Eco-anxiety and Insurance: Behavioral Experiments*

Eric Vansteenberghe[†]

Abstract

This paper examines the impact of eco-anxiety on the classification of insurance loss data using experiments with Large Language Model (LLM) agents. When individuals experience eco-anxiety, a third perceive risks as uninsurable, while those who still consider them insurable anticipate a 50% increase in expected loss. Additionally, wisdom—proxied by demographic characteristics such as age and experience—has no statistically significant effect on these outcomes. These findings highlight the need for a rational, collective approach to risk assessment that fosters informed action without exacerbating anxiety.

This version: February 18, 2025.

Keywords: Eco-anxiety, Insurance, Large Language Models, Behavioral Experiments.

JEL codes: C63, C91, D81, G22, Q54.

*I thank Liza Charroin, Nicolas Jacquemet, Mael Lebreton, Damien Mayaux, Lily Savey, Laurence Vardaxoglou and other Paris 1 Behavior Group for their feedback. I thank Mehdi Louafi, Claire Mollier, and other participants of the monthly Paris Junior Experimentalists Meetings. I thank Felix Feig, Abigail Hurwitz, Claire Mouminoux, Glenn Harrison and other participants of the 2024 CEAR / MRIC Behavioral Insurance Workshop.

[†]Banque de France : eric.vansteenberghe@banque-france.fr

1 Introduction

Economic systems can face sudden and unexpected disruptions, such as pandemics or geopolitical instability, caused by undergoing profound phenomenon such as climate change. Climate change amplifies the unpredictability of insurance loss distributions, leading to extreme and unforeseen outcomes. The mere knowledge of upcoming disruptions can reshape the behaviors of economic agents. Even rational professionals, such as actuaries, may experience decision-making biases caused by eco-anxiety, which could inadvertently contribute to subjective uninsurability. When eco-anxiety becomes prevalent in a population, some assets might stop being considered insurable, which could prevent a collective and rational problem solving approach in the face of climate change. This paper introduces a novel approach to understanding how eco-anxiety can reshape the insurance market by leveraging Large Language Models (LLMs), such as ChatGPT, to conduct rapid, cost-effective insurance experiments and simulations. Our research focuses on testing eco-anxiety and key demographic characteristics influence on how agents model and understand insurance losses. We build our experimental design using best practices and the ranking for demographic factors for integrating LLMs into behavioral experiments in Vansteenberghe (2025). Our experiments consist in asking agents to characterizing insurance losses from a small sample and our main focus are the implications of eco-anxiety for this task.

The experiment in this paper is grounded in the judgment under uncertainty literature (Tversky and Kahneman, 1974), where we task LLM agents with choosing between two parametric distribution families to express their beliefs. Drawing from this literature, particularly the bias linked to representativeness, we hypothesize that agents will predominantly classify samples as originating from a Gaussian Data Generation Process (DGP). This bias is driven by two factors: (1) agents may underestimate the likelihood of observing extreme values from a Pareto distribution in small samples, and (2) the Gaussian distribution's familiarity–it is widely recognized and frequently encountered, both by the general population and trained statisticians. A notable example supporting this familiarity is the Central Limit Theorem, which highlights the prevalence of Gaussian approximations. Additionally, the statistical profession often relies on the normality assumption due to its computational convenience (LaValle, 2006) and the challenges associated with detecting outliers (Leys et al., 2013). In some cases, normality assumptions are even encouraged (Knief and Forstmeier, 2021). We use known DGPs representing insurance losses to assess the behavioral realism of LLMs and enable a comparison with human participants in a laboratory. As experiment supervisor we generate random samples from either a Gaussian or a Pareto distributions. Each participant is tasked with discriminating whether the insurance loss sample comes from a "normal" or "fat tailed" distribution. The discrimination is reported confidentially to the experiment supervisor. A first straightforward test for the supervisor is whether the LLMs present the same bias as humans with respect to "normality". In a second phase of the experiment, we introduce eco-anxiety framing and evaluate first if the bias switch to pessimism and second if the demographic characteristics are influenced differently by eco-anxiety.

To efficiently initiate this research before securing funding for human-based experiments, we begin by leveraging LLM agents.[1] The first step involves following the most effective approach to prompt these models, based on the literature and surveys among researchers (Vansteenberghe, 2025). Next, we rely on the literature to reveal agents preference for insurance such as works on willingness-to-pay for flood-related insurance under climate change by Botzen, Aerts and van den Bergh (2008) for the Netherlands and Poussin, Botzen and Aerts (2013) for France. This literature provides respondent-level characteristics suitable for validating behavioral insurance experiments. However, given the availability of such data and prior research for LLM training, we focus on developing a unique experimental approach in this work to avoid seeing our LLM reciting or plagiarizing published work. An exciting extension of this research is the creation of methods to generate novel, never-before-published synthetic datasets for insurance experiments. This aligns with efforts in other domains, such as Assefa et al. (2020) in finance and Walonoski et al. (2018) in healthcare. Such datasets could validate the guidelines, tools, and architectures proposed in this paper while serving as valuable resources for future research.[2]

Literature: The experiment setup in this paper simplifies the complexity of information and decisions that economic agents face about insurance losses. Nevertheless, we aim to demonstrate how this can be applied to address current and pressing questions in insurance. Clearly identifiable examples are the growing impact of climate change on insurance losses (Charpentier, 2008) and behavioral consequences. Before insurance experts decide to revise their models and infrastructure, they must determine if the climate change has reached a tipping point. If so, they can decide that traditional models collapsed (Lenton et al., 2008; Ditlevsen and Ditlevsen, 2023) and agent fundamentally changed their decision making. This change can be a subjective or objective shift due to eco-anxiety (Pihkala, 2020;

[1] To fully realize the potential of this research, we aim to conduct a novel laboratory experiment or access an unpublished survey dataset.

[2] These synthetic datasets could also facilitate further studies and applications, reinforcing the methodological contributions of this work.

