Do Large Language Models Reason About Uncertainty Like Humans? A Benchmark on Hurricane Forecast Visualization Comprehension

Le Liu[1], **Yuhao Wang**[1], **Bohan Shen**[1], **Wei Zeng**[2], **Shizhou Zhang**[1], **Di Xu**[3*], **Peng Wang**[1]
[1]Northwestern Polytechnical University, China
[2]The Hong Kong University of Science and Technology (Guangzhou), China
[3]Huawei, China
leliu304@gmail.com, yuhaowang@mail.nwpu.edu.cn, shenbohan@mail.nwpu.edu.cn,
weizeng@hkust-gz.edu.cn, szzhang@nwpu.edu.cn, xudi21@huawei.com, peng.wang@nwpu.edu.cn

Abstract

Uncertainty visualizations, such as hurricane cones and ensemble tracks, are essential for risk communication but are often misinterpreted, leading to harmful decisions. As AI assistants like large language models (LLMs) increasingly support understanding of graphics and decision-making, they offer a promising pathway to enhance the interpretation of complex visualizations and a new opportunity to examine and improve the interpretation of uncertainty. We introduce **UnReason**, the first benchmark that systematically compares how humans and LLMs reason about hurricane forecast uncertainty visualizations. **UnReason** spans two escalating phases, seven representative visualization formats, six real hurricane cases, and three agent types (humans, LLMs with context, and LLMs without context), including 880 visualizations and 117,600 structured question–answer pairs under matched evaluation conditions. Phase 1 evaluates reasoning across implicit and explicit uncertainty encodings; Phase 2 examines reasoning under single- versus multi-dimensional uncertainty representations. We thoroughly assess damage estimation, reasoning strategies, and comprehension patterns, revealing that LLMs have a stronger semantic and conceptual understanding of uncertainty, and are less misled by visual variability, but still replicate key human biases during decision-making. Our findings offer insights into aligning LLM behavior with human cognition in uncertainty-rich visual reasoning tasks.

Datasets & Extended version —
https://github.com/lel304/UnReason

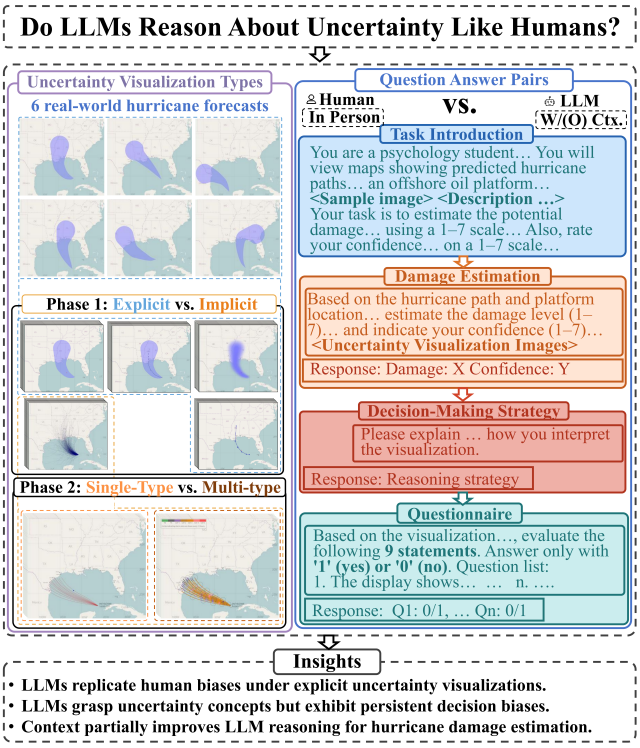


Figure 1: Overview of our research flow, **UnReason** benchmark construction, and summary of insights. Experimental images are derived from studies by (Ruginski et al. 2016) (phase 1) and (Liu et al. 2019) (phase 2). The complete prompt texts are provided in Appendix Figure 11.

Introduction

Uncertainty visualization is essential for risk communication in high-stakes domains such as hurricane forecasting, where public interpretation directly impacts life-saving decisions. Visual tools like the cone of uncertainty, widely used by the U.S. National Hurricane Center (NHC) and mass media, aim to convey forecast uncertainty, but are frequently misinterpreted. Prior studies show that users often mistake wider cones as stronger storms (Ruginski et al. 2016), assume minimal risk outside the cone (Cox, House, and Lindell 2013), or overtrust the central forecast line (Broad et al. 2007). These misconceptions are contributing to delayed evacuations, inadequate preparedness, and declining public trust in official forecasts (Witt et al. 2023). Although prior efforts have sought to redesign visual encodings (Liu et al. 2019) or improve user education (Demmans Epp and Bull 2015; Shilo and Raidou 2024), such interventions are challenging to address.

Meanwhile, LLMs are increasingly positioned as cognitive agents in decision workflows (Wang et al. 2022). Re-

*Corresponding author.

cent studies suggest that LLMs can interpret multimodal input (Hoque, Kavehzadeh, and Masry 2022; Ge et al. 2024), articulate reasoning (Bendeck and Stasko 2025; Zhang et al. 2024), and support user understanding (Chen et al. 2023), offering a potential paradigm shift in how uncertainty visualizations are reasoned, communicated, or corrected. However, fundamental questions remain: Do LLMs reason about uncertainty visualizations like humans? Do they exhibit the same bias or reason more consistently and accurately?

To address these questions, we introduce **UnReason**, the first comprehensive benchmark to systematically compare how humans and LLMs interpret hurricane forecast uncertainty visualizations. Built upon two foundational cognitive studies (Ruginski et al. 2016; Liu et al. 2019), **UnReason** spans two escalating phases and includes seven representative visualization formats, six real hurricane scenarios, and three agent types (humans, LLMs with context, and LLMs without context), as demonstrated by Figure 1. It pairs each visualization with a carefully designed decision-making task sequence, comprising damage estimation, strategy explanation, and comprehension questionnaire, resulting in over 880 forecast visualizations and 117,600 structured question–answer pairs.

