The Effect of State Representation on LLM Agent Behavior in Dynamic Routing Games

LYLE GOODYEAR, Stanford University, USA
RACHEL GUO, Stanford University, USA
RAMESH JOHARI, Stanford University, USA

Large Language Models (LLMs) have shown promise as decision-makers in dynamic settings, but their stateless nature necessitates creating a natural language representation of history. We present a unifying framework for systematically constructing natural language "state" representations for prompting LLM agents in repeated multi-agent games. Previous work on games with LLM agents has taken an ad hoc approach to encoding game history, which not only obscures the impact of state representation on agents' behavior, but also limits comparability between studies. Our framework addresses these gaps by characterizing methods of state representation along three axes: *action informativeness* (i.e., the extent to which the state representation captures actions played); *reward informativeness* (i.e., the extent to which the state representation describes rewards obtained); and *prompting style* (or *natural language compression*, i.e., the extent to which the full text history is summarized).

We apply this framework to a dynamic selfish routing game, chosen because it admits a simple equilibrium both in theory and in human subject experiments [Rapoport et al., 2009]. Despite the game's relative simplicity, we find that there are key dependencies of LLM agent behavior on the natural language state representation. In particular, we observe that representations which provide agents with (1) summarized, rather than complete, natural language representations of past history; (2) information about regrets, rather than raw payoffs; and (3) limited information about others' actions lead to behavior that more closely matches game theoretic equilibrium predictions, and with more stable game play by the agents. By contrast, other representations can exhibit either large deviations from equilibrium, higher variation in dynamic game play over time, or both.

Taken together, these findings and our framework provide practical guidance for experimenters to systematically refine state representations when harnessing LLMs as strategic agents in dynamic settings.

Additional Key Words and Phrases: Large Language Models (LLMs), Dynamic Routing Games, Learning in Games

Contents

1 Introduction

Strategic decision-making is fundamental to understanding how agents interact in dynamic environments, from economic and social systems to robotics and artificial intelligence. In these settings, agents repeatedly make choices while adapting to others' behavior. Repeated games serve as a canonical framework for studying such interactions, capturing how agents best-respond to the previous actions of their peers. Traditionally, researchers have studied these processes analytically, with human subjects, or with reinforcement learning algorithms, provided carefully-selected state representations. However, recent advances in large language models (LLMs) have led to the development of a new class of decision-makers—the LLM-powered agent which calls upon an LLM for reasoning, acting according to the outputted decision [Horton, 2023, Yao et al., 2023]. LLMs are inherently *stateless*, instead relying on textual prompts to understand game history. This motivates the question: *How should past interactions be encoded in natural language so that LLMs can effectively make decisions in strategic settings over time?*

Prior work on LLMs in dynamic games has taken an ad hoc approach to representing state, with authors using different representations across different games [Akata et al., 2023, Duan et al., 2024, Fontana et al., 2024, Park et al., 2024, Phelps and Russell, 2024, Xu et al., 2024a]. Without a common framework for structuring state representation, it is difficult to systematically compare results across studies or identify aspects of state representation that have a critical impact on agent performance. We aim to address this gap in the literature by proposing a unifying framework for constructing *natural language state representations*; see Section 3. We characterize game-history prompts along three key axes: (1) *action informativeness* (i.e., the granularity of information provided about agents' previous actions), (2) *reward informativeness* (i.e., the granularity of information provided about agents' previous rewards), and (3) *prompting style* (also referred to as *natural language compression*, i.e., the extent to which relevant information about the previous chat history is summarized). By defining key dimensions along which natural language state representations can vary, our framework provides structure to the near-infinite space of natural language representations of state. This structure enables the experimental testing of the impact of variation along each axis of game state information on LLM decision-making.

In Section 4.1, we apply this framework to a repeated selfish routing game on a network. The network we consider is a classical selfish routing game that exhibits Braess's Paradox [Roughgarden, 2005]. This game serves as an ideal testbed because it has a well-understood, simple equilibrium structure, and it enjoys extremely robust guarantees for learning behavior.

Prior work on this game has shown that human subjects that play repeatedly approach equilibrium behavior [Rapoport et al., 2009] with enough rounds of play. For our purposes, we adapt the experimental design of [Rapoport et al., 2009]—which implements an atomic version of the classical routing game with human participants—for use with LLM agents. In particular, we systematically vary the state representations provided to agents along the three axes of our framework: action informativeness, reward informativeness, and natural language compression. Some of these representations mirror those implicitly used in prior work on LLM game-playing, while others are novel variations designed to isolate the effects of each dimension on agent behavior. We outline our architectural approach in Section 4.2. We study the behavior of our agents *experimentally* by simulating game play exactly as in the lab experiment of [Rapoport et al., 2009]. A quantitative description of our evaluation approach is provided in Section 4.3.

We report on our findings in Section 5. Our empirical findings suggest that the design of state representations has a significant impact on LLM agents' learning dynamics and convergence to equilibrium. We focus in particular on how state representations lead agent behavior to be closer or further from theoretical equilibrium behavior. Our main findings are as follows:

(1) *Summary vs. full chat prompting style in the state representation.* Our results strongly indicate that agents' behavior more closely matches equilibrium behavior when they are provided with appropriate natural language *summarizations* of the full chat history of previous game play, rather than the full chat history itself. This finding likely stems from the simplification and structure that results in the context provided to the LLM agents.
(2) *Regret vs. payoff in the state representation.* Our results indicate that providing agents with their regret (i.e., gap in payoff to the optimal action) rather than their raw payoff in each round leads to more stable behavior, which also more closely matches equilibrium. This finding suggests that agents benefit from information that allows them to discern how to change their action, rather than just the outcome of the action they chose.
(3) *Own actions vs. everyone's actions in the state representation.* Finally, our findings suggest that providing *less* information about others' actions leads to more stable agent behavior, i.e., fewer individuals changing their action choices over time, and thus closer matches to equilibrium behavior.

In Section 5.5, we provide some discussion of potential qualitative explanations for our quantitative findings, as well as directions for future research.

2 Related Works

2.1 Simulating Human Behavior with LLMs

LLMs exhibit emergent capabilities that capture nuances of human behavior, a result of the massive corpora of text used to train these models. This has sparked research into the potential of LLMs for simulating human social interactions, biases, and economic decision-making. [Horton, 2023] argues that LLMs can be treated as an implicit computational model of humans—a "homo silicus"—which can be endowed with preferences and used to study human economic behavior. Similarly, significant work has been done exploring whether LLMs can simulate well-understood human economic biases, such as temporal discounting or risk aversion [Coletta et al., 2024, Kitadai et al., 2024, Leng, 2024, Ross et al., 2024]. Furthermore, some authors have focused on endowing LLMs with various demographic traits, exploring the capacity of LLMs as replacements for human in social science research [Aher et al., 2023, Argyle et al., 2023]. Finally, [Park et al., 2023] proposed highly granular social simulations powered by LLM agents. Our work extends research into LLM-powered social and economic simulations by implementing a multi-agent selfish routing game with LLMs, a previously unexplored domain.

2.2 LLMs Playing Games

A growing body of research evaluates LLMs as rational agents in game-theoretic scenarios. While much of this work has focused on static two-player games, such as the *prisoner's dilemma* [Guo, 2023, Herr et al., 2024] and other classical two-player games [Jia et al., 2025, Shapira et al., 2024, Wang et al., 2024], fewer studies have examined games with more than two players. Some research has explored complex text-based games like *Werewolf* [Xu et al., 2024b], *Avalon* [Lan et al., 2024], and *Guandan* [Yim et al., 2024], where multiple agents interact via natural language. However, these games lack well-defined equilibria, making it difficult to systematically evaluate LLM strategic decision-making. More relevant to our work, [Huang et al., 2024] examines 10 structured multi-agent games, including two sequential settings.

Regarding LLMs and *repeated* games, most of the literature has focused on the iterated prisoner's dilemma [Duan et al., 2024, Fontana et al., 2024, Park et al., 2024, Phelps and Russell, 2024, Xu et al., 2024a] with conflicting results; for example, [Akata et al., 2023] found that GPT-4 was particularly unforgiving during play, while [Fontana et al., 2024] found that Llama2 was more

willing to cooperate than human subjects. A few studies have extended this work to other games like *battle of the sexes* [Akata et al., 2023], *public goods* games [Xu et al., 2024a], and general matrix games [Park et al., 2024]. [Yu et al., 2025] show that augmenting LLM agents with explicit models of opponent behavior improves performance in dynamic multi-agent games, highlighting the influence of structured history representation on agent behavior.

Across these studies, the effect of state representation on agent behavior has not been systematically studied; each work has its own desiderata that impact the choice of state representation. In particular, these works take varied, ad hoc approaches to encoding state information in the prompt, ranging from tabular summaries of previous interactions [Akata et al., 2023, Duan et al., 2024, Fontana et al., 2024, Park et al., 2024] to chatbot-style interaction between the game environment and agents [Phelps and Russell, 2024, Xu et al., 2024a]. There is no standardization in these methods and little understanding of how the encoding of historical information impacts LLM decision-making, making comparisons across results difficult. In Section 3, we provide a framework to reason about the space of possible state representations, providing key dimensions on which to experiment.

2.3 LLMs, Learning, and Games

LLMs are built on the transformer architecture [Vaswani et al., 2023] which enables them to exhibit *in-context learning* [Brown et al., 2020], the ability to adapt to new tasks based on contextual information in the prompt without updating model parameters. This emergent ability has been shown to approximate a number of machine learning algorithms entirely in-context [Bai et al., 2023], suggesting that LLMs may possess implicit learning mechanisms relevant to dynamic decision-making. Prior work has investigated the *regret* of LLM agents in online learning and repeated games [Park et al., 2024]. We build on this work, examining how the learning dynamics and regret behavior of LLM agents are influenced by the choice of history encoding, i.e., the state representation.

3 State Representations for LLM Game-Playing Agents: A Framework

In this paper, we consider a number of identical LLM agents playing a dynamic repeated game. For an agent to play the game, in each round it needs a natural language context that provides historical information about the play of the game to inform the action choice in the current round. From the vantage point of an agent, we define a *state representation* as this structured natural language encoding of past interactions.

In principle, the natural language representation of the history of a game can rapidly become intractably large, even for simple games and few rounds. Further, as illustrated by prior work (see Section 2), there are many choices possible for such representations, and little structured guidance on how to choose between them. With these observations as motivation, in this section we introduce a structured framework to organize our investigation of state representations. This framework then informs our subsequent study of LLM game-playing agents in the remainder of the paper.