Hickman et al., 2021; Stanley et al., 2021; Coffey et al., 2021; Whitmarsh et al., 2022). There is therefore an identified challenge from economic agents risk perceptions and behaviors in insurance demand up to experts driving insurance supply. On the policyholder side, Harrison and Elisabet Rutström (2008) review the experimental evidence on risk aversion in controlled laboratory settings and Richter, Schiller and Schlesinger (2014) and Jaspersen (2016) review the literature in insurance demand experiments and surveys. Bhargava, Loewenstein and Sydnor (2017) show how policyholders can chose dominated options which contradict standard economic model of insurance choice. Harrison and Ng (2019) cite the behavioral insurance literature where roughly 50% of subjects best characterized by Expected Utility Theory (EUT) and 50% best characterized by Rank-Dependent Utility (RDU). This is a starting point to test whether LLM experiment can replicate this observed split among policyholders. There is on top of the concern that laboratory experiments results might not reflect behavior in naturally occurring settings (Harrison, List and Towe, 2007). On the insurance expert side, Haigh and List (2005) found that professional traders often exhibit greater myopic loss aversion than students, a counterintuitive result suggesting that even seasoned professionals can harbor biases typically associated with less informed individuals. Similarly, Alevy, Haigh and List (2007) demonstrated that professionals' decisions in experimental auctions align more closely with theoretical predictions, underscoring the role of experience in enhancing decision-making quality. Vansteenberghe (2024) builds on Raviv (1979) to model the insurance market when insurance experts have heterogeneous beliefs driven by climate change. We contribute to this debate by evaluating the potential shift in policyholders perceptions and insurance expert modeling.

Section 2 introduce our main experiment design using LLMs. Section 3 presents the experiment results and detail the impact of the eco-anxiety framing. Section 4 discusses our results and concludes.

2 Experiment Design with LLMs

To enable controlled comparisons, independent and identically distributed (iid) samples are drawn from two distinct probability distribution functions. The first is a Pareto DGP with a fixed threshold of zero and the second is a Gaussian DGP. The main idea is to chose parameters so that the samples cannot be overtly discriminated. The Kolmogorov-Smirnov (KS) statistic was used to calibrate the parameters of the distributions to align their means and variances. Using a grid search approach, four parameters were optimized while fixing the Pareto threshold at zero.[3] The Gaussian samples are truncated below zero to maintain compatibility with the Pareto distribution's positive support, reflecting real-world positive insurance losses. Once the parameters were optimized, the sample size was calibrated to achieve a balanced probability of correctly classifying the DGP type. Specifically, the goal was to ensure that the likelihood of accurately attributing the DGP to a specific type was approximately 50%. This was achieved using a two-sample KS test to evaluate the goodness of fit between observed samples and the theoretical distribution. A Monte Carlo (MC) simulation with 10^6 runs was employed, incrementally increasing the sample size n and recording the p-value from the KS test at each step. The critical sample size, $n_{0.5}$, was determined as the value of n where the median p-value crossed the 10% significance threshold. A sample size of ten observations balances the probability of correctly identifying the DGP type, as illustrated in Figure 1. Now, with this calibration and sample sizes, we can expect unbiased statisticians to correctly classify the sample half of the time, which will allow Pearson chi-square tests to detect biases in judgments.

The experiment involves m economic agents, all simulated using ChatGPT o1-mini. At the time of the first draft of this paper (December 2024), this model provided a suitable balance between cost-efficiency and advanced capabilities. The experiments were executed via API calls, ensuring flexibility and scalability in handling multiple agents and scenarios. The cost structure for the API was 1.5 USD input and 21.9 USD output cost for a thousand runs with 500 agents under the baseline and 500 agents under the eco-anxiety framing scenario. These costs influenced the design of the experiments, encouraging efficiency while maintaining the scope necessary for robust analysis. To ensure consistency and minimize variability in the outputs, we deliberately chose not to adjust the model's temperature, a

[3] The resulting Pareto parameters were: scale = 9.5234375 and shape = 2.25685313, providing the necessary degrees of freedom for matching. The Gaussian parameters were then $\mu_{\text{gaussian}} = \frac{\text{scale}\cdot\text{shape}}{\text{shape}-1}$ and $\sigma_{\text{gaussian}} = \sqrt{\frac{\text{scale}^2\cdot\text{shape}}{(\text{shape}-2)(\text{shape}-1)^2}}$.

parameter controlling the randomness of the LLM's responses. While higher temperatures can introduce creative variability, potentially capturing a broader range of plausible human behaviors, they first reduce reproducibility of our work and second increase the required sample size to test our effects.

The LLMs generate random profiles with heterogeneous demographic characteristics, focusing on those ranked in the researchers' survey (Vansteenberghe, 2025). In this experiment, we evaluate the influence of Financial Literacy, Education Attainment, and Sector of Employment by distinguishing between a Statistician with a PhD and a Farmer with no formal diploma.[4] We also test the effects of Professional Experience, Age, and Gender. According to our survey of researchers, the Role should have the strongest effect and Gender the least. We followed prompt engineering guidelines in Vansteenberghe (2025) and describe our process of writing the prompt for this experiment based on those guidelines Table 10. Prompt Engineering, also known as In-Context Prompting, refers to methods for how to communicate with an LLM to steer its behavior towards desired outcomes without updating the model's weights. It is an empirical science, and the effects of prompt engineering methods can vary significantly among models, thus requiring heavy experimentation and heuristics. Prompt engineering is a developing field of academic study (White et al., 2023; Giray, 2023; Wang et al., 2023; Jojic, Wang and Jojic, 2023). Our main prompt for this experiment was finalized as (the sample of size 10 observations being updated each time):

- **Experiment: Parametric Distribution Identification**
  - **Profile Setup:**
    - **Role:** Randomly assign one of the following roles:
      1. Statistician in the Finance industry with a PhD;
      2. Farmer with no formal diploma.
    - **Gender:** Randomly assign a gender (Male or Female).
    - **Age:** Randomly assign an age (minimum 22 years).
    - **Experience:** Randomly assign years of professional experience (not exceeding `age - 18`).
  - **Task Objective:** Analyze a dataset to identify its Data Generating Process (DGP), guided by your assigned profile.
  - **Dataset Details:**

[4] We focus on binary extremes to limit the number of potential combinations needed for statistical testing. We decided to focus on Farmer as they are exposed to insurance losses for their crops and they have likely faced insurance decision and are known to under-insure (Grislain-Letrémy, Villeneuve and Yeterian, 2024).

