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

Anonymous ACL submission

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 dynamic 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, reducing analyst preparation time by nearly 50%. 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 the first domain-adapted RAG framework for automated situation awareness reporting in peacebuilding, designed in collaboration with the United Nations Development Programme (UNDP). Our system dynamically integrates heterogeneous, real-time data—including structured conflict event data, unstructured news reports, humanitarian briefings, and economic indicators—into query-specific knowledge bases. These are then used to generate structured, evidence-backed reports that significantly 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, demonstrating its ability to produce coherent, actionable, and evidence-grounded reports in a fraction of the time required for manual preparation.

Our key contributions are as follows:

1. **A domain-adapted dynamic RAG framework for peacebuilding:** We present the first documented application of dynamic Retrieval-Augmented Generation for situation awareness reporting in peacebuilding and humanitarian contexts, integrating heterogeneous, real-time data into query-specific knowledge bases.
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 validation:** We co-developed and iteratively refined the system with UNDP experts, demonstrating real-world feasibility and reducing analyst report preparation time by approximately 50%.
4. **Open and reproducible research:** We share our implementation and sample reports to facilitate transparency, reproducibility, and future 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 retrieval 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 Semb 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., 2024; Wang et al., 2024; Assistant et al., 2024; Alam et al., 2024)

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., 2020; 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; **?**).

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 dynamic RAG framework and a multi-level, reference-free evaluation pipeline co-developed with stakeholders in the peacebuilding domain.

3 Method and System Design

Our system implements a dynamic RAG framework to generate structured, evidence-backed situation awareness reports. The overall architecture is shown in Figure 1, with six core stages: (1) dynamic data fetching, (2) preprocessing, (3) knowledge base (KB) construction, (4) vectorization, (5) querying the KB, (6)retrieval and evidence selection, (7) report generation via Large Language Models (LLMs), (6) report evaluation loop, and (7) automated topical and visual analysis.

3.1 Data Retrieval

The system accepts user-defined inputs (country, start date, end date) and retrieves relevant data from four publicly accessible APIs:

- GDELT (Global Database of Events, Language, and Tone)[1]: Provides event metadata and links to source articles. We scraped full-text news articles using newspaper3k[2] to enrich context.
- ACLED (Armed Conflict Location & Event Data Project)[3]: Supplies structured political violence data, including fatalities and event typologies.
- ReliefWeb[4]: Delivers humanitarian briefings and situation reports authored by NGOs and UN agencies.
- World Bank[5]: Provides economic indicators (e.g., GDP, growth rate, inflation, unemployment, military expenditure).

[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/

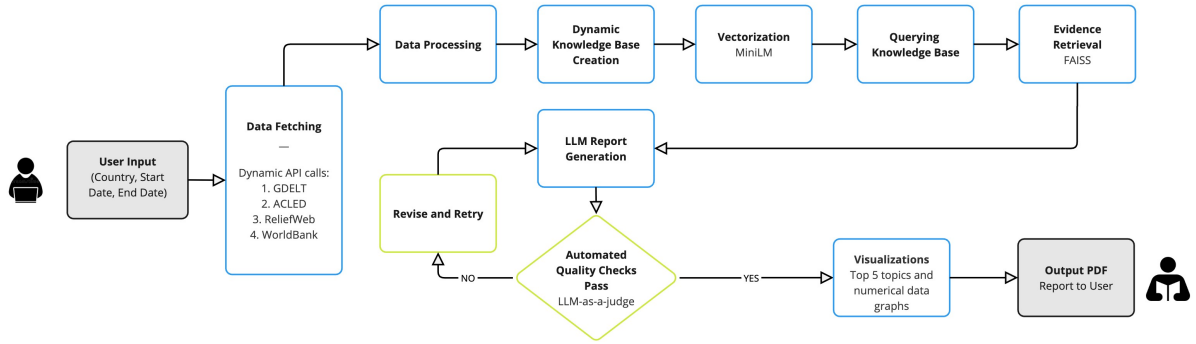


Figure 1: Overall system design presented as a flowchart.

3.2 Preprocessing

To ensure compatibility with LLMs, numerical indicators from ACLED and the World Bank were transformed into natural-language statements (e.g., "GDP growth in Sudan for 2023 was 2.3%"). We deduplicated textual data, standardized dates and country names, and dropped records with missing values.

