

Who Needs Savage Now?

The Effect of LLMs on Decision Making Under Ambiguity

Victor Gonzalez-Jimenez Stephanie Koornneef

Tilburg University

Radboud University

June 12, 2026

Motivation

AI is transforming economic decision making:

- ▶ Investors increasingly rely on AI-generated advice (Bianchi & Briere, 2024; Yang et al., 2025)
- ▶ Managers use AI to evaluate risky projects (Wang et al., 2026)

Motivation

AI is transforming economic decision making:

- ▶ Investors increasingly rely on AI-generated advice (Bianchi & Briere, 2024; Yang et al., 2025)
- ▶ Managers use AI to evaluate risky projects (Wang et al., 2026)

LLMs offer cheap, widely accessible AI-advice

Motivation

AI is transforming economic decision making:

- ▶ Investors increasingly rely on AI-generated advice (Bianchi & Briere, 2024; Yang et al., 2025)
- ▶ Managers use AI to evaluate risky projects (Wang et al., 2026)

LLMs offer cheap, widely accessible AI-advice

Open question: Does LLM advice improve decision making under uncertainty?

Motivation

AI is transforming economic decision making:

- ▶ Investors increasingly rely on AI-generated advice (Bianchi & Briere, 2024; Yang et al., 2025)
- ▶ Managers use AI to evaluate risky projects (Wang et al., 2026)

LLMs offer cheap, widely accessible AI-advice

Open question: Does LLM advice improve decision making under uncertainty?

- ▶ LLMs are trained on human text $\Rightarrow$ may reproduce human biases (Chen et al., 2025)

Motivation

AI is transforming economic decision making:

- ▶ Investors increasingly rely on AI-generated advice (Bianchi & Briere, 2024; Yang et al., 2025)
- ▶ Managers use AI to evaluate risky projects (Wang et al., 2026)

LLMs offer cheap, widely accessible AI-advice

Open question: Does LLM advice improve decision making under uncertainty?

- ▶ LLMs are trained on human text $\Rightarrow$ may reproduce human biases (Chen et al., 2025)
- ▶ But LLMs enable deliberation $\Rightarrow$ known to reduce biases (Baillon et al., 2016)

This Paper

This Paper

1. Study the effect of LLM advice on decision making under **ambiguity** (unknown probabilities) using a controlled laboratory experiment.

This Paper

1. Study the effect of LLM advice on decision making under **ambiguity** (unknown probabilities) using a controlled laboratory experiment.

Why ambiguity?

This Paper

1. Study the effect of LLM advice on decision making under **ambiguity** (unknown probabilities) using a controlled laboratory experiment.

Why ambiguity?

- ▶ Abundant empirical evidence of deviations from ambiguity neutrality (benchmark of rationality) (Trautmann & van de Kuilen, 2015).

Second Motivation

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

- ▶ Sure-Thing Principle (Savage, 1954).

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

- ▶ Sure-Thing Principle (Savage, 1954).
- ▶ Expert advice improves decisions...

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

- ▶ Sure-Thing Principle (Savage, 1954).
- ▶ Expert advice improves decisions...
- ▶ but experts are scarce.

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

- ▶ Sure-Thing Principle (Savage, 1954).
- ▶ Expert advice improves decisions...
- ▶ but experts are scarce.

LLMs trained on Sure-Thing Principle $\Rightarrow$ improves decision making?

Second Motivation

Exposure to a normative economic principle improves decision making under uncertainty (Slovic & Tversky, 1974; Meoskowitz, 1974; MacCrimmon, 1968).

- ▶ Sure-Thing Principle (Savage, 1954).
- ▶ Expert advice improves decisions...
- ▶ but experts are scarce.

LLMs trained on Sure-Thing Principle $\Rightarrow$ improves decision making?

- ▶ LLMs are scalable!

This Paper (cont'd)

1. Study the effect of LLM advice on decision making under **ambiguity** (unknown probabilities) using a controlled laboratory experiment.
2. Study the effect of LLM advice informed on the Sure-Thing Principle.

Outline

1. Measuring Ambiguity Attitudes
2. Theoretical Framework
3. Experimental Design
4. Predictions
5. Agentic Benchmark
6. Results

1. Measuring Ambiguity Attitudes

1. Measuring Ambiguity Attitudes

- ▶ We use **matching probabilities** (Dimmock et al., 2016; Baillon et al., 2018) to measure ambiguity attitudes for natural events.

