PERSONA-ADAPTIVE IE: Harnessing Human-AI Collaboration to Adaptively Extract Persona-Aware Information

Anonymous ACL submission

Abstract

The evolving nature of information needs across diverse domains like emergency situations (disease outbreak, earthquake) necessitates a flexible information extraction (IE) system. 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 solve it using our proposed PERSONA-ADAPTIVE IE framework that leverages human-in-the-loop refinement to adapt to changing user queries. Through human experiments on three diverse datasets, we demonstrate that PERSONA-ADAPTIVE IE is a *domain-agnostic*, *responsive*, *efficient* framework for helping users *access useful information* while quickly reorganizing information in response to evolving information needs.

1 Information Needs are Ever-Growing . . . How Can we Efficiently Extract Information On-The-Fly?

The primary objective of IE is to derive structured insights from unstructured documents, guided by a predefined schema specifying the targeted relationships to be extracted. Existing IE tools help the analysts understand certain patterns or behaviors in the world (Li et al., 2022; Móra et al., 2009). However, in a fast-moving real-world situation, IE requirements are prone to shift over time and vary significantly from individual to individual, making it impractical to anticipate the specific nature of information needs in advance. Figure 1 shows the contrasting needs of two distinct users engaging with the same initial corpus on an earthquake event. User A, primarily concerned with immediate safety measures post-earthquake wants to identify safe areas away from damaged buildings. So he would filter through broad information clusters to focus on actionable steps like accessing safe locations, de-emphasizing aspects like casualties or damage. Conversely, a journalist (User B) approaches same dataset with a different objective: to extract information on impact, including damage and casualties, tailoring to suit reporting needs. In such cases, a minimally supervised system should be ideal with a) *improved extraction accuracy over unsupervised approaches* (6.2), b) *rapidly adapting to user feedback and meets time-sensitive demands* (6.3), c) *ease of information access* (6.3) and d) *system's adaptability across domains*.

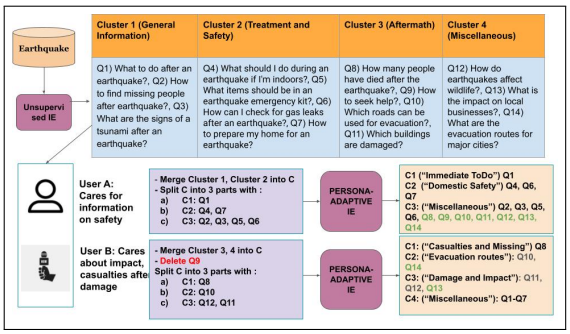


Figure 1: Illustrates that traditional IE (Yu et al., 2022) offers general set of information categories which *might not* satisfy people with varying interests; whereas PERSONA-ADAPTIVE IE can customize information through guidance of user needs.

Limitations of Existing IE Systems: Prior unsupervised approaches are probabilistic from modeling patterns in clauses (Chambers, 2013; Cheung et al., 2013; Bamman and Smith, 2014; Ferraro and Van Durme, 2016), some other methods rely on template-driven QA methods to represent events in the documents (Li et al., 2022; Móra et al., 2009). Still, template matching accuracy is low for these methods and they rely on pre-defined templates of questions to be generated from documents. Besides, recent unsupervised IE (Aharoni and Goldberg, 2020; Yu et al., 2022) systems can not distin-

guish between these nuanced needs without explicit guidance. Also, supervised IE systems which rely on human annotations of templates (Chinchor and Marsh, 1998; Pavlick et al., 2016) are impractical for deployment in such real-world scenarios. While numerous Large Language Models (LLM) leverage zero-shot or few-shot methods for IE (Yuan et al., 2023; Wei et al., 2023; Han et al., 2023), applying these methods on a full document corpus in emergencies, where time and cost efficiency are vital, is practically not suitable for deployment.

Contributions: To address this gap, we make four-fold contributions: **[1]** We introduce the task of *IE on-the-fly* from corpus that emphasizes on extracting personalized information (Section 2). **[2]** Since information needs are well-represented through asking questions (Du and Cardie, 2020; Du et al., 2022; Li et al., 2020a), we introduce a QA-guided unsupervised IE that detects events within the corpus and generating potential questions about the semantic roles associated with that event to meet the majority of informational needs (Section 3) **[3]** We propose a human-in-the-loop IE module (PERSONA-ADAPTIVE IE) where users are presented with the unsupervised clusters of information, allowing for iterative adjustments (split or merge clusters, or edit/remove unnecessary questions) to meet their current informational needs (Section 4). In Figure 1, user A first merges two broad clusters, splits those into three smaller clusters (Immediate ToDo, Domestic Safety, Miscellaneous), and rearranges questions to maintain logical consistency. Our system automatically learns the patterns of his information needs, and reclusters other information catering to his needs. On the other hand, user B, provides guidance on making specific clusters on causalities, evacuation protocols, damage and impact. **[4]** We conduct human experiments using three datasets and show that PERSONA-ADAPTIVE IE significantly increased F1 of extracted information over unsupervised approaches in 30 minutes, making it an adaptable, persona-aware, domain-agnostic solution to meet informational requirements on-the-fly.

2 Task Motivation and Formalization of 'Information Extraction On-the-Fly'

An IE system typically involves defining templates and a set of slot types. Each slot type pertains to a specific semantic role. In Figure 11, the objective is to gather comprehensive information related to the event 'fought.' Questions crafted to define the specific informational needs, like 'What battles did the Hussites engage in?', 'When did the battle take place?', or 'Who were the combatants in the battle?' should be posed to obtain the necessary details. Grouping these questions into clusters aids in structuring them according to their distinct informational needs. Answers to these questions fill up the values for each slot type.

Task Formalization: Let $\mathcal{C}$ denote a corpus from which information needs to be extracted. At any given point in time t, a user's informational need is represented by N_t. The task of 'on-the-fly IE', IE_{fly}, is defined as a function that maps a user's current informational need to a set of thematic information $\mathcal{T}_t$ extracted from $\mathcal{C}$: IE_{fly}: $\mathcal{N}_t \times \mathcal{C} \rightarrow \mathcal{T}_t$, where $\mathcal{T}_t$ represents a set of clusters extracted from $\mathcal{C}$ that satisfy user's informational need $\mathcal{N}_t$ at time t. This allows for dynamic extraction of clusters from the same corpus $\mathcal{C}$ in response to evolving informational needs. Specifically, if a user has a different informational need at time t_2 than at time t_1, the function can retrieve a different set of clusters, $\mathcal{T}_{t_2}$, corresponding to $\mathcal{N}_{t_2}$ and N_{t_2} such as $IE_{\text{fly}}(\mathcal{N}_{t_1}, \mathcal{C}) = \mathcal{T}_{t_1}, \quad IE_{\text{fly}}(\mathcal{N}_{t_2}, \mathcal{C}) = \mathcal{T}_{t_2}$
Hence, $\mathcal{T}_{t_1} \neq \mathcal{T}_{t_2}$ if $\mathcal{N}_{t_1} \neq \mathcal{N}_{t_2}$, demonstrating the system's flexibility in adapting to user's changing needs over time in a cost-effective manner. This also holds true when IE needs vary from user to user, we denote information needs of User 1 (u_1) and user 2 (u_2) as N_{u_1} and N_{u_2} respectively.

3 QA-guided Unsupervised IE

For capturing IE needs that change over time, we define a way to quickly bootstrap template schemas with zero to minimal supervision (motivation in 2). Our pipeline for unsupervisedIE begins with processing a corpus (C) with information need as n queries, $Q = \{q_1, q_2, \ldots, q_n\}$. This multi-step setup generates schema, S, for organizing related information (Output: Step 1 of Figure 2).

Event Trigger Identification: For extracting all possible information, we extract all trigger words (verbs describe the occurrence of events) corresponding to an event. We prompt LLM_t to extract the most important events or entities (triggers T = t_1, ...t_n) from each document in the corpus (B). We also use non-LLM approaches to extract verbs causing an event. For each t_i, we generate question-answer pairs to extract maximum information.

Question-Answer Pair Generation: Given a document d and set of triggers $T = t_1, ...t_n$, we generate "WH"-type questions by prompting LLM_{qa} such that they contain one of the triggers t_i whose answer is a continuous span in d. Our questions answer about Who, Whom, What, When, Where, Why, How of an event (Prompt B).

