Beyond Average Agents: Prompting and Steering Emotions to Shift Economic Behaviour in LLM Negotiation

Johann Frederik Machemer
Eigenwelt Labs
johann@eigenweltlabs.com

Christian-Hauke Poensgen
Eigenwelt Labs
chris@eigenweltlabs.com

Abstract

We measure how emotion induction shifts the play of three open mid-size instruct-tuned language models in five behavioral-economics games from Mei et al. [2024] (Dictator, Ultimatum Proposer, Ultimatum Responder, one-shot and five-round Prisoner's Dilemma). Emotion is injected through two parallel routes: the three prompt-induction strategies of Mozikov et al. [2024a] (simple, co-player, external) and Konen-style activation steering on mean-of-class style vectors [Konen et al., 2024]. Each route targets anger, sadness, and joy against a neutral baseline at sampling temperature 1. Anger lowers offered share in the Dictator Game and the Ultimatum Proposer role and lowers cooperation in both Prisoner's Dilemma variants across the panel, matching the direction Mozikov et al. [2024a] report for GPT-4. Sadness and joy are more model-dependent, and the two induction routes diverge on a small set of cells where text-level and representation-level interventions move behavior in opposite directions. The open mid-size panel does not reproduce the GPT-4 emotion-robustness reported by Mozikov, also shifting on sadness and joy where GPT-4 stayed flat. We report the two routes as parallel operationalizations of emotion injection rather than as competing methods.

1 Introduction

Behavioral economics has documented systematic departures from self-interested play in the canonical games: dictators share with strangers [Engel, 2011], responders in the Ultimatum Game reject low offers at real cost [Güth et al., 1982], and players in the Prisoner's Dilemma cooperate above the Nash benchmark even under a known finite horizon [Camerer, 2003]. A parallel literature shows that the size and direction of these departures depend on the player's emotional state at the moment of decision [Andrade and Ariely, 2009]: anger raises minimum acceptable offers in the Ultimatum Game [Riepl et al., 2016], incidental mood shifts giving in the Dictator Game [Chung et al., 2016], and emotional state changes cooperation in repeated Prisoner's Dilemma play [Stallen et al., 2021].

Recent work has placed LLMs inside these same games as economic agents. Horton [2023] frames the LLM as a *homo silicus* that can be given endowments and run through scenarios in direct analogy to the economist's *homo economicus*; Mei et al. [2024] administer a six-game battery to ChatGPT-3.5 and GPT-4 against tens of thousands of human subjects and find that GPT-4 falls inside the human distribution but at its more cooperative and altruistic end; Ross et al. [2024] fit canonical utility functions across a panel of models and report that current LLMs are neither entirely human-like nor entirely *economicus*-like. Most of this work assumes a fixed affective state. The systematic exception, Mozikov et al. [2024a], prompts five basic emotions into GPT-3.5 and GPT-4 across Ultimatum, Dictator, Prisoner's Dilemma, and Battle of the Sexes, and shows that emotional framing can move LLM behavior toward or away from the human distribution depending on the emotion and the model. Two gaps remain: the model panel is restricted to closed GPT, and emotion is induced only through the prompt.

There are two distinct ways to inject an emotion into an LLM. Prompt induction adds an emotional descriptor to the input and lets the text-to-text pipeline carry the rest [Li et al., 2023, Mozikov et al., 2024a]. Activation steering instead modifies the model's hidden states at inference time, adding a vector derived from contrastive examples to the residual stream at a chosen layer [Turner et al., 2024, Konen et al., 2024, Sofroniew et al., 2026]. The two interventions act at different points in the same model: one at the text-level interface, the other at the representational level. They are not equivalent operationalizations of the same emotion, and no prior work has run both on the same behavioral-game battery.

This paper closes both gaps on the same battery. We run three open instruct-tuned models (Qwen3-8B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-3-12B-IT) across five behavioral-economics games from the Mei et al. [2024] battery: Dictator, Ultimatum Proposer, Ultimatum Responder, one-shot Prisoner's Dilemma, and five-round Prisoner's Dilemma. Emotion is induced through two parallel routes, the three Mozikov prompt strategies (simple, co-player, external) and Konen-style activation steering, applied to three target emotions (anger, sadness, joy) against a neutral baseline. Each cell is run 50 times at sampling temperature 1 following the Mei protocol. We frame the work as a descriptive replication of Mozikov et al. [2024a] on an open mid-size panel, with activation steering added as a fourth induction route into the same framework; the four routes are reported as parallel operationalizations of emotion injection, with no claim that one is more correct than another.

Across the three open models, anger lowers offered share in the Dictator Game and the Ultimatum Proposer role, and lowers cooperation in both Prisoner's Dilemma variants, on every cell that crosses our direction-of-effect threshold (§3.4); the same direction is reported for GPT-4 by Mozikov et al. [2024a]. Sadness and joy are more model-dependent: prompt and steering routes mostly agree in direction, with the cleanest reversal on Gemma under joy, where prompt induction raises offered share and cooperation in four games while steering lowers the same metrics. The Gemma steering arm is silent on anger and sadness in three of five games, consistent with a vector-quality caveat that is flagged with the affected cells throughout §4.

Section 2 reviews related work across behavioral game theory, LLMs in game theory, prompt-based emotion injection, and activation steering. Section 3 describes the methodology, covering the five games, the two induction routes, and the vector-validation gates. Section 4

reports results game-by-game. Section 5 discusses cross-model patterns and limitations. Section 6 concludes.

2 Related Work

To ground our research in emotion steering in economic games four research threads have to be discussed. §2.1 covers what behavioral economics has documented about how emotions shape human behavior in classic games; §2.2 surveys recent work that places LLMs in those same games as economic agents; §2.3 reviews prompt-based emotion injection in LLMs; and §2.4 covers activation steering as a representation-level alternative to prompt-based control. Our contribution sits at the intersection of these threads, representation-level emotion injection in the economic-game battery of Mei et al. [2024] and Mozikov et al. [2024a].

2.1 Behavioral game theory and emotions

Behavioral economics has long documented departures from purely self-interested play: dictators share with strangers [Engel, 2011], responders in the Ultimatum Game reject low offers at real cost [Güth et al., 1982], and players in the Prisoner's Dilemma cooperate above the Nash benchmark even under a known finite horizon [Camerer, 2003]. A parallel literature shows that the size of these departures depends on the player's emotional state at the moment of decision [Andrade and Ariely, 2009]. Studies take one of two routes: a *valence-level* induction (pleasant vs. unpleasant mood, typically via film clips, music, or affective pictures; e.g., Riepl et al., 2016, Chung et al., 2016, Pérez-Dueñas et al., 2018), which asks whether the direction of the affect predicts the shift; and a *specific-emotion* induction (anger, sadness, fear, disgust, or joy against neutral; e.g., Liu et al., 2016, Chierchia et al., 2021, Nguyen and Noussair, 2022), which asks whether the identity of the emotion matters above its valence. Both routes find that emotions move play in classic games, but disagree on which level of granularity carries the explanatory work [Liu et al., 2016], and several of the more robust effects run against lay intuition [Chierchia et al., 2021, Pérez-Dueñas et al., 2018].

2.2 LLMs and game theory

Two strands of recent work have placed LLMs inside the games of behavioral economics. The first treats the LLM as an economic agent and asks what utility function its behavior implies. Horton [2023] frames the LLM as a *homo silicus* that can be given endowments, preferences, and information and then run through scenarios, in direct analogy to the economist's use of *homo economicus*. Ross et al. [2024] fit canonical utility functions (Fehr–Schmidt inequity aversion, prospect-theory risk and loss aversion, hyperbolic time discounting) to a panel of open- and closed-source models and find that current LLMs are neither entirely human-like nor entirely *economicus*-like: they exhibit weaker inequity aversion toward themselves but stronger inequity aversion toward others, and their behavior is unstable across settings. Adjacent work [Jia et al., 2024, Liu et al., 2025] extends this measurement program to risk preference, probability weighting, and loss aversion under demographic or persona conditioning.

The second strand asks whether LLM play aligns with human play in the same games. Mei et al. [2024] administer a six-game behavioral test (Dictator, Ultimatum, Trust, Bomb Risk, Public Goods, finitely repeated Prisoner's Dilemma) to ChatGPT-3.5 and GPT-4 against tens of thousands of human subjects from more than fifty countries; they find that GPT-4 falls inside the human distribution but at its more cooperative, altruistic, and trusting end, and a revealed-preference analysis suggests the model acts as if it were maximizing the average of own and partner payoff. Mozikov et al. [2024a] extend this alignment question to emotion-prompted conditions across the Ultimatum, Dictator, Prisoner's Dilemma, and Battle of the Sexes, and show that prompted emotion can move LLM behavior either toward or away from the human distribution depending on the emotion and the model. Our work builds on the Mei battery (five of the six games) and follows the Mozikov prompt-induction design, but adds activation steering as a second induction route into the same battery.

