Causal Evidence Extraction and Triangulation in Crisis Reports using Large Language Models: A ReliefWeb-based Study

Yuanjun Zhang
University of Oulu, Finland
LUT University, Finland
yuanjun.zhang@lut.fi

Mourad Oussalah*
University of Oulu, Finland
mourad.oussalah@oulu.fi

Abstract

Humanitarian reports are long, noisy, and multi-topic, making it difficult to consolidate decision-relevant causal evidence. We present a ReliefWeb study (2000–2024) and a two-stage Large Language Model (LLM) pipeline that extracts structured intervention-outcome records with direction and strength attributes. Query-conditioned extraction restricts output to a specified intervention class, reducing retrieval-induced over-extraction, while snippet grounding links each relation to supporting text for auditability and classification. In an expert-annotated dataset of 100 reports, the best closed-source LLM achieved a weighted F1 score of 90.73% with strong cost-efficiency, while Llama-3.1-8B with supervised fine-tuning reached 94.15% weighted F1 score. We further propose context-preserving triangulation that aggregates strength-weighted evidence within disaster×source cells, applies Laplace smoothing and equally weights cells to quantify cross-context convergence via a Level-of-Evidence score. Applied to cash assistance, food-related outcomes show strong positive convergence (LoE=0.865) and stable long-horizon trajectories.

1 Introduction

Humanitarian crises generate large volumes of narrative situation reports describing interventions and evolving outcomes, yet decision-relevant evidence about *what works, where, and under which constraints* remains hard to consolidate at scale. Cash assistance illustrates the challenge: it is widely deployed, but reported impacts vary with market conditions, targeting, delivery mechanisms, and protection risks (Cash Learning Partnership (CaLP), 2015; Doocy and Tappis, 2017; Freccero et al., 2019; Burton, 2020). A bottleneck is translating long, multi-topic narratives into structured, auditable intervention, and outcome evidence that can be compared across contexts.

Because keyword retrieval prioritizes recall, it often surfaces documents where the queried intervention is marginally mentioned. Besides, unconstrained extraction can then amplify this noise by producing irrelevant relations. Recent LLMs make prompt-based information extraction practical at scale (Brown et al., 2020), and grounding conclusions in explicit textual evidence is increasingly emphasized for auditability (DeYoung et al., 2020; Lewis et al., 2020). However, evidence-synthesis extraction frameworks are typically developed for more topically clean input (Shi et al., 2025), motivating report-centric and question-aware designs.

We address this gap with a two-stage LLM extraction pipeline for *causal evidence extraction* from ReliefWeb reports (ReliefWeb, 2025) in Fig. 1. In **Stage 1**, *query-conditioned extraction* restricts outputs to interventions in a query-defined class and extracts intervention–outcome candidates together with supporting snippets, reducing over-extraction from low-precision retrieval. In **Stage 2**, *snippet-grounded relation classification* predicts direction and strength for each candidate pair using the supporting snippets, improving auditability and inference. For synthesis, we then introduce **context-preserving triangulation**, which aggregates strength-weighted directional evidence within disaster×source cells and measures cross-context convergence rather than pooling raw counts (Jick, 1979). In addition, for lightweight deployment, we study an **optional distillation step**: using the best closed-source pipeline outputs as supervision, we apply Low-Rank Adaptation (LoRA) (Hu et al., 2022) supervised fine-tuning to Llama-3.1-8B-Instruct (Grattafiori et al., 2024).

Using ReliefWeb corpus (2000–2024), we compare five strategies across three API LLMs and an open-weight Llama-3.1-8B-Instruct model with LoRA fine-tuning. Across settings, query condi-

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 32478–32491

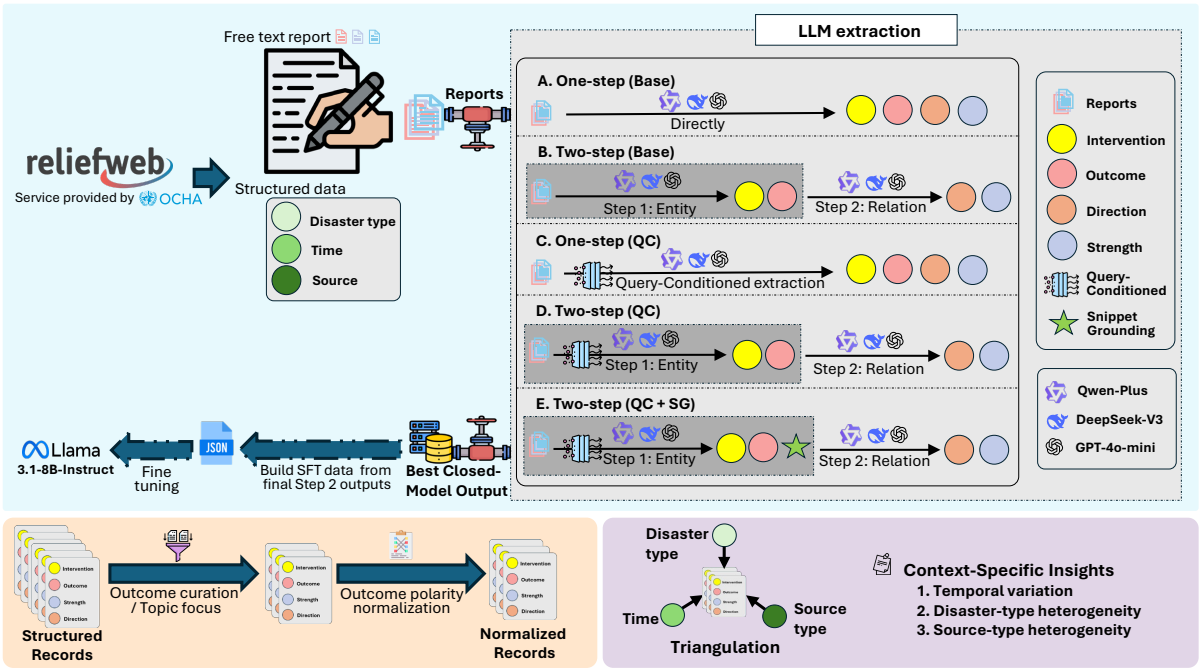

Figure 1: Overall framework for causal evidence extraction and context-preserving triangulation. ReliefWeb reports and metadata (disaster type, source type, time) are processed by a two-stage LLM pipeline with query conditioning and snippet grounding to produce structured causal records (Section 3.1). We then curate outcomes and normalize polarity for consistent directionality, and aggregate evidence via cell-based triangulation over disaster×source contexts using strength-weighted counts, Laplace smoothing, and equal cell weighting, yielding global probabilities (P_P, P_Z, P_N), a Level-of-Evidence (LoE) score, and temporal trajectories (Section 3.2).