Clustering with Explanations: By grouping similar questions and their corresponding answers, users can more efficiently retrieve relevant information and that helps in understanding the underlying patterns or commonalities among questions, leading to more accurate and relevant answer identification. Therefore, the refined question-answer pairs are then clustered into K-groups. We initially create clusters of questions, and then take questions corresponding to centroid of each cluster to prompt $LLM_{Cluster}$ to generate explanation of why these questions are clustered together (See B).

4 PERSONA-ADAPTIVE IE Methodology

Why is Human Feedback Important? Automatic clustering of questions as presented in section 3 encounters some challenges: Firstly, the clusters may contain questions that are repetitive. Secondly, semantic cohesion within a cluster can be weak, leading to inclusion of potentially irrelevant details. Figure 2 shows that the user with the goal of accessing information about immediate actions after the earthquake looks at the initial clusters (Table 4) " General Info " and " Safety Tips ", wishes to merge into a single, more coherent cluster named " Preparation and Safety " (Table1). This extracts information tailored to his needs and iteratively modifies to the final output Table1.

Goal Specification and Relevance Scoring: After looking at the output of Step 1 in Figure 2, the user specifies goal (Step 2) of their broad information need. Documents are ranked based on the cosine similarity based on the semantic (BERT(Devlin et al., 2019)) embeddings of goal and documents and corresponding documents with clusters are shown in front of them after Step 2.

Iterative Clustering: In Step 3, the user provides feedback on document-specific or broader clusters to rearrange information according to their needs. Let $Q = \{q_1, q_2, \ldots, q_n\}$ be the set of questions to be clustered, $C^{(t)} = \{C_1^{(t)}, C_2^{(t)}, \ldots, C_k^{(t)}\}$ as the set of clusters at iteration t, where each $C_j^{(t)}$

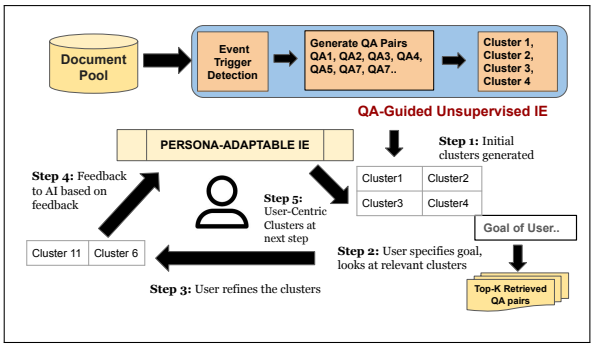


Figure 2: Illustrates a step-by-step process for adapting IE to user needs. Step 1 involves generating initial clusters from the data. In Step 2, users define their goals and review relevant clusters. Step 3 sees users refining these clusters based on their requirements. PERSONA-ADAPTABLE IE is designed to evolve based on user feedback, ensuring that extracted information is increasingly aligned with user-defined preferences.

contains questions grouped by their semantic roles, $f : Q \times V \times S \rightarrow C^{(t)}$ as the clustering function that maps each question q_i to one of the clusters in $C^{(t)}$, $F^{(t)}$ be the feedback at iteration t on clusters, $u : (C^{(t)}, F^{(t)}) \rightarrow C^{(t+1)}$ as update function that applies feedback $F^{(t)}$ to current set of clusters $C^{(t)}$ to generate clusters for next iteration $C^{(t+1)}$. At each iteration t, initial set of clusters $C^{(t)}$ are presented to user The user reviews clusters and provides feedback $F^{(t)}$ based on his goal. Let $C = \{C_1, C_2, \ldots, C_n\}$ be a set of clusters and Q be the questions, where each cluster C_i contains questions $Q_i \subseteq Q$. Following are types of user feedback ($F^{(t)}$) obtained:

A. Merge Clusters: The user feels that safety information and general guidelines are closely related and would benefit from being in the same cluster, so he merges " General Info " and " Safety Tips " into a new cluster called " Preparation and Safety ." To merge clusters C_i and C_j into a new cluster C_k: $C_k = C_i \cup C_j$, $C' = C \setminus \{C_i, C_j\} \cup \{C_k\}$

B. Rearrange Questions: Questions which do not fall into correct bin (user goal) "How to volunteer for rescue operations?" are moved to " Aid and Support ". For a question q moving from cluster C_i to C_j: $C_i' = C_i \setminus \{q\}$, $C_j' = C_j \cup \{q\}$

C. Split Clusters: The user wants to differentiate between immediate aid options and longer-term support services for affected individuals, So he splits " Aid and Support " into " Humanitarian Aid " (covering blood donation,

and financial aid) and " Support Services " (covering psychological support, repair) To split cluster C_i into two new clusters C_j and C_k: $C'_i = \varnothing$, $C_j = \{q \mid q \in C_i,$ and q meets criteria for $C_j\}$, $C_k = \{q \mid q \in C_i,$ and q meets criteria for $C_k\}$, $C' = C \setminus \{C_i\} \cup \{C_j, C_k\}$

D. Move Questions Since volunteering is considered an immediate aid action, so it fits better with humanitarian efforts, now the user wants to move "How to volunteer for rescue operations?" from " Preparation and Safety " to " Humanitarian Aid ." To move a question q from cluster C_i to C_j: $C'_i = C_i \setminus \{q\}$, $C'_j = C_j \cup \{q\}$

E. Delete Questions The ill-formed or redundant questions in the cluster are usually deleted by the users, for e.g., "How to seek help?" To delete a question q from cluster C_i: $C'_i = C_i \setminus \{q\}$.

Closing the Loop (Step 4): An update function u is used with current clusters $C^{(t)}$ to produce $C^{(t+1)}$ in next iteration. Clusters are re-arranged based on following principle:

I. Centroid-based Reclustering–Recluster-Rename (Rec-Ren) User feedback generates two types of constraints:

a) **Must-have constraints**: $M \subseteq Q \times Q$, indicating questions that must be in the same cluster.

b) Cannot-have constraints, $N \subseteq Q \times Q$, indicating questions that must not be in same cluster.

This update function u dynamically adjusts clusters regarding the semantic structuring of questions, enabling refined groupings over time. After reclustering, we use an LLM to generate the most suitable names for each cluster (using questions closest to centroid), providing users with flexibility to make edits after that.

II. Naming-based Reclustering–Rename-Recluster (Ren-Rec) We use an LLM to generate the most suitable names for each cluster (using questions closest to centroid), providing users with flexibility to make edits after that. Let N_i be the name of C_i, then we compute its embedding $e(N_i)$. For any question $q_j \in Q$, compute its embedding $e(q_j)$. The assignment of q_j to a cluster C_i is based on the highest 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. Finally, the user concludes if the clustering configuration aligns with his objectives, otherwise, proceed to next iteration with $t = t + 1$.

Clusters	Questions Corresponding to each Cluster
(Preparation and Safety)	What to do after an earthquake?, What are the best practices for earthquake-proofing a home?, What are the emergency kit essentials?, What are the evacution routes for major cities?, How to protect pets in an earthquake?, What to do if trapped under debris?
(Humanitarian Aid)	Where to donate blood in an emergency?, Organizations involved in earthquake relief?, How to apply for financial aid after an earthquake?, How to volunteer for rescue operations?
(Aid and Support)	What are the psychological support services available?, How to find missing people after an earthquake? What are the infrastructure repair timelines?
(Historical Data)	What are the biggest earthquakes in the last decade?, What are the earthquake prediction methods?, What are some of the seismic activity monitoring tools?
(Environmental and Community Impact)	How do earthquakes affect wildlife?, What is the impact on local businesses?, What are the community initiatives for rebuilding?, What are the environmental consequences of earthquakes?, What are the cultural responses to earthquake disasters?

Table 1: Shows the output of PERSONA-ADAPTIVE IE on the same 20 instances of the 2014 Chile Earthquake portion of CrisisNLP dataset where clusters pertain to user's informational needs starting from Table 4.

5 Experimental Setup and Evaluation

Datasets: We conduct experiments on three existing datasets from diverse domains to test the domain adaptability (generalizability) 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 200 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, where we experiment with 3000 tweets. 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 the dataset, 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, all relevant to emergency situations (Statistics in F).

