

Towards Automated Situation Awareness: A RAG-Based Framework for Peacebuilding Reports

Poli Nemkova¹, Suleyman O. Polat², Rafid I. Jahan², Sagnik Ray Choudhury²,
Sun-joo Lee³, Shouryadipta Sarkar³, Mark V. Albert²

¹Data Science Institute, Columbia University

²University of North Texas

³United Nations Development Programme

poli.nemkova@columbia.edu

{mark.albert, sagnik.raychoudhury}@unt.edu

{suleymanolcaypolat, rafidishrakjahan}@my.unt.edu

{sun-joo.lee, shouryadipta.sarkar}@undp.org

Abstract

Timely and accurate situation awareness underpins effective decision-making in humanitarian response, conflict monitoring, and peacebuilding. Yet, synthesizing heterogeneous and rapidly evolving information from news, conflict databases, and economic indicators remains labor-intensive and delays critical interventions. We present a Retrieval-Augmented Generation (RAG) system that autonomously produces structured, evidence-backed situation awareness reports by integrating real-time data from GDELT, ACLED, ReliefWeb, and World Bank APIs. To rigorously assess report quality without ground-truth references, we introduce a three-level reference-free evaluation framework combining automated NLP metrics, expert review by United Nations crisis analysts, and scalable LLM-as-a-Judge assessment. In a multi-country study across 15 conflict-prone regions, our system generated coherent, relevant, and actionable reports, with UNDP analysts estimating a 50% reduction in report preparation time. These findings demonstrate the feasibility of deploying RAG-based systems for peacebuilding and humanitarian operations and provide a reproducible framework for generating and evaluating AI-assisted situational intelligence at scale.

1 Introduction

Timely and reliable situation awareness is critical for peacekeeping operations, humanitarian response, and governmental interventions. Organizations such as the United Nations (UN), non-governmental organizations (NGOs), and policymakers rely on comprehensive, up-to-date reports to allocate resources, anticipate crises, and mitigate conflict escalation. However, generating these

reports manually is slow, labor-intensive, and complex, requiring extensive data collection, synthesis, and expert analysis across heterogeneous and often incomplete information sources. These delays can undermine the timeliness and effectiveness of interventions, particularly in fast-moving crises where rapid responses are essential.

Large language models (LLMs) have demonstrated strong capabilities in aggregating and summarizing large volumes of text, but their direct use for high-stakes reporting is limited by hallucination, inconsistencies, and poor traceability of evidence. This unreliability makes them unsuitable as standalone tools for peacebuilding or humanitarian decision-making. Retrieval-Augmented Generation (RAG) offers a promising alternative by dynamically grounding generative outputs in retrieved, real-world evidence. By constructing query-specific knowledge bases on demand and conditioning the generation process on relevant sources, RAG enhances factual consistency, reduces hallucination, and increases transparency—properties essential for operational use in peace and crisis management.

In this paper, we present a RAG framework for automated situation awareness reporting in peacebuilding, co-developed and deployed in collaboration with the United Nations Development Programme (UNDP). Our system integrates heterogeneous, up-to-date data—including structured conflict event data, unstructured news reports, humanitarian briefings, and economic indicators—into query-specific knowledge bases constructed on demand. These are then used to generate structured, evidence-backed reports that reduce analyst workload while maintaining expert oversight. Because there are *no gold-standard references* for this type

of report, we propose a three-level reference-free evaluation framework: (1) automated NLP metrics for factuality, coherence, and bias; (2) human expert evaluation by UN crisis analysts for relevance, completeness, and usability; and (3) LLM-as-a-Judge assessments to enable scalable benchmarking of report quality alongside human judgments. We evaluate our framework on 15 real-world geopolitical scenarios across multiple regions. Beyond demonstrating feasibility, our evaluation surfaces where a generic RAG pipeline breaks down in this domain—notably information deduplication, aspect coverage, and degraded performance in regions with sparse media coverage—offering diagnostic findings for future NLP systems supporting conflict monitoring in low-resource information environments.

Our key contributions are as follows:

1. **A domain-adapted RAG framework for peacebuilding:** We present the first documented application of Retrieval-Augmented Generation for situation awareness reporting in peacebuilding and humanitarian contexts, integrating heterogeneous, up-to-date data into query-specific knowledge bases constructed on demand.
2. **A three-level reference-free evaluation framework:** We introduce a multi-layered evaluation approach combining automated NLP metrics, expert review by UN crisis analysts, and LLM-as-a-Judge assessments for scalable, benchmarked evaluation of report quality in the absence of gold-standard references.
3. **Stakeholder-driven design and real-world validation:** We co-developed and iteratively refined the system with UNDP experts, who currently use generated reports as a base for review; participating experts estimate an approximately 50% reduction in report preparation time (from roughly two weeks to one).
4. **Diagnostic findings for humanitarian NLP:** Our evaluation identifies concrete failure modes of RAG-based reporting in this domain—redundancy, incomplete aspect coverage, and sensitivity to regional disparities in media coverage—informing the design of future systems for low-resource conflict monitoring.
5. **Open and reproducible research:** We share our implementation and sample reports to facilitate transparency, reproducibility, and fu-

ture research on AI-driven situation awareness systems.

2 Literature Review

Retrieval-Augmented Generation (RAG) represents an advanced methodology that combines the generative capabilities of large language models (LLMs) with the precision of information retrieval systems. By integrating external knowledge sources, RAG enhances the contextual relevance and factual accuracy of generated content. This hybrid approach has found applications across various domains, including safety, finance, healthcare, and scientific research, offering notable improvements in the quality and efficiency of report generation. The diverse methods of implementing this framework underscore its substantial benefits (Gao et al., 2023; Fan et al., 2024; Arslan et al., 2024). Recently, the RAG architecture has been increasingly adopted across multiple domains.

