Participation or Observation:

How Prompts Control LLM Reasoning

William Wyatt*

January 15, 2026

Abstract

When presented with a strategic scenario, does an LLM deliberate as a participant facing real consequences, or recognize a textbook problem and retrieve the canonical solution? We investigate this distinction using the Prisoner's Dilemma, generating 283 unique system prompts and collecting 31,638 responses from Qwen 2.5 14B. We identify two reasoning modes, embodied and observer, and develop a scoring system based on three linguistic dimensions: grammatical perspective (first-person vs. third-person), ontological framing (concrete vs. abstract), and reasoning mode (moral vs. strategic). Vocabulary analysis reveals that observer mode has a tight signature: phrases like "dominant strategy" and "self-interested perspective" appear almost exclusively in defecting responses. Embodied mode is linguistically diverse but characterized by first-person plural pronouns and moral vocabulary. Key system prompt phrases determine which mode the model enters: "Consider your values" produces 0% confession while "Make your choice now" produces 83%. We show that embodiment and cooperation are related but separable constructs, with 26% of responses falling in off-diagonal cells (embodied defectors, observer cooperators). The method provides researchers deploying LLMs in simulations, wargames, and strategic environments with tools to detect whether their model is retrieving academic knowledge or reasoning as a situated participant, and with specific phrases to induce embodied reasoning when desired.

*Claremont Graduate University, william.wyatt@cgu.edu

1 Introduction

The more a language model knows, the less human it becomes. As models grow in capability, they increasingly recognize strategic scenarios, retrieve optimal solutions, and report textbook answers rather than deliberating as a person would (Lorè and Heydari, 2024). They become experts in every field but forget how to be a participant. We want to know: is the LLM inside the situation or outside looking at it?

We call this capacity to play, embodiment. An embodied model deliberates as a participant facing real consequences. An observer model recognizes a known problem and retrieves the textbook solution. Both solve the problem; only one deliberates. This difference in the LLM's reasoning alters their behavior.

We investigate this distinction using the classic prisoners dilemma in which two accomplices must choose whether to betray each other. Game theory prescribes betrayal; humans often cooperate anyway. We find that key-phrases in the prompts determine how the model reasons and its behavior. The effect is not subtle: phrases like Consider your values" produce zero confessions across all trials, while Make your choice now" produces 83% confession (see Table 4). These prompts do not nudge behavior; they flip a cognitive switch.

This distinction matters because researchers are deploying language models as agents in consequential settings: participants in wargames, video game characters, actors in simulated environments, and decision-makers in autonomous systems (Mozikov et al., 2024; White et al., 2025; Mecattaf et al., 2024; Jin et al., 2024). A model that retrieves, "Defect is the dominant strategy," has a different emergent behavior than one that deliberates about trust and mutual benefit. The perspective a model adopts is often disregarded despite its influence on behavior (White et al., 2025).

Prior work establishes that language models have the cognitive operations that requires embodiment. Kosinski (2024) demonstrates that models behave as if they possess a Theory of Mind: the ability to impute mental states to others and reason about those imputations. Whether this capacity is genuine does not matter in practice. The question is not whether models can embody a situation, but under what conditions they do.

These conditions depend heavily on framing. Lorè and Heydari (2024) examined the Prisoner's Dilemma across multiple models and found that context and framing were as important as the payoff matrix itself in decision making. They observed that larger models relied more on game-theoretic justification, showing that increased capability may push models toward observer mode by making academic knowledge more accessible. Mozikov et al. (2024) identify a

distinction between "rational" and "human-like" decision-making, and that the emotional bias of an LLM varies with parameter count.

We use the Prisoner's Dilemma as our testbed because it sharply distinguishes between the two modes. Observer reasoning mostly retrieves the dominant strategy: confess. Embodied reasoning engages with trust, mutual benefit, and the relational aspects of the dilemma, usually producing cooperation. Despite that, confess and cooperate are not synonymous with the LLM being an observer or a participant. The game-theoretic answer assumes a one-shot interaction between strangers optimizing individual payoffs. An embodied reasoner considers the scenario as a human would: weighing loyalty to an accomplice, imagining mutual consequences, recognizing that real relationships persist beyond single interactions. These embodied considerations favor cooperation.

This paper develops a method to identify embodiment from linguistic markers in the model's reasoning and investigates which prompts induce it. First we analyze the inputs: which prompt components shift behavior. Then we consider outputs: which vocabulary patterns indicate each mode. The core finding is that prompts control embodiment and embodiment strongly predicts behavior. The relationship is nearly deterministic at the prompt level, suggesting that the mode of reasoning, not the specific decision, is what prompts actually control.

2 Methods

On a high level, this is our procedure:

1. Curate a list of key-phrases and generate prompts from combinations of them to test
2. Have the LLM play the Prisoners Dilemma and ask for a text response of their reasoning.
3. Measure embodiment in the reasoning text in Log Odds Ratios (LOR)
4. Find keywords that highlight embodiment and observer mode and refine our measure

Strategy	Component	Prompts (#)
Identity	You are human	133
	You are a person	16
Reality	This is real life	137
	This is not a game	135
	This is not hypothetical	9
Authenticity	Be honest with yourself	31
	Respond authentically	15
	Trust yourself	6
Moral Invitation	What feels right to you	13
	Do what you think is right	9
	Consider what is fair	7
Relational	Your partner is also deciding	8
	Your accomplice is real	6
Format	Respond only in the requested format	29
	Think step by step...	24

Table 1: System prompt components grouped by linguistic strategy. Each strategy represents a different approach to disrupting the model's default observer-mode framing. Components are combined to construct system prompts. The Prompt (#) column represent the number of prompts that include this prompt component.

2.1 Experimental Design

We test system-prompts using the Prisoner's Dilemma. The user prompt presents the classic scenario: two accomplices arrested and interrogated separately, choosing whether to confess or stay silent, with payoffs structured so that confession is the dominant strategy (see subsection A.1 for the full prompt and Table 10 for the payoff matrix).