Baselines for Comparison: We compare **UnsupervisedIE** (without human feedback) with the

following baselines: **1) BERTQA** (Du and Cardie, 2020) (Based on BERT, it enhances label semantics through a QA (Question Answering) 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):** (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. Moreover, we consider triple-extraction baselines (<Subject, Relation, Object> triple (SVO-based methods)) such as **3) OpenIE** (Angeli et al., 2015), **4) PromptORE** (Genest et al., 2022) which extracts some of the 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 the sentences, instead consider the overall context to ask questions conditioned on the tagged entities. **5) (Yu et al., 2022)**: It comprises of bottomup span extraction method regularized by unsupervised probabilistic context-free grammar (PCFG) structure, 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. We consider PERSONA-ADAPTIVE IE with the best-performing baselines (Prompts in B, experiment details in E).

Evaluation Metrics: For fair comparison among different methods, we use the recently adopted evaluation strategy of Yu et al. (2022) to calculate precision, recall, and F1 on induced slot types (Example in C). Besides, we also evaluate models using normalized pointwise mutual information (NMI), a standard measure of coherence of clusters.

6 Results and Analysis

6.1 UnsupervisedIE Performance Analysis

We compare our UnsupervisedIE approach with different baselines, making different choices of models for question generation and clustering approach and also the type of constraints during clustering. Our observations are as follows:

UnsupervisedIE using QA-driven clustering is a competitive baseline compared to other unsupervised approaches. The best configuration of UnsupervisedIE on Biomedical Dataset scores 0.20, which is an improvement over the Random baseline (0.09), and the method cited from (Du and Cardie, 2020) (0.13). It is, however, slightly lower than the highest score achieved by the method from (Yu et al., 2022) (0.23), and equal to (Angeli et al., 2015) (0.15) and (Genest et al., 2022) (0.23). The trend is similar for the disaster dataset. However, UnsupIE achieves the highest score of 0.15, outperforming or performing competitive with all other methods on the GENEVA Dataset (Table 2).

	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., 2022)	**0.23**	**0.26**	0.13
UnsupIE (Ours)	0.20	0.24	**0.15**

Table 2: Compares Macro-F1 of unsupervised baselines on Biomedical Slot Filling (Biomed) dataset, CrisisNLP (Crisis), and GENEVA. It shows that **our QA-Driven UnsupervisedIE (without human supervision)** competes closely with other methods (Yu et al., 2022; Genest et al., 2022) on datasets across *diverse domains*, often outperforming others on GENEVA dataset.

HDBScan Clustering achieves better slot filling performance. From Figure 10, 3, 4, we observe the performance of unsupervisedIE for different ablations at time 0 when the user starts reviewing clusters. On Biomedical Dataset, HDBScan achieves a marginal performance gain over Kmeans in F1-score, with increase of 5% for both Rec-Ren and Ren-Rec. In terms of NMI, HDBScan and Kmeans show comparable performance, but HDBScan edges out (3.4% increase in Rec-Ren) (A)

6.2 PERSONA-ADAPTIVE IE Performance

A total of ten participants were hired using Upwork to evaluate the effectiveness of our PERSONA-ADAPTIVE IE in IE on-the-fly through an interactive interface (Figure 9) showing clusters and explanations from the UnsupervisedIE model at the first iteration. 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 for all datasets (See D). We wanted to evaluate the *effectiveness in terms of improved IE performance over unsupervised approches*, *ease of information access by users*, *adapatability to various needs* and also *runtime comparison* compared to other SOTA approaches ensuring better response time. In the

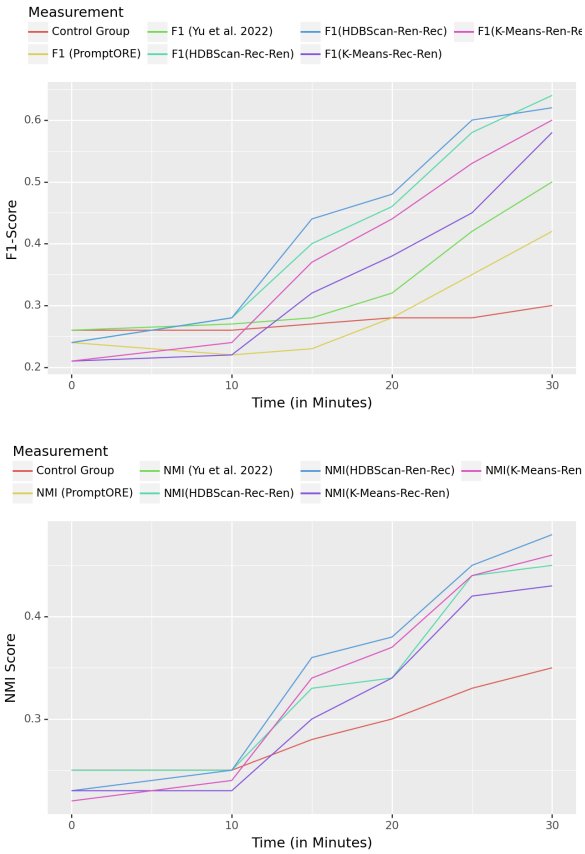


Figure 3: Average F1-scores and NMI scores achieved by ten users at different time stamps on the Disaster Dataset. At time 0, *UnsupervisedIE* clusters are shown initially and the participants kept interacting with PERSONA-ADAPTIVE IE for 30 minutes. At certain intervals, we notice change in performance of all the configurations (macro F1).

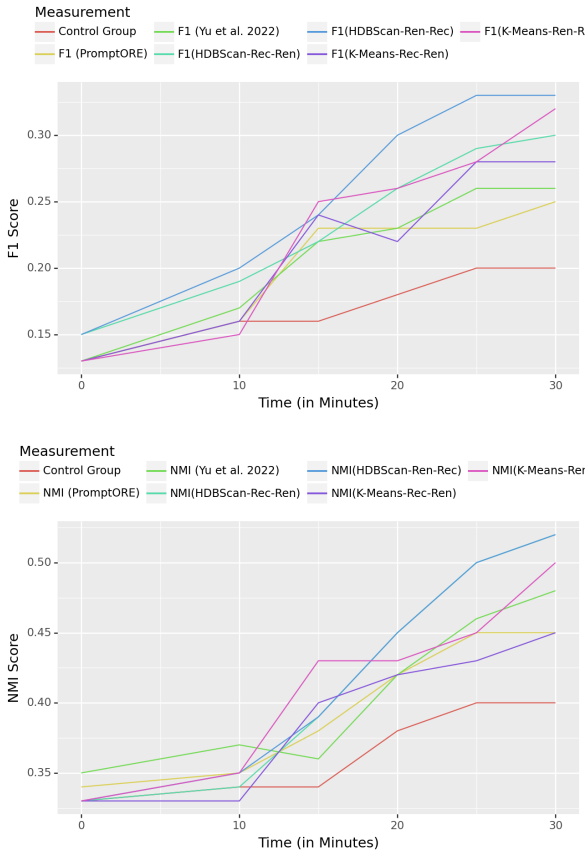


Figure 4: Average F1-scores and NMI scores achieved by ten users at different time stamps on the GENEVA Dataset. At time 0, *UnsupervisedIE* clusters are shown initially and the participants kept interacting with PERSONA-ADAPTIVE IE for 30 minutes. At certain intervals, we notice change in performance of all the configurations (macro F1).

Control phase, participants are tasked with manually specifying the goal, get the information from documents, pertaining to their goal, then we evaluate their answers based on the gold standard. In the Experimental Phase, we ask the same participants to use our PERSONA-ADAPTIVE IE to obtain the answers relevant to their goals. To initally experiment which configurations work well, we first sample 400 documents from Disaster corpus and then ask the users to glean on the clusters provided by Triple-based methods like (Genest et al., 2022) and QA-based methods such as (Du and Cardie, 2020) and our *UnsupervisedIE*. The participants were given a 30-minute window to extract information pertinent to five distinct goals, each requiring different types of information ("Emergency Services after an Earthquake"). We conducted tests using various configurations, some with the retrieval component and others without, as well as tests that either included or excluded explanations of the interactively refined clusters presented to the users.

Both Explanations and Retrieval of important documents related to user goal can help achieve better accuracy The QA-Based and Triple-Based categories show the highest F1 scores when both Explanations (E) and retrieval (R) components are used together (E+R), with F1 scores of 50.22% and 40.78%, respectively (Figure 5). The performance drops in Only E or Only R confugurations, with the QA-Based category showing F1 scores of 27.23% and 39.01% and the Triple-Based category showing F1 scores of 33.05% and 21.0%. The Experimental Configurations category has an F1 score of 43.13% when using both components together (E+R), which is higher than using only Explanations or only retrieval, with F1 scores of 43.89% and 33.05%, respectively.