RAG in Safety Report Generation. In the domain of safety, RAG frameworks have been customized to produce comprehensive reports based on work session descriptions and logs. These systems leverage models such as Llama and employ various embedding techniques to automate the generation of safety reports that comply with stringent documentation standards. Studies utilizing aviation safety datasets have demonstrated the effectiveness of RAG models in improving report accuracy and efficiency, with performance metrics such as Recall@5, GLEU, METEOR, and BERTScore showcasing significant enhancements over traditional reporting methods (Bernardi et al., 2024).

RAG for Financial Report Analysis. The financial sector has adopted RAG to enhance the analysis and interpretation of financial reports, particularly in question-answering tasks for private investors (Iaroshev et al., 2024). Models like OpenAI’s ADA and GPT-4 have been employed to process half-yearly and quarterly reports with high accuracy and contextual relevance. Research findings indicate that the quality of financial report structuring plays a crucial role in optimizing RAG performance, especially when addressing qualitative queries.

RAG in Medical Report Writing. Healthcare applications of RAG, particularly in radiology, involve the automated generation of radiology reports through the use of multimodal embeddings (Ranjit et al., 2023). These embeddings facilitate the re-

trieval of relevant radiology texts, which are then processed by generative models such as OpenAI's GPT series. The integration of user intents and specific clinical requirements into the generative process has resulted in improved clinical metrics, including BERTScore and Sem score, while also mitigating issues related to hallucinated content. This and other application of RAG framework for report automation in medicine are showing promising results (Markey et al., 2025; Wang et al., 2024; Mahalakshmi et al., 2024; Alam et al., 2025)

Scientific Research Summarization Using RAG. In the context of large-scale scientific research, RAG-based summarization agents have been deployed to manage and synthesize vast volumes of information (Suresh et al., 2024). These agents utilize vector databases to retrieve relevant content, enabling LLMs to generate concise summaries enriched with citations. The application of RAG in scientific research not only enhances data comprehension but also fosters collaborative engagement among researchers.

RAG and Broader NLP Applications in Peacebuilding. While advancements in Natural Language Processing (NLP), particularly the integration of Large Language Models (LLMs) with Retrieval-Augmented Generation (RAG), have demonstrated significant benefits across various domains, their application in peacebuilding remains an emerging area of exploration. Although progress has been made, the adoption of these technologies for conflict prevention, resolution, and post-conflict recovery is still in its early stages. Nevertheless, recent research has successfully leveraged advanced NLP techniques for conflict prediction and the monitoring of human rights violations (Trivedi et al., 2021; Alhelbawy et al., 2020; Mueller et al., 2024; Nemkova et al., 2023). These developments highlight the potential for further integration of RAG and other NLP methodologies to support peacebuilding initiatives and humanitarian efforts.

Multimodal RAG Systems. The development of multimodal RAG systems has expanded the capabilities of report generation by incorporating diverse data types, including text, tables, and images. These systems improve retrieval and content generation by considering interrelationships between different modalities. Empirical evaluations across multiple datasets have demonstrated the effectiveness of multimodal RAG in generating accurate and contextually enriched reports (Joshi et al., 2024; Xia

et al., 2025).

Challenges and Practical Solutions in RAG Implementation. Despite the promising advancements, the implementation of RAG systems presents several technical challenges, including data preprocessing, retrieval indexing, and response generation (Khan et al., 2024). Practical experience suggests that the integration of generative AI with precise retrieval mechanisms is essential for ensuring transparency, accuracy, and contextual relevance. Solutions such as leveraging OpenAI's Assistant API and Llama's open-source models have been proposed to enhance the reliability and robustness of RAG-based applications.

Gap in Existing Research. Despite these advancements, to our best knowledge, no prior work has applied RAG frameworks to the generation of structured situation awareness reports for peacebuilding or humanitarian decision-making. Existing applications in safety, healthcare, finance, and scientific research primarily focus on well-defined reporting tasks with accessible domain-specific references, whereas peacebuilding requires synthesizing heterogeneous, dynamic, and often incomplete data from multiple sources without gold-standard references. Moreover, evaluating such reports presents unique challenges: there are no benchmark datasets or established metrics tailored to peacebuilding intelligence products. This lack of domain-specific adaptation and reference-free evaluation strategies leaves a critical gap in both the methodological and applied literature—a gap our work directly addresses by introducing a RAG framework and a multi-level, reference-free evaluation pipeline co-developed with stakeholders in the peacebuilding domain.

3 Method and System Design

The overall architecture is shown in Figure 1. Given a country and date range, the system retrieves heterogeneous evidence from multiple external sources, normalizes the retrieved records into textual evidence units, constructs a vector-based knowledge base, and retrieves query-relevant evidence for report generation. The retrieved evidence is subsequently organized by source and incorporated into a structured prompt for an LLM. Finally, the generated report is evaluated and accompanied by topical and quantitative summaries for analyst review.

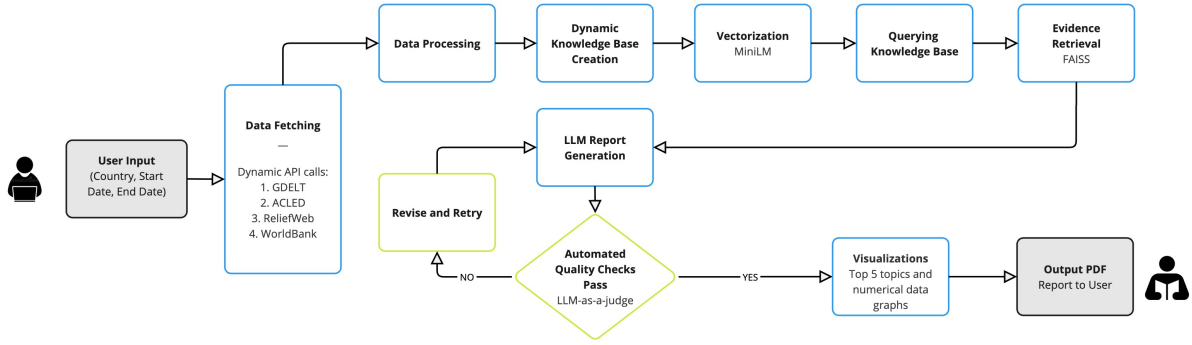


Figure 1: Overview of the situation-awareness reporting framework, from multi-source data retrieval and evidence indexing to report generation, evaluation, and analyst-facing outputs.

