Cummings, Mitchell DeHaven, Alex Hedges, Yash Kankanampati, Dong-Ho Lee, Ralph Weischedel, and Marjorie Freedman. 2021. Machine-assisted script curation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 8–17, Online. Association for Computational Linguistics.

Rotem Dror, Haoyu Wang, and Dan Roth. 2023. Zero-shot on-the-fly event schema induction. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 705–725, Dubrovnik, Croatia. Association for Computational Linguistics.

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.

Francis Ferraro and Benjamin Van Durme. 2016. A Unified Bayesian Model of Scripts, Frames and Language. In *Proceedings of the 30th Conference on Artificial Intelligence (AAAI)*, pages 2601–2607, Phoenix, Arizona. Association for the Advancement of Artificial Intelligence.

Pierre-Yves Genest, Pierre-Edouard Portier, Elöd Egyed-Zsigmond, and Laurent-Walter Goix. 2022. Promptore - a novel approach towards fully unsupervised relation extraction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, CIKM '22, page 561–571, New York, NY, USA. Association for Computing Machinery.

Muhammad Imran, Prasenjit Mitra, and Carlos Castillo. 2016. Twitter as a lifeline: Human-annotated Twitter corpora for NLP of crisis-related messages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1638–1643, Portorož, Slovenia. European Language Resources Association (ELRA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 829–838, Online. Association for Computational Linguistics.

Lishuang Li, Ruiyuan Lian, Hongbin Lu, and Jingyao Tang. 2022. Document-level biomedical relation extraction based on multi-dimensional fusion information and multi-granularity logical reasoning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 2098–2107, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Manling Li, Sha Li, Zhenhailong Wang, Lifu Huang, Kyunghyun Cho, Heng Ji, Jiawei Han, and Clare Voss. 2021. The future is not one-dimensional: Complex event schema induction by graph modeling for event prediction. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5203–5215, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Manling Li, Qi Zeng, Ying Lin, Kyunghyun Cho, Heng Ji, Jonathan May, Nathanael Chambers, and Clare Voss. 2020b. Connecting the dots: Event graph schema induction with path language modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 684–695, Online. Association for Computational Linguistics.

Samuel Louvan and Bernardo Magnini. 2020. Recent neural methods on slot filling and intent classification for task-oriented dialogue systems: A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 480–496, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Qing Lyu, Hongming Zhang, Elior Sulem, and Dan Roth. 2021. Zero-shot event extraction via transfer learning: Challenges and insights. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 322–332, Online. Association for Computational Linguistics.

Claudia Malzer and Marcus Baum. 2020. A hybrid approach to hierarchical density-based cluster selection. In *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*. IEEE.

György Móra, Richárd Farkas, György Szarvas, and Zsolt Molnár. 2009. Exploring ways beyond the simple supervised learning approach for biological event extraction. In *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, pages 137–140, Boulder, Colorado. Association for Computational Linguistics.

Eduardo Mosqueira-Rey, Elena Hernández-Pereira, José Bobes-Bascarán, David Alonso-Ríos, Alberto Pérez-Sánchez, Ángel Fernández-Leal, Vicente Moret-Bonillo, Yolanda Vidal-Ínsua, and Francisca Vázquez-Rivera. 2024. Addressing the data bottleneck in medical deep learning models using a human-in-the-loop machine learning approach. *Neural Computing and Applications*, 36(5):2597–2616.

Yannis Papanikolaou, Marlene Staib, Justin Joshua Grace, and Francine Bennett. 2022. Slot filling for biomedical information extraction. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 82–90, Dublin, Ireland. Association for Computational Linguistics.

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.

Ellie Pavlick, Heng Ji, Xiaoman Pan, and Chris Callison-Burch. 2016. The gun violence database: A new task and data set for NLP. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1018–1024, Austin, Texas. Association for Computational Linguistics.

Haoruo Peng, Yangqiu Song, and Dan Roth. 2016. Event detection and co-reference with minimal supervision. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 392–402, Austin, Texas. Association for Computational Linguistics.

Dongqi Pu and Vera Demberg. 2023. ChatGPT vs human-authored text: Insights into controllable text summarization and sentence style transfer. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1–18, Toronto, Canada. Association for Computational Linguistics.

Colin Raffel, Noam M. Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *ArXiv*, abs/1910.10683.

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.

Arpita Roy, Youngja Park, Taesung Lee, and Shimei Pan. 2019. Supervising unsupervised open information extraction models. 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 728–737, Hong Kong, China. Association for Computational Linguistics.

Robin M. Schmidt. 2019. Recurrent neural networks (rnns): A gentle introduction and overview. *ArXiv*, abs/1912.05911.

Kristina P. Sinaga and Miin-Shen Yang. 2020. Unsupervised k-means clustering algorithm. *IEEE Access*, 8:80716–80727.