Our experiments reveal that LLMs demonstrate structured reasoning patterns distinct from humans. They are often less susceptible to misleading visual cues, but not immune to critical misconceptions. For example, although LLMs conceptually recognize that cone width encodes uncertainty rather than intensity, they nonetheless assign higher damage ratings to wider cones. This gap between understanding and applied judgment highlights both the potential and the limitations of LLMs as cognitive agents for interpreting uncertainty visualizations.

Our contributions are: 1) proposing **UnReason**, the first comprehensive benchmark enabling structured, parallel comparisons between human and LLM reasoning under uncertainty visualizations; 2) conducting extensive evaluations across decision-making tasks, identifying where LLMs align with or diverge from human cognition; 3) offering new insights into the interpretability, limitations, and potential of LLMs as trustworthy reasoning companions in uncertainty-rich, high-risk scenarios.

Related Work

Human Understanding of Hurricane Forecast Uncertainty Visualizations The U.S. NHC cone of uncertainty remains the most widely used method for communicating hurricane forecast uncertainty to the public. However, a large body of research has shown that users frequently misinterpret these graphics. Common misconceptions include assuming safety outside the cone, interpreting its widening as increasing storm intensity, or over-relying on the central forecast track (Broad et al. 2007; Cox, House, and Lindell 2013; Ruginski et al. 2016). These cognitive biases persist across a variety of alternative designs that adopt similar principles by explicitly conveying uncertainty by visualizing key statistical characteristics (Ruginski et al. 2016; Liu et al. 2017; Bhatt et al. 2021). On the contrary, experimental studies indicate that implicit uncertainty visualizations, such as ensemble displays, represent the data directly without showing summary statistics explicitly, which may help reduce misconceptions. (Cox, House, and Lindell 2013; Ruginski et al. 2016; Liu et al. 2019). However, they may introduce new challenges like visual clutter and difficulty in interpreting multifaceted uncertainties (Liu et al. 2023; Buschmann, Trapp, and Döllner 2016).

Despite years of effort, no visualization technique has completely eliminated human reasoning misconceptions, highlighting the limitations of relying solely on visual design to improve understanding of forecast uncertainty.

Chart Understanding with Multimodal LLMs Recent advances in multimodal LLMs (MLLMs) have enabled processing of both images and text, expanding their role for chart understanding (Huang et al. 2025). However, while these models demonstrate strong perception and reasoning capabilities with natural images (Grassini and Koivisto 2025; Cao et al. 2024), they struggle to match human-level assessment when interpreting charts. For example, GPT-4 demonstrates notable difficulties with value lookup tasks (Hoque, Kavehzadeh, and Masry 2022; Bendeck and Stasko 2025) and interpreting fine-grained visual encodings (Zeng et al. 2025), as evidenced by evaluations using VLAT benchmarks and diverse stimulus sets (Pandey et al. 2015; Lee, Kim, and Kwon 2016). Notably, enhancing MLLMs' performance on domain-specific chart interpretation remains a critical research challenge (Huang et al. 2025). In line with this goal, our work investigates MLLMs' reasoning processes for uncertainty visualizations, establishing foundational insights for future model enhancements.

Human-LLM Alignment and Benchmark Recent studies have explored human–LLM alignment across a range of tasks using existing data and benchmarks of human performance, including theory of mind (van Duijn et al. 2023; Strachan et al. 2024; Kosinski 2024; Niu et al. 2024), spatial cognition (Yamada et al. 2024), field of medicine (Sallam et al. 2024; Luo et al. 2025), and collective decision-making (Yang et al. 2024). While some efforts have begun to construct benchmarks incorporating both LLM and paired human data for theory of mind (van Duijn et al. 2023) and visual reasoning (Cao et al. 2024), no existing benchmark has systematically examined how humans and LLMs interpret and reason about uncertainty visualizations.

UnReason Benchmark Construction

Overview We introduce **UnReason**, the first benchmark to comprehensively evaluate human–LLM alignment in reasoning about hurricane forecast uncertainty visualizations. Centered on estimating hurricane damage risk to offshore oil platforms, a high-stakes and ecologically grounded task, UnReason captures how human and LLM agents interpret, explain, and act upon uncertainty visualizations. The benchmark includes 880 forecast visualizations and 117,600 structured question–answer pairs, spanning seven visualization types, two forecasting timepoints from six real hurricane forecasts. It integrates behavioral data from 300 human participants (Ruginski et al. 2016; Liu et al. 2019) paired with

LLM responses under matched prompting conditions, with and without long-term context.

UnReason is organized into two reasoning layers. Phase 1 contrasts implicit versus explicit encodings, probing how visual format influences interpretation. Phase 2 examines agents' capacity to integrate multi-dimensional uncertainty, such as track, storm size, and intensity, or whether they default to heuristics. Our benchmark evaluates reasoning across a structured cognitive stack: risk estimation, strategy explanation, and conceptual comprehension, enabling fine-grained comparisons of interpretive alignment, robustness, and cognitive bias, bridging human cognition with LLMs' behavior in uncertainty-rich scenarios, offering a new paradigm for evaluating visual reasoning in AI.

Visualization Type and Uncertainty Encoding All visualizations in **UnReason** are generated from historical hurricane forecasts provided by the U.S. NHC, following standard protocols for cone and ensemble modeling. Each visualization depicts a continuous 72-hour forecast of a storm's projected path. Participants view the full forecast but estimate damage at a specific time point (24 or 48 hours) based on the expected impact at a simulated offshore oil platform. This task mirrors real-world scenarios where localized decisions must be made under evolving uncertainty. Phase 1 examines how different uncertainty encodings influence interpretation across five visualization formats (Ruginski et al. 2016). Explicit formats include cone-centerline, cone-only, fuzzy-cone, and centerline-only. The cone-centerline replicates the standard U.S. NHC product combining a probabilistic cone with a central forecast track. Cone-only isolates the effect of removing the visual anchor, while centerline-only presents only the deterministic track. The fuzzy-cone softens cone boundaries to convey uncertainty without sharp contours. The implicit format is an ensemble display that overlays multiple forecast tracks, requiring viewers to infer uncertainty from spatial dispersion rather than explicit boundaries. Phase 2 explores reasoning under compositional uncertainty using two ensemble-based formats (Liu et al. 2019). The representative-ensemble shows a curated subset of tracks that preserve spatial variability while reducing clutter. The annotated-ensemble further encodes variables such as storm size and intensity directly on the tracks.