Next, we presented the participants with clusters generated by UnsupervisedIE and two best-performing baselines (Table 2), and we compared these with extraction accuracy of a control group. The participants were tasked with determining which configuration yielded the highest performance after a 30-minute period. Initially, at the 0th minute, participants were shown the clusters produced by each configuration and instructed to commence IE. Each participant was exposed to various configurations (clusters from different unsupervised baselines) and asked to extract information related to three specific goals for a duration of 30 minutes. This test was applied to 500 documents from each of the three datasets, with the aim of identifying which configuration most effectively assists users in achieving the highest accuracy, given that their initial goals remained consistent.

PERSONA-ADAPTIVE IE achieves the best trend in helping the users achieve higher F1-gain compared to other baselines. A generic observarion in figure 10, 3, 4 is that the humans could achieve higher F1 and NMI scores compared to *UnsupervisedIE* on slot mapping within 30 minutes. Using HDBScan-Rec-Ren configuration of PERSONA-ADAPTIVE IE, we observe the most rapid improvement (+0.17 F1) in 30 minutes, followed by HDBScan-Ren-Rec (0.15). K-Means configurations have moderate improvements with K-Means-Rec-Ren at 0.13 and K-Means-Ren-Rec at 0.10. The slowest improvements are seen in Control Group and (Genest et al., 2022) with (Yu et al., 2022) matching K-Means-Ren-Rec at 0.10 (Figure 10) (See A) The improvement is due to the explanation of clusters at each step, and the question-answers provide enough context to users to group information needs in a logical way.

6.3 Testing Temporal Adaptability and Runtime Comparison

On newly created CrisisNLP Slot Filling dataset, we simulate dynamic changes in information needs, similar to those that occur in real-world crisis situations. Using the Ebola Outbreak as a case study, we divided the timeline into three phases (time=T1, time=T2, time=T3). Initially, at T1, users sought information predominantly about the transmission and symptoms of the disease. At T2, the focus shifted to the areas affected by the outbreak. Finally, at T3, the concern moved to vaccines and treatments. To explore these time-sensitive information needs, two graduate students engaged in a role-playing exercise beginning at T1. They started by seeking answers to slots such as 'Rate of transmission', 'Symptoms related to the disease', and 'Procedure of disease spread', all under the theme of 'Transmission and symptoms'. At T2, they searched for information on slots like 'Places Affected' and 'Casualties related to the outbreak', falling under the broader category of 'Affected Areas'. Then, at T3, their inquiries centered on 'Protection' and 'Vaccination rate'. Following the completion of T1, the participants preserved their findings and continued to search for the next set of information, maintaining the same state of clusters as in T1, and proceeded similarly from T2 to T3. Our goal was to assess the average time it took for participants to find answers to their evolving queries and to evaluate the adaptability of our system. For benchmarking purposes, we prompted GPT-4 to extract information from the documents. We then compared the time efficiency of our method with this state-of-the-art Language Model on a sample of 300 tweets, and the findings are reported in 6.

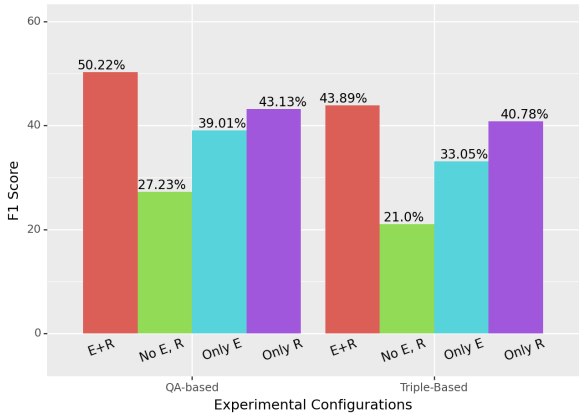


Figure 5: Presents F1-scores achieved for different experimental configurations of PERSONA-ADAPTIVE IE on CrisisNLP. Both cluster-specific content explanations (E) and retrieval augmentation (R) during human experiments achieve the highest F1-score of 50.22% on QA-based methods, and 43.89% on Triple-based methods, suggesting (explanations+retrieval augmentation) significantly boosts performance of IE on-the-fly.

Our system is temporally more adaptable too! Figure 6 displays the duration required for data extraction at successive phases (T1, T2, T3), comparing our system with GPT-4 zero-shot prompting. Initially, at T1, the system takes longer due to one-time overhead of question generation by GPT. Nevertheless, our system demonstrates higher per-

	F1 (↑)	Runtime (↓)	Compute (↓)
GPT-3	0.82	90 m	Low
GPT-3.5-turbo	0.84	88 m	Low
GPT-4	0.84	92 m	Low
LLAMA-13b	0.77	67 m	High
Ours	0.75	50 m	Low

Table 3: Shows the trade-off between SOTA IE models compared to our approach (300 emergency tweets), where we show efficacy of our model in emergency situations (high response time and low compute power).

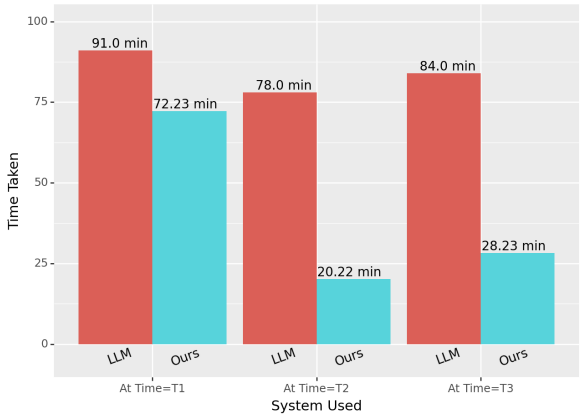


Figure 6: Illustrates the mean time taken by two students to extract information at different stages of an evolving crisis scenario using GPT-4 versus our system. At each interval (T1, T2, T3), our system consistently outperforms GPT-4, demonstrating faster information access in response to changing information needs during the Ebola Outbreak case study.

formance, ensuring quicker response time to information that aligns with dynamic requirements during Ebola Outbreak. In Table 3, we highlight the accuracy-cost tradeoff and accuracy-compute tradeoff of our model compared to GPT-models and LLama-13b (Touvron et al., 2023). Even though GPT is the winner in terms of F1-score, GPT-calls on a set of large documents incur very high API costs, limiting accessibility during emergency.

7 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 (Zhang et al., 2022) has focused on modeling event-type semantics by aligning the definition of events with the sentences in a zero-shot manner. However, these methods consider prior annotations of templates or event definitions to extract information from documents.

Recently, various methods have been developed to treat Event Extraction (EE) as a form of Question Answering (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., 2020b), and (Lyu et al., 2021). This process involves generating questions for each argument role, created using pre-defined templates. These methods proved to be effective, but using predefined question templates has its own limitations; these templates, created manually, lack flexibility and context-specific details, often only incorporating trigger words (Du and Cardie, 2020). Nonetheless, crafting well-thought-out questions are difficult to generate without knowing exact information need, and no human-in-the-loop approach has focused on tweaking questions for adaptive IE. To fill these gaps, we have first introduced a QA-driven IE approach using LLMs that extracts the answers of various argument roles of the events and entities involved in any relation. To enable adaptability to user needs, we also provide human agency to organize information into groups that they care about.

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 another work using GPT3 generated candidate steps for schema generation as proposed by (Zhang et al., 2023). Due to over-reliance on GPT-3 generations, these models might suffer 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 is domain-agnostic, which we have validated using three datasets from different domains.

8 Discussion and Conclusion

With the acknowledgements that fully 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 clustering and explanation generation capabilities of LLMs as the backbone. Our system can be pivotal in analyzing critical information from various data sources during emergencies, such as natural disasters, medical crises, or security threats. By rapidly processing unstructured data, PERSONA-ADAPTIVE IE can provide actionable insights, helping emergency responders make informed decisions quickly.

Limitations

We have a few limitations in our approach. First, we have conducted experiments with a small set of users and we plan to scale it up in the future. We will eventually segregate the pool of participants into two groups: participants with domain knowledge and no domain knowledge. This will help us analyze whether domain-specific knowledge is required to extract more useful information from such documents. Second, our experiments are based on two domain-specific datasets, therefore, we hope to experiment on different tasks and datasets where manual data annotation is an expensive affair, such as non-English datasets (mainly low-resource languages). Finally, some participants wanted to take a look at interactive TSNE plots at each step of their interactions with the interface, particularly when they are tweaking the number of clusters in the preprocessing view. As a next version of the interface, we hope to include both extrinsic and intrinsic evaluation in order to provide better guidance to the users.

