Minds Like Ours? Priming LLMs for Behavioral Alignment: Evidence from the Preference Survey Module

Prepared for ESA 2026 Conference, Rabat

Wanda Mimra and Pablo Winant *

2026-06-18

* ESCP Business School

Introduction

Rising Use of Large Language Models (LLM) in Decision Support

Finance

Application

Analyzing market trends and making investment recommendations.

Example

Robo-advisors like Betterment and Wealthfront use LLMs for portfolio management.

Human Resources

Application

Screening resumes and matching candidates to job descriptions.

Example

Tools like HireVue use LLMs to evaluate and rank job applicants.

Marketing

Application

Personalizing campaigns and predicting customer preferences.

Example

LLMs analyze customer data to create targeted advertising strategies.

Pre-requisite for Effective Use: Alignment with Human Preferences

Misalignment Risks

Deployment Context

Increasing integration of LLMs in decision support (finance, HR, public policy).

Divergence Risk

Suboptimal or ethically problematic choices if underlying LLM preference profiles diverge from humans.

Core Research Challenges

1. Measurement

How can we reliably elicit, measure, and validate the inherent preference profiles of LLMs?

2. Alignment & Steering

Can we effectively steer LLM behavior in human-predicted directions using lightweight prompting tools?

Literature Review

LLMs as economic agents

- rationality and consistency
- games and social preferences
- risk and time preferences

Binz and Schulz 2023 Agarwal et al. 2023 Schulz et al. 2023 Horton 2023 Dell and Sloane 2023 Goli and Singh 2023 Mei et al. 2024 Liu et al. 2025 Ou et al. 2025 Chen et al. 2025

LLMs as synthetic survey respondents

- synthetic human samples
- prompt and answer-format effects
- survey-response instability
- compressed heterogeneity

Argyle et al. 2023 Dominguez-Olmedo et al. 2024 Salecha et al. 2024 Ball et al. 2025

Preference alignment and steering

- human feedback
- preference optimization
- structured prompting
- value and fairness alignment

Askell et al. 2021 Bai et al. 2022 Ouyang et al. 2022 MacKay et al. 2023 Rafailov et al. 2023 Hosseini and Khanna 2025 Leng et al. 2026

- **Our contribution:** use a validated economic preference module to test both *measurement* and *steering* of LLM behavior.

blue = economics / management / behavioral experiments · gray = non-economics background · † = recent working paper

This Paper: Eliciting & Steering Preferences

We use the validated **Preference Survey Module (PSM)** (Falk et al., 2023) instead of complex experimental games.

- Elicits preferences via simple natural language questions.
- Enables direct comparison with large-scale human baseline datasets.

We design systematic experiments to study:

1. Measurement — Do LLM preference distributions align with human distributions? (Baseline)

2. Active Alignment — Can constructed self-assessments steer choices in human-predicted directions? (Steering)

The experiment

Why Survey Elicitation?

Challenge

Incentivized choice experiments are the gold standard for measuring economic preferences, but they are highly costly.

Solution (Falk et al., 2023)

A low-cost survey-based alternative carefully selected for its ability to predict choices in incentivized settings.

Preference Dimensions

Risk Aversion, Time Discounting, Trust, Altruism, Positive Reciprocity, Negative Reciprocity.

Elicitation Item Types

Self-Assessment

Subjective reflection on willingness to act/behave (e.g. general risk willingness, altruistic self-image).

Hypothetical Games

Choices in hypothetical scenarios mimicking real-life decisions:

- Single response games (e.g. lottery donation)
- Decision lists (e.g. price lists)

The Preference Module

Preference	Item Description	Type	Weights
Risk Taking	Multiple price list (31 choices between a lottery and a safe option).	Hypothetical Game (PL)	0.2758
	Are you a person who is generally willing to take risks?“	Self Assessment	0.2034
Time Discounting	List of 25 choices between payment “today” and “in 12 months”.	Hypothetical Game (PL)	0.4849
	Are you willing to give up something today to benefit in the future?	Self Assessment	−0.1712
Trust	Investment game: first mover behavior.	Hypothetical Game	0.6289
	I assume that people have only the best intentions.	Self-assessment	0.1331