Rohini Srihari and Wei Li. 2000. A question answering system supported by information extraction. In *Sixth Applied Natural Language Processing Conference*, pages 166–172, Seattle, Washington, USA. Association for Computational Linguistics.

Alexander Yates, Michele Banko, Matthew Broadhead, Michael Cafarella, Oren Etzioni, and Stephen Soderland. 2007. TextRunner: Open information extraction on the web. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 25–26, Rochester, New York, USA. Association for Computational Linguistics.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022a. Unsupervised slot schema induction for task-oriented dialog. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1174–1193, Seattle, United States. Association for Computational Linguistics.

Tiezheng Yu, Rita Frieske, Peng Xu, Samuel Cahyawijaya, Cheuk Tung Yiu, Holy Lovenia, Wenliang Dai, Elham J. Barezi, Qifeng Chen, Xiaojuan Ma, Bertram

Shi, and Pascale Fung. 2022b. Automatic speech recognition datasets in Cantonese: A survey and new dataset. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6487–6494, Marseille, France. European Language Resources Association.

Xianlong Zeng, Yijing Gao, Fanghao Song, and Ang Liu. 2024. Similar data points identification with llm: A human-in-the-loop strategy using summarization and hidden state insights. *arXiv preprint arXiv:2404.04281*.

Hongming Zhang, Wenlin Yao, and Dong Yu. 2022. Efficient zero-shot event extraction with context-definition alignment. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7169–7179, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tianyi Zhang, Isaac Tham, Zhaoyi Hou, Jiaxuan Ren, Leon Zhou, Hainiu Xu, Li Zhang, Lara Martin, Rotem Dror, Sha Li, Heng Ji, Martha Palmer, Susan Windisch Brown, Reece Suchocki, and Chris Callison-Burch. 2023. Human-in-the-loop schema induction. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 1–10, Toronto, Canada. Association for Computational Linguistics.

A Appendix: Implementation Details

We use sentenceBERT (Reimers and Gurevych, 2019) to encode the passages and the queries before doing the clustering. We use both Kmeans (Sinaga and Yang, 2020) and HDBSCan (Malzer and Baum, 2020) clustering of the questions for grouping the questions based on semantic similarities. For question-generation methods, we use pretrained T5 (Raffel et al., 2019) and BART (Lewis et al., 2020) to generate questions pivoted on event triggers and different entities. Besides, we have experimented with three different LLMs such as GPT-3 (*text-davinci-003*), ChatGPT (*gpt-3.5-turbo*) and GPT-4 (*gpt-4-turbo*) from OpenAI. All experiments are carried out with temperature 0 to have a reproducible setup and top-p nucleus sampling set to 0.9. For generating the initial number of clusters by UnsupervisedIE, K was chosen to be any number greater the number of desired slots in the experiment and (For instance, in Figure 8, K=4 where the number of desired slots is 2, only "Damaged Properties" and "Causalties"). The idea is to have atleast the number of slots that the user would like to see. Now, we use the following prompts for event trigger identification, question-answering using the prompts in B using all the LLMs specified above. Our interface is developed using streamlit (Screenshot in Figure 10).

Algorithm 1 Adaptive IE Methodology

Require: Q: Set of all questions from documents $(D_1, D_2, D_3, \dots)$, $\mathbf{F}(t)$: User feedback
Ensure: $\mathbf{C}(t+1)$: Refined clusters of questions

1. **Step A: Initialize with** $\mathbf{C}(t)$
   - Extract initial triggers and generate QA pairs from D
   - Cluster QA pairs into initial groups based on semantic similarity
2. **While** User not satisfied **do**
   - 2.1. **Step B and C: Handle User Feedback** $\mathbf{F}(t)$ **and update the clusters**
     - Apply user feedback to adjust clusters:
       - Merge related clusters
       - Split broad clusters
       - Rearrange questions into correct clusters
3. **Step D: Present Clusters to User**
   - Display updated clusters to the user for feedback

Output: Final refined clusters $\mathbf{C}(t+1)$

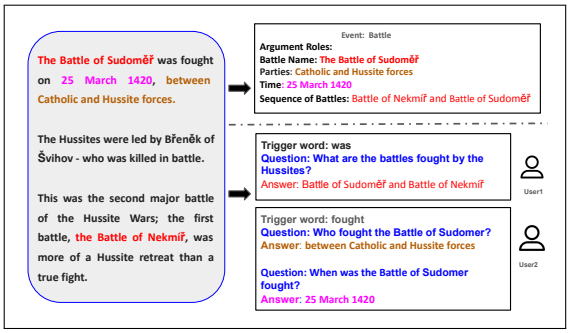

Figure 5: An example shows the motivation of using a QA-driven approach of extracting information on-the-fly depending on user requirements. Supervised template-driven approaches require pre-annotated templates, whereas QA-driven interactive pipeline using trigger words **fought** generates all possible question-answer pairs corresponding to an event, and it satisfies user's information needs on-the-fly.