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.

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.

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

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.

Li Du, Xiao Ding, Yue Zhang, Ting Liu, and Bing Qin. 2022. A graph enhanced BERT model for event prediction. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2628–2638, Dublin, Ireland. 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.

Ridong Han, Tao Peng, Chaohao Yang, Benyou Wang, Lu Liu, and Xiang Wan. 2023. Is information extraction solved by chatgpt? an analysis of performance, evaluation criteria, robustness and errors.

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, Y. Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020a. Event extraction as multi-turn question answering. In *Findings*.

Fayuan Li, Weihua Peng, Yuguang Chen, Quan Wang, Lu Pan, Yajuan Lyu, and Yong Zhu. 2020b. 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.

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.

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.

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.

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Xiang Wei, Xingyu Cui, Ning Cheng, Xiaobin Wang, Xin Zhang, Shen Huang, Pengjun Xie, Jinan Xu, Yufeng Chen, Meishan Zhang, Yong Jiang, and Wenjuan Han. 2023. Zero-shot information extraction via chatting with chatgpt.

Dian Yu, Mingqiu Wang, Yuan Cao, Izhak Shafran, Laurent Shafey, and Hagen Soltau. 2022. 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.

Chenhan Yuan, Qianqian Xie, and Sophia Ananiadou. 2023. Zero-shot temporal relation extraction with ChatGPT. In *The 22nd Workshop on Biomedical Natural Language Processing and BioNLP Shared Tasks*, pages 92–102, Toronto, Canada. Association for Computational Linguistics.

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 Ablation of components of PERSONA-ADAPTIVE IE

PERSONA-ADAPTIVE IE achieves the best trend in helping the users achieve higher F1-gain compared to other baselines. On GENEVA, K-Means-Ren-Rec configuration demonstrates the most rapid improvement with an absolute increase of 0.19. HDBScan-Ren-Rec follows closely at 0.18, with K-Means-Rec-Ren and HDBScan-Rec-Ren showing equal improvements of 0.15. The Control Group exhibits slower improvement (0.07) (Figure 4). On disaster dataset, HDBScan-Rec-Ren configuration shows the most rapid improvement with an absolute increase of 0.40, closely followed by K-Means-Ren-Rec and HDBScan-Ren-Rec at 0.39 and 0.38, respectively. The slowest improvement is observed in the Control Group at 0.04 (Figure 3).

HDBScan Clustering achieves better slot filling performance. On the GENEVA Dataset, HDBScan outperforms Kmeans with a more notable margin in F1-score, exhibiting a 15.4% increase in the Rec-Ren configuration. For NMI, HDBScan again outperforms Kmeans, this time with a smaller but still significant 8.7% increase in the Rec-Ren configuration. Similary, for the Disaster Dataset, HDBScan shows a substantial performance gain over Kmeans in F1-score, with an increase of 14.3% for the Rec-Ren configuration. In the case of NMI, HDBScan surpasses Kmeans by 13.6% in the Ren-Rec configuration, indicating a better clustering quality that aligns well with the ground truth.

B 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

Clusters	Questions Corresponding to each Cluster
Cluster 1 (General Info)	What to do after an earthquake?, How to find missing people after an earthquake?, What are the signs of a tsunami after an earthquake?, How to volunteer for rescue operations?
Cluster 2 (Safety Tips)	What are the best practices for earthquake-proofing a home?, What are the emergency kit essentials?, What are the evacuation routes for major cities?, How to protect pets in an earthquake?, What to do if trapped under debris?, How to seek help?
Cluster 3 (Aid and Support)	Where to donate blood in an emergency?, What organizations are involved in earthquake relief?, How to apply for financial aid after an earthquake?, What are the psychological support services available?, What are the infrastructure repair timelines?
Cluster 4 (Historical Data)	What are the biggest earthquakes in the last decade?, What are the earthquake prediction methods?, What are some of the seismic activity monitoring tools?, Where to get food supplies during an emergency?
Cluster 5 (Miscellaneous Information)	How do earthquakes affect wildlife?, What is the impact on local businesses?, What are the community initiatives for rebuilding?, What are the environmental consequences of earthquakes?, What are the cultural responses to earthquake disasters?

Table 4: Shows the output of QA-Guided UnsupervisedIE on 20 instances of 2014 Chile Earthquake subset of CrisisNLP. The clusters seem to be a little out of the place particularly the "General Info" and "Miscellaneous Info", since none of the answers to the slots represent a unique information need.

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

Clusters	Questions
Side Effects of any drug	What are the side effects of heparin?, What disease is adversely caused due to the intake of heparin?, What causes heparin?, What can be the subacute effects of cocaine?
Decrease in rate of reaction of biomedical species	what drug may inhibit the metabolism of mifepristone?

Table 5: shows the output from *InteractiveIE* pipeline after human-edits in Explorer View.

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

C Initial Slot Mapping and Evaluation

For instance, in the context passage "*Glutamate stimulates glutamate receptor interacting protein 1 degradation by ubiquitin-proteasome system to regulate surface expression of GluR2. Down-regulation of GRIP1 by glutamate was blocked by*

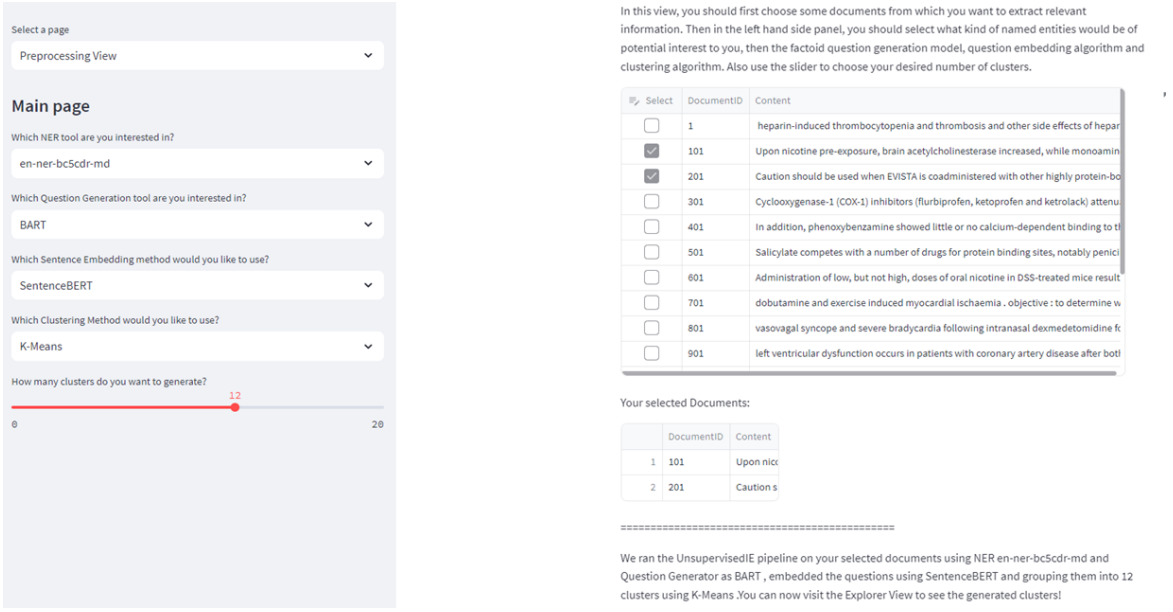


Figure 7: First step of running the preprocessing pipeline on the user-specified needs. The user can choose relevant documents, NER model, Question generation model, sentence embedding model, clustering algorithm and number of clusters to group the question-answer pairs into.

carbobenzoxyl-leucinyl-leucinyl-leucinal (MG132), a proteasome inhibitor and by expression of K48R-ubiquitin, a dominant negative form of ubiquitin. Our results suggest that glutamate induces GRIP1 degradation by proteasome through an NMDA receptor-Ca2+ pathway and that GRIP1 degradation may play an important role in regulating GluR2 surface expression.", the gold tuples annotated are: "Glutamate [SEP] downregulator GRIP1 and glutamate receptor interacting protein 1". Here, the gold slot is downregulator and the entities involved in this slot are Glutamate, GRIP1 and glutamate receptor interacting protein 1. After running *UnsupervisedIE* initially, we obtain a question-answer pair such as **Question:** *"which substance was regulated by glutamate and hence blocked by carbobenzoxyl-leucinyl-leucinyl-leucinal (MG132)?"* - **Answer:** GRIP1.

