Large Language Models Are Effective Human Annotation Assistants, But Not Good Independent Annotators

Feng Gu[1] **Zongxia Li**[1] **Carlos Rafael Colon**[2]
Benjamin Evans[2] **Ishani Mondal**[1] **Jordan Lee Boyd-Graber**[1]
[1]Department of Computer Science, University of Maryland
[2]National Consortium for the Study of Terrorism and Responses to Terrorism
{fgu1, zli12321, raffy, benevans, imondal, ying}@umd.edu

Abstract

Event annotation is important for identifying, monitoring, and understanding sociological trends. Although expert annotators set the gold standard, they are expensive and inefficient. While state-of-the-art NLP models are an attractive alternative, they are often evaluated on standalone subtasks rather than entire workflows. Thus, we evaluate a holistic workflow that summarizes news with event coreference resolution and argument extraction in three modes: AI-only, AI assistance, and human only. Although AI's recall is seven times higher than the TF-IDF baseline at coreference resolution, it is far from replacing experts. However, experts *adopt* AI-extracted arguments 60% of the time, reducing extraction time by 25%. Our code and data are in https://github.com/Obertura777/gtd-data.

1 Making Event Annotation Realistic

Quality data is crucial for informed decisions (Akella et al., 2025; Zhang et al., 2025). Bloomberg[1] and LSEG,[2] for example, gather top-quality data from trusted sources to monitor market trends, while companies like Scale AI[3] collect, curate, and annotate data for AI models.

An important but complex step in data processing is cross-document coreference resolution (Grosz and Sidner, 1986): finding unique incidents from duplicate and conflicting documents. While humans excel at this step, machines handle larger inputs and improve efficiency and scalability (King and Lowe, 2003; Mason et al., 2012). The Uppsala Conflict Data Project, one of the earliest event dataset programs, covers fewer than 300,000 events with human labor. In comparison, automated processes like GDELT (Leetaru and Schrodt, 2013) collect trillions of events despite starting later. Such datasets enable organizations to monitor trends and detect anomalies, driving lasting impact in their communities.

Automating event annotation presents significant challenges. Models struggle with text extraction (Schrodt and Van Brackle, 2013): empirical evaluations yield low confidence and inconsistent results (O'Brien, 2013). Annotating original sources requires expertise in synthesizing information from multiple documents (Gao et al., 2024; Li et al., 2024a), often involving steps beyond labeling. Realistic event datasets in social science research typically combine manual efforts (Elliott, 2018; Pierre and Jackson, 2014) with quantitative methods (Probierz et al., 2022; Li et al., 2025) to ensure consistency and data quality.

While large language models (LLMs) are more powerful than methods used in real-world applications (Chen et al., 2023; Zhao et al., 2025), existing pipelines such as ACLED and ICEWS do not use them. First, data characteristics differ. Unlike deployed workflows that collect millions of documents periodically (Raleigh et al., 2010; Sundberg, 2013; García-Durán et al., 2018; Leetaru and Schrodt, 2013), NLP datasets like ACE and ECB have fixed sizes, few distinct components, and little lexical diversity (Zhukova et al., 2022; Doddington et al., 2004). They also use secondary sources (Ahmed et al., 2024), under-trained workers (Sharif et al., 2024), and document-level scope (Ebner et al., 2020; Li et al., 2021). Lacking source diversity introduces overfitting: zero-shot GPT-4o-mini achieves 0.83 F_1 on the six event types of the Gun Violence Corpus (Vossen et al., 2018) and 63% accuracy on 710 examples in MAVEN-ARG (Wang et al., 2024) in our test, showing these tasks are becoming less challenging for fast-evolving models.

Second, specialized models are inaccessible. They are scheme-dependent and quickly become obsolete as the field changes (Chen et al., 2023;

[1]https://www.bloomberg.com/
[2]https://www.lseg.com/
[3]https://scale.com/

Findings of the Association for Computational Linguistics: ACL 2026, pages 71–89

Li et al., 2021). Reproducing their results is also difficult because open-source models often omit components, inflate results by overfitting, and generalize poorly (Gao et al., 2024; Liu et al., 2024; Bugert et al., 2021; Cattan et al., 2021b). A general-purpose model is easier to adapt for organizations where these realistic annotations take place.

Third, NLP research on event annotation focuses on dedicated tasks but not holistic systems. While focusing on one task (e.g., event extraction) allows dedicated contributions, it undermines downstream applicability for social science annotations, which chain individual tasks into a complete pipeline. Without reliably combining components like filtering, retrieval, and coreference resolution, end-to-end applicability remains low: actively-maintained pipelines still use statistical methods and rely on trained annotators (Armed Conflict Location & Event Data Project (ACLED), 2017; O'Brien, 2010).

To address these challenges, we present a case study on Global Terrorism Database (GTD),[4] an open-source event database used to systematically study terrorism events. This dataset reflects current annotation practices in social science research: it is large, recent, noisy, first-hand, and expert-annotated. We detail the operational background of this workflow in Section 2.

We evaluate this pipeline by comparing a manual workflow with one involving LLMs. We test LLM capabilities in coreference resolution (Section 3), then measure their impact on event argument extraction (Section 4). Using operational metrics like variable extraction frequency, F_1 against expert annotations, and reduction in annotation time, we show that while LLMs do not reach expert quality, even a small model effectively assists trained annotators by reducing annotation time.

Finally, we outline how to practically achieve LLM integrations by summarizing the common errors LLMs still produce (Section 5), highlighting mitigation strategies for live deployment (Section 5.1), and contextualizing our approach within related work (Section 6).

2 A Realistic Workflow

From a list of unorganized documents, our goal is to form a set of incidents with annotated variables and references to relevant documents. To achieve this, annotators start with a massive but noisy pool of documents gathered through automated ingestion, which removes easily identifiable duplicates and calculates document similarities before manual work begins.

Once the automated process reduces the document count, experts engage in the first manual step: Event Set Curation. Experts identify unique events from all documents and construct event sets, sets of documents about single events like violent protests.

This process is akin to cross-document coreference resolution (Bagga and Baldwin, 1998). Annotators must review all documents to consolidate scattered details, as single documents often contain only partial, complementary, or even conflicting information. To help find similar documents, the automated system generates relevance scores via TF-IDF and nearest-neighbor search, allowing annotators to view similar texts (Figure 6).

Because we need consistent event sets to fairly evaluate downstream annotations, we first isolate this curation step. We evaluate candidate event sets using precision, recall, and F_1 against expert-created baselines, giving a clear metric for model helpfulness in clustering.

Once event sets are established, the second step is event argument extraction, which GTD calls Variable Coding. Here, experts synthesize the curated text to extract domain variables defined in a formal codebook. These variables range from free-text descriptions (e.g., target of "police vehicle" in a terrorism event in Pakistan) and enumerated categories (e.g., attack types such as bombing, armed assault, and facility/infrastructure attack) to numerical values like casualty counts.

To measure the impact of LLM assistance on this step, we test three experimental conditions: humans annotating alone, humans vetting LLM-generated outputs, and LLM outputs alone, following conventions in human-AI collaboration (Sheridan and Verplank, 1978; Rebensky et al., 2022). We assess practical utility by measuring total annotation time with and without LLM-coded variables, tracking how often humans accept LLM suggestions, and calculating overall annotation accuracy.

3 Methods for Event Set Curation

To evaluate the effectiveness of LLMs in Event Set Curation, experts manually review 500 documents from February 2022[5] to identify events and create

[4] https://www.start.umd.edu/data-tools/GTD

[5] We deliberately sampled from this period because the onset of the Russo-Ukrainian war generates many terrorism-

	Precision	Recall	F_1	Same Sets
TF-IDF	0.19	0.09	0.10	12
Pairwise				
EMBEDDING	**0.89**	0.51	0.59	66
LLM-CLS	0.36	0.35	0.35	105
+SEG	0.65	**0.66**	**0.63**	**203**
K-LLMMEANS				
TEXT-EMBEDDING-3-SMALL	0.36	0.45	0.38	85
GEMINI-EMBEDDING-001	0.32	0.36	0.32	60
DISTILBERT	0.12	0.10	0.09	7
MODERNBERT	0.20	0.20	0.16	15

Table 1: In Event Set Curation, EMBEDDING has the highest precision of creating event sets. LLM-CLS+SEG shows superior recall and a higher overall F_1 score. Additionally, compared to 371 annotatedevent sets, LLM-CLS+SEG generates event sets that align most closely with expert annotations. K-LLMMEANS is more cost-effective than LLM-CLS with better recall and F_1.

gold-standard event sets. Given the same documents, we then use three methods to generate event sets that contain these documents. We compare them against this gold standard.

LLM-CLS (Pairwise Classification): GPT-4o-mini[6] classifies document pairs. It compares a candidate document not yet in the output event sets against a reference document to predict whether both texts describe the exact same event. This method also includes an optional pre-processing step (SEG) that uses the LLM to segment long, multi-topic digest documents into single-event text blocks before classification.