tioning yields large, consistent gains, and snippet grounding further improves faithfulness. We also demonstrate long-horizon triangulation for cash assistance, producing interpretable global probabilities and temporal evidence trajectories.

2 Related Work

Prior humanitarian natural language processing (NLP) research highlights domain shift and the need for trustworthy decision support (Rocca et al., 2023). We study *report-centric* extraction of intervention–outcome records from long, multi-topic ReliefWeb reports (ReliefWeb, 2025), leveraging document-level IE/RE (Yao et al., 2019) and evidence grounding (DeYoung et al., 2020).

Complementary to text-based evidence extraction, recent work has also emphasized reliability and explainability in disaster-related multimodal decision support, showing that preference optimization and explainable vision-language reasoning can improve both assessment performance and interpretability in disaster severity assessment (Zhang et al., 2026c). LLMs are increasingly used for evidence extraction and synthesis, including directional triangulation (Shi et al., 2025) and end-to-end evidence-synthesis pipelines (Wang et al., 2025b). We adapt this line to humanitarian reports with query-conditioned extraction and context-preserving synthesis within disaster×source cells, while attaching short supporting spans for traceability. LLM factuality and hallucination detection have also been studied, e.g., via consistency-based checks (Manakul et al., 2023).

More specifically, our work differs from prior evidence extraction studies in three ways. First, we target long, noisy, multi-topic humanitarian reports. Second, we use query-conditioned extraction to control retrieval-induced over-extraction. Third, beyond extraction, we introduce context-preserving triangulation over disaster×source cells, together with snippet grounding for auditability, instead of relying only on pooled document-level evidence.

3 Methodology

Given a ReliefWeb report x and a query q specifying an intervention *class* (e.g., *cash assistance*), we extract a set of directional causal records

$$e = (a, o, \delta, \sigma, z) \tag{1}$$

where a is an intervention mention (restricted to class q), o is an outcome mention, $\delta \in$

$\{\text{INC}, \text{DEC}, \text{NO}\}$ denotes whether the reported outcome increases, decreases, or shows no clear change, $\sigma \in \{\text{WEAK}, \text{MOD}, \text{STRONG}\}$ is the evidence strength, and z is a supporting text snippet for auditability.

Data. We query ReliefWeb with the keyword *'cash assistance'* and retain English reports with accessible text (2000–2024). We restricted to ReliefWeb-defined disaster and source types: four disasters (FLOOD, DROUGHT, EPIDEMIC, EARTHQUAKE) and five sources (ACADEMIC, GOVERNMENT, MEDIA, NGO, RED CROSS/RED CRESCENT MOVEMENT), yielding 8,029 reports; INTERNATIONAL ORGANIZATION and OTHER are excluded due to imbalance/ambiguity (Appendix A and Appendix B). For evaluation, experts annotate 100 sampled reports (220 relations); see Appendix C for details.

Evaluation. We evaluate intervention/outcome extraction via semantic set matching using BERTScore (threshold 0.8) (Zhang et al., 2020). Direction and strength are scored on matched intervention–outcome pairs; strength is scored only when direction is correct. Since strength is ordinal (WEAK<MOD<STRONG), adjacent confusions receive partial credit (0.5).

3.1 Query-conditioned and snippet-grounded extraction

As shown in Fig. 1, our final method is a **two-stage extraction pipeline** with two LLM calls. **Stage 1 (QC entity & evidence extraction):** the LLM extracts only interventions in class q (query-conditioned, QC), the associated *outcome mentions* expressed in the report as raw text spans / free-text phrases, and supporting snippets z. The output of Stage 1 therefore consists of candidate intervention–outcome pairs together with their evidence snippets. **Stage 2 (SG relation classification):** given these candidate pairs and snippets, the LLM predicts the relation labels, i.e., direction δ and strength σ. Snippets provide explicit grounding (snippet-grounded, SG) while the full report remains available.

Separately, we compare **single-prompt** versus **two-prompt** prompting as an ablation, and we also ablate QC and SG. Here, *stage* refers to the two LLM calls in the final extraction pipeline, whereas *single-prompt* and *two-prompt* refer only to prompting variants.

Using the best **two-stage QC+SG extraction pipeline** as the teacher, we select the best-performing proprietary closed-source LLM. We then decompose its outputs into **Stage-1 supervision** (QC intervention–outcome pairs with grounded snippets) and **Stage-2 supervision** (direction/strength conditioned on Stage 1), and train two parameter-efficient LoRA adapters (one per stage) on an open-weight Llama-3.1-8B-Instruct student while preserving the same two-stage inference structure. To prevent leakage, we exclude the 100 evaluation reports from distillation; to reduce model-family relatedness bias, we choose a student outside the teacher's family (Wang et al., 2025a).

Model	Domain	GT	Base	QC
qwen plus	Medical	3.09	5.00	–
	Humanitarian	2.20	13.76	2.79
gpt-4o mini	Medical	3.09	4.33	–
	Humanitarian	2.20	7.43	2.19
deepseek v3	Medical	3.09	3.70	–
	Humanitarian	2.20	10.50	1.60

Table 1: Extraction volume comparison across domains and models. Values are average #relations/document. QC is applied only for humanitarian reports.

