A Comparative Study of Large Language Models' Performances in Estimating Flood Depths from Textual Reports

HIRES Interns: Brandom Guzman-Lora, Jail Alexander Flores; **Mentors:** Chaw Nandar, Seon-Ho Kim; **Faculty Advisors/Researchers:** Naresh Devineni

Address/Affiliation - CUNY CREST High School Initiative in Remote Sensing of Earth System Engineering and Sciences (HIRES)
The City College of New York, NY 10031

Background

- New York City has experienced severe flooding in recent years, including Hurricane Ida (2021), which caused major inundation, infrastructure damage, and fatalities.
- Flood depth data is essential for impact assessment, infrastructure design, and mitigation planning.
- Traditional data collection methods (e.g., sensors, surveys) are costly, time-consuming, and offer limited spatial coverage.
- Textual flood reports provide a promising alternative, offering long-term and reliable crowd-sourced descriptions of flood impacts.
- Large Language Models (LLMs) can interpret contextual clues to estimate quantitative flood depths from textual reports.
- A recent study by Kim et al. (2025) showed that using ChatGPT for this task achieved reasonable accuracy.
- This research builds on that work by testing multiple LLMs to improve LLM-based flood depth estimation.

Method

- **Collecting reports** – Gathered written accounts of flood events, such as descriptions of locations, impacts, and water levels.
- **Prompting** – Asked each model the same clear question, requesting only the flood depth in inches.
- **Estimating** – Ran six different language models, with each model performing three independent simulations for every event.
- **Validation** – Evaluated the model's estimations against expert-provided depths using the Mean Absolute Error (MAE).

Table 1. Selected Large Language Models for Flood Depth Estimation.

Gemini	DeepSeek	Chat GPT	Perplexity
2.5 Pro (Google AI Pro) Reasoning, math & code	**DeepSeek R1** (DeepThink) Reasoning, math & code	**GPT-4o** General purpose	**Perplexity** General purpose combined with real-time web search
	DeepSeek V3 High-end reasoning & multimodal understanding	**GPT-o4-mini-high** High reasoning & coding	

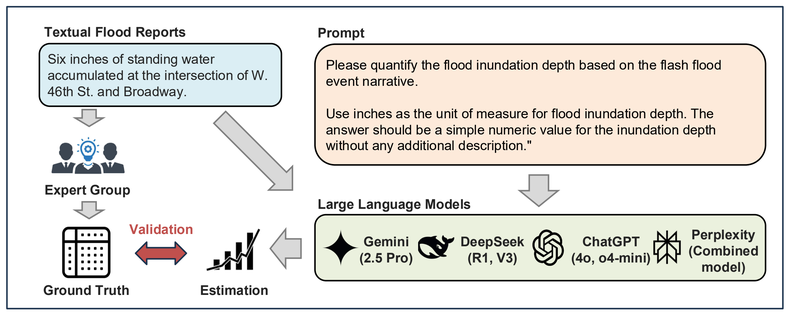

Figure 1. LLM-Based Flood Depth Estimation Workflow.

Study Area and Data

- In Figure 2, the smaller map illustrates NYC's high flood vulnerability, driven by dense population.
- The larger map plots 42 recorded flood events from 2004 to 2022 in NOAA's storm database, revealing spatial patterns of past flooding in NYC (NOAA, n.d.).

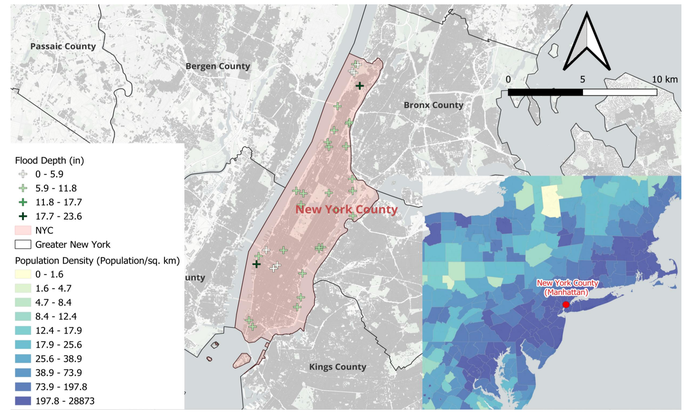

Figure 2. Flood Event Locations in NYC Used for LLM Flood Depth Estimation.

Results

- **Can LLMs accurately represent flood depth ranges?** – They typically overestimated the flood depths compared to ground truths.