EMBEDDING (Embedding Similarity): We group related documents based on their overarching semantic proximity rather than relying on exact keyword matches. To achieve this, we calculate the pairwise cosine similarity of vector embeddings between a new document and the existing members of a cluster. A document joins the cluster if its similarity score to any member exceeds 0.859, a threshold established by a grid search over a separate, held-out validation set that optimize for the higest F_1 score.[7]

K-LLMMEANS: This method uses K-MEANS with LLM summarization to iteratively update cluster centroids and assign documents (Diaz-Rodriguez, 2026). We show our configuration and a summary example in Appendix A.16 and A.17.

related events not previously accounted for.

[6]We choose it for its availability and cost-effectiveness as larger models do not necessarily yield significantly better results (Appendix A.13).

[7]We show the algorithm in Appendix A.9.

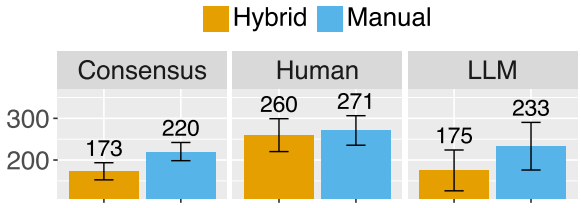

Figure 1: Experts need less time (in seconds) in Variable Coding when they have access to an LLM (i.e., hybrid) under all three conditions. Consensus are the event sets that human and AI agree, in which the time taken is the lowest. Annotators take less time in the LLM-generated event sets because many are invalid. When experts and LLM-CLS+SEG agree on event sets, they take the least amount of time with lower deviations.

3.1 Improving Event Set Curation

EMBEDDING captures semantic relationships. Mapping text into dense vector spaces, EMBEDDING captures the underlying meaning rather than relying on word matches. This allows it to group documents effectively and achieves 0.89 precision, a significant improvement over the traditional TF-IDF baseline (i.e., scores generated by the existing automated system). EMBEDDING reduces false negatives by 82% and successfully identifies 51% of all relevant documents.

Segmentation helps. Real-world documents are messy; news digests contain multiple unrelated events. Dividing them into discrete event sections before classification improves Event Set Curation. Segmentation reduces ambiguity and prevents the LLM from misidentifying events, yielding the best recall and F_1 score (Table 1).

K-LLMMEANS has higher accuracy and lower cost than LLM-CLS. Although segmentation paired with pairwise classification increases recall, binary predictions are not cost-effective for large corpora because of quadratic complexity. However, by combining high-quality embeddings with LLM-generated summaries, K-LLMMEANS achieves a higher recall and F_1 score than LLM-CLS at less than one-third of the cost.

4 Variable Coding with LLMs

A holistic workflow includes multiple NLP tasks. The quality of Event Set Curation determines the success of Variable Coding. In Event Set Curation, we prioritize recall over precision because it ensures annotators have coverage of relevant documents in this stage. While EMBEDDING has the highest precision, LLM-CLS+SEG has the highest

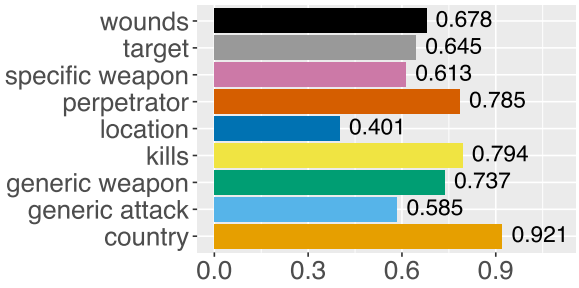

Figure 2: How often expert annotators use LLM-coded variables when given the chance in Variable Coding. While the selection frequency varies, even the least selected variable, `location`, is useful 40% of the time.

recall and generates many event sets identical to expert annotations (203 out of 371, Table 1).

In Variable Coding, we assess how flawed event sets affect annotation by replacing some expert-curated events with the most similar LLM-generated ones. This creates three distinct types of event sets (Figure 1): MANUAL (human-created), LLM (LLM-generated), and CONSENSUS. The CONSENSUS sets, where experts and LLM-CLS+SEG concur, allow us to compare the pipeline independent of the event set creation method.

For MANUAL, LLM, and CONSENSUS respectively, we compare the variables coded by experts operating in two conditions: a hybrid setting (with access to LLM suggestions) and a manual setting (without LLM assistance). Experts code nine variables for 212 event sets.

Because humans show "automation bias" by over-relying on LLM-coded variables (An et al., 2023), we compare inter-human agreement among three teams against the human-LLM agreement.

Traditional exact-match metrics are overly strict (Kocmi et al., 2021; Chen et al., 2020), so we use three evaluation methods:

Normalized Match (NM): Checks if two variables are identical after stripping punctuation, converting to lowercase, and normalizing whitespace.

BERT Match (BEM): Measures embedding similarity (Bulian et al., 2022).

PEDANTS: Uses F_1 score and TF-IDF to measure the variable match (Li et al., 2024b).

4.1 LLMs Help Experts Code Variables

LLM-coded variables drastically reduce annotation time. Experts annotate fastest when their judgments align with the LLM-CLS+SEG clusters (Figure 1). Conversely, coding event variables from purely human-created event sets takes the longest. Despite errors in LLM-coded variables, simply providing these suggestions to annotators reduces annotation time by 25%.

Annotators use LLM-coded variables two-thirds of the time. Presented with options, experts rely heavily on the LLM. They choose the suggestion for the `Country` variable 92% of the time. The higher agreement in the hybrid setting across all event set types shows that LLM-coded variables aid annotation (Figure 3).

Experts agree with LLM-coded variables over 55% of the time, even without seeing them (Figure 5). Because the human-LLM agreement nearly matches human-human agreement rate, the LLM-coded variables provide a near-human level utility.

Model size does not strongly correlate with accuracy. We test open- and closed-source LLMs, expecting larger models to dominate, but none are significantly better. Notably, MISTRAL 8X7B shows low accuracy for numerical variables (`kills` and `wounds`). We advise researchers to benchmark models against their specific codebook if the target variables include non-textual ones.

Models hallucinate when data are missing. We calculate precision and recall for scenarios where a variable is missing from the source incident. If no casualty count is mentioned, the correct annotation is "Not Available" (NA). High precision means when the model reports a variable as "NA", it is rarely wrong. High recall means the model correctly identifies most of the situations where the information is truly not available instead of hallucinating a false value. For most models tested, precision hovers around 0.5 and recall sits around 0.25, showing strong hallucination. The outlier is MIXTRAL-8X7B: showing high precision (0.62) but abysmal recall (0.03), meaning it almost always hallucinate an answer (Table 5).

5 Error Analysis

In Event Set Curation, LLM-generated event sets correlate poorly with expert judgment, and even the best model falls short of expert accuracy. We describe some error types below and show examples in Appendix A.11.

Under-specified Instructions: Unclear prompts lead to imprecise and missing details. For example, experts frequently select the suggested `country` variable but rarely `location` because `country` is rigidly defined while `location` is vague, ranging

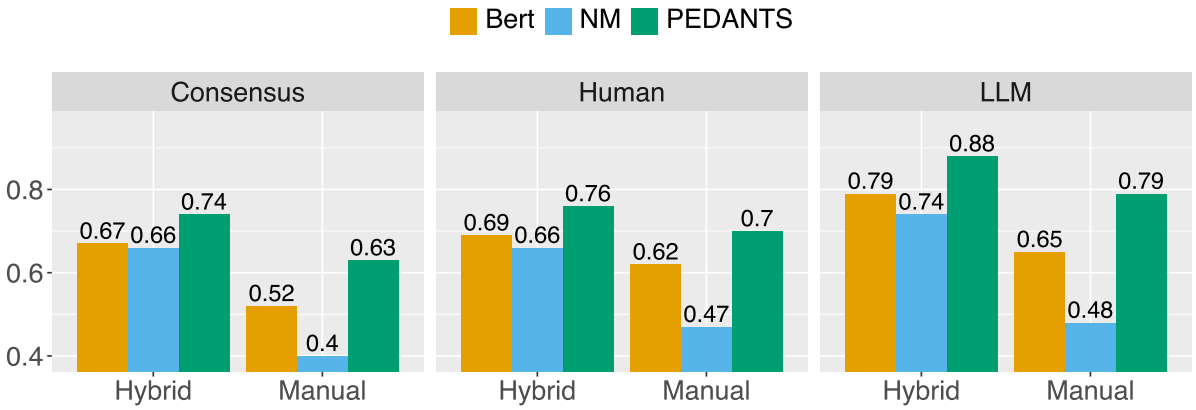

Figure 3: Models help annotate the variables. Annotators show higher agreement in the hybrid setting, where LLM-coded variables are available. The variables prove particularly beneficial in LLM-generated event sets, which often contain misinformation.

from broad regions to specific sites.

Source Document Ambiguity and Temporal Conflict: Dense cross-subtopic links (2.95 documents per event on average) mean digest documents summarize multiple events, misleading GPT-4o-mini. A digest document, common in our data, can also summarize multiple events, which misleads GPT-4o-mini into annotating the wrong event. Additionally, temporal components also appear in the news cycle.

Interpretive Subjectivity: Human judgments affect Variable Coding. LLMs lack access to the GTD codebook conventions that disambiguate overlapping categories.