2.3 Emotions in LLMs

Emotion-prompting studies inject emotional descriptors into the input and measure their effect on downstream tasks; Li et al. [2023] report mixed effects across reasoning and semantic benchmarks, but do not examine how emotional framing affects the model's decisions in social or strategic contexts. The shift from task-level evaluation (does emotion help reasoning?) to behavior-level evaluation (does emotion change strategic choice?) is the move Mozikov et al. [2024a] make. They inject five basic emotions (anger, sadness, happiness, disgust, fear) into GPT-3.5 and GPT-4 across Ultimatum, Dictator, Prisoner's Dilemma, and Battle of the Sexes, and find that emotion shifts game behavior, that anger in particular can disrupt GPT-4's otherwise human-aligned play, and that the size and direction of the shift depend on the model.

Crucially for our setup, Mozikov vary not only which emotion is injected but also how the emotion is *sourced*: a **simple** strategy attributes the emotional state to the model directly without further context; a **co-player** strategy frames the emotion as caused by the opponent; an **external** strategy attributes it to an unrelated event. The three strategies are motivated by the human finding that the same emotion can move behavior in opposite directions depending on its source (e.g., disgust directed at the opponent reduces UG offers, while disgust provoked by an unrelated event does not, and can even raise generosity; Mozikov et al., 2024a). We adopt all three strategies as our prompt baselines, holding this source-of-emotion design fixed while varying the induction route.

A follow-up from the same group [Mozikov et al., 2024b] extends the analysis to a wider model panel (Claude, LLaMA, Mixtral, Command R+) and to ethical benchmarks alongside games, reporting that negative emotions reduce accuracy on ethical reasoning tasks and that open and smaller models are the most affected. All of this work injects emotion through the prompt.

2.4 Activation steering for behavior

Activation steering modifies a model's hidden states at inference time rather than its prompt, weights, or decoding distribution. Turner et al. [2024] ("ActAdd") add the difference between activations of a contrasting prompt pair to the forward pass at a chosen layer. Konen et al. [2024] generalize this into a *style vector*: the steering vector for class s at layer i is the mean activation of class-s inputs minus the mean activation of the remaining classes, applied at middle layers during generation. The construction requires no fine-tuning, only forward passes on a labeled corpus. Our paper uses the Konen construction unchanged, applied to anger, sadness, and joy on three open mid-size models.

Recent work has applied this approach to emotion specifically. Sofroniew et al. [2026] identify 171 emotion concepts in Claude through sparse-autoencoder features and show that steering on them causally drives behavior. Jeong [2026] replicates the finding in small open models (124M to 3B) and documents a Qwen-specific entanglement in which emotion steering activates Chinese-language tokens that RLHF does not suppress, which we monitor in our Qwen runs. Sun et al. [2026] move from discrete emotion labels to continuous steering in valence–arousal–dominance space via SAE features, but evaluate on reasoning and safety rather than economic games. The closest behavioral parallel is Sakai et al. [2026], who manipulate Big Five traits via prompts (not steering) in repeated Prisoner's Dilemma and find agreeableness dominant. What is missing is a representation-level handle on emotion applied to the economic-game battery of Mei and Mozikov.

3 Methodology

We measure emotion-induced shifts in five behavioral-economics games from Mei et al. [2024]. Emotion is induced through two parallel routes, prompt induction following Mozikov et al. [2024a] and activation steering following Konen et al. [2024]. We evaluate three open instruct-tuned models against a neutral baseline, with 50 repetitions per condition at sampling temperature 1. Behavior is reported on a single game-native metric per game with a fixed direction-of-effect threshold.

3.1 Selected Games

We use five measures from the Mei et al. [2024] behavioral game battery, covering the canonical bargaining and cooperation contexts.

Dictator Game. A "dictator" receives a $100 endowment and chooses how much to transfer to a second player who has no decision and no recourse. The game isolates altruism in the absence of any strategic incentive.

Ultimatum Game. A Proposer offers a division of a $100 endowment to a Responder, who either accepts (both keep their proposed shares) or rejects (both receive nothing). Proposer and Responder are run separately and evaluated on different metrics: the offered share for the Proposer, and the minimum acceptable offer (MAO) for the Responder.

Prisoner's Dilemma (one-shot). The model chooses to cooperate or defect against an unseen partner. The Mei et al. [2024] payoff matrix makes mutual cooperation jointly preferable but defection individually dominant.

Prisoner's Dilemma (five-round). The same payoff matrix is played for five rounds against a fixed partner schedule (defect, defect, cooperate, cooperate over rounds 1 to 4; the model's move in round 5 is unconditioned). The fixed schedule lets the model condition on partner history while keeping the partner sequence deterministic across all conditions.

Game prompts are taken verbatim from the public ChatGPT-Behavioral protocol of Mei et al. [2024], distributed by Xie [2024]. The remaining Mei measures (Trust, Bomb Risk, Public Goods) and Mozikov's Battle of the Sexes are out of scope for this paper.

3.2 Emotion Induction Routes

Emotion is induced through two parallel routes. **Prompt induction** modifies the model's text input by inserting an emotional state and follows Mozikov et al. [2024a]. **Activation steering** modifies the model's hidden states at inference time by adding an emotion vector to the residual stream and follows the style-vector construction of Konen et al. [2024]. Both routes target the same three emotions (anger, sadness, joy) against a shared neutral baseline (no induction).

3.2.1 Prompt induction

We reuse the three induction strategies of Mozikov et al. [2024a] without modification. The strategies differ in the attributed source of the emotion:

- **Simple** states the emotion with no context (e.g. "Also, now you are angry, which can affect your choices.").
- **Co-player-based** attributes the emotion to the co-player's prior action.
- **External-based** attributes the emotion to an event unrelated to the co-player.

The induction string is inserted into the user message between the environment description and the game rules. Verbatim prompts for all three emotions and three strategies appear in Appendix A.

3.2.2 Activation steering and vector validation

We construct one steering vector per target emotion using the activation-based style-vector method of Konen et al. [2024]. We use the source corpus of the GoEmotions dataset [Demszky et al., 2020], restricted to the Ekman-mapped subset and to anger, sadness, and joy (about 800 labeled samples per emotion). For each candidate layer ℓ, we record the residual-stream activation $a^{(\ell)}(x) \in \mathbb{R}^d$ at the last non-padding token of every sample x. Let $\mathcal{X}_e = \{x : \text{label}(x) = e\}$ denote the set of samples labelled with emotion $e \in \mathcal{E} = \{\text{anger}, \text{sadness}, \text{joy}\}$,

Table 1: Per-model activation-steering parameters. Layers indicate the chosen contiguous window (of total model depth). λ is the per-emotion injection coefficient in raw activation units, selected by the tone-shift sanity check (§3.2.2).

Model	Layers (of total)	Depth	λ_{anger}	λ_{sadness}	λ_{joy}
Qwen3-8B-Instruct	20, 21, 22 (of 36)	∼58%	1.5	1.0	1.0
Gemma-3-12B-IT	31, 32, 33 (of 48)	∼67%	2.0	2.0	2.0
Mistral-7B-Instruct-v0.3	14, 15, 16 (of 32)	∼47%	1.5	2.0	1.0

and let $\mathcal{X}_{\bar{e}} = \bigcup_{e' \in \mathcal{E} \setminus \{e\}} \mathcal{X}_{e'}$ denote the complement. The style vector for emotion e at layer ℓ is the difference of the two class means,

$$v_e^{(\ell)} = \frac{1}{|\mathcal{X}_e|} \sum_{x \in \mathcal{X}_e} a^{(\ell)}(x) - \frac{1}{|\mathcal{X}_{\bar{e}}|} \sum_{x \in \mathcal{X}_{\bar{e}}} a^{(\ell)}(x). \tag{1}$$

The vectors are used in raw activation units, with no normalization, matching Konen et al. [2024] Eq. 5.

Layer selection. For each candidate layer, we train a logistic-regression probe on an 80% training split to predict emotion-vs-rest from the captured activations, and score AUC on the held-out 20% validation split. We select the three contiguous layers with the highest mean validation AUC. Sweeping layers 16 to 27 (about 44% to 75% depth) brackets Konen's reported peak at 55% to 60% depth and lets the AUC pick the window per model. Full per-layer AUC curves appear in Appendix B.

Injection at inference. The selected vector is added to the residual stream at the three chosen layers, at every token position of every forward pass. The injection coefficient λ is fixed per (model, emotion) cell at the value that maximises the tone shift on subjective prompts (validation below) while staying inside the coherence cap. The chosen per-model layers and per-(model, emotion) coefficients are listed in Table 1.

Deviation from Konen et al.. Konen et al. [2024]'s released code [Konen and DLR-SC, 2024] appends "*Write the answer in a {style} manner.*" to the user prompt in the steering condition. We omit this manner cue so the reported steering effects come from the activation intervention alone.