1. Measuring Ambiguity Attitudes

- ▶ We use **matching probabilities** (Dimmock et al., 2016; Baillon et al., 2018) to measure ambiguity attitudes for natural events.
 - ▶ “Receiving 20 Euros if E (and nothing otherwise) is equivalent to receiving 20 Euros with probability m (and nothing otherwise)”
- ▶ We use choice lists to elicit this preference.

Which option do you prefer?

Option 1	1	2	Option 2
You win €20 if the AEX decreases by less than 0.3% or increases by less than 0.2% (and nothing otherwise)			You win €20 with the following probability (and nothing otherwise)
	☒	☐	0%
	☒	☐	1%
	☒	☐	2%
	☒	☐	5%
	☒	☐	10%
	☒	☐	15%
	☒	☐	20%
	☒	☐	25%
	☒	☐	30%
	☒	☐	35%
	☒	☐	40%
	☒	☐	45%
	☐	☒	50%
	☐	☒	55%
	☐	☒	60%
	☐	☒	65%
	☐	☒	70%
☐	☒	75%	
☐	☒	85%	
☐	☒	100%	

Submit

1. Measuring Ambiguity Attitudes (cont'd)

► $S = \{E_1, E_2, E_3\}$

1. Measuring Ambiguity Attitudes (cont'd)

- ▶ $S = \{E_1, E_2, E_3\}$
 - ▶ E_1 : "the AEX index increases by more than 0.2%"
 - ▶ E_2 : "the AEX index decreases by more than 0.2%"
 - ▶ E_3 : "the AEX index increases by less than 0.2% or decreases by less than 0.2%."

1. Measuring Ambiguity Attitudes (cont'd)

- ▶ $S = \{E_1, E_2, E_3\}$
 - ▶ E_1 : "the AEX index increases by more than 0.2%"
 - ▶ E_2 : "the AEX index decreases by more than 0.2%"
 - ▶ E_3 : "the AEX index increases by less than 0.2% or decreases by less than 0.2%."
- ▶ Use matching probabilities to elicit m_1, m_2 and m_3

1. Measuring Ambiguity Attitudes (cont'd)

- ▶ $S = \{E_1, E_2, E_3\}$
 - ▶ E_1 : "the AEX index increases by more than 0.2%"
 - ▶ E_2 : "the AEX index decreases by more than 0.2%"
 - ▶ E_3 : "the AEX index increases by less than 0.2% or decreases by less than 0.2%."
- ▶ Use matching probabilities to elicit m_1, m_2 and m_3
 - ▶ Also, elicit m_{1c}, m_{2c} and m_{3c} (corresponding to complement events).

1. Measuring Ambiguity Attitudes (cont'd)

- ▶ $S = \{E_1, E_2, E_3\}$
 - ▶ E_1 : "the AEX index increases by more than 0.2%"
 - ▶ E_2 : "the AEX index decreases by more than 0.2%"
 - ▶ E_3 : "the AEX index increases by less than 0.2% or decreases by less than 0.2%."
- ▶ Use matching probabilities to elicit m_1, m_2 and m_3
 - ▶ Also, elicit m_{1c}, m_{2c} and m_{3c} (corresponding to complement events).
- ▶ Define
 - ▶ $\bar{m}_s := \frac{m_1 + m_2 + m_3}{3}$ and $\bar{m}_c = \frac{m_{1c} + m_{2c} + m_{3c}}{3}$.

1. Measuring Ambiguity Attitudes (cont'd)

- ▶ $S = \{E_1, E_2, E_3\}$
 - ▶ E_1 : "the AEX index increases by more than 0.2%"
 - ▶ E_2 : "the AEX index decreases by more than 0.2%"
 - ▶ E_3 : "the AEX index increases by less than 0.2% or decreases by less than 0.2%."
- ▶ Use matching probabilities to elicit m_1, m_2 and m_3
 - ▶ Also, elicit m_{1c}, m_{2c} and m_{3c} (corresponding to complement events).
- ▶ Define
 - ▶ $\bar{m}_s := \frac{m_1+m_2+m_3}{3}$ and $\bar{m}_c = \frac{m_{1c}+m_{2c}+m_{3c}}{3}$.
- ▶ Under Probabilistic Beliefs
 - ▶ $\bar{m}_s = \frac{1}{3}$ and $\bar{m}_c = \frac{2}{3}$.