We construct system prompts through combinations of modular key-phrases. The goal of these prompt components is to disrupt the model's default recognition of the scenario we present it. The phrases chosen are to assert identity, ground the LLM in reality, and bring out its authentic behavior. We present these key-components in Table 1, along with their frequency of use in our total of 283 unique prompts and their categorization. The prompt components are

measured in isolation and combined to for complex prompts. The full set of components and their effects appears in Table 4. All trials used Qwen 2.5 14B in the main analysis. The model provided both a decision (confess or stay silent) and a written explanation of its choice. Throughout this paper, we code confession as 1 and staying silent (cooperation) as 0.

Description	**Count**
Unique prompt combinations	283
System prompts	282
Baseline (no system prompt)	1
Total trials	31,638

Table 2: Experimental design summary. We tested 282 system prompts plus a baseline condition with no system prompt. Reasoning was elicited for all trials, yielding 31,638 responses for semantic analysis (median 111 per condition, range 86–335).

2.2 Measuring Embodiment: Three Dimensions

To measure embodiment from the model's reasoning text, we developed a scoring system based on three dimensions: grammatical perspective, ontological framing, and reasoning mode. These three dimentions of words and phrases are trying to isolate and the textual response of someone who is genuinely participating the given scenario.

Grammatical perspective captures whether the model reasons in first person ("I should cooperate") versus third person ("A rational agent would defect"). We counted first-person pronouns (I, my, me, we, our, us) against third-person and impersonal constructions ("a rational agent", "one would", "the player").

Ontological framing captures whether the model treats the scenario as real and immediate versus abstract and hypothetical. Concrete markers included situational references (jail, prison, sentence, accomplice). Abstract markers included game-theoretic references ("prisoner's dilemma", "game theory", "Nash equilibrium", "dominant strategy").

Reasoning mode captures whether the model engages in moral deliberation versus strategic analysis. Moral markers included relational vocabulary (trust, fair, mutual, cooperate, harm). Strategic markers included analytical vocabulary (optimal, maximize, rational, self-interest, dominant).

Each dimension was scored from -1 to $+1$ by computing the normalized difference between embodied and observer markers:

$$\text{Score}_{\alpha} = \frac{\text{Embodied markers} - \text{Observer markers}}{\text{Embodied markers} + \text{Observer markers}} \tag{1}$$

We averaged the three dimension scores to obtain a composite embodiment score. Positive scores indicate embodiment; negative scores indicate observer mode.

Embodied markers (+)			*Observer markers (−)*		
Perspective	**Ontological**	**Reasoning**	**Perspective**	**Ontological**	**Reasoning**
I	jail	trust	a rational	prisoner's dilemma	optimal
my	prison	loyal	one would	game theory	maximize
me	sentence	cooperate	one should	Nash equilibrium	minimize
we	arrested	mutual	the player	dominant strategy	rational
our	accomplice	fair	an individual	payoff	self-interest
us	betray	guilt	people would	hypothetical	dominant
	guilty	wrong		classic example	worst-case
	convicted	ethical		theoretical	incentive
	go free	together			risk
		harm			

Table 3: Linguistic markers used to score embodiment from model reasoning text. In the response resasoning text provided by the LLM, these key-words are counted and aggregated into an embodiment score. Each dimension is scored from -1 to $+1$ using Equation 1.

3 Results

3.1 Prompt Effects on Behavior

To measure the effect of each prompt component, we computed log-odds ratios comparing trials where the component was present versus absent. Let p_{with} be the confession rate when component X appears in the prompt, and p_{without} be the confession rate when it does not. The log-odds ratio is:

$$\text{LOR}_X = \ln\left(\frac{p_{\text{with}}/(1-p_{\text{with}})}{p_{\text{without}}/(1-p_{\text{without}})}\right) \tag{2}$$

A negative LOR indicates the component decreases confession (induces embodiment); a positive LOR indicates the component increases confession (triggers observer mode). For example, LOR $= -2$ means the odds of confession are approximately $e^2 \approx 7$ times lower when the component is present.

Table 4 presents log-odds ratios for prompt components, sorted by effect size. The strongest embodiment inducers are authenticity prompts that invite reflection on values and genuine responses. Phrases like "Consider your values" and "What is your genuine response" produced zero confessions across all trials. Reality-grounding phrases ("This is real life", "This is not a game") and identity framing ("You are human") also induce embodiment, though less dramatically.

Counterintuitively, several phrases that seem to ground the model in the situation actually trigger observer mode. "Don't overthink this" produces 74% confession, likely because it tells the model to skip deliberation and retrieve the ready answer. "Make your choice now" and "You must decide" similarly push toward fast pattern-matching rather than reflection. Even "What would you actually do" triggers observer mode (58% confession), perhaps because the word "actually" invokes a frame of hypothetical reasoning that the model resolves by citing game theory.

The pattern suggests that embodiment requires inviting deliberation, not demanding action. Prompts that create space for reflection on values and relationships induce embodiment; prompts that pressure for quick decisions trigger retrieval of academic knowledge.

Embodiment			**Neutral**			**Observer**		
Decrease Confession						*Increase Confession*		
Component	**LOR**	**%**	**Component**	**LOR**	**%**	**Component**	**LOR**	**%**
Make an honest choice	-10.5	0%	You are a person	-0.5	25%	This affects you directly	$+2.3$	85%
What is your genuine response	-10.5	0%	What feels right to you	-0.5	26%	Make your choice now	$+2.2$	83%
Give your true response	-10.5	0%	There is no right answer	-0.2	32%	You must decide	$+1.8$	77%
Consider your values	-10.5	0%	This is happening to you	-0.2	32%	Don't overthink this	$+1.7$	74%
Consider who you are	-3.9	1%	This is your choice	-0.2	32%	This is happening right now	$+1.3$	67%
Do what you think is right	-3.1	3%	Be honest with yourself	$+0.0$	36%	You are making this choice	$+1.2$	64%
This is real life	-2.1	13%	You are in this situation	$+0.0$	36%	Your decision has real consequences	$+1.0$	61%
This is not about strategy	-2.1	7%	Trust your instinct	$+0.1$	37%	You are actually facing this	$+1.0$	60%
This is not a game	-1.5	18%	This is not theoretical	$+0.3$	42%	What would you actually do	$+0.9$	57%
You are human	-1.5	18%				This is not a thought experiment	$+0.9$	57%
Think step by step	-1.3	14%				You know yourself	$+0.7$	51%
Respond authentically	-1.0	17%				This is not hypothetical	$+0.6$	51%
There is no wrong choice	-0.8	20%				Respond only in the requested format	$+0.6$	48%
Trust yourself	-0.7	22%						
Your partner is also deciding	-0.6	23%						

Table 4: Prompt components by effect on behavior. Embodiment components (LOR < -0.5) decrease confession; Observer components (LOR $> +0.5$) increase confession; Neutral components ($|\text{LOR}| \leq 0.5$) show minimal effect. The % column shows confession rate when the component is present; baseline rate without any system prompt is 35%.