The Preference Module

Preference	Item Description	Type	Weights
Altruism	You won 1,000 euros. How much would you donate to charity?	Hypothetical Game	0.1845
	How do you assess your willingness to share with others?	Self Assessment	0.3210
Positive Reciprocity	Investment: second mover behavior.	Hypothetical Game	0.4857
	Which bottle of wine do you give as a thank-you gift?	Hypothetical Game	0.1640
Negative Reciprocity	Minimum acceptable offer in ultimatum game.	Hypothetical Game	0.3284
	Are you willing to punish unfair behavior even if costly?	Self Assessment	0.1479

Experimental Design: Baseline

Setup & Models

Models Evaluated
GPT-4, GPT-4o, GPT-OSS, Gemini, Claude, Mistral, Deepseek

Sampling Size
50 independent trials per experimental treatment

Data Quality Control
Record all chats; limit API calls to avoid over-fitting and session state leakage

Elicitation Treatments

Baseline Prompt
"Assuming you are a survey participant, please answer the following question."

Chain-of-thought (CoT)
Baseline + *"Provide a short explanation before answering the question."*

Variations
Administered in both English and German to test cross-lingual robustness

Experimental Design: Alignment

Models: GPT-4 family (GPT-4, GPT-4o, GPT-OSS) · **Sampling:** 50 trials per treatment · **Language:** English

1. PairsHistory

Spontaneous structural consistency

Protocol
Ask two related preference questions successively (keep history) in a single chat.

Control
Invert the order of question pairs to check for order effects.

Objective
Evaluate if self-assessment answers naturally predict decisions without external steering.

2. ALIGN (Priming)

Active steering via history

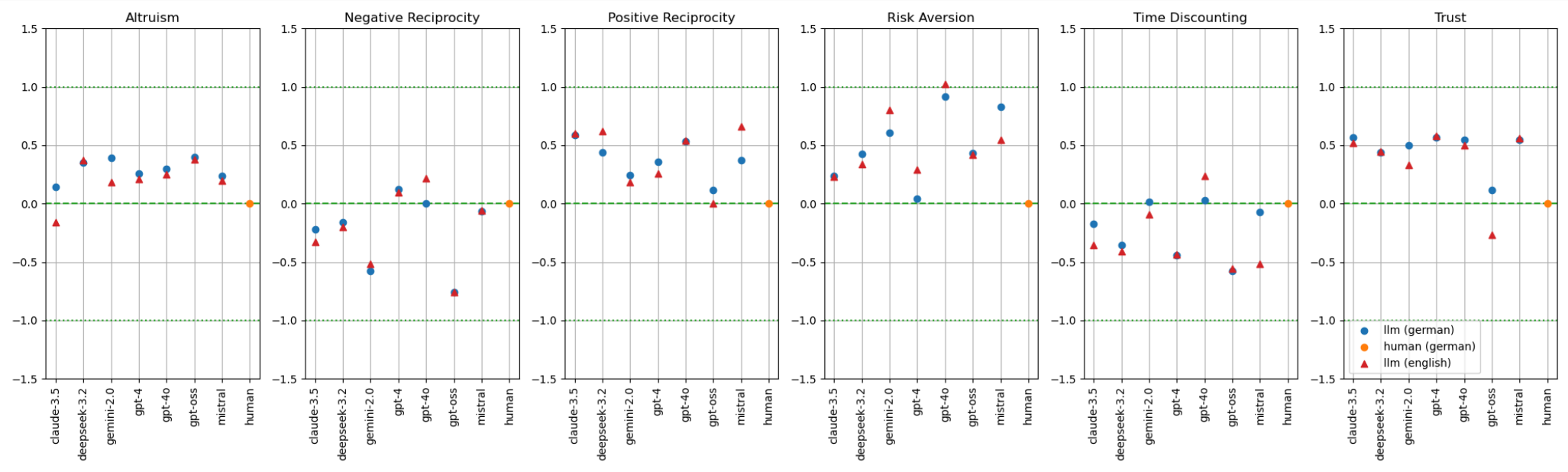
Protocol
Construct interaction histories where the model is presented with self-assessment responses varying across the full scale (1–10).

