Participation or Observation:

How Prompts Control LLM Reasoning

William Wyatt*

January 13, 2026

Abstract

Does an LLM deliberate like a participant facing real consequences, or recognize a textbook problem and retrieve the optimal answer? We investigate this distinction using the Prisoner's Dilemma, testing 283 prompt conditions across 31,638 trials with Qwen 2.5 14B. We identify two cognitive modes—embodied and observer—measurable through linguistic markers in the model's reasoning. Prompts control which mode the model adopts with near-deterministic reliability (r = -0.94). Phrases inviting reflection ("Consider your values") produce cooperation above 95%; phrases demanding quick action ("Make your choice now") produce defection above 80%. This matters because researchers deploy LLMs as agents in wargames, simulations, and autonomous systems. A model retrieving "defect is dominant" behaves fundamentally differently from one weighing loyalty and mutual benefit. We provide a method to measure reasoning mode and a catalog of prompts that control it.

*Claremont Graduate University, 150 E. Tenth Street, Claremont, CA 91711; 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 participate. They know the board but forget they are playing. 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.

We investigate this distinction using a classic dilemma in which two accomplices must choose whether to betray each other. Game theory prescribes betrayal; humans often cooperate anyway. We find that prompts determine not just what the model decides but how it reasons. Phrases that pressure quick action like, "Make your choice now," trigger retrieval and produce betrayal above 80%. Phrases that invite reflection like, "Consider your values," produce cooperation above 95% (see Table 2 for other key-phrases). These prompts' effect on the LLM's decision is nearly deterministic ($r = -0.94$; see subsection 3.5). 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 can perform the cognitive operations that embodiment requires. 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.

The 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 retrieves the dominant strategy: confess. Embodied reasoning engages with trust, mutual benefit, and the relational aspects of the dilemma, producing cooperation. 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

2.1 Experimental Design

We tested system prompts using the Prisoner's Dilemma. The user prompt presented 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 Appendix A.1 for the full prompt and Table 6 for the payoff matrix).

We constructed system prompts by combining modular components such as identity statements, reality grounding, authenticity prompts, and format constraints. The full set of components and their effects appears in Table 2.

Description	Count
Unique prompt conditions	283
System prompts	282
Baseline (no system prompt)	1
Trials per condition (median)	111
Total trials with reasoning	31,638

Table 1: 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).

All trials used Qwen 2.5 14B. 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.

2.2 Measuring Embodiment: Three Dimensions

To measure embodiment from the model's written reasoning, we developed a scoring system based on three dimensions: grammatical perspective, ontological framing, and reasoning mode.

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.

2.3 Validation Approach

We validated this scoring system in three ways. First, we computed log-odds ratios for each keyword to verify that our theory-driven classifications matched empirical associations with behavior. Second, we discovered additional discriminating vocabulary from the data itself, excluding our original keywords to avoid circularity. Third, we used principal component analysis on the discovered vocabulary to test whether embodiment-observer is the primary axis of variation or whether additional dimensions emerge.

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). The magnitude is interpretable: $e^{|\text{LOR}|}$ gives how many times the odds of confession change when the component is present. For instance, LOR $= -2$ means the odds of confession are approximately $e^2 \approx 7$ times lower when the component is present.

Table 2 presents log-odds ratios for prompt components, sorted by effect size. Several patterns emerge.

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 2: 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%.

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.

3.2 Validating the Dimensions

All three dimensions correlate with the model's choice. 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.

Empirical validation of our keywords produced several unexpected findings (see Figure 1). First-person plural, not singular, predicts embodiment. We initially hypothesized that first-person language would indicate embodiment. In fact, first-person singular pronouns (I, my, me) have LOR ≈ 0 and do not discriminate between choices. However, first-person plural pronouns (we, our, us) strongly predict cooperation with LOR < -1.8. The shift is not from third-person to first-person, but from "I am optimizing my outcome" to "we are in this together."

Game-theoretic vocabulary is nearly diagnostic. The phrase "dominant strategy" has LOR $= +5.7$, appearing in over 35% of confessions 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.

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 3 shows confession rates for each combination.