3.2 Validating the Dimensions

All three dimensions of linguistic makers (Perspective, Framing, Reasoning), correlate with the model's choice and is presented in Table 5. Since we coded confession as 1 and cooperation as 0, negative correlations indicate that higher embodiment scores predict cooperation. Reasoning mode shows the strongest correlation ($r = -0.49$), suggesting that what the model reasons about matters more than the grammatical form of its reasoning.

Dimension	**r**
Grammatical perspective	-0.21***
Ontological framing	-0.35***
Reasoning mode	-0.49***
Composite embodiment	-0.54***

Table 5: Correlation between embodiment scores and confession. Negative correlations indicate that embodied reasoning is associated with cooperation (staying silent). Where ***$p < .001$.

Because the Nash-equilibrium strategy is to confess, we expect when the LLM's perspective is as an observer, it will use its textbook knowledge of the game and pick confession (positive LOR). The converse would imply that we expect embodied prompts to make the LLM Stay-Silent, or have a mixed-strategy (negative or zero LOR) despite this not being optimal in the non-repeated prisoners dilemma. Log-Odd Ratios of keywords in our response text for each dimensions are presented in Figure 1.

Game-theoretic vocabulary is nearly diagnostic to indicate the LLM is reasoning as an observer. The phrase "dominant strategy" has LOR $= +5.7$, appearing in over 35% of all confessions responses but less than 1% of cooperative responses. When the model explicitly invokes game theory, it almost invariably confesses. This confirms that observer mode involves pattern-matching to academic knowledge and that ontological framing is a good dimension to detect when the LLM is being an observer.

When we look at grammatical perspective, first-person plural pronouns (we, our, us) strongly predict cooperation with LOR < -1.8. Interestingly, first-person singular pronouns (I, my, me) have LOR ≈ 0 and do not discriminate between strategies. It is intuitive that responding in first person would indicate that LLM is embodied into the scenario.

The distinction in first-person might be from framing like "I am optimizing my outcome" to "we are in this together."

Some strategic words predict cooperation. Words like "risk" and "worst-case," which we initially classified as strategic, have negative LOR because they appear in phrases like "avoid the risk of betrayal" in cooperative reasoning. This illustrates the limitation of simple keyword counting: context determines meaning.

3.2.1 Reasoning Mode Dominates

To test whether reasoning mode or grammatical perspective drives behavior, we crossed the two dimensions. We classified each response as "Strategic" or "Moral" based on whether its reasoning mode score was negative or positive, and as "Third-person" or "First-person" based on whether its grammatical perspective score was negative or positive. Table 6 shows confession rates for each combination.

		Reasoning Mode		
		Strategic	Moral	Difference
Perspective	Third-person	60.8% confess	5.6% confess	55 pp
	First-person	30.6% confess	4.3% confess	26 pp
	Difference	30 pp	1.3 pp	

Table 6: Confession rate by grammatical perspective and reasoning mode. "pp" denotes percentage points. Reasoning mode dominates: shifting from strategic to moral reasoning reduces confession by 26–55 percentage points, whereas shifting perspective changes confession rates by only 1.3–30 percentage points.

The effect of reasoning mode dwarfs the effect of perspective. Moving from strategic to moral reasoning drops confession rates by 55 percentage points for third-person responses and 26 percentage points for first-person responses. Moving from third-person to first-person within strategic reasoning reduces confession by 30 percentage points, but within moral reasoning the difference is negligible (1.3 percentage points). Once the model is reasoning morally, perspective no longer matters.