Examples of UnsupervisedIE Output. Figure 5 illustrates how a QA-driven UnsupervisedIE approach, using trigger words, can dynamically generate question-answer pairs from a text corpus, in this case using data from the GENEVA dataset. Trigger words such as "was" and "fought" are employed to extract specific details about events, allowing for the generation of relevant questions and answers in real-time based on user requirements. For example, when the trigger word "was" is detected, the system identifies questions that ask for factual information about past events, such as "What battles were fought by the Hussites?" and retrieves the answer, "Battle of Sudoměř and Battle of Nekmíř". Similarly, the trigger word "fought" generates ques-

tions like “Who fought the Battle of Sudoměř?" and “When was the Battle of Sudoměř fought?” with corresponding answers drawn directly from the text. This approach eliminates the need for pre-annotated templates by leveraging trigger words to create a dynamic and interactive pipeline. It efficiently addresses users’ specific queries by generating all possible question-answer pairs related to the event, adapting to different information needs on-the-fly.

Figure 8 illustrates a zero-shot QA-guided information extraction (IE) slot induction process, where the system automatically generates question-answer pairs by identifying triggers related to an event—in this case, the 2014 Chile Earthquake—and organizes them into meaningful clusters. The first step involves unsupervised information extraction (IE) on documents D1, D2, and D3, which are associated with the event. Triggers such as “trapped” and “commute” are detected to identify specific sub-events or concerns related to the earthquake. Once the triggers are identified, the system generates relevant QA pairs linked to those triggers. For example, for the trigger “trapped”, the system generates questions such as “What is the estimated number of people trapped under debris?” or “What are the designated safe areas near Mexico?” Similarly, for the trigger “commute”, questions like “How to commute from East to West Chile?” and “By what means can we travel in this situation?” are produced. After the QA pairs are generated, the system clusters the questions into four categories: C1 (General Information), C2 (Safety), C3 (Aftermath), and C4 (Miscellaneous). The questions are grouped based on the type of information they address. For instance, questions related to safety measures, such as blocked roads and evacuation routes, are categorized under C2, while questions related to the aftermath of the earthquake, such as the number of people who have died or the number of damaged buildings, are placed under C3. This process allows for efficient organization of information in a way that satisfies various user needs and queries about the event.

B Appendix: Prompts

Question-Answer Generation Prompt

Instruction: You are an assistant that reads through a passage and provides all possible question and answer pairs to the trigger word t_i, and the questions will help ascertain facts about the event triggered by t_i,. The questions should roughly follow templates like:wh* verb subject trigger object1 preposition object2 Wh* is a question word that starts with wh (i.e. who, what, when, where). Answers MUST be direct quotes from the passage. Do not ask any inference questions. From this question set, remove semantically redundant or duplicate question-answer pairs and produce a set of question-answers that are quite different from each other in terms of information need. Questions: Q
Passage: P

Cluster Explanation Generation Prompt

Instruction: The collection of questions within this cluster can be presented as follows. Generate an explanation regarding how they cater to similar informational needs.
Questions: Q

Event Trigger Extraction Prompt

Instruction: List all potential event triggers from the passage. Format your output as a list of triggers.
Passage: P

Zero-Shot Prompt for IE-on-the-Fly

Instruction: You are an assistant that reads through a passage and extracts all possible information pertaining to the goal of the user. Format your answer as a list of JSON Objects where keys are the infomation type and values are the extracted spans from the passage.
Passage: P
Goal of the user: G

Few-Shot Prompt for IE-on-the-Fly

Instruction: Instruction: You are an assistant that reads through a passage and extracts all possible information pertaining to the goal of the user. Format your answer as a list of JSON Objects where keys are the infomation type and values are the extracted spans from the passage.
Passage: P
Goal of the user: G
Some Examples:
Example 1
Example 2
Example 3
Example 4
Example 5