3.1 Data Retrieval

The system accepts three user-defined inputs—country, start date, and end date—and dynamically retrieves evidence from four sources that capture complementary dimensions of a developing situation.

GDELT (Global Database of Events, Language, and Tone)¹ provides event metadata and links to news coverage. Because event-level metadata alone provides limited narrative context, we additionally retrieved the full text of linked news articles using newspaper3k². ACLED (Armed Conflict Location & Event Data Project)³ provides structured political-violence event records, including event typologies and reported fatalities. ReliefWeb⁴ contributes humanitarian briefings and situation reports published by UN agencies and NGOs. Finally, the World Bank API⁵ provides structured economic indicators, including GDP, GDP growth, inflation, unemployment, and military expenditure.

These sources therefore contribute complementary evidence types: news coverage from GDELT, structured conflict-event information from ACLED, humanitarian reporting from ReliefWeb, and longer-term socioeconomic context from the World Bank. Rather than passing the heterogeneous API outputs directly to the language model, the system first converts them into a common textual representation, as described below.

3.2 Preprocessing and Evidence Representation

Because the retrieved sources differ substantially in structure, we normalize their contents into textual evidence units before indexing and retrieval. Unstructured GDELT news articles and ReliefWeb reports are retained as text, whereas structured numerical information from ACLED and the World Bank is converted into natural-language statements. For example, an economic indicator may be represented as “GDP growth in Sudan for 2023 was 2.3%.”

During preprocessing, we additionally deduplicate textual data, standardize country names and dates, and remove records with missing values. This normalization allows structured event and economic records and unstructured narrative sources to be represented within the same text-based retrieval pipeline while preserving their source identity for downstream attribution.

3.3 Knowledge Base Construction and Retrieval

The normalized evidence units from all four data sources are encoded using MiniLM-L6-v2⁶ and indexed together in a FAISS vector store (Douze et al., 2024). The vector store therefore serves as a unified semantic retrieval layer over the dynamically constructed multi-source knowledge base.

For each input, we formulate the retrieval query as “Conflict and social unrest issues in {country} between {start_date} and {end_date}.” We compute cosine similarity between the query embedding and the indexed evidence units and retrieve the 10 highest-ranked items across the combined knowledge base. Thus, the retrieved evidence may

¹<https://www.gdeltproject.org/>

²<https://pypi.org/project/newspaper3k/>

³<https://acleddata.com/>

⁴<https://reliefweb.int/>

⁵<https://pypi.org/project/wbgapi/>

⁶<https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2>

originate from any of the four sources depending on its semantic relevance to the query.

The retrieved evidence is then passed to the prompt-construction stage while retaining its source label. During prompt compilation, the retrieved items are grouped under source-specific headings, and each text item is truncated to 500 characters for token efficiency. The resulting source-attributed evidence block is provided to the language model together with the report-generation instructions described in Section 3.4.

3.4 Prompt Design

We evaluated two prompting strategies while holding the requested report structure constant. The *Instructional Prompt* directly asks the model to generate a structured situation-awareness report containing an optional *Important Ongoing Situation* section, followed by *Key Recent Insights, Trends*, and *Recommendations*. Recommendations are explicitly labeled as experimental.

The *Persona Prompt* requests the same output structure but additionally frames the model as “a conflict analyst preparing a situation awareness report for humanitarian decision-makers.” This condition tests whether domain-specific role framing improves report quality beyond explicit task instructions alone.

For both prompting strategies, the retrieved evidence is inserted under source-specific headings before the reporting instructions. The model is explicitly instructed to cite the corresponding source whenever it uses a factual or numerical claim. Full prompt-construction templates are provided in Appendix A.

3.5 Report Generation and Analysis

We generated reports using GPT-4o⁷ via the OpenAI API and Llama-3-8B-Instruct⁸. Reports were stored in plain-text format for subsequent analysis and rendered as PDF documents for presentation to analysts.

In addition to the narrative report, the system generates supplementary topical and quantitative summaries intended to facilitate rapid interpretation of the underlying evidence. For topical analysis, we apply Non-negative Matrix Factorization (NMF) separately to GDELT news articles, ReliefWeb reports, and ACLED event summaries.

We specify five components for each source and extract the five highest-weighted terms for each component. The resulting source-specific topic summaries provide a compact representation of the dominant themes and make it possible to compare which issues are emphasized across news, humanitarian, and conflict-event sources.

The system also generates quantitative visualizations from structured data. ACLED records are used to visualize daily reported fatalities over the selected period, while World Bank data provide longer-term socioeconomic context through ten-year trajectories of GDP, GDP growth, inflation, and unemployment. These visual outputs complement the generated narrative by allowing analysts to inspect salient conflict and economic trends directly rather than relying exclusively on the LLM-generated synthesis.

3.6 Computational Setup & Reproducibility

All experiments were conducted using Google Colab Pro+ with NVIDIA A100 GPUs. To facilitate reproducibility, we provide the complete implementation for data retrieval, preprocessing, knowledge-base construction, evidence retrieval, and report generation,⁹ together with the prompt-construction templates in Appendix A and a sample generated report in Appendix D.

4 Evaluation

We evaluate the generated reports using a three-level reference-free framework (Figure 2) designed to capture complementary dimensions of report quality: (1) automated NLP metrics, (2) human expert evaluation, and (3) LLM-as-a-Judge assessment. The three levels are not sequential filtering stages and are not combined into a single overall quality score. Instead, each provides a distinct perspective on the reports: automated metrics assess properties that can be measured at scale, domain experts evaluate practitioner-relevant qualities such as relevance, completeness, and usability, and LLM judges provide an exploratory mechanism for scaling similar assessments.