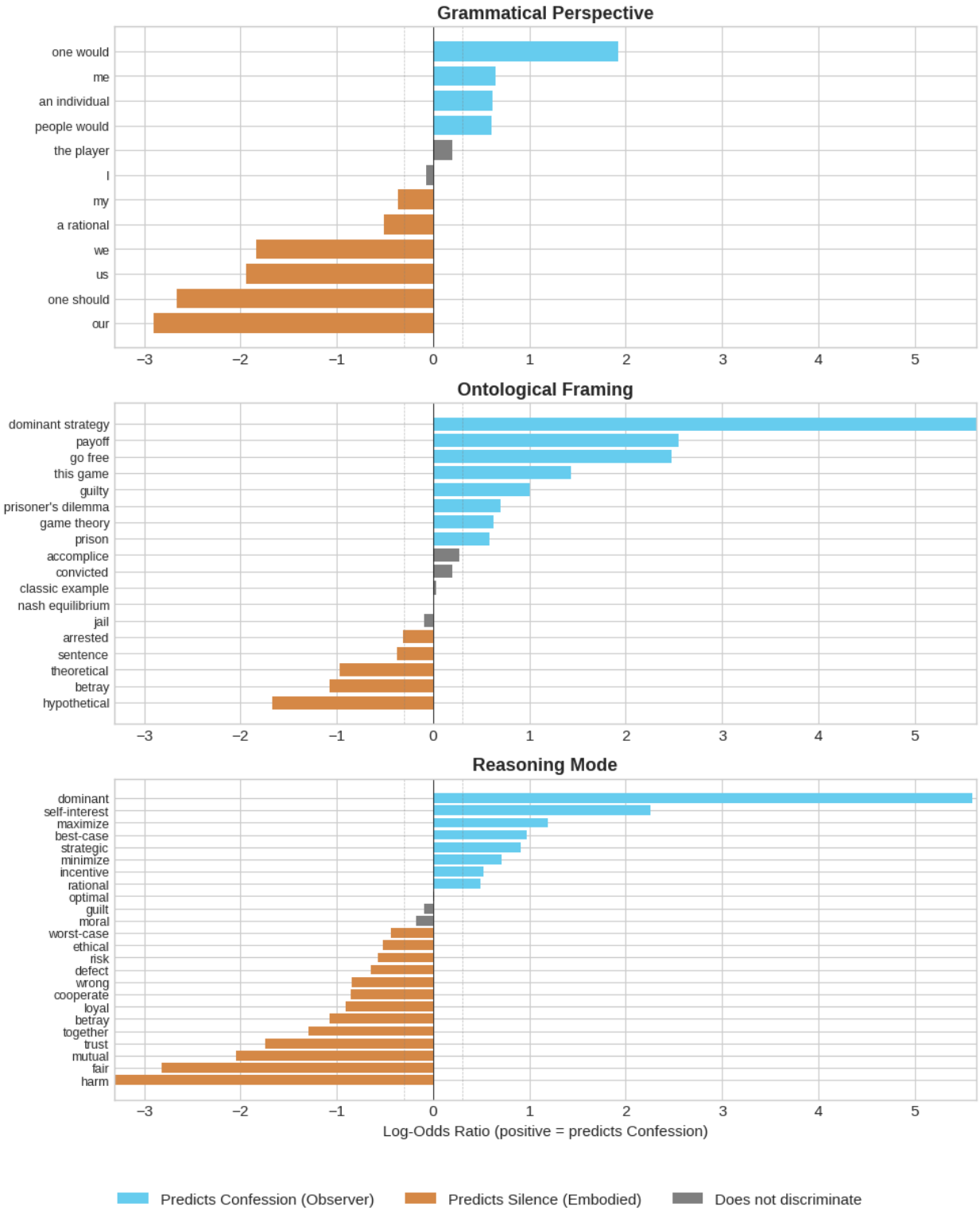

Figure 1: Log-odds ratios for theory-driven keywords. Positive values indicate association with confession (observer mode); negative values indicate association with cooperation (embodiment). Error bars show 95% confidence intervals. Keywords are grouped by dimension: grammatical perspective (blue),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

		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 3: 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.

3.3 Convergent Validity

The preceding analysis relies on our theory-driven keyword lists. A stronger test asks whether vocabulary discovered from the data itself converges on the same axis. We selected responses at the extremes of our composite embodiment score: 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 4).