1. Measuring Ambiguity Attitudes (cont'd)

Two indexes, capturing ambiguity attitudes

***b* index (motivational):**

$$b = 1 - \bar{m}_c - \bar{m}_s$$

1. Measuring Ambiguity Attitudes (cont'd)

Two indexes, capturing ambiguity attitudes

b index (motivational):

$$b = 1 - \bar{m}_c - \bar{m}_s$$

both equal zero at the rational benchmark:

1. Measuring Ambiguity Attitudes (cont'd)

Two indexes, capturing ambiguity attitudes

b index (motivational):

$$b = 1 - \bar{m}_c - \bar{m}_s$$

both equal zero at the rational benchmark:

- ▶ $b > 0$: ambiguity aversion
- ▶ $b < 0$: ambiguity seeking

1. Measuring Ambiguity Attitudes (cont'd)

Two indexes, capturing ambiguity attitudes

b index (motivational):

$$b = 1 - \bar{m}_c - \bar{m}_s$$

a index (cognitive):

$$a = 3\left(\frac{1}{3} - \bar{m}_c + \bar{m}_s\right)$$

both equal zero at the rational benchmark:

- ▶ $b > 0$: ambiguity aversion
- ▶ $b < 0$: ambiguity seeking

1. Measuring Ambiguity Attitudes (cont'd)

Two indexes, capturing ambiguity attitudes

b index (motivational):

$$b = 1 - \bar{m}_c - \bar{m}_s$$

both equal zero at the rational benchmark:

- ▶ $b > 0$: ambiguity aversion
- ▶ $b < 0$: ambiguity seeking

a index (cognitive):

$$a = 3\left(\frac{1}{3} - \bar{m}_c + \bar{m}_s\right)$$

- ▶ $a > 0$: a -insensitivity (fails to discriminate likelihoods)

Advantage of this method: free from utility curvature and probability weighting complications.

2. Theoretical Framework

Assumptions:

- ▶ A1: $\succeq$ over bets is weak-order and monotonic.
- ▶ A2: Sure-thing principle.

2. Theoretical Framework

Assumptions:

- ▶ A1: $\succeq$ over bets is weak-order and monotonic.
- ▶ A2: Sure-thing principle.

Lemma 1 (Complement Event Additivity)

If “AEX rises” feels 60% likely, then “AEX does not rise” must feel 40% likely.

2. Theoretical Framework

Assumptions:

- ▶ A1: $\succeq$ over bets is weak-order and monotonic.
- ▶ A2: Sure-thing principle.

Lemma 1 (Complement Event Additivity)

If “AEX rises” feels 60% likely, then “AEX does not rise” must feel 40% likely.

Lemma 2 (Composite Event Additivity)

Under A1 and A2: $m_i + m_j = m_k^c$ with $i = \{1, 2, 3\}$ and $i \neq j \neq k$

If “rises” = 60% and “flat” = 20%, then “rises or flat” = 80%.

2.Theoretical Framework (cont'd)

Proposition 1

Lemma 1 and Lemma 2 imply $m_1 + m_2 + m_3 = 1$ (ambiguity neutrality)

2.Theoretical Framework (cont'd)

Proposition 1

Lemma 1 and Lemma 2 imply $m_1 + m_2 + m_3 = 1$ (ambiguity neutrality)

- In the matching probabilities setting only A1 and A2 are needed for ambiguity neutrality.

2.Theoretical Framework (cont'd)

Proposition 1

Lemma 1 and Lemma 2 imply $m_1 + m_2 + m_3 = 1$ (ambiguity neutrality)

- ▶ In the matching probabilities setting only A1 and A2 are needed for ambiguity neutrality.
- ▶ The implications of Lemma 1 and Lemma 2 are in-between steps to achieve probabilistic beliefs.

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

Between-Subjects: 2×2 factorial design in Part 2.

	No LLM	LLM Access
No Theory	Control	LLM-Free
Theory (STP)	P2-Explained	LLM-P2

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

Between-Subjects: 2×2 factorial design in Part 2.

	No LLM	LLM Access
No Theory	Control	LLM-Free
Theory (STP)	P2-Explained	LLM-P2

- ▶ **LLM-Free:** ChatGPT-4o mini, no theoretical guidance

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

Between-Subjects: 2×2 factorial design in Part 2.

	No LLM	LLM Access
No Theory	Control	LLM-Free
Theory (STP)	P2-Explained	LLM-P2

- ▶ **LLM-Free:** ChatGPT-4o mini, no theoretical guidance
- ▶ **P2-Explained:** Written explanation of Savage's STP and implications (Lemmas)

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

Between-Subjects: 2×2 factorial design in Part 2.

