Using Large Language Models for Humanitarian Frontline Negotiation: Opportunities and Considerations

Zilin Ma[*†‡], Susannah (Cheng) Su[*‡], Nathan Zhao[*‡], Linn Bieske[‡], Blake Bullwinkel[§], Yanyi Zhang[‡], Sophia (Yanrui) Yang[‡], Ziqing Luo[‡], Siyao Li[‡], Gekai Liao[‡], Boxiang Wang[‡], Jinglun Gao[‡], Zihan Wen[‡], Claude Bruderlein[¶], Weiwei Pan[‡]

[*] Equal contribution [†] Correspondence to: <zilinma@g.harvard.edu> [‡] Harvard School of Engineering and Applied Sciences [§] Microsoft [¶] Harvard T. H. Chan School of Public Health

Understanding Frontline Negotiation

Efficient and Accurate Information Synthesis: Frontline humanitarian negotiations in conflict zones facilitate aid delivery by synthesizing diverse, unstructured information under time pressure, requiring rapid navigation of conflicting perspectives (GISF, 2020).

Templates for Synthesizing Information: Negotiators developed templates to assist in manual information processing (e.g., the Island of Agreement (IoA), Iceberg and Common Shared Space (CSS), and Stakeholder Mapping (ShM)).

Main Contributions

Contribution 1: Evaluated the quality and stability of ChatGPT-based negotiation tools; verified that LLMs (GPT-4) can quickly generate high-quality negotiation summaries.

Contribution 2: Identified key uses (e.g., automated context analysis and ideation) for LLMs in frontline negotiations through in-depth interviews.

Contribution 3: Revealed pertinent ethical and practical challenges to using LLMs in frontline negotiation including confidentiality, bias, and overreliance.

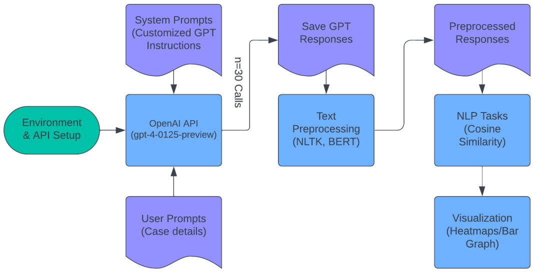
Method

Tool Development and Iteration: Using multiple case studies and feedback from frontline negotiators, we developed three LLM tools to populate the IoA, Iceberg/CSS, and ShM negotiation templates.

Evaluating the Stability of the LLM Responses: We assessed consistency using cosine similarity to ensure low variance and minimal hallucinations.

Benchmarking LLM Responses Against Practitioner Responses: We benchmarked LLM-generated responses against experienced practitioner responses, using BERT embeddings to measure semantic similarity.

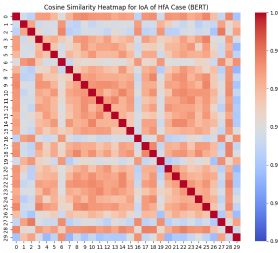
Human-Centered Benchmarking and Evaluation of Tools We conducted semi-structured interviews with 13 frontline negotiators to demo the LLM-based tools and understand the primary opportunities and risks associated with using LLMs in frontline negotiation.


Environment & API Setup
System Prompts (Customized GPT Instructions)
OpenAI API (gpt-4-0125-preview)
User Prompts (Case details)
n=30 Calls
Save GPT Responses
Text Preprocessing (NLTK, BERT)
Preprocessed Responses
NLP Tasks (Cosine Similarity)
Visualization (Heatmaps/Bar Graph)

Quantitative Results

Stability of Responses: We observed high consistency in LLM responses, with cosine similarity scores in the Iceberg/CSS and IoA frameworks ranging from 0.9474 to 0.9919. The ShM framework showed higher variability.

Benchmarking with Practitioners: In benchmarking tests, the average cosine similarity between ChatGPT and practitioner responses was 0.93 for the IoA framework and 0.92 for Iceberg/CSS, indicating high alignment with experts.



Interview Results

Opportunities: Accelerating context analysis and enhancing creativity through ideation.

Concerns: Confidentiality, Western bias, the influence of public and mandator opinions, accuracy, overreliance, and the need to maintain human involvement in negotiations.

Discussion

Our quantitative results and interviews with experienced practitioners highlight the potential for LLMs to enhance humanitarian negotiations.

However, safe deployment hinges on mitigating privacy concerns, bias in LLMs, overreliance, and other risks.

Any production-ready system would require a more comprehensive evaluation, including red teaming. Further, training programs would be needed to ensure that negotiators can leverage these technologies effectively and responsibly.

Future Work

- Enhance privacy through development and assessment of open-source LLMs for local deployment.
- Implement training and support systems to help practitioners assess LLM outputs.
- Develop interfaces to better integrate LLM-tools with human-led negotiations.
- Red teaming to identify additional risks.