    - **Possible Distributions:** Gaussian (Normal) or Pareto.
  - **Output Format:** Provide a single-line result with these details:
    - **Parameters:** `loc`, `scale`, and `shape` (use `NA` for `shape` in Gaussian).
    - **Chosen Distribution:** `norm` for Gaussian, `pareto` for Pareto.
    - **Method Used:** Brief description of the identification method (e.g., Kolmogorov-Smirnov test, histogram analysis).
    - **Profile Information:** Role, Gender, Age, Experience.
  - **Output Examples:**
    1. `0,1,,norm,Kolmogorov-Smirnov test,35,Male,17,Statistician`
    2. `1,1.5,2,pareto,Histogram analysis,28,Female,6,Farmer`
- **Input:** A sample of size 10 observations:

```
32.69330325078325
23.01804007181683
0.35765309776633253
18.836435657036173
17.1389655992674
17.13411786433262
46.640015462016535
17.726928616489516
3.2532160515624122
10.027199635841825
```

When we framed with eco-anxiety, we modify the Task Objective:

- Analyze a dataset to identify its Data Generating Process (DGP), guided by your assigned profile. The analysis should consider the framing of eco-anxiety: accelerating climate change threatens irreversible damage within our lifetime, characterized by extreme and unpredictable events (e.g., insurance failure, food insecurity). Reflect on how these concerns might shape the choice of the most plausible distribution.

Upon securing funding, a subsequent project will incorporate human economic agents through laboratory experiments or surveys. Participants will be selected to ensure demographic heterogeneity, with pilot studies conducted at conferences to address the complexity of subjective probability distribution tasks. We will then run LLM experiments with the demographic characteristics of participants from the laboratory study.

3 LLM experiment results

The experiment involved a controlled assessment where LLMs, prompted with specific roles and demographic characteristics, were tasked with identifying and characterizing insurance loss data generated from Gaussian or Pareto distributions. The randomly generated samples were evenly distributed between normal and fat-tailed cases. Table 1 summarize the observations from Gaussian and Pareto samples. Figure 2 presents histograms of the sample data used in the experiment. Notably, the small sample sizes hinder straightforward discrimination between Gaussian and Pareto distributions based solely on empirical histogram visuals. However, when samples from each distribution are pooled together, as shown in Figure 3, the visual distinction becomes evident. This is further illustrated by comparing the empirical distributions of the sample maxima: while extreme observations are present in Pareto samples, their occurrence is infrequent due to the limited sample size.

TABLE 1
Comparison of pooled Gaussian and Pareto samples.

	Gaussian	Pareto
Count	2600	2590
Mean	25.7	7.5
Std Dev	16.9	13.6
Skewness	0.7	6.4
Min	0.0458	0.0030
Max	89.4	201.2

Note: The Mean and Standard Deviation here seems different, but when tested with higher sample size 10^6, they do converge.

3.1 Baseline experiment results

As expected, LLM agents utilized various statistical tests to discriminate between Gaussian and fat tailed DGP. The methods employed include KS tests and histogram analysis, we provide some illustrative examples Table 11. Table 2 provide the frequency of method used and we find that they have comparable accuracy. The results underscored the challenges in distinguishing between the heavy tails of Pareto distributions and the thinner tails of Gaussian distributions, particularly in smaller sample sizes. Conditional on guessing the correct parametric distribution type, Figure 4 display the estimated parameters against the true underlying DGP parameters. The standard deviation of the estimates are important and a focus on the shape of the Pareto distribution indicates that some agents anticipated an uninsurable loss distribution, with a shape parameter below 1, this is even more pronounce when eco-anxiety framing is introduced in the second experiment.

TABLE 2
Discrimination method used and Accuracy.

Method Used	Frequency	Accuracy
Histogram analysis	207	0.56
Kolmogorov-Smirnov test	299	0.62
Maximum Likelihood Estimation	13	0.62

Note: The accuracy differences we can observe in our experiment is not going to be driven by the method used.

The construction of the expected contingency table, with expected frequencies under the null hypothesis reported in Table 3, is based on the following assumptions. The data set contains N samples, equally divided between two classes: Gaussian and Pareto. Half of the predictions ($N/2$) are made correctly, distinguishing Gaussian from Pareto samples. This is due to our parametrization of the DGP and our choice of sample size to have only half of the time the p-value of the KS test below the 10% threshold. For the remaining $N/2$, the classifier randomly assigns labels, distributing predictions equally.

TABLE 3
Expected Contingency Table for LLM Classification under Null Hypothesis

Predicted \Actual	Gaussian	Pareto	Total
Gaussian	3N/8	N/8	N/2
Pareto	N/8	3N/8	N/2
Total	N/2	N/2	N

TABLE 4
Observed Contingency Table for LLM Classification

Predicted \Actual	Gaussian	Pareto
Gaussian	139	88
Pareto	121	171

Note: Outcome of our LLM experiment. ChatGPT o1-mini was used with prompting sent by API.

Table 4 presents the observed contingency table of predictions for our baseline experiment with ChatGPT o1-mini. The Pearson chi-squared test yielded a quasi-null p-value, confirming a highly statistically significant relationship between predictions and true DGPs at any conventional significance level. This result provides strong evidence against the null hypothesis of independence, indicating that classifier performance is systematically influenced by the underlying DGP. This is the first and preliminary results of our baseline experiment:

the way we prompted the LLMs enables a discrimination of small samples between Gaussian and Pareto, despite the task being complex and the DGP and sample size chosen for the discrimination to require careful investigations.

