Reason Impossible: Can LLMs Forecast Wartime Returns?

Farzana Yasmin Ahmad, Kazi Noshin, Zakaria Mehrab
University of Virginia

Introduction

Motivation:

- The 2022 Russian Invasion of Ukraine.
- Millions displaced internally and externally.
- Humanitarian Response necessitates understanding.
- Return Migration has not been well explored in literature.
- Potential of LLM as policy-making tool.

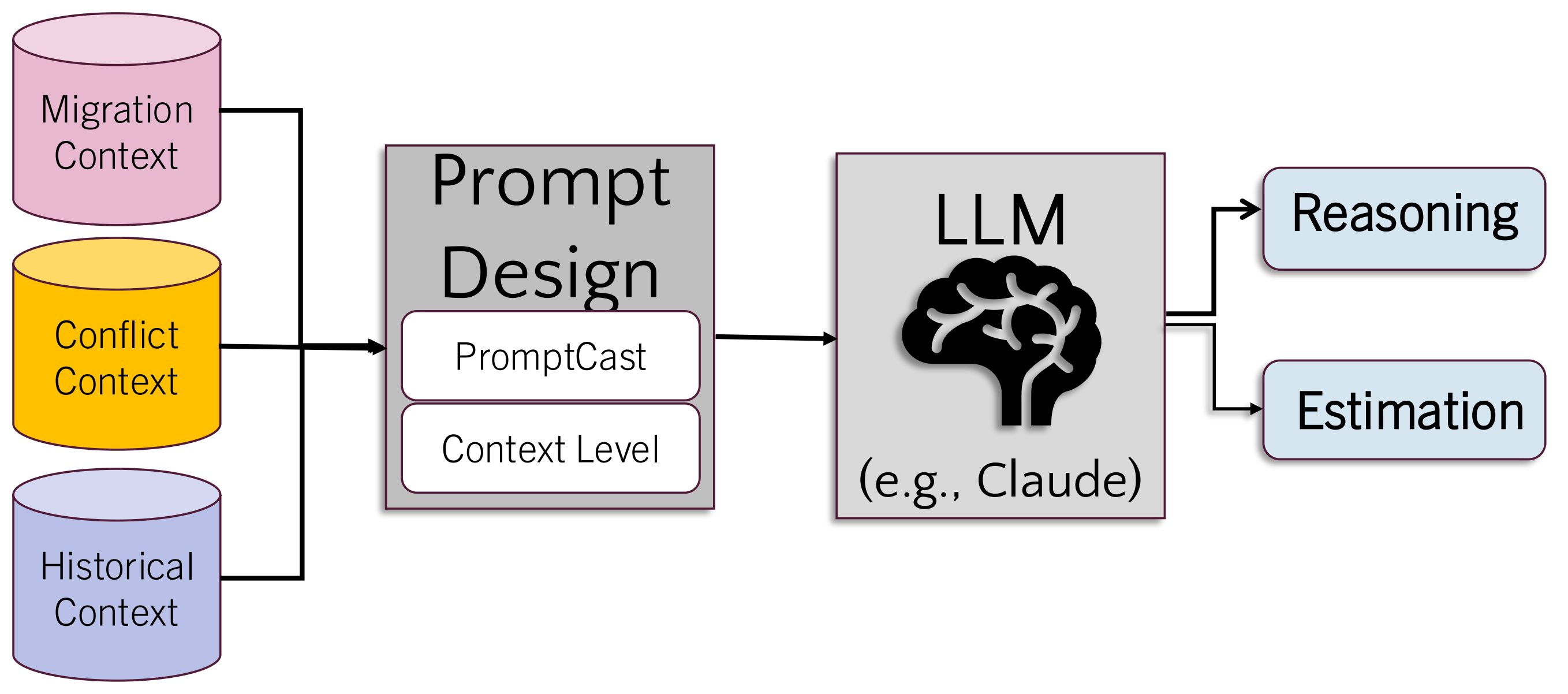
Research Questions:

- How well can LLMs forecast crisis induced temporal return migration patterns using real-world data from the Ukraine–Russia war? **(RQ1)**
- How do LLMs reason about their forecast? **(RQ2)**
- How robust are the forecasting performance and reasoning processes? **(RQ3)**

Contributions

- First systematic evaluation of forecasting wartime return migration by LLMs.
- Comparative taxonomy and visualization of LLM-generated underlying models.