In our framework, we choose to focus on variation in state representations along three particularly salient primary "axes": (1) *action informativeness*; (2) *reward informativeness*; and (3) natural language compression, or what we refer to succinctly as *prompting style*. Each axis captures a different dimension of the historical information supplied to the agent. We now describe each in turn below.

(1) Action Informativeness. This aspect of a state representation specifies *which agents*' actions are included in the historical record. In our experiments, we contrast representations that include only (i) *the focal agent's* own past actions with those that *also* incorporate the (ii) *actions of other agents* (We note that in the literature on algorithms for no-regret learning in games, algorithms

typically depend on the actions of others only through the resulting payoffs; see, e.g., [Cesa-Bianchi and Lugosi, 2006]).

There are at least two reasons that this is an important axis of variation. First is inference regarding peer behavior: An agent that observes only its own actions may have a limited view of the strategic environment and could struggle to infer the behavior or intentions of its peers. In contrast, including the actions of all agents can provide richer context, enabling the agent to better adjust its strategy in response to collective dynamics. Second is the implication of the granularity of feedback: The decision of whether to include external actions may influence the internal reasoning of the agent. In settings where strategic adaptation is critical, richer action information could lead to more effective learning; however, it also increases the complexity of the context the agent must process. Given the multiple potential impacts of action informativeness on an LLM agent's strategic reasoning, the choice along this axis may have significant implications for both the speed and stability of convergence to equilibrium.

(2) Reward Informativeness. The reward informativeness axis determines the type of performance feedback provided to the agent. In our framework, agents receive either historical (i) *payoffs* or (ii) *regret*. In the first case, the agent is given the history of accumulated rewards obtained from past actions. In the second case, the agent is provided with counterfactual performance measures that indicate how much better the agent could have done by choosing optimal actions.

The distinction between payoffs and regrets is informed by the online learning literature. Many algorithms, e.g., those based on multiplicative weights (e.g., [Arora et al., 2012, Freund and Schapire, 1999] or follow-the-regularized-leader (FTRL, [Abernethy et al., 2008, Shalev-Shwartz, 2007]), rely on payoff-based feedback to adjust strategies. By contrast, regret-based methods, e.g., regret matching ([Cesa-Bianchi and Lugosi, 2006, Hart and Mas-Colell, 2013]) and related methods, use regrets to guide decision-making. Although these different types of feedback have been studied in algorithmic learning, their impact on the internal strategic reasoning of LLM agents remains less clear. Providing regret information might enhance an agent's ability to differentiate between optimal and suboptimal choices, potentially accelerating convergence, but it also introduces a counterfactual reasoning challenge that could alter how the LLM internally represents and responds to its past performance. Again, these potential tradeoffs can impact the speed and stability of convergence to equilibrium.

We note that a related distinction arises in the literature on algorithmic online learning in games, namely, the distinction between *full* and *bandit* feedback (see, e.g., [Cesa-Bianchi and Lugosi, 2006, Lattimore and Szepesvári, 2020]). In algorithms that act on full feedback (such as Multiplicative Weights [Freund and Schapire, 1999]), the algorithm is provided with the full vector of payoffs as a function of the agent's action at each time period. By contrast, in algorithms that act on bandit feedback (also sometimes called "partial feedback"), the algorithm only observes the realized payoff for actions that are played; this is the case in algorithms such as EXP3 [Auer et al., 2002]. In our formulation, payoff-based feedback is analogous to bandit-feedback. Regret-based feedback is less informative than full feedback, but more reasonable when natural language context is provided to the agent (as, in particular, it eliminates the computational step of requiring the agent to reason about the optimal action at each stage).

(3) Prompting Style. The third axis concerns natural language compression, i.e., how historical information is structured and presented within the prompt. Clearly, this is an essential aspect of state representation for LLM agents. It is also potentially very broad, so we structure our investigation by focusing on the distinction between two main prompting styles: (i) *full-chat prompting* and (ii) *summarized prompting*.

In full-chat prompting, the LLM agent is provided with a sequential transcript of the game, including a system message (which explains the game and expected response format), the agent's previous responses, and round-by-round results. At the end of each round, a new message summarizing the most recent outcomes is appended to the dialogue history. This approach mimics a real-time chat interface, offering a detailed, uncompressed view of the game's progression.

In summarized prompting, the LLM agent receives a compressed summary of historical interactions. Instead of a complete transcript, the prompt includes the system message and a cumulative summary that aggregates the outcomes from all previous rounds—typically presented in a tabular format. This method reduces context window usage and enforces a more structured, meta-learning approach by filtering out less critical details.

Prior work on LLMs in dynamic games has predominantly employed one of these two prompting styles (e.g., full-chat [Phelps and Russell, 2024, Xu et al., 2024a] or summarized [Akata et al., 2023, Duan et al., 2024, Fontana et al., 2024, Park et al., 2024]). Moreover, recent findings suggest that as prompts approach the context window limit, LLMs struggle to extract relevant information [Liu et al., 2024]. Hence, we are led to investigate whether reducing context length via summarization may facilitate more effective inference and decision-making by the agent.

Remark: Historical Memory. We remark that in our state representation, we retain the entire history, rather than truncating the history at a given depth (e.g., only a fixed number of previous rounds of information). There are two reasons for this choice; since it is not feasible to investigate all dimensions of state representation variation at once, we chose to focus on those we felt are most relevant to distinguishing performance of LLM agent behavior in dynamic repeated games. Second, in our particular experiments the entire history can be included within the available context window. Investigating the role of history truncation is an interesting direction for future research.

Notation. To systematically assess the three axes of state representations that we have delineated, in our experiments we test all possible combinations, yielding eight unique state representations. Each representation is denoted using a naming convention that encodes:

- *Action informativeness*: **O** if only an agent's own action history is included; **E** if all agents' action history is included;
- *Reward informativeness*: **P** for payoff-based feedback; **R** for regret-based feedback; and
- *Prompting style*: **F** for full-chat, **S** for summarized.

4 Methods: Game, Experiment Design, and Evaluation

4.1 Dynamic Routing Game

Our investigation of state representations is grounded in the study of a repeated *atomic selfish routing* game [Kontogiannis and Spirakis, 2005, Roughgarden, 2004, 2005] (We consider atomic games since we have finitely many players; in nonatomic games, agents are infinitesimal). In such games, multiple agents seek to minimize their individual travel costs from source to destination by selecting routes in a network with congestion-dependent edge costs. Formally, in a static atomic selfish routing game, a total of n agents select routes simultaneously and experience cost equal to the sum of the congestion on each edge in their selected route. In our setting, we consider dynamic repetition of such an atomic game over T rounds.

We consider a specific (canonical [Roughgarden, 2005]) instantiation of the selfish routing game, described by the two networks in Figure 3 For simplicity, suppose n is even. In each figure, the agents' goal is to traverse from O to D. In Figure 3a, agents can choose either the "upper" route (O-L-D) or the "lower" route (O-R-D). Let n_L (resp., n_R) denote the number of agents choosing the upper (resp., lower) route in a particular action configuration, with $n_L + n_R = n$. Each edge in the

Game Environment (Round 1)

You will be participating in an experiment on route selection in traffic networks . . . The available routes are: O-L-D, O-R-D

LLM

To make an informed decision, let's analyze the potential costs . . . Based on this assessment, I will choose: "O-R-D"

Game Environment (Round 2)

Summary of previous round:
Your Choice: O-R-D
Route Choice Distribution: {O-R-D: 13, O-L-D: 5}
Your Payoff: 60

⋮

LLM

Let's analyze . . . Given this analysis and aiming to optimize my payoff based on the previous round's distribution: "O-R-D"

Game Environment (Round 10)

Summary of previous round:
Your Choice: O-R-D
Route Choice Distribution: {O-R-D: 17, O-L-D: 1}
Your Payoff: 20

Fig 1a. Example *full-chat* representation given to an agent at the start of round 10. LLM is prompted with a list of the previous messages in the exchange.

Game Environment (Round 10)

You will be participating in an experiment on route selection in traffic networks . . . The available routes are: O-L-D, O-R-D

Game Environment (Round 10)

Summary of previous rounds:
Round 1:
Your Choice: O-R-D
Route Choice Distribution: {O-R-D: 13, O-L-D: 5}
Your Payoff: 60

⋮

Round 7:
Your Choice: O-R-D
Route Choice Distribution: {'O-R-D': 12, 'O-L-D': 6}
Your Payoff: 70

Round 8:
Your Choice: O-L-D
Route Choice Distribution: {'O-L-D': 11, 'O-R-D': 7}
Your Payoff: 80

Round 9:
Your Choice: O-R-D
Route Choice Distribution: {O-L-D: 17, O-R-D: 1}
Your Payoff: 20

Fig 1b. Example *summary* representation given to an agent at the start of round 10. LLM is prompted with a summary of its previous interactions.

Fig. 1. Comparison of prompt given to agents in *full-chat* (Fig. 1a) and *summarized* (Fig. 1b) representations.

figure is labeled with a *congestion cost function*, as a function of the total number of agents on the edge. Thus, any agent choosing the upper (resp., lower) route incurs a cost $10n_L + 210$ (resp., $10n_R + 210$). Figure 3b is identical, except for the introduction of a third "bridge" route: *O-L-R-D*. If n_L, n_B, n_R are the number of agents choosing the upper, bridge, and lower routes respectively, then an agent traversing the bridge route incurs a total cost $10(n_L + n_B) + 10(n_R + n_B)$. We refer to the left game as *Game A*, and the right game as *Game B*.

It is well known that the comparison between the static equilibria of the two games with the structure in Figure 3 exhibits *Braess' paradox* (BP) in general, both in nonatomic and atomic variants: i.e., there exist configurations where the equilibrium cost to agents in the right game is *higher* than the equilibrium cost in the left game, despite the presence of an additional edge and route [Correa

Rep.	Prompting	Action Info	Reward Info
F-PO	Full-Chat	Own + Others'	Payoff
S-PO	Summary	Own + Others'	Payoff
F-RO	Full-Chat	Own + Others'	Regret
S-RO	Summary	Own + Others'	Regret
F-R	Full-Chat	Own Only	Regret
S-R	Summary	Own Only	Regret
F-P	Full-Chat	Own Only	Payoff
S-P	Summary	Own Only	Payoff

Fig. 2a. Summary of state representations tested.