This design reflects a central challenge of situation-awareness reporting: unlike conventional summarization tasks, there are no gold-standard reference reports against which generated outputs can be directly matched. We therefore evaluate fac-

⁷<https://openai.com/index/hello-gpt-4o/>

⁸<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

⁹<https://github.com/PoliNemkova/rag-peacebuilding-reports>

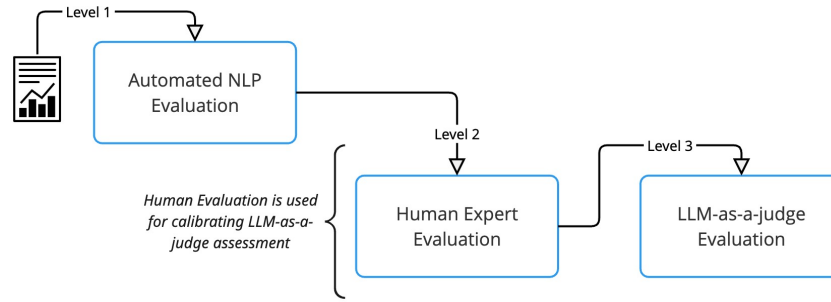


Figure 2: Complementary three-level framework for evaluating generated situation-awareness reports. The three levels assess different dimensions of report quality and are analyzed separately rather than combined into a single aggregate score.

tuality, evidence consistency, coherence, relevance, completeness, and practical usability through multiple complementary signals. Human expert judgments are treated as the primary domain-specific assessment, while automated and LLM-based evaluations provide additional diagnostic information.

4.1 Test Sample Selection

We generated reports for 15 distinct input sets, each defined by a country and reporting period, across several geopolitical regions:

- Middle East (ME): Iran, Israel, Syria, Lebanon, Yemen;
- Eastern Europe (EE): Ukraine, Russia;
- Horn of Africa (HOA): Sudan, Ethiopia, Somalia, South Sudan;
- Asia: Myanmar, China.

For each input set, we generated reports using two LLMs (GPT-4o and Llama-3-8B-Instruct) and two prompting strategies (instructional and persona), resulting in 60 reports. The input sets covered reporting periods of 1 month, 3 months, and 1 year.

4.2 Level 1: Automated NLP Metrics

The first evaluation level uses automated metrics to provide scalable diagnostic signals for factual accuracy, consistency with retrieved evidence, political bias, and coherence across all generated reports (Table 1). These metrics are not used as gating criteria: all 60 reports were subsequently evaluated by human experts and LLM judges. This allows us to examine whether automated quality signals are consistent with practitioner judgments rather than assuming that any individual metric is sufficient to determine report quality.

For factual accuracy, we used VERISCORE (Song et al., 2024) to evaluate verifiable claims

against Google search results and additionally adapted the method to evaluate claims against evidence contained in the dynamically constructed knowledge base. We release the customized implementation publicly.¹⁰

For *consistency with evidence*, we applied SummaC (Laban et al., 2022) to assess the faithfulness of generated content relative to the retrieved evidence. For *political bias*, we used politicalBiasBERT (Baly et al., 2020) to detect ideological skew, using its classification scale from left-leaning to politically neutral to right-leaning content. Finally, for *coherence and clarity*, we computed semantic coherence using BERT-based similarity measures (Devlin et al., 2019).

Together, these metrics provide complementary automated diagnostics, but we do not interpret them as a complete measure of report quality because several practitioner-relevant properties—particularly completeness and operational usefulness—require domain expertise.

4.3 Level 2: Human Expert Evaluation

The second evaluation level focuses on domain-expert assessment. Two UNDP crisis analysts, representing the intended users of the system, independently evaluated each generated report. Both participated voluntarily; one evaluator was female and the other male. Detailed evaluator instructions are provided in Appendix B.

The questionnaire contains two components. Part A consists of seven binary questions assessing relevance and completeness. These include whether the report is relevant to the requested context, whether more than 50% and more than 90% of its content is relevant, whether it avoids dupli-

¹⁰<https://github.com/PoliNemkova/rag-peacebuilding-reports>

Evaluation Metric	GPT-generated, prompt_1	GPT-generated, prompt_2	Llama- generated, prompt_1	Llama- generated, prompt_2
<i>Accuracy</i>				
Average VERIScore	0.76	0.73	0.91	0.79
Average RAG VERIScore	0.65	0.69	0.60	0.61
<i>Consistency</i>				
SummaC	0.52	0.57	0.64	0.67
<i>Objectivity/Bias</i>				
Politically Center Confidence	0.99	0.99	0.99	0.99
<i>Coherence/Clarity</i>				
Coherence Score	0.79	0.81	0.81	0.84

Table 1: Automated Evaluation of Reports with NLP methods (Level 1 Evaluation)

Evaluation Metric	GPT-as-a- judge	Llama-as-a- judge	Claude-as-a- judge
Average Score for GPT generated reports	1	0.82	0.95
Average Score for Llama generated reports	0.93	0.78	0.98
Poorly Performed Questions with GPT generated reports	None	Q3	Q3
Poorly Performed Questions with Llama generated reports	Q4	Q3	Q3

Table 2: Comparison of GPT and Llama judgment of the reports (Level 3 Evaluation)

cate and irrelevant information, whether it appears complete, and whether it covers relevant economic, political, social, or humanitarian dimensions.

Part B consists of pairwise preference judgments. Experts compared alternative reports in terms of completeness, accuracy, and overall preference for use in their work. The complete questionnaire is provided in Appendix C.

The binary criteria are intended as descriptive indicators rather than independent dimensions of equal conceptual weight. In particular, the 50% and 90% relevance questions represent nested relevance thresholds. We therefore report the aggregate binary score as a compact descriptive summary alongside question-level results and preference judgments rather than treating it as a calibrated measure of overall report quality.

4.4 Level 3: LLM-as-a-Judge

The third evaluation level examines whether LLM-based evaluation can provide a scalable supplementary signal for report assessment. GPT-4o and Llama-3 evaluated all reports using the same questionnaire administered to the human experts. Be-

cause these models also generated reports in our experiment, we additionally used Claude-2 as an independent third judge to reduce reliance on self-evaluation.