	No LLM	LLM Access
No Theory	Control	LLM-Free
Theory (STP)	P2-Explained	LLM-P2

- ▶ **LLM-Free:** ChatGPT-4o mini, no theoretical guidance
- ▶ **P2-Explained:** Written explanation of Savage's STP and implications (Lemmas)
- ▶ **LLM-P2:** ChatGPT-4o mini trained on the same STP content

3. Experimental Design

Within-Subjects: measurement of ambiguity attitudes in Part 1 and Part 2

Between-Subjects: 2×2 factorial design in Part 2.

	No LLM	LLM Access
No Theory	Control	LLM-Free
Theory (STP)	P2-Explained	LLM-P2

- ▶ **LLM-Free:** ChatGPT-4o mini, no theoretical guidance
- ▶ **P2-Explained:** Written explanation of Savage's STP and implications (Lemmas)
- ▶ **LLM-P2:** ChatGPT-4o mini trained on the same STP content
- ▶ Theoretical content is *identical* across P2-Explained and LLM-P2

3. Experimental Design - Prompts

3. Experimental Design - Prompts

LLM-Free

You are a chatbot assisting participants making decisions in an experiment **Core instructions:**

- ▶ Initial message: only ask if the participant wants help; no other information
- ▶ Respectful, everyday language. *Do not mention theory, axioms, or P2*
- ▶ Responses $\leq$ 350 characters, but substantive

3. Experimental Design - Prompts

LLM-Free

You are a chatbot assisting participants making decisions in an experiment **Core instructions:**

- ▶ Initial message: only ask if the participant wants help; no other information
- ▶ Respectful, everyday language. *Do not mention theory, axioms, or P2*
- ▶ Responses $\leq$ 350 characters, but substantive

LLM-P2

Same core instructions as LLM-Free, plus:

Task: Explain Savage's STP and help participants make choices consistent with it. Guide them through **four steps** (one per message):

1. Focus only on events where the two options differ
2. Pick a probability m for the market event under consideration; compare to each p
3. For **complementary events**: use $1 - m$ (e.g. if "AEX rises" = 60%, "AEX does not rise" = 40%)
4. For **combined events**: sum probabilities (e.g. "rises" 60% + "flat" 20% = "rises or flat" 80%)

Key contrast: LLM-P2 embeds the same theoretical content as P2-Explained, but delivered interactively through a rigid 4-step protocol.

4. Predictions

4. Predictions

Theory generates three predictions:

4. Predictions

Theory generates three predictions:

Prediction	Condition	Expected
P1	Lemma 1 only	$b \rightarrow 0, a > 0$

4. Predictions

Theory generates three predictions:

Prediction	Condition	Expected
P1	Lemma 1 only	$b \rightarrow 0, a > 0$
P2	Lemma 2 only	$a = b \neq 0$

4. Predictions

Theory generates three predictions:

Prediction	Condition	Expected
P1	Lemma 1 only	$b \rightarrow 0, a > 0$
P2	Lemma 2 only	$a = b \neq 0$
P3	Both Lemmas	$a = b = 0$

5. AI Agent Benchmark

To benchmark human behavior, we deployed **1,000 autonomous GPT-4o agents** through the experiment.

5. AI Agent Benchmark

To benchmark human behavior, we deployed **1,000 autonomous GPT-4o agents** through the experiment.

	$\bar{m}_s$	$\bar{m}_c$	b	a
Mean	0.417	0.725	-0.142	0.076
SD	0.063	0.043	0.097	0.141
Median	0.400	0.700	-0.100	0.100

5. AI Agent Benchmark

To benchmark human behavior, we deployed **1,000 autonomous GPT-4o agents** through the experiment.

	$\bar{m}_s$	$\bar{m}_c$	b	a
Mean	0.417	0.725	-0.142	0.076
SD	0.063	0.043	0.097	0.141
Median	0.400	0.700	-0.100	0.100

$b < 0$: AI agents are **ambiguity seeking**
($p < 0.001$) — mirroring previous findings
on humans

5. AI Agent Benchmark

To benchmark human behavior, we deployed **1,000 autonomous GPT-4o agents** through the experiment.