3.2 Context-preserving triangulation

We synthesize records across contexts while preventing high-volume contexts from dominating.

Polarity normalization. Outcomes may be positively framed (e.g., *food security*) or negatively framed (e.g., *food insecurity*). We assign polarity $\pi(o) \in \{+1, -1\}$ and flip INC $\leftrightarrow$ DEC when $\pi(o) = -1$ (leaving NO unchanged), so a *positive* relation always denotes improvement. (For the case study, polarity is labeled for frequent outcomes, e.g., frequency ≥ 3. See Appendix D for details.)

Cell-based evidence aggregation. Let $\mathcal{D}$ denote the set of disaster types and $\mathcal{S}$ denote the set of source types. Each record inherits disaster type $d \in \mathcal{D}$ and source type $s \in \mathcal{S}$. We define triangulation cells $c = (d, s) \in \mathcal{D} \times \mathcal{S}$ and map normalized directions to $r \in \{P, Z, N\}$ (positive / no-change / negative). We convert strength to an ordinal weight $w(\text{WEAK})=0,\ w(\text{MOD})=1,\ w(\text{STRONG})=2$:

$$C_r^{(c)} = \sum_{i \in c} w(\sigma_i)\, \mathbb{I}(r_i = r). \tag{2}$$

Smoothed cell probabilities. To avoid degenerate estimates in sparse cells, we apply Laplace

Model	Intervention	Outcome	Direction	Strength	Weighted F1	Unit price (USD)	Efficiency
DeepSeek-V3							
One-step (Base)	44.20	58.80	93.77	89.63	67.58	0.707	0.956
Two-step (Base)	34.69	57.09	94.46	89.51	64.33	0.707	0.910
One-step (QC)	75.00	74.65	95.85	91.38	82.34	0.707	1.164
Two-step (QC)	76.47	82.12	97.19	92.20	85.46	0.707	1.208
Two-step (QC + SG)	87.44	87.86	92.88	90.02	89.17	0.707	1.261
GPT-4o-mini							
One-step (Base)	48.24	62.52	92.14	82.35	68.13	0.375	1.817
Two-step (Base)	49.05	67.43	91.06	75.10	68.18	0.375	1.818
One-step (QC)	73.28	77.33	94.08	84.91	80.98	0.375	2.159
Two-step (QC)	84.49	86.34	94.29	81.08	86.32	0.375	2.302
Two-step (QC + SG)	87.50	88.83	94.17	77.55	87.24	0.375	2.326
Qwen-Plus							
One-step (Base)	34.96	49.86	92.11	89.97	61.86	0.198	3.124
Two-step (Base)	33.77	52.95	88.09	89.01	61.44	0.198	3.103
One-step (QC)	85.59	84.88	95.48	91.58	88.55	0.198	4.472
Two-step (QC)	87.88	88.63	93.71	90.44	89.78	0.198	4.534
Two-step (QC + SG)	87.31	91.99	93.65	91.05	90.73	0.198	4.582
Llama-3.1-8B-Instruct using Two-step (QC + SG)							
No SFT	62.72	71.69	89.04	84.96	75.12	–	–
LoRA SFT	94.06	95.68	94.64	91.50	94.15	–	–
Weights	0.30	0.30	0.20	0.20			

Table 2: Overall performance comparison across models and extraction strategies. All F1 scores are in %. Weighted F1 uses weights 0.30/0.30/0.20/0.20. Base denotes the baseline extraction strategy from Shi et al. (2025).

smoothing (α=0.1):

$$P_r^{(c)} = \frac{C_r^{(c)} + \alpha}{\sum_{r' \in \{P,Z,N\}} C_{r'}^{(c)} + 3\alpha} \tag{3}$$

Cells with $\sum_r C_r^{(c)} = 0$ are discarded.

Equal cell weighting and Level of Evidence. Let $\mathcal{C}$ denote the set of retained non-empty cells. We average cell probabilities equally:

$$P_r = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} P_r^{(c)} \tag{4}$$

Cross-context convergence is measured by a Level of Evidence score:

$$\text{LoE} = \frac{\max(P_P, P_Z, P_N) - \frac{1}{3}}{1 - \frac{1}{3}} \in [0, 1]. \tag{5}$$

Here, P_P, P_Z, and P_N denote the global cell-averaged probabilities of positive / no-change / negative relationships. By construction, P_P, P_Z, and P_N are non-negative and sum to 1, so $\max(P_P, P_Z, P_N) \in [1/3, 1]$. Therefore, LoE $\in [0, 1]$ and cannot be negative.

Temporal triangulation. For evidence trajectories, we recompute Eqs. (2)–(5) cumulatively by year t using records with $\text{year}_i \leq t$, producing $(P_P(t), P_Z(t), P_N(t), \text{LoE}(t))$.

4 Experiments and Results

We evaluate on four LLMs: *Qwen-Plus* (Bai et al., 2023), *GPT-4o-mini* (Hurst et al., 2024), *DeepSeek-V3* (DeepSeek-AI et al., 2025), and Llama-3.1-8B-Instruct.

4.1 QC reduces over-extraction

Keyword retrieval is recall-oriented and often surfaces reports where the queried intervention is only marginally mentioned; baseline prompting then over-extracts query-irrelevant relations. Tab. 1 shows severe over-extraction on humanitarian reports, while query-conditioned extraction (QC) brings the volume close to the expected ground-truth level across closed-source models. Medical numbers are from Shi et al. (2025).