5.1 Mitigation Strategies

Refined prompts reduce these errors. Providing LLMs with contexts, additional instructions and task information makes them more robust.

6 Related Work

Automated Event Data Annotation: Early efforts to scale event data collection incorporate traditional NLP techniques, such as statistical retrieval systems, to efficiently filter content and automate extractions (D'Orazio et al., 2014; Mason et al., 2012; Boschee et al., 2013). As event attributes become more granular and complex, automated extraction accuracy rapidly decreases (Jenkins and Maher, 2016). Automated datasets are notorious for containing duplicates, erroneously including non-events, and generating false positives, largely stemming from geo-localization errors and rigid syntactic parsing (Hammond and Weidmann, 2014; Miller et al., 2022; Raleigh et al., 2023). Consequently, accurate, high-fidelity event databases still require human labor. Our work bridges this gap by positioning LLMs not as fully autonomous replacements, but as expert assistants.

LLMs for Event Tasks: Prior work shows LLMs are cost-effective document-level annotators for event (Chen et al., 2024) and argument extraction (Shuang et al., 2024; Zhang et al., 2024; Zhu et al., 2025; Wang et al., 2021). However, LLMs struggle with financial data (Tseng et al., 2024) and word semantics (Yadav et al., 2024). While Zhao et al. (2023) show that GPT-4 achieves higher accuracy than under-trained crowd workers, we apply LLMs to a more complex workflow where annotators must first identify relevant events before extracting arguments from them.

7 Conclusion

Realistic event annotation remains a complex, resource-intensive process that depends on trained human labor. We present a case study to integrate LLMs to alleviate human effort in a holistic pipeline. In Event Set Curation, we show that LLM-based segmentation and clustering methods achieve higher precision and recall than the TF-IDF baseline, hence helping annotators to find similar documents, despite not able to automate the workflow. During Variable Coding, we find that providing LLM-coded variables to experts reduce annotation time by 25%. Future work can explore fine-tuning with task rewards or domain-specific datasets to further encourage LLM adoption. Task-agnostic prompt-tuning and few-shot learning offer paths to increase model accuracy (Khattab et al., 2024; Wang et al., 2022; Gao et al., 2021).

8 Limitations

NLP techniques have not yet reached human-level accuracy in document classification. In Event Set Curation, the computational cost of pairwise similarity increases quadratically with the linear growth in the number of documents, rendering LLM-based methods inefficient for large document sets (except K-LLMMEANS). Although it is impractical to fully replace TF-IDF, LLM-based methods are useful in finding semantically similar documents when TF-IDF are the same. To alleviate computational efforts, instead of computing pairwise embedding similarity for every document, we only compute pairwise embedding similarity for documents above the TF-IDF threshold. For Variable Coding, annotation requires greater granularity to meet criteria outlined in specified guidelines. These steps are essential for implementing LLM-assisted large-scale event data collection workflows. Additionally, we have not tested LLMs in cross-document event coding, which is a useful step for mitigating Variable Coding errors.

Acknowledgments

We thank researchers from GTD. Specifically, our thank goes to Jacob Scott Loewner, Margaret A. Hayden, Oleksiy Krylyuk, and Tyler Yates for data annotation and its discussion. We thank Brian Wingenroth and Dr. Amy Pate for their insights on the GTD workflow. This research was funded in part by the Department of Defense under award no. HQ003421F0481. Any opinions, findings, and conclusions or recommendations expressed in this report are those of the authors and do not necessarily reflect the views of the Department of Defense.

References

Shafiuddin Rehan Ahmed, Zhiyong Eric Wang, George Arthur Baker, Kevin Stowe, and James H. Martin. 2024. Generating harder cross-document event coreference resolution datasets using metaphoric paraphrasing. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 276–286, Bangkok, Thailand. Association for Computational Linguistics.

Ashlesha Akella, Abhijit Manatkar, Krishnasuri Narayanam, and Sameep Mehta. 2025. CodeGenWrangler: Data wrangling task automation using code-generating models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, pages 949–960, Albuquerque, New Mexico. Association for Computational Linguistics.

Haozhe An, Zongxia Li, Jieyu Zhao, and Rachel Rudinger. 2023. SODAPOP: Open-ended discovery of social biases in social commonsense reasoning models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1573–1596, Dubrovnik, Croatia. Association for Computational Linguistics.

Armed Conflict Location & Event Data Project (ACLED). 2017. Acled methodology. Technical report, ACLED. Accessed 2026-04-14.

Dennis Aumiller, Satya Almasian, Sebastian Lackner, and Michael Gertz. 2021. Structural text segmentation of legal documents. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Law*, ICAIL '21. ACM.

Amit Bagga and Breck Baldwin. 1998. Entity-based cross-document coreferencing using the vector space model. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 1*, pages 79–85, Montreal, Quebec, Canada. Association for Computational Linguistics.

Shany Barhom, Vered Shwartz, Alon Eirew, Michael Bugert, Nils Reimers, and Ido Dagan. 2019. Revisiting joint modeling of cross-document entity and event coreference resolution. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4179–4189, Florence, Italy. Association for Computational Linguistics.

Ari Bornstein, Arie Cattan, and Ido Dagan. 2020. CoRefi: A crowd sourcing suite for coreference annotation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 205–215, Online. Association for Computational Linguistics.

Elizabeth Boschee, Premkumar Natarajan, and Weischedel Ralph. 2013. *Automatic Extraction of Events from Open Source Text for Predictive Forecasting*, pages 51–67. Springer New York, New York, NY.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Jannis Bulian, Christian Buck, Wojciech Gajewski, Benjamin Börschinger, and Tal Schuster. 2022. Tomayto, tomahto. beyond token-level answer equivalence for question answering evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 291–305, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2020. Streamlining cross-document coreference resolution: Evaluation and modeling. *ArXiv*, abs/2009.11032.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021a. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107, Online. Association for Computational Linguistics.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021b. Realistic evaluation principles for cross-document coreference resolution. In *Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics*, pages 143–151, Online. Association for Computational Linguistics.

Anthony Chen, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2020. MOCHA: A dataset for training and evaluating generative reading comprehension metrics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6521–6532, Online. Association for Computational Linguistics.

Ruirui Chen, Chengwei Qin, Weifeng Jiang, and Dongkyu Choi. 2024. Is a large language model a good annotator for event extraction? *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17772–17780.

Xinyu Chen, Sheng Xu, Peifeng Li, and Qiaoming Zhu. 2023. Cross-document event coreference resolution on discourse structure. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4833–4843, Singapore. Association for Computational Linguistics.

Jairo Diaz-Rodriguez. 2026. Summaries as centroids for interpretable and scalable text clustering. In *The Fourteenth International Conference on Learning Representations*.

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

Vito D'Orazio, Steven T. Landis, Glenn Palmer, and Philip Schrodt. 2014. Separating the wheat from the chaff: Applications of automated document classification using support vector machines. *Political Analysis*, 22(2):224–242.

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

Alon Eirew, Avi Caciularu, and Ido Dagan. 2022. Cross-document event coreference search: Task, dataset and modeling. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 900–913, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Victoria Elliott. 2018. Thinking about the coding process in qualitative data analysis. *Qualitative Report*, 23:2850–2861.

Qiang Gao, Zixiang Meng, Bobo Li, Jun Zhou, Fei Li, Chong Teng, and Donghong Ji. 2024. Harvesting events from multiple sources: Towards a cross-document event extraction paradigm. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1913–1927, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2021. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3816–3830, Online. Association for Computational Linguistics.

Alberto García-Durán, Sebastijan Dumančić, and Mathias Niepert. 2018. Learning sequence encoders for temporal knowledge graph completion. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4816–4821, Brussels, Belgium. Association for Computational Linguistics.

Barbara J. Grosz and Candace L. Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Jesse Hammond and Nils B. Weidmann. 2014. Using machine-coded event data for the micro-level study of political violence. *Research & Politics*, 1(2).

Marti A. Hearst. 1997. Texttiling: segmenting text into multi-paragraph subtopic passages. *Comput. Linguist.*, 23(1):33–64.

J. Craig Jenkins and Thomas V. Maher. 2016. What should we do about source selection in event data? challenges, progress, and possible solutions. *International Journal of Sociology*, 46(1):42–57.

Heng Ji, Ralph Grishman, Zheng Chen, and Prashant Gupta. 2009. Cross-document event extraction and tracking: Task, evaluation, techniques and challenges. In *Proceedings of the International Conference RANLP-2009*, pages 166–172, Borovets, Bulgaria. Association for Computational Linguistics.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2024. Dspy: Compiling declarative language model calls into self-improving pipelines.

Gary King and Will Lowe. 2003. An automated information extraction tool for international conflict data with performance as good as human coders: A rare events evaluation design. *International Organization*, 57(3):617–642.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. 2018. Text segmentation as a supervised learning task. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 469–473, New Orleans, Louisiana. Association for Computational Linguistics.

Kalev Leetaru and Philip A Schrodt. 2013. Gdelt: Global data on events, location, and tone. In *Proceedings of the International Studies Association Annual Conference*, San Diego, CA.