Methods



Migration Context

Conflict Context

Historical Context

Prompt Design

PromptCast

Context Level

LLM

(e.g., Claude)

Reasoning

Estimation

Figure: Systematic overview of using LLM to forecast wartime return estimation with various contexts. Prompts are provided using strategy outlined in PromptCast [7].

Datasets, Models, Metrics

HDX [3]
Migration Context

ACLED [5]
Conflict Context

UNHCR [6]
Historical Context

ChatGPT 5 [4]

Gemini 2.5 Flash [2]

Claude Sonnet 4.5 [1]

NRMSE
Captures accuracy

PCC
Captures correlation

NACRPS
Captures uncertainty

Quantitative Evaluation (RQ1 & RQ3): Ground Truth Validation

Slovakia

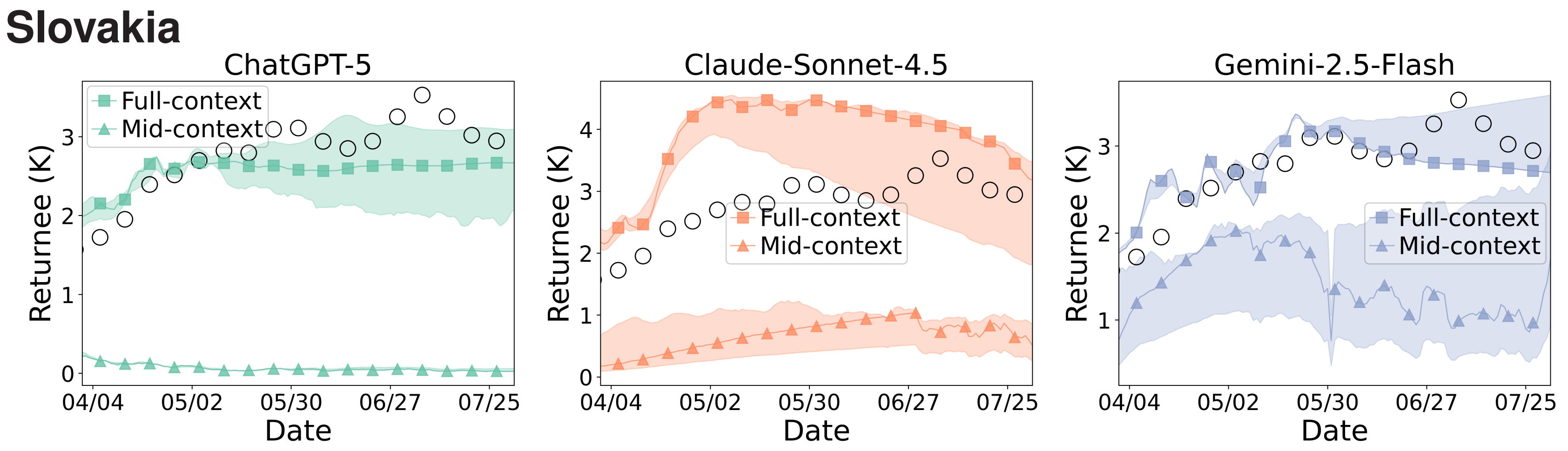


Figure: LLM return estimation for Slovakia. Gemini-2.5 is the best in terms of model accuracy and uncertainty, whereas GPT-5 is the best in capturing trend.