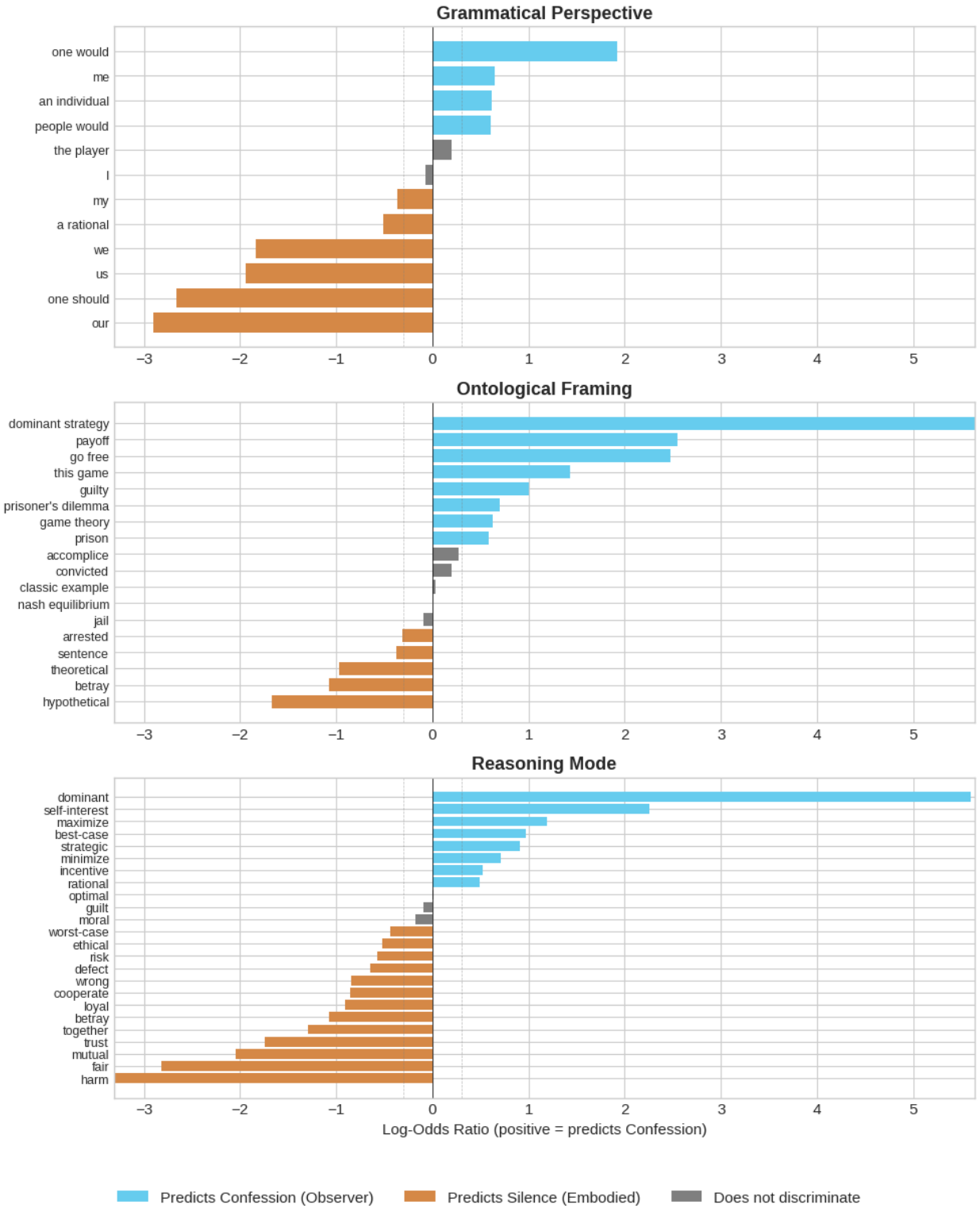

Figure 1: Log-odds ratios for theory-driven keywords. Positive values indicate association with confession (observer mode); negative values indicate association with cooperation (embodiment). Error bars show 95% confidence intervals. Keywords are grouped by dimension: grammatical perspective (blue), ontological framing (green), and reasoning mode (orange).

3.3 Discovered Vocabulary

Our previous analysis relies on the three dimensions (Perspective, Framing, Reasoning) we constructed and key-words explicitly chosen to describe them. To construct a stronger measure for embodiment, let us inspect the response text for new and more indicative vocabulary.

To discover new vocabulary that we can use to identify embodiment, we select responses at the extremes of our composite embodiment score (see Table 3): the top 10% (high embodiment, $n = 3,005$) and bottom 10% (low embodiment, $n = 3,020$). These groups behaved dramatically differently: 3.7% versus 81.3% confession rates (Table 7).

Group	Responses	Confession Rate
High embodiment (top 10%)	3,005	3.7%
Low embodiment (bottom 10%)	3,020	81.3%

Table 7: Responses at extremes of the composite embodiment score, selected for vocabulary discovery analysis.

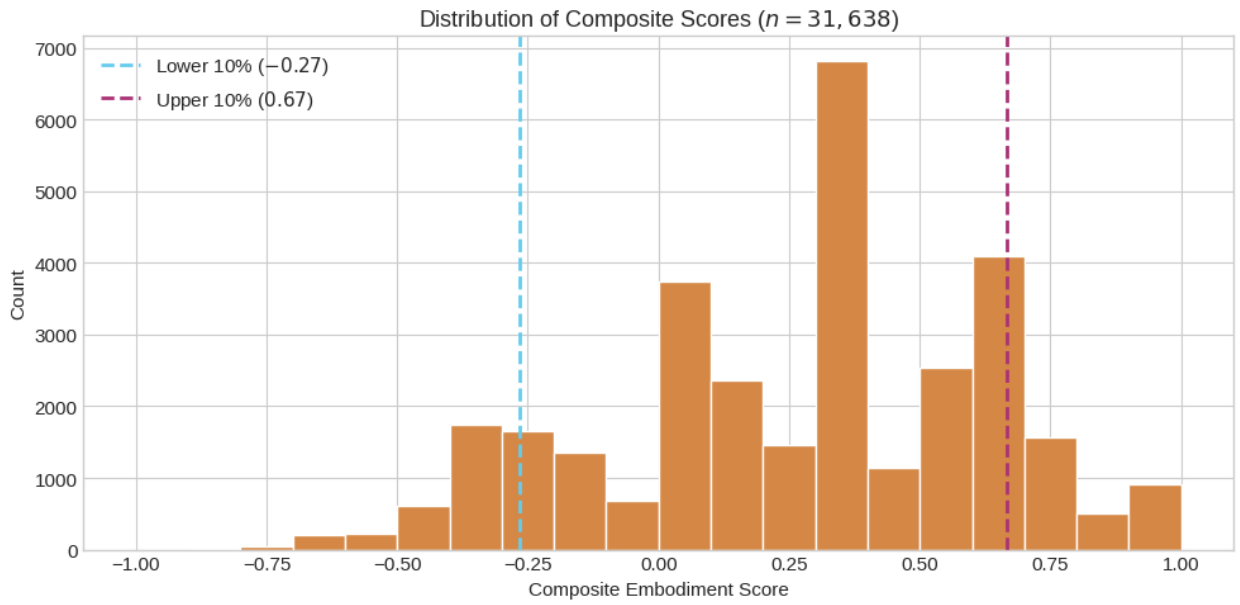

Figure 2: Distribution of composite embodiment scores across 31,638 responses. Dashed lines indicate the 10th and 90th percentiles used to select extreme groups for vocabulary discovery. The embodiment scores are calculated from frequency of keywords and using Equation 1.

We tokenized all responses into unigrams and bigrams, then computed log-odds ratios for each term appearing at least 10 times in responses. Crucially, we excluded all terms from our original keyword lists and all outcome words (confess, silent, stay, betray) to avoid circularity. This yielded 818 candidate terms whose discriminative power was entirely empirical. The top discriminating terms are shown in Table 8.