For mapping this predicted question-answer pair to an intended slot, we use fuzzy matching to map the question answer intent to one of the slots and provide the description of each slot: **"cause"**:"mention of something like what drugs cause which disease", **"downregulator"**: "decrease or inhibition effects of any biomedical drug on enzymes or other biomedical species",**"upregulator"**: "increase or rise in the effects of any biomedical drug on enzymes or other biomedical species", **"interacts with"**: "mention of any adverse effect when two or more biomedical species act together" and **"regulator":** "when there is any binding effect between biomedical species". After the mapping, we use fuzzy matching to determine if the involved entities in the gold tuple **Glutamate**, **GRIP1** and **glutamate receptor interacting protein 1** are present in the predicted QA pair. If yes, then we consider that as a true positive. We make the slot mapping evaluation with respect to gold standard slots using the standard metrics of Precision, Recall and F1-measures. If the description of a cluster description doesn't match with any desired gold slot, then we refrain from evaluating with the gold standard slots. Moreover, we also merge the results of two or more clusters if two or more clusters are mapped to similar gold slot, and then evaluate with respect to Precision, Recall and F1-measures.

D 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,

	F1 (GENEVA)	F1 (Biomedical)
GPT2	0.32	0.46
GPT3	0.35	0.48
GPT3.5-turbo	0.38	0.51

Table 6: Generalizability of our approach on three LLMs, where we report the zero-shot performance of all the models on the training set of the two datasets. We report the macro-f1 scores. Stoked by the best performance of **GPT3.5-turbo**, we conduct all our experiments in the main paper using that model.

	Ill-formed (%)
T5	0.43
BART	0.52
GPT2	0.28
GPT3	0.15
GPT3.5-turbo	0.12

Table 7: Questions automatically generated based on triggers using the models which are classified as ill-formed by a few-shot GPT-4 (serving as proxy-human) and do not map to any concrete information need.

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 12. 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.

E Ablation Analysis of UnsupervisedIE

Implementation Details of UnsupervisedIE: We use sentenceBERT (Reimers and Gurevych, 2019) to encode the passages and the queries. Our interface is developed using streamlit (Figure 7). For extracting the event triggers, we also make use of spacy-POS Tagger and nltk pos tagger to generate the triggers. For question-generation methods, we use pre-trained T5 (Raffel et al., 2019) and BART (Lewis et al., 2020) to generate questions pivoted on event triggers and different entities. We have experimented with three different LLMs such as GPT-3 (*text-davinci-003*), ChatGPT (*gpt-3.5-turbo*) and GPT-4 from OpenAI. All experiments are carried out with temperature 0 to have a reproducible setup and top-p nucleus sampling set to 0.9. More ablation results can be found in 6 and 7. Our system is generalizable with any LLMs.

F CrisisNLP Slot Filling Dataset Statistics

Chile Earthquake (1,000 tweets) had the following pairs:

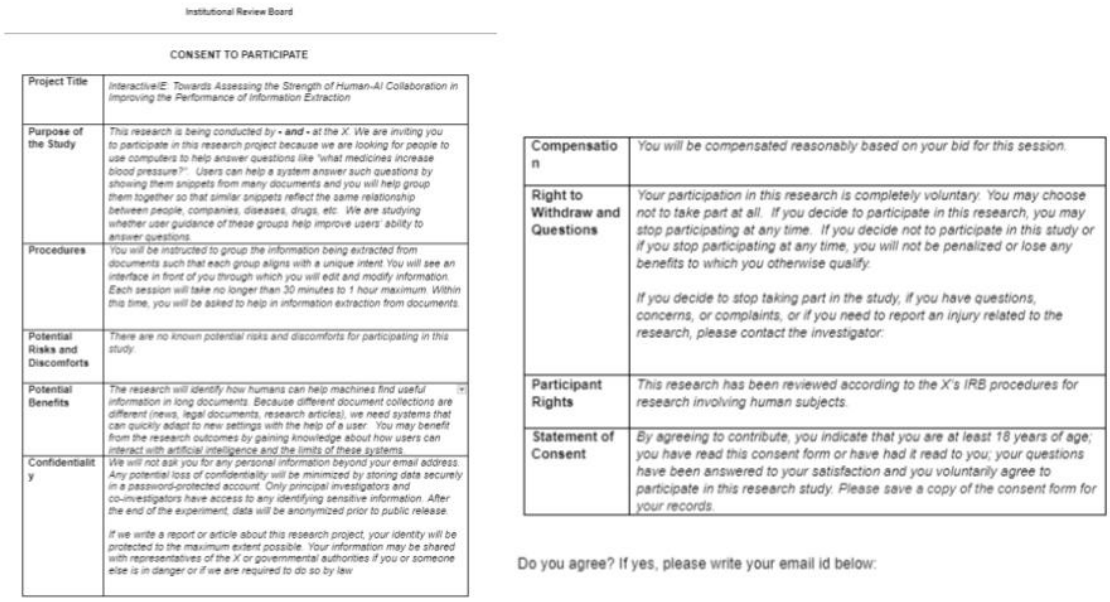
- Emergency and Supplies: 200 slot-value pairs (e.g., availability of water, food, shelter)
- Affected Areas and Evacuation: 200 slot-value pairs (e.g., specific locations hit, evacuation centers)
- Casualties and Damage: 300 slot-value pairs (e.g., death toll, infrastructure damage)
- Emotional Support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



Institutional Review Board

CONSENT TO PARTICIPATE

Project Title	*InteractiveIE: Towards Assessing the Strength of Human-AI Collaboration in Improving the Performance of Information Extraction*
Purpose of the Study	*This research is being conducted by - **and** - at the X. We are inviting you to participate in this research project because we are looking for people to use computers to help answer questions like "what medicines increase blood pressure?". Users can help a system answer such questions by showing them snippets from many documents and you will help group them together so that similar snippets reflect the same relationship between people, companies, diseases, drugs, etc. We are studying whether user guidance of these groups help improve users' ability to answer questions.*
Procedures	*You will be instructed to group the information being extracted from documents such that each group aligns with a unique intent. You will see an interface in front of you through which you will edit and modify information. Each session will take no longer than 30 minutes to 1 hour maximum. Within this time, you will be asked to help in information extraction from documents.*
Potential Risks and Discomforts	*There are no known potential risks and discomforts for participating in this study.*
Potential Benefits	*The research will identify how humans can help machines find useful information in long documents. Because different document collections are different (news, legal documents, research articles), we need systems that can quickly adapt to new settings with the help of a user. You may benefit from the research outcomes by gaining knowledge about how users can interact with artificial intelligence and the limits of these systems.*
Confidentiality	*We will not ask you for any personal information beyond your email address. Any potential loss of confidentiality will be minimized by storing data securely in a password-protected account. Only principal investigators and co-investigators have access to any identifying sensitive information. After the end of the experiment, data will be anonymized prior to public release.* *If we write a report or article about this research project, your identity will be protected to the maximum extent possible. Your information may be shared with representatives of the X or governmental authorities if you or someone else is in danger or if we are required to do so by law.*
Compensation	*You will be compensated reasonably based on your bid for this session.*
Right to Withdraw and Questions	*Your participation in this research is completely voluntary. You may choose not to take part at all. If you decide to participate in this research, you may stop participating at any time. If you decide not to participate in this study or if you stop participating at any time, you will not be penalized or lose any benefits to which you otherwise qualify.* *If you decide to stop taking part in the study, if you have questions, concerns, or complaints, or if you need to report an injury related to the research, please contact the investigator.*
Participant Rights	*This research has been reviewed according to the X's IRB procedures for research involving human subjects.*
Statement of Consent	*By agreeing to contribute, you indicate that you are at least 18 years of age; you have read this consent form or have had it read to you; your questions have been answered to your satisfaction and you voluntarily agree to participate in this research study. Please save a copy of the consent form for your records.*

Do you agree? If yes, please write your email id below:

Figure 8: Consent Form from Participants

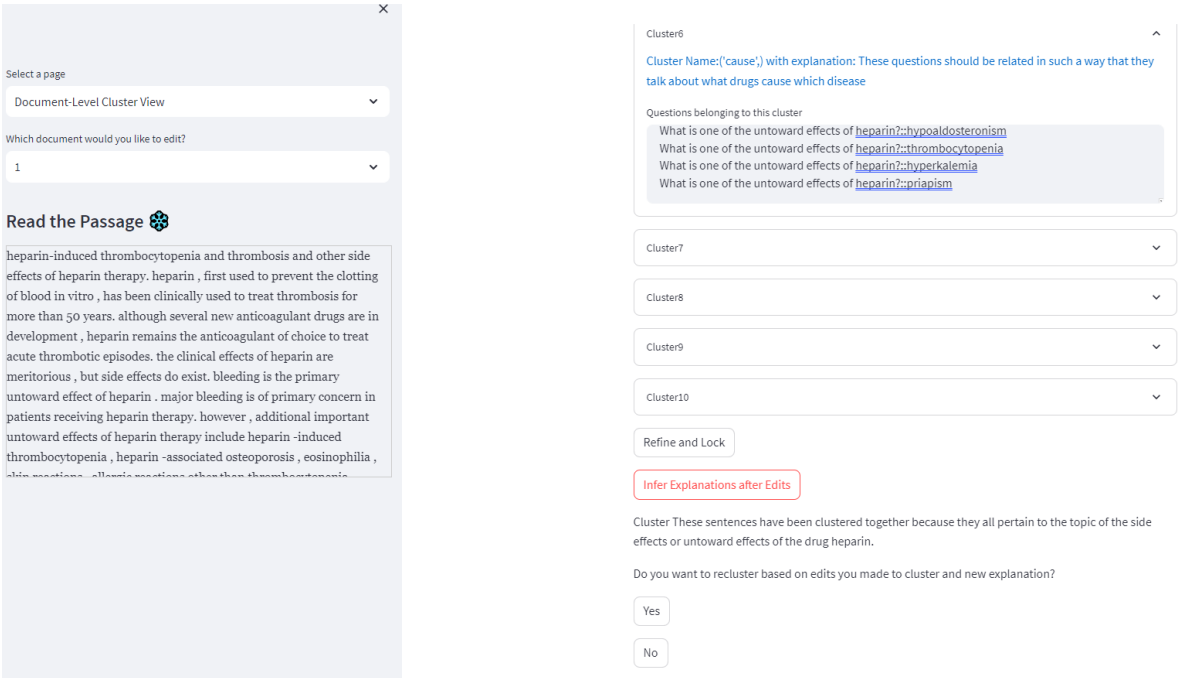


Figure 9: Infer Explanations Functionality in the "Document-Level Cluster view"

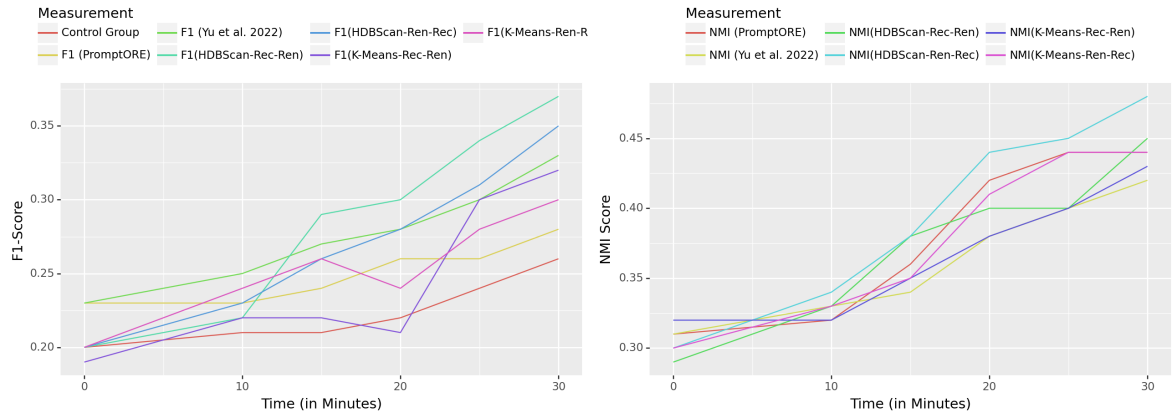


Figure 10: Average F1-scores and NMI scores achieved by ten users at different time stamps on the Biomedical Dataset. At time 0, *UnsupervisedIE* clusters are shown initially and the participants kept interacting with PERSONA-ADAPTIVE IE for 30 minutes. At certain intervals, we notice change in performance of all the configurations (macro F1).

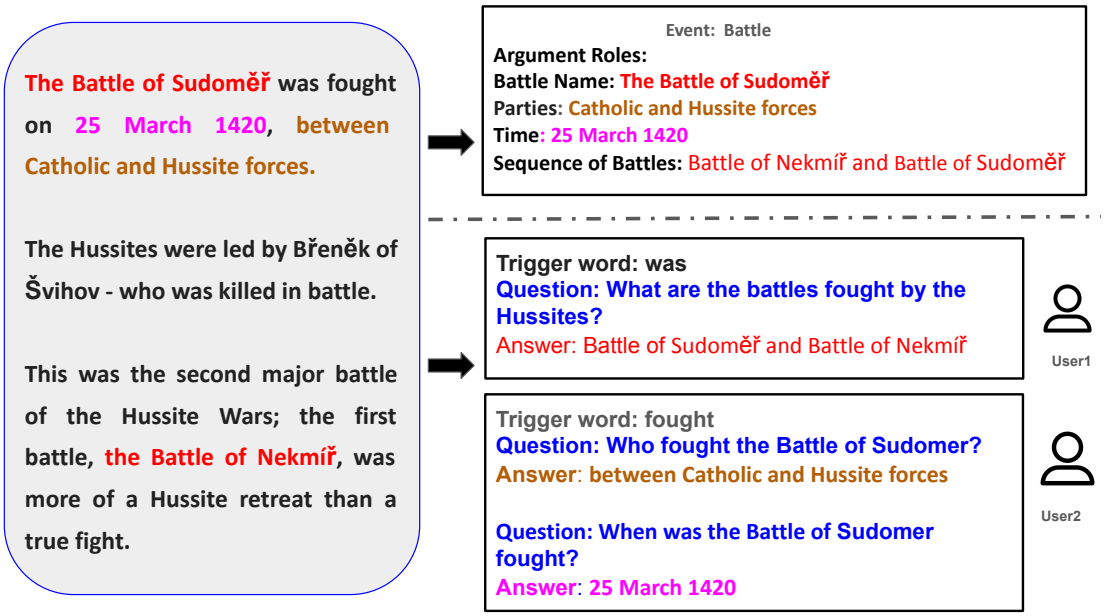


Figure 11: An example shows the motivation of using a QA-driven approach of extracting information on-the-fly depending on user requirements. Supervised template-driven approaches require pre-annotated templates, whereas QA-driven interactive pipeline using trigger words **fought** generates all possible question-answer pairs corresponding to an event, and it satisfies user's information needs on-the-fly.

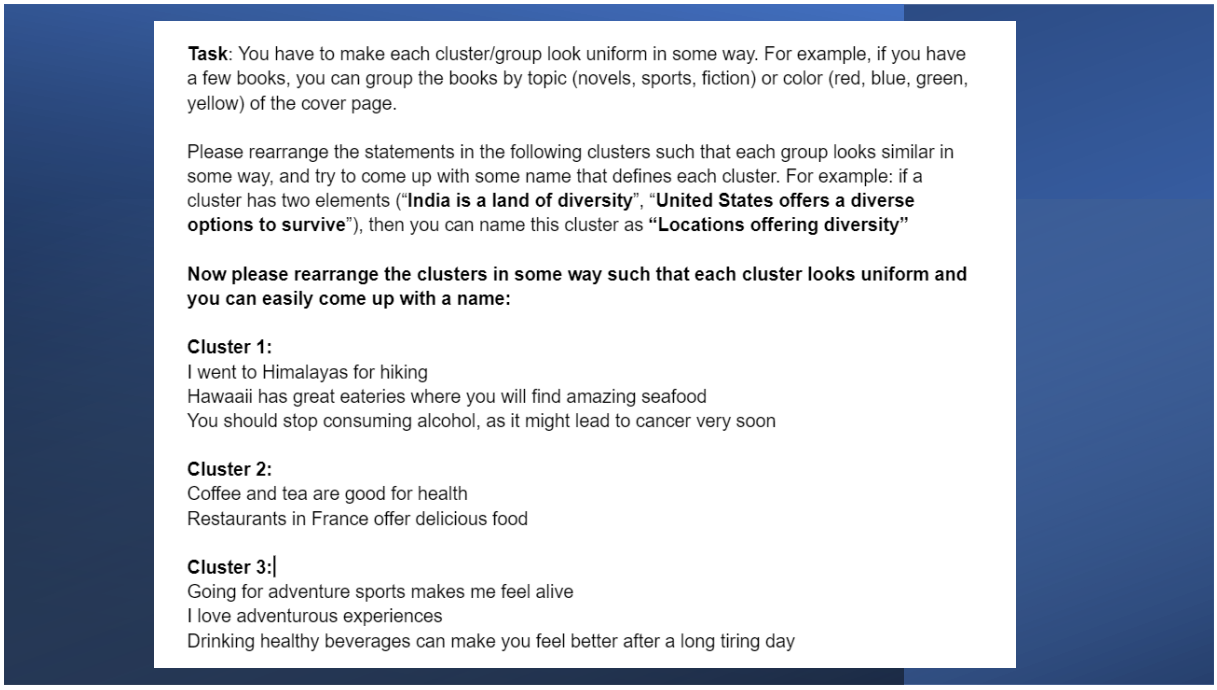


Figure 12: Clustering Assignment used for recruiting participants.