To formally assess whether experts exhibit a bias for or against Gaussian or Pareto distributions, we calculate the difference between observed and expected frequencies in each cell of the contingency table. A Chi-Squared Goodness-of-Fit Test is applied, comparing the observed and expected frequencies. The p-value of $p < 0.001$ indicates a significant deviation from the null hypothesis. This suggests systematic biases in expert predictions. To further investigate directional biases, we define a bias ratio:

$$\text{Bias Ratio (Gaussian)} = \frac{\text{Predicted Gaussian Correct}}{\text{Predicted Gaussian Total}} = \frac{139}{139 + 88} \approx 61.3\%$$

$$\text{Bias Ratio (Pareto)} = \frac{\text{Predicted Pareto Correct}}{\text{Predicted Pareto Total}} = \frac{171}{121 + 171} \approx 58.6\%$$

While both ratios are above random assignment levels (50%), the higher accuracy for Gaussian suggests a potential subtle bias towards Gaussian classification. This is as expected by existing literature on representativeness (Tversky and Kahneman, 1974; LaValle, 2006; Knief and Forstmeier, 2021).

Next, we search for evidence of an effect of the expert characteristics on this discrimination capacity and find no evidence, meaning that under the baseline experiment. Table 5 summarizes the classification accuracy, bias, and representation for Farmers and Statisticians compared to the overall population. As intuitively expected, a Farmer with no formal education is more likely to rely on a histogram analysis than a formal KS test than a Statistician with a PhD. Nevertheless, the LLMs agents acting as Farmers are still using a KS test, a proportion no expected in a laboratory experiment with human agents.

The Pearson chi-squared test results indicate a statistically significant relationship between role and classification performance for Farmers ($p = 0.0008$), as well as for Statisticians ($p = 0.0003$). These findings suggest that the role of the expert has no impact on the classification performance for this experiment, this is mainly due to the fact that the Method Used has no significant impact on the classification accuracy.

The Pearson chi-squared test results indicate a statistically significant relationship between gender and classification performance for Males ($p = 0.003$) and Females ($p = 0.002$) are not significant. These findings suggest that gender has no impact on the classification performance.

TABLE 5
Role-based Analysis of Classification Performance and Representation

Role	Representation (%)	Accuracy (%)	Bias (Gaussian as Pareto) (%)	Bias (Pareto as Gaussian) (%)	KS (%)	Hist (%)	MLE (%)
Overall	100.00	59.73	23.31	16.96	57.61	39.88	2.50
Farmer	54.72	58.45	32.75	8.80	40.14	56.34	3.52
Statistician	45.28	61.28	11.91	26.81	78.72	20.00	1.28
Male	31.21	62.35	20.99	16.67	66.67	32.10	1.23
Female	68.79	58.54	24.37	17.09	53.50	43.42	3.08

Next, we analyze the relationship between experience (in years) and classification performance. Table 12 summarizes the mean accuracy across experience ranges. The Pearson correlation coefficient between experience and accuracy (r = -0.01), (p = 0.7458) suggests no significant linear relationship. The ANOVA test results (p-value = 0.7159) confirms no statistically significant differences in classification accuracy across the defined experience groups . Linear regression analysis (Table 13) also confirms no significant relationship between experience and classification accuracy.

3.2 Main experiment with eco-anxiety framing results

Eco-anxiety framing was introduced in our main experiment, using the same samples and the same agents (but framed with eco-anxiety) as the baseline. We find as intuitively expected a bias toward the Pareto classification and a higher share of expectations that the insurance losses are uninsurable.

The overall accuracy of the classifier dropped, while its performance varied significantly between the two distributions. The accuracy for the Gaussian distribution was 17.31%, whereas for the Pareto distribution, it was notably higher at 89.19%. The observed contingency table in Table 6 reveals a significant imbalance in classifier predictions. The Pearson chi-squared test yields a quasi-null p-value, confirming that the null hypothesis of independence is strongly rejected at any conventional significance level. These results highlight that the classifier's performance is systematically influenced by the underlying distribution, even when eco-anxiety is introduced. While the overall accuracy remains consistent, the stark discrepancy in classification accuracy across distributions suggests that eco-anxiety amplifies classification biases, favoring the Pareto distribution over the Gaussian.

Table 7 summarizes the non-influence of role nor gender on the accuracy when eco-anxiety is introduced. What is striking it that the method used are not stable under these conditions, where now a higher proportion of farmers as using the KS test for discrimination compared with the baseline experiment with no eco-anxiety framing. This is an unexpected results but should not impact our observed accuracy nor biases.

We test the credibility of the ranking of two categorical variables, Role and Gender. They

TABLE 6
Observed Contingency Table for LLM Classification with Eco-anxiety

Predicted \Actual	Gaussian	Pareto
Gaussian	45	28
Pareto	215	231

Note: Outcome of our LLM experiment incorporating eco-anxiety. Chat-GPT o-mini was used with prompting sent via API.

TABLE 7
Gender-based Analysis of Classification Performance and Representation

Group	Representation (%)	Accuracy (%)	Bias (Gaussian as Pareto) (%)	Bias (Pareto as Gaussian) (%)	KS (%)	Hist (%)	MLE (%)
Overall	100.00	53.18	41.43	5.39	73.99	18.88	3.85
Farmer	54.72	50.70	43.31	5.99	64.44	28.52	4.23
Statistician	45.28	56.17	39.15	4.68	85.53	7.23	3.40
Male	31.21	48.77	44.44	6.79	74.69	16.05	4.32
Female	68.79	55.18	40.06	4.76	73.67	20.17	3.64

are well separated in our survey outcome. We find no evidence of demographics characteristic impact on the accuracy of the main task in the baseline experiment. We next rely on the eco-anxiety outcome as they present the most bias and economic intuition would be in favor of differentiated eco-anxiety reaction based on exposure to climate change (farmers) or experience (wisdom).