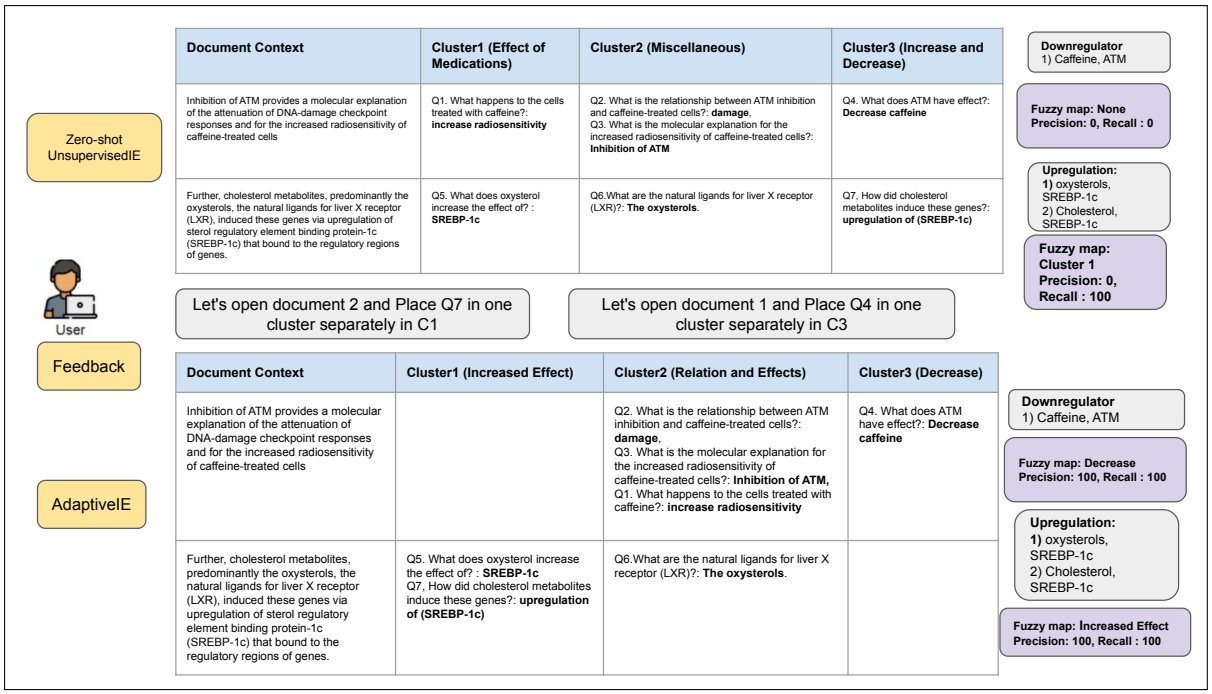

Figure 6: This figure illustrates a feedback-driven ADAPTIVE IE workflow, starting with Zero-Shot Unsupervised IE to generate initial clusters of questions based on the context of documents. Clusters such as "Effect of Medication", "Miscellaneous" and "Increase and Decrease" are formed but lack semantic refinement. User feedback is then applied, where specific questions (e.g., Q4 and Q7) are manually reassigned to better-defined clusters. The updated clusters (e.g., "Increased Effect", "Relation and Effects", "Decrease") achieve greater semantic coherence, enabling ADAPTIVE IE to improve the precision and recall of information extraction through a "fuzzy map" refinement process. The figure demonstrates the effectiveness of incorporating user feedback to enhance the semantic alignment of clusters, improving the precision and recall of extracted information. It showcases how ADAPTIVE IE dynamically adapts clusters and aligns them with user intent, highlighting the value of human-in-the-loop processes in achieving more accurate and meaningful information extraction results.

C Appendix: Human Study Recruitment

Our user study was not limited to the individuals who are well-versed in the concepts of Machine Learning or Natural Language Processing, we wanted to verify if the participants can understand what does a semantically coherent cluster look like. For this, we recruited those participants with their native language as English. Out of ten, only four of the participants had prior experience on NLP. In order to familiarize them with the clustering task, we asked them to solve a simple assignment as described in figure 7. We have recruited those participants who could successfully complete the task without any difficulty. Prior to the study, we collected consent forms for the workers to agree that their answers would be used for academic purposes. All the involved participants gave their consent to disclose their interactions with the interface. Moreover, they were fairly compensated based on the amount they had proposed for this particular task. During the actual study, we provided some examples of passages and gold slots to make them understand the context. We ensured that the documents we have used for uploading in the interface were different from the ones shown to them for making themselves familiar with the task and setup.

D Appendix: CrisisNLP Slot Filling Dataset Statistics

We repurpose CrisisNLP to create a slot filling dataset for emergency domain. Using GPT-4, we initially identified precise information from each tweet, ensuring it matched predetermined categories. For instance, in "Emergency Aids" category, we focused on extracting specific details like locations of emergency and availability of emergency supplies, organizing this information into slot-value pairs. Manual examination was conducted to guarantee the accuracy of the dataset, which involved removing entries that were not relevant, finally creating a dataset comprising 3,000