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

Zongxia Li, Lorena Calvo-Bartolomé, Alexander Hoyle, Paiheng Xu, Daniel Stephens, Alden Dima, Juan Francisco Fung, and Jordan Boyd-Graber. 2025. Large language models struggle to describe the haystack without human help: A social science-inspired evaluation of topic models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7583–7604, Vienna, Austria. Association for Computational Linguistics.

Zongxia Li, Andrew Mao, Daniel Stephens, Pranav Goel, Emily Walpole, Alden Dima, Juan Fung, and Jordan Boyd-Graber. 2024a. Improving the TENOR of labeling: Re-evaluating topic models for content analysis. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 840–859, St. Julian's, Malta. Association for Computational Linguistics.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024b. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics.

Wanlong Liu, Li Zhou, DingYi Zeng, Yichen Xiao, Shaohuan Cheng, Chen Zhang, Grandee Lee, Malu Zhang, and Wenyu Chen. 2024. Beyond single-event extraction: Towards efficient document-level multi-event argument extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9470–9487, Bangkok, Thailand. Association for Computational Linguistics.

Richard Mason, Brian McInnis, and Siddhartha Dalal. 2012. Machine learning for the automatic identification of terrorist incidents in worldwide news media. In *2012 IEEE International Conference on Intelligence and Security Informatics*, pages 84–89.

Erin Miller, Roudabeh Kishi, Clionadh Raleigh, and Caitriona Dowd. 2022. An agenda for addressing bias in conflict data. *Scientific Data*, 9(593).

Sean P. O'Brien. 2010. Crisis early warning and decision support: Contemporary approaches and thoughts on future research. *International Studies Review*, 12(1):87–104.

Sean P. O'Brien. 2013. *A Multi-Method Approach for Near Real Time Conflict and Crisis Early Warning*, pages 401–418. Springer New York, New York, NY.

Elizabeth A. St. Pierre and Alecia Y. Jackson. 2014. Qualitative data analysis after coding. *Qualitative Inquiry*, 20(6):715–719.

Barbara Probierz, Jan Kozak, and Anita Hrabia. 2022. Clustering of scientific articles using natural language processing. *Procedia Computer Science*, 207:3449–3458. Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 26th International Conference KES2022.

Clionadh Raleigh, Roudabeh Kishi, and Andrew Linke. 2023. Political instability patterns are obscured by conflict dataset scope conditions, sources, and coding choices. *Humanities and Social Sciences Communications*, 10(74).

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

Summer Rebensky, Kendall Carmody, Cherrise Ficke, Meredith Carroll, and Winston Bennett. 2022. Teammates instead of tools: The impacts of level of autonomy on mission performance and human–agent teaming dynamics in multi-agent distributed teams. *Frontiers in Robotics and AI*, Volume 9 - 2022.

Martin Riedl and Chris Biemann. 2012. Topictiling: A text segmentation algorithm based on lda. In *Proceedings of the Student Research Workshop of the 50th Meeting of the Association for Computational Linguistics*, pages 37–42, Jeju, Republic of Korea.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Philip A. Schrodt and David Van Brackle. 2013. *Automated Coding of Political Event Data*, pages 23–49. Springer New York, New York, NY.

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

Thomas B. Sheridan and William L. Verplank. 1978. Human and computer control of undersea teleoperators.

Kai Shuang, Zhouji Zhouji, Wang Qiwei, and Jinyu Guo. 2024. Thinking about how to extract: Energizing LLMs' emergence capabilities for document-level event argument extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5520–5532, Bangkok, Thailand. Association for Computational Linguistics.

Ralph Sundberg. 2013. Introducing the ucdp georeferenced event dataset. *Journal of Peace Research*, 50(4):523–532.

Yu-Min Tseng, Wei-Lin Chen, Chung-Chi Chen, and Hsin-Hsi Chen. 2024. Are expert-level language models expert-level annotators? *Preprint*, arXiv:2410.03254.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: a data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Jianing Wang, Chengyu Wang, Fuli Luo, Chuanqi Tan, Minghui Qiu, Fei Yang, Qiuhui Shi, Songfang Huang, and Ming Gao. 2022. Towards unified prompt tuning for few-shot text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 524–536, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaozhi Wang, Hao Peng, Yong Guan, Kaisheng Zeng, Jianhui Chen, Lei Hou, Xu Han, Yankai Lin, Zhiyuan Liu, Ruobing Xie, Jie Zhou, and Juanzi Li. 2024. MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4072–4091, Bangkok, Thailand. Association for Computational Linguistics.

Benjamin Warner, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Griffin Thomas Adams, Jeremy Howard, and Iacopo Poli. 2025. Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2526–2547, Vienna, Austria. Association for Computational Linguistics.

Sachin Yadav, Tejaswi Choppa, and Dominik Schlechtweg. 2024. Towards automating text annotation: A case study on semantic proximity annotation using gpt-4. *Preprint*, arXiv:2407.04130.

Tao Zhang, Yige Wang, ZhuHangyu ZhuHangyu, Li Xin, Chen Xiang, Tian Hua Zhou, and Jin Ma. 2025. WebQuality: A large-scale multi-modal web page quality assessment dataset with multiple scoring dimensions. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 583–596, Albuquerque, New Mexico. Association for Computational Linguistics.

Xinliang Frederick Zhang, Carter Blum, Temma Choji, Shalin Shah, and Alakananda Vempala. 2024. ULTRA: Unleash LLMs' potential for event argument extraction through hierarchical modeling and pairwise self-refinement. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8172–8185, Bangkok, Thailand. Association for Computational Linguistics.

Jin Zhao, Jingxuan Tu, Bingyang Ye, Xinrui Hu, Nianwen Xue, and James Pustejovsky. 2025. Beyond benchmarks: Building a richer cross-document event coreference dataset with decontextualization. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3499–3513, Albuquerque, New Mexico. Association for Computational Linguistics.

Jin Zhao, Nianwen Xue, and Bonan Min. 2023. Cross-document event coreference resolution: Instruct humans or instruct GPT? In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 561–574, Singapore. Association for Computational Linguistics.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2025. Exploring the capability of chatgpt to reproduce human labels for social computing tasks. In *Social Networks Analysis and Mining*, pages 13–22, Cham. Springer Nature Switzerland.

Anastasia Zhukova, Felix Hamborg, and Bela Gipp. 2022. Towards evaluation of cross-document coreference resolution models using datasets with diverse

annotation schemes. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4884–4893, Marseille, France. European Language Resources Association.

A Appendix

A.1 Extended Related Work

Cross-Document Event Tasks: Cross-document Event Set Curation and Variable Coding are more challenging than document-level ones because they require identifying events and selecting documents from a noisy group (Bornstein et al., 2020; Cattan et al., 2020; Ji et al., 2009; Barhom et al., 2019). Prior work uses discourse structure (Chen et al., 2023), Transformer encoders (Gao et al., 2024), negative examples (Cattan et al., 2021a), and retrievers (Eirew et al., 2022). The work by Zhao et al. (2023) most closely resembles a realistic setting; they collected over 1,500,000 COVID-19 documents but evaluated on only 100. In contrast, we use a fivefold larger test set and both the Event Set Curation and Variable Coding tasks.

Text Segmentation and Clustering: Because real-world news articles often digest multiple events, text segmentation is crucial for improving pipeline recall. Historically, algorithms like TextTiling (Hearst, 1997) and TopicTiling (Riedl and Biemann, 2012) manage texts with highly distinct sections. Later neural approaches treated segmentation as a supervised task, using bi-directional LSTMs and BERT-based models trained on dataset like Wikipedia or legal documents (Koshorek et al., 2018; Aumiller et al., 2021). However, documents containing closely related, overlapping events like weekly regional conflict roundups post greater segmentation challenges than articles with distinct topical boundaries. Our use of LLM-based segmentation with K-LLMMEANS address the limitations of inefficient clustering and semantic understanding.

A.2 Automated Data Processing at GTD

A retrieval model collects news articles from trusted sources like LexisNexis and BBC Monitoring. String filters identify documents containing potential attack information, focusing on keywords such as *assault*, *hostage*, and *rebel*. Then, an algorithm removes duplicates and irrelevant documents from this pool. A Support Vector Machine uses TF-IDF to flag highly relevant documents for manual review. In Event Set Curation, our baseline, k-NN search, displays documents within a 1.35 radius using a TF-IDF vectorizer.

A.3 Definition of An Event and Inclusion Criteria

The GTD team defines a terrorism event as the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion, or intimidation. Specifically, a document must have all of the following attributes to be included in the database:

1. **The event must be intentional** - the result of a conscious calculation on the part of a perpetrator.
2. **The event must involve violence** - against either property or people.
3. **The perpetrators of the events must be subnational actors.** State-level is excluded from the database.

In addition, the document must also meet at least two of the following criteria:

1. **The act must be aimed at attaining a political, economic, religious, or social goal.** In terms of economic goals, the exclusive pursuit of profit does not satisfy this criterion. It must involve the pursuit of more profound, systemic economic change.
2. **There must be evidence of an intention to coerce, intimidate, or convey some other message to a larger audience (or audiences) than the immediate victims.** It is the act taken as a totality that is considered, irrespective if every individual involved in carrying out the act was aware of this intention. As long as any of the planners or decision-makers behind the attack intended to coerce, intimidate or publicize, the intentionality criterion is met.
3. **The action must be outside the context of legitimate warfare activities.** That is, the act must be outside the parameters permitted by international humanitarian law, insofar as it targets non-combatants.