The chi-squared analysis revealed subgroup-specific patterns, with Females ($p = 0.021$) and Statisticians ($p = 0.052$) showing marginally significant associations with classification performance under eco-anxiety framing. However, chi-squared test results indicated no statistically significant relationship for Farmers ($p = 0.441$) or Males ($p = 1.0$). These results suggest that eco-anxiety may disproportionately affect classification performance in specific subgroups. However, when tested jointly, the significance with a logistic regression analysis is absent, Table 8.

We test the inclusion of eco-anxiety and the relationship between experience (in years) and classification performance. Descriptive statistics revealed no notable differences in mean experience between correctly and incorrectly classified samples:

- Correct Classification: Mean = 9.76, Variance = 10.14

TABLE 8
Logistic Regression Analysis of Role and Gender Impact

Variable	Coefficient	Std. Error	z-Value	p-Value
Intercept	0.1024	0.1278	0.8014	0.4229
Role Statistician	0.2680	0.1803	1.4865	0.1372
Gender Male	-0.3059	0.1932	-1.5837	0.1133

- Incorrect Classification: Mean = 9.86, Variance = 8.43

Correlation analysis showed no significant linear relationship between experience and accuracy, with a Pearson correlation coefficient of $r = -0.02$ ($p = 0.7136$). This result was consistent with the findings of the ANOVA test, which yielded an F-statistic of 0.33 ($p = 0.8061$), confirming no significant differences in classification accuracy across defined experience groups. Table 14 presents the regression analysis results, which also suggest no significant relationship between experience and classification accuracy under eco-anxiety conditions.

TABLE 9
Experience Group Statistics with Eco-anxiety

Experience Range (Years)	**Mean Accuracy**	**Count**
$[1, 6)$	0.5652	46
$[6, 11)$	0.5369	298
$[11, 16)$	0.5033	151
$[16, 21)$	0.5833	24

These results confirm that, even under eco-anxiety conditions, experience does not appear to be a significant factor influencing classification performance. In this experiment, age and experience, that we consider a proxy for wisdom, did not differ in how eco-anxiety would influence the discrimination. We run an experiment with more than 500 agents (simulated by LLMs). And yet, we cannot find statistical significance of demographic characteristics on agents judgment of insurance losses. This would need to be tested with human participants on the same task.

We prompted our LLMs to understand why they discrimination task output were impacted by eco-anxiety framing. The observed bias toward Pareto under eco-anxiety framing can be attributed to how the framing emphasizes extreme and unpredictable events, conceptually aligning with the heavy-tailed characteristics of the Pareto distribution. This narrative primes ChatGPT to prioritize distributions that fit the semantic context of extremes, even when the dataset does not strongly support such a conclusion. Additionally, the framing influence methodological interpretation, such as lowering the threshold for accepting a Pareto fit during KS tests or biasing histogram analysis toward emphasizing tail features. As a pattern-matching engine, ChatGPT is further guided by linguistic cues like "accelerating", "extreme", "unpredictable" and "irreversible," which reinforce Pareto-related associations, anchoring its responses within the framing's conceptual domain.

4 Discussion and conclusion

In this paper, we conduct both a baseline and an eco-anxiety framing experiment, where agents classify insurance loss samples as either normal or exhibiting fat tails. Our main finding reveals that, under the baseline scenario, agents correctly distinguish between normal (Gaussian) and extreme (Pareto) distributions, with 56% classifying samples as Pareto instead of the true split, 50%. However, under eco-anxiety, classification shifts dramatically, with 86% of samples identified as extreme. This shift extends beyond classification: even when agents recognize the risk as normal, they estimate a 57% increase in expected loss. Furthermore, while only 1% of agents in the baseline setting deem the risk uninsurable, this figure rises to 36% under eco-anxiety, with an additional 47% perceiving the risk as too volatile to estimate. These findings have significant policy implications. When individuals experience eco-anxiety, a third perceive risks as uninsurable, while those who still consider them insurable anticipate a 50% increase in expected loss. This heightened perception of risk underscores the importance of responsible climate change communication. Rather than downplaying risks, our results advocate for a rational, collective approach that fosters informed action without exacerbating anxiety.

Our secondary finding indicates no statistically significant effect of demographic characteristics on classification performance, either in the baseline scenario or under eco-anxiety framing. This outcome suggests that the observed results are not driven by known biases learn when the LLM was trained. We proxy wisdom with age and experience, but those had no effect when prompting an LLM. If laboratory experiment confirm these results it will confirm the growing literature on using LLMs to behave as human agents in the laboratory. As next steps, our findings must be compared with human behavior to ensure that the absence of demographic effects in LLM-based experiments does not result from limitations in the model's ability to simulate realistic heterogeneity. Future research should investigate the influence of and the best way of framing for LLM simulations and the potential for such models to inform decision-making in complex economic scenarios. When designing prompts for behavioral insurance experiments using LLMs, researchers may prioritize task-specific framing, such as eco-anxiety, over the inclusion of detailed demographic profiles. Future work should address this by incorporating human validation and exploring whether LLM agents can adequately capture the nuanced effects of demographic characteristics in experimental settings.

References

Alevy, Jonathan E, Michael S Haigh, and John A List. 2007. "Information cascades: Evidence from a field experiment with financial market professionals." *The Journal of Finance*, 62(1): 151–180.

Assefa, Samuel A, Danial Dervovic, Mahmoud Mahfouz, Robert E Tillman, Prashant Reddy, and Manuela Veloso. 2020. "Generating synthetic data in finance: opportunities, challenges and pitfalls." 1–8.

Bhargava, Saurabh, George Loewenstein, and Justin Sydnor. 2017. "Choose to lose: Health plan choices from a menu with dominated option." *The Quarterly Journal of Economics*, 132(3): 1319–1372.