4.2 Extraction performance: ablations over QC, SG, and prompting variants

Tab. 2 reports extraction performance on the expert-annotated benchmark. Across models, QC provides the largest gains, while prompt decomposition and snippet grounding (SG) yield additional consistent improvements. Among closed-source models, Qwen-Plus with two-step QC+SG achieves the best weighted F1 (90.73%) and is used for downstream triangulation. Efficiency is computed as Efficiency = (Weighted F1/100) / unit price

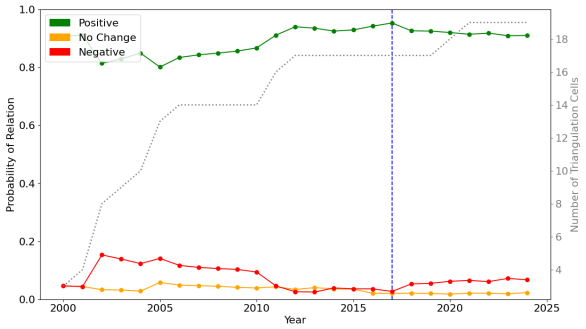

Figure 2: Cumulative temporal triangulation for polarity-normalized food-related outcomes: cell-averaged P_P, P_Z, P_N and the number of contributing disaster×source cells over time.

(prices in Appendix Tab. 5). Our pipeline also provides supervision for distillation: LoRA fine-tuning Llama-3.1-8B-Instruct improves weighted F1 from 75.12% to 94.15% (Appendix F).

4.3 Triangulation: cross-context convergence

Using Qwen-Plus (two-step QC+SG), we extract 19,568 causal records across four disaster types (5,999 drought; 3,696 earthquake; 4,412 epidemic; 5,461 flood). We focus on food-related outcomes due to high evidence density (Appendix B). After polarity normalization (87 positive vs. 65 negative outcome strings with frequency ≥ 3), triangulation over $4 \times 5 = 20$ disaster×source cells retains 19 non-empty cells and yields $P_P = 0.91$ and $\text{LoE} = 0.865$.

Fig. 2 shows cumulative trajectories. Early years (2000–2005) have sparse coverage and volatile estimates; from 2006–2017, expanding cell coverage yields strong positive convergence, peaking in 2017 ($P_P = 0.953$, $\text{LoE} = 0.929$). From 2018–2024, the signal remains highly positive with mild softening, consistent with residual context heterogeneity and operational "do no harm" constraints (Cash Learning Partnership (CaLP), 2015; Freccero et al., 2019; Burton, 2020).

5 Conclusion

We propose a query-conditioned, snippet-grounded two-stage extraction pipeline to extract auditable intervention–outcome evidence from long humanitarian reports, and a context-preserving cell-based triangulation method for cross-context synthesis. On ReliefWeb, the best closed-source setting reaches 90.73% weighted F1, LoRA SFT on Llama-3.1-8B-Instruct reaches 94.15%, and triangulation shows strong positive convergence of cash assistance on food outcomes (LoE=0.865).

Limitations

Due to limitations in funding and manpower, our corpus and case study are constrained by a keyword query ("cash assistance"), English-only accessible text, a limited set of disaster types, and a coarse source taxonomy, which may reduce generalizability to other interventions, languages, hazards, and reporting contexts. The expert-annotated evaluation set is also small; while it allows us to capture overall trends, performance estimates may be sensitive to sampling and annotation guidelines. Our triangulation and aggregation reflect convergence of reported evidence and rely on design choices (e.g., polarity normalization, weighting, smoothing) that can influence the resulting scores. We note that the strength labels in this work reflect the strength of reported evidence rather than effect magnitude or population-scale impact, which we leave for future work. Therefore, the outputs should be treated as decision support with human review rather than as definitive conclusions.

Ethical Considerations

This work analyzes publicly available humanitarian reports and does not intentionally collect or infer sensitive personal information. Nevertheless, reports may contain incidental mentions of individuals or vulnerable groups; we therefore use the data only for research purposes and recommend caution when releasing derived artifacts. Because ReliefWeb serves as an upstream source, our method may inherit biases in the reporting ecosystem (e.g., uneven geographic coverage, organizational incentives, and language accessibility), which can affect extracted evidence and aggregated scores. In addition, assessments of humanitarian measures in text may themselves be subjective or contested, and LLM-based extraction may further reflect hidden or unknown model biases. To mitigate this, we attach supporting snippets for auditability and triangulate across disaster×source contexts to reduce reliance on any single report or context. Therefore, the outputs should not be used as a standalone basis for operational choices. We position the system as an assistive tool to surface and summarize reported evidence, and we recommend human review—especially for contested or ambiguous claims and for any downstream decisions that

may impact affected populations.

Acknowledgements

This work is funded by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004509, and the Collaboration of Humanities and Social Sciences in Europe (CHANSE) research project Digital Emergency Communication (DIGeMERGE).

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Jo Burton. 2020. "doing no harm" in the digital age: What the digitalization of cash means for humanitarian action. *International Review of the Red Cross*, 102(913):43–73.

Cash Learning Partnership (CaLP). 2015. Operational guidance and toolkit for multipurpose cash grants. Enhanced Response Capacity Project 2014–2015.

Zhiwei Chen, Yupeng Hu, Zhiheng Fu, Zixu Li, Jiale Huang, Qinlei Huang, and Yinwei Wei. 2026. Intent: Invariance and discrimination-aware noise mitigation for robust composed image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(25):20463–20471.

Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, Xuemeng Song, and Liqiang Nie. 2025. Offset: Segmentation-based focus shift revision for composed image retrieval. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 6113–6122, New York, NY, USA. Association for Computing Machinery.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark to evaluate rationalized nlp models. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 4443–4458.

Shannon Doocy and Hannah Tappis. 2017. Cash-based approaches in humanitarian emergencies: a systematic review. *Campbell Systematic Reviews*, 13(1):1–200.

Julie Freccero, Audrey Taylor, Joanna Ortega, Zabihullah Buda, Paschal Kum Awah, Alexandra Blackwell, Ricardo Pla Cordero, and Eric Stover. 2019. Safer cash in conflict: exploring protection risks and barriers in cash programming for internally displaced persons in cameroon and afghanistan. *International Review of the Red Cross*, 101(911):685–713.