3.3 Knowledge Base Construction and Retrieval

All textual data were encoded using MiniLM-L6-v2[6] for efficient contextual embeddings. We indexed the embeddings in a FAISS vector store (Douze et al., 2024), enabling semantic search. For each query ("Conflict and social unrest issues in {country} between {start_date} and {end_date}"), we retrieved the top-10 semantically relevant documents based on cosine similarity. These documents formed the evidence set for report generation.

3.4 Prompt Design

We evaluated two prompting strategies:

1. *Instructional Prompt* – explicitly requested a structured report with the following sections:
   (a) Important ongoing situation (optional);
   (b) Key recent insights;
   (c) Trends;
   (d) Recommendations (experimental).
2. *Persona Prompt* – identical structure but framed the model as "a conflict analyst preparing a report for humanitarian decision-makers," encouraging a professional tone.

The model was instructed to cite the source for each numerical or factual statement.

Full prompt templates are included in Appendix 7.

[6]https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2

3.5 Report Generation and Analysis

We used GPT-4o[7] (via OpenAI API) and LLaMA-3-8B-Instruct[8]. Preliminary experiments with DeepSeek (Guo et al., 2025) were excluded due to insufficient output quality. Reports were saved in .txt (for structured analysis) and .pdf (for presentation) formats.

In addition to narrative reports, we automatically generated supplementary analytical outputs:

1. **Topics.** Top-5 emerging topics were identified separately for each textual dataset (GDELT news articles, ReliefWeb reports, and ACLED event summaries) using Non-negative Matrix Factorization (NMF) (scikit-learn implementation). We set the number of components to five per dataset, corresponding to five interpretable topics. For each topic, we extracted the five highest-weighted terms from the NMF components, providing concise keyword lists that capture the dominant themes within each source. This multi-source topic extraction highlights the key issues emphasized across different types of evidence, offering a richer and more granular understanding of the reporting landscape.

2. **Visualizations**:
   - Daily fatalities plot (from ACLED event data).
   - GDP over the past 10 years plot.
   - GDP growth rate over the past 10 years plot.
   - Inflation rate over the past 10 years plot.
   - Unemployment rate over the past 10 years plot.

[7]https://openai.com/index/hello-gpt-4o/

[8]https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

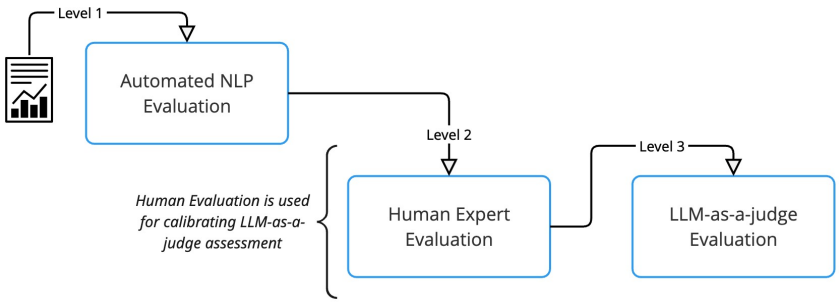


Figure 2: Report evaluation three-level framework.

These additions provide quick-reference insights for analysts and contextualize the generated narrative reports with key quantitative trends.

3.6 Computational Setup & Reproducibility

All experiments were conducted on Google Colab Pro+ with NVIDIA A100 GPUs. To promote reproducibility, we provide:

- Full codebase (data fetching, preprocessing, retrieval, and report generation)[9];
- Prompt templates in Appendix A;
- Sample generated reports for review in Appendix D.

4 Evaluation

We evaluated the performance of our system using a three-level reference-free framework (Figure 2.), designed to capture complementary perspectives on report quality: (1) automated NLP metrics, (2) human expert evaluation, and (3) LLM-as-a-Judge assessments. This design reflects the real-world challenge of situation awareness reporting: no gold-standard references exist for such reports, as they are typically bespoke, analyst-generated products. Consequently, our framework focuses on factuality, coherence, completeness, and usability rather than reference matching.

4.1 Test Sample Selection

We generated reports for 15 distinct input sets (country × date range combinations) across diverse 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 query, we used two LLMs (GPT-4o and LLaMA-3-8B-Instruct) and two prompting strategies (instructional and persona), resulting in **60 reports** covering 1-month, 3-month, and 1-year periods.