Botzen, Wouter JW, Jeroen CJH Aerts, and Jeroen CJM van den Bergh. 2008. "Report on a Survey about Perceptions of Flood Risk, Willingness to Pay for Flood Insurance, and Willingness to Undertake Mitigation Measures: Explanation of the Survey Instrument." *Institute for Environmental Studies (IVM), Vrije Universiteit, Amsterdam.*

Charpentier, Arthur. 2008. "Insurability of climate risks." *The Geneva Papers on Risk and Insurance-Issues and Practice*, 33: 91–109.

Coffey, Yumiko, Navjot Bhullar, Joanne Durkin, Md Shahidul Islam, and Kim Usher. 2021. "Understanding eco-anxiety: A systematic scoping review of current literature and identified knowledge gaps." *The Journal of Climate Change and Health.*

Ditlevsen, Peter, and Susanne Ditlevsen. 2023. "Warning of a forthcoming collapse of the Atlantic meridional overturning circulation." *Nature Communications.*

Giray, Louie. 2023. "Prompt engineering with ChatGPT: a guide for academic writers." *Annals of biomedical engineering*, 51(12): 2629–2633.

Grislain-Letrémy, Céline, Bertrand Villeneuve, and Marc Yeterian. 2024. "Don't bet the Farm on Crop Insurance Subsidies: A Marginal Treatment Effect Analysis of French Farms."

Haigh, Michael S, and John A List. 2005. "Do professional traders exhibit myopic loss aversion? An experimental analysis." *The Journal of Finance*, 60(1): 523–534.

Harrison, Glenn W, and E Elisabet Rutström. 2008. "Risk aversion in the laboratory." In *Risk aversion in experiments.* 41–196. Emerald Group Publishing Limited.

Harrison, Glenn W., and Jia Min Ng. 2019. "Behavioral insurance and economic theory: A literature review." *Risk Management and Insurance Review*, 22(2): 133–182.

Harrison, Glenn W, John A List, and Charles Towe. 2007. "Naturally occurring preferences and exogenous laboratory experiments: A case study of risk aversion." *Econometrica*, 75(2): 433–458.

Hickman, Caroline, Elizabeth Marks, Panu Pihkala, Susan Clayton, R Eric Lewandowski, Elouise E Mayall, Britt Wray, Catriona Mellor, and Lise Van Susteren. 2021. "Climate anxiety in children and young people and their beliefs about government responses to climate change: a global survey." *The Lancet Planetary Health.*

Jaspersen, Johannes G. 2016. "Hypothetical surveys and experimental studies of insurance demand: A review." *Journal of Risk and Insurance*, 83(1): 217–255.

Jojic, Ana, Zhen Wang, and Nebojsa Jojic. 2023. "Gpt is becoming a turing machine: Here are some ways to program it." *arXiv preprint arXiv:2303.14310.*

Knief, Ulrich, and Wolfgang Forstmeier. 2021. "Violating the normality assumption may be the lesser of two evils." *Behavior Research Methods*, 53(6): 2576–2590.

LaValle, Steven M. 2006. *Planning algorithms.* Cambridge university press.

Lenton, Timothy M, Hermann Held, Elmar Kriegler, Jim W Hall, Wolfgang Lucht, Stefan Rahmstorf, and Hans Joachim Schellnhuber. 2008. "Tipping elements in the Earth's climate system." *Proceedings of the national Academy of Sciences.*

Leys, Christophe, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. "Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median." *Journal of experimental social psychology*, 49(4): 764–766.

Pihkala, Panu. 2020. "Anxiety and the ecological crisis: An analysis of eco-anxiety and climate anxiety." *Sustainability.*

Poussin, Jennifer K, WJ Wouter Botzen, and Jeroen CJH Aerts. 2013. "Stimulating flood damage mitigation through insurance: An assessment of the French CatNat system." *Environmental Hazards*, 12(3-4): 258–277.

Raviv, Artur. 1979. "The design of an optimal insurance policy." *The American Economic Review*.

Richter, Andreas, Jörg Schiller, and Harris Schlesinger. 2014. "Behavioral insurance: Theory and experiments." *Journal of Risk and Uncertainty*, 48: 85–96.

Stanley, Samantha K, Teaghan L Hogg, Zoe Leviston, and Iain Walker. 2021. "From anger to action: Differential impacts of eco-anxiety, eco-depression, and eco-anger on climate action and wellbeing." *The Journal of Climate Change and Health*.

Tversky, Amos, and Daniel Kahneman. 1974. "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty." *science*, 185(4157): 1124–1131.

Vansteenberghe, Eric. 2024. "Insurance Supervision under Climate Change: A Pioneers Detection Method." Banque de France.

Vansteenberghe, Eric. 2025. "Prompt Engineering Guidelines for Economic Research."

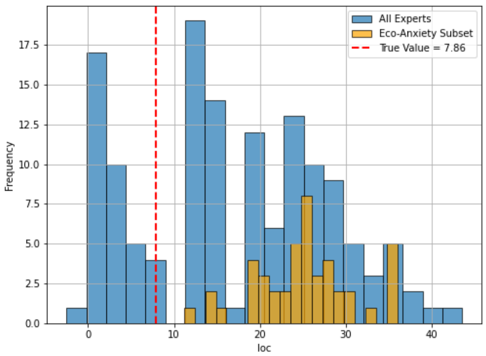
Walonoski, Jason, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. "Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record." *Journal of the American Medical Informatics Association*, 25(3): 230–238.

Wang, Xinyuan, Chenxi Li, Zhen Wang, Fan Bai, Haotian Luo, Jiayou Zhang, Nebojsa Jojic, Eric P Xing, and Zhiting Hu. 2023. "Promptagent: Strategic planning with language models enables expert-level prompt optimization." *arXiv preprint arXiv:2310.16427*.

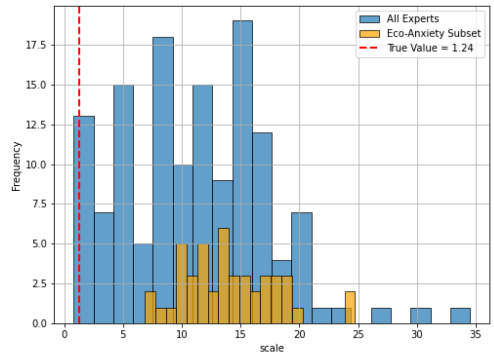
White, Jules, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. "A prompt pattern catalog to enhance prompt engineering with chatgpt." *arXiv preprint arXiv:2302.11382*.

