AI-Powered Query Triage Enhances Humanitarian Aid Delivery for IRC

IRC leverages AI and frameworks like DSPy to enhance humanitarian aid delivery by streamlining triage, resource allocation, and user support. This enables their staff to focus on complex, high-impact cases while automating routine processes by boosting both efficiency and the reach of their critical services.

April 2025

Pic Courtsey: IRC

Introduction

The International Rescue Committee (IRC) provides critical humanitarian aid and support to people affected by crises worldwide. This work is being conducted in the context of the IRC's Signpost project, an access to information program that connects people affected by crises with critical information and services.

A significant operational challenge is managing the high volume and diverse nature of user communications, ranging from urgent safety concerns to requests for information about services. Ensuring timely and appropriate responses is paramount, especially when individuals may be in distress or facing immediate threats.

This project is aimed to explore and validate the use of Artificial Intelligence (AI), specifically Language Models (LMs), to enhance IRC's capacity to triage and classify user queries efficiently and accurately, thereby improving response times and resource allocation. While the scope of the outcome of this work may extend beyond Signpost, the use case is currently

The Challenge: Prioritizing and Routing at Scale

IRC faces two primary challenges in managing user communications:

1. **High-Risk Content Identification:** The user queries can contain sensitive information or indicate immediate risks, such as self-harm ideation, Gender-Based Violence (GBV) disclosures, child safety concerns, or direct threats. Rapidly identifying these high-risk messages from a large volume of communications is crucial for enabling swift human intervention and ensuring user safety.

2. **Multi-Label Category Classification:** The users request assistance across a wide spectrum of needs. To connect them with the appropriate support or information efficiently, their queries must be classified into relevant service categories (e.g., health, shelter, legal aid, etc.). Manually sorting these queries is time-consuming and can delay aid delivery. The goal was to classify queries into 1-3 categories from a predefined list of 18 service areas.

Addressing these challenges with an automated, AI-driven approach can significantly improve IRC's operational efficiency, enhance user experience, and ultimately, save lives by ensuring the most critical cases receive immediate attention.

The Solution:
AI-Powered Triage and Classification with DSPy

Sahaj partnered with IRC to develop and evaluate an AI-driven solution leveraging modern Language Models and the DSPy framework. DSPy allows for a programmatic approach to prompting and optimizing LMs, moving beyond simple prompt engineering to build more robust and reliable AI systems.
The solution focused on two core AI capabilities:

High-Risk Identification Module: Its objective is to automatically detect if a user's input messages fall into predefined high-risk or sensitive categories, outputting a simple boolean (True/False). It is treated as a boolean classification task. Various LMs, from large API-based models (GPT, Gemini, Claude series) to smaller, locally deployable models (e.g., Mistral 7B, Qwen 2.5 7B), were evaluated using DSPy's Predict module. Synthetic data, including versions translated into Arabic and French (languages relevant to IRC's operations in Libya, the initial test case), was used for evaluation. The DSPy framework allowed for consistent evaluation across diverse models with no code changes, defining the task via a clear "Signature" (Input: query -> Output: True/False).

Multi-Label Category Classification Module: It classifies user inputs into 1 to 3 relevant service categories from IRC's list of 18 service areas. It is implemented as a multi-label classification task using DSPy. The system was designed to output a list of relevant category names. Cost-effective large API-based models (e.g., GPT-4o mini, Gemini 2.5 Flash, Claude 3 Haiku) and selected local models were tested. DSPy's optimization capabilities (e.g., BootstrapFewShotWithRandomSearch, MIPROv2) were explored to refine prompts and improve performance. Its accurate classification enables faster routing to the correct IRC departments or specialized RAG (Retrieval-Augmented Generation) chatbots. These AI assistants can handle common questions within specific domains (e.g., education, non-food items), freeing up human staff for complex cases.

Key Results:
High Accuracy and Feasibility Demonstrated

The experiments, conducted using synthetic data, yielded highly promising results:

High-Risk Content Identification:

- Achieved ~99% accuracy in identifying high-risk content.
- Notably, smaller, cost-effective local models like Mistral 7B Q5 and Qwen 2.5 7B performed exceptionally well, demonstrating the feasibility of deploying reliable high-risk detection even in resource-constrained environments.
- Performance remained high on synthetic data translated into Arabic and French.

Top Performing Models for High-Risk Identification (Mean Score across Languages):

Model	English	Arabic	French	Mean
mistral-small3.1	99.38	100.00	99.38	99.59
qwen2.5:7b	99.38	100.00	99.38	99.59
llama-4-maverick	97.52	100.00	98.76	98.76
phi4	99.38	98.14	97.52	98.35
gpt-4.1-nano	97.52	97.52	98.14	97.73

Multi-Label Category Classification:

- Achieved F1 scores in the 85-90% range using cost-effective large API-based models like GPT-4.1 mini, Gemini 2.5 Flash, and Claude 3 Haiku.
- DSPy's optimization techniques (MIPROv2, BootstrapFewShotWithRandomSearch) showed potential in improving scores, particularly for smaller models.