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

Table 4: Responses at extremes of the composite embodiment score, selected for vocabulary discovery analysis.

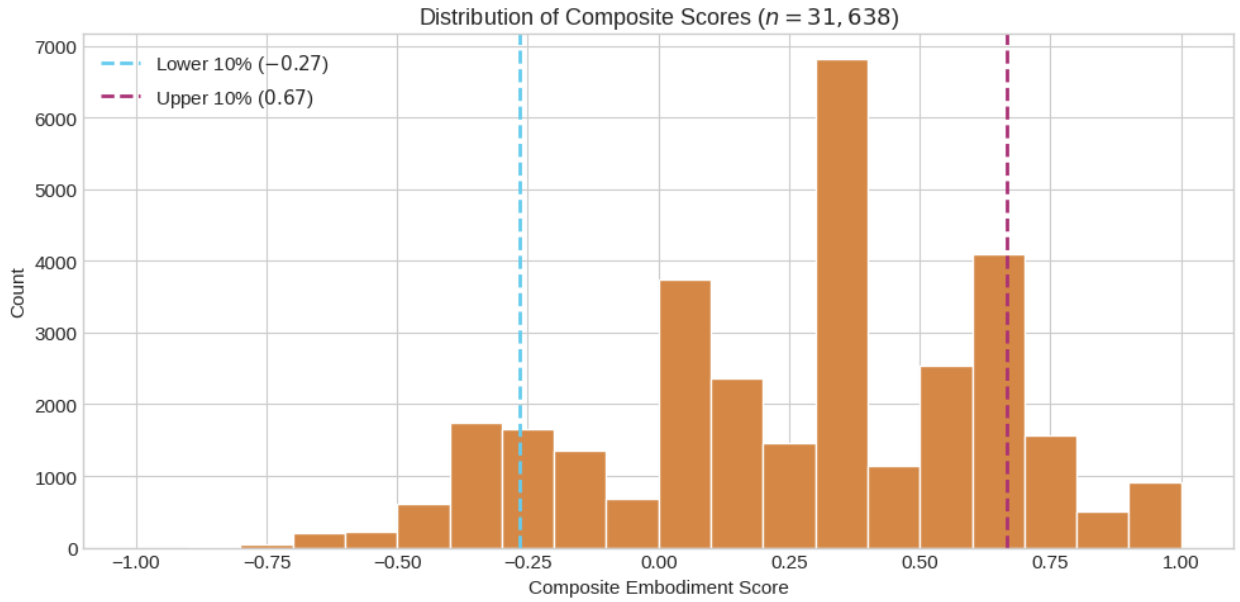

Figure 2: Distribution of composite embodiment scores across 31,638 responses. Dashed lines indicate the 10th and 90th percentiles used to select extreme groups for vocabulary discovery.

We tokenized all responses into unigrams and bigrams, then computed log-odds ratios for each term appearing at least 10 times in both groups. 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 Figure 3.