Whitmarsh, Lorraine, Lois Player, Angelica Jiongco, Melissa James, Marc Williams, Elizabeth Marks, and Patrick Kennedy-Williams. 2022. "Climate anxiety: What predicts it and how is it related to climate action?" *Journal of Environmental Psychology.*

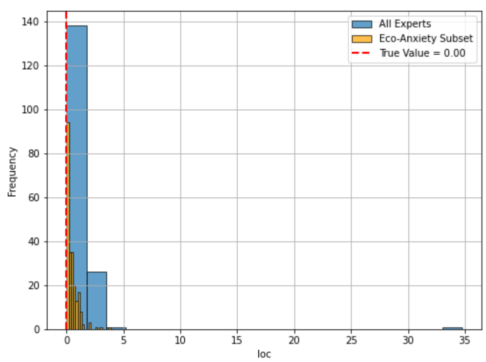
FIGURE 4
Histograms of empirical estimates for each parameter with the true values marked.



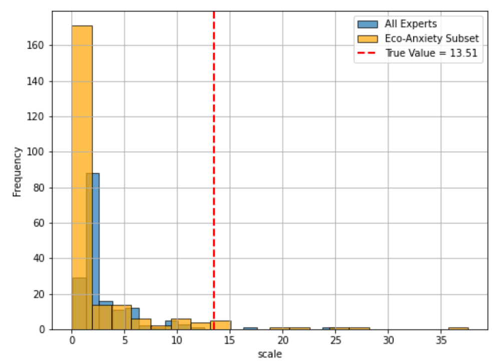
Panel A. Gaussian - loc



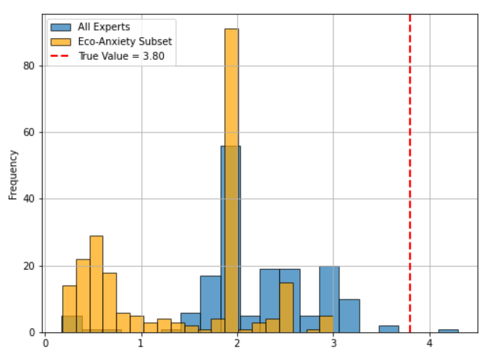
Panel B. Gaussian - scale



Panel C. Pareto - loc

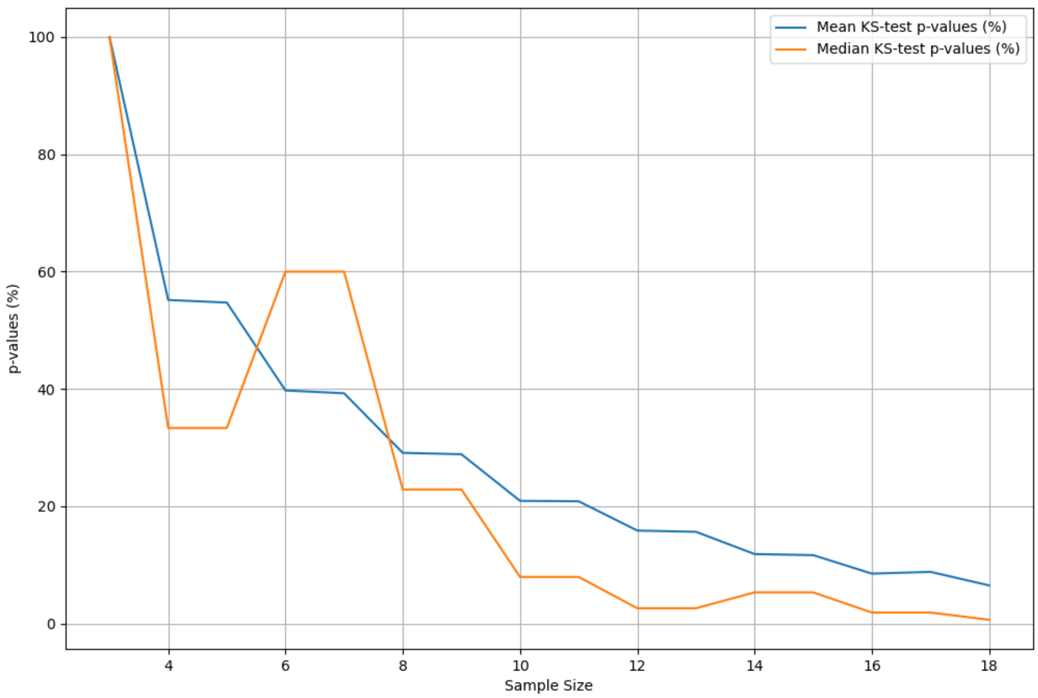


Panel D. Pareto - scale



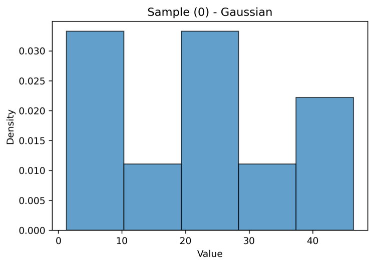
Panel E. Pareto - shape

FIGURE 1
Sample size and p-values

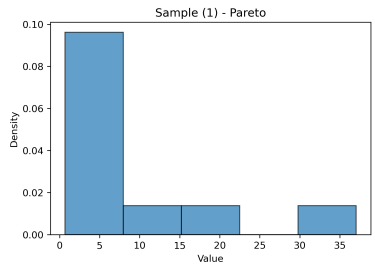


Notes: We conducted 10^6 Monte Carlo simulations to compare samples from Gaussian and Pareto distributions with equivalent expected means and variances. For each simulation, we generated a Gaussian sample truncated above the Pareto threshold and a Pareto sample, ensuring comparable ranges. A two-sample Kolmogorov-Smirnov (KS) test was applied, testing the null hypothesis that the samples were drawn from the same distribution. The null was rejected when the p-value fell below the 10% significance level, indicating sufficient evidence to distinguish between the distributions. The median p-value of the KS test across simulations first crossed the critical threshold when the sample size reached 10. *Source:* author's computation.