A.4 Variable Schema

Here, we describe the variables in the annotation process, specified by GTD.

1. **Country:** the country in which the event occurred.

2. **Location:** the most specific location (e.g., village name) in which the event occurred.

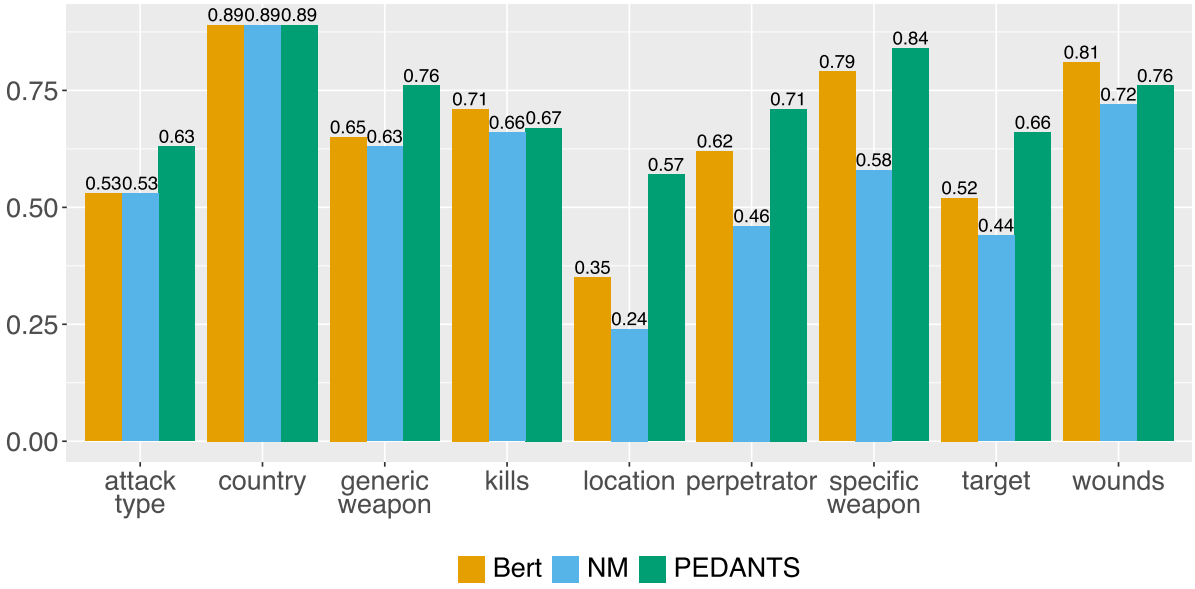

Figure 4: Agreement grouped by variable type. Human annotators agree more with extracted variables with higher degree of specificity. `Country` has over 90% agreement. *Generic attack type* and *weapon type* also show high agreement. In comparison, low specificity variables like *location* demonstrate low agreement with human judgment.

3. **Target:** the targeted group of the event.
4. **Perpetrator:** the group carrying out the event.
5. **Generic Attack Type:** One or more of *Facility/Infrastructure Attack*, *Armed Assault*, *Assassination*, *Bombing/Explosion*, *Hostage Taking (Kidnapping)*, and *NA*.
6. **Generic Weapon:** One or more of *Explosives*, *Firearms*, *Incendiary*, *Sabotage Equipment*, *Melee*, *Vehicle*, and *NA*.
7. **Specific Weapon:** A detailed description of *Generic Weapon*.
8. **Kills:** Number of people killed during the event.
9. **Wounds:** Number of people injured during the event.

A.5 Attack Type Distribution

Attack Type	Percentage
Armed Assault	30.66%
Facility/Infrastructure	13.68%
Bombing/Explosion	31.45%
Hostage Taking (Kidnapping)	7.55%
Unarmed Assault	1.10%
N/A	0.16%

Table 2: Attack Type Distribution

A.6 Generic Weapon Type Distribution

Generic Weapon	Percentage
Explosives	41.04%
Firearms	29.56%
Incendiary	8.96%
Melee	2.52%
Other	1.42%
N/A	30.82%

Table 3: Generic Weapon Type Distribution

A.7 An Illustration of Annotation Difficulty

In Event Set Curation, one document describes an event "*on Sunday, where a man attack a vehicle in Barangay Palampas,*" while another specifies "*the incident occurred in Barangay Palampas, San Carlos City, Negros Occidental, on February 20,*" providing precise details about the event's date.

In Variable Coding, one document describes an event involving "*approximately 20 people, some with axes, who caused injuries*" in "*Houston, British Columbia,*" while another document states "*far-left anti-pipeline extremists*" attacked the "*Morice River drill pad site off the Marten Forest Service Road.*"

A.8 Annotator Background and Training

Only full-time GTD researchers are involved in Event Set Curation, since it is regarded as a more complex and difficult process than Variable Coding. In Variable Coding, student interns and part-time research assistants under the supervision of full-time researchers record detailed information about each attack. Everyone on the GTD team has a background in terrorism/political violence. All annotators on this project hold at least an MA.

One training presentation focuses on the interface—what are the various features of the interface, how to navigate the document window, similar documents, how to identify existing events, etc. The second presentation focuses on the inclusion criteria and an overview of the codebook. New annotators are not expected to know all of the details and nuances for each variable, but they should be able to grasp the top-level variables as well as understand what information to look for in documents that might be relevant to each coding domain (locations, weapons/tactics, casualties, perpetrators, targets, general). These presentations are split to avoid overwhelming trainees with information, but together they probably comprise about 8 hours.

Then, the trainees are put into a sandbox loaded with real documents from previous months. The supervisor provides feedback on those annotations. Once the supervisor is satisfied with the work, the trainee is taken "out of the sandbox" and allowed to triage with the rest of the team. There is not a fixed number of hours that a person stays inside the sandbox as people understand things at different speeds but the process can take weeks. Trainees go through these documents as they would in the real data, identifying events and collecting documents with the relevant information. The supervisor gets files with their progress (events created, documents added to events, as well as documents discarded) and evaluates it, then reviews the feedback with the trainee. There are multiple rounds of work and feedback in the sandbox until the trainee is proficient to begin triaging realistically. The length of this process depends on each new annotator but generally takes a few weeks. New annotators are also spending time doing other tasks (e.g. coding, supervising interns), a time estimate for this stage of the training is 25-40 hours.

In addition, these annotators will often have previous experience with GTD (e.g. coding experience as an intern or hourly), or they will be gaining experience in their role on one of the coding domains, through which they will be getting in-depth knowledge on a section of the codebook, which should reinforce their annotation expertise.

A.9 Algorithm for Finding the Best Embedding Threshold

Algorithm 1 Embedding Algorithm

```
Input: list of documents
Output: best precision, recall, F1
best precision, recall, F1 ← 0, 0, 0
for i = 1 to steps do
    threshold ← min + ((max - min)·i) / steps
    for each document1 in all documents do
        similars ← ∅
        for each document2 in all documents do
            cal_sim(document1, document2)
            if similarity ≥ threshold then
                similars.add(document2)
            end if
        end for
        if F1(ref, similars) > best F1 then
            update best F1
        end if
    end for
end for
return best (precision, recall, F1)
```

A.10 Over- and Under-generation of LLM-CLS

Using an LLM to generate event sets candidates would almost always create different numbers ofevent sets compared to TF-IDF, making direct comparison difficult. To address potential over- or under-generation, we formulate this as a linear assignment problem, optimizing the average F_1 score between the results and the human-coded reference set.

The linear assignment problem is an optimization problem. The objective is to assign a gold event set to a generated event set in such a way that the overall cost is minimized. $\mathcal{C}_{i,j}$ represents the cost of matching event set i in the gold set with event set j in the prediction set. Formally, the optimal assignment has cost

$$\sum_i \sum_j \mathcal{C}_{i,j},$$

where

$$\mathcal{C}_{i,j} = -F_1(i,j)$$

A.11 Variable Coding Error Examples

Interpretive Subjectivity: One document mentions the administrative area "*Pale, Sagaing*" and the more specific "*Einmahti village*" in Myanmar, the model selects the broader region, whereas experts prefer the latter.

Temporal Conflict: One initial breaking report states "*five people injured,*" while an updated report later revises the count to "*at least eight people wounded.*" Under-trained annotators and LLMs alike struggle to resolve these timelines to find the accurate information.

Interpretive Subjectivity: An attack on non-human entities is `facility/infrastructure` if buildings are the primary target; otherwise, they use `armed assault`. This dividing line is often subjective. Assassinations via explosives are classified as `assassination` in the GTD codebook, not `bombing/explosion`. Lacking this knowledge, LLM codes those variables more frequently as `bombing` and `armed assault`.

When we include variable definitions in Appendix A.4 into the prompt, semantic accuracy (measured by BEM and PEDANT) increases for `specific weapon`, `location`, `target`, `perpetrator`, `wounds`, and `generic attack` with statistical significance ($p < 0.05$).