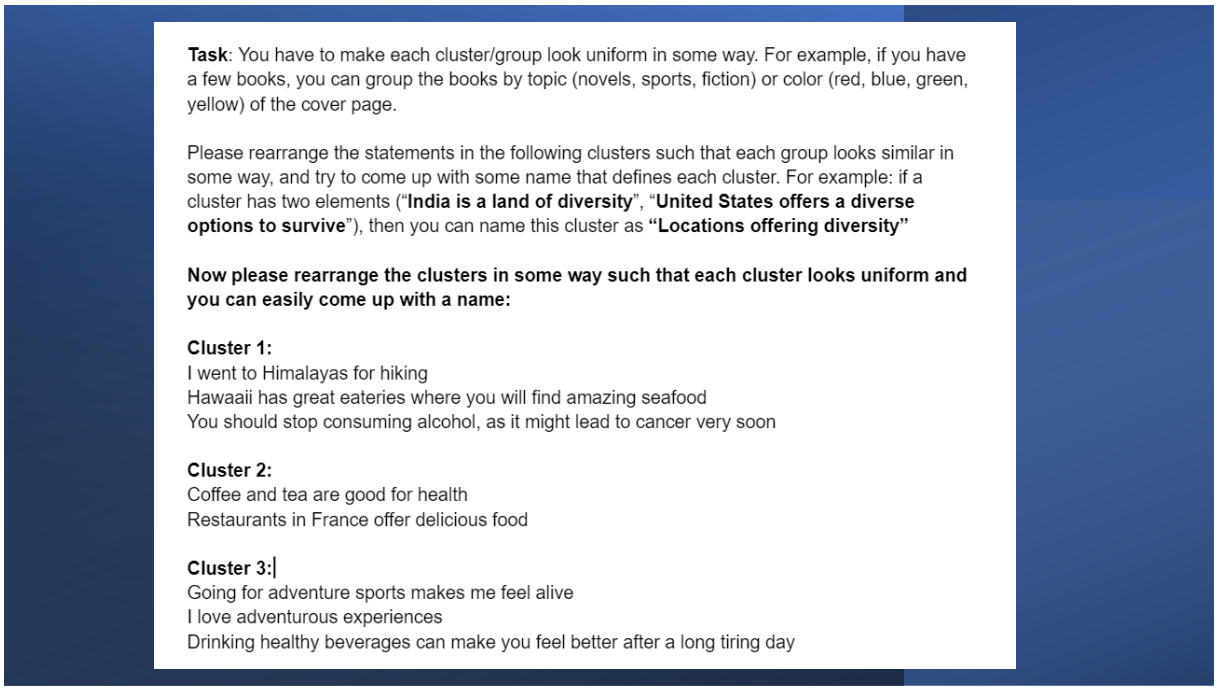

Figure 7: This figure presents a clustering assignment where users are instructed to rearrange statements within given clusters to create uniform groups that can be easily named. Users are encouraged to organize the statements based on shared themes or topics, such as "Locations offering diversity" or "Healthy habits", and assign an appropriate cluster name. The exercise demonstrates the importance of semantic coherence in clustering and naming, highlighting the role of user understanding in refining and improving clusters. The task emphasizes how human intuition and semantic reasoning can enhance the interpretability and uniformity of clusters, showcasing the importance of meaningful grouping in organizing information.

tweets from Chile Earthquake, Ebola Outbreak, Typhoon and 6,940 slot-value pairs. Chile Earthquake (1,000 tweets) had the following pairs:

- Emergency and Supplies: 200 slot-value pairs (e.g., availability of water, food, shelter)
- Affected Areas and Evacuation: 200 slot-value pairs (e.g., specific locations hit, evacuation centers)
- Casualties and Damage: 300 slot-value pairs (e.g., death toll, infrastructure damage)
- Emotional Support and Prayers: 300 slot-value pairs (e.g., messages of hope, calls for assistance)

For the Ebola Outbreak, the slot-value pair focus on medical supplies, affected individuals, regions with outbreaks, and awareness efforts.

- Medical Supplies and Aid: 250 slot-value pairs (e.g., availability of medicines, medical teams)
- Affected Individuals: 250 slot-value pairs (e.g., number of cases, recovery rates)
- Regions with Outbreaks: 250 slot-value pairs (e.g., specific towns or districts affected)
- Awareness and Education: 250 slot-value pairs (e.g., preventive measures, symptoms)

For the Typhoon, the focus was meteorological data, evacuation information, relief efforts, and infrastructure damage.