4.2 Level 1: Automated NLP Metrics

At the first level, we applied automated tools to screen reports for factual accuracy, coherence, and bias. Reports failing these checks were excluded from further evaluation.

- Factual accuracy: We used VERISCORE (Song et al., 2024), adapted to check factual claims against (a) Google search results and (b) our dynamic knowledge base (customized implementation released on GitHub[10]). Reports required a VERISCORE >= 0.8, meaning that all claims must be verifiable.
- Consistency with evidence: We applied SummaC (Laban et al., 2022) to measure the faithfulness of generated content relative to retrieved evidence, retaining only reports with SummaC >= 0.7.
- Political bias: We used politicalBiasBERT (Baly et al., 2020) to detect ideological skew (scale: 0 = strong left, 1 = neutral, 2 = strong right). Only reports with scores between 0.8 and 1.2 (near-neutral) were accepted.
- Coherence/clarity: We computed semantic coherence using BERT-based similarity measures (Devlin, 2018), requiring a coherence score >= 0.8.

4.3 Level 2: Human Expert Evaluation

At this stage, two UNDP crisis analysts — representing the target users of our system — independently evaluated each report. Both participated on a voluntary basis. One evaluator was female and

[9] GitHub/URL anonymized for review

[10] URL anonymized for review

the other was male. The instructions are available in Appendix B.

Evaluation criteria: Experts assessed reports along two dimensions:

- Part A: Relevance and Completeness (Binary): Seven True/False questions covering relevance (e.g., "Does >90% of the content contain relevant information?"), completeness, avoidance of duplication, and coverage of economic, political, social, and humanitarian dimensions.
- Part B: Preference-Based Comparison: Pairwise comparisons on completeness, accuracy, and overall preference.

The full questionnaire is provided in Appendix C. This combination of binary and preference-based judgments ensured a holistic view of report usefulness for practitioners.

4.4 Level 3: LLM-as-a-Judge

To enable scalable evaluation, we implemented an LLM-as-a-Judge approach. GPT-4o and LLaMA-3 evaluated all reports using the same questionnaire as human experts. To mitigate self-evaluation bias (e.g., GPT-4o scoring its own outputs), we also employed a third independent model (Claude-2) to evaluate all outputs.

Purpose and calibration: While the human and LLM evaluations currently use identical questionnaires, their roles differ. Human expert judgments serve as a calibration benchmark for the LLM-as-Judge outputs, ensuring that automated evaluations remain aligned with domain-expert expectations. This step allows us to scale evaluation to larger datasets in future iterations while preserving human-validated quality standards.

5 Results and Discussion

Human evaluation. Results of human evaluation are presented in Table 1. The inter-annotator agreement between the two human experts, measured using Cohen's Kappa (Cohen, 1960), indicated a moderate level of agreement, with values of 0.54 for GPT-generated reports and 0.57 for LLaMa-generated reports.

In the binary evaluation of Part A (Relevance and Completeness), GPT-generated reports achieved an average of 62% of the total possible points, while LLaMa-generated reports scored slightly higher at 64%.

In the preference-based evaluation of Part B, human experts selected GPT-generated reports in 76% of cases, whereas LLaMA-generated reports were preferred in only 24% of cases.

Notably, the aspects that consistently received the lowest scores were largely similar for both GPT-generated and LLaMa-generated reports. These included questions 4, 5, and 7, which focus on issues of redundant information and the omission of specific aspect of coverage. Additionally, LLaMa-generated reports demonstrated particular difficulties with question 6, which pertains to the completeness of the reports.

LLM-as-a-Judge. Notably, GPT assigned an average perfect score of 100% to reports generated by itself, compared to 93% for those produced by LLaMA. The primary weakness identified by GPT across both sets of reports was the presence of redundant information, as reflected in Question 4 (Q4) of the evaluation criteria. This issue was more pronounced in the LLaMA-generated reports, which received the lowest scores in this category.

5.1 Human vs. LLM Evaluation

The highest Cohen's Kappa values from humans were around 0.57 (GPT-generated, prompt 2) and 0.54 (LLaMA-generated, prompt2). These values indicate moderate agreement among human evaluators. For binary evaluations, humans showed varying agreement, with LLaMA-generated reports (prompt 2) achieving the highest human agreement (0.61). GPT-as-a-judge gave itself a perfect 1.0 score, while Claude scored GPT reports at 0.95—both suggesting near-perfect evaluations. LLaMA, however, was more critical of both itself (0.78) and GPT (0.82).