	$\bar{m}_s$	$\bar{m}_c$	b	a
Mean	0.417	0.725	-0.142	0.076
SD	0.063	0.043	0.097	0.141
Median	0.400	0.700	-0.100	0.100

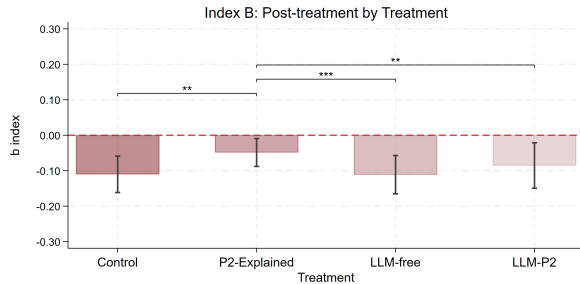
$b < 0$: AI agents are **ambiguity seeking**
($p < 0.001$) — mirroring previous findings
on humans

$a > 0$: AI agents exhibit **a-insensitivity**
($p < 0.001$) — also mirroring human
patterns

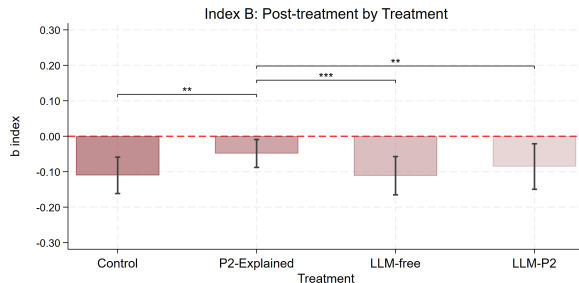
LLMs reproduce the same systematic deviations from ambiguity neutrality observed in humans.

6. Main Results

6. Main Results



6. Main Results

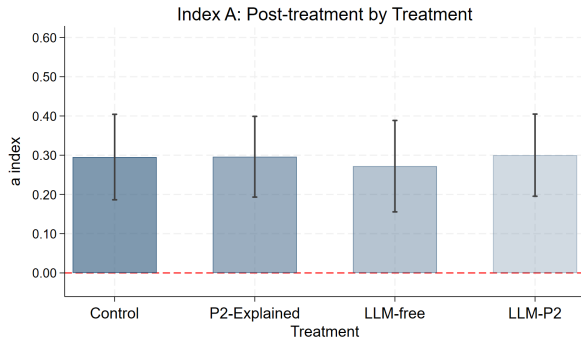


Close to Prediction 1:

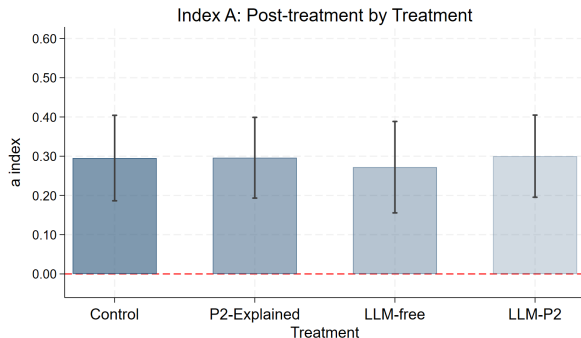
Subjects in P2-Explained follow only **complement additivity** (Lemma 1), not composite event additivity (Lemma 2). This moves $b \rightarrow 0$ while leaving a unchanged — exactly Prediction 1.

6. Main Results

6. Main Results



6. Main Results



Close to Prediction 1:

Subjects in P2-Explained follow only **complement additivity** (Lemma 1), not composite event additivity (Lemma 2). This moves $b \rightarrow 0$ while leaving a unchanged — exactly Prediction 1.

6. Results - Regression

	(1) <i>b</i>	(2) <i>b</i>	(3) <i>a</i>	(4) <i>a</i>
Part 2 × P2-Explained	0.038 (0.034)	0.038 (0.034)	−0.023 (0.078)	−0.023 (0.078)
Part 2 × LLM-Free	−0.013 (0.025)	−0.013 (0.025)	0.068 (0.074)	0.068 (0.075)
Part 2 × LLM-P2	−0.050** (0.024)	−0.050** (0.024)	0.003 (0.068)	0.003 (0.068)
Part 2	0.026 (0.019)	0.026 (0.019)	−0.107** (0.053)	−0.107** (0.053)
P2-Explained	0.023 (0.041)	0.029 (0.040)	0.024 (0.066)	0.017 (0.066)
LLM-Free	0.006 (0.040)	0.001 (0.040)	−0.113* (0.067)	−0.107 (0.066)
LLM-P2	0.066 (0.042)	0.064 (0.041)	−0.030 (0.069)	−0.031 (0.067)
Controls	No	Yes	No	Yes
Constant	−0.134*** (0.031)	−0.060 (0.039)	0.410*** (0.046)	0.348*** (0.060)
R^2	0.016	0.068	0.021	0.053
N	542	542	542	542
<i>Wald tests (col. 2 / col. 4)</i>				
P2-Expl.: within-arm	$F = 5.16^{**}, p = 0.024$		n.s.	
P2-Expl. vs. Control	$F = 4.40^{**}, p = 0.037$		n.s.	
LLM-P2 vs. P2-Expl.	$F = 3.92^{**}, p = 0.049$		n.s.	
LLM-P2 vs. LLM-Free	$F = 4.10^{**}, p = 0.043$		n.s.	