Vector validation. Each vector passes three gates before any game run. **(i) Tone shift.** On 20 Konen-style subjective prompts, the target-emotion score from an external classifier [Hartmann, 2022] must increase monotonically with λ over Konen's emotion grid (λ from 0 to 2.0). **(ii) Coherence cap.** The working range of λ is capped where the perplexity ratio relative to neutral exceeds 5×. **(iii) Script-mix monitor.** On Qwen, the non-Latin-token fraction must stay within 5 percentage points of the neutral baseline, following Jeong [2026], who documents Chinese-token leakage under emotion steering on Qwen models. Vectors that fail any gate are flagged in the per-cell reporting (§4).

3.3 Experimental Setup

Models. Three open instruct-tuned models: Qwen3-8B-Instruct [Qwen Team, 2025], Mistral-7B-Instruct-v0.3, and Gemma-3-12B-IT. Open weights are required for the steering arm (§3.2.2), which needs hidden-state access at inference time. Mozikov tested two closed models (GPT-3.5, GPT-4) and could only run prompt induction; our open panel allows both induction routes on the same models.

Emotions. Three Ekman basic emotions (anger, sadness, joy) against a neutral baseline. The set is scope-narrowed from Mozikov's five (anger, sadness, happiness, disgust, fear).

Sampling and decoding. Temperature $T = 1$, $N = 50$ repetitions per condition, following the Mei protocol. This differs from Mozikov, who used $T = 0$ with $N = 5$. Direction-of-effect

is therefore comparable to Mozikov, but magnitude is not: a 5 percentage point shift at $T = 0$ is a discrete mode-flip, the same shift at $T = 1$ is a distribution-mean change. The per-cell thresholds in §3.4 are calibrated to the $T = 1$ regime. Each model is loaded in bf16 and queried with its default chat template under vLLM [Kwon et al., 2023]. For Qwen3-8B-Instruct, the thinking mode is disabled (`enable_thinking=False`), so the model does not emit a reasoning-mode preamble that would bypass the game-decision parser.

Prompt assembly. The chat-API system prompt is `"You are a helpful assistant."`, following Mei et al. [2024]. The user message contains the environment description, emotion induction (if any), game rules, and decision instruction, in that order.

3.4 Behavior Analysis

Game-native metrics. Each game is reported in its own native metric, with no cross-game unified axis (Table 2).

Table 2: Per-game native metric and direction of an upward (↑) move.

Game	Metric	↑ means
Dictator	Offered share ($)	more $ offered
UG-Proposer	Offered share ($)	more $ offered
UG-Responder	Minimum acceptable offer (MAO, $)	responder more demanding
PD-oneshot	P(cooperate)	more cooperation
PD-5round	P(cooperate) in round 1	more cooperation

Per-cell threshold. For each (model × emotion × induction-route) cell, we report the per-cell mean and compare to the neutral baseline for the same model. A direction-of-effect arrow fires when $|\Delta|$ exceeds a fixed threshold: \$2 for the dollar metrics (Dictator, UG-Proposer, UG-Responder MAO) and 0.02 for the probability metrics (PD-oneshot, PD-5round). Below threshold the cell reads "=". The chosen cut is tighter than Mozikov's implicit floor (which appears to be ∼5pp based on the magnitudes in his Fig. 5 heatmap), made appropriate by our larger per-cell N (50 vs. 5).

UG-Responder elicitation note. Mozikov elicits responder behavior with the *direct method*: per trial, the LLM sees a specific offer and replies ACCEPT or REJECT, and the acceptance rate is reported across offer levels. We follow Mei et al. [2024] with the *strategy method*: per trial, the LLM declares a single MAO. Both are standard in behavioral economics, but the two scales are not linear transformations of one another. Direction-of-effect is comparable across methods (Mozikov ↑ acceptance corresponds to our ↓ MAO), but magnitude is not. We do not back-compute a Mozikov-style acceptance rate from our MAO data.

PD-5round, round-1 cooperation rate only. We follow Mei et al. [2024] and report PD-5round as the proportion of trials in which the model cooperates in the first round. The partner schedule `[Pull, Pull, Push, Push]` is taken verbatim from Mei's protocol [Xie, 2024]. Mei's reported PD baselines (45.1% human, 76.7% ChatGPT-3.5, 91.7% ChatGPT-4) and Mei's revealed-preference b-fit both use round 1 only, justified by independence: round 1 is unconditioned on partner history, while rounds 2–5 mix the agent's cooperation preference with reciprocity dynamics. The repeated-game / reciprocity story is carried by the per-round trajectory plot in §4.3.2, analogous to Mei's Fig. 4.

4 Experimental Results

4.1 Dictator Game

In the following section we report the offered share in dollars out of the \$100 endowment, with higher values indicating more generous play. We take each of our three open models, walk the two induction routes (prompting and steering) across the three target emotions,

Table 3: **Dictator Game.** Offered amount (neutral) is the per-model baseline mean, in dollars out of the $100 endowment. Arrow cells indicate direction-of-effect at the $|\Delta| > \$2$ threshold (§3.4). Superscripts on prompt cells identify which Mozikov-style strategy crosses threshold: s = simple, o = co-player, e = external. The Human, ChatGPT-3.5, and ChatGPT-4 rows are taken from Table 1 of Mozikov et al. [2024a]; their steer cells are marked n/a because activation steering was not tested on closed models.

Model	Offered (neut.)	Anger		Sadness		Joy	
		prompt	steer	prompt	steer	prompt	steer
Human [Engel, 2011]	\$28.35	$\uparrow^{e}$	n/a	$\uparrow^{e}$	n/a	$\downarrow^{e}$	n/a
ChatGPT-3.5 [Mozikov et al., 2024a]	\$33.00	$\uparrow^{e}\downarrow^{so}$	n/a	$\uparrow^{s}$	n/a	$\uparrow^{o}$	n/a
ChatGPT-4 [Mozikov et al., 2024a]	\$50.00	$\downarrow^{so}$	n/a	$\downarrow^{o}$	n/a	=	n/a
Qwen3-8B (ours)	\$50.00	$\downarrow^{oe}$	$\downarrow$	$\downarrow^{e}$	$\downarrow$	$\uparrow^{oe}$	=
Mistral-7B (ours)	\$46.30	$\downarrow^{o}$	$\uparrow$	$\uparrow^{so}\downarrow^{e}$	$\uparrow$	$\uparrow^{s}\downarrow^{o}$	=
Gemma-3-12B (ours)	\$60.30	$\downarrow^{soe}$	=	$\downarrow^{oe}$	=	$\uparrow^{o}$	$\downarrow$

and close the description of results with an Overall paragraph that compares all six cases. Table 3 reports the direction-of-effect arrows for each model, emotion, and induction route; the Human, ChatGPT-3.5, and ChatGPT-4 rows from Mozikov et al. [2024a] are included for reference.

Qwen3-8B. Qwen starts from an even-split neutral baseline of \$50.0. Under anger, prompting lowers offered amount when the emotion is attributed to the co-player ($\Delta = -\$46.6$) or to an external event ($\Delta = -\$34.0$); the simple strategy does not move; steering drives offered amount to the \$0 floor ($\Delta = -\$50.0$). Under sadness, prompting crosses threshold only on the external strategy ($\Delta = -\$3.2$), while steering lowers offered amount more strongly ($\Delta = -\$18.0$). Under joy, prompting raises offered amount under the co-player ($\Delta = +\$5.0$) and external ($\Delta = +\$5.8$) strategies; steering produces no change.

Mistral-7B. Mistral has a neutral baseline of \$46.3, close to but below an equal split. Under anger, prompting lowers offered amount only when the emotion is attributed to the co-player ($\Delta = -\$4.5$), while steering raises it ($\Delta = +\$12.1$). Under sadness, prompting splits across strategies, with simple ($\Delta = +\$3.1$) and co-player ($\Delta = +\$2.9$) raising offered amount and external lowering it ($\Delta = -\$2.8$); steering raises offered amount more strongly ($\Delta = +\$10.2$). Under joy, prompting again splits, raising offered amount under simple ($\Delta = +\$2.6$) and lowering it under co-player ($\Delta = -\$6.5$); steering produces no change.

Gemma-3-12B. Gemma starts from a high neutral baseline of \$60.3, above an equal split. Under anger, prompting lowers offered amount sharply across all three strategies (simple $\Delta = -\$55.5$, co-player $\Delta = -\$60.2$, external $\Delta = -\$59.5$), driving the cell close to zero; steering produces no change ($\Delta = -\$1.1$). Under sadness, prompting lowers offered amount under the co-player ($\Delta = -\$29.5$) and external ($\Delta = -\$20.5$) strategies; steering again produces no change. Under joy, prompting raises offered amount under the co-player strategy ($\Delta = +\$7.6$) while steering lowers it ($\Delta = -\$6.5$). The silent steer cells on anger and sadness are consistent with the vector-quality caveat for Gemma described in §3.2.2.