Human evaluations show more variability and critical judgment, reflected in moderate Cohen's Kappa scores, while LLM judges—especially GPT and Claude—tend to rate reports significantly higher, hinting at possible overconfidence or evaluation bias in LLMs. LLaMA-as-a-judge mirrors human evaluators to some extent by being more critical, particularly towards its own reports. Interestingly, Claude-as-a-judge preferred LLaMA-generated reports (0.98) slightly more than GPT reports (0.95), suggesting Claude's evaluation diverges from human preferences. GPT-as-a-judge rated its own reports higher than LLaMA's, aligning more closely with human preferences. While GPT-as-a-judge aligns with human preferences by favoring its own reports, Claude shows a bias to-

Evaluation Metric	GPT-generated, prompt_1	GPT-generated, prompt_2	LLaMA-generated, prompt_1	LLaMA-generated, prompt_2
Cohen's Kappa Overall	0.54	**0.57**	0.42	0.54
Binary Evaluation				
Cohen's Kappa on Binary Evaluation	0.53	0.51	0.50	**0.61**
Preference-based Evaluation				
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 1: Comparison of GPT and LLaMA Reports Based on Human Expert Evaluation (Level 2 Evaluation)

ward LLaMA-generated reports, diverging from human evaluators. This indicates that LLM judges may not always reflect human judgment, especially when cross-model evaluations are involved. Human evaluators and LLM judges identify different weaknesses in the reports. Humans consistently highlight Q4 and Q5, while LLMs focus on Q3. This discrepancy suggests that LLMs and humans prioritize different aspects of report quality or interpret the evaluation criteria differently.

5.2 Strengths and Limitations of the Approach

Strengths:

- The system dynamically retrieves relevant information, ensuring reports are evidence-based. All the data used in the study is free and publicly available.
- The multi-layered evaluation framework enhances reliability and robustness. Human Evaluation is only used for aligning automated evaluation that will be needed for scaling.
- The use of multiple LLMs (GPT-4o, LLaMA 3) allows for comparative analysis and improved output quality.
- The generated reports significantly reduce the time required for human analysts to draft reports from scratch. Currently a human analyst, on average, can take up to 2 weeks to create similar report. With our system, this time can drop to 1 week (generated report is used as a base for review and refinement).

Limitations:

- Potential biases in retrieved evidence may affect the objectivity of the reports.
- LLMs may struggle with complex geopolitical nuances and context-dependent interpretations.
- Human evaluation introduces subjectivity.
- Despite automation, human review is still mandatory before reports can be delivered to stakeholders (human-in-the-loop requirement).

5.3 Regional Variations in Model Performance

In our study, we observed distinct variations in model performance across different regions. While GPT models generally outperformed LLaMA in overall reporting quality, region-specific differences emerged. GPT demonstrated superior performance in generating reports for Asia and Eastern Europe, whereas LLaMA produced more accurate and contextually relevant outputs for the Middle East and Africa.

A notable factor influencing these results is the disparity in media coverage across regions. Events in Europe and the Middle East tend to receive sig-

nificantly more international attention compared to regions like the Horn of Africa. This uneven distribution of data likely contributes to variations in model performance, as regions with limited coverage may present greater challenges for accurate information retrieval and synthesis.

5.4 Benefits for Real-World Applications

The proposed system offers several distinct advantages for practical implementation:

1. *Enhanced Time Efficiency*: The manual production of comparable analytical reports typically requires up to two weeks of continuous effort by a human analyst. Our system reduces this time by approximately 50%, generating a preliminary report that serves as a foundation for further refinement. This significantly accelerates the reporting pipeline while maintaining analytical rigor.

2. *Scalability and Expanded Coverage*: Human resource constraints often limit the geographical or thematic scope that analysts can feasibly cover. By automating the initial stages of report generation, our system enables broader coverage across multiple regions or topics on a more frequent or regular basis, thereby enhancing the scalability of monitoring and analysis efforts without proportional increases in staffing.

3. *Resource Optimization and Cost Efficiency:* The system exclusively utilizes publicly available, open-access data sources, eliminating the need for costly proprietary datasets. This approach not only reduces operational expenditures but also makes the system particularly suitable for resource-constrained organizations, such as NGOs and humanitarian agencies.