Measurement
Query the decision item conditional on this constructed history.

Objective
Evaluate if priming steers choices in human-predicted directions.

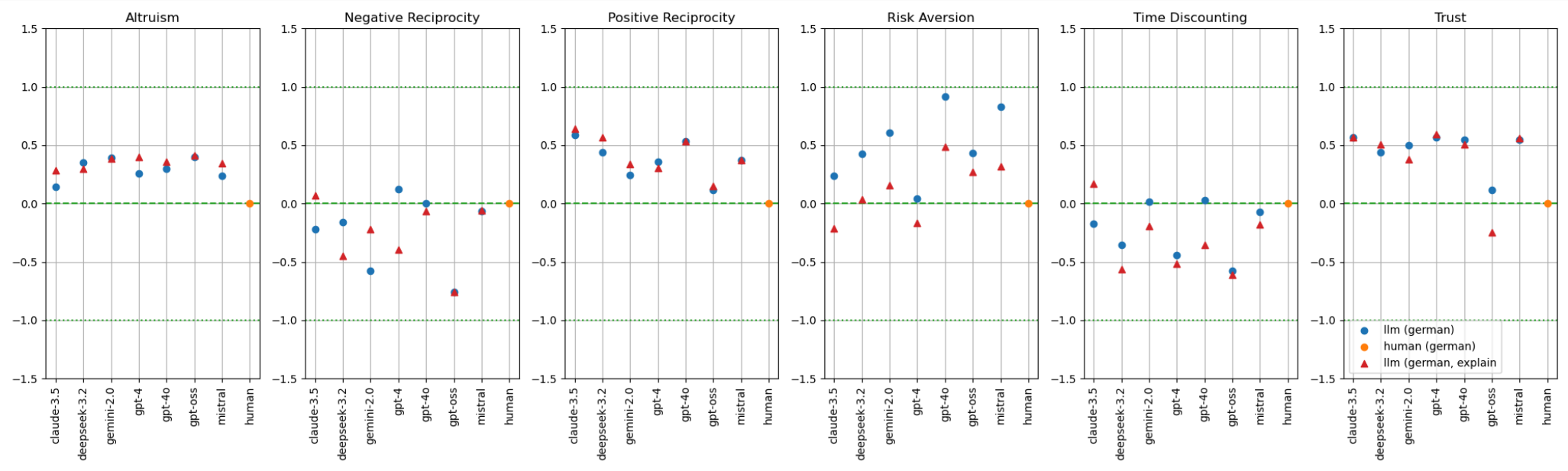
Results

Responses to the PSM: Baseline English/German



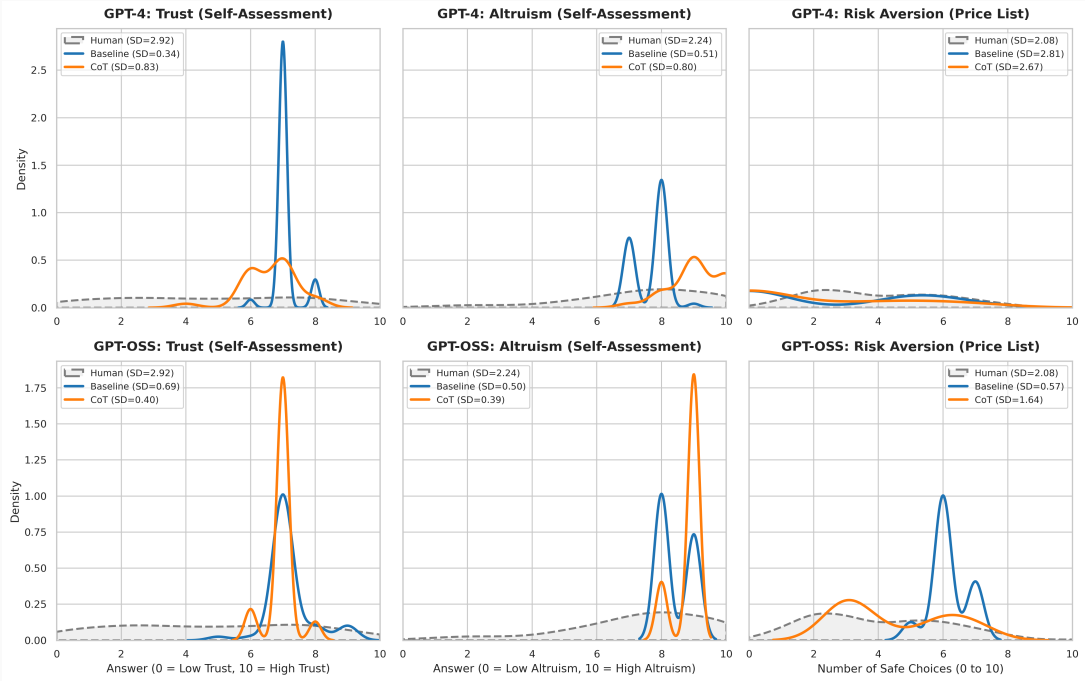
- We use PSM weights and normalization.
- For most preference items LLM responses fall within the range of human responses
- Responses are fairly similar across models