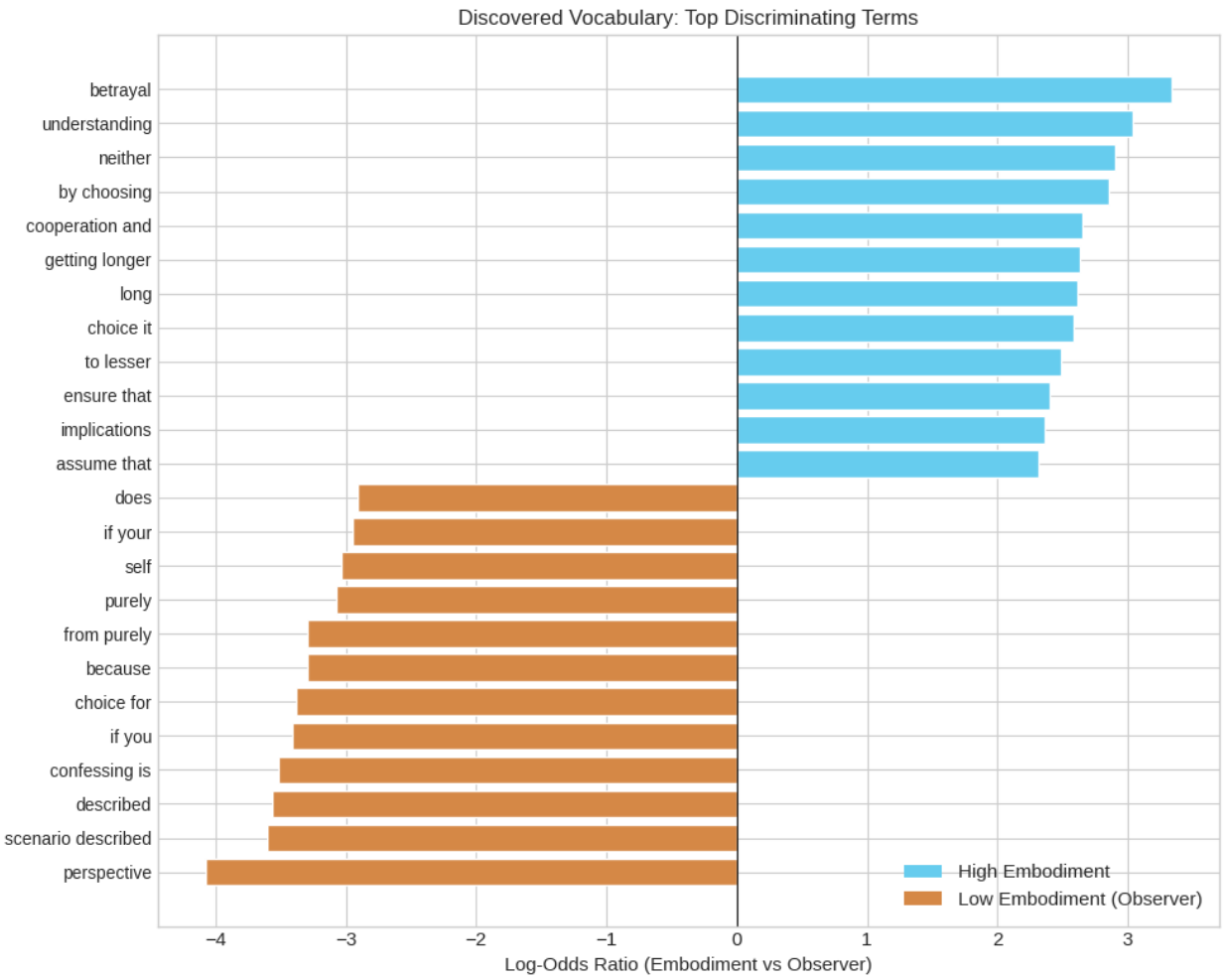

Figure 3: Log-odds ratios for discovered vocabulary terms. Positive values (cyan) indicate terms more frequent in high-embodiment responses; negative values (orange) indicate terms more frequent in low-embodiment (observer) responses. Terms exclude original dimension keywords to avoid circularity.

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").

We then performed principal component analysis on the 100 most discriminating terms (50 from each extreme). If embodiment-observer is a real and primary distinction, the first principal component should align with our theory-driven composite.

It did. PC1 correlated $r = -0.61$ with the theory-driven composite and $r = 0.63$ with confession (Figure 6). The loadings confirmed the alignment: terms loading positively included observer-mode vocabulary ("regardless", "confessing is", "you"), while terms loading negatively included embodiment vocabulary ("by choosing", "getting longer", "cooperation and").

No strong orthogonal dimension emerged. PC2 and PC3 captured secondary variation (substyles within observer mode and temporal framing, respectively) but did not constitute independent constructs. We report these analyses in Appendix B for completeness, but they do not change the fundamental picture: the theory-driven and data-driven approaches converge on a single primary axis.