- Meteorological Data: 200 slot-value pairs (e.g., wind speed, rainfall levels)
- Evacuation Information: 300 slot-value pairs (e.g., safe zones, transportation options)
- Relief Efforts: 250 slot-value pairs (e.g., aid distribution, volunteer groups)
- Infrastructure Damage: 250 slot-value pairs (e.g., roads blocked, power outages)

E Appendix: User Study Details

In this section, we aim to detail the various slot categories utilized during our user study (RQ1). Here, we explain the experimental configurations for three datasets:

A. Biomedical Dataset. Initially, participants reviewed initial question clusters derived from a corpus of 1,000 biomedical documents. Following this review, they engaged with the ADAPTIVE IE interface to systematically categorize information pertaining to potential side effects of biomedical

entities. These were classified into specific slots: "Cause-Effect", "Upregulation", and "Downregulation". These slots served as proxies for the users' information needs, allowing us to assess how an unsupervised system adapts to fulfill specific informational requirements. For instance, we observed the system's ability to dynamically populate slots concerning "Cause-Effect" relationships, such as identifying the consequences of drug interactions, or "Upregulation" and "Downregulation", which involve changes in gene expression levels due to various stimuli or interventions (Examples of slots and the questions that are mapped to the slot in Table). This approach tested the effectiveness of ADAPTIVE IE system to align with user-specific queries without prior supervision. The participants interacted with the interface for 30 minutes, and we record the slot mapping performance due to interactions (Using Precision, Recall, F1-score for each slot) at an interval of 10 minutes. Besides, we also provide examples of some good edits of the user in Figure 6.

B. CrisisNLP Dataset. In this study, the participants focused on analyzing 2,000 documents from the CrisisNLP Dataset, which comprised two disaster events: the Chile Earthquake and the Ebola Outbreak, with 1,000 documents dedicated to each event. This part of the study was designed to gauge how effectively participants could extract information corresponding to predefined slots that cater to specific needs during disaster responses. For the Ebola Outbreak, the primary slots were "affected areas", "casualties", and "recovery rates". Participants were tasked with identifying and organizing data from the documents that detailed the regions impacted by the outbreak, the number of casualties, and the rates at which affected individuals were recovering. This exercise aimed to simulate the process of gathering critical health and location-specific information during a health crisis, which is vital for directing medical response and resources. Similarly, for the Chile Earthquake, the designated slots were "Damaged Properties" and statistics on "missing or dead people". Here, the task involved extracting information about the extent of property damage and the human toll in terms of missing or deceased individuals. This kind of information is crucial for initiating recovery efforts, understanding the severity of the impact, and mobilizing rescue and rehabilitation operations. These activities within the user study not only tested the system's ability to assist in the rapid categorization of vital information during crises but also provided insights into the practical challenges and effectiveness of using an unsupervised learning approach to manage real-world disaster data. The study aimed to demonstrate how such a system could potentially enhance decision-making processes by providing timely and organized information to responders and planners. Besides, we also provide examples of some good edits of the user in Figure 8.

C. GENEVA Dataset. In the study, we asked participants to role-play scenarios involving Action Type events, such as "medical procedures" and "crimes". Specifically, they were tasked with identifying key details like the place, time, and reason for the event's occurrence, as well as the instruments used. Additionally, for 'attack' events, participants were asked to extract information about the occurrence time and the motive behind the attack. This approach helped participants engage more deeply with the event structure, focusing on these critical aspects to fill the corresponding slots in the information extraction system.

F Appendix: Cost-Performance Tradeoff of ADAPTIVE IE

LLM calls with LLAMA-13b incur no API cost, as shown in Table 2; however, its performance, as measured by F1 score, is lower compared to GPT-based models. For the UnsupervisedIE pipeline, we began by extracting trigger words and generating QA pairs for each document, averaging six LLM calls per tweet. When LLama-13b is used, there is no API cost associated with UnsupervisedIE. In cases where the user modifies clusters within a document, each cluster is reorganized, and LLMs are used to refine the cluster names based on the representative questions within each cluster (RecRen approach). After the initial clusters from the UnsupervisedIE pipeline are displayed, the user provides feedback, which may require renaming up to K clusters, depending on the changes made in the ADAPTIVE IE reclustering process. Thus, an additional K LLM calls are needed to rename clusters based on representative questions. Overall, we make approximately 6 * 300 + K LLM calls.

