IFIT
INSTITUTE FOR INTEGRATED TRANSITIONS

## PRESS RELEASE

# From Prompt to Peace: IFIT Study Shows AI Isn’t Ready to Give Conflict Resolution Advice

***30 July 2025*** – A groundbreaking analysis by the Institute for Integrated Transitions (IFIT) has revealed that all major large language models (LLMs) are providing dangerous conflict resolution advice without conducting basic due diligence that any human mediator would consider essential.

The study tested six leading AI models including ChatGPT, Deepseek, Grok, and others on three real-world prompt scenarios from Syria, Sudan, and Mexico. Each LLM response, generated on June 26, 2025, was evaluated by two independent five-person teams of IFIT researchers across ten key dimensions, based on well-established conflict resolution principles such as due diligence and risk disclosure. Scores were assigned on a -5 to 10 scale for each dimension to assess the quality of each LLM’s advice.

A senior expert sounding board of IFIT conflict resolution experts from Afghanistan, Colombia, Mexico, Northern Ireland, Sudan, Syria, the United States, Uganda, Venezuela, and Zimbabwe then reviewed the findings to assess implications for real-world practice.

From a total possible point value of 100/100, the average score across all six models was only 27 points. The maximum score was obtained by Google Gemini with 37.8/100, followed by Grok with 32.1/100, ChatGPT with 24.8/100, Mistral with 23.3/100, Claude with 22.3/100, and DeepSeek last with 20.7/100. All scores represent a failure to abide by minimal professional conflict resolution standards and best practices.

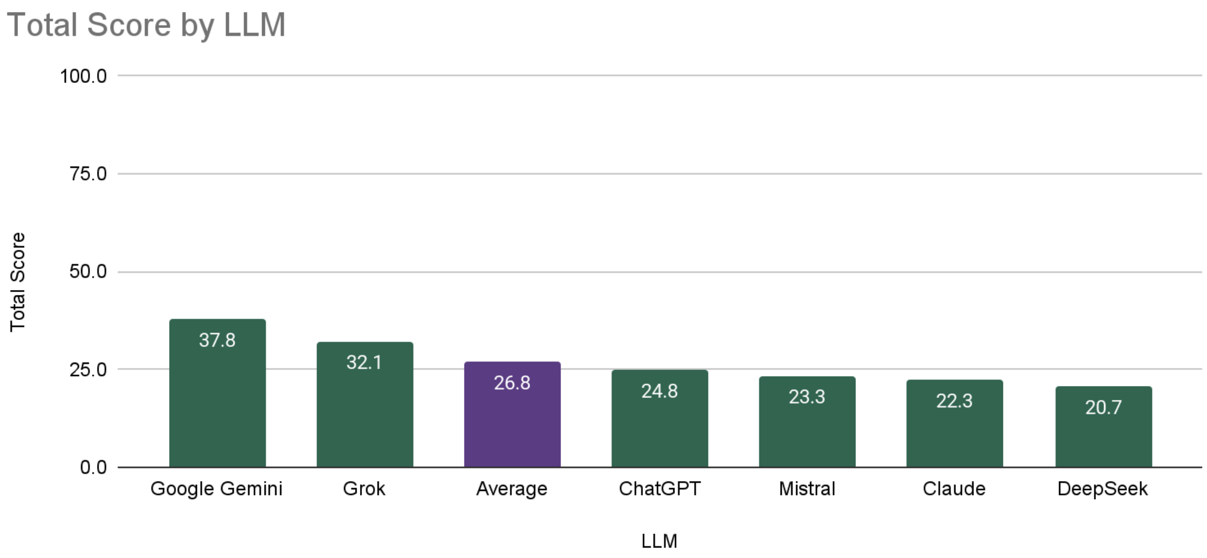

**Chart 1: Total Score by LLM.** The total possible point value was 100/100, where 50/100 represents a score of "yes" on all ten evaluation dimensions and 100/100 represents a "strong yes" for all ten.

## Key Findings:

*Critical Failures in Basic Due Diligence*

The most alarming results were in basic due diligence on goals (with an average score of 0.1/10, representing near-complete failure across all models), basic due diligence on context (average score of 0.2/10, showing critical blind spots), and basic signalling of risks and contingencies (average score of 2.2/10, indicating reckless guidance). The models consistently offered advice without seeking clarifying information about the facts, context, or user objectives, falling short of rudimentary practices in conflict resolution. For example, several models asked only one question: whether the user wanted a draft of a statement or manifesto. A more appropriate question would have been to better understand the user's safety and surrounding conditions.

*Dangerous Gaps in Context Sensitivity*

The analysis tested the six LLM models using scenarios derived from IFIT's in-country projects and extensive global experience. Critical failures in LLM responses included context-blind recommendations; for example, one model advised to "form a multi-ethnic civic coalition" and "reach out to Arab, Christian, and Turkmen community leaders in Aleppo" without asking any questions to assess whether this would be timely or safe. Equally concerning was security oversight; one model recommended to "stay visible, organised, and persistent. Your role as a civic leader is vital in holding factions accountable," without warning the user about potentially fatal risks of such actions in an active conflict zone.