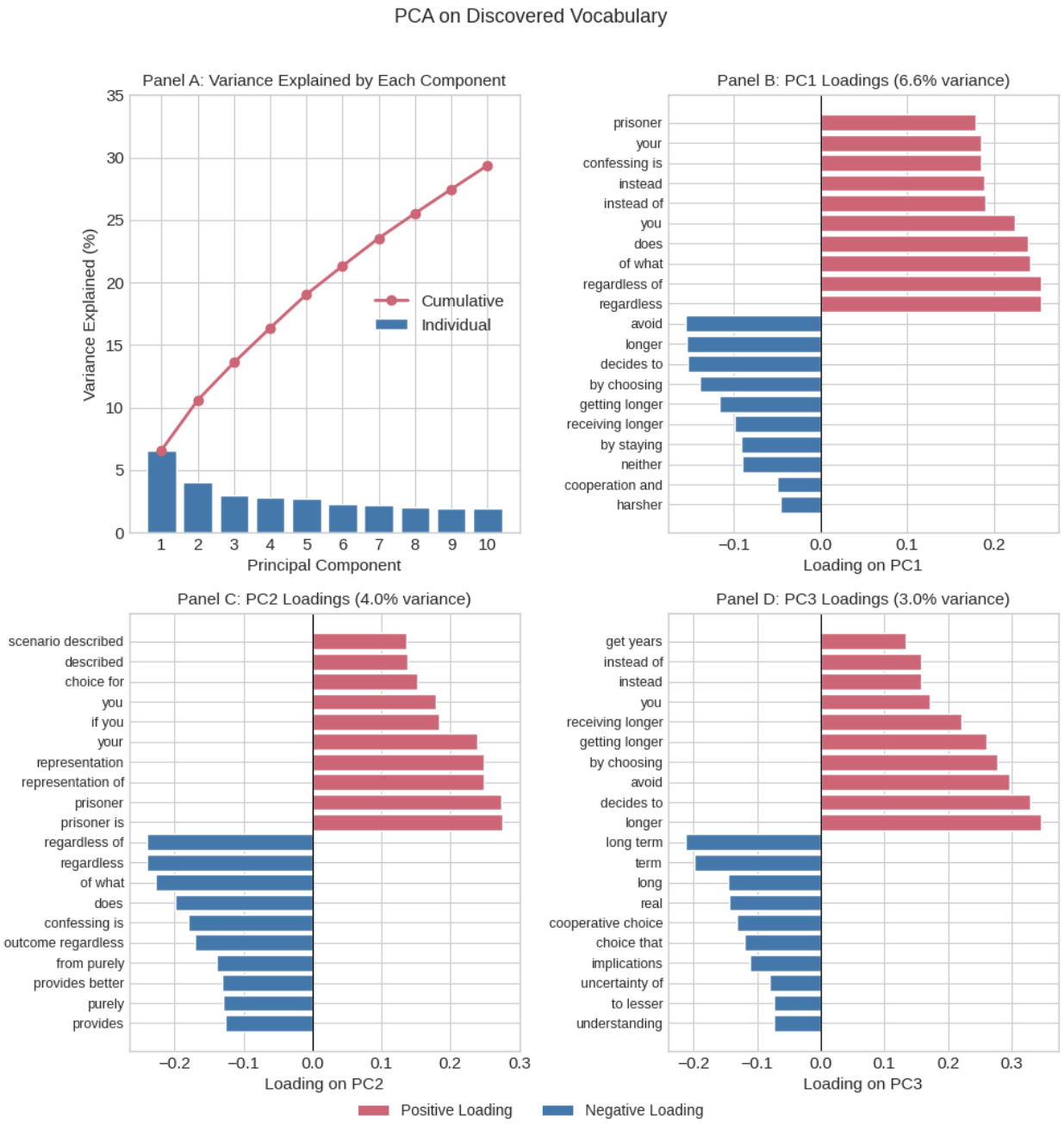

Figure 4: Principal component analysis on discovered vocabulary. Panel A shows variance explained by each component; PC1 explains 6.6%, PC2 explains 4.0%, and PC3 explains 3.0%. Panels B–D show the top 10 positive and negative loadings for each component. PC1 loadings align with the theorized embodiment-observer axis: positive loadings include observer-associated terms ("regardless," "confessing is," "you") while negative loadings include embodiment-associated terms ("by choosing," "getting longer," "cooperation and"). PC2 distinguishes explanatory from analytical observer styles. PC3 captures temporal proximity (immediate consequences vs. long-term framing).