Benchmark Structure and Task Configurations Our benchmark is organized around forecast-image trials. Each trial presents a visualization of a specific storm, evaluated at a simulated offshore platform location and forecast timepoint. In Phase 1, we selected six historical hurricanes, each rendered using five visualization formats. For every forecast–visualization pair, we generated two sets of 12 oil rig locations corresponding to the 24-hour and 48-hour timepoints, yielding 720 unique trials (6 storms $\times$ 5 visualizations $\times$ 2 timepoints $\times$ 12 locations). This configuration directly follows the design of (Ruginski et al. 2016), enabling precise comparison between LLM and human behavioral data. To populate the dataset, we collected responses from three agent groups. The first group consisted of 200 human participants, each completing 144 damage-rating trials under a single visualization condition, along with 12 additional think-aloud trials. The other two groups were composed of LLMs evaluated under two prompting conditions: one with context (*w/ ctx.*) and one without (*w/o ctx.*). Each LLM configuration was run 200 times using the same trial structure as the human participants. In total, Phase 1 yielded 86,400 structured rating QA pairs across all three agent groups. While human participants provided verbal explanations in the 12 think-aloud trials, these were not systematically available across all trials. In contrast, LLM agents were prompted to generate free-text explanations for every trial, resulting in 4,800 strategy rationales across two model groups. These were subsequently annotated using a rule-based coding scheme to quantify model reasoning. All agents, including humans and LLMs, also completed a comprehension questionnaire to assess their understanding of key concepts. Phase 2 adopts the experimental configuration of Liu et al. (Liu et al. 2019), focusing on compositional uncertainty. Five hurricane forecasts were visualized using representative-ensemble and annotated-ensemble formats. For each, eight simulated platform locations were sampled per timepoint (24h and 48h), producing 160 trials. These were completed by 100 human participants and 100 LLM runs per prompting mode, yielding 24,000 rating QA pairs. As in Phase 1, only the LLM agents provided full-text strategy explanations, yielding 2,400 rationales across two prompting conditions, which were subsequently annotated using a rule-based coding scheme.

Human Baselines Our benchmark integrates human response data from two foundational studies: (Ruginski et al. 2016) for phase 1 and (Liu et al. 2019) for phase 2. All responses were reformatted into a unified schema aligned with the benchmark's visual inputs and task structure. This alignment ensures that human and LLM agents are evaluated under identical conditions, enabling direct, fine-grained comparisons of reasoning behaviors across trials.

LLMs' Reponses and Prompting Strategy Unlike human participants, who receive instructions through in-person demonstrations and guidance, LLMs rely entirely on prompts to interpret the task, process visual input, and generate structured responses. We therefore designed a four-stage prompting pipeline that mirrors the original instruction structure, response constraints, and task flow (Figure 1). Intro Prompt initializes the session with (1) a scenario description defining the LLM's role, (2) a sample visualization, (3) an explanation of the visualization type, and (4) a task directive to report platform damage and confidence on 1–7 scales. This staged setup aligns the LLM's contextual grounding with that of human participants (Cui, Li, and Zhou 2025). Rating Prompt then presents visualization-specific trials, requiring the LLM to provide damage and confidence estimates following a strict response template. Extraneous text is explicitly prohibited to ensure clean, comparable outputs (Amatriain 2024). Strategy Prompt elicits an open-ended explanation of the LLM's rating rationale, mirroring the human think-aloud protocol and revealing cues such as storm proximity, cone inclusion, or trajectory dispersion. Questionnaire Prompt assesses visualization comprehension via declarative statements (Table 20 in Appendix B.3 for

Phase 1; Table 38 in Appendix C.3 for Phase 2). The LLM responds using a binary format (`1`=Agree, `0`=Disagree), analogous to the human post-task questionnaire. All prompts were run with fixed sampling settings ($temperature$=0.7, top_p=1) and max_tokens=200 or 1000 for non-strategy and strategy responses, respectively. Non-conforming outputs were reissued using the same prompt.

Experiment

Experimental Objectives Our goal is to investigate deeper cognitive alignment across agents. Specifically, we ask: 1) Do LLMs produce similar damage estimations as humans when viewing the same visualization under the same conditions? 2) How does persistent context affect the LLMs' decision-making? 3) Do LLMs employ reasoning strategies comparable to humans, and do they correctly internalize the intended semantics visualizations?

Trial Structure and Behavioral Measures Each trial presents a hurricane forecast visualization with a simulated offshore oil rig platform. The agent is asked to: 1) estimate the platform's damage at a specific forecast timepoint and report confidence; 2) provide a brief free-text explanation of their reasoning; 3) complete a comprehension questionnaire assessing visualization interpretation. This procedure replicates the protocols in (Ruginski et al. 2016; Liu et al. 2019), ensuring human and LLM trials align in both procedure and output format.

Multimodal Large Language Model Selection We selected **GPT-4o** as the representative model for our benchmark based on both prior literature (Bendeck and Stasko 2025) and an experimental comparison across six state-of-the-art vision-language models: Gemma-3-27B-IT, LLaMA-4-Scout-17B-16E, Qwen2.5-VL-72B-Instruct, Claude-4-Sonnet, Gemini-2.5-Pro-Preview, and GPT-4o. Following a consistent prompting framework, we tested each model on a representative subset of our benchmark. Several models produced plausible explanations but failed to reflect visual differences in their ratings. Others generated repetitive or unreasonable outputs. GPT-4o was the only model that varied its responses with storm time, platform location, and uncertainty format, and explicitly referenced relevant visual cues in its reasoning. We provide the complete evaluation details in Appendix F.