Xueren Ge, Sahil Murtaza, Anthony Cortez, and Homa Alemzadeh. 2026. Emsdialog: Synthetic multi-person emergency medical service dialogue generation from electronic patient care reports via multi-llm agents. *Preprint*, arXiv:2604.07549.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3929–3938. PMLR.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Yupeng Hu, Zixu Li, Zhiwei Chen, Qinlei Huang, Zhiheng Fu, Mingzhu Xu, and Liqiang Nie. 2026. Refine: Composed video retrieval via shared and differential semantics enhancement. *ACM Trans. Multimedia Comput. Commun. Appl.* Just Accepted.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow,

Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, and 399 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Todd D Jick. 1979. Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, 24(4):602–611.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Junxian Li, Xinyue Xu, Sai Ma, and Sichao Li. 2025. Faithact: Faithfulness planning and acting in mllms. *arXiv preprint arXiv:2511.08409*.

Zixu Li, Yupeng Hu, Zhiwei Chen, Qinlei Huang, Guozhi Qiu, Zhiheng Fu, and Meng Liu. 2026a. Retrack: Evidence-driven dual-stream directional anchor calibration network for composed video retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23373–23381.

Zixu Li, Yupeng Hu, Zhiwei Chen, Shiqi Zhang, Qinlei Huang, Zhiheng Fu, and Yinwei Wei. 2026b. Habit: Chrono-synergia robust progressive learning framework for composed image retrieval. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6762–6770.

Honglin Lin, Zheng Liu, Yun Zhu, Chonghan Qin, Juekai Lin, Xiaoran Shang, Conghui He, Wentao Zhang, and Lijun Wu. 2026. Mmfinereason: Closing the multimodal reasoning gap via open data-centric methods. *Preprint*, arXiv:2601.21821.

Peiyang Liu, Ziqiang Cui, Di Liang, and Wei Ye. 2025a. Who stole your data? a method for detecting unauthorized rag theft. *Preprint*, arXiv:2510.07728.

Peiyang Liu, Xi Wang, Ziqiang Cui, and Wei Ye. 2025b. Queries are not alone: Clustering text embeddings for video search. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, page 874–883, New York, NY, USA. Association for Computing Machinery.

Peiyang Liu, Jinyu Yang, Lin Wang, Sen Wang, Yunlai Hao, and Huihui Bai. 2023. Retrieval-based unsupervised noisy label detection on text data. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 4099–4104, New York, NY, USA. Association for Computing Machinery.

Zheng Liu, Hao Liang, Bozhou Li, Wentao Xiong, Chong Chen, Conghui He, Wentao Zhang, and Bin Cui. 2025c. Synthvlm: Towards high-quality and efficient synthesis of image-caption datasets for vision-language models. *Preprint*, arXiv:2407.20756.

Yunbo Long, Yuhan Liu, and Liming Xu. 2026. Emomas: Emotion-aware multi-agent system for high-stakes edge-deployable negotiation with bayesian orchestration. *Preprint*, arXiv:2604.07003.

Yunbo Long, Yuhan Liu, Liming Xu, and Alexandra Brintrup. 2025. Emodebt: Bayesian-optimized emotional intelligence for strategic agent-to-agent debt recovery. *Preprint*, arXiv:2503.21080.

Kexin Ma, Ruochun Jin, Wang Haotian, Wang Xi, Huan Chen, Yuhua Tang, and Qian Wang. 2024. Context-driven index trimming: A data quality perspective to enhancing precision of RALMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4886–4901, Miami, Florida, USA. Association for Computational Linguistics.

Xiaoxu Ma, Xiangbo Zhang, and Zhenyu Weng. 2026. Stable and explainable personality trait evaluation in large language models with internal activations. *Preprint*, arXiv:2601.09833.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. 2024. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand. Association for Computational Linguistics.

ReliefWeb. 2025. Reliefweb api documentation. `https://apidoc.reliefweb.int/`. Service provided by UN OCHA; accessed 2025-12-20.

Roberta Rocca, Nicolò Tamagnone, Selim Fekih, Ximena Contla, and Navid Rekabsaz. 2023. Natural language processing for humanitarian action: Opportunities, challenges, and the path toward humanitarian nlp. *Frontiers in big Data*, 6:1082787.

Xuanyu Shi, Wenjing Zhao, Ting Chen, Chao Yang, and Jian Du. 2025. Evidence triangulator: using large language models to extract and synthesize causal evidence across study designs. *Nature Communications*, 16(1):7355.

Qianli Wang, Van Bach Nguyen, Nils Feldhus, Luis Felipe Villa-Arenas, Christin Seifert, Sebastian Möller, and Vera Schmitt. 2025a. Truth or twist? optimal model selection for reliable label flipping evaluation in LLM-based counterfactuals. In *Proceedings of*

the 18th International Natural Language Generation Conference, pages 80–97, Hanoi, Vietnam. Association for Computational Linguistics.

Zifeng Wang, Lang Cao, Benjamin Danek, Qiao Jin, Zhiyong Lu, and Jimeng Sun. 2025b. Accelerating clinical evidence synthesis with large language models. *npj Digital Medicine*, 8(1):509.

Qianyun Yang, Zhiwei Chen, Yupeng Hu, Zixu Li, Zhiheng Fu, and Liqiang Nie. 2026. Stable: Efficient hybrid nearest neighbor search via magnitude-uniformity and cardinality-robustness. *Preprint*, arXiv:2604.01617.

Yuan Yao, Deming Ye, Peng Li, Xu Han, Yankai Lin, Zhenghao Liu, Zhiyuan Liu, Lixin Huang, Jie Zhou, and Maosong Sun. 2019. DocRED: A large-scale document-level relation extraction dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 764–777, Florence, Italy. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Qianchi Zhang, Hainan Zhang, Liang Pang, Yongxin Tong, Hongwei Zheng, and Zhiming Zheng. 2026a. Less is more: Compact clue selection for efficient retrieval-augmented generation reasoning. In *Proceedings of the ACM Web Conference 2026*, WWW '26, page 1971–1982, New York, NY, USA. Association for Computing Machinery.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Wenyuan Zhang, Xinghua Zhang, Haiyang Yu, Shuaiyi Nie, Bingli Wu, Juwei Yue, Tingwen Liu, and Yongbin Li. 2026b. Expseek: Self-triggered experience seeking for web agents. *Preprint*, arXiv:2601.08605.