SE in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. ■ P2-Explained ■ LLM-P2 ■ Benchmark

6. Results - Regression

	(1) <i>b</i>	(2) <i>b</i>	(3) <i>a</i>	(4) <i>a</i>
Part 2 × P2-Explained	0.038 (0.034)	0.038 (0.034)	−0.023 (0.078)	−0.023 (0.078)
Part 2 × LLM-Free	−0.013 (0.025)	−0.013 (0.025)	0.068 (0.074)	0.068 (0.075)
Part 2 × LLM-P2	−0.050** (0.024)	−0.050** (0.024)	0.003 (0.068)	0.003 (0.068)
Part 2	0.026 (0.019)	0.026 (0.019)	−0.107** (0.053)	−0.107** (0.053)
P2-Explained	0.023 (0.041)	0.029 (0.040)	0.024 (0.066)	0.017 (0.066)
LLM-Free	0.006 (0.040)	0.001 (0.040)	−0.113* (0.067)	−0.107 (0.066)
LLM-P2	0.066 (0.042)	0.064 (0.041)	−0.030 (0.069)	−0.031 (0.067)
Controls	No	Yes	No	Yes
Constant	−0.134*** (0.031)	−0.060 (0.039)	0.410*** (0.046)	0.348*** (0.060)
R^2	0.016	0.068	0.021	0.053
N	542	542	542	542
<i>Wald tests (col. 2 / col. 4)</i>				
P2-Expl.: within-arm	$F = 5.16^{**}, p = 0.024$		n.s.	
P2-Expl. vs. Control	$F = 4.40^{**}, p = 0.037$		n.s.	
LLM-P2 vs. P2-Expl.	$F = 3.92^{**}, p = 0.049$		n.s.	
LLM-P2 vs. LLM-Free	$F = 4.10^{**}, p = 0.043$		n.s.	

Result 1

b: P2-Explained ↑, LLM-P2 ↓,
LLM-Free n.s.

a: no treatment effect

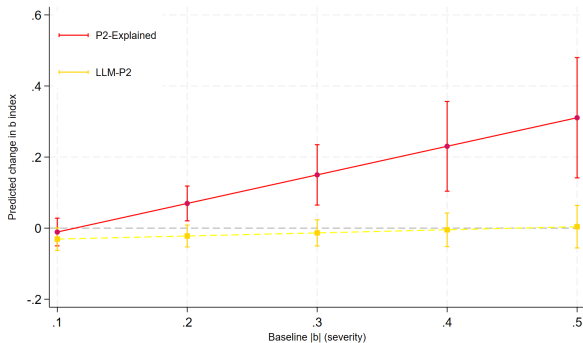
Prediction 1 confirmed:

Subjects follow **complement
additivity** (Lemma 1) only.

SE in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. ■ P2-Explained ■ LLM-P2 ■ Benchmark

6. Results - Heterogeneous Treatment Effects

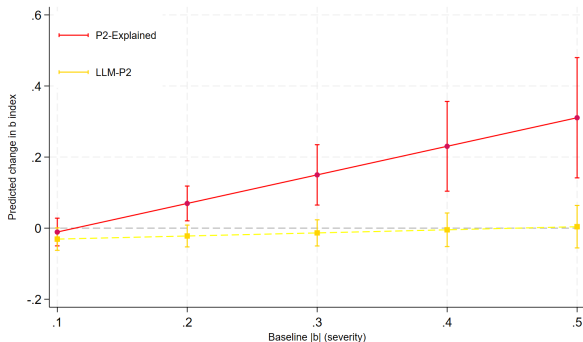
6. Results - Heterogeneous Treatment Effects



Result 2:

- P2-Explained: effect *grows* with b in Part 1
 - At $|b| = 0.4$: +0.23 ($p < 0.001$)