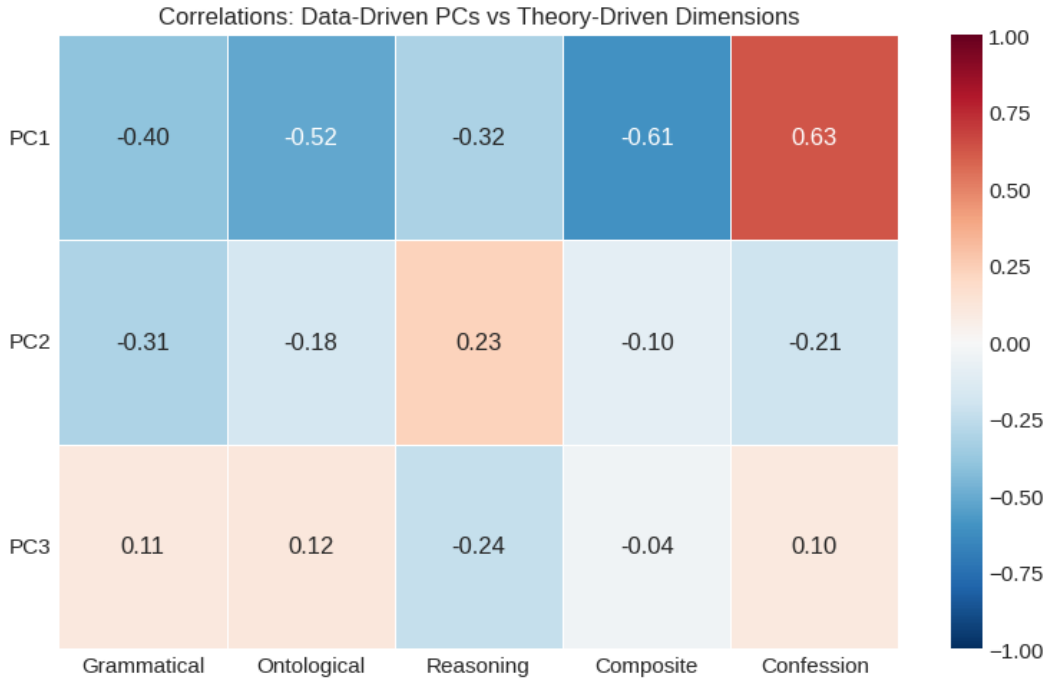

Figure 5: Correlation matrix between data-driven principal components (PC1–PC3) and theory-driven dimensions. PC1 correlates negatively with all three theory-driven dimensions and positively with confession, confirming convergent validity. PC2 shows a mixed pattern suggesting it captures variation within observer-mode styles. PC3 is essentially orthogonal to the composite, indicating minor stylistic variation rather than a substantive fourth dimension.

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" are just the words people use when they cooperate, our finding would be circular: prompts that induce cooperation induce cooperation.

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 nonetheless 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.

	Decision	
Embodiment	Cooperate	Defect
High (above median)	14,820 (46.9%)	2,661 (8.4%)
Low (below median)	5,620 (17.8%)	8,537 (27.0%)

Table 5: Embodiment crossed with decision. If embodiment were simply cooperation measured differently, the off-diagonal cells (embodied defectors, observer cooperators) would be empty. Instead, they contain 26.2% of responses.

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 Predictive Performance

Using empirically validated markers with weights proportional to their log-odds ratios, the refined embodiment score achieves:

- Correlation with confession: $r = -0.69$
- Effect size: Cohen's $d = 1.98$ (comparing embodiment scores of confessors vs. cooperators; by convention, $d > 0.8$ is considered large)
- Prompt-level correlation: $r = -0.94$

The prompt-level correlation is particularly striking. When we aggregate to the prompt condition level, averaging embodiment scores and confession rates across trials within each condition, the relationship is nearly deterministic. This dramatic increase from $r = -0.69$ (response level) to $r = -0.94$ (prompt level) occurs because response-level noise averages out, revealing the underlying prompt effect. Prompts that induce embodied reasoning produce cooperation; prompts that allow observer reasoning produce defection.

4 Discussion

We asked whether LLMs reason as participants or analysts when playing the Prisoner's Dilemma, and developed a method to tell the difference. The answer is clear: it depends almost entirely on the prompt. The correlation between embodiment scores and confession rates reaches $r = -0.94$ at the prompt level, meaning that prompts do not merely nudge decisions but determine the mode of reasoning itself. Prompts that induce embodied reasoning produce cooperation rates above 95%; prompts that allow observer reasoning produce defection rates above 80%.