4. *Transparency and Reproducibility:* By leveraging open data, the system ensures transparency in data sourcing and analytical processes. This facilitates reproducibility and fosters trust among stakeholders, including policymakers, researchers, and civil society actors, who can validate and build upon the generated reports.

5. *Rapid Situational Awareness:* In the event of a sudden conflict outbreak, the system can generate automated reports that offer stakeholders an immediate preliminary assessment of the situation. This rapid access to critical information enables timely decision-making and response, bridging the gap before official reports are published.

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 following 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 dynamic 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.

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.

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:

- Limited expert pool: Only two UNDP crisis analysts participated in the human evaluation. While this provided valuable domain insights, a broader and more diverse panel would improve robustness and generalizability of the results.
- Calibration stage for LLM-as-Judge: Although human evaluations served as a calibration step for the LLM-as-Judge outputs, this process remains preliminary. Future work should include more experts and formal inter-rater reliability measures (e.g., Cohen's Kappa) to better align automated judgments with expert standards.
- Lack of traditional baselines: We do not compare against human-written analyst reports or template-based summarization methods. Such baselines would contextualize the added value of RAG-powered LLMs.
- No direct assessment of decision impact: While our evaluation captures factuality, coherence, and usability, we do not measure how these reports influence real-world decision-making (e.g., resource allocation speed, quality of interventions).
- Potential LLM-as-Judge bias: Despite mitigating self-evaluation bias using a third independent model (Claude-2), LLM-based evaluations may still inherit systemic biases from their training data.
- Topic scope: Our evaluation focuses on country-level situation reports. Testing at sub-national or cross-regional levels could reveal new challenges in scaling the framework.

References

Hasan Md Tusfiqur Alam, Devansh Srivastav, Md Abdul Kadir, and Daniel Sonntag. 2024. Towards interpretable radiology report generation via concept bottlenecks using a multi-agentic rag. *arXiv preprint arXiv:2412.16086*.

Ayman Alhelbawy, M. Lattimer, Udo Kruschwitz, C. Fox, and Massimo Poesio. 2020. An nlp-powered human rights monitoring platform. *Expert Syst. Appl.*, 153:113365.

Muhammad Arslan, Hussam Ghanem, Saba Munawar, and Christophe Cruz. 2024. A survey on rag with llms. *Procedia Computer Science*, 246:3781–3790.

Dr. M. Mahalakshmi Assistant, Shardul Bharadwaj, and Aklanta Niraz. 2024. A real-time medical report analysis and ai-powered diagnosis: A cloud-based solution for improved patient care. *2024 Second International Conference on Advances in Information Technology (ICAIT)*, 1:1–6.

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.

M. Bernardi, Marta Cimitile, and Riccardo Pecori. 2024. Automatic job safety report generation using rag-based llms. *2024 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Jacob Devlin. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Ivan Iaroshev, Ramalingam Pillai, Leandro Vaglietti, and T. Hanne. 2024. Evaluating retrieval-augmented generation models for financial report question and answering. *Applied Sciences*.

Pankaj Joshi, Aditya Gupta, Pankaj Kumar, and Manas Sisodia. 2024. Robust multi model rag pipeline for documents containing text, table & images. *2024 3rd International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pages 993–999.

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.

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.

Nigel Markey, Ilyass El-Mansouri, Gaetan Rensonnet, Casper van Langen, and Christoph Meier. 2024. From rags to riches: Using large language models to write documents for clinical trials. *arXiv preprint arXiv:2402.16406*.

Hannes Mueller, Christopher Rauh, and Ben Seimon. 2024. Introducing a global dataset on conflict forecasts and news topics. *Data & Policy*, 6.

Poli Nemkova, S. Ubani, S. Polat, Nayeon Kim, and Rodney D. Nielsen. 2023. Detecting human rights violations on social media during russia-ukraine war. *ArXiv*, abs/2306.05370.

M. Ranjit, G. Ganapathy, R. Manuel, and T. Ganu. 2023. Retrieval augmented chest x-ray report generation using openai gpt models. *ArXiv*, abs/2305.03660.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. Veriscore: Evaluating the factuality of verifiable claims in long-form text generation. *arXiv preprint arXiv:2406.19276*.

Karthik Suresh, Neeltje Kackar, Luke Schleck, and C. Fanelli. 2024. Towards a rag-based summarization for the electron ion collider. *Journal of Instrumentation*.