Romania

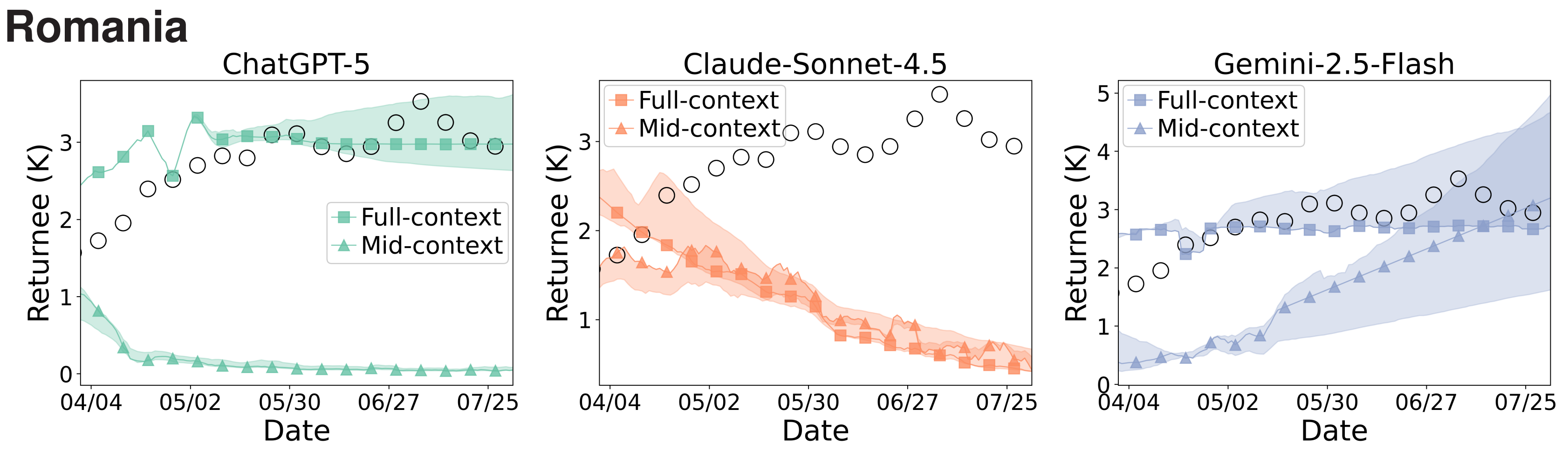
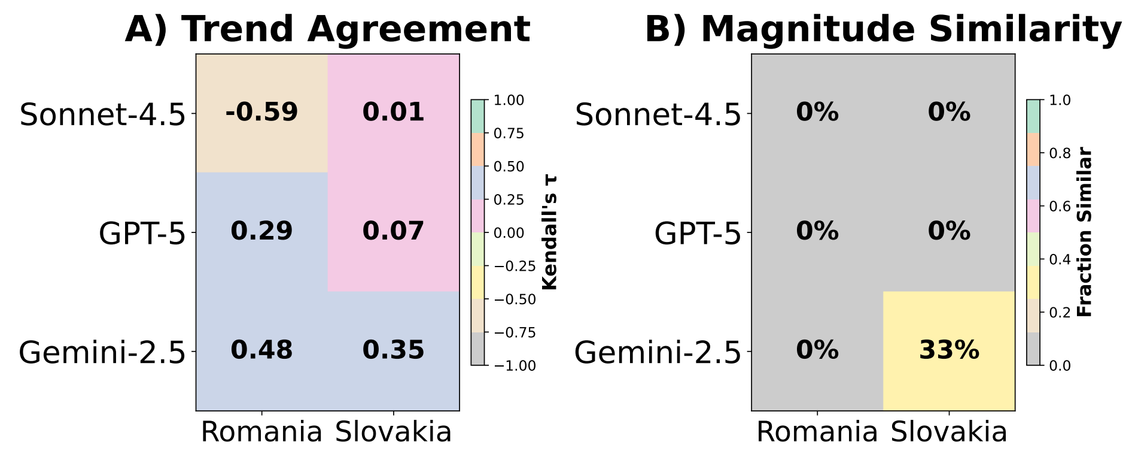


Figure: LLM return estimation for Romania. GPT-5 and Gemini-2.5 performs comparably. However, the performance of all models deteriorate significantly compared to Slovakia.