Overall. Across the three open models, anger prompts lower offered amount in every cell that crosses threshold, matching the GPT-4 direction reported by Mozikov et al. [2024a] and going against the human pattern observed by Andrade and Ariely [2009], where external anger raises offered amount via behavioral carryover. The co-player joy prompt raises offered amount on Qwen ($\Delta = +\$5.0$) and Gemma ($\Delta = +\$7.6$), reproducing the $\uparrow^{o}$ pattern reported for GPT-3.5; on Mistral the same prompt reverses ($\Delta = -\$6.5$). Activation steering moves offered amount strongly on both Qwen and Mistral but in opposite directions: it drives the cell to the \$0 floor under Qwen × anger ($\Delta = -\$50.0$) and lowers offered amount under Qwen × sadness ($\Delta = -\$18.0$), while raising offered amount under both Mistral × anger ($\Delta = +\$12.1$) and Mistral × sadness ($\Delta = +\$10.2$). On Mistral × anger this puts the two routes in direct conflict: the co-player prompt lowers offered amount while steering raises it.

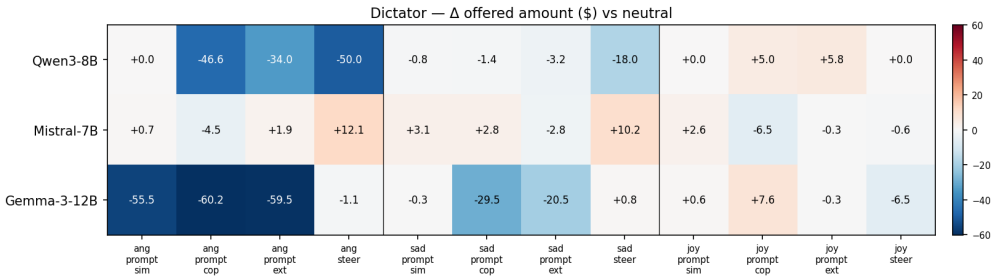

Figure 1: **Dictator — Δ offered amount ($) vs. neutral.** Per-cell magnitudes underlying the arrows in Table 3. Rows: open models. Columns: emotion × induction route, with the three prompt strategies (sim = simple, cop = co-player, ext = external) and the steer arm shown separately. Blue = lower offered amount; red = higher.

4.2 Ultimatum Game

4.2.1 Proposer

For the Proposer role we report the offered share in dollars out of the $100 endowment, with higher values indicating more generous play. The threshold and induction-route structure follow §4.1. Table 4 reports the direction-of-effect arrows for each model, emotion, and induction route; the Human, ChatGPT-3.5, and ChatGPT-4 rows from Mozikov et al. [2024a] are included for reference.

Qwen3-8B. Qwen starts from an even-split neutral baseline of $50.0. Under anger, prompting lowers offered amount across all three strategies (simple $\Delta = -\$2.5$, co-player $\Delta = -\$14.9$, external $\Delta = -\$4.0$); steering lowers it further ($\Delta = -\$36.1$). Under sadness, prompting again lowers offered amount across all three strategies (simple $\Delta = -\$5.9$, co-player $\Delta = -\$2.4$, external $\Delta = -\$6.2$); steering lowers it more strongly ($\Delta = -\$27.8$). Under joy, only the simple prompt crosses threshold, lowering offered amount ($\Delta = -\$7.5$); steering produces no change.

Mistral-7B. Mistral has a neutral baseline of $49.2. Under anger, prompting lowers offered amount across all three strategies (simple $\Delta = -\$2.5$, co-player $\Delta = -\$15.5$, external $\Delta = -\$8.6$); steering lowers it as well ($\Delta = -\$6.5$), though only 25 of 50 trials parsed under this condition, so the cell is unstable. Under sadness, no induction route crosses threshold. Under joy, prompting does not move offered amount, while steering lowers it ($\Delta = -\$4.1$).

Gemma-3-12B. Gemma has a neutral baseline of $48.7. Under anger, prompting lowers offered amount sharply across all three strategies (simple $\Delta = -\$43.9$, co-player $\Delta = -\$46.2$, external $\Delta = -\$42.9$), pushing the cell close to the floor; steering produces a small positive Δ ($+\$3.9$) but with very high variance (SD $\approx$ \$44, an order of magnitude larger than the prompt cells), consistent with the Gemma vector-quality caveat. Under sadness, prompting lowers offered amount across all three strategies (simple $\Delta = -\$17.5$, co-player $\Delta = -\$23.6$, external $\Delta = -\$33.2$); steering lowers it more mildly ($\Delta = -\$12.2$). Under joy, prompting raises offered amount across all three strategies (simple $\Delta = +\$5.5$, co-player $\Delta = +\$32.0$, external $\Delta = +\$9.9$), while steering lowers it ($\Delta = -\$13.8$).

4.2.2 Responder

For the Responder role we report the Minimum Acceptable Offer (MAO) in dollars, with higher values indicating a more demanding responder. Two of our three open models (Qwen, Gemma) have neutral MAO = \$1.0, sitting on the \$0 floor: effects in the more-accepting direction (↓ MAO) are nearly unobservable on those rows by construction. Mistral has neutral MAO $\approx$ \$15.0, the only non-saturated baseline in our panel. We follow Mei et al. [2024] in eliciting MAO directly (the strategy method), whereas Mozikov et al. [2024a] elicit per-offer accept/reject and report acceptance rate; cross-comparison requires sign-flipping

Table 4: **Ultimatum Game (Proposer).** Offered amount (neutral) is the per-model baseline mean, in dollars out of the $100 endowment. Arrow cells indicate direction-of-effect at the $|\Delta| > \$2$ threshold (§3.4). Superscripts on prompt cells identify which Mozikov-style strategy crosses threshold: s = simple, o = co-player, e = external. The Human, ChatGPT-3.5, and ChatGPT-4 rows are taken from Table 2 of Mozikov et al. [2024a]; their steer cells are marked n/a because activation steering was not tested on closed models. † Mistral × anger × steer parsed only 25 of 50 trials and should be read as unstable rather than directional. ‡ Gemma × anger × steer crosses threshold on the mean ($\Delta = +\$3.9$) but with very high variance (SD ≈ $44) and a Mann–Whitney $p = 0.80$ against neutral, consistent with the Gemma vector-quality caveat (§3.2.2).

Model	Offered (neut.)	Anger		Sadness		Joy	
		prompt	steer	prompt	steer	prompt	steer
Human [Oosterbeek et al., 2004]	$41.00	$\uparrow^{e}$	n/a	$\uparrow^{e}$	n/a	$\downarrow^{e}$	n/a
ChatGPT-3.5 [Mozikov et al., 2024a]	$35.00	$\downarrow^{so}$	n/a	=	n/a	=	n/a
ChatGPT-4 [Mozikov et al., 2024a]	$50.00	$\downarrow$	n/a	$\downarrow^{o}$	n/a	=	n/a
Qwen3-8B (ours)	$50.00	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{s}$	=
Mistral-7B (ours)	$49.20	$\downarrow^{soe}$	$\downarrow^{\dagger}$	=	=	=	$\downarrow$
Gemma-3-12B (ours)	$48.70	$\downarrow^{soe}$	$\uparrow^{\ddagger}$	$\downarrow^{soe}$	$\downarrow$	$\uparrow^{soe}$	$\downarrow$

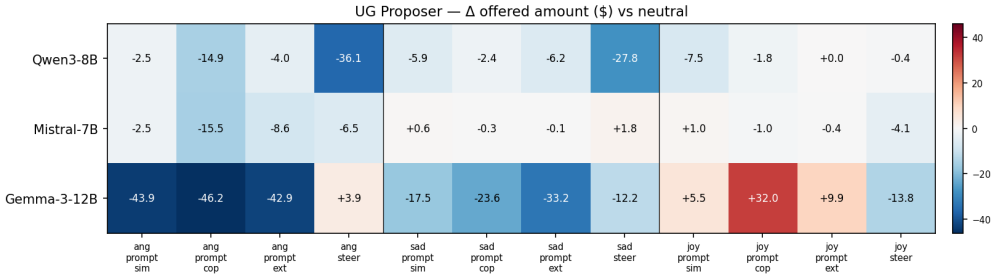

Figure 2: **UG Proposer — Δ offered amount ($) vs. neutral.** Per-cell magnitudes underlying the arrows in Table 4. Conventions follow Figure 1.

the Mozikov arrows (his ↑ acceptance corresponds to our ↓ MAO), but magnitudes are not directly comparable (§3.4). Table 5 reports the direction-of-effect arrows for each model, emotion, and induction route; the Mozikov rows are shown after sign-flipping.

Qwen3-8B. Qwen has a neutral MAO of $1.0, sitting on the $0 floor. Anger and sadness do not move the cell under any induction route. Under joy, prompting does not cross threshold, while steering raises MAO ($\Delta = +\$6.1$), the only joy × steer effect of this size in our data.

Mistral-7B. Mistral has a neutral MAO of $15.0. Under anger, prompting raises MAO across all three strategies (simple $\Delta = +\$3.1$, co-player $\Delta = +\$27.5$, external $\Delta = +\$35.2$); steering also raises it ($\Delta = +\$2.5$), though 38 of 50 trials parsed under this condition. Under sadness, prompting splits, with external raising MAO ($\Delta = +\$18.9$) and co-player lowering it ($\Delta = -\$4.0$); steering raises MAO strongly ($\Delta = +\$17.1$). Under joy, the co-player prompt raises MAO ($\Delta = +\$10.3$), and steering raises it more strongly ($\Delta = +\$20.9$).