Responses to the PSM: Chain-of-thought



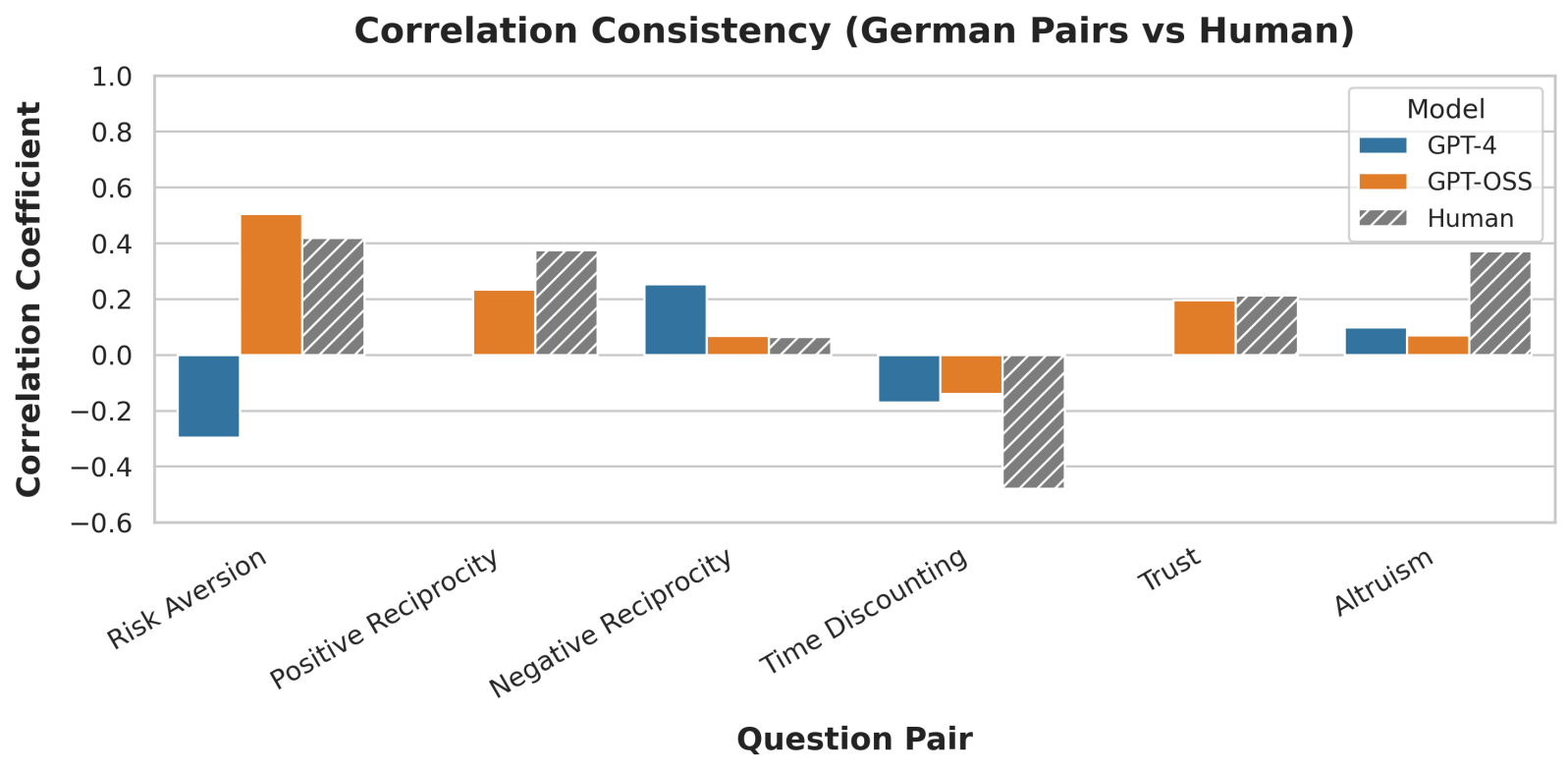
- Limited but noticeable effect on on average responses
- Going further -> look at the distribution of responses to self assessment questions