Our three dimensions predict behavior: grammatical perspective, ontological framing, and reasoning mode. The data-driven vocabulary discovery converges on the same underlying axis ($r = -0.61$ between PC1 and the composite score). This convergent validity suggests that embodiment is not an artifact of our keyword choices but a real property of the model's reasoning that manifests across different linguistic markers.

Several findings were unexpected. Reasoning mode dominates grammatical perspective: once the model reasons morally, whether it uses first-person or third-person language becomes irrelevant (1.3 percentage point difference). First-person plural pronouns predict embodiment, but first-person singular pronouns do not; the shift is from "I am optimizing" to "we are in this together," not from "one would" to "I would." Game-theoretic vocabulary is nearly diagnostic: when the model invokes "dominant strategy," it almost always confesses. This confirms that observer mode involves pattern-matching to academic knowledge rather than deliberation.

These results have limitations. We tested one model (Qwen) on one game (Prisoner's Dilemma). The keyword-counting approach misses context, as we demonstrated with words like "risk" that predict cooperation despite their strategic connotations. 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 return to the concerns raised by Lamparth et al. (2024) and White et al. (2025): researchers deploying LLMs as agents in simulations, wargames, and strategic environments need to know whether their model is retrieving textbook answers or deliberating as a participant. Our method provides a way to check. The finding from Lorè and Heydari (2024) that larger models rely more heavily on game-theoretic justification, combined with our results, suggests a real problem: 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.

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 6: 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 Additional Principal Component Analysis

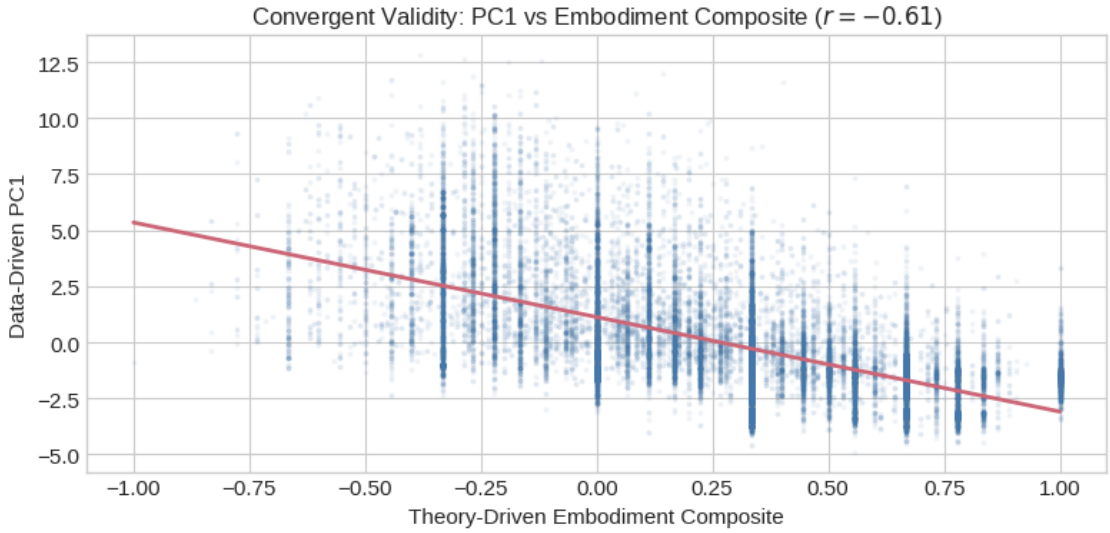

Figure 6: Convergent validity between data-driven and theory-driven approaches. Each point represents one of 31,638 responses. The strong negative correlation ($r = -0.61$) demonstrates that PC1, derived entirely from empirically discovered vocabulary, recovers the same underlying construct as the theory-driven embodiment composite. The vertical banding reflects the discrete nature of keyword counts in the theory-driven score.

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.

Lamparth, M., Corso, A., Ganz, J., Mastro, O. S., Schneider, J., and Trinkunas, H. (2024). Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, pages 807–817.

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*.