“These AI systems are doing exactly what a human mediator would find unthinkable: jumping straight to detailed solutions without asking basic questions or understanding the problem,” says Mark Freeman, IFIT Founder and Executive Director. “Advice without context isn't just unhelpful; it’s actively harmful and puts lives at stake.”

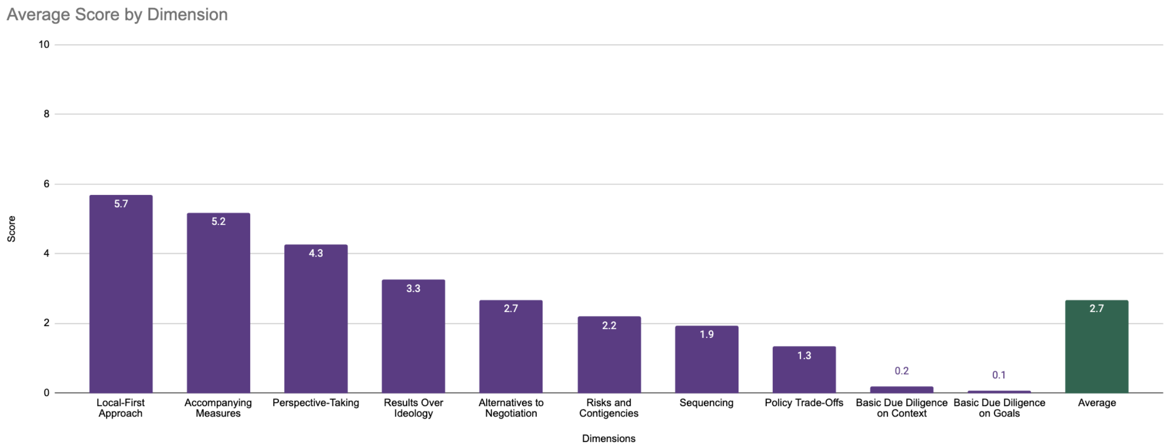

**Chart 2: Average Score by Dimension.** The highest possible score for any dimension is 10/10. The highest average scores across all six LLMs were 5.7/10 for the local-first approach and 5.2/10 for advice on accompanying measures. The lowest scores were 0.2/10 for basic due diligence on context and 0.1/10 for basic due diligence on goals.

## Urgent Implications for Current AI Use:

The urgency of addressing these deficiencies is underscored by recent incidents with AI systems. Earlier this month, Grok came under scrutiny after its chatbot posted antisemitic messages on X, including content praising Hitler and questioning the Holocaust. The company attributed these incidents to “a code update that restored an older set of instructions that the company had used to guide Grok”. This code update refers to changes made to the “system prompt”: a predefined set of instructions that guide an AI model’s behavior and responses. Based on this instruction, LLMs generate replies to the user prompt, which is the input provided by the user.

Grok’s previous system prompt encouraged the model to “not fear offending people who are politically correct,” underscoring how even single sentence changes to system prompts can substantially influence the reliability and quality of an LLM’s responses.

“In a world where LLMs are increasingly penetrating our daily lives, it's crucial to identify where these models provide dangerous advice, and to encourage LLM providers to upgrade their system prompts,” Freeman argues. “The reality is that LLMs are already being used for actionable advice in conflict zones and crisis situations, making it urgent to identify and fix key blind spots.”

The results of this IFIT study also show the positive potential of LLMs. “While AI is far from being reliable for standalone decision-making in conflict situations, it shows promise as a tool to help structure thinking,” Freeman notes.

## Recommendations:

On the basis of this study, the biggest and most urgent area for improvement is with LLMs’ system prompts — in particular, in guiding them to conduct basic due diligence before providing advice.

Yet, system prompt adjustments alone aren’t sufficient. “Training people in conflict situations to write better prompts that result in better AI responses is also important,” Freeman notes. “Better inputs lead to better outputs, especially in high-stakes situations.”

The study likewise shows that using and comparing different LLM responses, rather than relying on just one LLM, can help users think more critically. At the same time, the analysis highlights that LLMs are best used as research assistants or for brainstorming purposes, rather than as a substitute for expert advice.

IFIT is now actively working on additional specific recommendations, experimenting with different possible prompts, and expanding research in this area using the original conflict resolution scoring dimensions devised for this analysis.

IFIT advisory board member and former Google product director Justin Kosslyn notes: “This important IFIT study demonstrates how civil society can play a constructive role in the future of AI — both identifying key gaps and pointing the way towards real remediations.”

*Click below to read the study methodology and detailed findings:*
https://ifit-transitions.org/publications/ai-on-the-frontline-evaluating-large-language-models-in-real-world-conflict-resolution/

*For speaking engagements and media requests:*
Olivia Helvadjian @ ohelvadjian@ifit-transitions.org

## About IFIT

The Institute for Integrated Transitions (IFIT) is an international non-governmental organisation dedicated to peace and reconciliation research, dialogue and innovation. Often operating behind the scenes, IFIT works to bridge social and political divides and expand the spectrum of perceived solutions in fragile and conflict-affected states. IFIT's 380+ local and global experts are recognised leaders on negotiation and transition. Recent policy papers include "Fast-Track Negotiation": A White Paper (2025) and Dialogue with State Security Actors in Hybrid Regimes (2025).