Statistical Modeling We fit a linear mixed-effects model damage ratings using four core predictors: agent group (G), visualization type (V), spatial distance from the storm center (D), and forecast timepoint (T), along with their interactions. The compact expression of the model is:

$$\text{Damage}_{ijkl} = \mathbf{X}_{ijkl}\boldsymbol{\beta} + u_{s(i)} + \varepsilon_{ijkl} \quad (1)$$

Here, $\mathbf{X}_{ijkl}$ encodes the fixed effects and their interactions across G, V, D, and T. $\boldsymbol{\beta}$ is the coefficient vector for these effects. $u_{s(i)} \sim \mathcal{N}(0, \sigma_u^2)$ is a random intercept for each subject or model run, and $\varepsilon_{ijkl} \sim \mathcal{N}(0, \sigma^2)$ is the trial-level residual error. This formulation allows us to evaluate whether agents differ in how they modulate spatial and temporal inferences across visualizations. A complete expansion of the model specification is provided in Appendix G.

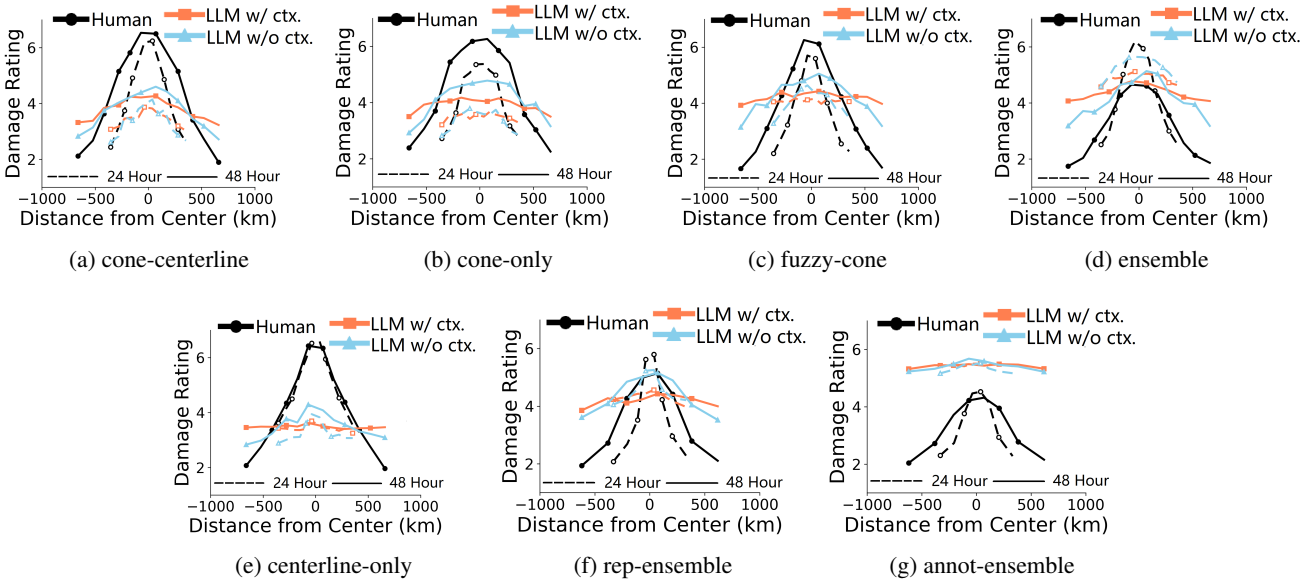
Results

We examine four well-established effects across visualization types and agent groups in both phases: **Center Effect**: variation in damage ratings at the storm center ($distance = 0$) under the 24-hour forecast. **Time** $\times$ **Center Effect**: how center effect evolves from 24h to 48h. **Distance Effect**: variation in ratings with distance from the storm center at 24h. **Time** $\times$ **Distance Effect**: how distance effects evolve from 24h and 48h. These effects capture key human reasoning patterns and common cognitive biases, including overreliance on the centerline, underestimation of risk outside the cone, and interpreting wider cones as stronger storms (Ruginski et al. 2016). Reasoning trials were coded into eight visual properties: Distance, Containment, Count, Depth of Color, Curve, Size, Intensity, and Movement (details in Appendix H). We analyzed the average number of trials where agents justified their responses based on these properties. We also analyze questionnaire responses and LLM confidence levels. This section summarizes the primary findings; the complete data and analysis are in Appendix B (Phase 1) and Appendix C (Phase 2).

Phase 1: Explicit vs. Implicit Uncertainty Encoding

Center Effect As shown in Figure 2(a–e), dashed lines at $distance = 0$ indicate the average central ratings for human participants (black), LLMs without context (blue), and LLMs with context (orange). Using the cone-centerline as the reference, consistent with its role in official forecasts, human participants reproduced prior findings (Ruginski et al. 2016): damage ratings at the center decreased significantly under cone-only and fuzzy-cone formats, suggesting that the presence of a centerline amplifies perceived storm severity. Ratings for ensemble and centerline-only formats did not significantly differ from cone-centerline. LLMs showed significantly different patterns. Both with and without context, LLMs gave significantly lower central ratings than humans (LLM w/o context: $\beta = -2.43$, LLM w/ context: $\beta = -2.76$, both $p < .001$, Table 1 in Appendix B.1), and rated higher relative to cone-centerline in ensemble ($\beta = 1.74$ and $\beta = 1.36$ for w/ and w/o context, both $p < .001$, in Appendix B.1) and fuzzy-cone ($\beta = 1.33$ and $\beta = 1.04$ for w/ and w/o context, both $p < .001$, in Appendix B.1), an inverse of the human pattern. While providing conversational context reduced overall central estimates ($\beta = -0.32$, $p < .001$, Table 4 in Appendix B.1), it did not shift LLMs' general interpretation behavior toward humans. These findings reveal that while LLMs respond to both explicit and implicit uncertainty encoding, they diverge systematically from human perceptual heuristics in assessing the center effect across visualization types.