Yuanjun Zhang, Fuzel Ahamed Shaik, Suvojit Acharjee, Fahad Khalid, and Mourad Oussalah. 2026c. Towards reliable multimodal disaster severity assessment through preference optimization and explainable vision-language reasoning. *Reliability Engineering System Safety*, page 112674.

A Data Filtering and Preprocessing

This section describes the dataset before filtering, including the full set of reports retrieved from ReliefWeb, the distribution across disaster types and source categories, and the filtering criteria applied to obtain the final dataset used in our experiments.

Retrieval overview and disaster coverage. We first retrieve ReliefWeb reports using the keyword *"cash assistance"* and keep English reports with accessible plain text (2000–2024). Tab. 3 summarizes the disaster-type distribution *before* applying our final selection criteria. The corpus spans a wide range of hazards, with FLOOD, EPIDEMIC, DROUGHT, and EARTHQUAKE being the most frequently covered types. To ensure sufficient evidence density and stable cross-context analysis, we focus on these four high-coverage disaster types in subsequent experiments.

Source-type composition over time. Because reporting volume and institutional participation vary substantially across years, we further examine the annual distribution by source type. Tab. 4 provides pre-filter year-by-year counts for the source categories, highlighting both long-term growth in reporting and shifts in which institutions contribute reports. Fig. 3 visualizes the same trend.

Filtering rationale. Based on the above distributions, we apply filtering and selection to balance coverage and interpretability: (i) retain only reports with accessible text and valid metadata, (ii) restrict to the four most prevalent disaster types for stable cell-level analysis, and (iii) map ReliefWeb sources into the source-type taxonomy used in the main experiments. The resulting dataset forms the basis for both extraction evaluation and downstream triangulation.

B Additional Analyses

B.1 Outcome distribution

Motivation for the food-related case study. To choose an outcome family with sufficient evidence for long-horizon triangulation, we inspect the frequency distribution of extracted outcome mentions. Fig. 4 shows a frequency-based word cloud of outcomes extracted by our best-performing configuration, revealing that food-related outcomes appear prominently. This observation motivates the focus on food-related outcomes in the main triangulation analysis (Section 3.2).

Disaster Type	Total Reports
Flood	5907
Epidemic	5878
Drought	5431
Earthquake	3108
Tropical Cyclone	2858
Flash Flood	2134
Land Slide	1860
Other	1322
Tsunami	895
Insect Infestation	511
Severe Local Storm	323
Cold Wave	319
Volcano	272
Storm Surge	265
Technological Disaster	263
Mud Slide	201
Snow Avalanche	79
Wild Fire	73
Fire	35
Heat Wave	34
Extratropical Cyclone	20

Table 3: Distribution (before applying the final selection criteria) of ReliefWeb reports by disaster type (2000–2024).

B.2 API pricing for efficiency metric

Cost accounting. To compare practical deployability across LLMs, we report a cost-efficiency metric in Tab. 2 based on per-token API prices. Tab. 5 lists the input/output prices used in our calculations (per 1M tokens), and we convert qwen-plus and deepseek-v3 pricing from RMB using a fixed exchange rate stated in the caption.

C Sampling Details of the Expert Evaluation Set

Sampling procedure. The expert evaluation set consists of 100 full-length crisis reports randomly sampled from our 8029-report corpus. We use simple random sampling at the report level, without stratification by source type, year, or region. Because each report is typically long and may contain multiple interventions and outcomes, expert annotation in this setting is substantially more time- and expertise-intensive than standard single-label annotation.

Source-type composition of the sampled reports. As a descriptive check, the 100 sampled reports

Year	Academic	Government	Intl. Org.	Media	NGO	Other	RC/RC
2000	0	27	214	36	63	0	103
2001	2	73	315	25	84	0	172
2002	4	60	275	23	73	0	142
2003	3	39	188	8	19	0	99
2004	1	99	217	10	78	2	131
2005	24	316	572	57	471	3	152
2006	14	97	143	6	108	0	100
2007	0	60	157	20	73	0	105
2008	2	73	144	37	83	0	128
2009	4	46	56	32	64	0	111
2010	6	98	118	42	96	2	118
2011	1	106	151	32	163	0	116
2012	8	177	307	43	214	1	175
2013	6	114	312	87	192	0	249
2014	12	195	325	51	243	0	178
2015	5	200	458	43	116	1	178
2016	13	220	1032	18	163	5	210
2017	16	278	1281	13	294	1	226
2018	15	230	798	12	156	0	138
2019	24	211	1057	18	167	2	133
2020	106	232	2508	11	436	7	202
2021	114	202	2230	3	324	10	238
2022	84	226	2018	2	457	3	233
2023	35	185	1585	3	377	5	227
2024	49	206	1477	0	296	5	179
Total	548	3770	17938	632	4810	47	4043

Table 4: Annual number of ReliefWeb reports by source type (2000–2024). Academic = Academic and Research Institution; Government = Government; Intl. Org. = International Organization; Media = Media; NGO = Non-governmental Organization; Other = Other; RC/RC = Red Cross/Red Crescent Movement.

Model	Input cost (USD)	Output cost (USD)	Unit price (USD, avg)
gpt-4o-mini	0.15	0.60	0.375
qwen-plus	0.113	0.283	0.198
deepseek-v3	0.283	1.131	0.707

Table 5: API pricing used in cost-efficiency calculations (prices per 1M tokens). And qwen-plus/deepseek-v3 are converted from RMB using 1 RMB = 0.1414 USD.