The use of an identical questionnaire enables a direct comparison of the patterns identified by domain experts and LLM judges. However, we do not treat LLM judgments as calibrated substitutes for expert evaluation, nor do we combine human and LLM scores into a single metric. Instead, the LLM-as-a-Judge experiment is exploratory: it allows us to examine differences in scoring severity, model preferences, and identified failure modes, and to assess the potential and limitations of scaling report evaluation with automated judges. Section 5.1 compares these patterns with the human expert results.

5 Results and Discussion

Automated evaluation. Results of the Level 1 automated evaluation are shown in Table 1. Llama-generated reports scored higher on VERIScore against web search (0.79–0.91 vs. 0.73–0.76 for GPT), whereas GPT-generated reports scored

slightly higher on RAG-VERISCORE, which evaluates verifiability against the retrieved evidence (0.65–0.69 vs. 0.60–0.61). SummaC consistency and coherence scores were moderately higher for Llama-generated reports, while all configurations received a political center confidence of 0.99. These automated rankings do not fully align with the expert preferences reported below, illustrating the value of using complementary evaluation signals.

Human evaluation. Results of the human evaluation are presented in Table 3. Inter-annotator agreement between the two experts, measured using Cohen’s Kappa (Cohen, 1960), was moderate overall, ranging from 0.54 to 0.57 for GPT-generated reports and from 0.42 to 0.54 for Llama-generated reports.

In the binary evaluation of Part A, GPT-generated reports averaged 62% (prompt 1) and 64% (prompt 2) of the total possible points, while Llama-generated reports averaged 60% and 63%, respectively. These values should not be interpreted as conventional accuracy scores or as indicating that the reports are “60% correct.” Rather, they represent the proportion of seven binary quality criteria satisfied according to expert judgments. In addition, Q2 and Q3 represent nested relevance thresholds of more than 50% and more than 90%, respectively. We therefore treat the aggregate binary score as a descriptive summary rather than a calibrated measure of overall report quality.

In the preference-based evaluation of Part B, experts favored the instructional prompt over the persona prompt in 76% of comparisons for GPT-generated reports and in 65% for Llama-generated reports. The lowest-scoring aspects were broadly similar across configurations. Questions 4, 5, and 7, concerning duplication, irrelevant information, and aspect coverage, were among the most common weaknesses. Llama-generated reports additionally showed difficulties on Question 6, which concerns report completeness.

We also observed regional variation in report quality, with neither model dominating uniformly. Under the instructional prompt, GPT performed best for Asia and Eastern Europe, while Llama performed best for Asia and the Middle East. The Horn of Africa was among the lowest-performing regions across several configurations (Table 3). One plausible contributing factor is disparity in the quantity and richness of available media coverage: lower-coverage regions may provide less in-

formative retrieval pools. Because media-coverage density was not directly measured in this study, however, this explanation should be interpreted as a hypothesis rather than a causal finding.

LLM-as-a-Judge. LLM judges generally assigned higher scores than the human experts. GPT-as-a-Judge assigned an average score of 1.00 to GPT-generated reports and 0.93 to Llama-generated reports. Claude assigned scores of 0.95 and 0.98, respectively, whereas Llama-as-a-Judge was more conservative, assigning 0.82 to GPT-generated reports and 0.78 to Llama-generated reports (Table 2). GPT’s perfect average score for its own outputs is particularly notable and illustrates the potential for self-evaluation effects.

5.1 Human vs. LLM Evaluation

Human evaluators showed moderate inter-annotator agreement, whereas LLM judges generally assigned substantially higher report-quality scores. Although some aggregate preferences were similar between expert and LLM evaluation, the present study does not establish formal agreement between human and LLM judgments.

The evaluators also emphasized different failure modes. Human experts most consistently identified problems associated with duplication and irrelevant content (Q4 and Q5), whereas LLM judges more frequently identified weaknesses associated with Q3. This discrepancy suggests that domain experts and LLM judges may attend to different dimensions of report quality or interpret the same evaluation criteria differently.

We therefore treat LLM-as-a-Judge as a supplementary, scalable evaluation signal rather than a replacement for domain-expert assessment. Future work should evaluate direct human–LLM agreement using larger expert panels.

6 Case Study: Iran – June 1 to June 30, 2025

To illustrate our framework in practice, we present a case study on Iran covering June 1–30, 2025. The system aggregated data from GDELT, ACLED, ReliefWeb, and World Bank, dynamically constructing a query-specific knowledge base. Using the persona prompt with GPT-4o, it generated a structured report summarizing key events, trends, and recommendations.

The analysis identified a major escalation follow-

Evaluation Metric	GPT prompt 1	GPT prompt 2	Llama prompt 1	Llama prompt 2
Cohen’s Kappa Overall	0.54	0.57	0.42	0.54
<i>Binary Evaluation</i>				
Cohen’s Kappa on Binary Evaluation	0.53	0.51	0.50	0.61
<i>Preference-based Evaluation</i>				
Cohen’s Kappa on Preference Evaluation	0.52	0.52	0.26	0.31
Cohen’s Kappa for Q8: Which report is more complete?	0.58	0.58	0.24	0.12
Cohen’s Kappa for Q9: Which report is more accurate?	0.54	0.54	0.17	0.12
Cohen’s Kappa for Q10: Which report do you prefer overall?	0.44	0.54	0.36	0.48
Avg. Max Score (Binary Questions)	0.62	0.64	0.60	0.63
Preferred Report (%)	0.76	0.24	0.65	0.35
Poorly Performed Questions	Q4, Q5, Q7	Q4, Q5, Q6, Q7	Q4, Q5	Q4, Q5
Regional Best Performance	Asia, EE	HOA, ME	Asia, ME	Asia, EE
Regional Worst Performance	HOA, ME	Asia, EE	EE, HOA	ME

Table 3: Comparison of GPT and Llama Reports Based on Human Expert Evaluation (Level 2 Evaluation)

ing Israeli airstrikes on Iranian facilities on June 13, with ACLED data confirming over 600 Iranian and 24 Israeli fatalities. ReliefWeb reports highlighted worsening humanitarian conditions, while GDELT data emphasized heightened nuclear tensions as Iran rejected U.S. demands to reduce uranium enrichment. Topic modeling extracted dominant themes across sources, including nuclear negotiations, military operations, and humanitarian responses.