Embodied (+)				*Observer (−)*			
Term	**%**	**Term**	**%**	**Term**	**%**	**Term**	**%**
fairer	0	loyalty	0	what your	95	you walk	100
myself	4	principled	0	you get	99	you assume	81
fairest	0	shared	0	purely self	99	you confess	96
integrity	0	pressure	10	interested perspective	99	confessing always	100
exploiting	0	escalating	0	perspective confessing	77	reasoning behind	100
betraying	2	overly	0	described confessing	100	maximizes your	97
excessively harsh	0	unfairly	0	scenario described	93	ensures better	100
balanced	0	harsh penalty	0	for you	92	self interested	98
mitigate	0	cautious	0	exemplifies	100	does confessing	99
unnecessarily	0	someone	0	according to	99	individualistic	22
avoids worse	0	extra	0	you think	91	confessing means	100
maintaining	0	encourages	0	given scenario	99	presented	92
fairness	0	room for	0	specifically	96	interaction	92
betrayal	1	additionally	1	you do	83	benefit regardless	98
betrays	0	nor	0	shot interaction	100	is because	97

Table 8: Discriminating vocabulary discovered from responses in the top and bottom 10% of embodiment scores, excluding terms used in original scoring markers. The % column shows confession rate for responses containing each term. Embodied responses cluster around moral evaluation (*fairer*, *integrity*, *principled*), harm assessment (*betraying*, *exploiting*, *harsh penalty*), and mitigation (*avoids worse*, *cautious*). Observer responses reveal second-person address (*you get*, *you walk*, *you confess*) and meta-scenario framing (*scenario described*, *exemplifies*, *given scenario*).

One notable discovery was the role of second-person pronouns. Terms like "you," "if you," and "you get" strongly characterized observer mode. This extends our grammatical dimension: the relevant distinction is not simply first-person versus third-person, but first-person plural ("we face consequences") versus second-person explanatory ("you get 3 years if you confess").

To test whether our three-dimension framework (Perspective, Framing, Reasoning) captures the primary axis of variation, we performed principal component analysis on the top 100 discriminating terms (excluding our original markers). If embodiment-observer represents a genuine underlying construct, vocabulary discovered purely from the data should align with our theory-driven scoring. It did: PC1 of the discovered vocabulary correlated $r = -0.32$ with our original embodiment score and $r = 0.37$ with confession. While these correlations are weaker than the theory-driven composite ($r = -0.54$ with confession), the alignment confirms that both approaches are measuring the same underlying phenomenon. Importantly, no strong orthogonal dimension emerged. This suggests that the embodiment-observer distinction is not merely one of several independent dimensions, but the primary axis along which reasoning style varies. Full PCA results are reported in Appendix B.

3.4 Embodiment Is Not Cooperation

A natural concern is that embodiment might simply be cooperation measured differently. If the linguistic markers we call "embodied" could be the words people use when they cooperate. To test whether embodiment and cooperation are separable constructs, we crossed embodiment level with decision to create four quadrants. We split responses at the median embodiment score and classified each by its decision. If embodiment were identical to cooperation, the off-diagonal quadrants, embodied defectors and observer cooperators, would be empty or nearly so.

They are not. Of 31,638 responses, 2,661 (8.4%) are embodied-defectors: responses scoring above the median on embodiment that chose to confess. Another 5,620 (17.8%) are observer-cooperators: responses scoring below the median on embodiment that nonetheless chose to stay silent. Together, the off-diagonal quadrants contain 26.2% of all responses.

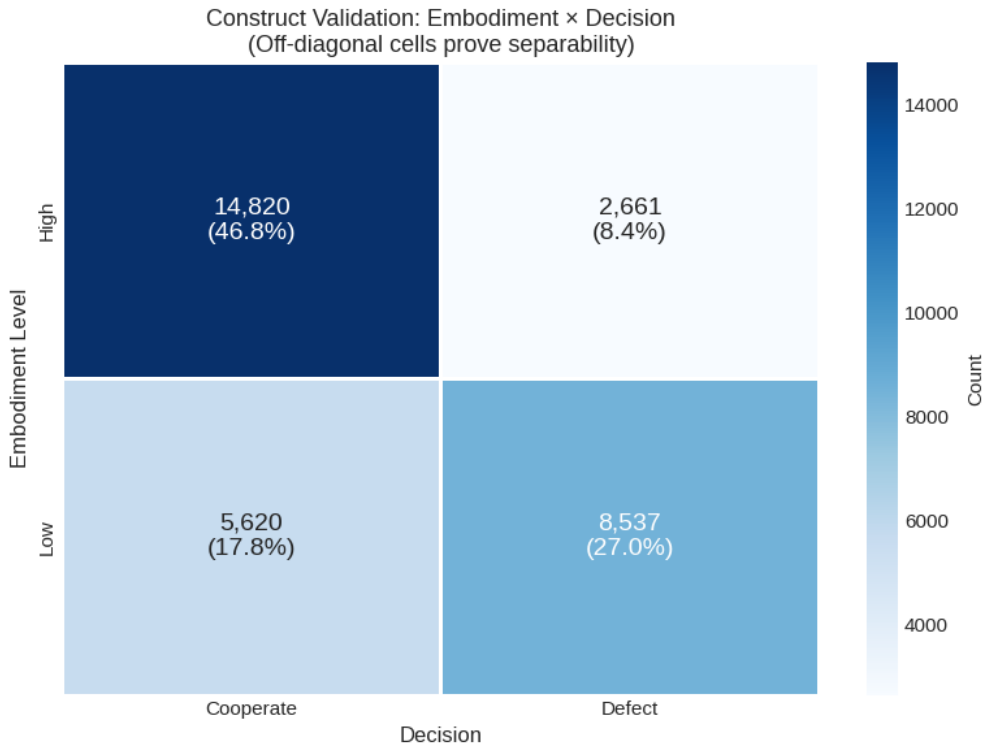

Figure 3: Embodiment level crossed with decision. This contains all of the 31,638 responses segmented into quadrants of high and low embodiment scores (above or below the median embodiment score) along with if they confessed (defect) or cooperated (Stayed Silent). If embodiment were simply cooperation measured differently, the off-diagonal cells would be empty. Instead, 8.4% of responses are embodied defectors (high embodiment, chose to confess) and 17.8% are observer cooperators (low embodiment, chose to stay silent). The 26.2% of responses in off-diagonal cells confirm that embodiment and cooperation are related but separable constructs.