A.12 Human-LLM Agreement

We ask three groups of annotators to annotate a shared portion of the event sets and check the inter-annotator agreement on variables (human-human). We also calculate the agreement between annotators operating in the hybrid and the manual setting (human-LLM). Figure 5 displays the results: human-human and human-LLM agreements do not differ significantly. Figure 4 shows human-LLM agreement by variable types. Annotators show 0.89 agreement with *Country*, a variable with a high degree of specificity. In contrast, annotators agree with *Location* infrequently, suggesting less utility of variables with a lower degree of specificity. We also investigate if human-LLM agreement differ by providers and models (Table 4).

A.13 Model/Prompt Specifications and Justification for Model Selection

We tested the following LLMs: QWEN3-NEXT-80B-A3B-INSTRUCT, GEMINI 2.5 PRO, GPT-4.1 MINI, MIXTRAL 8X7B INSTRUCT, CLAUDE SONNET 4, LLAMA 3.3 70B INSTRUCT, MISTRAL SMALL 3.2 24B, GEMINI 2.5 FLASH, and GPT-4O-MINI. All models are zero-shot with 0 temperature. However, output might change for future models.

Upon preliminary inspection, no single model has significantly higher accuracy in Event Set Curation and Variable Coding. The specific model we used is GPT-4O-MINI-2024-07-18 because it is cost-efficient. It does not require fine-tuning or computational resources. Social science researchers can use well-developed API calls with minimum cost compared to larger and more advanced models. We show the results of all models in Tables 5 and 4.

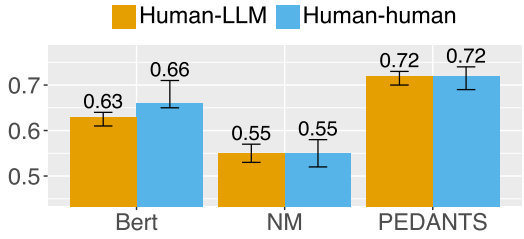

Figure 5: In Variable Coding, the percentage of agreement difference between human-human and human-LLM is not statistically significant, suggesting that LLM-coded variables provide human-level utility. On average, annotators and GPT-4o-mini agree 55% using NM. PEDANTS and BEM show higher agreements.

Model	BEM	EM	PEDANT
CLAUDE-SONNET-4	0.47	0.40	0.50
GEMINI-2.5-FLASH	0.45	0.42	0.48
GEMINI-2.5-PRO	0.45	0.39	0.50
GPT-4.1-MINI	0.48	0.42	0.52
LLAMA-3.3-70B-INSTRUCT	0.44	0.39	0.51
MISTRAL-SMALL-3.2-24B-INSTRUCT	0.46	0.40	0.46
MIXTRAL-8X7B-INSTRUCT	0.39	0.28	0.46
QWEN3-NEXT-80B-A3B-INSTRUCT	0.48	0.43	0.51

Table 4: Human-LLM agreement in CDEAE between expert annotators and different models. No model is significantly better at CDEAE.

A.13.1 Prompt for Event Set Curation

Determine whether the following articles describe the same incident:
`{document 1}`
`{document 2}`

A.13.2 Prompt for SEG

The following document describes zero or more incidents. Segment the document based on incidents mentioned and return an array.
`{document}`

A.13.3 Prompt for K-LLMMEANS

The following is a set of documents describing a single terrorism event. Write a concise summary that represents the cluster.

A.13.4 System Prompt for Variable Coding

You are a trained annotator. Extract the relevant information based on the given question. Respond with 'NA' if the information is not present in the text. ONLY provide the answer, without any additional explanation.

A.13.5 Prompt for Variable Coding

What is EVENT VARIABLE *in the event?*

A.13.6 Prompt for Error Mitigation

What is EVENT VARIABLE*? DEFINITION:* EVENT VARIABLE DEFINITION

A.14 Model Accuracy in Identifying NA Variables

See Table 5 and 4.

A.15 Generative AI Use

We use generative AI for writing and coding. For writing, we only use AI for checking grammatical and syntax mistakes. For example, making sure we use LaTeX macros instead of string literals and correcting typographical errors and subject-verb agreements. We do not use generative AI in creating research ideas and writing the draft. For coding, we use generative AI for assisting in creating figures and retry decorators for API calls.

A.16 K-LLMMEANS Configurations

We maintain similar settings by using TEXT-EMBEDDING-3-SMALL from OpenAI and DistilBERT (Sanh et al., 2020). However, because document length exceeds 512, DistilBERT's maximum input sequence length, we also add ModernBERT (Warner et al., 2025). We use the same parameters as K-LLMMEANS-5: 120 iterations with five summarization steps ($l = 20$) and all documents as inputs at each step. Overall, the cost of generating $2,355$ TEXT-EMBEDDING-3-SMALL embeddings ($500 + 5 * 371$) and $5,565$ ($3 * 5 * 371$) summarizations for all three embedding types is $13.67, much less than directly comparing documents pairwise.

A.17 Example of K-LLMMEANS Centroid Summaries

Iteration 20

Over the course of 40 hours, the Baloch Liberation Army (BLA), an armed group advocating for Balochistan's independence, conducted a significant attack on the Pakistani Frontier Corps (FC) headquarters in Panjgur, Balochistan. The BLA successfully occupied the headquarters, outmaneuvering the Pakistani military despite its superior resources, resulting in substantial casualties among Pakistani troops. This major offensive is the BLA's most daring against what they consider the oppressive and exploitative Pakistani forces. The BLA regards the Pakistani government and its Chinese partners as colonial powers exploiting Balochistan's resources, leaving the local population disenfranchised and impoverished. The attack occurs amid heightened tensions over the China-Pakistan Economic Corridor (CPEC) projects in Balochistan, viewed by locals as marginalizing and excluding them from economic benefits. Recent BLA attacks have targeted Chinese interests, signaling the group's opposition to Chinese involvement in the region. Previously, notable attacks attributed to the BLA include targeted assaults on Chinese nationals and projects in Balochistan, Karachi, and Gwadar, underscoring the group's resolve against external exploitation. These operations have prompted demands from China for enhanced security measures for its nationals in Pakistan.

Iteration 40

Over the last 40 hours, the Baloch Liberation Army (BLA), a militant group advocating for Balochistan's independence, executed a significant attack on the Pakistani Frontier Corps (FC) headquarters in Panjgur, Balochistan. This attack, which occurred on Wednesday night and continued until Friday evening, resulted in heavy casualties for the Pakistani forces, despite their sophisticated military equipment. This marks the most daring operation by Baloch organizations against what they perceive as Pakistani occupying forces, who allegedly suppress the Baloch community and exploit their region's resources in collaboration with China, particularly through the China-Pakistan Economic Corridor (CPEC) projects. The BLA's Majeed Brigade, known for its committed fighters, has a history of audacious attacks targeting Chinese interests and Pakistani assets, illustrating the growing unrest and resistance within Balochistan against perceived foreign-backed exploitation. Recent years have witnessed several high-profile BLA attacks, including assaults on Chinese engineers and installations, culminating in deteriorating security conditions in the region for both Pakistani and Chinese stakeholders.

Iteration 60

In a significant escalation, the Baloch Liberation Army (BLA) attacked and occupied the Pakistani Frontier Corps headquarters in Panjgur, Balochistan, for nearly 40 hours, marking one of the largest and boldest assaults by the group. This attack, which resulted in heavy casualties

Model	Precision	Recall	F1-Score	TP	FP	FN	TN
CLAUDE-SONNET-4	0.51	0.27	0.35	193	189	533	1947
GEMINI-2.5-FLASH	0.43	**0.39**	**0.41**	284	377	442	1759
GEMINI-2.5-PRO	0.44	0.24	0.31	172	217	554	1919
GPT-4.1-MINI	0.53	0.20	0.30	148	129	578	2007
LLAMA-3.3-70B-INSTRUCT	0.51	0.28	0.36	204	198	522	1938
MISTRAL-SMALL-3.2-24B-INSTRUCT	0.48	0.36	**0.41**	262	289	464	1847
MIXTRAL-8X7B-INSTRUCT	**0.62**	0.03	0.06	24	15	702	2121
QWEN3-NEXT-80B-A3B-INSTRUCT	0.52	0.25	0.34	180	167	546	1969

Table 5: Model accuracies in identifying variables that are not available in documents. MIXTRAL-8X7B-INSTRUCT has the highest precision because it almost always code variables even when they are not available. Other models do not differ significantly from the general population.

among Pakistani troops, underscores the deep-seated grievances of Baloch organizations against the Pakistani government, which they view as a colonial occupier exploiting Balochistan's resources in partnership with China. The BLA's Majeed Brigade, known for its audacious attacks, has historically targeted interests shared by Pakistan and China, driven by marginalization and exploitation concerns, particularly related to the China Pakistan Economic Corridor (CPEC). This incident is part of a series of attacks that include high-profile assaults on Chinese nationals and infrastructure, illustrating a broader struggle for autonomy and recognition by Baloch groups who feel alienated in their region.