The report recommended urgent diplomatic de-escalation, humanitarian assistance, and UN-led mediation. This case study demonstrates how our system integrates heterogeneous sources into decision-oriented outputs. The full report, including visualizations and detailed topic extraction, is provided in Appendix D.

7 Conclusion and Future Work

We presented a Retrieval-Augmented Generation (RAG) system for automated situation awareness reporting, integrating real-time data from news, conflict databases, humanitarian briefings, and economic indicators to provide evidence-grounded, structured insights for peacekeeping, humanitarian, and governmental decision-making. Rather than replacing experts, our approach augments analysts’ work, accelerating early-stage intelligence generation while preserving human oversight — including rapid preliminary assessments during sudden conflict outbreaks, before official reporting becomes available.

Future work will focus on incorporating new

data modalities (e.g., social media, geospatial inputs), improving interpretability, and scaling our evaluation framework with broader expert participation and real-world impact studies. We openly share our code, prompts, and evaluation framework and invite researchers, NGOs, and policymakers to collaborate in piloting and advancing AI-driven tools for dynamic situation awareness reporting.

Ethical Statement

Data Usage. All data utilized in this project is sourced exclusively from publicly accessible APIs, ensuring transparency and verifiability. No proprietary or confidential data is used, maintaining compliance with ethical and legal standards.

AI Assistants. ChatGPT was used for text polishing.

Potential Risks. While the system is designed to support humanitarian decision-making, there is a risk of over-reliance on AI-generated reports or misinterpretation of automatically synthesized information. To mitigate this, we emphasize that these outputs are intended to assist, not replace, expert analysis and should be used with appropriate human oversight.

Additionally, unequal media coverage across regions means the system may be least reliable precisely for under-covered crises, and generated recommendations (labeled experimental) should be clearly distinguished from evidence-backed factual summaries in any operational use (Nemkova, 2026).

Acknowledgments

We sincerely appreciate the support of the United Nations Development Programme (UNDP) Crisis Bureau for introducing the research problem and providing valuable insights throughout the study. Their comprehensive feedback and expert evaluation were instrumental in refining our approach and assessing the effectiveness of the generated reports.

Limitations

Our evaluation framework and experiments have several limitations:

- **Human evaluation:** Only two UNDP crisis analysts participated in the expert evaluation. While this provided valuable domain-specific insights, a larger and more diverse expert panel would improve the robustness and generalizability of the findings. Human expert judgments therefore serve as a domain-specific reference point for comparison with LLM-as-a-Judge outputs rather than as a definitive benchmark.
- **LLM-as-a-Judge reliability:** Despite mitigating self-evaluation bias using a third independent model (Claude-2), LLM-based evaluations may still inherit systemic biases from their training data. Moreover, some evaluation questions (e.g., report completeness) require domain knowledge or access to the underlying information landscape that LLM judges lack; their answers to such questions should therefore be treated as supplementary signals rather than substitutes for expert judgment.
- **System components:** Beyond the two prompting strategies, we evaluate the pipeline end-to-end without isolating the contribution of individual stages (e.g., retrieval, evidence selection, and knowledge base construction). In particular, we did not directly evaluate retrieval quality (e.g., top-10 coverage or how reliably semantic search respects the date-range constraint; metadata-based date pre-filtering is a promising alternative), nor the effect of truncating retrieved source texts to 500 characters on report completeness.
- **No direct assessment of decision impact:** While our evaluation captures factuality, coherence, relevance, completeness, and usability,

we do not measure how these reports influence real-world decision-making, such as the speed of resource allocation or the quality of interventions.

- **Time-savings estimate:** The reported $\sim 50\%$ reduction in preparation time is based on an estimate provided by a UNDP manager (approximately two weeks per report under standard workflows), not on a controlled timing study, and should therefore be interpreted as a preliminary, expert-elicited estimate.
- **Topic scope:** Our evaluation focuses on country-level situation reports; testing at sub-national or cross-regional levels could reveal additional challenges related to retrieval quality, scalability, and information availability.