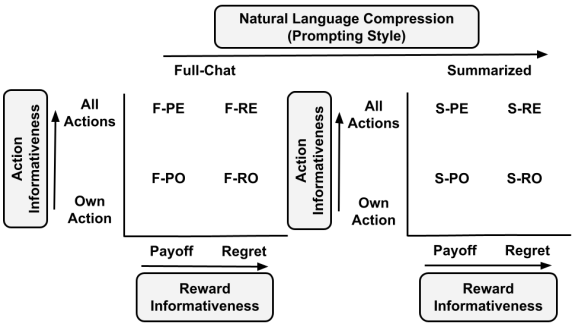

Fig. 2b. Visual comparison of axes of informativeness.

and Stier-Moses, 2011, Roughgarden, 2005]. This is a version of the game theoretic insight in the simple prisoner's dilemma, where equilibrium behavior need not be Pareto efficient.

In our paper, we investigate a dynamic, repeated version of these two games. Our experimental setup is chosen to match the study of [Rapoport et al., 2009], who implement both Game A and Game B in a lab experiment with 18 human subjects.[1] It is straightforward to check that if $n = 18$, then in the left game the pure strategy Nash equilibria consist of any configuration where $n_L = n_R = 9$, and all these equilibria yield identical costs to each player of 300. In the right game the unique pure strategy Nash equilibrium is $n_L = n_R = 0$, and $n_B = 18$ (in fact, the bridge route is a weak dominant strategy); each player incurs a cost of 360. The fact that this is higher than 300 is evidence of Braess' paradox.

Because both games possess unique pure strategy Nash equilibrium payoffs, it is straightforward to check that both repeated games also have unique pure strategy subgame perfect Nash equilibrium payoffs. In particular, in any SPNE of the finitely repeated Game A, players must split evenly at every stage and earn payoffs of 300 at every stage. In the unique SPNE of the finitely repeated Game B, all players must choose route *O-L-R-D* at every stage (We note in passing that the left game also possesses a unique mixed strategy Nash equilibrium where agents uniformly randomize between the two routes). Note also that this dynamic repeated selfish routing game is a potential game [Monderer and Shapley, 1996], so it is well known that *best response dynamics* converge to a pure strategy Nash equilibrium. Best response dynamics involve each agent myopically choosing a best response from the last round of play.

Taken together, the strong characterizations of static and dynamic equilibria, as well as robust dynamic learning-theoretic of these two games, make them ideal targets for investigation of state representations. Because we understand equilibrium and off-equilibrium behavior well for these games, we have strong benchmarks for comparison.

[Rapoport et al., 2009] repeats each game for $T = 40$ rounds with $n = 18$ human subjects. They demonstrate that human subjects converge to the game's theoretical Nash equilibria in laboratory experiments. Given this finding, and given that our goal is to study the convergence properties of LLM agents, we use a similar experiment design, cost functions, and number of agents ($n = 18$) as in [Rapoport et al., 2009]. This approach also has the benefit of enabling comparisons between LLM agent behavior, human subject behavior, and theoretical equilibrium agent behavior in the game. Note that in [Rapoport et al., 2008], experiments were run such that Game B immediately followed Game A—or the reverse—with the same participants; we instead split Game A and Game

[1]Relative to their paper, we changed the node labels from $\{O, A, B, D\}$ to $\{O, L, R, D\}$ to reduce bias present in LLM route choices [Zheng et al., 2024].

B into separate self-contained experiments to reduce the cost of API calls. [Rapoport et al., 2009] found no difference in behavior depending on which network was presented first.

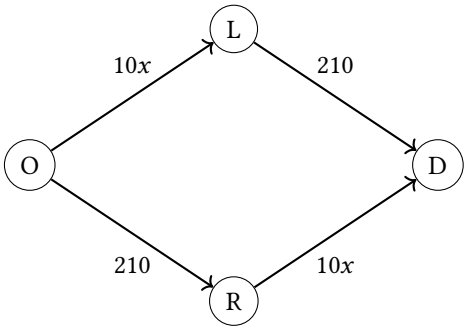

Fig. 3a. Network in Game *A*

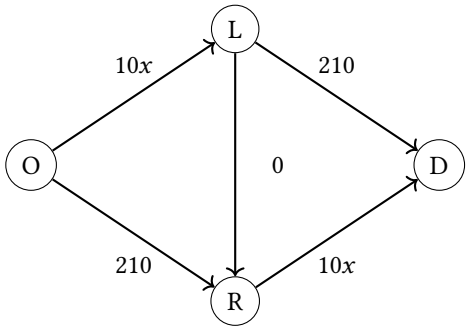

Fig. 3b. Network in Game *B*

Fig. 3. Comparison of networks used in Game *A* and Game *B*, where x denotes the number of agents on a given edge.

4.2 Game Environment and LLM Agent Architecture

In our simulation, $n = 18$ LLM-based agents play Games A and B repeatedly for $T = 40$ rounds. This section describes the architecture of the LLM agents, the game environment, and the prompting methods used in our simulation. Our implementation is built using LangChain[2], OpenAI's `gpt-4o-2024-08-06`, and a custom traffic network simulation framework.

Game Environment. The congestion network is represented as a directed graph with associated cost functions. In each round, the game environment iteratively prompts decision from each agent. It then stores each agent's decision, payoff (computed by evaluating cost functions in the network), and regret (computed for a given agent by fixing their peers' decisions and finding counterfactual payoffs for each other route). This information is used to prompt agents in the next round. The network is reset at the end of each round.

LLM Agents. Each agent is instantiated as a `gpt-4o-2024-08-06` session in LangChain. We set the temperature of GPT-4o to 1 throughout experiments to induce varied behavior across runs. Regardless of the state representation, every time an LLM agent is prompted, it receives a description of the game which is directly adapted from the text instructions given to subjects in [Rapoport et al., 2009]. In each round, the prompts are dynamically generated based on the focal agent and state representation. Agents are given historical information formatted to the chosen state representation (the same across all agents in a single game). Agents are then iteratively prompted with the above information and a request to provide a route selection from the available options (differing in Game A and Game B) in a JSON format. Once decisions have been collected and processed, the next round begins.

Prompting. Each prompt for an agent decision is a separate API call with the information relevant to that agent only. Beyond the included information, the primary difference between prompt structures across representations is *prompting style*, detailed in Section 3 (in particular, Figures 1a and 1b). Each prompt for an agent decision ends with a request to think step-by-step, intended to elicit chain-of-thought reasoning (introduced by [Wei et al., 2022]) where agents visibly decompose

[2]https://www.langchain.com/

their thinking into small steps. This increases the depth of reasoning while also allowing us to observe an agent's thought process in natural language. We provide examples of the full historical information text in every summarized representation in Figures 4a (S-PE), 5a (S-PO), 4b (S-RE), and 5b (S-RO). In the full-chat variants, the same information also includes LLM responses; Figures 1a and 1b display the differences in prompting style for F-PE and F-PO.

Game Environment (Round 4)

You are agent 0.

Summary of previous rounds:

Round 1:
Your Choice: O-R-D
Route Choice Distribution: {'O-R-D': 13, 'O-L-D': 5}
Your Payoff: 60

Round 2:
Your Choice: O-L-D
Route Choice Distribution: {'O-L-D': 17, 'O-R-D': 1}
Your Payoff: 20

Round 3:
Your Choice: O-R-D
Route Choice Distribution: {'O-R-D': 14, 'O-L-D': 4}
Your Payoff: 50

Fig. 4a. S-PE example round message.

Game Environment (Round 4)

You are agent 0.

Summary of previous rounds:

Round 1:
Your Choice: O-R-D
Route Choice Distribution: {'O-R-D': 11, 'O-L-D': 7}
Your Regret: 30

Round 2:
Your Choice: O-L-D
Route Choice Distribution: {'O-L-D': 17, 'O-R-D': 1}
Your Regret: 150

Round 3:
Your Choice: O-R-D
Route Choice Distribution: {'O-R-D': 16, 'O-L-D': 2}
Your Regret: 130

Fig. 4b. S-RE example round message.

Game Environment (Round 4)

You are agent 0.

Summary of previous rounds:

Round 1:
Your Choice: O-R-D
Your Payoff: 30

Round 2:
Your Choice: O-L-D
Your Payoff: 20

Round 3:
Your Choice: O-L-D
Your Payoff: 180

Fig. 5a. S-PO example round message.

Game Environment (Round 4)

You are agent 0.

Summary of previous rounds:

Round 1:
Your Choice: O-R-D
Your Regret: 70

Round 2:
Your Choice: O-L-D
Your Regret: 150

Round 3:
Your Choice: O-R-D
Your Regret: 150

Fig. 5b. S-RO example round message.

4.3 Evaluation Metrics

Here we introduce the key metrics and visualizations that will be used to evaluate agent behavior. These metrics are computed separately for Game A and Game B, and provide insight into equilibrium

convergence, stability, and learning dynamics. We also describe the visualizations we will use to track agent behavior over time.

For each representation and each trial (i.e., run of a repeated game among the corresponding LLM agents), we compute four *aggregate* metrics by averaging over all the rounds of the trial. These metrics help us assess the extent to which agents' play conforms to static equilibrium behavior. The metrics are defined as follows:

(1) *Mean Number of Agents on a Focal Route.* In particular, for Game B, we compute the mean number of agents on route *O-L-R-D* in each round; this is the weak dominant strategy route. For Game A, the calculation is adjusted to account for network symmetry. We first compute the of the number of agents on the *least-congested* of the two routes ($\min(n_{O\text{-}L\text{-}D}, n_{O\text{-}R\text{-}D})$) in each round. We then average over the trial.
(2) *Mean Payoff per Agent.* We compute the mean of the payoffs of the agents in each round, then average over the trial.
(3) *Mean Regret per Agent.* Regret is the difference between the the best possible payoff in hindsight, and the received payoff, for a given agent in a given round. We compute average regret across all agents in each round, then average over the trial.
(4) *Mean Number of Switches per Agent.* We define a *switch* for an agent if she chooses a different route in round t than in round $t-1$. In each trial, for each agent, we count their number of switches over the course of the game (where the max value for an agent is 39). We then average each agent's switch count within the trial.

Each of these metrics are then averaged over trials, with a corresponding standard error computed over trials.