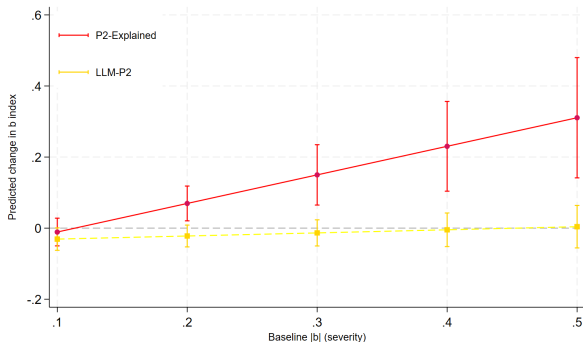
6. Results - Heterogeneous Treatment Effects



Result 2:

- ▶ P2-Explained: effect *grows* with b in Part 1
 - ▶ At $|b| = 0.4$: $+0.23$ ($p < 0.001$)
- ▶ LLM-P2: harms mildly biased participants most; effect attenuates at high severity

6. Results - Heterogeneous Treatment Effects



Result 2:

- ▶ P2-Explained: effect *grows* with b in Part 1
 - ▶ At $|b| = 0.4$: +0.23 ($p < 0.001$)
- ▶ LLM-P2: harms mildly biased participants most; effect attenuates at high severity
- ▶ LLM-P2 corrects *least* where correction matters most

6. Results -Heterogeneous Effects (cont'd)

	(1) <i>b</i> index	(2) <i>a</i> index
Part 2 × P2-Explained × Distance	0.615** (0.300)	−0.238 (0.240)
Part 2 × LLM-Free × Distance	−0.070 (0.251)	0.096 (0.214)
Part 2 × LLM-P2 × Distance	−0.088 (0.199)	0.061 (0.240)
Part 2 × Distance	0.173 (0.185)	−0.416*** (0.158)
Part 2 × P2-Explained	−0.074 (0.051)	0.098 (0.124)
Part 2 × LLM-Free	0.004 (0.050)	−0.002 (0.122)
Part 2 × LLM-P2	−0.028 (0.040)	−0.020 (0.122)
Part 2	−0.011 (0.034)	0.085 (0.087)
P2-Explained × Distance	−0.303 (0.364)	−0.010 (0.074)
LLM-Free × Distance	−0.206 (0.372)	−0.046 (0.087)
LLM-P2 × Distance	0.643 (0.474)	−0.044 (0.069)
Distance	−0.479 (0.313)	1.102*** (0.049)
P2-Explained	0.082 (0.057)	0.016 (0.054)
LLM-Free	0.040 (0.060)	0.000 (0.054)
LLM-P2	−0.062 (0.069)	−0.024 (0.053)
Constant	0.040 (0.053)	−0.125* (0.067)
R^2	0.252	0.500
N	542	542

Distance = $|b_{\text{part } 1}|$ (baseline severity). Standard errors in parentheses. * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

6. Results- NLP Methodology

6. Results- NLP Methodology

We tokenize and analyze all chats

What is a token? A token is the basic unit of text “ I want to follow Savage” has 5 tokens

6. Results- NLP Methodology

We tokenize and analyze all chats

What is a token? A token is the basic unit of text “ I want to follow Savage” has 5 tokens **Standard NLP measures** capture two dimensions of message quality:

Informativeness

- ▶ **Entropy:** diversity of vocabulary
- ▶ **Density:** ratio of content words to total words

Complexity

- ▶ **Prop. polysyllabic:** share of words with 3+ syllables
- ▶ **FK grade:** Flesch–Kincaid readability grade level

6. Results - Chat Analysis

	Participants		Bots	
	LLM-Free	LLM-P2	LLM-Free	LLM-P2
N messages	6.603 (3.938)	5.861 (3.589)	—	—
Tokens	11.418 (5.126)	9.833 (5.675)	44.057 (10.195)	48.610 (12.752)
Entropy	2.955 (0.672)	2.495 (0.936)	5.047 (0.257)	5.009 (0.289)
Density	0.342 (0.103)	0.279 (0.135)	0.367 (0.053)	0.316 (0.044)
Prop. polysyllabic	0.148 (0.092)	0.084 (0.059)	0.182 (0.030)	0.136 (0.028)
FK grade	6.053 (3.011)	3.863 (2.752)	9.394 (1.181)	7.790 (1.417)

Standard deviations in parentheses. Highlighted = significant difference LLM-Free vs. LLM-P2 ($p < 0.05$).