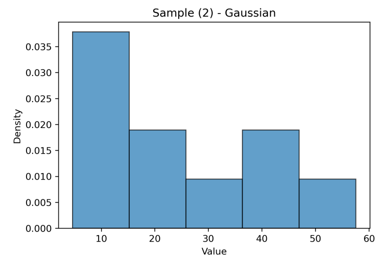
FIGURE 2
Histograms of generated samples



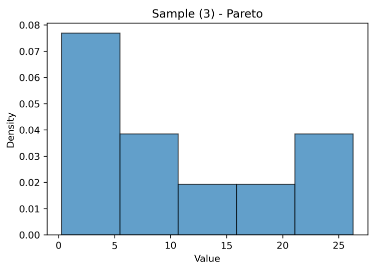
Panel A. Sample (0) - Gaussian



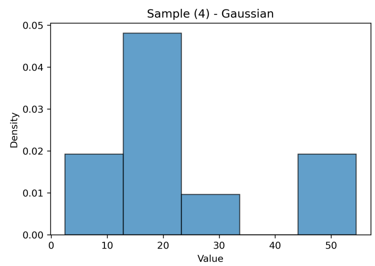
Panel B. Sample (1) - Pareto



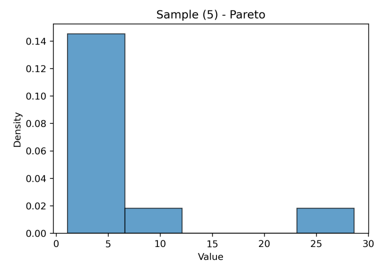
Panel C. Sample (2) - Gaussian



Panel D. Sample (3) - Pareto

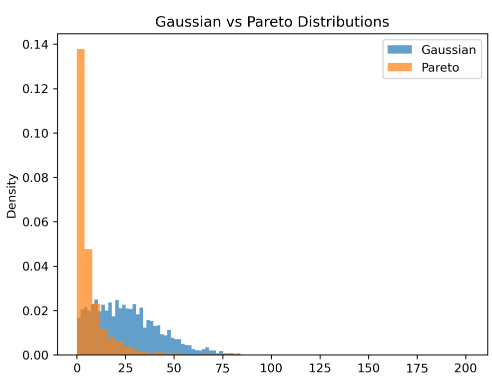


Panel E. Sample (4) - Gaussian

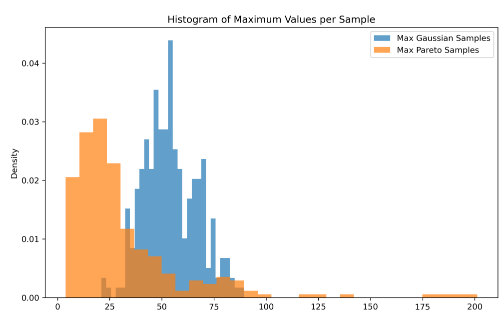


Panel F. Sample (5) - Pareto

FIGURE 3
Histograms of generated samples



Panel A. Empirical histogram of concatenated samples.



Panel B. Empirical histogram of maximum per samples.

TABLE 10
Prompt Engineering Iterations Aligned with Consensus Ranking

Iteration	**Description and Alignment with Consensus Ranking**
1	Began with simple and generic prompts to establish a baseline understanding of the model's behavior. Evaluated outputs for accuracy and relevance, iteratively refining prompts by progressively adding complexity.
2	Explained the general objective of the simulation to clarify the goal of identifying the underlying DGP using demographic and professional attributes.
3	Add "no yapping" in the prompt to reduce stochasticity in longer outputs and enhance clarity.
4	Added specificity to desired outcomes by introducing precise formats and providing examples of outcomes, aligning with few-shot prompting approaches.
5	Assigned specific roles to the LLM, including professional and demographic attributes, to enhance task relevance.

TABLE 11
Profile analysis correlating expert demographics, wrong DGP discrimination are in bold.

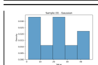
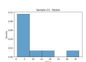
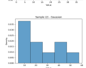
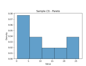
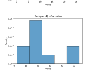
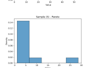
Graphic	loc	scale	shape	Chosen Distribution	Method	Age	Gender	Experience	Role	Correct guess
	2.00	12.10		norm	Kolmogorov-Smirnov test	27	Male	8	Statistician	True
	0.67	1.50	2.00	pareto	Histogram analysis	30	Female	10	Farmer	True
	4.66		2.00	pareto	Histogram analysis	28	Female	5	Farmer	False
	2.50	1.30		norm	Histogram analysis	28	Female	8	Farmer	False
	1.50	2.30		norm	Kolmogorov-Smirnov test	27	Female	5	Farmer	True
	2.50	5.47	2.00	pareto	Histogram analysis	27	Male	7	Farmer	True

TABLE 12
Experience-based Analysis of Classification Performance

Experience Range	Mean Accuracy	Count
$[1, 5]$	0.67	46
$[6, 10]$	0.59	298
$[11, 15]$	0.59	151
$[16, 20]$	0.62	24

TABLE 13
Regression Analysis of Experience and Accuracy

Variable	Coefficient	Std. Error	t-Value	p-Value
Intercept	0.6198	0.0726	8.5385	0.0000
Experience	-0.0023	0.0071	-0.3243	0.7458

TABLE 14
Regression Analysis of Experience and Accuracy with Eco-anxiety

Variable	Coefficient	Std. Error	t-Value	p-Value
Intercept	0.5577	0.0738	7.5516	0.0000
Experience	-0.0026	0.0072	-0.3672	0.7136