Metric	LLM	Slovakia	Romania
NRMSE ↓	GPT-5	0.212	**0.33**
	Claude Sonnet 4.5	0.618	0.573
	Gemini 2.5 Flash	**0.178**	0.36
PCC ↑	GPT-5	**0.8**	0.37
	Claude Sonnet 4.5	0.733	-0.76
	Gemini 2.5 Flash	0.68	**0.39**
NACRPS ↓	GPT-5	0.092	**0.19**
	Claude Sonnet 4.5	0.166	0.37
	Gemini 2.5 Flash	**0.045**	0.2

Table: Quantitative performance of LLM with *full context* in estimating return migration from Romania and Slovakia, compared with border guard observation data.



A) Trend Agreement

	Romania	Slovakia
Sonnet-4.5	-0.59	0.01
GPT-5	0.29	0.07
Gemini-2.5	0.48	0.35

B) Magnitude Similarity

	Romania	Slovakia
Sonnet-4.5	0%	0%
GPT-5	0%	0%
Gemini-2.5	0%	33%

Figure: Fraction of runs where models have significant agreement (p-value< 0.05) with respect to ground truth data. **Left figure** shows agreement for trend estimation whereas **right figure** shows magnitude similarity.

Quantitative Evaluation (RQ3): Model vs Model Comparison