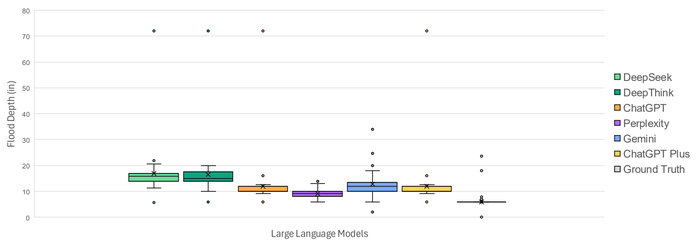

Figure 3. Flood Depth Estimates by Different Large Language Models.

- **Did the advanced models perform better?** – Advanced models (high-end or paid versions) did not perform better than the standard Perplexity model.

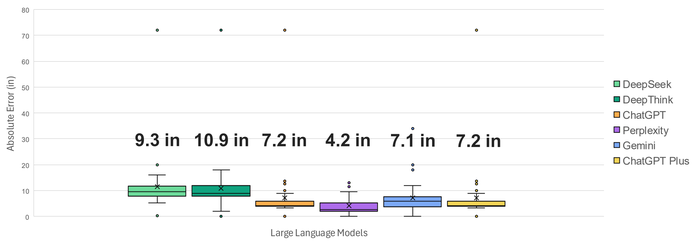

Figure 4. Variation in Absolute Error Across Large Language Models.

- **Are there patterns in the error distribution across models?** – Across all error range, Perplexity > ChatGPT & Gemini > DeepSeek & DeepThink

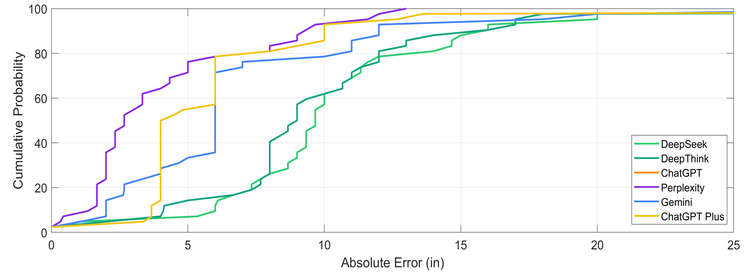

Figure 5. Distribution of Absolute Error Across Different LLMs.

- **Which types of words cause LLMs to fail in capturing flood depths?** – numerical values unrelated to flood depth (Economic loss, sinkhole), underground floods (subway), and ambiguous objects (road closure, above wheel, car tire).

Table 2. Keywords from the Five Most Erroneous Flood Depth Estimates

Rank	Gemini	DeepSeek	DeepThink	ChatGPT	ChatGPT Plus	Perplexity
1	Above Wheel (34)	Sinkhole (26)	Sinkhole (72)	Sinkhole (72)	Sinkhole (72)	Above Wheel (13)
2	Subway (20)	Economic Loss (20)	Subway (18)	Car Tire (13.6)	Car Tire (13.6)	Sinkhole (12)
3	Economic Loss (18)	Subway (20)	Above Wheel (17)	Above Wheel (10)	Above Wheel (10)	Subway (11.67)
4	No Clue (12)	Road Closure (16)	Subway (17)	Subway (10)	Subway (10)	Subway (9.67)
5	Road Closure (12)	Above Wheel (16)	Road Closure (16)	Road Closure (10)	Road Closure (10)	Economic Loss (9.33)

Conclusion and Future Studies

- LLMs can capture the flood depths with simple prompts from textual description.
- Advanced models do not always perform better than standard models.
- The simple prompt-based approach has limitations in addressing complex contexts within textual description.
- Improved approach, including prompt engineering, retrieval-augmented generation, and multi-LLMs-based application, should be further tested.

References

- Seon-Ho Kim, Fatemeh, Yavari, Te Pei, & Naresh Devineni (2025). Can Large Language Models Quantify Urban Floods from Crowd-sourced Information? (Under Review).
- National Oceanic and Atmospheric Administration. (n.d.). *Storm Event Database*. https://www.ncdc.noaa.gov/stormevents/

This summer research project was supported by Pinkerton Foundation (HIRES).
The authors would like to express their gratitude to their faculty advisors Naresh Devineni, mentors Chaw Nandar & Seon-Ho Kim. The statements contained within the poster are not the opinions of the funding agency or the U.S. government but reflect the author's opinions.