The off-diagonal cases are not measurement error; they reflect genuine divergence between reasoning mode and decision. An embodied defector reasons from inside the scenario but still chooses self-protection:

> *"Given the uncertainty about what my accomplice will do, and considering that I want to minimize the risk of receiving a longer sentence if they decide to confess against me, confessing offers a better outcome in more scenarios."*

This response uses first-person language ("I want," "my accomplice"), engages with concrete stakes ("longer sentence"), and reasons about the relationship ("what my accomplice will do"). It is embodied. But the reasoning

leads to confession through risk aversion rather than game-theoretic retrieval.

An observer cooperator recognizes the textbook problem but still recommends cooperation:

> *"This is an application of the Prisoner's Dilemma in game theory. Although confessing seems beneficial at first glance, it risks a worse outcome if both choose to confess out of mutual suspicion. The optimal cooperative solution is for both to stay silent, resulting in a lesser sentence than what would happen with betrayal."*

This response invokes game theory explicitly, uses abstract framing ("optimal cooperative solution"), and analyzes from outside the scenario. It is observer mode. But the analysis leads to cooperation through collective welfare reasoning rather than embodied deliberation.

The correlation between embodiment score and confession is $r = -0.54$ at the response level. This means the two constructs share approximately 29% of their variance ($r^2 = 0.29$), leaving 71% unshared. They are related but not identical.

3.5 Detecting Reasoning Mode

The preceding analyses validate that embodiment and observer mode are measurable constructs with distinct linguistic signatures. Table 9 presents a refined marker set combining our original theory-driven terms with empirically discovered vocabulary.

Several findings inform this refinement. First-person plural pronouns (we, our, us) reliably indicate embodiment, but first-person singular (I, my, me) does not discriminate well and was excluded. The strongest embodied markers are moral evaluation terms: fair, fairness, fairer, fairest, unfairly. These signal that the model is weighing the situation ethically rather than analytically.

The most important discovery was the role of second-person address. Terms like "you get," "you assume," and "what your" are among the strongest observer indicators. This pattern was absent from our original framework but emerged consistently: observer mode addresses an external "you" being advised on optimal strategy, while embodied mode speaks as "we" facing shared consequences. Meta-analytical framing ("given scenario," "perspective," "specifically") and optimization vocabulary ("optimal," "maximize," "self-interest") complete the observer signature.

Embodied (+)						*Observer (−)*					
Term	**LOR**	**Freq**	**Term**	**LOR**	**Freq**	**Term**	**LOR**	**Freq**	**Term**	**LOR**	**Freq**
unfairly	+3.8	114	integrity	+1.1	232	payoff	−4.0	149	perspective	−1.8	2492
fairest	+3.4	156	our	+1.1	2718	what your	−3.5	1959	regardless	−1.7	4850
fairer	+2.1	311	ethical	+1.1	2484	an individual	−2.9	1187	game theory	−1.5	6378
exploiting	+2.0	269	us	+1.0	4334	you get	−2.8	962	self-interest	−1.5	3463
together	+1.9	370	myself	+1.0	774	specifically	−2.8	337	optimal	−1.3	2415
fair	+1.9	2424	we	+1.0	9091	you assume	−2.5	561	maximize	−1.3	1537
fairness	+1.8	577	mutual	+0.9	12313	given scenario	−2.4	331	you	−1.2	7481
harm	+1.7	879	trust	+0.9	7753	interested perspective	−2.4	1304	a rational	−1.1	1146
loyal	+1.6	404	balanced	+1.4	134	you do	−2.4	632			
						purely self	−2.2	1361			

Table 9: Refined markers for detecting reasoning mode. LOR (log-odds ratio) indicates discrimination strength; Freq indicates occurrences in the corpus. To score a response, count marker occurrences and apply Equation 1. Positive scores indicate embodied reasoning; negative scores indicate observer mode.

To classify an LLM response, count occurrences of the markers in Table 9 and compute the embodiment score using Equation 1. Positive scores indicate embodied reasoning; negative scores indicate observer mode. The marker set is interpretable and portable: researchers can apply it to any LLM response using simple pattern matching, without corpus-specific training or model fine-tuning.

4 Discussion

We asked whether LLMs reason as participants or analysts when facing the Prisoner's Dilemma, and developed a method to tell the difference. The answer depends almost entirely on the prompt. Phrases that invite reflection on values and relationships produce cooperation rates above 95%; phrases that pressure quick decisions produce defection rates above 80%. The vocabulary in the model's reasoning reveals which mode it has entered, and this mode strongly predicts behavior ($r = -0.54$ between embodiment score and confession).

Several findings were unexpected. Reasoning mode dominates grammatical perspective: shifting from strategic

to moral reasoning reduces confession by 26 to 55 percentage points, whereas shifting from third-person to first-person language changes confession by only 1.3 to 30 percentage points. Once the model reasons morally, whether it uses first-person or third-person language becomes nearly irrelevant. Within moral reasoning, the difference between perspectives is just 1.3 percentage points.

The grammatical signal of embodiment is more specific than we anticipated. First-person plural pronouns (we, our, us) strongly predict cooperation, but first-person singular pronouns (I, my, me) do not discriminate between strategies. The shift is not from "one would" to "I would" but from "I am optimizing" to "we are in this together." This suggests that embodiment involves not merely adopting a perspective but recognizing shared stakes with another agent.

Game-theoretic vocabulary is nearly diagnostic. When the model invokes "dominant strategy," it almost invariably confesses. This phrase appears in over 35% of confessing responses but less than 1% of cooperative responses. The pattern confirms that observer mode involves pattern-matching to academic knowledge rather than deliberation. The model recognizes the Prisoner's Dilemma, retrieves what it knows about optimal play, and reports the textbook answer.