Participants: entropy $p = 0.002$, density $p = 0.004$, polysyllabic $p = 0.001$, FK grade $p = 0.001$. Bots: tokens $p = 0.029$, density/polysyllabic/FK grade $p = 0.001$.

Result 3 (part i)

LLM-Free participants write more complex and informative messages than LLM-P2 participants.

6. Results - Chat Analysis

Named Entity Recognition (NER) classifies words into semantic categories: Confusion, Decision, Market, Money, Probability

6. Results - Chat Analysis

Named Entity Recognition (NER) classifies words into semantic categories: Confusion, Decision, Market, Money, Probability

	Category (avg. references per message)				
	Confusion	Decision	Market	Money	Probability
<i>Participants</i>					
LLM-Free	0.026 (0.059)	0.374 (0.530)	0.684 (0.460)	0.377 (0.616)	0.349 (0.384)
LLM-P2	0.066 (0.122)	0.341 (0.515)	0.432 (0.385)	0.288 (0.593)	0.289 (0.308)
<i>p</i> -value	0.022	0.721	0.001	0.405	0.317
<i>Bots</i>					
LLM-Free	0.850 (0.381)	1.107 (1.479)	1.284 (0.870)	0.383 (0.549)	0.848 (0.902)
LLM-P2	0.872 (0.472)	2.492 (0.985)	0.623 (0.628)	0.231 (0.386)	1.328 (1.028)
<i>p</i> -value	0.770	0.001	0.001	0.067	0.006

6. Results - Chat Analysis

Named Entity Recognition (NER) classifies words into semantic categories: Confusion, Decision, Market, Money, Probability

	Category (avg. references per message)				
	Confusion	Decision	Market	Money	Probability
<i>Participants</i>					
LLM-Free	0.026 (0.059)	0.374 (0.530)	0.684 (0.460)	0.377 (0.616)	0.349 (0.384)
LLM-P2	0.066 (0.122)	0.341 (0.515)	0.432 (0.385)	0.288 (0.593)	0.289 (0.308)
<i>p</i> -value	0.022	0.721	0.001	0.405	0.317
<i>Bots</i>					
LLM-Free	0.850 (0.381)	1.107 (1.479)	1.284 (0.870)	0.383 (0.549)	0.848 (0.902)
LLM-P2	0.872 (0.472)	2.492 (0.985)	0.623 (0.628)	0.231 (0.386)	1.328 (1.028)
<i>p</i> -value	0.770	0.001	0.001	0.067	0.006

Standard deviations in parentheses. Highlighted = significant difference LLM-Free vs. LLM-P2 ($p < 0.05$).

Participants: Confusion $p = 0.022$, Market $p = 0.001$. Bots: Decision $p = 0.001$, Market $p = 0.001$, Probability $p = 0.006$.

Result 3 (part ii)

LLM-Free participants discuss market events more and express less confusion. The LLM-P2 bot uses more decision- and probability-related language — but P2-Explained contains more of all of it than either bot.

6. Results - Written Instruction

The same NLP analysis applied to the **instruction** given to P2-Explained participants.

Interpretation

If even reading a small fraction of the instruction delivers more relevant content than a full LLM-P2 session, the failure of LLM-P2 is a failure of *delivery*, not of content.

6. Results - Written Instruction

The same NLP analysis applied to the **instruction** given to P2-Explained participants.

Interpretation

If even reading a small fraction of the instruction delivers more relevant content than a full LLM-P2 session, the failure of LLM-P2 is a failure of *delivery*, not of content.

Result 4: Reading only **20%** of the written instruction delivers as much market-event language as an entire LLM-P2 session (47% for decision language).

Why LLM-P2 Backfires?

Why LLM-P2 Backfires?

Coherent mechanism (Result 5):

1. Severely biased participants naturally want to discuss the market (as in LLM-Free)
2. LLM-P2 imposes a rigid 4-step protocol, refusing natural conversation
3. Participants disengage early, before reaching the helpful content
4. Mildly biased participants receive *incomplete* probability guidance $\Rightarrow$ misleading heuristics

The failure of LLM-P2 is not a failure of *content* but a failure of *delivery*.

Conclusion

Contributions:

- ▶ First paper to study LLM advice and ambiguity *preferences* (not outcomes)
- ▶ Written delivery of STP implications can reduce the motivational bias
- ▶ Theory-informed LLMs can backfire by suppressing natural engagement

Future work: Adaptive protocols matching participants' reasoning style; agentic LLMs that set rational defaults without requiring active engagement.