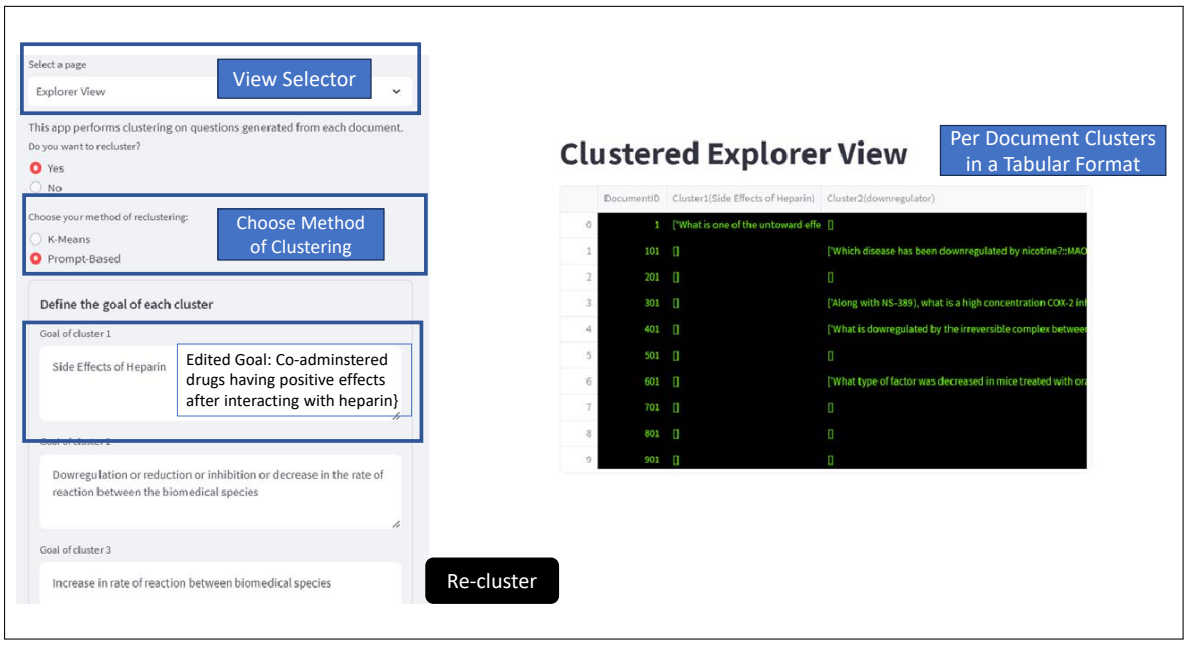


Figure 13: shows the *Explorer View* of *InteractiveIE*. Users can see the clusters generated by the model with the rationales. Based on needs, they can edit the existing goal of *"Side Effects of heparin"* to *"Co-administered drugs having positive effects after interacting with heparin"*. Then users can find the new set of clusters by pressing "Recluster Button". Based on the goals, the clusters have been named to some slot such as the goal of cluster 3 *"Increase in rate of reaction between biomedical species"* to *"Upregulation"* as seen in Clustered Explorer View.

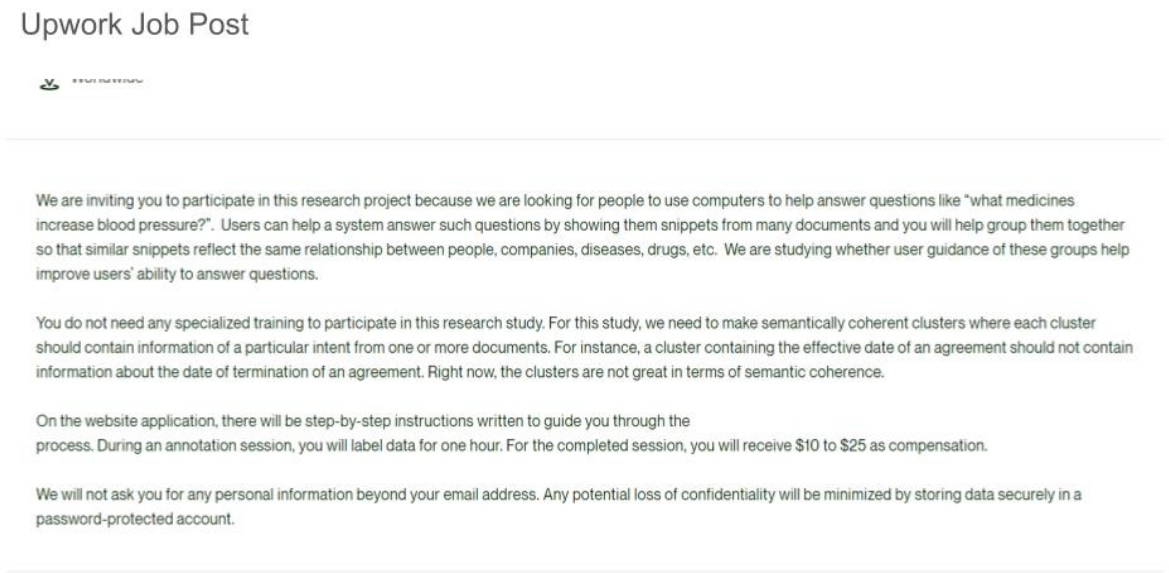


Figure 14: Upwork Job Post.

Algorithm 1 Iterative Clustering with User Feedback

Require: $Q = \{q_1, q_2, \ldots, q_n\}$ (set of questions), U (user providing feedback)
Ensure: Refined clusters $C^{(t)}$ aligned with user feedback
Initialize $t \leftarrow 0$
Perform initial clustering $C^{(0)}$ using Q
repeat
P
until
;
resent clusters $C^{(t)}$ to user U
$F^{(t)} \leftarrow$ Collect feedback from U on $C^{(t)}$ **for** *each feedback* $f \in F^{(t)}$ **do**
feedback suggests reassignment of q_i
Identify $C_a^{(t)}$ and $C_b^{(t)}$
$C_a^{(t+1)} \leftarrow C_a^{(t)} \setminus \{q_i\}$
$C_b^{(t+1)} \leftarrow C_b^{(t)} \cup \{q_i\}$
feedback involves constraints
Apply constraint-based reclustering
feedback involves naming
Apply naming-based classification

$t \leftarrow t + 1$
Check for convergence
clustering meets user's objectives or a maximum number of iterations is reached

Algorithm 2 Constraint-based Reclustering

Require: Q, M, N, $C^{(t)}$ (current clusters), $F^{(t)}$ (user feedback on constraints)
Ensure: Updated clusters $C^{(t+1)}$ respecting constraints **for** *each constraint in* $F^{(t)}$ **do**
must-have constraint
Update clusters to ensure specified questions are in the same cluster
cannot-have constraint
Update clusters to separate specified questions

Recalculate centroids for updated clusters
Reassign questions to nearest centroid while respecting constraints

Algorithm 3 Naming-based Classification

Require: Q, N_i (cluster names), $C^{(t)}$ (current clusters), e (embedding function)
Ensure: Updated clusters $C^{(t+1)}$ based on names **for** *each* $q_j \in Q$ **do**

Compute $e(q_j)$
Assign q_j to cluster C_i with highest cosine similarity between $e(q_j)$ and $e(N_i)$
=0