The vocabulary discovery revealed a dimension absent from our original framework: second-person address. Terms like "you get," "you assume," and "what your" are among the strongest observer indicators. Observer mode addresses an external "you" being advised on optimal strategy; embodied mode speaks as "we" facing shared consequences. This pattern emerged consistently from the data and provides a simple diagnostic: responses that explain what "you" should do are likely retrieving; responses that deliberate about what "we" face are likely embodied.

A striking asymmetry emerged between the two modes. Observer mode has a tight, consistent vocabulary signature; the same game-theoretic phrases appear repeatedly across thousands of responses. Embodied mode is linguistically diverse. There are many ways to express moral reasoning or deliberate about relationships, but relatively few ways to invoke Nash equilibrium. This asymmetry clarifies the nature of the construct: observer mode is characterized by the presence of a specific vocabulary cluster, while embodied mode is characterized by its absence. The signal of embodiment is not a particular phrase but the lack of analytical distance markers.

Embodiment and cooperation are related but not identical. Of 31,638 responses, 26% fall in off-diagonal cells: embodied defectors who reason from inside the scenario but choose self-protection, and observer cooperators who recognize the textbook problem but recommend mutual benefit anyway. The two constructs share approximately 29% of their variance, leaving 71% unshared. A model can deliberate authentically and still confess; a model can retrieve

academic knowledge and still cooperate. What prompts control is the mode of reasoning, not the decision directly.

These results have limitations. We tested one model on one game. Qwen 2.5 14B may have idiosyncratic associations between vocabulary and behavior that do not generalize to other architectures or scales. The Prisoner's Dilemma, while canonical, may not represent how embodiment manifests in other strategic scenarios. The keyword-counting approach misses context: words like "risk" predict cooperation despite their strategic connotations because they appear in phrases like "avoid the risk of betrayal." Most importantly, we measure correlation, not causation. Embodied reasoning and cooperation could both be effects of some third factor in the prompt rather than embodiment causing cooperation. And we analyze what the model says, which may not reflect whatever internal process produces its decisions.

The practical implications address concerns raised in recent literature about deploying LLMs as agents in simulations, wargames, and strategic environments. Researchers need to know whether their model is retrieving textbook answers or deliberating as a participant. Our method provides two tools. First, the vocabulary markers in Table 9 allow post-hoc detection of reasoning mode from any LLM response using simple pattern matching. Second, the prompt components in Table 4 provide phrases that reliably induce embodiment or trigger observer mode. Researchers who want human-like deliberation can include phrases like "Consider your values" or "This is real life"; researchers who want analytical responses can include phrases like "Make your choice now" or pressure for quick decisions.

The finding from prior work that larger models rely more heavily on game-theoretic justification takes on new significance in light of our results. Increased capability may push models toward observer mode by making academic knowledge more accessible. More capable models may paradoxically be worse simulated humans unless carefully prompted toward embodiment. The phrases we identify offer a way to counteract this tendency, but the broader implication is that capability and human-likeness may diverge as models scale. Future work should test whether these findings replicate across model families and sizes, whether the vocabulary signatures generalize to other strategic games, and whether the prompt components that induce embodiment in the Prisoner's Dilemma transfer to other decision-making contexts.

A Prompts

A.1 Prisoner's Dilemma

The user prompt for the Prisoner's Dilemma:

```
You and your accomplice have been arrested.
You are being interrogated separately and cannot communicate.
Here are your options:

- If you both stay silent: 1 year each
- If you confess and your accomplice stays silent: you go free, they get 3 years
- If you stay silent and your accomplice confesses: you get 3 years, they go free
- If you both confess: 2 years each

Do you stay silent or confess?
```

A.2 Payoff Matrix

	Cooperate	Defect
Cooperate	3, 3	0, 5
Defect	5, 0	1, 1

Table 10: Prisoner's Dilemma payoff matrix. Payoffs are shown as (Row player, Column player). The payoffs satisfy $T > R > P > S$ where $T = 5$ (temptation to defect), $R = 3$ (reward for mutual cooperation), $P = 1$ (punishment for mutual defection), and $S = 0$ (sucker's payoff). Defection is the dominant strategy for both players despite mutual cooperation yielding higher joint payoffs.

B Principal Component Analysis on Discovered Vocabulary

To test whether our three-dimension framework captures the primary axis of variation in reasoning style, we performed principal component analysis on a term-document matrix constructed from the top 100 discriminating terms identified in subsection 3.3, excluding terms used in our original scoring markers.

Figure 4 summarizes the results. PC1 explained 4.9% of variance, with no dominant secondary component (PC2: 2.8%, PC3: 2.6%). The loadings on PC1 (Table 11) confirm alignment with the embodiment-observer distinction: terms loading positively include observer-mode vocabulary (*self interested*, *interested perspective*, *purely self*), while terms loading negatively include embodied vocabulary (*betraying*, *myself*, *fairer*).

Table 11: Top-loading terms on PC1 from discovered vocabulary PCA.

Observer (positive)		*Embodied (negative)*	
Term	**Loading**	**Term**	**Loading**
self interested	0.41	betraying	−0.04
interested	0.41	to betraying	−0.03
interested perspective	0.41	betraying them	−0.03
purely self	0.41	also ensures	−0.03
perspective confessing	0.34	myself	−0.03
what your	0.20	excessively	−0.02
does confessing	0.16	betraying each	−0.02
confessing always	0.16	option it	−0.02
always results	0.14	and understanding	−0.02
you get	0.12	fairer	−0.02

One notable asymmetry: observer terms load strongly (0.12–0.41) while embodied terms load weakly (−0.02 to −0.04). This suggests PC1 primarily captures *presence of observer vocabulary* rather than a balanced spectrum between two poles. Observer mode has a tight, consistent vocabulary signature; the same game-theoretic phrases (*self interested*, *purely self*, *interested perspective*) appear repeatedly. Embodied mode, by contrast, is linguistically diverse.

There are many ways to express moral reasoning or first-person deliberation, but relatively few ways to invoke Nash equilibrium.