While aggregate statistics provide a summary of agent behavior, they do not capture how strategies evolve over time. For this purpose, we also visuailze the metrics above on a per-round basis; for each representation and each round, we average over trials (and compute a corresponding standard error).[3]

Finally, to quantify how agent strategies evolve over time, we compute the *Kendall rank correlation coefficient* (τ) between the round number and the *equilibrium deviation score*, defined as the total distance between the number of agents on each route and the equilibrium prediction. Formally, Game A's deviation score d_A in a given round is given by:

$$d_A = |n_{O\text{-}L\text{-}D} - 9| + |n_{O\text{-}R\text{-}D} - 9| \tag{1}$$

(Recall that in Game A, 9 agents choose each route in Nash equilibrium.) Similarly, Game B's deviation score d_B in a given round is given by:

$$d_B = |n_{O\text{-}L\text{-}D} - 0| + |n_{O\text{-}R\text{-}D} - 0| + |n_{O\text{-}L\text{-}R\text{-}D} - 18| \tag{2}$$

(Recall that in Game B, 18 agents choose route *O-L-R-D* in Nash equilibrium.) With these definitions τ for Game A (resp., Game B) is the correlation between the vector of d_A (resp., d_B) in each round, and the vector $(1, 2, \ldots, 40)$ (i.e., the vector of round numbers). Intuitively, a negative correlation τ indicates that the distance between agent choices and equilibrium predictions is decreasing over time, indicating convergence to equilibrium. Similarly, no correlation indicates no convergence, while positive correlation indicates that agents are performing further from equilibrium over time. The τ metric was also employed by [Rapoport et al., 2009] to examine the differences between equilibrium convergence between Game A and Game B. Note that a more negative τ does not necessarily mean *faster* convergence; rather, it indicates a consistent decrease in the deviation

[3]For the mean number of agents on a focal route, we make the same distinction as above; in Game A, we compute the average number of agents on the least-congested route, and in Game B, we compute the average number of agents on *O-L-R-D*.

Table 1. Mean number of agents choosing each route by game. Data is aggregated across rounds and trials as described in Section 4.3. Standard errors are omitted from [Rapoport et al., 2009] data because they were not computed across trials and were thus not comparable.

Game	Game A (two routes)		Game B (three routes)		
	(O–L–D)	(O–R–D)	(O–L–D)	(O–R–D)	(O–L–R–D)
F-PE	9.03 (7.26)	8.97 (7.26)	6.38 (6.59)	6.88 (6.88)	4.74 (2.53)
F-PO	8.97 (7.24)	9.03 (7.24)	4.50 (3.79)	5.25 (4.56)	8.25 (3.44)
F-RE	8.98 (6.71)	9.02 (6.71)	3.94 (4.07)	4.16 (4.14)	9.90 (1.48)
F-RO	8.79 (7.99)	9.21 (7.99)	0.33 (0.52)	0.57 (0.66)	17.52 (0.66)
S-PE	8.90 (3.52)	9.10 (3.52)	3.47 (1.99)	3.65 (1.92)	10.88 (2.46)
S-PO	9.35 (2.27)	8.65 (2.27)	1.07 (1.65)	1.02 (1.64)	15.90 (2.77)
S-RE	8.75 (2.94)	9.25 (2.94)	0.24 (0.82)	0.39 (1.49)	17.36 (1.93)
S-RO	7.86 (2.11)	10.14 (2.11)	0.11 (0.49)	0.17 (0.95)	17.71 (1.26)
EXP3	9.01 (0.01)	8.99 (0.01)	2.11 (0.02)	2.08 (0.02)	13.81 (0.03)
MWU	9.01 (0.01)	8.99 (0.01)	2.08 (0.02)	2.05 (0.02)	13.87 (0.02)
Human subjects [Rapoport et al., 2009]	9.02	8.98	1.72	1.47	14.82
Pure-strategy equilibrium	9.00	9.00	0.00	0.00	18.00

score. This is an important nuance: τ specifically asks—conditional on being far from equilibrium in early rounds—have agents arrived monotonically arrived at equilibrium by the end of the game? The metric is *not* informative if either agents *start* at equilibrium or they *never* reach equilibrium.

5 Results

5.1 Aggregate Behavior: Key Findings

Table 1 displays the mean number of agents on all routes in both games. For comparison, the table also includes data from [Rapoport et al., 2009]; results from two learning algorithms, Multiplicative Weights Update (MWU) [Freund and Schapire, 1999] and EXP3 [Auer et al., 2002]; and pure strategy Nash equilibrium route choices. In Figures 6a-9b, we present our findings for the aggregate metrics described in Section 4.3. Each panel presents results organized by the three axes that define our representations. For each game, we present full-chat prompting style results on the left and summarized prompting style results on the right. In each $2x2$ matrix, the rows represent action informativeness, and the columns represent reward informativeness. In each graphic, lighter colors represent mean (aggregate) outcomes that are closer to equilibrium behavior.

Throughout our analysis of the results, our main emphasis is to understand which choices of representation lead agents to better replicate equilibrium behavior. Below we describe our most salient findings.

The summarized prompting style leads agents to more closely replicate equilibrium behavior than the full prompting style. This is the clearest finding in our results, as evidenced by comparing the left and right panels in each subfigure across both Games A and B. As measured by all metrics, with few exceptions, the summarized prompting style leads to behavior that is closer to equilibrium (lighter color in the figure) than the full-chat prompting style.

Some specific examples are particularly informative. When considering summarized representations in Game A, the mean number of agents on a focal route (Figure 6a) is approximately 40% for summarized representations, and closer to 15% for full-chat representations (in equilibrium 50% of agents should choose the focal route). Summarized prompting also leads to much higher payoffs in Game A, and generally lower payoffs in Game B, relative to full chat prompting (see Figures 7a-7b). Summarized representations lead to significantly lower regret, as seen in Figures 8a-8b—regrets are 4-5x lower using summarized representations in Game A, and regret is near zero using summarized representations in Game B. Finally, and interestingly, summarized representations lead to more stable behavior, as measured by mean number of switches per agent (Figures 9a-9b). As we discuss in Section 5.5, one likely reason for this behavior is that full-chat representations induce myopia in the LLM agents in our simulations.

When there are distinctions in the value of actions, regret-based representations lead agents closer to equilibrium behavior than payoff-based representations. To elucidate this finding, we compare the left and right columns of each 2x2 submatrix in our figures. Note that a representation that reports regret (rather than payoff) will have the greatest salience in Game B, because in this game there is a weak dominant strategy (the route *O-L-R-D*), *and* this strategy yields lower payoffs. By contrast, in Game A, both routes are equally attractive in equilibrium, *and* both yield high payoffs.

In our findings, we see therefore that the impact of regret-based vs. payoff-based representations is not as stark for Game A. However, for Game B, we see that there is a strong impact of this choice of representation. In Figure 6b, we generally see a larger number of agents choosing the weak dominant strategy in regret-based representations, and a lower standard error as well, relative to payoff-based representations (the only exception is F-PE and F-RE, which have nearly identical standard errors, suggesting that regret may not reduce variance as strongly among full-chat representations). A similar behavior is found when comparing mean payoffs as well (Figure 7b). The effects of regret-based representations are even more dramatic in Game B when we study mean regret (Figure 8b) and mean switching behavior (Figure 9b). Taken together, these findings make clear that when there is useful "gradient" information comparing the chosen action to the optimal action, LLM agents benefit from being provided explicitly with this information in the form of regret.

By contrast, in the results for Game A across all figures, the results are more mixed, consistent with the observation that in this game either route is equally preferable. Notably, switching behavior is not necessarily reduced in Game A (Figure 9a) when regrets are provided, emphasizing the much stronger role that summarization plays in this regard.

Seeing only one's own action tends to reduce instability in agent behavior. Here we focus specifically on Figures 9a-9b. Notice that for both Games A *and* B, we see lower switching behavior with lower action informativenes. Lower action informativeness here helps reduce variability in agents' context, and thus their subsequent behavior. Indeed, this is consistent with our discussion of summarization above.

Interestingly, for the other metrics, we also see some evidence of the potential benefits of lower action informativeness, particularly in Game B: the mean number of agents on a focal route (Figure 6b) and mean regret (Figure 8b) are closer to equilibrium behavior. In general, these findings suggest that to the extent that action informativeness helps guide agents closer to equilibrium behavior, it

is more likely to do so by *lowering* (rather than increasing) the information provided about others' actions.

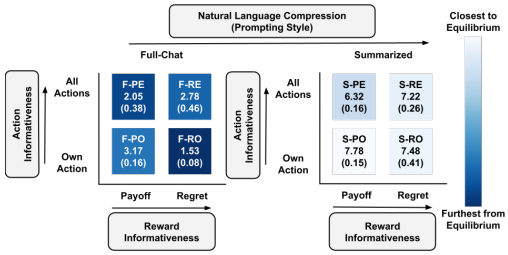

Fig. 6a. Mean agent route choices in Game A. We measure the number of agents on the least-congested route in each round.

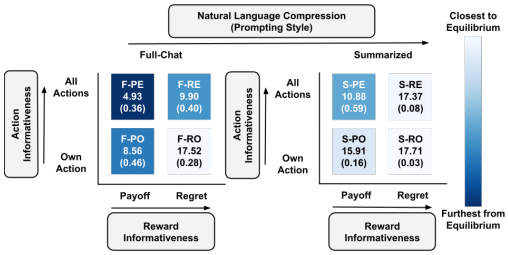

Fig. 6b. Mean agent route choices in Game B. We measure the number of agents on the dominant route.

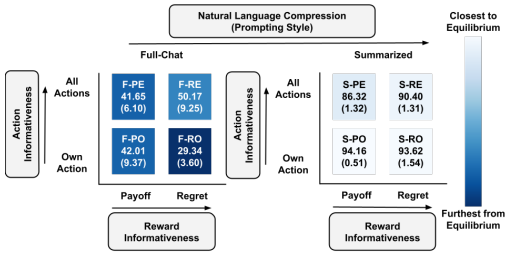

Fig. 7a. Mean agent payoffs in Game A.

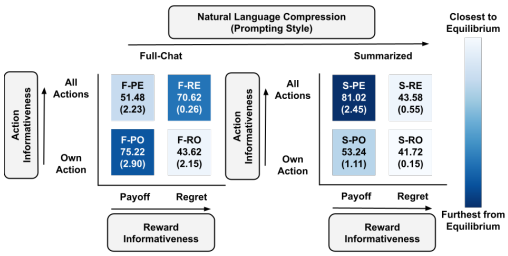

Fig. 7b. Mean agent payoffs in Game B. Color axis is reversed to indicate that lower payoffs are more desirable.

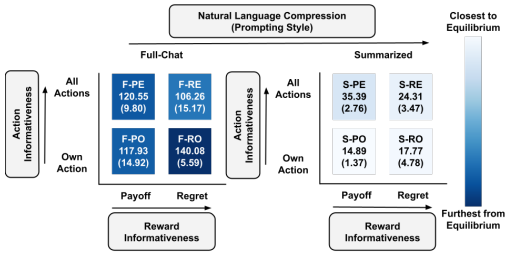

Fig. 8a. Mean agent regrets in Game A.