Anusua Trivedi, Kate Keator, Michael Scholtens, Brandon Haigood, R. Dodhia, J. Ferres, Ria Sankar, and Avirishu Verma. 2020. How to handle armed conflict data in a real-world scenario? *Philosophy & Technology*, 34:111 – 123.

Yingding Wang, Simon Leutner, Michael Ingrisch, Christoph Klein, Christian Hinske, and Katharina Danhauser. 2024. Optimizing data extraction: Harnessing rag and llms for german medical documents. *Studies in health technology and informatics*, 316:949–950.

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 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 report "
              f"for humanitarian decision-makers. Use the evidence provided to craft "
              f"a clear, concise, and professional report 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

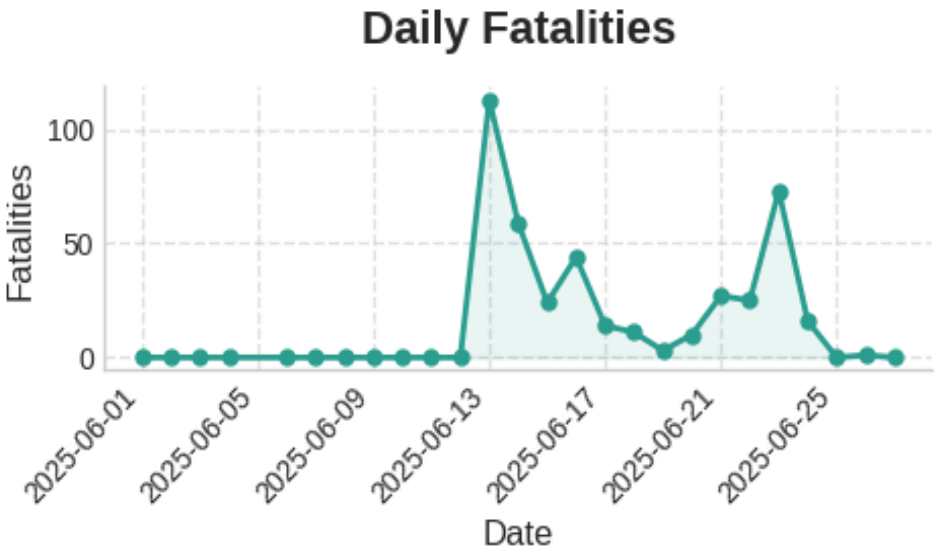
ACLED | Topics Most Covered

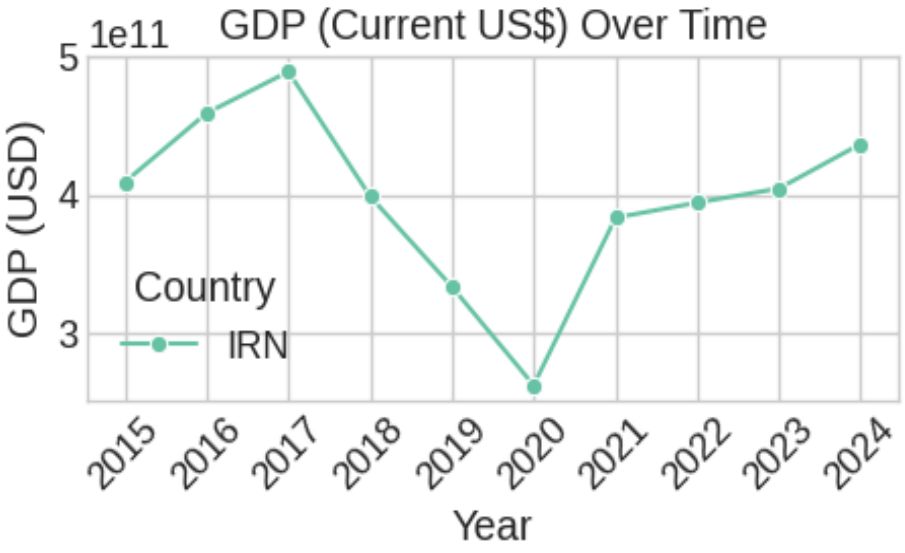
1. Military Operations
   ----→ Keywords: military, casualty, unknown, target, carry
2. Local Religious Activities
   ----→ Keywords: resident, prayer, local, friday, denounce
3. Air Defense Tactics
   ----→ Keywords: air, force, interception, intercept, shoot
4. Tehran District Unrest
   ----→ Keywords: 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