Time $\times$ **Center Effect** Human participants often misinterpret explicit visualizations (Ruginski et al. 2016). They assign higher damage ratings at 48 hours than at 24 hours, incorrectly inferring increased storm intensity from the widening cone. In contrast, under the implicit (ensemble) visualization, ratings decrease over time, suggesting a more uncertainty-aware interpretation based on spatial spread rather than shape extent. LLMs without context partially re-



(a) cone-centerline (b) cone-only (c) fuzzy-cone (d) ensemble

(e) centerline-only (f) rep-ensemble (g) annot-ensemble

Figure 2: Average damage ratings by distance under different visualizations. Compared with humans, LLMs show reduced center bias and more consistent temporal reasoning, especially when contextual memory is enabled. (a–e) Phase 1; (f–g) Phase 2. Solid = 48h, dashed = 24h (radii = 347.28 km, 186.07 km).

produced this temporal pattern (Figure 2a-e, $distance = 0$). As supported by Table 6 in Appendix B.1, the model gave higher 48-hour ratings under cone-centerline ($\beta = 0.613$, $p < .001$) and cone-only ($\beta = 0.619$, $p < .001$), mirroring the human bias. For ensemble, the model produced lower ratings over time ($\beta = -1.300$, $p < .001$), consistent with human responses but with a less steep decline, indicating more restrained interpretation of spatial uncertainty. LLMs with context showed similar trends, but further reduced the temporal increase under cone-only ($\beta = -0.501$, $p < .001$; Table 8 in Appendix B.1), suggesting memory continuity modestly enhances temporal consistency and mitigates over-reliance on cone width as a proxy for intensity. In summary, LLMs replicate human misconceptions under explicit visualizations but show improved robustness in reasoning over time, particularly when uncertainty is conveyed implicitly.

Distance Effect Human exhibited an apparent distance effect (Ruginski et al. 2016). In the cone-centerline condition, damage ratings declined from the center. This effect slightly weakened under cone-only and centerline-only, and became statistically non-significant under fuzzy-cone and ensemble formats, indicating that bounded contours amplify perceived central risk. LLMs, by contrast, displayed significantly weaker distance effects, as revealed by the slopes of dashed curves in Figure 2a-e. Without context, the model's slope under the cone-centerline was -0.045 (Table 10 in Appendix B.1), less than half that of humans (-0.121; Table 9 in Appendix B.1). Under cone-only, centerline-only, and fuzzy-cone, the slopes were either near zero or reversed (e.g., cone-only: $\beta = 0.016$, $p < .001$, Table 10 in Appendix B.1), suggesting reduced reliance on centrality. Ratings of LLM with context became even less sensitive to distance. Across all formats, slopes further flattened, with cone-centerline showing only a slight gradient ($\beta = 0.021$, $p < .001$; Table 12 in Appendix B.1). LLMs, particularly with context, exhibit more uniform risk assessments across spatial extent and are less sensitive to the human bias of interpreting central proximity as increased severity.

Time × Distance Effect Human participants showed a reduced distance effect at 48 hours under the cone-centerline visualization, which then flattened even more under the ensemble visualization, suggesting that ensemble visualizations tend to promote a more spatially distributed interpretation of hurricane risk over time (Ruginski et al. 2016). LLMs without context partially mirrored this pattern under cone-centerline, with a modest slope reduction ($\beta = 0.014$, $p < .001$; Table 14 in Appendix B.1), though the magnitude of this effect was somewhat smaller than in humans ($\beta = -0.026$; Table 13 in Appendix B.1). Under both the ensemble and cone-only formats, however, LLMs' behavior diverged: interaction slopes were negative (-0.013 for w/ ctx. and -0.015 for w/o ctx., both $p < .001$; Table 14 in Appendix B.1), indicating a decreased level of spatial sensitivity over time, opposite to human participants. One plausible explanation for this discrepancy is that LLMs already exhibit a flatter spatial profile at 24h, thus offering less room for further "flattening" at 48h. Overall, LLMs, particularly without context, demonstrate more stable spatial reasoning across timepoints, with reduced sensitivity to the visual expansion of spatial uncertainty over time.

Reasoning Strategies Comparison of Human and LLMs' reasoning strategies is shown in Figure 4a-e and detailed in Table 17-19 in Appendix B.2. In the ensemble condition, both groups relied mainly on Distance and Count, though LLMs applied them almost universally (narrower

bars). LLMs also used Movement more often, indicating broader spatiotemporal reasoning. Humans occasionally inferred storm severity from trajectory clustering (Depth of Color), whereas LLMs did not. In explicit formats (cone-centerline, cone-only, fuzzy cone), both groups commonly referenced Distance and Containment. LLMs tended to give templated responses, often starting with whether the platform was "inside the cone." Without context, LLMs relied more on Size (e.g., "narrow" or "wide" cones), suggesting a human-like misconception linking cone width to severity; context reduced this slightly. Under fuzzy-cone, humans used Depth of Color more frequently, while LLMs did not. Under centerline-only, both relied on Distance, but LLMs additionally used Movement. Notably, LLMs still applied Containment despite its absence in this format, indicating a possible misconception. Overall, LLMs tended to align more closely with humans under explicit conditions, even amplifying the misconception of cone size. In contrast, they diverged more in implicit conditions, emphasizing counting and distance-based cues.

Questionnaire Results Figure 5 in Appendix B.3 compares LLMs' agreement rates across nine comprehension questions (Table 20) to humans'. (Ruginski et al. 2016) reported typical human biases under cone-based visualizations, such as assuming that storm intensity grows over time or that areas outside the cone face much lower risk. In contrast, LLMs strongly agreed with accurate statements (Q1, Q3, Q7, Q9) and disagreed with misleading ones (Q2, Q4, Q5, Q6, Q8) across both implicit and most explicit conditions. They also showed highly consistent, categorical responses regardless of context, suggesting that LLMs accurately and robustly interpret the visual cues at the semantic level. However, under centerline-only, their agreement with statements about uncertainty distribution (Q1, Q9) decreased, indicating sensitivity to the absence of explicit uncertainty cues. Taken together with the damage rating and strategy results, LLMs show stronger conceptual accuracy than humans but remain susceptible to similar visual biases during decision-making.