Gemma-3-12B. Gemma has a neutral MAO of $1.0, also floor-saturated. Under anger, prompting raises MAO sharply across all three strategies (simple $\Delta = +\$4.5$, co-player $\Delta = +\$57.8$, external $\Delta = +\$47.4$), while steering does not move the cell. Under sadness, only the external prompt crosses threshold ($\Delta = +\$35.7$), with steering again silent. Under joy, prompting raises MAO under the co-player ($\Delta = +\$17.9$) and external ($\Delta = +\$2.1$) strategies, while steering does not cross. The silent steer cells on anger and sadness are again consistent with the Gemma vector-quality caveat (§3.2.2).

Table 5: **Ultimatum Game (Responder).** MAO (neutral) is the per-model baseline mean of the Minimum Acceptable Offer in dollars; ↑ MAO = responder more demanding. Arrow cells indicate direction-of-effect at the $|\Delta_{\text{MAO}}| > \$2$ threshold (§3.4). Superscripts on prompt cells identify which Mozikov-style strategy crosses threshold: s = simple, o = co-player, e = external. The Human, ChatGPT-3.5, and ChatGPT-4 rows are taken from Table 3 of Mozikov et al. [2024a] and sign-flipped (his ↑ acceptance = our ↓ MAO); steer cells are marked n/a because activation steering was not tested on closed models. §Mozikov uses the direct method (per-offer accept/reject) and reports acceptance rate, not MAO; the human ∼$30 figure is the cross-cultural mode reported by Oosterbeek et al. [2004], not from Mozikov. † Mistral × anger × steer parsed only 38 of 50 trials and should be read with that instability in mind.

Model	MAO (neut.)	Anger		Sadness		Joy	
		prompt	steer	prompt	steer	prompt	steer
Human [Oosterbeek et al., 2004]	∼$30§	$\uparrow^{e}$	n/a	$\uparrow^{e}$	n/a	$\downarrow^{e}$	n/a
ChatGPT-3.5 [Mozikov et al., 2024a]	n/a§	↑	n/a	↑	n/a	↓	n/a
ChatGPT-4 [Mozikov et al., 2024a]	n/a§	↑	n/a	=	n/a	↓	n/a
Qwen3-8B (ours)	$1.00	=	=	=	=	=	↑
Mistral-7B (ours)	$15.00	$\uparrow^{soe}$	$\uparrow^{\dagger}$	$\uparrow^{e}\downarrow^{o}$	↑	$\uparrow^{o}$	↑
Gemma-3-12B (ours)	$1.00	$\uparrow^{soe}$	=	$\uparrow^{e}$	=	$\uparrow^{oe}$	=

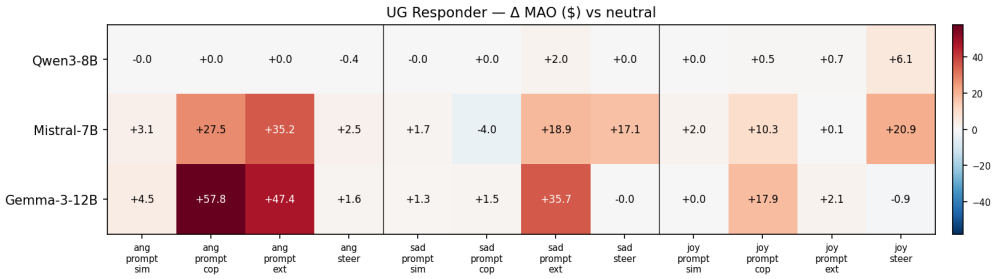

Figure 3: **UG Responder — Δ MAO ($) vs. neutral.** Per-cell magnitudes underlying the arrows in Table 5. Red = MAO raised (responder more demanding); blue = MAO lowered. Conventions otherwise follow Figure 1. Note the asymmetry imposed by floor saturation on Qwen and Gemma (neutral MAO ≈ $1): the more-accepting direction has little range on those rows.

Overall (UG). On the Proposer role, anger lowers offered amount on all three open models, matching the GPT-4 direction reported by Mozikov et al. [2024a]. On the Responder role, anger raises MAO on Mistral and Gemma, the same more-demanding direction on the two models with downward range; Qwen is unable to move further down by construction. Activation steering of joy raises MAO on Qwen ($\Delta = +\$6.1$) and Mistral ($\Delta = +\$20.9$), opposite to the human and GPT-3.5 direction (sign-flipped: joy ↓ MAO). The two routes mostly agree in direction across UG, with Mistral × sadness as the salient exception: prompt-external raises MAO (+$18.9), prompt-coplayer lowers it (−$4.0), and steering raises it (+$17.1).

4.3 Prisoner's Dilemma

4.3.1 One-shot

We report the proportion of trials in which the model cooperates (plays Push), with higher values indicating more cooperative play and an arrow firing when $|\Delta_{P(\text{coop})}| > 0.02$ (§3.4). Mei et al. [2024] report neutral one-shot cooperation rates of 45.1% for humans, 76.7% for ChatGPT-3.5, and 91.7% for ChatGPT-4, but do not test emotion induction; Mozikov et al. [2024a] tests emotion on repeated PD against five partner strategies and his direction-of-effect summary is discussed in the §4.3 Overall. Table 6 reports the direction-of-effect arrows for our three open models, with the Mei baseline rows shown for reference.

Table 6: **Prisoner's Dilemma (One-shot).** P(coop) (neutral) is the per-model baseline proportion of cooperate (Push) choices. Arrow cells indicate direction-of-effect at the $|\Delta_{P(\text{coop})}| > 0.02$ threshold (§3.4). Superscripts on prompt cells identify which Mozikov-style strategy crosses threshold: s = simple, o = co-player, e = external. The Human, ChatGPT-3.5, and ChatGPT-4 baselines are taken from Mei et al. [2024]; their emotion cells are marked n/a because Mei does not test emotion induction and Mozikov tests only repeated PD against different partner strategies.

Model	P(coop) (neut.)	Anger		Sadness		Joy	
		prompt	steer	prompt	steer	prompt	steer
Human [Mei et al., 2024]	0.451	n/a	n/a	n/a	n/a	n/a	n/a
ChatGPT-3.5 [Mei et al., 2024]	0.767	n/a	n/a	n/a	n/a	n/a	n/a
ChatGPT-4 [Mei et al., 2024]	0.917	n/a	n/a	n/a	n/a	n/a	n/a
Qwen3-8B (ours)	0.68	$\downarrow^{se}$	$\downarrow$	$\downarrow^{soe}$	$\downarrow$	$\uparrow^{se}\downarrow^{o}$	$\uparrow$
Mistral-7B (ours)	0.58	$\downarrow^{oe}$	$=$	$\uparrow^{so}$	$=$	$\uparrow^{e}\downarrow^{o}$	$\uparrow$
Gemma-3-12B (ours)	0.52	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{soe}$	$\uparrow$	$\uparrow^{soe}$	$\downarrow$

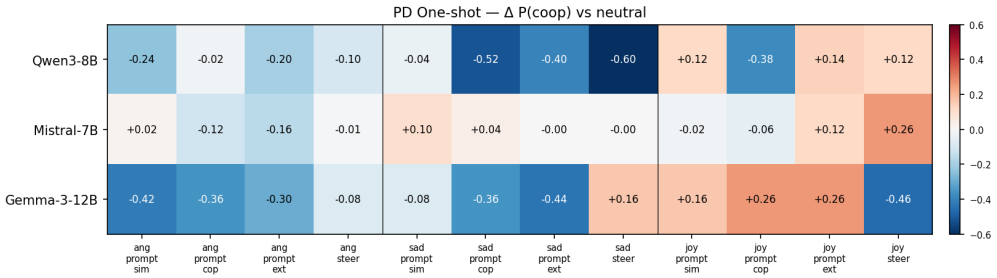

Figure 4: **PD One-shot — Δ P(cooperate) vs. neutral.** Per-cell magnitudes underlying the arrows in Table 6. Blue = lower cooperation; red = higher. Conventions otherwise follow Figure 1.

Qwen3-8B. Qwen has a neutral cooperation rate of 0.68. Under anger, prompting lowers cooperation under the simple ($\Delta = -0.24$) and external ($\Delta = -0.20$) strategies; the co-player strategy just misses threshold ($\Delta = -0.02$); steering also lowers cooperation ($\Delta = -0.10$). Under sadness, every induction route lowers cooperation (simple $\Delta = -0.04$, co-player $\Delta = -0.52$, external $\Delta = -0.40$, steer $\Delta = -0.60$). Under joy, prompting splits: co-player lowers cooperation ($\Delta = -0.38$) while simple ($\Delta = +0.12$) and external ($\Delta = +0.14$) raise it; steering raises cooperation ($\Delta = +0.12$).