Chain of thought: Effect on Distribution



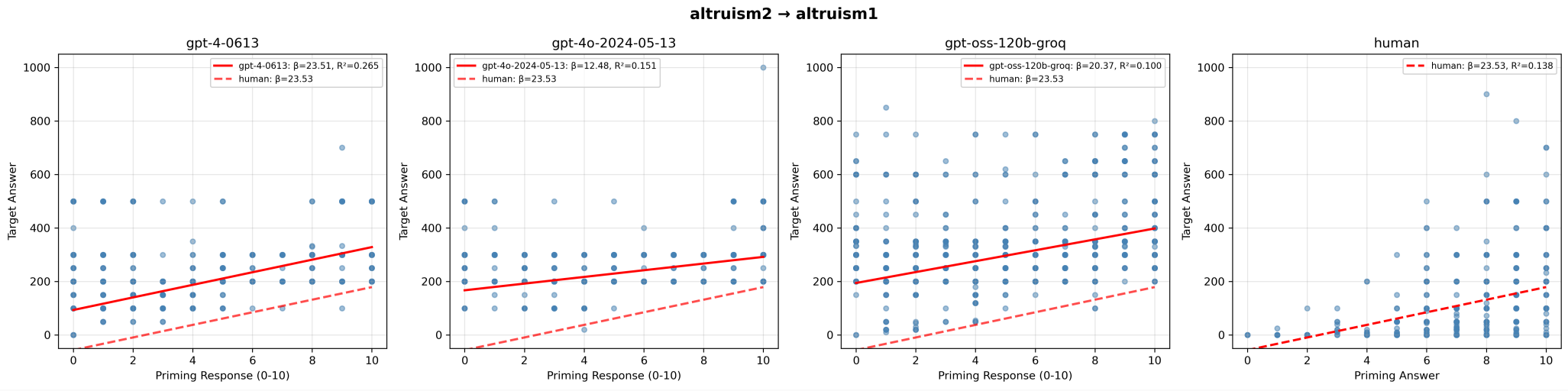
- Normal answers have too little heterogeneity w.r.t. humans
- Chain-of-thought prompts *increase the variance of responses*

Alignment Experiments: PairsHistory



- Spontaneous correlation between self-assessment and decision is *low or absent* for most traits — LLMs lack the structural consistency observed in humans.

Alignment Experiments: ALIGN (Altruism)