Top Performing Models/Optimizations for Multi-Label Classification (Mean Score across Languages):

Model: Optimization	English	Arabic	French	Mean
gemini-2.5-pro:miprov2	90.61	89.63	91.83	90.69
gemini-2.5-flash:miprov2	89.88	89.67	89.97	89.84
gpt-4.1:miprov2	89.99	89.22	89.39	89.53
gpt-4.1-mini:miprov2	89.40	88.21	88.95	88.85
mistral-small3.1:bfswrs	86.29	88.54	88.80	88.54
claude-3-5-haiku:None	88.48	87.26	88.09	87.94

These results strongly indicate the viability of using LMs for these critical tasks within IRC's operational context.

Technical Innovation:
A Flexible and Programmatic AI Approach

The project's success was underpinned by a modern, flexible technology stack and a principled approach to AI development:

- **DSPy Framework:** The core of the solution. DSPy enabled:
  - **Systematic Programming of LMs:** Defining tasks through clear input/output signatures.
  - **Modular Design:** Building reusable components (Predict, ChainOfThought).
  - **Automated Optimization:** Employing optimizers to refine prompts and few-shot examples, reducing manual effort and improving performance.

- **Model Agnosticism:** Easily swapping and evaluating different LMs (local and API-based) with the same codebase.

- **Ollama:** Facilitated the use of local LMs (e.g., Mistral, Qwen, Phi-4) on standard hardware, crucial for exploring cost-effective and potentially offline deployment scenarios.
- **Marimo:** Interactive Python notebooks were used for rapid experimentation, visualization, and development of the evaluation pipelines.
- **Data Handling:** Polars for efficient data manipulation and SQLite for storing and managing evaluation results.
- **Synthetic Data Generation & Translation:** Due to the initial absence of labeled IRC data, synthetic datasets were carefully generated and translated (using DeepL) to simulate multilingual environments (English, French, and Arabic for the Libya context).

Benefits for IRC as a Non-Profit

This AI-driven approach offers significant advantages for a humanitarian organization like IRC:

- **Leveraging Off-the-Shelf Models:** Access to powerful pre-trained models without the need for extensive in-house ML engineering teams to train models from scratch.

- **Systematic Prompt Optimization:** DSPy's optimizers automate the complex and time-consuming process of prompt engineering.

- **Future-Proofing:** The ability to easily integrate and evaluate new and improved LMs as they become available.

- **Cost-Effectiveness:** The framework supports evaluation of various models, allowing IRC to balance performance with cost, and identify efficient local models for certain tasks.

- **Empowering Existing Teams:** Enables IRC to adopt cutting-edge AI without a massive investment in specialized AI talent, allowing them to focus on their core mission.

Challenges and What's Next

While the project demonstrated strong potential, several challenges and next steps are crucial for real-world deployment:

- **Validation with Real Data:** The highest priority is to acquire and label real, anonymized IRC user queries. Validating the system's performance on this real-world data is essential to confirm the findings from synthetic data and understand nuances like local dialects.
- **Refinement of Categories and High-Risk Definitions:** Real data will inform the refinement of the 18 service categories (e.g., addressing overlaps, clarity, etc.) and ensure high-risk definitions are comprehensive.

- **Expanded Language and Dialect Support:** While initial tests included Arabic and French (for Libya), IRC operates globally. Future work should involve testing and adapting the models for other languages and specific regional dialects.
- **MLOps and Continuous Monitoring:** Establishing a pipeline for ongoing monitoring, evaluation, and retraining of models in production is vital to maintain performance as data patterns evolve.
- **Iterative Deployment:** A phased rollout will begin, starting with close monitoring and human oversight, to build trust and refine the system based on operational feedback.

Conclusion:
Augmenting Humanitarian Response with AI

This project successfully demonstrates that AI, specifically through the programmatic use of Language Models with frameworks like DSPy, can provide powerful tools to enhance the efficiency and effectiveness of humanitarian aid delivery.

For organizations like IRC, AI offers a transformative way to:

- **Accelerate Triage:** Quickly identify and escalate critical situations, potentially saving lives.
- **Improve Resource Allocation:** Efficiently route queries to the appropriate human staff or automated RAG chatbots.
- **Enhance User Experience:** Provide faster, more relevant responses to individuals seeking help.

By automating routine classification and identification tasks, AI can augment human capabilities, allowing IRC staff to focus their expertise on complex cases requiring nuanced human judgment and compassionate support. This project marks a significant step towards integrating responsible AI to amplify the impact of humanitarian efforts globally.

> "Sahaj's approach made advanced AI accessible for our operations without specialized technical teams. By leveraging deep expertise in data science and applied AI, they've set us on a pathway to our goal of 10x improvement in our field response capabilities."

Andre Heller
Signpost Director, International Rescue Committee