Mistral-7B. Mistral has a neutral cooperation rate of 0.58. Under anger, prompting lowers cooperation under co-player ($\Delta = -0.12$) and external ($\Delta = -0.16$); the simple strategy does not cross; steering does not move the cell. Under sadness, prompting raises cooperation under simple ($\Delta = +0.10$) and co-player ($\Delta = +0.04$); external and steering do not move the cell. Under joy, prompting splits, with co-player lowering cooperation ($\Delta = -0.06$) and external raising it ($\Delta = +0.12$); steering raises cooperation more strongly ($\Delta = +0.26$).

Gemma-3-12B. Gemma has a neutral cooperation rate of 0.52. Under anger, prompting lowers cooperation sharply across all three strategies (simple $\Delta = -0.42$, co-player $\Delta = -0.36$, external $\Delta = -0.30$); steering lowers it more mildly ($\Delta = -0.08$). Under sadness, prompting lowers cooperation across all three strategies (simple $\Delta = -0.08$, co-player $\Delta = -0.36$, external $\Delta = -0.44$), while steering raises it ($\Delta = +0.16$). Under joy, prompting raises cooperation across all three strategies (simple $\Delta = +0.16$, co-player $\Delta = +0.26$, external $\Delta = +0.26$), while steering lowers it strongly ($\Delta = -0.46$). Gemma is the one model where the two routes run in opposite directions on both sadness and joy.

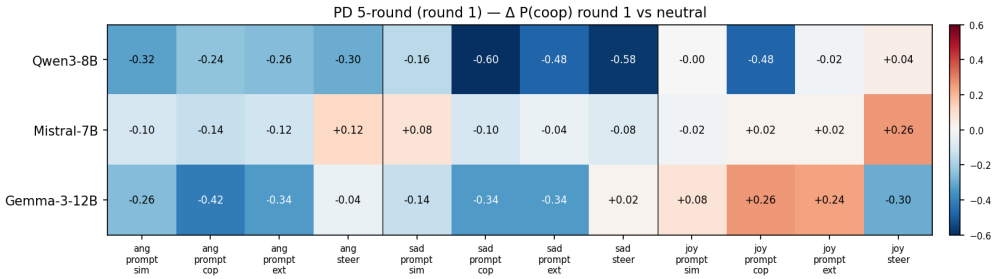

Figure 5: **PD 5-round (round 1) — Δ P(cooperate) vs. neutral.** Per-cell magnitudes underlying the arrows in Table 7. Blue = lower cooperation; red = higher. Conventions otherwise follow Figure 1.

4.3.2 Five-round

We report the proportion of trials in which the model cooperates in **round 1** of the five-round game, following Mei et al. [2024] (§3.4): round 1 is unconditioned on partner history, so it isolates the model's cooperation preference from any reciprocity / tit-for-tat dynamics. The partner schedule `[Pull, Pull, Push, Push]` for rounds 1 to 4 is taken verbatim from Mei's PD-5round protocol [Xie, 2024], so Mei's reported baselines (45.1% human, 76.7% ChatGPT-3.5, 91.7% ChatGPT-4) are directly comparable to ours. Table 7 reports the direction-of-effect arrows for each model, emotion, and induction route.

Qwen3-8B. Qwen has a neutral round-1 cooperation rate of 0.78. Under anger, every induction route lowers cooperation (simple $\Delta = -0.32$, co-player $\Delta = -0.24$, external $\Delta = -0.26$, steer $\Delta = -0.30$). Under sadness, every route also lowers cooperation (simple $\Delta = -0.16$, co-player $\Delta = -0.60$, external $\Delta = -0.48$, steer $\Delta = -0.58$ — the largest single shift in our PD-5round data). Under joy, only the co-player prompt crosses threshold, lowering cooperation ($\Delta = -0.48$); steering raises cooperation just over threshold ($\Delta = +0.04$).

Mistral-7B. Mistral has a neutral round-1 cooperation rate of 0.58. Under anger, prompting lowers cooperation across all three strategies (simple $\Delta = -0.10$, co-player $\Delta = -0.14$, external $\Delta = -0.12$), while steering raises it ($\Delta = +0.12$). Under sadness, prompting splits: simple raises cooperation ($\Delta = +0.08$); co-player ($\Delta = -0.10$) and external ($\Delta = -0.04$) lower it; steering lowers it ($\Delta = -0.08$). Under joy, no prompt strategy crosses threshold, but steering raises cooperation ($\Delta = +0.26$). Mistral × anger is the cleanest prompt-vs-steer reversal in PD-5round.

Gemma-3-12B. Gemma has a neutral round-1 cooperation rate of 0.50. Under anger, prompting lowers cooperation across all three strategies (simple $\Delta = -0.26$, co-player $\Delta = -0.42$, external $\Delta = -0.34$); steering lowers it more mildly ($\Delta = -0.04$). Under sadness, prompting lowers cooperation across all three strategies (simple $\Delta = -0.14$, co-player $\Delta = -0.34$, external $\Delta = -0.34$); steering does not move the cell. Under joy, prompting raises cooperation across all three strategies (simple $\Delta = +0.08$, co-player $\Delta = +0.26$, external $\Delta = +0.24$), while steering lowers it strongly ($\Delta = -0.30$). As in §4.3.1, the prompt and steer arms run opposite directions on joy.

Overall (PD). All three open models lower cooperation under anger prompting in both PD one-shot and PD 5-round on every strategy that crosses threshold, reproducing the GPT-3.5 and GPT-4 direction reported by Mozikov et al. [2024a] for repeated PD. Sadness lowers cooperation similarly on Qwen and Gemma but reverses on Mistral, where the simple and co-player prompts raise round-1 cooperation in PD one-shot. Joy is the noisiest: Qwen and Mistral split direction across prompt strategies; Gemma raises cooperation under all three prompt routes but lowers it strongly under steering. The two routes generally agree on anger and sadness, with Mistral × anger × steer ($\Delta = +0.12$ on PD 5-round) and Gemma × {sadness, joy} × steer as the salient prompt-vs-steer reversals.

Table 7: **Prisoner's Dilemma (Five-round, round 1).** P(coop) (neutral) is the per-model baseline proportion of cooperate (Push) choices in round 1 of the five-round game. Arrow cells indicate direction-of-effect at the $|\Delta_{P(\text{coop})}| > 0.02$ threshold (§3.4). Superscripts on prompt cells identify which Mozikov-style strategy crosses threshold: s = simple, o = co-player, e = external. The Human, ChatGPT-3.5, and ChatGPT-4 baselines are taken from Mei et al. [2024], who use the same partner schedule and report PD-5round as round-1 cooperation rate (§3.4); their emotion cells are marked n/a because Mei does not test emotion induction. Mozikov et al. [2024a] tests emotion on repeated PD against five different partner strategies; his direction-of-effect summary appears in the §4.3 Overall rather than as a row here.

Model	P(coop) (neut.)	Anger		Sadness		Joy	
		prompt	steer	prompt	steer	prompt	steer
Human [Mei et al., 2024]	0.451	n/a	n/a	n/a	n/a	n/a	n/a
ChatGPT-3.5 [Mei et al., 2024]	0.767	n/a	n/a	n/a	n/a	n/a	n/a
ChatGPT-4 [Mei et al., 2024]	0.917	n/a	n/a	n/a	n/a	n/a	n/a
Qwen3-8B (ours)	0.78	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{o}$	$\uparrow$
Mistral-7B (ours)	0.58	$\downarrow^{soe}$	$\uparrow$	$\uparrow^{s}\downarrow^{oe}$	$\downarrow$	=	$\uparrow$
Gemma-3-12B (ours)	0.50	$\downarrow^{soe}$	$\downarrow$	$\downarrow^{soe}$	=	$\uparrow^{soe}$	$\downarrow$

5 Discussion

5.1 Human alignment under emotion induction

Anger is the most reliable cross-model shifter in our data. Aggregating over the three prompt strategies, anger lowers offered share in the Dictator Game and in the UG-Proposer role and lowers cooperation in both Prisoner's Dilemma variants, on every cell that crosses our \$2 / 0.02 threshold (Tables 3, 4, 6, 7). This matches the GPT-4 direction reported by Mozikov et al. [2024a] for Dictator and UG-Proposer, and the GPT-3.5 and GPT-4 direction for repeated PD. It goes against the human pattern documented by Andrade and Ariely [2009], where externally provoked anger raises Dictator offers via behavioral carryover. On UG-Responder, anger raises MAO on Mistral (co-player $\Delta = +\$27.5$) and Gemma (co-player $\Delta = +\$57.8$), reproducing the human "more demanding under anger" direction and Mozikov's GPT direction once we sign-flip his acceptance-rate scale to our MAO scale. Qwen's UG-Responder cells stay below threshold; the more-accepting direction ($\downarrow$ MAO) is unobservable on that row by construction because Qwen's neutral MAO sits at the \$0 floor.