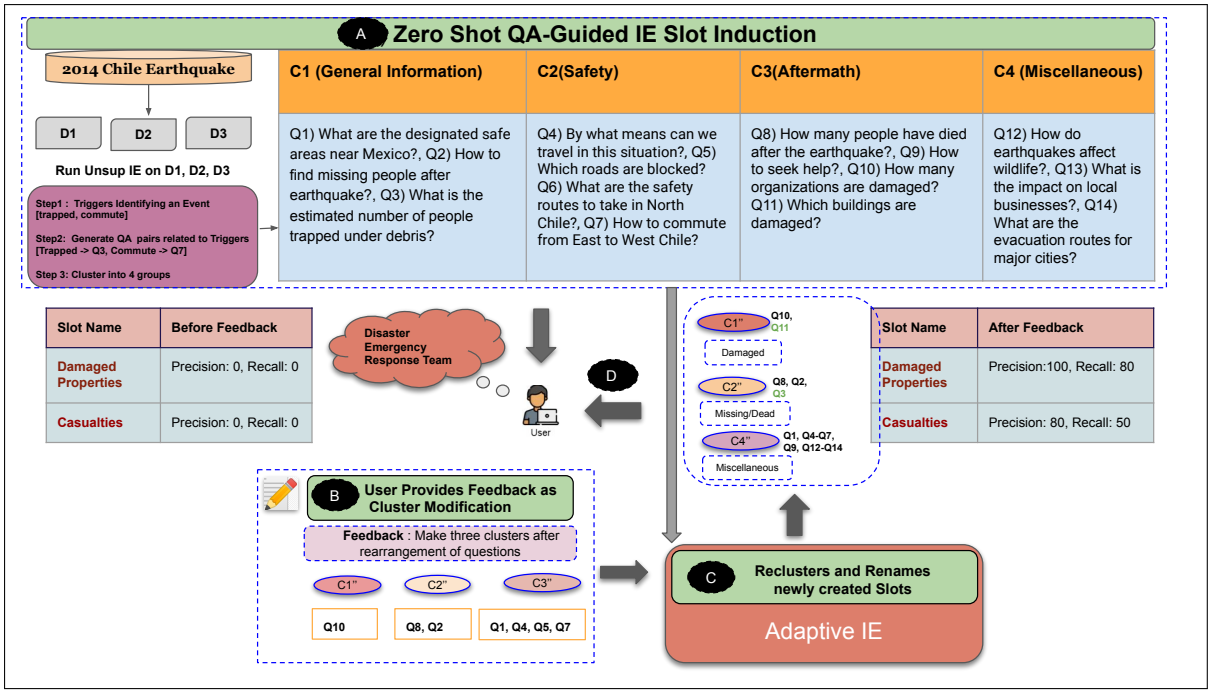

Figure 8: This figure illustrates the workflow of a Zero-Shot QA-Guided Information Extraction (IE) system, focusing on slot induction for emergency scenarios, such as the 2014 Chile Earthquake. The process begins with unsupervised IE applied to datasets (D1, D2, D3), generating question-answer (QA) pairs clustered into groups based on semantic similarity. Users provide feedback to refine clusters, leading to reclustering and renaming of slots, improving precision and recall metrics significantly through the ADAPTIVE IE framework. The figure demonstrates how user feedback can effectively guide the refinement of clusters and improve slot-based information extraction, showcasing the adaptability of the system in aligning semantic relationships to achieve higher precision and recall in real-world crisis scenarios. This highlights the value of combining human-in-the-loop processes with adaptive systems for improving information extraction workflows.