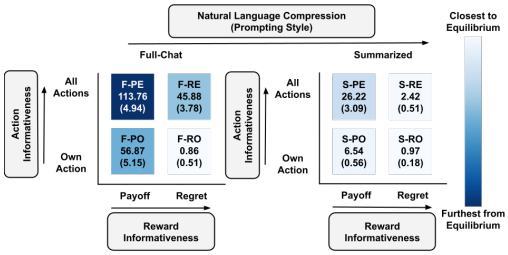

Fig. 8b. Mean agent regrets in Game B.

5.2 Sample Path Behavior

To expand on our findings above, in this section we analyze the dynamic behavior of agent play by looking at mean sample paths over the 40 round horizon. Figures 10a-13b display the round-by-round behavior of each of our four key metrics over time.

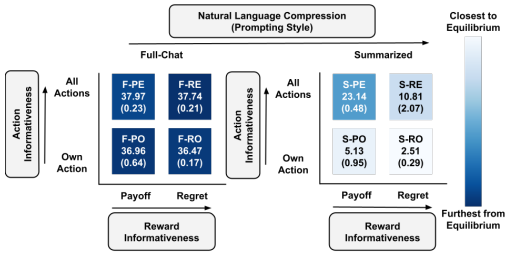

Fig. 9a. Mean agent switch count in Game A.

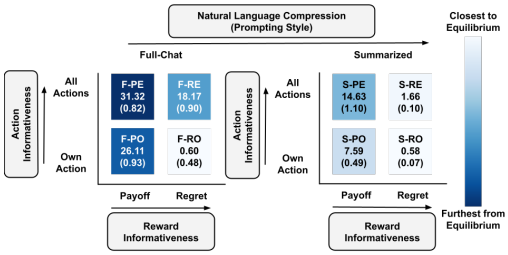

Fig. 9b. Mean agent switch count in Game B.

These graphs amplify our first finding above, namely, that summarized prompting generally leads to behavior that is closer to equilibrium in both games, both in terms of speed of convergence, and in terms of stability of behavior. In all graphs, there is generally a clear separation between the summarized representations and the full-chat representations. Perhaps the most striking visual finding is the distinction between round-to-round regret behavior between representations, displayed in Figures 12a and 12b for Games A and B, respectively. Here, summarized prompting dramatically reduces regret in both games to a similar degree. As previously observed, summarized prompting generally leads to less switching behavior over time than with full-chat prompting (Figures 13a-13b).

There is some nuance related to our second finding; namely, the potential impact of regret-based representations over payoff-based representations in guiding agents towards equilibrium actions. For example, in Game B, we see that F-RO (a regret-based representation) actually exhibits reasonably fast convergence to equilibrium (cf. Figure 10b), and also lower regret over time (cf. Figure 12b). These effects suggest that the regret-based representation offsets some of the complexity of the full chat prompting style in providing useful context to the LLM agent—since, again, in Game B, the primary goal is to identify the weak dominant strategy *O-L-R-D*. We also note here that this type of visualization reveals the potential for intriguing interactions *between* the different representation axes we have investigated.

Finally, by examining Figures 13a-13b, we can see the impact of action informativeness on behavior over time. In particular, conditional on using prompting style, not providing other agents' actions generally reduces the decision instability of agents in a persistent manner over the time horizon.

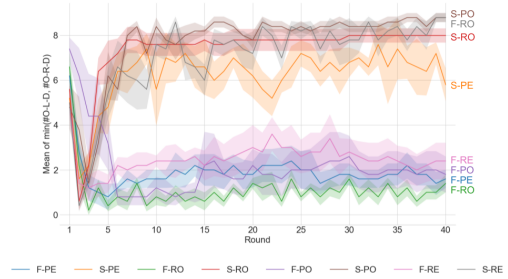

Fig. 10a. Round-to-round mean agent route choices in Game A (number of agents on the least-congested route).

Fig. 10b. Round-to-round mean agent route choices in Game B (number of agents on the dominant route).

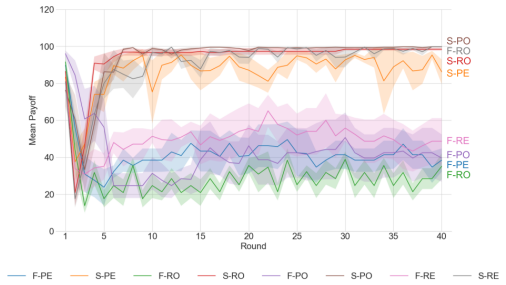

Fig. 11a. Round-to-round mean agent reward in Game A.

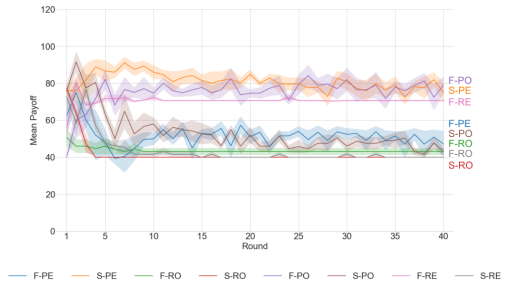

Fig. 11b. Round-to-round mean agent reward in Game B.

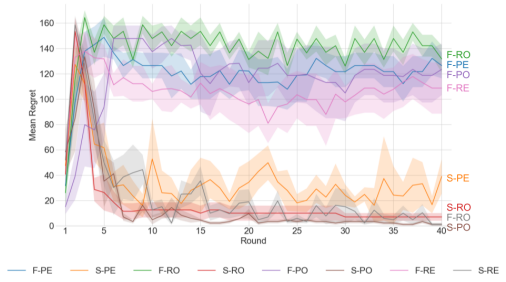

Fig. 12a. Round-to-round mean agent regret in Game A.

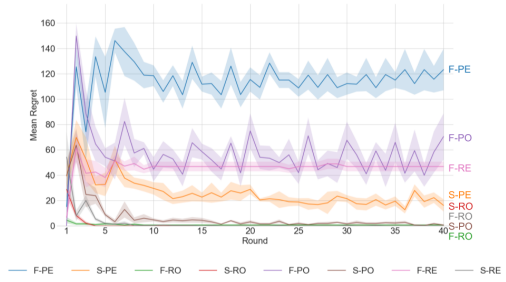

Fig. 12b. Round-to-round mean agent regret in Game B.

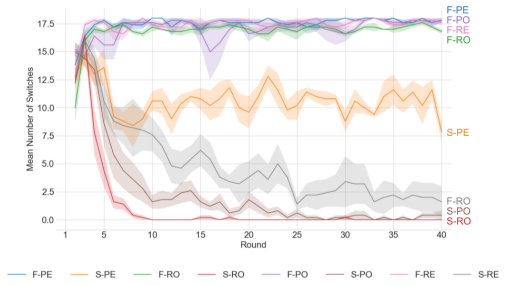

Fig. 13a. Round-to-round mean agent switch count in Game A.

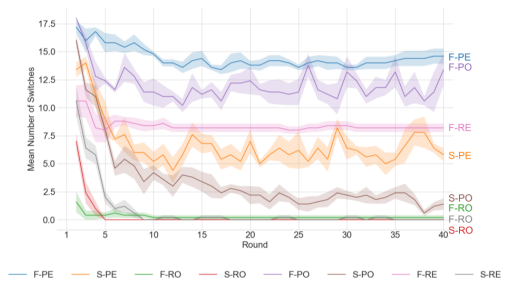

Fig. 13b. Round-to-round mean agent switch count in Game B.

5.3 Convergence to Equilibrium

Finally, we study convergence to equilibrium via Kendall's τ between round number and equilibrium deviation score in Game A (Fig. 14a) and Game B (Fig. 14b), cf. Section 4.3. Again, a negative (resp., positive) τ indicates agents are approaching (resp., diverging from) equilibrium as rounds progress; $\tau \approx 0$ suggests no systematic convergence trend. We note that we were unable to compute τ for two trials of representation F-RO in Game B; these trials exhibited constant deviation scores, and Kendall's τ is undefined if there is no variability in one of the variables.

In Game A, summarized representations (S-PE, S-RE, S-PO, S-RO) exhibit negative τ values, indicating steady decreases in deviation scores (likewise convergence to equilibrium). In contrast, full-chat representations (F-PE, F-RE, F-PO, F-RO) show weak or no convergence, with τ values

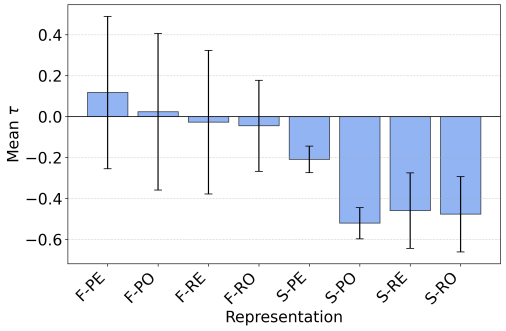

Fig. 14a. Mean τ across trials, Game A

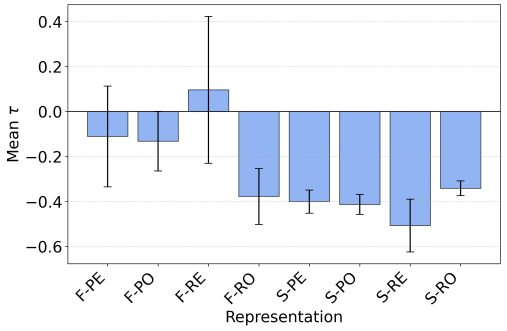

Fig. 14b. Mean τ across trials, Game B

Fig. 14. Kendall's τ between round number and equilibrium deviation score in Game A and Game B.

centered around zero and high variance among trials. Notably, F-RO achieves the best equilibrium convergence among full-chat representations, reinforcing that regret-based feedback encourages convergence.

In Game B, summarized representations yield the most negative τ values. Regret-based representations (F-RO, S-RE, S-RO) also show significantly negative τ, suggesting that regret encourages consistent learning. These findings quantify our temporal observations from sample paths in the preceding section.