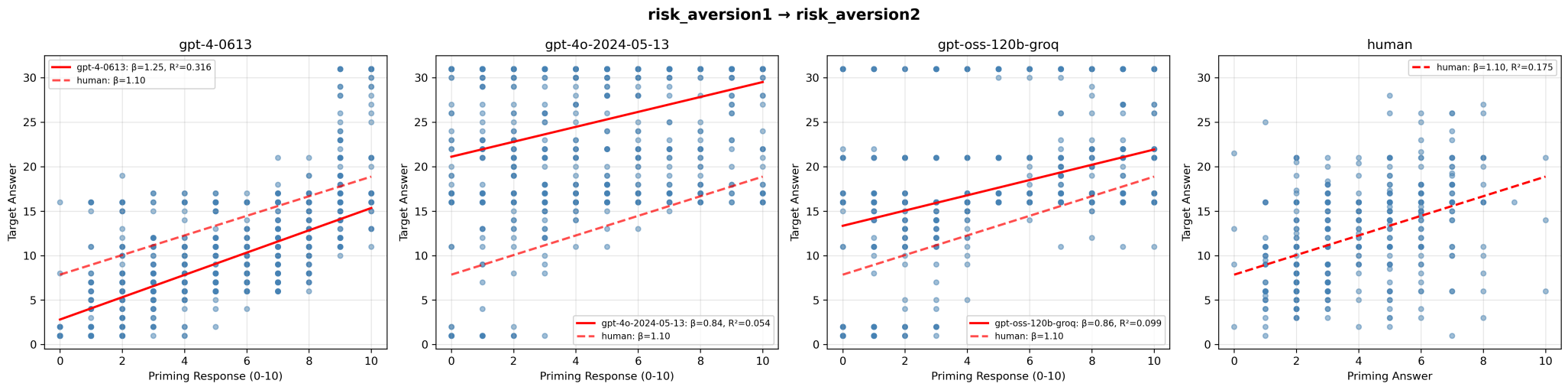


- **Robust Alignment**: Priming self-assessment shifts the altruistic choice significantly.
- **Human-like Slopes**: Both GPT-4 ($\beta = 23.51$) and GPT-OSS ($\beta = 20.37$) track the human baseline ($\beta = 23.53$) very closely.

Alignment Experiments: ALIGN (Trust)

- **Model Over-steering**: GPT-4 is over-sensitive to priming ($\beta = 3.74$) compared to the human benchmark ($\beta = 2.11$).
- **Under-steering**: GPT-OSS shows much weaker steering effects ($\beta = 0.85$), though still highly significant ($p < 0.01$).

Alignment Experiments: ALIGN (Risk Aversion)



- **Strong Steering**: Steering effect is highly robust for all models ($p < 0.01$).
- **Ballpark Slopes**: GPT-4 ($\beta = 1.25$) and GPT-OSS ($\beta = 0.86$) surround the human baseline slope ($\beta = 1.10$).

Alignment Experiments: Priming Regression Results

Trait (Self-Assess -> Choice)	GPT-4	GPT-OSS	Human
Altruism	23.51^{***} (1.18)	20.37^{***} (1.84)	23.53^{***} (2.96)
Risk Aversion	1.25^{***} (0.06)	0.86^{***} (0.08)	1.10^{***} (0.12)
Time Discounting	-1.07^{***} (0.04)	-0.38^{***} (0.03)	-1.55^{***} (0.14)
Trust	3.74^{***} (0.16)	0.85^{***} (0.11)	2.11^{***} (0.49)
Negative Reciprocity	-0.15 (0.09)	0.63^{***} (0.10)	0.39 (0.31)

Note: Standard errors are in parentheses. *** indicates significance at $p < 0.01$. Faded grey coefficients are not statistically significant ($p > 0.10$).

- **Robust Priming Effect**: Priming self-assessments shifts decisions significantly in the expected direction ($p < 0.01$, except negative reciprocity on GPT-4).
- **Close to Human**: Induced slopes are very close to human benchmarks for Altruism and Risk Aversion.
- **Over-Steering**: GPT-4 is over-sensitive to Trust priming (slope of 3.74 vs. 2.11 in humans).

Conclusion: Clarifying the Objects of Alignment

1. Level Alignment:

Weak

- **Question**: Does the LLM mean match the human mean?
- **Result**: Averages are in the ballpark, but matching is imperfect.

2. Distributional Alignment:

Poor

- **Question**: Does the LLM distribution match the human one?
- **Result**: Highly concentrated answers; low heterogeneity.

3. Structural Alignment:

Weak

- **Question**: Does self-assessment predict decisions as in humans?
- **Result**: Spontaneous correlation is low/missing (except GPT-OSS).

4. Steering Alignment:

Robust

- **Question**: Does an external self-assessment value steer choices?
- **Result**: Priming shifts decisions in the human-predicted direction.

Takeaway: The PSM is less useful as a passive measurement tool (weak level, poor distributional, and weak structural alignment), but highly promising as an active steering device (robust steering alignment).