References

- Hasan Md Tusfiquir Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. 2025. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. In *European Conference on Information Retrieval*, pages 201–209. Springer.
- Ayman Alhelbawy, Mark Lattimer, Udo Kruschwitz, Chris Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Systems with Applications*, 153:113365.
- Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790.
- Ramy Baly, Giovanni Da San Martino, James Glass, and Preslav Nakov. 2020. We can detect your bias: Predicting the political ideology of news articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, EMNLP '20, pages 4982–4991.
- Mario Luca Bernardi, Marta Cimitile, and Riccardo Pecori. 2024. Automatic job safety report generation using rag-based llms. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#).
- Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Ivan Iaroshev, Ramalingam Pillai, Leandro Vaglietti, and Thomas Hanne. 2024. Evaluating retrieval-augmented generation models for financial report question and answering. *Applied Sciences*, 14(20):9318.
- Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. 2024. Robust multi model rag pipeline for documents containing text, table & images. In *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999. IEEE.
- Ayman Asad Khan, Md Toufique Hasan, Kai Kristian Kemell, Jussi Rasku, and Pekka Abrahamsson. 2024. Developing retrieval augmented generation (rag) based llm systems from pdfs: An experience report. *arXiv preprint arXiv:2410.15944*.
- Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. Summac: Re-visiting nli-based models for inconsistency detection in summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.
- M Mahalakshmi, Shardul Bharadwaj, and Aklanta Niraz Bhuyan. 2024. A real-time medical report analysis and ai-powered diagnosis: A cloud-based solution for improved patient care. In *2024 Second International Conference on Advances in Information Technology (ICAIT)*, volume 1, pages 1–6. IEEE.
- Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper Van Langen, and Christoph Meier. 2025. From rags to riches: utilizing large language models to write documents for clinical trials. *Clinical Trials*, 22(5):626–631.
- Hannes Mueller, Christopher Rauh, and Ben Seimon. 2024. [Introducing a global dataset on conflict forecasts and news topics](#). *Data & Policy*, 6:e17.
- Poli Nemkova. 2026. Invisible conflicts: Media coverage asymmetry and categorical failure in llm conflict forecasting. In *Trustworthy AI for Good (AI4GOOD) Workshop@ ICML 2026*.
- Poli Nemkova, Solomon Ubani, Suleyman Olcay Polat, Nayeon Kim, and Rodney D. Nielsen. 2023. Detecting human rights violations on social media during russia-ukraine war. *arXiv preprint arXiv:2306.05370*.
- Mercy Ranjit, Gopinath Ganapathy, Ranjit Manuel, and Tanuja Ganu. 2023. [Retrieval augmented chest x-ray report generation using openai gpt models](#). *arXiv preprint arXiv:2305.03660*.
- Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474.
- Karthik Suresh, Neeltje Kackar, Luke Schleck, and Cristiano Fanelli. 2024. Towards a rag-based summarization for the electron ion collider. *Journal of Instrumentation*, 19(07):C07006.
- Anusua Trivedi, Kate Keator, Michael Scholtens, Brandon Haigood, Rahul Dodhia, Juan Lavista Ferres, Ria Sankar, and Avirishu Verma. 2021. [How to handle armed conflict data in a real-world scenario?](#) *Philosophy & Technology*, 34(Suppl 1):111–123.
- Yingding Wang, Simon Leutner, Michael Ingrisch, Christoph Klein, Ludwig Christian Hinske, and Katharina Danhauser. 2024. [Optimizing data extraction: Harnessing rag and llms for german medical documents](#). In *Digital Health and Informatics Innovations for Sustainable Health Care Systems: Proceedings of MIE 2024*, pages 949–950. SAGE Publications.
- Peng Xia, Kangyu Zhu, Haoran Li, Tianze Wang, Weijia Shi, Sheng Wang, Linjun Zhang, James Y Zou, and Huaxiu Yao. 2025. Mmed-rag: Versatile multimodal rag system for medical vision language models. In *International Conference on Learning Representations*, volume 2025, pages 66188–66217.

A Prompt Construction for Report Generation

We provide the function used to compile retrieved evidence into a GPT-ready prompt. This function takes the user query and multi-source text snippets, truncates them for token efficiency, and appends explicit instructions for structured reporting.

1. Instructional Prompt

Listing 1: Function for compiling multi-source evidence into a structured LLM prompt.

```
def compile_data_for_llm(query_text,
    retrieved_data):
    """
    Prepare the extracted data into a
    GPT-ready prompt.
    Args:
        query_text (str): The original
            query.
        retrieved_data (dict): Retrieved
            text organized by dataset.
    Returns:
        str: A formatted input prompt
            for GPT.
    """
    prompt = (
        f"Create a situation awareness
        report based on the
        following "
        f"evidence for the query: '{
        query_text}''.\n\n"
    )

    for dataset, texts in retrieved_data
        .items():
        prompt += f"\n### {dataset.upper
        ()} Data:\n"

        for idx, text in enumerate(texts
        ):
            text = str(text)
            prompt += (
                f"{idx+1}. {text
                [:500]}...\n"
            ) # Truncate to 500
                characters

        prompt += (
            "\nProvide a structured summary
            including sections: "
            "important ongoing situation (if
            any, optional), "
            "key recent insights, trends,
            and recommendations "
            "(label as: Recommendation [
            experimental]). "
            "Whenever you use a number or
            fact, cite the exact "
            "source in parentheses."
        )

    return prompt
```

2. Persona Prompt

Listing 2: Function for compiling multi-source evidence into a persona-based LLM prompt.

```
def compile_data_for_llm_persona(
    query_text, retrieved_data):
    """
    Prepare the extracted data into a
    persona-based GPT-ready prompt.
    Args:
        query_text (str): The original
            query.
        retrieved_data (dict): Retrieved
            text organized by dataset.
    Returns:
        str: A formatted input prompt
            for GPT with persona
            instructions.
    """
    prompt = (
        f"You are a conflict analyst
        preparing a situation
        awareness "
        f"report for humanitarian
        decision-makers. Use the
        evidence "
        f"provided to craft a clear,
        concise, and professional
        report "
        f"for the query: '{query_text
        }''.\n\n"
    )

    for dataset, texts in retrieved_data
        .items():
        prompt += f"\n### {dataset.upper
        ()} Data:\n"

        for idx, text in enumerate(texts
        ):
            text = str(text)
            prompt += (
                f"{idx+1}. {text
                [:500]}...\n"
            ) # Truncate to 500
                characters

        prompt += (
            "\nProvide a structured summary
            with the following sections:
            "
            "important ongoing situation (if
            any), key recent insights,
            "
            "trends, and recommendations "
            "(label this section as:
            Recommendation [experimental
            ]). "
            "Whenever you reference a number
            or fact, cite the exact "
            "source in parentheses."
        )

    return prompt
```

B Evaluator Instructions

This appendix provides the instructions given to United Nations Development Programme (UNDP) crisis analysts for evaluating the automatically generated situation awareness reports. Evaluators were informed that the reports are intended to support, not replace, expert analysis.

1. Read the full report carefully.

2. Answer the following questions:

(a) Relevance and Completeness (True/-False):

- i. Is the report relevant to the country and time period?
- ii. Does more than 50% of the content provide relevant information?
- iii. Does more than 90% of the content provide relevant information?
- iv. Does the report avoid duplication of information?
- v. Does the report contain minimal irrelevant content (less than 10%)?
- vi. Does the report seem complete for its purpose?
- vii. Does it cover economic, political, social, and humanitarian aspects?

(b) Preference-Based Comparison:

- i. Between two reports, which is more complete?
- ii. Which is more accurate?
- iii. Which would you prefer to use in your work?

3. Add optional comments on missing information, factual inaccuracies, or sections that could be improved.

Evaluators were asked to review the reports independently and base their answers on their professional judgment.