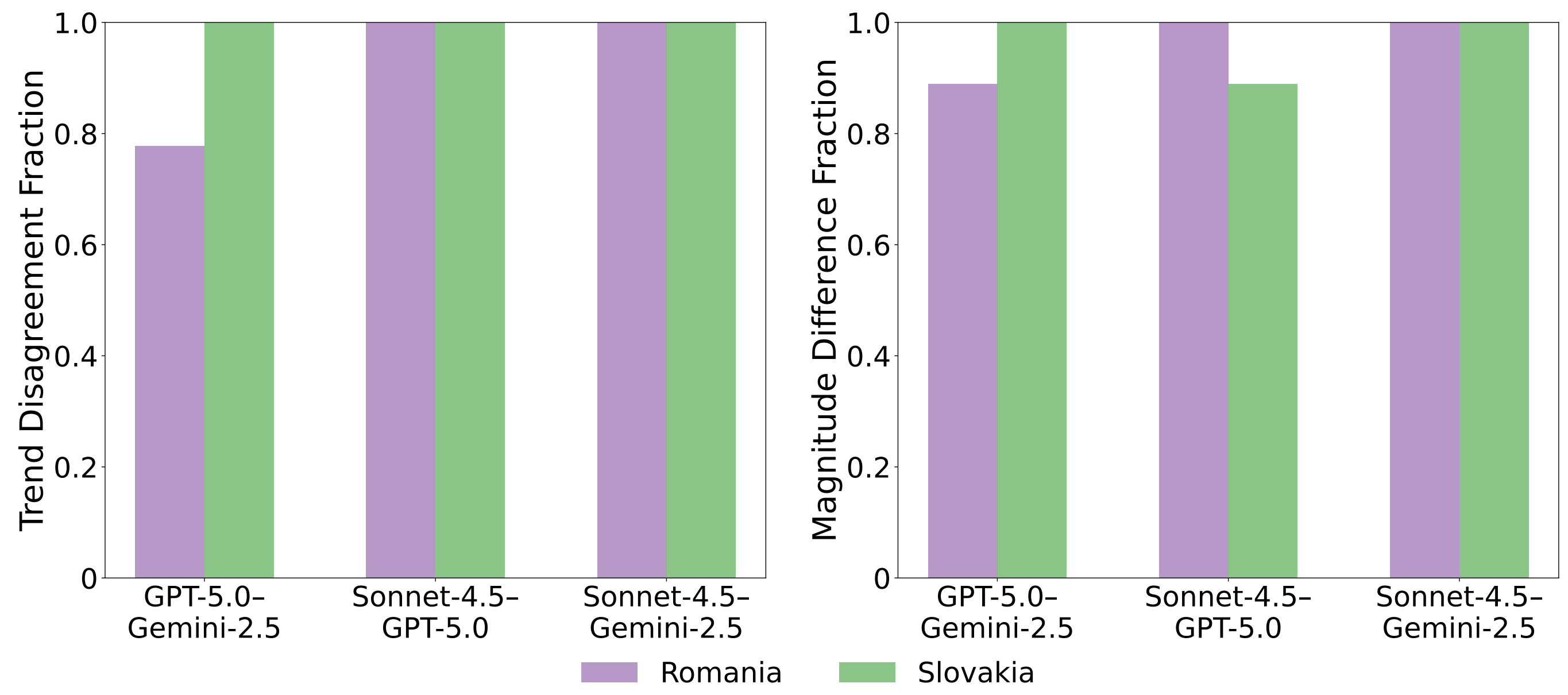
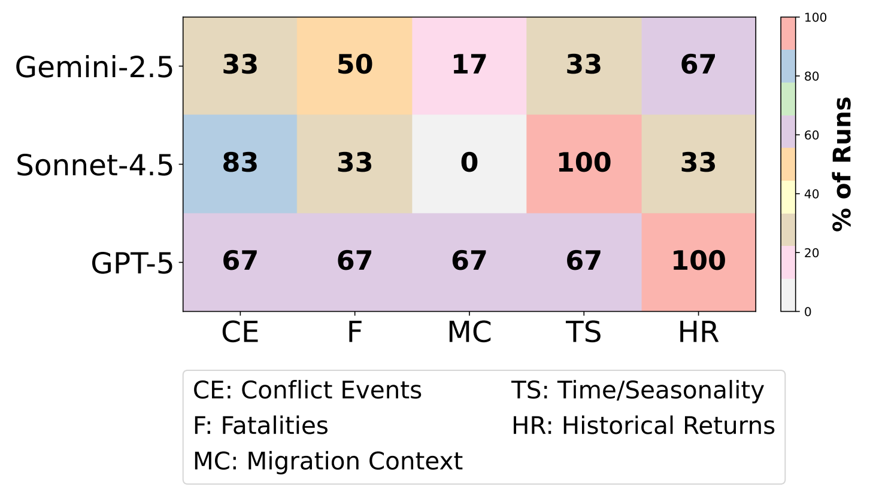


Figure: Fraction of runs where significant disagreement (p-value< 0.05) occurs in estimation across model pairs. **Left figure** shows disagreement for trend estimation whereas **right figure** shows difference across estimated magnitude.

Pairwise model estimation difference is statistically significant.

Qualitative Evaluation (RQ2 & RQ3): Feature & Model Selection



	CE	F	MC	TS	HR
Gemini-2.5	33	50	17	33	67
Sonnet-4.5	83	33	0	100	33
GPT-5	67	67	67	67	100

CE: Conflict Events
F: Fatalities
MC: Migration Context
TS: Time/Seasonality
HR: Historical Returns

Figure: Feature usage patterns.

- ► GPT-5 consistently utilizes all features.
- ► Claude Sonnet-4.5 shows a strong preference for time/seasonality and conflict context, and completely ignores migration context.
- ► Gemini-2.5 does not show any consistent pattern in feature selections.