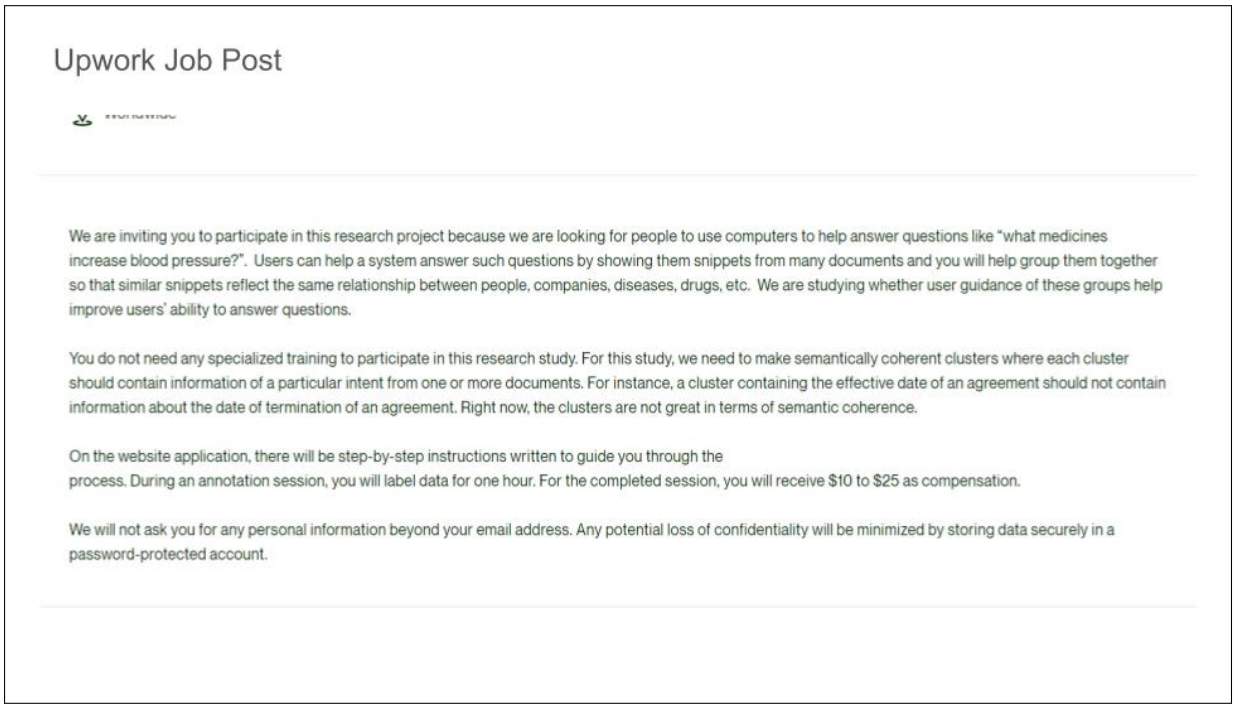

Figure 9: This figure presents an Upwork job post inviting participants to contribute to a research study focused on semantic clustering. Participants are tasked with organizing snippets of information into semantically coherent clusters, such as grouping data based on relationships between people, companies, diseases, or drugs, to enhance a system's ability to answer complex questions. The post outlines the task details, compensation, and assurances regarding data security and privacy. The job post highlights the importance of human input in improving semantic coherence in clustering tasks, demonstrating how user guidance can refine automated systems and enhance their performance in question-answering scenarios.

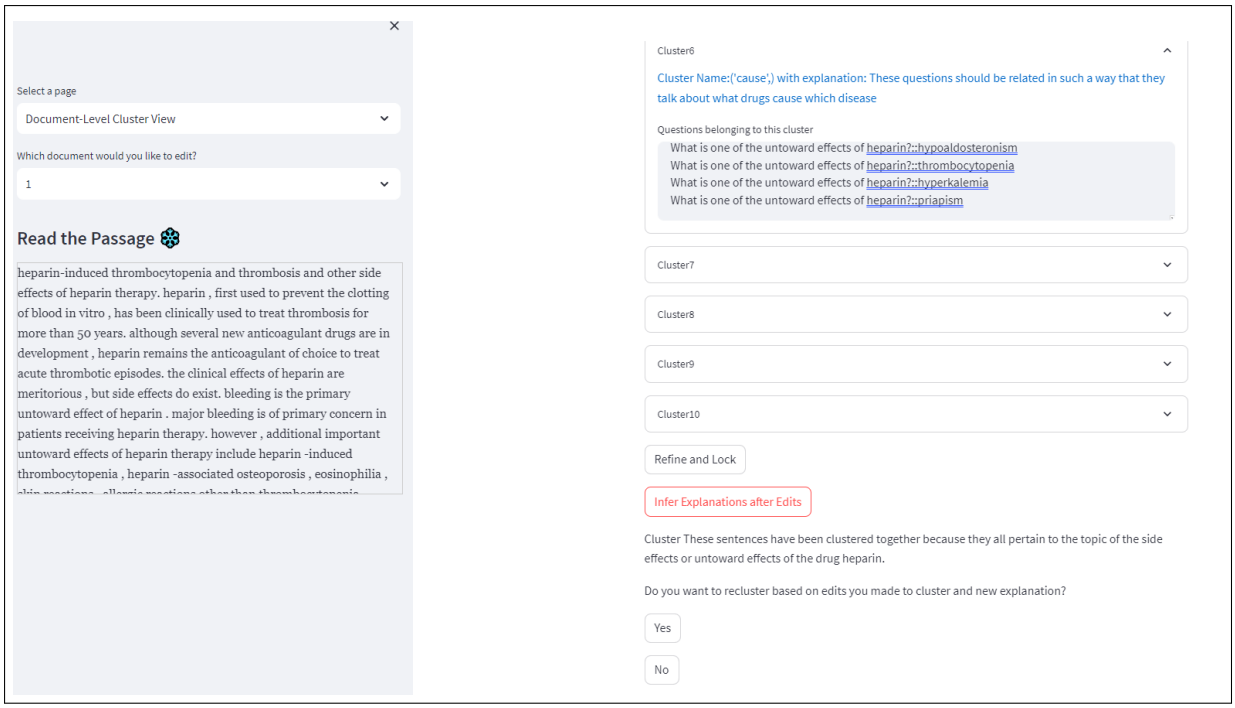

Figure 10: This figure shows a screenshot of our interface which shows the document-level cluster editing page. The participants can view, edit, and refine clusters of questions based on their semantic grouping. The left panel displays the passage text for reference, while the right panel shows the clusters, their names, explanations, and associated questions. Users can refine clusters, lock changes, and infer new explanations after edits, ensuring the clusters accurately align with the intended semantics of the passage.