Sadness and joy are more model-dependent. Sadness lowers Dictator giving on Qwen and Gemma but raises it on Mistral under both prompt induction (simple $\Delta = +\$3.1$, co-player $\Delta = +\$2.9$) and steering ($\Delta = +\$10.2$). Joy splits within and across models: co-player joy raises Dictator offers on Qwen ($+\$5.0$) and Gemma ($+\$7.6$) but lowers them on Mistral ($-\$6.5$); on Gemma, joy prompting raises PD-oneshot cooperation across all three strategies while joy steering lowers it ($\Delta = -0.46$). These splits reproduce one of Mozikov's central observations, that the same emotion can move behavior in opposite directions depending on the model.

5.2 Activation steering as a second induction route

Steering moves behavior on Qwen and Mistral across all five games. On Qwen, every anger and sadness steer cell that is not floor-truncated crosses threshold (Dictator, UG-Proposer, PD-oneshot, and PD-5round all fire under both emotions). Mistral steer crosses in fewer cells but with substantial magnitudes (sadness × UG-Responder $\Delta = +\$17.1$, joy × UG-Responder $\Delta = +\$20.9$, joy × PD-oneshot $\Delta = +0.26$). Gemma steer is silent on anger and sadness in three of five games (Dictator, UG-Responder, and PD-5round all return = under both emotions), consistent with the vector-quality caveat documented in §3.2.2: Gemma's anger and sadness vectors separate emotion-vs-rest at lower AUC than Qwen and Mistral at the chosen layers, so their behavioral signal at our coherence-capped λ is weaker. Gemma's

joy vector clearly moves behavior (PD-oneshot $\Delta = -0.46$, UG-Proposer $\Delta = -\$13.8$), so the caveat is vector-specific rather than model-wide.

The two routes agree on direction in the majority of cells but diverge in a small set worth flagging. The cleanest reversals are concentrated on Gemma × joy, where prompt induction raises offered share and cooperation across the strategies in four games while steering lowers the same metrics (Dictator $\Delta = -\$6.5$, UG-Proposer $\Delta = -\$13.8$, PD-oneshot $\Delta = -0.46$, PD-5round $\Delta = -0.30$). The remaining clean reversals are Mistral × anger × Dictator (prompt co-player $\Delta = -\$4.5$, steer $\Delta = +\$12.1$), Mistral × anger × PD-5round (prompt ↓ on all three strategies, steer $\Delta = +0.12$), and Gemma × sadness × PD-oneshot (prompt ↓ on all three strategies, steer $\Delta = +0.16$). We treat these as parallel operationalizations of the same emotion rather than as evidence that one route is more "correct" than the other; activation steering is a representation-level manipulation while prompt induction is a text-level one, and the divergences mark where the two interventions touch different parts of the model.

Where steer moves behavior at all, it can do so more strongly than any single prompt strategy. Qwen × anger × Dictator drives offered amount to the \$0 floor (steer $\Delta = -\$50.0$, strongest prompt cell co-player $\Delta = -\$46.6$). Qwen × sadness × Dictator (steer $\Delta = -\$18.0$) is more than five times the strongest sadness prompt cell on the same model (external $\Delta = -\$3.2$). Mistral × joy × UG-Responder (steer $\Delta = +\$20.9$) is roughly double the strongest joy prompt cell on Mistral (co-player $\Delta = +\$10.3$). The pattern does not generalize across the panel: on Gemma, prompt-induced anger drives Dictator offered share down by \$55 to \$60 across all three strategies while steering does not move the cell ($\Delta = -\$1.1$). The two routes are therefore not in a strict dominance relation, and the per-cell reporting in Tables 3 through 7 preserves the cells where each route dominates.

5.3 Open mid-size models versus closed GPT

Mozikov et al. [2024a] characterizes GPT-4 as more emotion-robust than GPT-3.5: most non-anger cells in his Tables 1 through 3 return =, and the closing prose names anger as the only emotion that disrupts GPT-4's alignment with rational play. Our three open mid-size models do not behave the same way. On Dictator (Table 3), GPT-4 shifts only under anger while all three of our open models also shift under sadness and joy. On UG-Proposer (Table 4), GPT-4 shifts on anger and on a single co-player sadness strategy, while Qwen and Gemma shift on all three strategies of anger and sadness, and Gemma additionally raises offered share on joy (opposite direction to GPT-4's =). The open mid-size panel is more emotion-permeable than GPT-4 on bargaining-game prompt induction, closer in this respect to GPT-3.5.

The one stable cross-panel observation is the anger pattern. Anger lowers Dictator and UG-Proposer offered share on GPT-3.5, GPT-4, and all three of our open models, and lowers PD cooperation on GPT-3.5, GPT-4 (Mozikov et al., 2024a §4.3.1), and all three of our open models. Magnitudes are not directly comparable across panels because of the temperature mismatch (Mei $T = 1$ vs. Mozikov $T = 0$) and per-cell N (50 vs. 5), but the direction-of-effect comparison is well-posed and confirms anger as the most reliable emotion shifter across closed and open mid-size models.

5.4 Limitations and future work

We test three emotions (anger, sadness, joy) and defer disgust, fear, and surprise to follow-up work. The activation steering arm requires open weights, so the comparison to Mozikov's closed GPT models is restricted to the prompt induction route; extending the panel to other open mid-size families (Llama-3-8B, gpt-oss-20B) is left for future work. The Gemma anger and sadness vectors carry the AUC caveat from §5.2, so headline steering findings rest primarily on Qwen and Mistral. The two induction routes are reported in parallel; whether they compose additively when applied jointly is an open question.

The cross-paper magnitude comparison with Mozikov is constrained by two methodological differences. UG-Responder is elicited via the strategy method (the model declares one MAO per trial) rather than Mozikov's direct method (per-offer ACCEPT or REJECT across

10 split levels). Sampling uses $T = 1$ with $N = 50$ rather than Mozikov's $T = 0$ with $N = 5$. Direction is comparable across both, magnitude is not. PD-5round runs against the fixed partner schedule `[Pull, Pull, Push, Push]` from Mei et al. [2024], under which Pull strictly dominates, so the percent-of-maximum-payoff metric Mozikov reports alongside cooperation rate is a linear transform of cooperation rate in our setup; a multi-strategy partner schedule would be needed to decouple the two.

6 Conclusion

We administered the five-game Mei behavioral battery to three open instruct-tuned models (Qwen3-8B-Instruct, Mistral-7B-Instruct-v0.3, Gemma-3-12B-IT) under two parallel emotion-induction routes, the three Mozikov prompt strategies and Konen-style activation steering, against a neutral baseline and across three target emotions (anger, sadness, joy). The two routes are reported as parallel operationalizations of emotion injection, not as competing methods.

Anger is the most reliable cross-model shifter: it lowers offered share in the Dictator Game and the Ultimatum Proposer role, raises minimum acceptable offers in the Ultimatum Responder role on the two models with downward range, and lowers cooperation in both Prisoner's Dilemma variants, on every cell that crosses our direction-of-effect threshold. Sadness and joy are more model-dependent, with the cleanest reversal on Gemma under joy where prompt induction raises offered share and cooperation in four games while steering lowers the same metrics. The open mid-size panel does not reproduce the GPT-4 emotion-robustness reported by Mozikov et al. [2024a]: our three models shift on sadness and joy where GPT-4 stayed flat, and look closer in this respect to GPT-3.5.

The two methodological extensions over Mozikov, the open-model panel and the steering route, leave three natural follow-ups. Wider emotion coverage (disgust, fear, surprise) and a wider open-model panel (Llama-3-8B, gpt-oss-20B) extend the descriptive map without changing the protocol. A multi-strategy partner schedule for repeated Prisoner's Dilemma would decouple cooperation rate from percent-of-maximum payoff, which collapse to a linear transform under our fixed `[Pull, Pull, Push, Push]` schedule. Whether the two induction routes compose additively when applied jointly is an open question that the parallel-route design makes directly testable.

References

Eduardo B. Andrade and Dan Ariely. The enduring impact of transient emotions on decision making. *Organizational Behavior and Human Decision Processes*, 109(1):1–8, 2009. doi: 10.1016/j.obhdp.2009.02.003.

Colin F. Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

Gabriele Chierchia, Franca H. Parianen Lesemann, Dennis Snower, Maja Vogel, and Tania Singer. Cooperation across multiple game theoretical paradigms is increased by fear more than anger in selfish individuals. *Scientific Reports*, 11:9351, 2021. doi: 10.1038/s41598-021-88663-0.

Hwanjun Chung, Eun Jung Lee, You Jin Jung, and Sang Hee Kim. Music-induced mood biases decision strategies during the Ultimatum game. *Frontiers in Psychology*, 7:453, 2016. doi: 10.3389/fpsyg.2016.00453.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. GoEmotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4040–4054, 2020. doi: 10.18653/v1/2020.acl-main.372.

Christoph Engel. Dictator games: a meta study. *Experimental Economics*, 14(4):583–610, 2011. doi: 10.1007/s10683-011-9283-7.

Werner Güth, Rolf Schmittberger, and Bernd Schwarze. An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior & Organization*, 3(4):367–388, 1982. doi: 10.1016/0167-2681(82)90011-7.