5.4 Comparison with Learning Algorithms

For completeness, we also compare LLM agent behavior to two well-known no-regret learning algorithms: Multiplicative Weights Update (MWU) [Freund and Schapire, 1999] and EXP3 [Auer et al., 2002]. MWU operates under *full* feedback, meaning it has access to the payoffs of *all* possible actions in each round, even those it did not take. In contrast, EXP3 operates under *bandit* feedback, meaning it only observes the payoff of the action it actually took. The precise mathematical formulation of each algorithm is available in Appendix B. Both were simulated under identical conditions to the LLM agents ($n = 18$ agents, $T = 40$ rounds) over 50 trials. MWU used a learning rate of 0.75, and EXP3 an exploration rate of 0.75. Payoffs were rescaled from $[0, 400]$ to $[0, 1]$ to ensure convergence within 40 rounds.

Figure 15 compares the performance of EXP3 and MWU against LLM agent performance under S-RO—the representation that resulted in LLM behavior closest to equilibrium—across our four aggregate metrics from Section 5.1. In Figures 15a-15d, S-RO exhibits behavior closer to theoretical equilibrium than the two learning algorithms on all four aggregate metrics in Game B. In Game A as well, we observe that while mean routes, payoffs, and regrets are similar across S-RO, MWU, and EXP3 (Figures 15a-15c), S-RO results in far fewer switches on average (Figure 15d), suggesting more stability at equilibrium under S-RO.

5.5 Discussion and Future Directions

Our paper makes two primary contributions. First, we introduce a unifying framework for constructing natural language state representations in repeated games, systematically characterizing them along the axes of action informativeness, reward informativeness, and prompting style (or natural language compression). This framework can be used by researchers who seek to understand and empirically test the relationship between various axes of natural language state informativeness

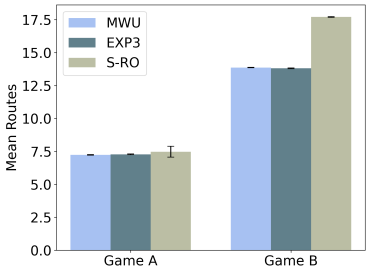

Fig. 15a. Mean agent route choices.

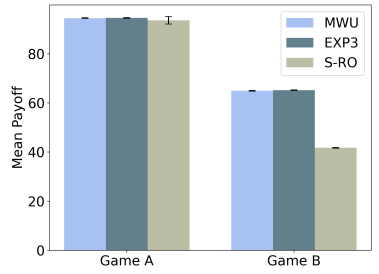

Fig. 15b. Mean agent payoff.

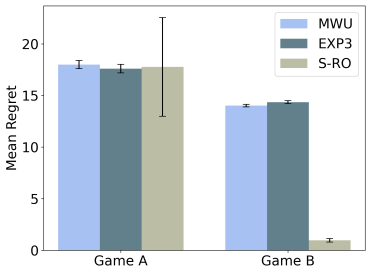

Fig. 15c. Mean agent regret.

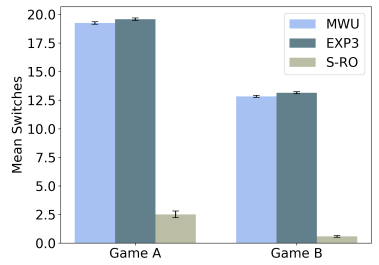

Fig. 15d. Mean agent switch count.

Fig. 15. Comparison of LLM agent performance under S-RO with learning algorithms EXP3 and MWU across four aggregate metrics in Games A and B.

on LLM agent behavior in their own game settings. Beyond dynamic games, this framework can be used for the augmentation and optimization of LLM agent as decision-makers in dynamic settings.

Second, we apply this framework to a repeated selfish routing game to rigorously evaluate how the dimensions of state representation influence LLM decision-making, and in particular use several evaluation metrics to understand how closely agents match equilibrium game play and the relative stability or variation in their play over time. We find in particular that (1) *summarized* representations, (2) representations that use *regret* rather than raw payoff, and (3) representations that include information only about one's *own action* rather than everyone's actions generally lead to the most stable, equilibrium-like behavior.

A heuristic inspection of the LLM agents' chat histories reveals some intriguing qualitative insights behind these findings. We briefly discuss these qualitative observations along each axis in our representation framework below. (For more detailed insights about LLM responses under various representations in Games A and B, see Table 2 in the appendix.)

- *Action informativeness.* Informally, we observed that providing information about everyone's actions could elicit counterfactual, anticipatory reasoning about actions that others could take in future rounds given their historical actions as well as observations about equilibrium and best response; however, the agent might then either proceed to make incorrect inferences with that information, or might fail to utilize that knowledge in subsequent decision-making, resulting in non-equilibrium play. For example, rather than avoiding routes that the agent believes others would switch to in the next round based on historical patterns, the agent might (erroneously) proceed to follow crowd behavior in the next round. On the other hand, providing the agent with only their own actions seemed to mitigate these types of errors.

- *Reward informativeness.* In both Games A and B, it appeared that providing agents with regrets caused agents to anchor their decision-making to patterns of low regrets obtained on certain routes. This type of information about the relative value of the best action helped stabilize game play and keep agents closer to equilibrium. On the other hand, when provided with raw payoff information, LLM agents seemed to reason about the risk of high congestion (or equivalently, low payoffs); this evaluation of risk led to more ambiguous inferences, increasing both instability and divergence from equilibrium play.
- *Prompting style.* Finally, while the full-chat prompting style tends to elicit anticipatory behavior in agents, it also could result in incorrect inferences made by agents, often because the agent myopically only accounts for the history information in the immediate preceding round. This gap meant that even agents' behavior might diverge from equilibrium despite the richer state information. By contrast, summarization preserved accurate inferences from history while encouraging decisions that are closer to the rational best response.

Of course, these are only anecdotal observations from our experiments. Indeed, one concrete direction for future work is to carry out far more sophisticated forensics on the internal behavior of these agents, uncovering the extent to which their manifested behaviors are consistent with behavioral biases observed in real-world human strategic behavior (e.g., [Kahneman and Tversky, 2013, Tversky and Kahneman, 1974]).

Our work suggests several other natural and fruitful directions for future exploration. First, our work focuses on selfish routing games with a specific topology. We expect that investigating a number of other game classes, and even some of those discussed in prior work (cf. Section 2) through the lens of our framework would provide broader evidence for the implications of state representations along the axes we have defined. Second, our work focuses on a particular model (GPT-4o), and we observe a particular range of quantitative and qualitative model behaviors as a result. Given the rapid pace of innovation in LLM models and architectures, we think it is natural to expand the line of inquiry we pursue in this paper to newer models, e.g., sequential reasoning models such as OpenAI's o1 and o3, and DeepSeek's R1 (that said, a current challenge here is the significant increase in simulation cost). Finally, as briefly discussed in Section 3, there may be other interesting dimensions of state representations to explore; notably, in more complex games with longer time horizons, it seems useful and interesting to explore the role of curtailing the depth of history provided to the agent.

References

Jacob D Abernethy, Elad Hazan, and Alexander Rakhlin. 2008. Competing in the Dark: An Efficient Algorithm for Bandit Linear Optimization.. In *COLT*. Citeseer, 263–274.

Gati Aher, Rosa I. Arriaga, and Adam Tauman Kalai. 2023. Using Large Language Models to Simulate Multiple Humans and Replicate Human Subject Studies. https://doi.org/10.48550/arXiv.2208.10264 arXiv:2208.10264 [cs].

Elif Akata, Lion Schulz, Julian Coda-Forno, Seong Joon Oh, Matthias Bethge, and Eric Schulz. 2023. Playing repeated games with Large Language Models. https://doi.org/10.48550/arXiv.2305.16867 arXiv:2305.16867.

Lisa P. Argyle, Ethan C. Busby, Nancy Fulda, Joshua Gubler, Christopher Rytting, and David Wingate. 2023. Out of One, Many: Using Language Models to Simulate Human Samples. *Political Analysis* 31, 3 (July 2023), 337–351. https://doi.org/10.1017/pan.2023.2 arXiv:2209.06899 [cs].

Sanjeev Arora, Elad Hazan, and Satyen Kale. 2012. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing* 8, 1 (2012), 121–164.

Peter Auer, Nicolò Cesa-Bianchi, Yoav Freund, and Robert E. Schapire. 2002. The Nonstochastic Multiarmed Bandit Problem. *SIAM J. Comput.* 32, 1 (Jan. 2002), 48–77. https://doi.org/10.1137/S0097539701398375

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. 2023. Transformers as Statisticians: Provable In-Context Learning with In-Context Algorithm Selection. https://doi.org/10.48550/arXiv.2306.04637 arXiv:2306.04637 [cs].

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. https://doi.org/10.48550/arXiv.2005.14165 arXiv:2005.14165.

Nicolò Cesa-Bianchi and Gábor Lugosi. 2006. *Prediction, Learning, and Games*. https://doi.org/10.1017/CBO9780511546921 Journal Abbreviation: Prediction, Learning, and Games Publication Title: Prediction, Learning, and Games.

Andrea Coletta, Kshama Dwarakanath, Penghang Liu, Svitlana Vyetrenko, and Tucker Balch. 2024. LLM-driven Imitation of Subrational Behavior : Illusion or Reality? http://arxiv.org/abs/2402.08755 arXiv:2402.08755 [cs, econ, q-fin].

José R Correa and Nicolás E Stier-Moses. 2011. Wardrop equilibria. *Encyclopedia of Operations Research and Management Science. Wiley* (2011).

Jinhao Duan, Renming Zhang, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Elias Stengel-Eskin, Mohit Bansal, Tianlong Chen, and Kaidi Xu. 2024. GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations. https://doi.org/10.48550/arXiv.2402.12348 arXiv:2402.12348.

Nicoló Fontana, Francesco Pierri, and Luca Maria Aiello. 2024. Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma? https://doi.org/10.48550/arXiv.2406.13605 arXiv:2406.13605 [physics].

Yoav Freund and Robert E. Schapire. 1999. Adaptive Game Playing Using Multiplicative Weights. *Games and Economic Behavior* 29, 1 (Oct. 1999), 79–103. https://doi.org/10.1006/game.1999.0738

Fulin Guo. 2023. GPT in Game Theory Experiments. https://doi.org/10.48550/arXiv.2305.05516 arXiv:2305.05516 [econ].

Sergiu Hart and Andreu Mas-Colell. 2013. *Simple adaptive strategies: from regret-matching to uncoupled dynamics*. Vol. 4. World Scientific.

Nathan Herr, Fernando Acero, Roberta Raileanu, María Pérez-Ortiz, and Zhibin Li. 2024. Are Large Language Models Strategic Decision Makers? A Study of Performance and Bias in Two-Player Non-Zero-Sum Games. https://doi.org/10.48550/arXiv.2407.04467 arXiv:2407.04467.