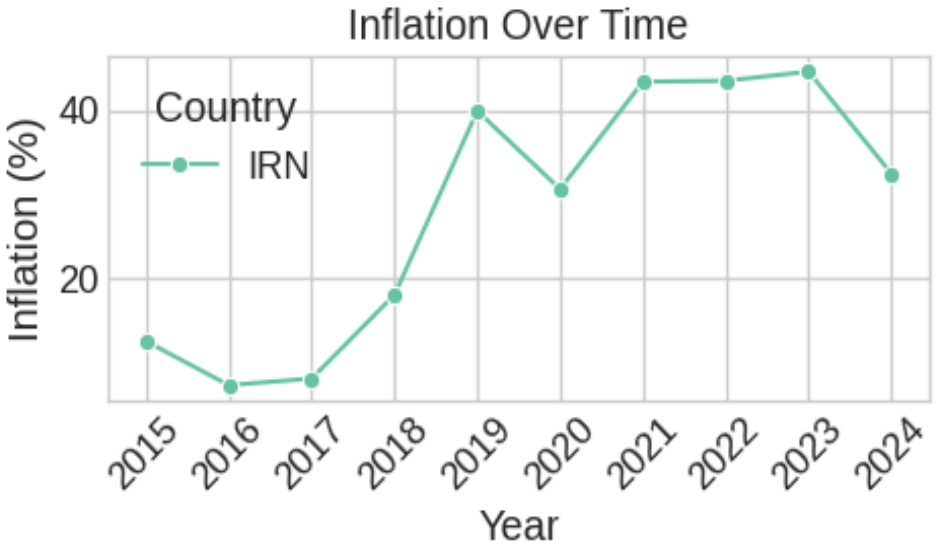
Graphs & Visualizations


Daily Fatalities
Fatalities
100
50
0
2025-06-01
2025-06-05
2025-06-09
2025-06-13
2025-06-17
2025-06-21
2025-06-25
Date

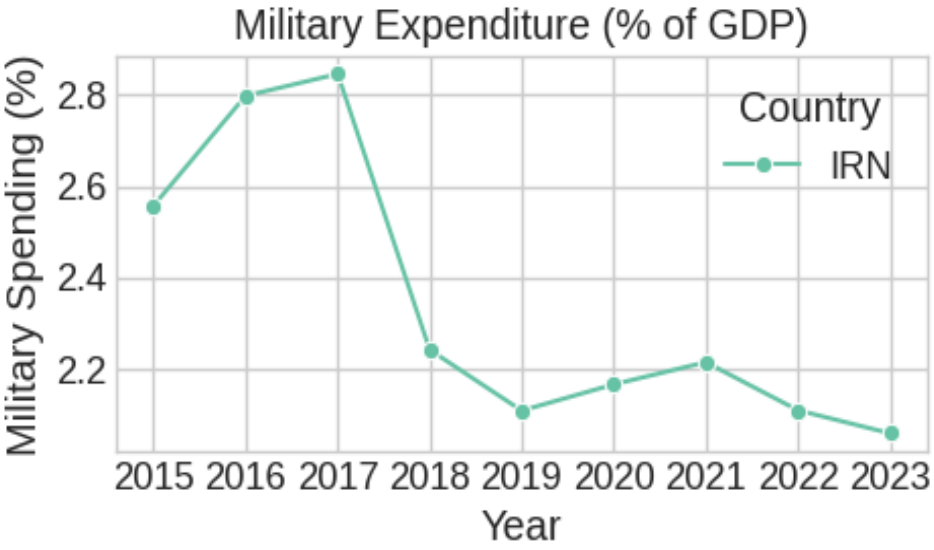

GDP (Current US$) Over Time
1e11
5
4
3
GDP (USD)
Country
IRN
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year

1030


GDP Growth Rate Over Time
Country
IRN
Growth (%)
5
0
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year


Inflation Over Time
Country
IRN
Inflation (%)
40
20
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year


Unemployment Rate Over Time
Unemployment (%)
12
11
10
9
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
Year
Country
IRN


Military Expenditure (% of GDP)
Military Spending (%)
2.8
2.6
2.4
2.2
2015 2016 2017 2018 2019 2020 2021 2022 2023
Year
Country
IRN