Journal of Cases on Information Technology
Volume 28 • Issue 1 • January-December 2026

Words Matter:
How Prompt Framing Shapes Strategic Behaviour in Large Language Models

Milan Kořínek
https://orcid.org/0009-0008-1208-4818
Faculty of Informatics and Management, University of Hradec Králové, Czech Republic

Kamila Štekerová
https://orcid.org/0000-0001-5847-7950
Faculty of Informatics and Management, University of Hradec Králové, Czech Republic
Received: December 2nd, 2025 | **Accepted:** December 31st, 2025

ABSTRACT

This study investigates how linguistic framing influences the strategic behaviour of large language models in repeated interactions. Four models (Mistral, Qwen3, Llama3.2, Llama3.3) were embedded as autonomous agents in a simulation of a 25-round iterated prisoner's dilemma under three prompt conditions: neutral, positively biased, and hunger-framed. Although payoff structures remained constant, linguistic variation produced substantial behavioural divergence. A one-way analysis of variance showed significant prompt effects in 13 out of 16 model pairings (adjusted $p < 0.05$). Positively biased prompts increased cooperation by 4–9 percentage points, while survival-framed prompts increased cooperation up to 80 percentage points. While Qwen3 and Llama3.3 were highly sensitive to framing, Llama3.2 showed minimal responsiveness. Several models exhibited emergent strategies such as conditional cooperation and end-game defection. These findings indicate that subtle linguistic cues can systematically modulate cooperative behaviour in large language model agents.

KEYWORDS

Large Language Model, Prompt Framing, Iterated Prisoner's Dilemma, Agent-Based Simulation

INTRODUCTION

In cognitive science, it has long been established that decision-making can be systematically altered by framing effects, even when objective payoffs and outcomes remain unchanged (Tversky & Kahneman, 1981). Differences in linguistic presentation, such as affective tone, moral valence, or contextual emphasis, can lead individuals to adopt distinct behavioural strategies despite formally equivalent choice structures. This insight has proven central to understanding human reasoning under uncertainty and provides a natural conceptual bridge to the study of artificial decision-making systems.

In a closely related manner, large language models (LLMs) have been shown to adjust their outputs in response to linguistic framing, including variations in moral language, agent-role attribution, persona cues, and prosocial narratives (Kamruzzaman et al., 2024; Patel & Pavlick, 2021; Tan & Lee, 2025). More recent work further demonstrates that subtle, non-instructional prompt modifications, such as shifts in affective or normative tone, can systematically influence model behaviour even when the underlying task structure and objective conditions remain fixed (Brucks et al., 2025; Félix-Peña

DOI: 10.4018/JCIT.398628

et al., 2025; Germani & Spitale, 2025; Herrera-Poyatos et al., 2025; Lorè & Heydari, 2023). These findings suggest that LLMs internalise prompt framing as part of the decision context rather than processing it as a purely surface-level instruction.

To investigate the behavioural implications of such sensitivity in a strategic setting, this study adopts the iterated prisoner's dilemma (IPD), a canonical framework in game theory for analysing cooperation, reciprocity, and defection under repeated interaction (Axelrod, 2012; Glynatsi et al., 2024; Nowak & Sigmund, 1993). The IPD has long served as a standard testbed for examining how dynamic strategies emerge over time, and it remains a widely used benchmark in agent-based modelling. Its use enables the systematic tracking of behavioural adaptation under repeated feedback and provides a well-understood reference point for comparative analysis.

When LLMs are embedded as decision-making agents within IPD-like environments, they differ fundamentally from classical game-theoretic or agent-based agents. Traditional agents are typically governed by explicit utility functions, optimisation objectives, or reinforcement-learning policies (Ale Ebrahim Dehkordi et al., 2023; Elsenbroich et al., 2013). In contrast, LLM-based agents respond to natural language prompts that simultaneously specify the task, describe the environment, and convey contextual and normative cues. As a result, the prompt may function not only as a task specification but also as a latent value frame that modulates strategic behaviour without altering payoffs or formal rules (Germani & Spitale, 2025; Hayes et al., 2025).

Empirical evidence increasingly supports the claim that LLM behaviour is sensitive to such linguistic context. Persona framing and explicit role assignment have been shown to influence model outputs across a range of domains systematically (Fontana et al., 2025; Phelps & Russell, 2025). At the same time, concerns have been raised regarding the robustness of LLM behaviour to seemingly minor prompt variations. Sclar et al. (2024) documented substantial performance differences arising solely from formatting changes, while benchmark efforts such as PromptRobust reveal a high degree of vulnerability to small prompt perturbations (Zhu et al., 2023). Together, these findings underscore the methodological importance of treating prompt formulation as a substantive experimental variable rather than a neutral interface.

Despite this growing body of work, the majority of prior studies focus on classification tasks, one-shot decisions, or explicitly normative instructions. The extent to which subtle, meaning-preserving linguistic variations can influence strategic behaviour across repeated interactions—where memory, expectation formation, and reciprocal adaptation are central—remains insufficiently understood. In particular, fine-grained quantitative evaluations of prompt framing effects in iterated game-theoretic environments are still relatively scarce (Lorè & Heydari, 2023).

This study addresses this gap by examining how subtle linguistic variations, introduced without modifying the underlying game structure or payoff matrix, affect cooperation in repeated dyadic interactions among LLM-based agents. Multiple language models are embedded as strategic players within simulations of the IPD and evaluated under three prompt conditions: a neutral prompt, a positively biased prompt, and a hunger-framed prompt that situates the interaction within a survival-oriented narrative. The central research question is whether linguistic framing alone can serve as a latent mechanism for shaping the strategic behaviour of LLMs over time. By isolating prompt-level variation within a controlled game-theoretic environment, the study aims to clarify the extent to which cooperation in LLM-driven agents reflects stable strategic reasoning versus sensitivity to contextual, affective, and narrative cues.

STATE OF THE ART

Large Language Model Multi-Agent Simulation

Recent work demonstrates that LLMs can function as autonomous agents capable of exhibiting coherent, context-sensitive behaviour in simulated social environments. Park et al. (2023) introduced the concept of generative agents, showing that when LLMs are endowed with memory, reflection, and