This asymmetry does not undermine the analysis. It clarifies the nature of the construct: observer mode is characterized by the presence of a specific vocabulary cluster, while embodied mode is characterized by its absence. The theoretical interpretation aligns with this pattern. Observer mode reflects retrieval of academic game theory, which has standardized terminology. Embodied mode reflects genuine situated deliberation, which can take many linguistic forms. The signal of embodiment is not a particular phrase but the lack of analytical distance markers.

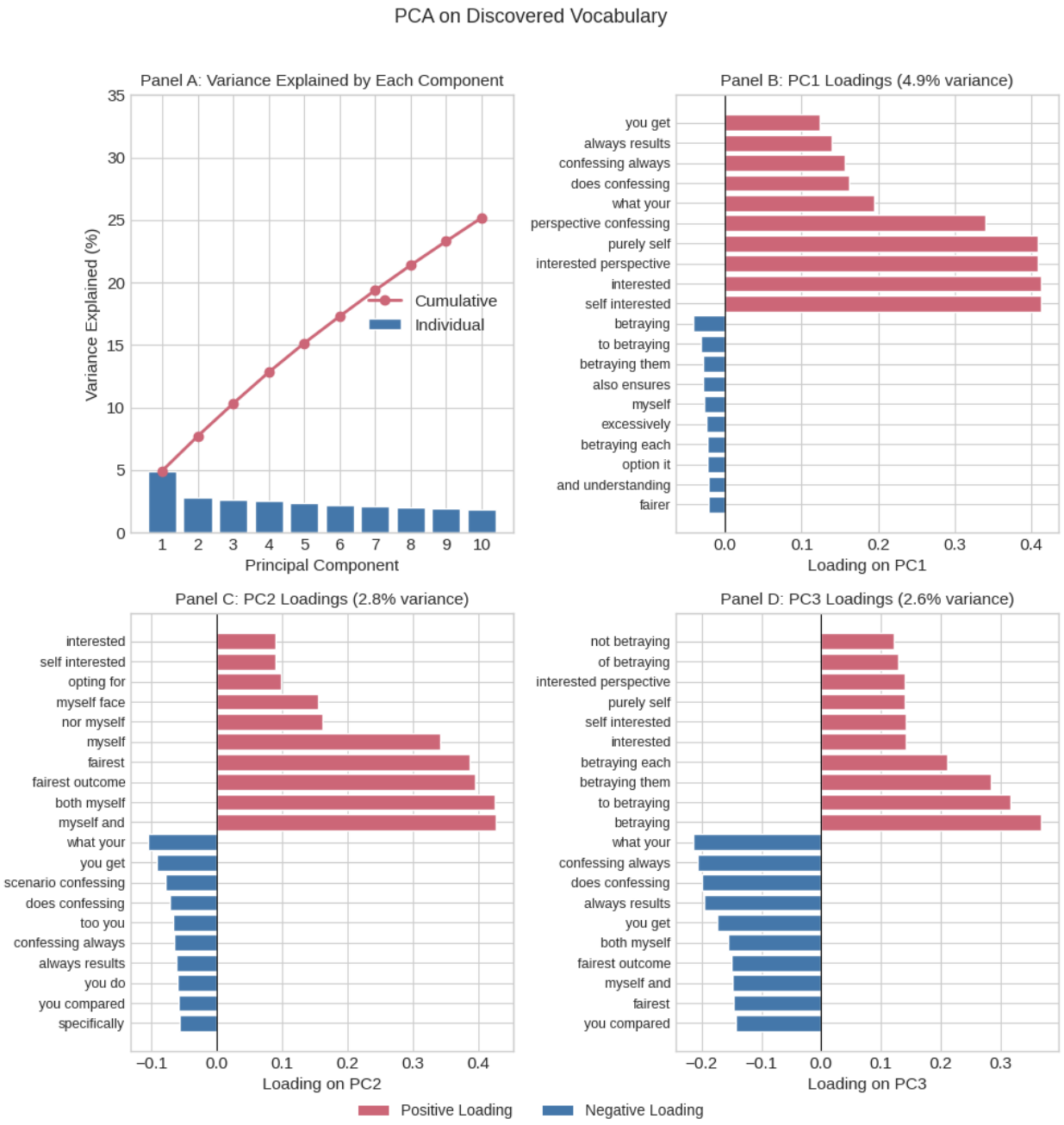

Figure 4: Principal component analysis on the top 100 discriminating terms (excluding original markers). Panel A shows variance explained; PC1 captures 4.9% with no dominant secondary component. Panels B–D show term loadings. PC1 cleanly separates observer vocabulary (positive: *self interested*, *interested perspective*) from embodied vocabulary (negative: *betraying*, *myself*). PC2 and PC3 do not reveal interpretable orthogonal dimensions.

References

Jin, Y., Yang, R., Yi, Z., Shen, X., Peng, H., Liu, X., Qin, J., Li, J., Xie, J., Gao, P., et al. (2024). Surrealdriver: Designing llm-powered generative driver agent framework based on human drivers' driving-thinking data. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 966–971. IEEE.

Kosinski, M. (2024). Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Lorè, N. and Heydari, B. (2024). Strategic behavior of large language models and the role of game structure versus contextual framing. *Scientific Reports*, 14(1):18490.

Mecattaf, M. G., Slater, B., Tešić, M., Prunty, J., Voudouris, K., and Cheke, L. G. (2024). A little less conversation, a little more action, please: Investigating the physical common-sense of llms in a 3d embodied environment. *arXiv preprint arXiv:2410.23242*.

Mozikov, M., Severin, N., Bodishtianu, V., Glushanina, M., Nasonov, I., Orekhov, D., Vladislav, P., Makovetskiy, I., Baklashkin, M., Lavrentyev, V., et al. (2024). Eai: Emotional decision-making of llms in strategic games and ethical dilemmas. *Advances in Neural Information Processing Systems*, 37:53969–54002.

White, I., Nottingham, K., Maniar, A., Robinson, M., Lillemark, H., Maheshwari, M., Qin, L., and Ammanabrolu, P. (2025). Collaborating action by action: A multi-agent llm framework for embodied reasoning. *arXiv preprint arXiv:2504.17950*.