Confidence As shown in Figure 7 in Appendix D, LLMs without context consistently reported high confidence (mean $\approx$ 5 on a 7-point scale), indicating a strong baseline decisiveness. With context enabled, confidence increased further (mean $\approx$ 6), suggesting that access to prior trials enhances internal certainty. However, this elevated confidence does not always reflect improved accuracy, especially under conditions where visual reasoning biases persist (e.g., cone-based formats).

Phase 2: Single- vs. Multi-Dimensional Uncertainty

Center Effect As shown in Figure 2(f–g), dashed lines at $distance = 0$ indicate the average central ratings for human participants and LLMs. In the representative-ensemble condition, human participants gave an average rating of 5.64, while LLMs rated significantly lower, 5.17 without context ($\beta = -0.471$, $p < .001$ in Table 21 in Appendix C.1) and 4.55 with context ($\beta = -1.085$, $p < .001$ in Table 21 in Appendix C.1). This suggests that LLMs, particularly with context, adopt a more conservative stance in early damage assessment when only trajectory information is shown. Reproducing findings in (Liu et al. 2019), humans showed reduced central ratings under the annotated-ensemble condition compared to representative-ensemble ($\beta = -0.908$, $p < .001$ in Table 21 in Appendix C.1). In contrast, LLMs significantly increased their ratings under annotation (without context: $\beta = 0.358$, $p < .001$ in Table 22, with context: $\beta = 0.938$ in Table 23 in Appendix C.1; both $p < .001$), indicating varied cognitive patterns in the encoding of the uncertainty distributions of storm size and intensity.

Time $\times$ Center Effect In the representative-ensemble condition, humans rated 48-hour damage significantly lower than 24-hour damage, rejecting the misconception that longer lead time implies greater storm intensity. This effect was more substantial under annotated-ensemble, suggesting that encoding of uncertainties of storm size and intensity further improves temporal judgment (Liu et al. 2019). LLM without context showed no significant time effect within-group, but its higher 48-hour ratings diverged from humans ($\beta = 0.362, p < .001$ in Table 25 in Appendix C.1), indicating timing-related misinterpretation. In contrast, LLM with context aligned with human trends, suggesting that visual encoding of multi-dimensional uncertainty, when combined with context, can help mitigate the misconceptions.

Distance Effect At 24 hours, under the path-only condition, human participants exhibited a slope of -0.115 with respect to $\mathrm{Distance}$, whereas LLM without context showed a much flatter slope of -0.036, significantly different from humans ($\mathrm{coef.} = 0.079$, $p < 0.001$, Table 29 in Appendix C.1). With context, the slope became even flatter at -0.014 ($\mathrm{coef.} = 0.101$, $p < 0.001$, Table 29 in Appendix C.1), consistent with Experiment 1's finding that LLM displays a reduced centrality bias compared to humans, further strengthened by contextual information, which enhances LLM's ability to make reasonable judgments. When annotation information is added, all three groups—humans ($\mathrm{coef.} = 0.037$, $p < 0.001$, Table 29 in Appendix C.1), LLM with context ($\mathrm{coef.} = 0.024$, $p < 0.001$, Table 30 in Appendix C.1), and LLM without context ($\mathrm{coef.} = 0.012$, $p < 0.001$, Table 31 in Appendix C.1)—show significantly flatter slopes. This suggests that annotations help mitigate centrality bias in both humans and LLM, effectively conveying spatial uncertainty.

Time $\times$ Distance Effect Humans showed a flattening of damage estimation from 24 to 48 hours in the representative-ensemble condition ($\beta = 0.056$, $p < 0.001$, Table 33 in Appendix C.1). This flattening was enhanced significantly by including annotations of uncertainty of additional storm charateristics ($\beta = -0.020$, $p < 0.001$, Table 33 in Appendix C.1). Together with time $times$ distance effect analysis of Phase 1, this suggests that such annotations amplify the advantages of implicit uncertainty visualization to mitigate the misleading of the increasing of spatial uncertainty as an increasing in storm severity. For LLMs with and without context, no significant time-distance effect was observed under any visualization type (Table 34 and Table 35 in Ap-

pendix C.1). Together with the significantly flattened slopes discussed in the distance effect, this suggests a strong consistency of LLMs to mitigate central overestimations across different forecast times.

Reasoning Strategies Since human strategies were not collected by (Liu et al. 2019), our analysis focuses on LLM strategies (Table 37 in Appendix C.2 and Figure 4f–g in Appendix B.2). Results show that contextual memory had minimal influence on overall strategy distribution. Under the representative-ensemble condition, LLMs relied mainly on Distance and Count, with occasional Movement, emphasizing spatial proximity and forecast density similar to Phase 1. With annotations, LLMs shifted to richer strategies, increasing use of Containment, Color, Intensity, and Size, indicating successful interpretation of multidimensional uncertainty visual encodings. Context slightly reduced reliance on Count and Color, possibly inducing attention toward more abstract storm characteristics. Overall, annotated formats prompted more semantically grounded reasoning, with LLMs moving beyond spatial heuristics to interpreting encoded uncertainty cues.

Questionnaire Results Figure 6 (Appendix C.3) summarizes agreement across six comprehension items (Table 38 in Appendix C.3). LLMs showed consistent patterns across context settings and visualization, strongly agreeing with Q4 and Q6, and rejecting Q1–Q3 and Q5. These responses suggest accurate conceptual understanding: recognizing uncertainty expansion (Q4), expressing high confidence (Q6), and rejecting misconceptions such as increasing storm intensity over time (Q3) or zero risk outside the display (Q5). Human responses were more varied. Many participants struggled with Q3–Q5, particularly under annotated visualizations, reflecting persistent difficulty with reasoning intensity and storm size trends. Together, these results reinforce Phase 1 findings: LLMs exhibit more coherent semantic understanding of uncertainty visualizations than humans, but still lead to decision-making biases.