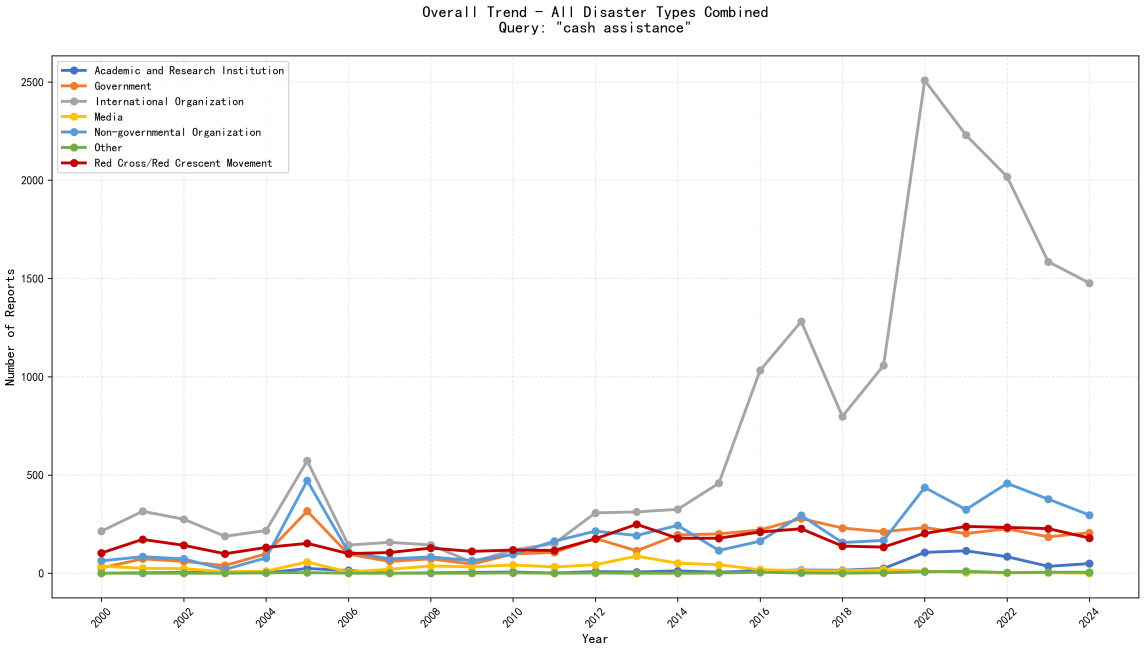

Figure 3: Annual number of ReliefWeb reports by source type.

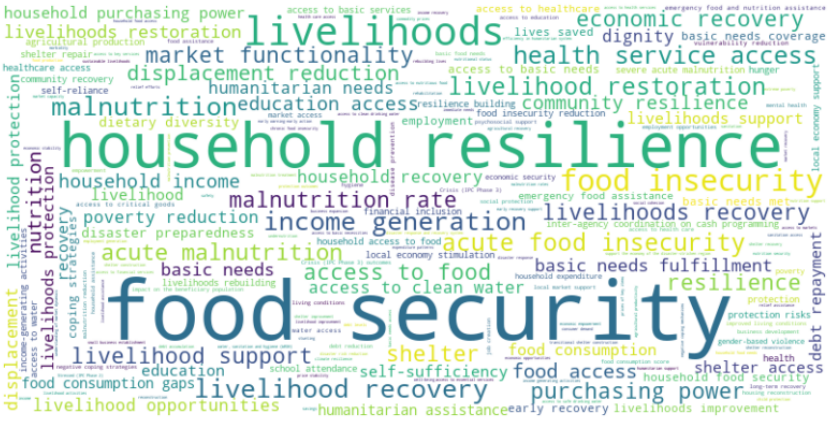

Figure 4: Frequency-based word cloud of extracted outcomes (used to motivate the food-related case study).

cover multiple major source categories in the corpus, including Red Cross/Red Crescent Movement (34 reports), Non-governmental Organization (32 reports), Government (25 reports), Media (5 reports), and Academic and Research Institution (4 reports). Since the evaluation set is obtained through simple random sampling rather than stratified sampling, we do not expect its source-type proportions to exactly match the corpus-level distribution. Instead, this breakdown is reported to show that the sampled reports span the major source categories represented in the corpus.

D Polarity Labeling for Frequent Outcomes

Rationale for labeling polarity only for frequent outcomes. In the triangulation case study, we label polarity only for outcome terms occurring at least 3 times. This decision is motivated by the frequency distribution of outcome terms. Specifically, outcome terms with frequency ≥ 3 comprise only 941 unique vocabularies, yet they cover 13554 out of 19568 samples (69.27%). By contrast, outcome terms occurring only 1–2 times include 5080 unique vocabularies, but together they cover only 6014 out of 19568 samples (30.73%). This distribution reflects a typical long-tail pattern: a relatively small set of frequent outcomes covers the majority of the samples, whereas a large number of rare outcomes contributes limited sample coverage. Given our effort under limited manpower, we prioritize the high-frequency outcomes in polarity labeling. We further clarify that these polarity labels are used only in the triangulation case study and do not affect the main extraction evaluation.

E Supplementary Discussion and Targeted Future Directions

Our results show that query-conditioned extraction and snippet-grounded relation classification can make long humanitarian reports more auditable and more amenable to context-preserving triangulation. Several broader methodological directions also suggest how this framework could be strengthened in future work.

Retrieval reliability and clue selection remain key bottlenecks for long-document humanitarian evidence extraction. Our current corpus construction relies on keyword-based retrieval, which is practical but can still surface reports in which the queried intervention is only marginally discussed. More broadly, retrieval-augmented modeling has long motivated modular and interpretable use of external evidence (Guu et al., 2020). For our setting, future versions of the pipeline could incorporate retrieval-quality control and contextual filtering before extraction. Context-driven index trimming may help remove query-inconsistent retrieved contexts, while compact clue selection can reduce reasoning cost by retaining only the minimum sufficient evidence (Ma et al., 2024; Zhang et al., 2026a). Query clustering and hybrid search further suggest richer retrieval over both free text and structured metadata such as source type, disaster type, and time (Liu et al., 2025b; Yang et al., 2026). Adaptive retrieval and self-reflective evidence use are also relevant: rather than always applying the same fixed pipeline depth, a future system could decide when extra retrieval or verification is necessary and remain robust when some retrieved content is irrelevant (Asai et al., 2024; Zhang et al., 2026b; Yoran et al., 2024).