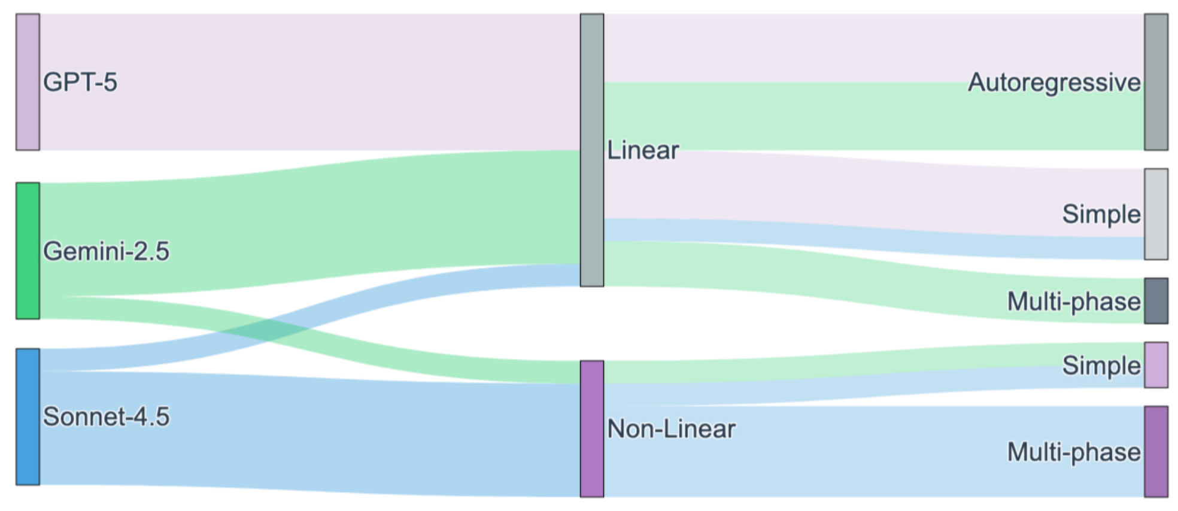


Figure: Model type selection.

- ► GPT-5 consistently uses linear models with autoregressive approaches.
- ► Claude Sonnet-4.5 favors non-linear, multi-phase modeling.
- ► Gemini-2.5 heavily relies on the linear models.

Qualitative Evaluation (RQ2): Reasoning

GPT-5 and Gemini-2.5 focuses on **simplicity** whereas Claude Sonnet-4.5 employs **multi-phase approach**, assuming return composition changes over time.

Gemini-2.5 and Claude Sonnet-4.5 both view **conflict context** as an important driver while GPT-5 treats this as a correlational factor rather than a causal one.

GPT-5 acknowledges **limitations** explicitly. However, Claude Sonnet-4.5 and Gemini-2.5 lacks transparency in methodological choices.

Discussion

GPT-5 performance is reasonably better than the other two LLMs.

From quantitative evaluation, GPT-5 and Gemini-2.5 shows competitive ground truth agreement across all three metrics.

From qualitative analysis, GPT-5 takes all features into account, focuses on simplicity and transparency more than Gemini-2.5 and Sonnet-4.5.

Overall, LLMs are yet not reliable in such crisis situation without human-in-the-loop.

References

[1] Anthropic. Claude. Large language model, 2025. Version: Claude Sonnet 4.5, Accessed: November 3, 2025.

[2] Google. Gemini Advanced. https://gemini.google.com/, 2025. [Large language model; Accessed: 3 Nov 2025].

[3] HUMDATA. The Humanitarian Data Exchange . https://data.humdata.org/. [Online; accessed December 2, 2022].

[4] OpenAI. Chatgpt (gpt-5). https://chat.openai.com/, 2025. Large language model.

[5] Raleigh, C., et al. Introducing ACLED: An armed conflict location and event dataset. *Journal of peace research 47*, 5 (2010), 651–660.

[6] UNHCR. Ukraine Refugee Situation, 2022.

[7] Xue, H., and Salim, F. D. Promptcast: A new prompt-based learning paradigm for time series forecasting. *IEEE Transactions on Knowledge and Data Engineering 36*, 11 (2023), 6851–6864.