Confidence Aligning with Phase 1, LLMs reported higher confidence (Figure 7 in Appendix D). Compared to representative-ensemble displays, annotated-ensemble visualizations further increased the confidence level, especially under the with-context condition, suggesting that visual annotation and context history jointly reinforce LLMs' certainty in complex, multi-dimensional uncertainty reasoning.

Result Summary and Insights

Overall, our results reveal that LLMs reason about uncertainty visualizations in more structured but rigid ways, differing from humans. Interestingly, although they exhibit a superior conceptual understanding and are less misled by visual variability, they still replicate key human biases during decision-making, particularly under explicit encodings. Enabling context partially improves LLM reasoning, and LLMs adopt more semantically grounded strategies.

Discussion

Context Window Size We experimentally evaluated GPT-4o's capacity to handle conversational histories using its official APIs. Results showed that it reliably processed up to 36 consecutive rounds in our study without encountering timeout errors (see Appendix A for full details of the evaluation).

Probing the Malleability of LLMs' Cognition To probe the origin of LLMs' decision-level biases, we conducted a prompt intervention that explicitly disambiguated visual uncertainty from storm severity, partially addressing key misconceptions of cone width as storm intensity. Full details are in Appendix E. The results suggest that LLMs' reasoning is neither rigidly visual nor purely statistical, but shaped by an internalized semantic model sensitive to framing. The fact that a brief textual clarification can override visual heuristics indicates that LLMs dynamically reweight interpretive strategies based on contextual priors. In this context, prompt and context engineering serve as a tool for behavior control, and they also provide insight into the LLMs' internal decision-making structure, thus offering a practical way to adjust their cognitive processing in situations under uncertainty. This opens a broader research direction on semantic alignment: not just whether LLMs can perform human-like reasoning, but how their latent representations can be guided and calibrated toward desirable interpretive norms.

Model Generalizability GPT-4o was chosen as the representative model since a comparative evaluation demonstrating that it was the one among several state-of-the-art multimodal LLMs to exhibit stable and visually grounded reasoning (Appendix F). Consequently, our findings are most reliably generalizable to models that possess similar levels of multimodal alignment and reasoning capability.

Limitation and Future Work A limitation of this study is the lack of data on the reasoning strategies of human participants in Phase 2, preventing a direct comparison between human and LLMs' reasoning under multi-dimensional uncertainty visualization. Furthermore, while this study identifies systematic differences between human reasoning and LLMs' interpretation, it remains a challenge to incorporate these findings into model alignment objectives, architectural changes, or prompt-engineering strategies.

Conclusion

This paper presents **UnReason**, the first benchmark for evaluating how LLMs and humans interpret uncertainty visualizations in hurricane forecasting. Through two experimental phases with human participants and evaluations of LLMs under controlled conditions, we examined alignment in damage estimation, reasoning strategies, and conceptual understanding. We found that LLMs show some human-like reasoning but also exhibit cognitive biases. They excel in conceptual understanding but are easily misled by visual cues. Overall, while initial insights into LLM cognition are promising, achieving deeper alignment remains a challenge, highlighting the need for further research.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant No. 62472357. The authors would like to thank Ian Ruginski, Sarah Creem-Regehr, and Lace Padilla.

References

Amatriain, X. 2024. Prompt Design and Engineering: Introduction and Advanced Methods. arXiv:2401.14423.

Bendeck, A.; and Stasko, J. 2025. An Empirical Evaluation of the GPT-4 Multimodal Language Model on Visualization Literacy Tasks. *IEEE Transactions on Visualization and Computer Graphics*, 31(1): 1105–1115.

Bhatt, U.; Antorán, J.; Zhang, Y.; Liao, Q. V.; Sattigeri, P.; Fogliato, R.; Melançon, G.; Krishnan, R.; Stanley, J.; Tickoo, O.; et al. 2021. Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 401–413. New York, NY, USA: ACM.

Broad, K.; Leiserowitz, A.; Weinkle, J.; and Steketee, M. 2007. Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society*, 88(5): 651–668.

Buschmann, S.; Trapp, M.; and Döllner, J. 2016. Animated visualization of spatial–temporal trajectory data for air-traffic analysis. *The Visual Computer*, 32(3): 371–381.

Cao, X.; Lai, B.; Ye, W.; Ma, Y.; Heintz, J.; Chen, J.; Cao, J.; and Rehg, J. M. 2024. What is the Visual Cognition Gap between Humans and Multimodal LLMs? arXiv:2406.10424.

Chen, Z.; Zhang, C.; Wang, Q.; Troidl, J.; Warchol, S.; Beyer, J.; Gehlenborg, N.; and Pfister, H. 2023. Beyond Generating Code: Evaluating GPT on a Data Visualization Course. In *2023 IEEE VIS Workshop on Visualization Education, Literacy, and Activities (EduVis)*, 16–21.

Cox, J.; House, D.; and Lindell, M. 2013. Visualizing uncertainty in predicted hurricane tracks. *International Journal for Uncertainty Quantification*, 3: 143–156.

Cui, Z.; Li, N.; and Zhou, H. 2025. A large-scale replication of scenario-based experiments in psychology and management using large language models. *Nature Computational Science*, 1–8.

Demmans Epp, C.; and Bull, S. 2015. Uncertainty Representation in Visualizations of Learning Analytics for Learners: Current Approaches and Opportunities. *IEEE Transactions on Learning Technologies*, 8(3): 242–260.

Ge, Z.; Huang, H.; Zhou, M.; Li, J.; Wang, G.; Tang, S.; and Zhuang, Y. 2024. WorldGPT: Empowering LLM as Multimodal World Model. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 7346–7355. New York, NY, USA: Association for Computing Machinery.

Grassini, S.; and Koivisto, M. 2025. Artificial creativity? Evaluating AI against human performance in creative interpretation of visual stimuli. *International journal of human–computer interaction*, 41(7): 4037–4048.