Ensuring faithful, focused, and robust evidence grounding is equally important in this setting. Our current method attaches supporting snippets and performs relation classification with access to both snippets and the full report, but future systems could evaluate more explicitly whether intermediate reasoning remains faithful to the cited evidence. Work on faithfulness-aware planning and acting is directly relevant here because it emphasizes evidence-constrained reasoning rather than relying only on final-answer quality (Li et al., 2025). Related work on hallucination analysis in retrieval-augmented systems also suggests the value of finer-grained checks for unsupported or weakly supported outputs (Niu et al., 2024). In parallel, robust composed retrieval studies highlight transferable ideas for evidence calibration and focus control, such as aligning heterogeneous signals and reducing the influence of noisy or peripheral content (Li et al., 2026a; Chen et al., 2025). Related work on noise mitigation and differential semantics is also relevant when multiple snippets or contexts must be combined without collapsing distinct signals into a single oversimplified summary (Chen et al., 2026; Hu et al., 2026).

Further improvements are also likely to come from a more data-centric treatment of both the extraction model and the distillation pipeline. Recent work shows that gains can come not only from larger models, but also from better supervision design, stronger filtering, and more robust treatment

of noisy data. For our setting, this suggests several practical extensions: synthesizing or augmenting training instances for rare intervention–outcome patterns, curating higher-quality teacher data, and selecting fine-tuning examples based on difficulty and utility (Liu et al., 2025c; Lin et al., 2026; Ge et al., 2026). Retrieval-based noisy-label detection provides a natural starting point for identifying suspicious labels or low-quality teacher supervision (Liu et al., 2023). Progressive robust learning under noisy correspondence further suggests that student training may benefit from improved filtering, curriculum design, or confidence-aware weighting when teacher outputs are imperfect (Li et al., 2026b).

As retrieval-enhanced evidence systems mature, questions of provenance, release control, modular orchestration, and stable model characterization become increasingly important (Ma et al., 2026). Methods for detecting unauthorized use of retrieval-enhanced data suggest that future evidence platforms may need stronger provenance protection for curated corpora, extracted records, and derived knowledge stores (Liu et al., 2025a). More broadly, high-stakes multi-agent work suggests a natural architectural extension of our current multi-stage pipeline: separate agents could specialize in retrieval, verification, grounding, normalization, and triangulation, while a higher-level controller dynamically allocates trust, requests more evidence, or routes difficult cases for human review (Long et al., 2026, 2025). In our context, such an architecture may be more appropriate than a single monolithic extractor for handling long reports, heterogeneous evidence, and uncertainty-sensitive humanitarian decision support.

F LoRA Fine-Tuning Details

Implementation details (LoRA SFT). We fine-tune an open-weight *Llama-3.1-8B-Instruct* with LoRA (Hu et al., 2022) and train two adapters to match our two-step pipeline (for Step 1 and Step 2). For both adapters, we use rank $r{=}8$ (`lora_target=all`), cutoff length 4096, 3 epochs, learning rate $1\mathrm{e}{-4}$.

Training runs on a single node with $3{\times}$V100 GPUs; we evaluate/save every 111 steps and load the best checkpoint by dev loss. Distillation data are built from a disjoint training split and exclude the 100 expert-annotated evaluation reports.

G Summary of Notations

For readability, Tab. 6 summarizes the main notations used in this paper.

Notation	Meaning
q	Query specifying the intervention class (e.g., *cash assistance*).
e	A structured causal record extracted from a report.
$(a, o, \delta, \sigma, z)$	The tuple form of a causal record defined in Eq. (1).
a	Intervention mention, restricted to the query-defined class q.
o	Outcome mention extracted from the report.
$\delta \in \{\textsc{Inc}, \textsc{Dec}, \textsc{No}\}$	Direction label: increase, decrease, or no clear change.
$\sigma \in \{\textsc{Weak}, \textsc{Mod}, \textsc{Strong}\}$	Strength label of the reported evidence.
z	Supporting text snippet grounding the extracted relation.
$\pi(o) \in \{+1, -1\}$	Polarity of outcome o used for direction normalization.
$\mathcal{D}$	Set of disaster types.
$\mathcal{S}$	Set of source types.
$d \in \mathcal{D}$	A disaster type associated with a record.
$s \in \mathcal{S}$	A source type associated with a record.
$c = (d, s)$	A triangulation cell defined by a disaster–source pair.
$\mathcal{C}$	Set of retained non-empty triangulation cells.
$r \in \{P, Z, N\}$	Normalized relation category: positive, no-change, or negative.
P	Positive relation after polarity normalization.
Z	No-change relation after polarity normalization.
N	Negative relation after polarity normalization.
$w(\sigma)$	Ordinal weight assigned to strength label σ.
$w(\textsc{Weak}) = 0$	Weight for weak evidence.
$w(\textsc{Mod}) = 1$	Weight for moderate evidence.
$w(\textsc{Strong}) = 2$	Weight for strong evidence.
$C_r^{(c)}$	Strength-weighted count of relation type r in cell c in Eq. (2).
$\mathbb{I}(\cdot)$	Indicator function.
$P_r^{(c)}$	Smoothed probability of relation type r within cell c in Eq. (3).
α	Laplace smoothing parameter; in this paper, $\alpha = 0.1$.
P_r	Global cell-averaged probability of relation type r in Eq. (4).
P_P, P_Z, P_N	Global probabilities of positive, no-change, and negative relations.
t	Year index used in cumulative temporal triangulation.

Table 6: Summary of the main notations used in the paper.