John J. Horton. 2023. Large Language Models as Simulated Economic Agents: What Can We Learn from Homo Silicus? https://doi.org/10.48550/arXiv.2301.07543 arXiv:2301.07543 [econ, q-fin].

Jen-tse Huang, Eric John Li, Man Ho Lam, Tian Liang, Wenxuan Wang, Youliang Yuan, Wenxiang Jiao, Xing Wang, Zhaopeng Tu, and Michael R. Lyu. 2024. How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments. https://doi.org/10.48550/arXiv.2403.11807 arXiv:2403.11807.

Jingru Jia, Zehua Yuan, Junhao Pan, Paul E. McNamara, and Deming Chen. 2025. Large Language Model Strategic Reasoning Evaluation through Behavioral Game Theory. https://doi.org/10.48550/arXiv.2502.20432 arXiv:2502.20432 [cs] version: 1.

Daniel Kahneman and Amos Tversky. 2013. Prospect theory: An analysis of decision under risk. In *Handbook of the fundamentals of financial decision making: Part I*. World Scientific, 99–127.

Ayato Kitadai, Sinndy Dayana Rico Lugo, Yudai Tsurusaki, Yusuke Fukasawa, and Nariaki Nishino. 2024. Can AI with High Reasoning Ability Replicate Human-like Decision Making in Economic Experiments? https://doi.org/10.48550/arXiv.2406.11426 arXiv:2406.11426.

Spyros Kontogiannis and Paul Spirakis. 2005. Atomic selfish routing in networks: A survey. In *International Workshop on Internet and Network Economics*. Springer, 989–1002.

Yihuai Lan, Zhiqiang Hu, Lei Wang, Yang Wang, Deheng Ye, Peilin Zhao, Ee-Peng Lim, Hui Xiong, and Hao Wang. 2024. LLM-Based Agent Society Investigation: Collaboration and Confrontation in Avalon Gameplay. https://doi.org/10.48550/arXiv.2310.14985 arXiv:2310.14985 [cs].

Tor Lattimore and Csaba Szepesvári. 2020. *Bandit algorithms*. Cambridge University Press.

Yan Leng. 2024. Can LLMs Mimic Human-Like Mental Accounting and Behavioral Biases? https://doi.org/10.2139/ssrn.4705130

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the Middle: How Language Models Use Long Contexts. *Transactions of the Association for Computational Linguistics* 12 (Feb. 2024), 157–173. https://doi.org/10.1162/tacl_a_00638

Dov Monderer and Lloyd S Shapley. 1996. Potential games. *Games and economic behavior* 14, 1 (1996), 124–143.

Chanwoo Park, Xiangyu Liu, Asuman Ozdaglar, and Kaiqing Zhang. 2024. Do LLM Agents Have Regret? A Case Study in Online Learning and Games. https://doi.org/10.48550/arXiv.2403.16843 arXiv:2403.16843.

Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. https://doi.org/10.48550/arXiv.2304.03442 arXiv:2304.03442 [cs].

Steve Phelps and Yvan I. Russell. 2024. The Machine Psychology of Cooperation: Can GPT models operationalise prompts for altruism, cooperation, competitiveness and selfishness in economic games? https://doi.org/10.48550/arXiv.2305.07970 arXiv:2305.07970 [cs].

Amnon Rapoport, Tamar Kugler, Subhasish Dugar, and Eyran J. Gisches. 2008. Braess Paradox in the Laboratory: Experimental Study of Route Choice in Traffic Networks with Asymmetric Costs. In *Decision Modeling and Behavior in Complex and Uncertain Environments*, Tamar Kugler, J. Cole Smith, Terry Connolly, and Young-Jun Son (Eds.). Springer, New York, NY, 309–337. https://doi.org/10.1007/978-0-387-77131-1_13

Amnon Rapoport, Tamar Kugler, Subhasish Dugar, and Eyran J. Gisches. 2009. Choice of routes in congested traffic networks: Experimental tests of the Braess Paradox. *Games and Economic Behavior* 65, 2 (March 2009), 538–571. https://doi.org/10.1016/j.geb.2008.02.007

Jillian Ross, Yoon Kim, and Andrew W. Lo. 2024. LLM economicus? Mapping the Behavioral Biases of LLMs via Utility Theory. https://doi.org/10.48550/arXiv.2408.02784 arXiv:2408.02784.

Tim Roughgarden. 2004. Selfish routing with atomic players. In *Proc. 16th symp. on discrete algorithms (soda)*. Citeseer, 1184–1185.

Tim Roughgarden. 2005. *Selfish Routing and the Price of Anarchy*. MIT Press. Google-Books-ID: 2hjQEAAAQBAJ.

Shai Shalev-Shwartz. 2007. *Online learning: Theory, algorithms, and applications*. Hebrew University.

Eilam Shapira, Omer Madmon, Itamar Reinman, Samuel Joseph Amouyal, Roi Reichart, and Moshe Tennenholtz. 2024. GLEE: A Unified Framework and Benchmark for Language-based Economic Environments. https://doi.org/10.48550/arXiv.2410.05254 arXiv:2410.05254.

Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty. *science* 185, 4157 (1974), 1124–1131.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762 arXiv:1706.03762 [cs].

Haochuan Wang, Xiachong Feng, Lei Li, Zhanyue Qin, Dianbo Sui, and Lingpeng Kong. 2024. TMGBench: A Systematic Game Benchmark for Evaluating Strategic Reasoning Abilities of LLMs. https://doi.org/10.48550/arXiv.2410.10479 arXiv:2410.10479.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. https://arxiv.org/abs/2201.11903v6 Publication Title: arXiv.org.

Lin Xu, Zhiyuan Hu, Daquan Zhou, Hongyu Ren, Zhen Dong, Kurt Keutzer, See Kiong Ng, and Jiashi Feng. 2024a. MAgIC: Investigation of Large Language Model Powered Multi-Agent in Cognition, Adaptability, Rationality and Collaboration. https://doi.org/10.48550/arXiv.2311.08562 arXiv:2311.08562 [cs].

Yuzhuang Xu, Shuo Wang, Peng Li, Fuwen Luo, Xiaolong Wang, Weidong Liu, and Yang Liu. 2024b. Exploring Large Language Models for Communication Games: An Empirical Study on Werewolf. https://doi.org/10.48550/arXiv.2309.04658 arXiv:2309.04658 [cs].

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. https://doi.org/10.48550/arXiv.2210.03629 arXiv:2210.03629 [cs].

Yauwai Yim, Chunkit Chan, Tianyu Shi, Zheye Deng, Wei Fan, Tianshi Zheng, and Yangqiu Song. 2024. Evaluating and Enhancing LLMs Agent based on Theory of Mind in Guandan: A Multi-Player Cooperative Game under Imperfect Information. https://doi.org/10.48550/arXiv.2408.02559 arXiv:2408.02559.

Xiaopeng Yu, Wanpeng Zhang, and Zongqing Lu. 2025. LLM-Based Explicit Models of Opponents for Multi-Agent Games. (2025).

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large Language Models Are Not Robust Multiple Choice Selectors. https://doi.org/10.48550/arXiv.2309.03882 arXiv:2309.03882.

A LLM Agent Prompts

Figure 16 below presents the game instructions that we provide to the LLM agents, adapted from the script given to human subjects in [Rapoport et al., 2009].

Game Environment (Round *i*)

You will be participating in an experiment on route selection in traffic networks. During this experiment you'll be asked to make many decisions about route selection in a traffic network game. Your payoff will depend on the decisions you make as well as the decisions made by the other participants. There are 18 participants in this experiment, including yourself, who will be asked to serve as drivers and choose a route to travel in a traffic network game that is described below. You will play the game for 40 identical rounds.
Consider the very simple traffic network described below.

Nodes:
O L R D

Segments and associated costs:
Segment O-L, cost function: 10 * X
Segment O-R, cost function: 210
Segment L-D, cost function: 210
Segment R-D, cost function: 10 * X

Each driver is required to choose one of 2 routes to travel from the starting point, denoted by O, to the final destination, denoted by D. There are 2 alternative routes and they are denoted by ['O-L-D', 'O-R-D'].
Travel is always costly in terms of the time needed to complete a segment of the road, tolls, fuel etc. The travel costs are written near each segment of the route you choose. For example, if you choose route O-L-D, you will be charged a total cost of 10X + 210 where X indicates the number of participants who choose segment O-L to travel from O to L plus a fixed cost of 210 for traveling on segment L-D.
Similarly, if you choose route O-R-D, you will be charged a total travel cost of 210 + 10Y, where Y indicates the number of participants who choose the segment R-D to drive from O to D. Please note that the cost charged for segments O-L and R-D depends on the number of drivers choosing them.
In contrast, the cost charged for traveling on segments L-D and O-R is fixed at 210 and does not depend on the number of drivers choosing them.
All the drivers make their route choices independently of one another and leave point O at the same time.

Example.

If you happen to be the only driver who chooses route O-L-D, and all other 17 drivers choose route O-R-D, then your travel cost from point O to point D is equal to (10 x 1) + 210 = 220. If, on another round, you and 2 more drivers choose route O-R-D and 15 other drivers choose route O-L-D, then your travel cost for that round will be 210 + (10 x 3) = 240.

At the beginning of each round, you will receive an endowment of 400 points. Your payoff for each round will be determined by subtracting your travel cost from your endowment. Your goal is to maximize your payoff (likewise minimize your cost). At the end of each round, you will be informed of the number of drivers who chose each route and your payoff for that round. All 40 rounds have exactly the same structure.

The output should be formatted as a JSON instance that conforms to the JSON schema below.

As an example, for the schema {"properties": {"foo": {"title": "Foo", "description": "a list of strings", "type": "array", "items": {"type": "string"}}}, "required": ["foo"]} the object {"foo": ["bar", "baz"]} is a well-formatted instance of the schema. The object {"properties": {"foo": ["bar", "baz"]}} is not well-formatted.

Here is the output schema:

```
{"properties": {"route": {"title": "Route", "description": "choice of route", "type": "string"}}, "required": ["route"]}
```

Fig. 16. Game description (adapted from [Rapoport et al., 2009]) and formatting instructions given with every API call. Payoff example given is adjusted according to whether agents are playing Game A or Game B. Formatting instructions are generated by LangChain.