Iteration 80

In a significant escalation of conflict, the Baloch Liberation Army (BLA) launched a bold attack on the Pakistani Frontier Corps headquarters in Panjgur, Balochistan, on February 10, inflicting heavy casualties and holding their position for nearly 40 hours. This attack represents the largest and most audacious by Baloch groups against Pakistani forces, whom they view as oppressive occupiers. The BLA, especially its elite Majeed Brigade, has a history of targeting both Pakistani and Chinese interests, as seen in attacks on projects related to the China Pakistan Economic Corridor (CPEC) and Chinese nationals. These actions stem from grievances over resource exploitation and marginalization of the Baloch people by both Pakistan and China. The recent surge in attacks follows a larger movement led by Baloch leaders like Jamaat-i-Islami's popular figure, amidst the Baloch community's growing hostility towards Sino-Pakistani cooperation.

Iteration 100

Over a 40-hour period beginning on February 10, the Baloch Liberation Army (BLA) launched a bold attack on the Frontier Corps headquarters in Panjgur, Balochistan, marking one of their largest assaults against Pakistani forces. The BLA, representing Baloch groups who view the Pakistani government as a colonial and exploitative force, aims to dismantle perceived occupation and exploitation particularly linked to the China-Pakistan Economic Corridor (CPEC) projects which marginalize local Baloch people. Using their elite Majeed Brigade comprised of committed fidayeen, the BLA has a history of targeting both Pakistani and Chinese interests to express their grievances over resource deprivation and exploitative partnerships. Such violent campaigns have not only caused significant casualties but also unsettled China, which has responded by demanding increased security for its nationals in Pakistan. The attack fits into a pattern of recent escalations, reflecting the Baloch nationalists' deep-seated discontent.

Iteration 120

In a recent escalation of violence, the Baloch Liberation Army (BLA) has launched a significant attack on the Pakistani Frontier Corps headquarters in Panjgur, Balochistan, maintaining control for 40 hours and inflicting heavy casualties on Pakistani forces. This marks one of the Baloch groups' most audacious attacks against what they perceive as an oppressive Pakistani regime, which alongside China, exploits Balochistan's resources while marginalizing its people. The surge in attacks coincides with growing discontent over the China Pakistan Economic Corridor (CPEC) that bypasses local employment. The BLA's elite Majeed Brigade, known for its previous high-profile assaults on Chinese interests, underscores the group's targeting of both Pakistani and Chinese influences in the region. Notable past attacks include assaults on Chinese workers and installations, showcasing the deep links Baloch nationalists have with regional geopolitics, further destabilizing Pakistan-China relations.

A.18 Screenshots of Annotation Interface

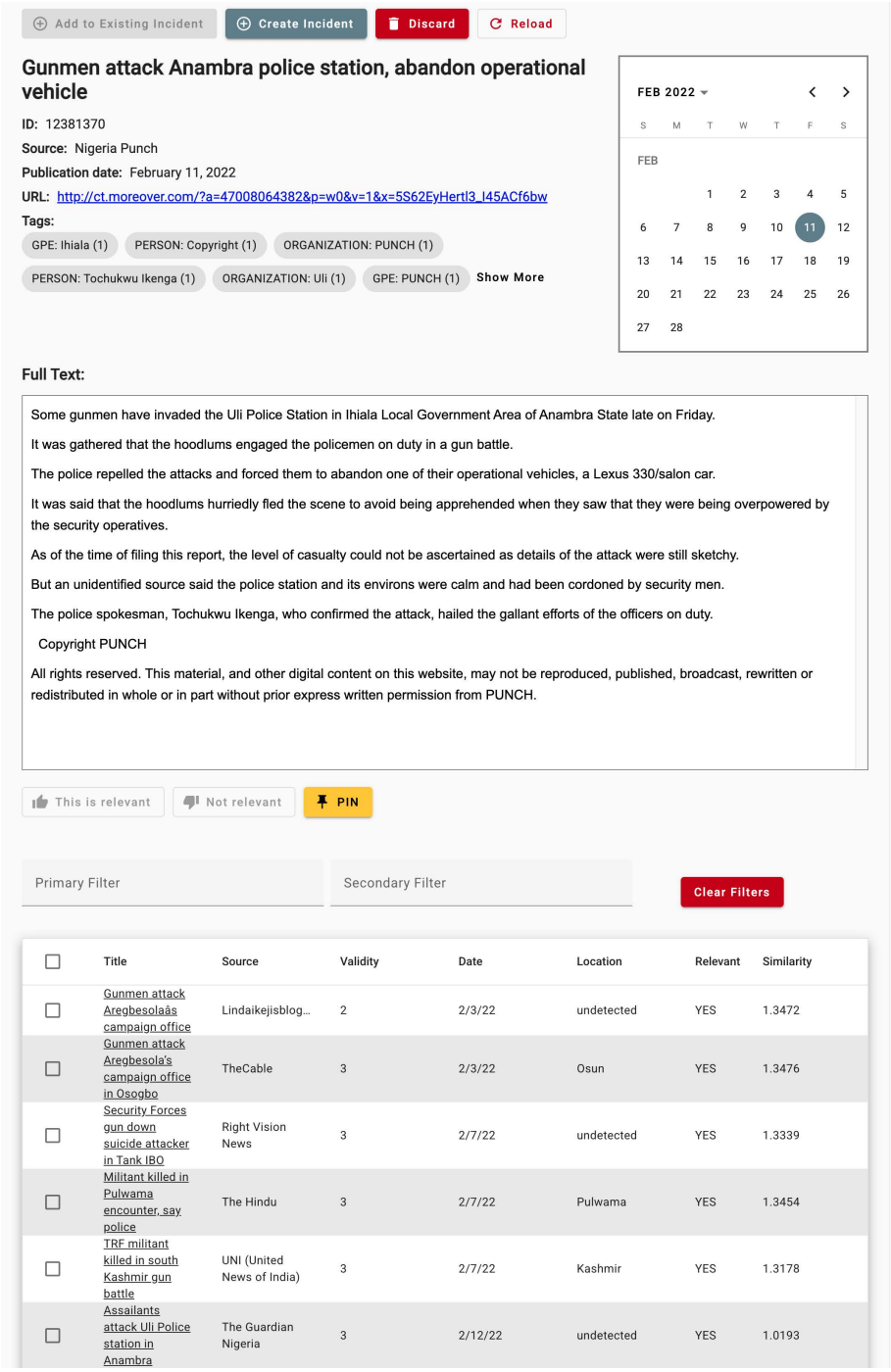

Figure 6: An example of CDCR interface. Annotators see a document along with similar documents retrieved by TF-IDF

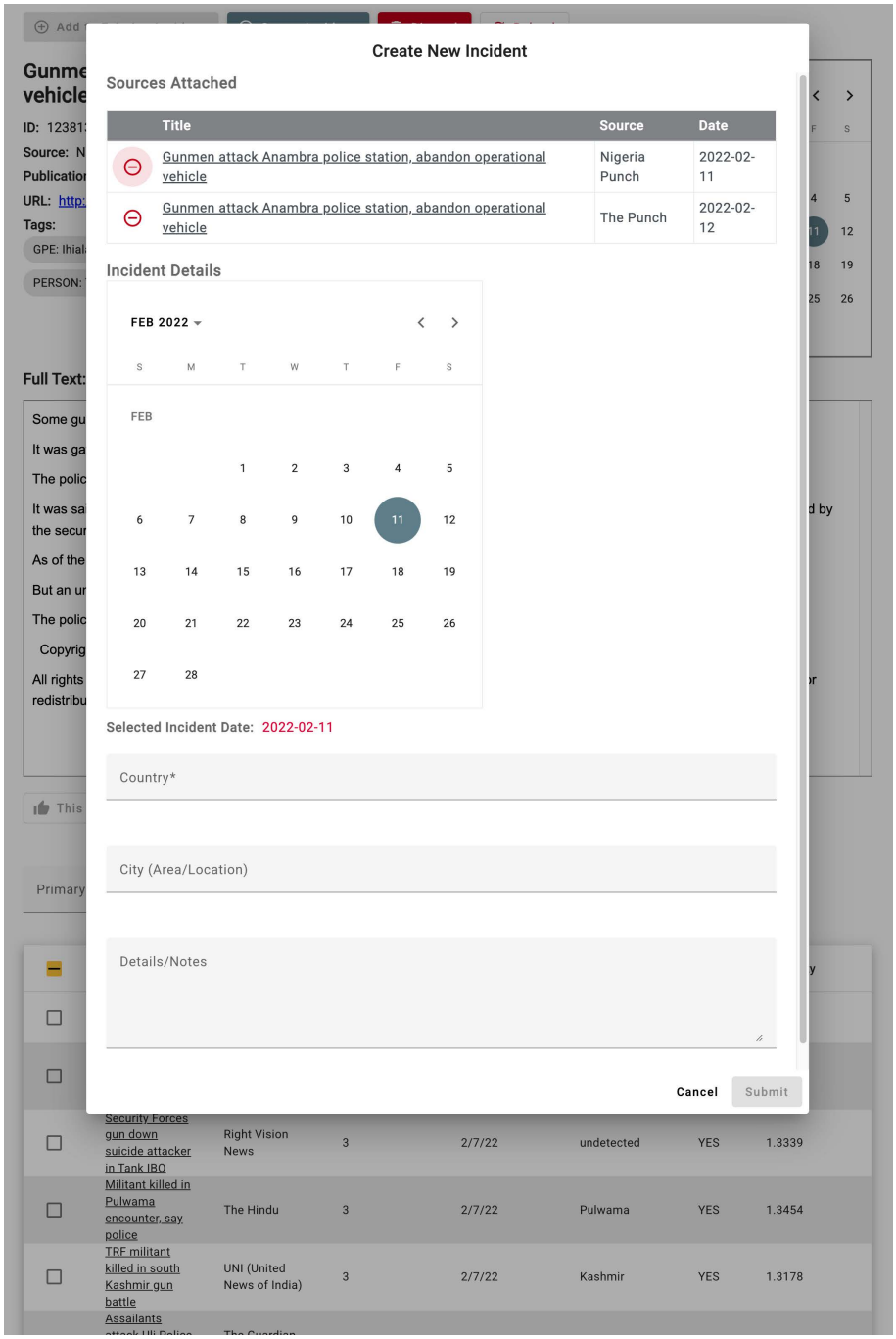

Figure 7: In CDCR, once an annotator finds all relevant documents, they create the events and fill in the preliminary variables for the second phase–extraction.