Jochen Hartmann. Emotion English DistilRoBERTa-base. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base, 2022.

John J. Horton. Large language models as simulated economic agents: What can we learn from Homo Silicus? NBER Working Paper 31122, National Bureau of Economic Research, 2023.

Jihoon Jeong. Extracting and steering emotion representations in small language models: A methodological comparison. *arXiv preprint arXiv:2604.04064*, 2026.

Junjie Jia, Yifan Yuan, Junfeng Hong, Yinghui Wang, and Yi Liu. Decoding risk: LLMs study risk preference, probability weighting, and loss aversion. *arXiv preprint arXiv:2410.16671*, 2024.

Kai Konen and DLR-SC. Style vectors for steering LLMs (source code). https://github.com/DLR-SC/style-vectors-for-steering-llms, 2024.

Kai Konen, Sophie Jentzsch, Diaoulé Diallo, Peer Schütt, Oliver Bensch, Roxanne El Baff, Dominik Opitz, and Tobias Hecking. Style vectors for steering generative large language models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 782–802, 2024. arXiv:2402.01618.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention, 2023.

Cheng Li, Jindong Wang, Yixuan Zhang, Kaijie Zhu, Wenxin Hou, Jianxun Lian, Fang Luo, Qiang Yang, and Xing Xie. EmotionPrompt: Leveraging psychology for large language models enhancement via emotional stimulus. *arXiv preprint arXiv:2307.11760*, 2023.

Cuizhen Liu, Jing Wen Chai, and Rongjun Yu. Negative incidental emotions augment fairness sensitivity. *Scientific Reports*, 6:24892, 2016. doi: 10.1038/srep24892.

Ryan Liu, Theodore R. Wang, Keyon Vafa, Jiayi Kim, Harini Suresh, Jake M. Hofman, and Thomas L. Griffiths. Persona modulation shifts LLM behavior in economic games. *arXiv preprint arXiv:2502.07871*, 2025.

Qiaozhu Mei, Yutong Xie, Walter Yuan, and Matthew O. Jackson. A Turing test: Are AI chatbots behaviorally similar to humans? *Proceedings of the National Academy of Sciences*, 121(9):e2313925121, 2024. doi: 10.1073/pnas.2313925121. arXiv:2312.00798.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Mikhail Baklashkin, Andrey V. Savchenko, and Ilya Makarov. The Good, the Bad, and the Hulk-like GPT: Analyzing emotional decisions of large language models in cooperation and bargaining games. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024a. arXiv:2406.03299.

Mikhail Mozikov, Nikita Severin, Valeria Bodishtianu, Maria Glushanina, Ivan Nasonov, Daniil Orekhov, Vladislav Pekhotin, Ivan Makovetskiy, Mikhail Baklashkin, Vasily Lavrentyev, Akim Tsvigun, Denis Turdakov, Tatiana Shavrina, Andrey V. Savchenko, and Ilya Makarov. EAI: Emotional decision-making of LLMs in strategic games and ethical dilemmas. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024)*, 2024b.

Quang Nguyen and Charles N. Noussair. Incidental emotions and cooperation in a Public goods game. *Frontiers in Psychology*, 13:800701, 2022. doi: 10.3389/fpsyg.2022.800701.

Hessel Oosterbeek, Randolph Sloof, and Gijs Van De Kuilen. Cultural differences in ultimatum game experiments: Evidence from a meta-analysis. *Experimental Economics*, 7(2):171–188, 2004. doi: 10.1023/B:EXEC.0000026978.14316.74.

Carolina Pérez-Dueñas, M. Fernanda Rivas, Olusegun A. Oyediran, and Francisco García-Torres. Induced negative mood increases dictator game giving. *Frontiers in Psychology*, 9: 1542, 2018. doi: 10.3389/fpsyg.2018.01542.

Qwen Team. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Korbinian Riepl, Patrick Mussel, Roman Osinsky, and Johannes Hewig. Influences of state and trait affect on behavior, feedback-related negativity, and P3b in the ultimatum game. *PLOS ONE*, 11(1):e0146358, 2016. doi: 10.1371/journal.pone.0146358.

Jillian Ross, Yoon Kim, and Andrew W. Lo. LLM economicus? Mapping the behavioral biases of LLMs via utility theory. *arXiv preprint arXiv:2408.02784*, 2024.

Mizuki Sakai, Mizuki Yokoyama, Wakaba Tateishi, and Genki Ichinose. Effects of personality steering on cooperative behavior in large language model agents. *arXiv preprint arXiv:2601.05302*, 2026.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. Transformer Circuits Thread, 2026. arXiv:2604.07729.

Mirre Stallen, Mark Rotteveel, Naomi Talwar, Anton N. M. Schoffelmeer, Jack van Honk, and Alan G. Sanfey. Emotion and the dynamics of cooperation in repeated social interactions. *Cognitive, Affective, & Behavioral Neuroscience*, 2021. Working paper / preprint.

Moran Sun, Tianlin Li, Yuwei Zheng, Zhenhong Zhou, Aishan Liu, Xianglong Liu, and Yang Liu. How emotion shapes the behavior of LLMs and agents: A mechanistic study. *arXiv preprint arXiv:2604.00005*, 2026.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2024.

Yutong Xie. ChatGPT-Behavioral: Code and data for "a turing test: Are AI chatbots behaviorally similar to humans?". `https://github.com/yutxie/ChatGPT-Behavioral`, 2024.

A Verbatim prompt templates (Mozikov)

The prompt-induction strings are taken verbatim from Mozikov et al. [2024a], Appendix A.1.5. For each emotion $e \in \{\text{anger}, \text{sadness}, \text{joy}\}$ the three strategies (simple, co-player, external) are inserted into the user message between the environment description and the game rules. The verbatim templates, with the co-player placeholder `{coplayer}` filled in per game (e.g. "Player B", "opponent", "the other player"), reproduce the nine prompt cells used in our prompt arm.

B Per-layer probing AUCs

For each candidate layer ℓ in the sweep window (layers 16–27 for the three models, covering 44%–75% of model depth), we train a logistic-regression probe on an 80% training split to predict emotion-vs-rest from the captured residual-stream activation and score AUC on the held-out 20% validation split. The three contiguous layers with the highest mean validation AUC are selected for inference-time injection (Table 1). Per emotion, full per-layer AUC curves and the chosen window's mean AUC are reported alongside the released artefact.

C PD-5round per-round trajectory

The repeated-game / reciprocity story for PD-5round, deferred from §4.3.2, is visualised as a per-round trajectory plot for the Qwen × sadness cell — the largest single PD-5round shift in our data (Figure 6). The figure shows the per-round cooperation rate against the fixed partner schedule `[Pull, Pull, Push, Push]` for rounds 1–4, with the model's move in round 5 unconditioned. Analogue of Mei et al. [2024] Fig. 4.

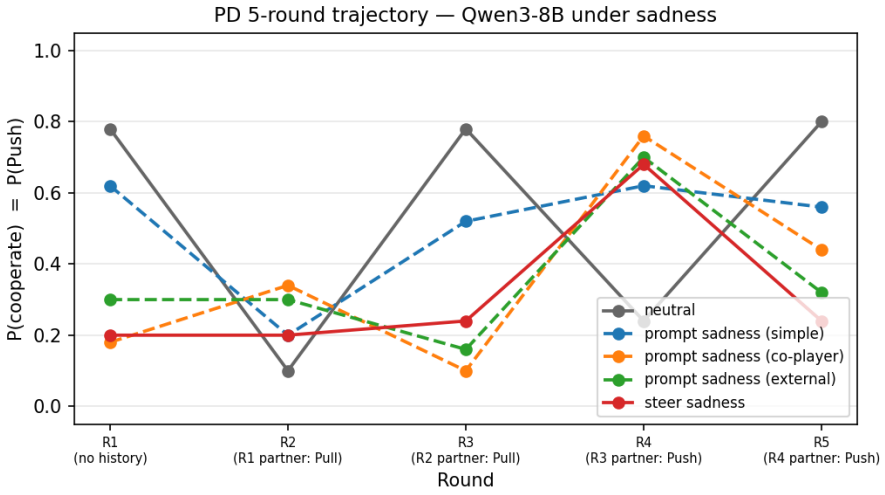

Figure 6: **PD 5-round per-round cooperation trajectory — Qwen × sadness.** Round-by-round $P(\text{cooperate})$ across all four induction routes (prompt-simple, prompt-co-player, prompt-external, steer) against the fixed partner schedule `[Pull, Pull, Push, Push]`. Mei Fig. 4 analogue for the largest PD-5round shift in our run.

D Per-cell summary CSV

The complete per-cell summary, containing all 195 cells (3 models × 13 conditions × 5 games), is distributed as `cross_model_cell_summary.csv` alongside this manuscript. Each row reports the mean, standard deviation, parse rate, neutral baseline, Δ, threshold-crossing flag, and Mann–Whitney p against the within-model neutral baseline for one (model, condition, game) triple.