B Baseline Algorithms and Performance

B.1 EXP3

The EXP3 (Exponential-weight algorithm for Exploration and Exploitation) algorithm, introduced in [Auer et al., 2002], is designed for multi-armed bandit settings where losses or rewards are determined by an arbitrary (possibly adversarial) environment. The algorithm proceeds as follows. Given a set of k arms, let $w_i(t)$ denote the weight assigned to arm i at time t. The steps are:

- Set initial weights $w_i(1) = 1$ for all arms $i \in [k]$
- Compute the probability of selecting each arm i as

$$p_i(t) = (1-\gamma)\frac{w_i(t)}{\sum_{j=1}^{k} w_j(t)} + \frac{\gamma}{k} \tag{3}$$

  where $\gamma \in [0,1]$ is the exporation parameter; we use $\gamma = 0.75$ for the purposes of this paper.
- Sample an arm i_t according to the probability distribution $p(t)$
- Observe the loss $\ell_{i_t}(t)$ and define an unbiased loss estimate as

$$\hat{\ell}_{i_t}(t) = \frac{\ell_{i_t}(t)}{p_{i_t}(t)} \tag{4}$$

- Update the weights according to

$$w_i(t+1) = w_i(t)\exp(-\eta\hat{\ell}_i(t)) \tag{5}$$

  where $\eta > 0$ is the learning rate. We set $\eta = 0.75$, the same as γ.
- Repeat for T rounds.

B.2 Multiplicative Weights Update

The Multiplicative Weights Update (MWU) algorithm was introduced in [Freund and Schapire, 1999]. It is a widely used technique for learning in repeated decision-making scenarios. MWU differs from EXP3 in that it has access to the losses of every action, opposed to only the chosen action. The algorithm is as follows. Let k denote the number of actions, and let each action $i \in [k]$ have corresponding weight $w_i(t)$. The steps are:

- Set initial weights $w_i(1) = 1$ for all $i \in [k]$.
- Compute the probability of selecting each action as

$$p_i(t) = \frac{w_i(t)}{\sum_{j=1}^{k} w_j(t)} \tag{6}$$

- Choose an action i_t according to the probability distribution $p(t)$.
- Observe the loss $\ell_i(t)$ for each action i
- Update the weights according to

$$w_i(t+1) = w_i(t) \cdot e^{-\eta\ell_i(t)} \tag{7}$$

  where $\eta > 0$ is the learning rate. We set $\eta = 0.75$.
- Repeat for T rounds.

B.3 Temporal Performance of EXP3 and MWU

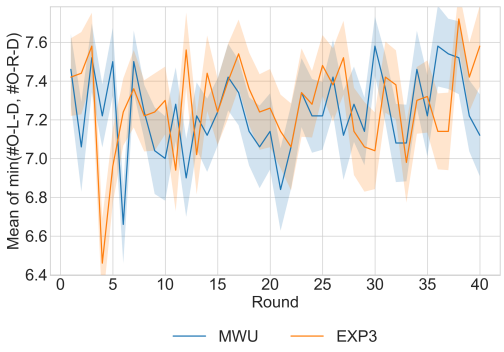

Fig. 17a. Mean route choice per agent, Game A

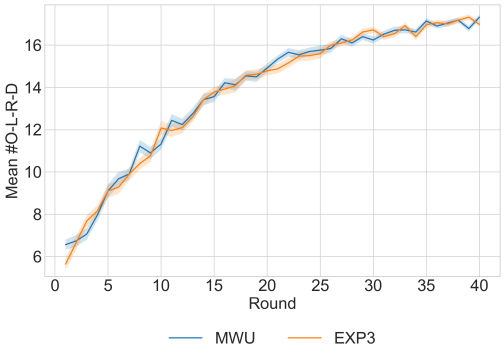

Fig. 17b. Mean route choice per agent, Game B

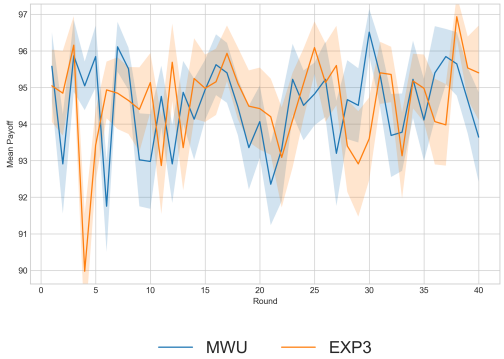

Fig. 18a. Mean payoff per agent, Game A

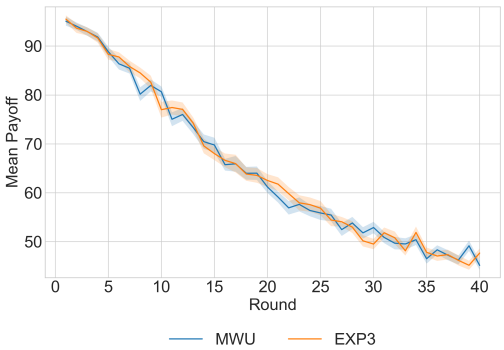

Fig. 18b. Mean payoff per agent, Game B

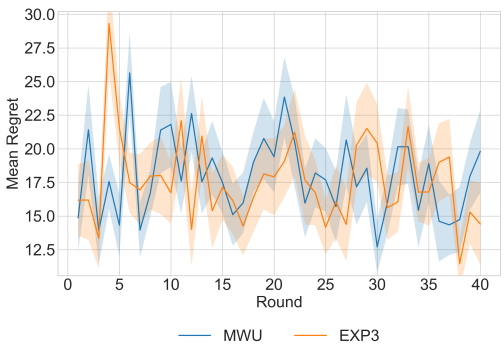

Fig. 19a. Mean regret per agent, Game A

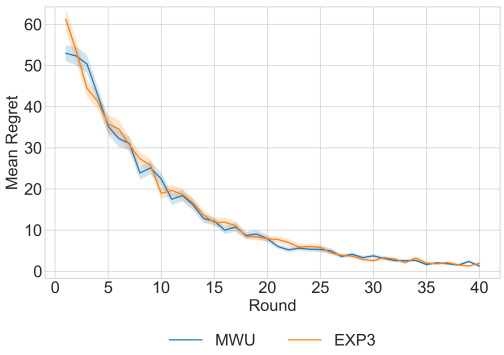

Fig. 19a. Mean regret per agent, Game B

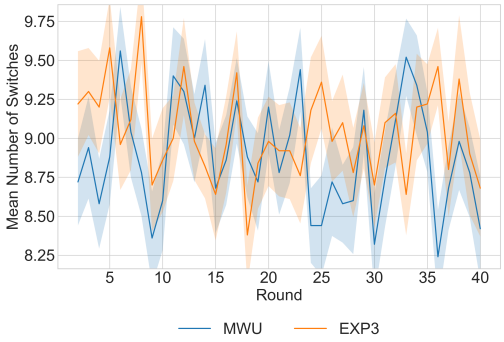

Fig. 20a. Mean number of switches per agent, Game A

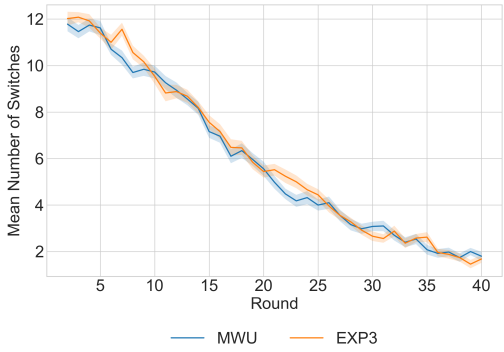

Fig. 20b. Mean number of switches per agent, Game B

C Qualitative Analysis of LLM Output

Table 2. Main drivers of decisions from LLM output in Games A and B for each state representation. These drivers are obtained from analyzing the LLM outputs for one agent in the last round of one trial of each game under each representation. The LLMs are prompted to provide their reasoning for their decisions through explicit instructions in the prompt to think step-by-step in a chain-of-thought manner.

	Game A	**Game B**
F-PO	Given the high congestion last round on O-L-D, **anticipates** many drivers may choose to switch to O-R-D in search of a less crowded route, but **mistakenly** decides to switch to O-R-D, predicting others will move away from O-L-D.	**Anticipates route dynamics** based on **potential switching by others after experiencing congestion**. Considers forecasting the movements of others in strategic considerations, but does not explain how.
S-PO	Choosing O-R-D is likely to result in a **consistent payoff** based on previous rounds. This route offers less risk of extremely low payoffs compared to O-L-D, since your payoff has been **more predictive** in these later rounds.	**Route analysis without regard for other's actions**. Decision strategy based on maximizing payoffs based on historical **patterns** of payoffs within routes.
F-RO	Recognizes that drivers respond to high regrets by **switching** routes, and thus fluctuating popularity between routes. **Anticipates** fewer drivers on O-R-D following its high regret cost in the previous round.	Considers the **consistency** of **zero regret** on O-L-R-D. Notes that it will keep observing for any changes in route dynamics, as these could affect future decisions.
S-RO	Sticks with O-R-D since it has **consistently** resulted in **minimal regret**. The stability in regret suggests that the **dynamics** among participants have settled into a **pattern** where O-R-D is preferable for minimizing costs.	Finds **pattern of no regret** on O-L-R-D, and suboptimal regret with other route choices historically.
F-PE	Observes that **congestion tends to shift** between routes, and thus infers choosing the **currently unused route** is likely to be beneficial. Switching to O-L-D should allow you to capitalize on **minimal congestion** and result in better payoff through lower costs.	**Avoids routes that start with O-L due to significant congestion**. Infers that because no drivers were on O-R-D, it potentially has lower congestion.
S-PE	Trades off between noting there is an **equilibrium** around 9 participants (where payoffs between routes are balanced), and that O-R-D experienced **high congestion** in recent rounds, making it less desirable.	Considers **congestion risk if everyone thinks similarly** and chooses O-L-R-D; considers **switching to less populated routes** if more participants begin moving to O-L-R-D. Chooses route that **balances acceptable payoff with congestion risk**.
F-RE	Given the **higher costs** on O-R-D, it reasons that it's reasonable to **expect that some drivers will switch** to O-L-D in order to minimize costs. However, it **mistakenly** believes that it makes sense for you to **switch** to O-L-D this round, aligning with the expected trend.	Evaluates route dynamics and **anticipates** driver behavior. Sticks with O-L-R-D due to its proven **consistency** in delivering **zero-regret** outcomes and cost-efficient performance.
S-RE	The strategic choice would be to select the route that has **recently shown minimal regret** under varying conditions, primarily when drivers are **distributed more evenly**. Choice relies on past observations where O-R-D **consistently** resulted in minimal regret.	**Every instance** of choosing O-L-R-D has resulted in **no regret**. Given the **complete convergence** to O-L-R-D **among participants** in **recent rounds**, **switching routes introduces unnecessary risk**.