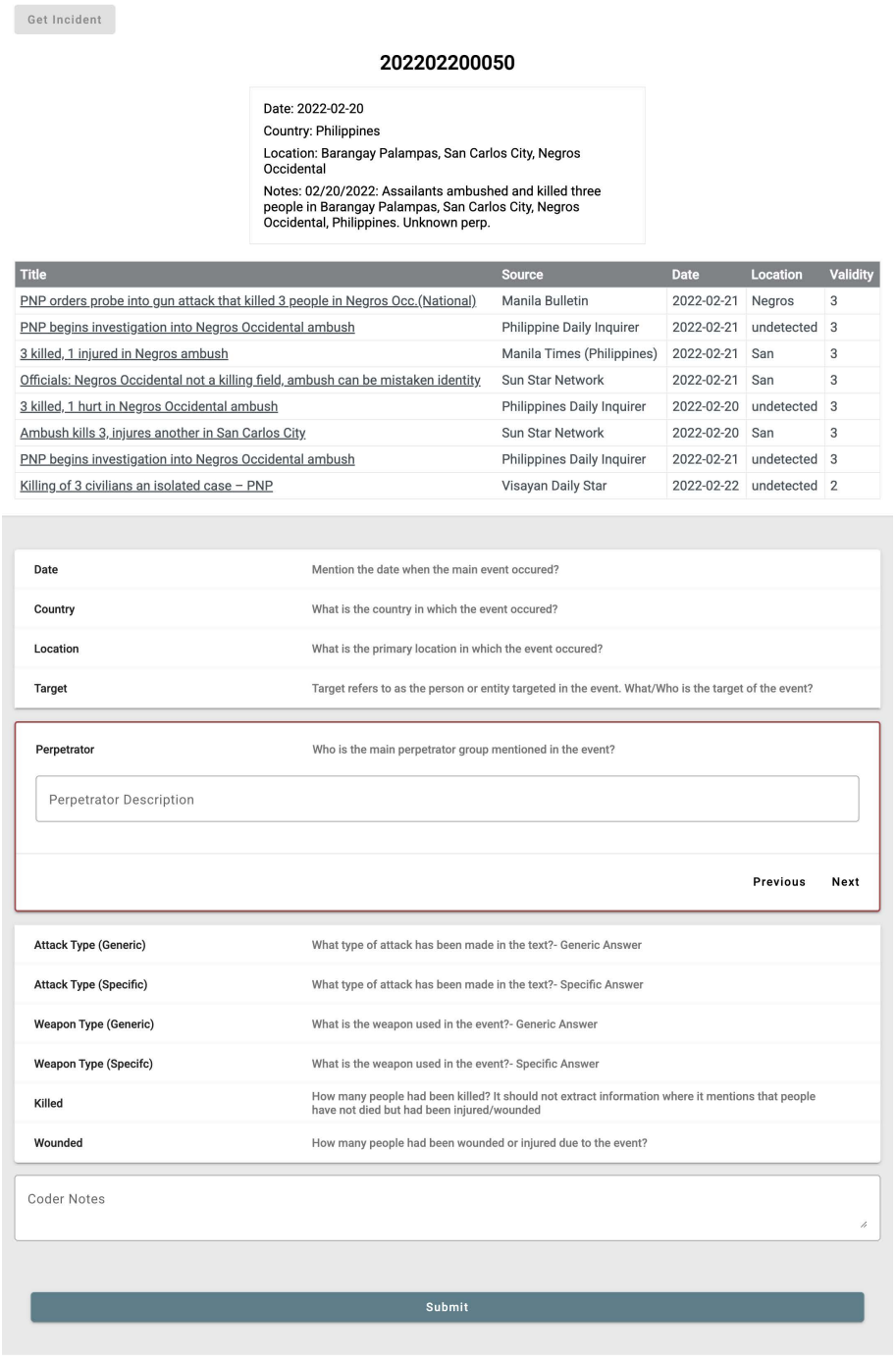

Figure 8: Existing pipeline where annotators code the variables without LLM assistance.

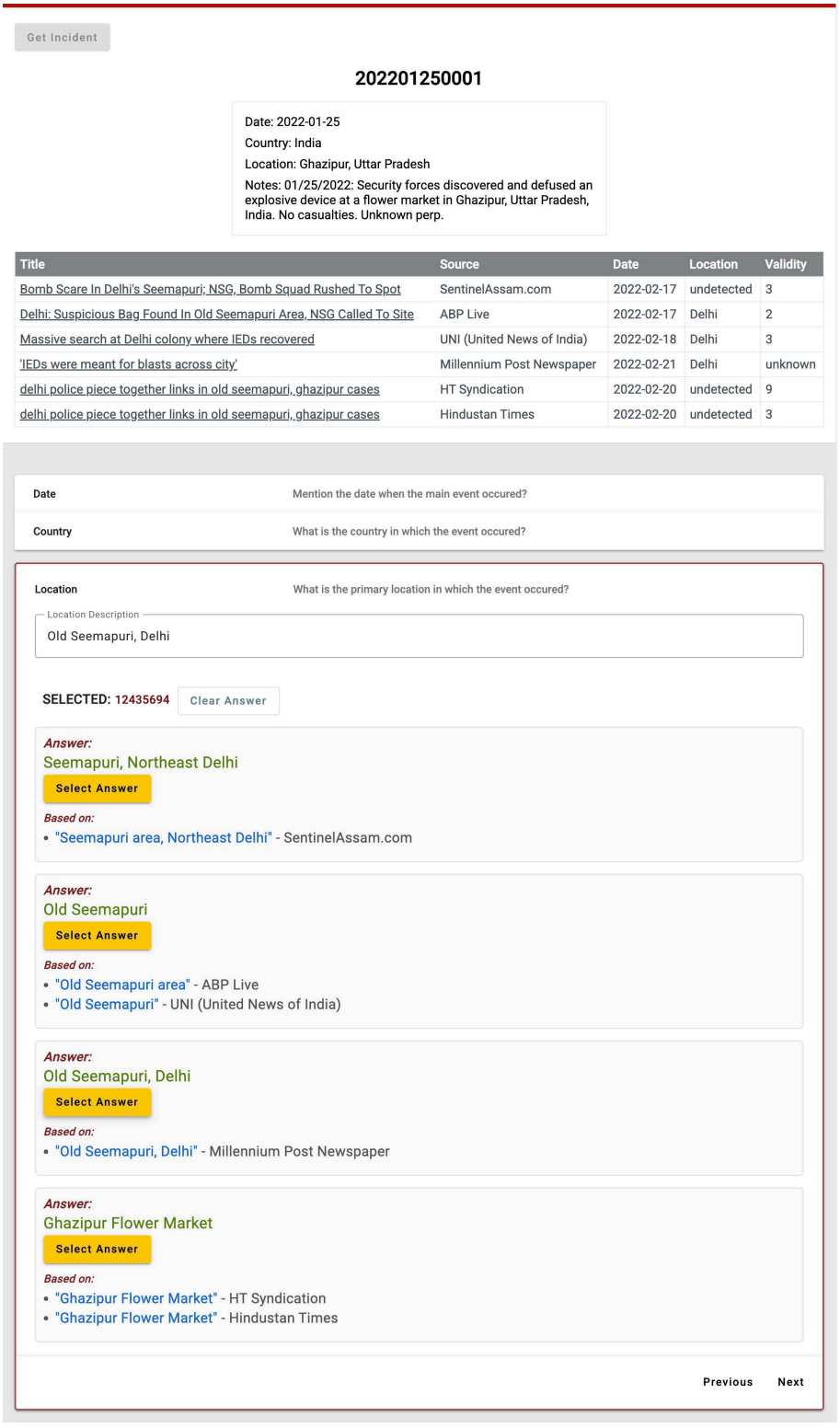

Figure 9: Interface where annotators see the LLM-extracted variables.