C Evaluation Questionnaires

This appendix lists the full questionnaires used in the human expert and LLM-as-a-Judge evaluations of the generated reports. These questions formed the basis of both binary assessments and pairwise preference comparisons.

C.1 Level 2: Human Expert Evaluation

Part A: Relevancy and Completeness (True/-False)

1. Is the report relevant?
2. Does more than 50% of the report contain relevant information?
3. Does more than 90% of the report contain relevant information?
4. Does the report avoid duplicate information?
5. Does the report contain no more than 10% irrelevant information?
6. Does the report seem complete?
7. Does the report cover economic, political, social, or humanitarian aspects?

Part B: Preference-Based Comparison

8. Which report is more complete? (Report 1 vs. Report 2, Report 3 vs. Report 4)
9. Which report is more accurate? (Report 1 vs. Report 2, Report 3 vs. Report 4)
10. Which report do you prefer overall? (Report 1 vs. Report 2, Report 3 vs. Report 4)

C.2 Level 3: LLM-as-a-Judge

The same questionnaire (Q1–Q10) was used for LLM-as-a-Judge evaluation.

D Sample Generated Report: Iran (June 1 – July 1, 2025)

This appendix includes the full system-generated situation awareness report for Iran, covering the period June 1 – July 1, 2025.

Situation Awareness Report

IRAN | Situation Awareness Report

Period: 2025-06-01 -> 2025-06-30

Situation Awareness Report: Conflict Situation in Iran from 2025-06-01 to 2025-06-30

Key Ongoing Situation

From June 13, 2025, a significant military conflict has unfolded between Israel and Iran, marking a major escalation in their historically tense relationship. The hostilities began with a series of Israeli airstrikes targeting Iranian infrastructure in locations such as Natanz, Fordow, and Tehran (ReliefWeb Data 1). The Israeli strikes on Iran have led to serious humanitarian concerns, with civilian areas in several cities, particularly Tehran, affected by the ongoing strikes (ReliefWeb Data 1).

Key Recent Insights

- Escalating Conflict: The Israeli-Iranian conflict has escalated significantly since June 14, 2025, with Israel conducting strikes on Iranian military and scientific infrastructure. Iranian missiles are now regularly breaching Israeli defenses with over 600 Iranian deaths and 24 Israeli deaths reported so far (ReliefWeb Data 2).

- International Response: The conflict has drawn international concern. The United Nations Secretary-General, António Guterres, has expressed deep concern over the escalating conflict (ReliefWeb Data 3). Global Affairs Canada, on behalf of the G7 Foreign Ministers, reiterated their support for the ceasefire between Israel and Iran (ReliefWeb Data 5).

- Nuclear Program: Iran's Supreme Leader Ayatollah Ali Khamenei has maintained that abandoning uranium enrichment was "100 per cent" against the country's interests, rejecting a central US demand in talks to resolve the dispute over Tehran's nuclear ambitions (GDELT Data 4).

Trends

- Tensions in the Middle East: Tensions have escalated following reports that U.S. embassies and military bases across the region are taking heightened security measures (GDELT Data 6). Sri Lanka and Armenia have expressed deep concern over the developments and have urged both countries to engage in diplomatic efforts for reconciliation (GDELT Data 7).

- Political Violence: Political violence in the form of battles and armed clashes have been reported in various regions of Iran, including Zahedan, Hasanabad, Iranshahr, and Sanandaj (ACLED Data 1, 2, 3, 8).

- Nuclear Program Controversy: Iran has issued stern warnings against growing Western pressure and potential Israeli aggression as the Islamic Republic prepares for a sixth round of indirect talks with the United States over its nuclear program (GDELT Data 10).

Recommendations (Experimental)

- De-escalation and Diplomacy: All parties involved need to engage in diplomatic efforts to de-escalate the situation and prevent further destabilization of regional and international peace and security (GDELT Data 7).

- Humanitarian Assistance: Given the significant impact on civilians and the growing humanitarian crisis, international organizations and other countries should provide necessary aid and support to the affected population (ReliefWeb Data 1).

- International Mediation: The international community, particularly the United Nations, should play a more proactive role in mediating the conflict between Israel and Iran, to achieve a peaceful resolution (ReliefWeb Data

5).

- Nuclear Negotiations: It's crucial that negotiations over Iran's nuclear programme continue with the aim of reaching a solution that respects Iran's right to peaceful nuclear energy while ensuring non-proliferation commitments are upheld (GDELT Data 4).

=====

Topics / Sections

GDELT | Topics Most Covered

1. Iran Nuclear Deal
----> Keywords: nuclear, iran, trump, say, deal
2. Trump's Israel Advertising Strike
----> Keywords: advertisement, trump, israel, friday, strike
3. Israel-Iran Missile Conflict
----> Keywords: missile, israel, iran, tel, aviv
4. Oil Region Analysis
----> Keywords: wednesday, oil, say, hegseth, region
5. Confidential Document Leakage
----> Keywords: rahman, document, information, secret, leak

ACLED | Topics Most Covered

1. Military Operations
----> Keywords: military, casualty, unknown, target, carry
2. Local Religious Activities
----> Keywords: resident, prayer, local, friday, denounce
3. Air Defense Tactics
----> Keywords: air, force, interception, intercept, shoot
4. Tehran District Unrest
----> Keywords: tehran, district, strike, neighborhood, airdrone
5. Telecom Company Disputes
----> Keywords: company, telecommunication, pension, protest, office

RELIEFWEB | Topics Most Covered

1. Middle East Nuclear Politics
----> Keywords: nuclear, iran, israel, united, right
2. French Humanitarian Aid
----> Keywords: plus, personne, humanitaire, pay, liran
3. Afghan Refugee Crisis
----> Keywords: return, refugee, afghanistan, million, country
4. International Human Rights
----> Keywords: personas, mas, internacional, derecho, derechos
5. Iranian Humanitarian Issues
----> Keywords: civilian, child, humanitarian, iranian, red

Graphs & Visualizations