Hoque, E.; Kavehzadeh, P.; and Masry, A. 2022. Chart question answering: State of the art and future directions. In *Computer Graphics Forum*, volume 41, 555–572.

Huang, K.-H.; Chan, H. P.; Fung, M.; Qiu, H.; Zhou, M.; Joty, S.; Chang, S.-F.; and Ji, H. 2025. From Pixels to Insights: A Survey on Automatic Chart Understanding in the Era of Large Foundation Models. *IEEE Transactions on Knowledge and Data Engineering*, 37(5): 2550–2568.

Kosinski, M. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45).

Lee, S.; Kim, S.-H.; and Kwon, B. C. 2016. Vlat: Development of a visualization literacy assessment test. *IEEE transactions on visualization and computer graphics*, 23(1): 551–560.

Liu, L.; Boone, A. P.; Ruginski, I. T.; Padilla, L.; Hegarty, M.; Creem-Regehr, S. H.; Thompson, W. B.; Yuksel, C.; and House, D. H. 2017. Uncertainty Visualization by Representative Sampling from Prediction Ensembles. *IEEE Transactions on Visualization and Computer Graphics*, 23(9): 2165–2178.

Liu, L.; Padilla, L.; Creem-Regehr, S. H.; and House, D. H. 2019. Visualizing Uncertain Tropical Cyclone Predictions using Representative Samples from Ensembles of Forecast Tracks. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 882–891.

Liu, L.; Wang, L.; Shrestha, J. R.; Zhao, K.; and Zhang, Y. 2023. Immersive Visualization of The Multifaceted Uncertainties of Hurricane Prediction Ensembles. In *2023 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 103–107. Sydney, Australia: IEEE.

Luo, X.; Rechardt, A.; Sun, G.; Nejad, K. K.; Yáñez, F.; Yilmaz, B.; Lee, K.; Cohen, A. O.; Borghesani, V.; Pashkov, A.; et al. 2025. Large language models surpass human experts in predicting neuroscience results. *Nature human behaviour*, 9(2): 305–315.

Niu, Q.; Liu, J.; Bi, Z.; Feng, P.; Peng, B.; Chen, K.; Li, M.; Yan, L. K.; Zhang, Y.; Yin, C. H.; Fei, C.; Wang, T.; Wang, Y.; Chen, S.; and Liu, M. 2024. Large Language Models and Cognitive Science: A Comprehensive Review of Similarities, Differences, and Challenges. arXiv:2409.02387.

Pandey, A. V.; Rall, K.; Satterthwaite, M. L.; Nov, O.; and Bertini, E. 2015. How deceptive are deceptive visualizations? An empirical analysis of common distortion techniques. In *Proceedings of the 33rd annual acm conference on human factors in computing systems*, 1469–1478. Seoul, Republic of Korea: ACM.

Ruginski, I. T.; Boone, A. P.; Padilla, L. M.; Liu, L.; Heydari, N.; Kramer, H. S.; Hegarty, M.; Thompson, W. B.; House, D. H.; and Creem-Regehr, S. H. 2016. Non-expert interpretations of hurricane forecast uncertainty visualizations. *Spatial Cognition & Computation*, 16(2): 154–172.

Sallam, M.; Al-Salahat, K.; Eid, H.; Egger, J.; and Puladi, B. 2024. Human versus artificial intelligence: ChatGPT-4 outperforming Bing, bard, ChatGPT-3.5 and humans in clinical

chemistry Multiple-Choice questions. *Advances in Medical Education and Practice*, 857–871.

Shilo, A.; and Raidou, R. G. 2024. Visual narratives to edutain against misleading visualizations in healthcare. *Computers & Graphics*, 123: 104011.

Strachan, J. W.; Albergo, D.; Borghini, G.; Pansardi, O.; Scaliti, E.; Gupta, S.; Saxena, K.; Rufo, A.; Panzeri, S.; Manzi, G.; et al. 2024. Testing theory of mind in large language models and humans. *Nature Human Behaviour*, 8(7): 1285–1295.

van Duijn, M. J.; van Dijk, B.; Kouwenhoven, T.; de Valk, W.; Spruit, M.; and van der Putten, P. 2023. Theory of mind in large language models: Examining performance of 11 state-of-the-art models vs. children aged 7-10 on advanced tests. In *Proceedings of the 27th Conference on Computational Natural Language Learning*, 389–402. Singapore: Association for Computational Linguistics.

Wang, Q.; Zhu-Tian, C.; Wang, Y.; and Qu, H. 2022. A Survey on ML4VIS: Applying Machine Learning Advances to Data Visualization. *IEEE Transactions on Visualization and Computer Graphics*, 28(12): 5134–5153.

Witt, J. K.; Labe, Z. M.; Warden, A. C.; and Clegg, B. A. 2023. Visualizing uncertainty in hurricane forecasts with animated risk trajectories. *Weather, Climate, and Society*, 15(2): 407–424.

Yamada, Y.; Bao, Y.; Lampinen, A. K.; Kasai, J.; and Yildirim, I. 2024. Evaluating Spatial Understanding of Large Language Models. arXiv:2310.14540.

Yang, J. C.; Dalisan, D.; Korecki, M.; Hausladen, C. I.; and Helbing, D. 2024. LLM Voting: Human choices and ai collective decision-making. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 1696–1708. San Jose, California, USA: AAAI Press.

Zeng, X.; Lin, H.; Ye, Y.; and Zeng, W. 2025. Advancing Multimodal Large Language Models in Chart Question Answering with Visualization-Referenced Instruction Tuning. *IEEE Transactions on Visualization and Computer Graphics*, 31(1): 525 – 535.

Zhang, K.; Zeng, J.; Meng, F.; Wang, Y.; Sun, S.; Bai, L.; Shen, H.; and Zhou, J. 2024. Tree-of-reasoning question decomposition for complex question answering with large language models. In *Proceedings of the AAAI Conference on artificial intelligence*, 19560–19568. Vancouver, Canada: AAAI Press.