Reasonably reasoning AI agents avoid game-theoretic failures in zero-shot, provably

Enoch Hyunwook Kang*
Foster School of Business, University of Washington

Abstract

As online platform markets become increasingly mediated by autonomous AI agents, a basic question is whether such markets generate stable strategic outcomes. In repeated strategic environments, Nash equilibrium provides a conservative minimum benchmark for such stability: failure to approach it already indicates a basic game-theoretic failure. Yet evidence on off-the-shelf LLM agents is mixed, and it remains unclear whether independently deployed agents converge to equilibrium behavior without universal post-training. In this paper, we provide an affirmative answer: we study infinitely repeated games and show that if agents 1) act Bayesian, i.e., learn beliefs about others' strategies from observed play and 2) asymptotically best respond to those beliefs, then along almost every realized play path, continuation behavior eventually becomes weakly close to a Nash equilibrium of the continuation game. We then connect this result to modern LLM agents by showing that their posterior-sampling behavior naturally induces this asymptotic best-response property. We further extend the analysis to the case in which stage payoffs are unknown ex ante and each agent observes only its own privately realized stochastic payoffs, obtaining the same convergence guarantee. Finally, we evaluate these implications in five repeated-game environments. The results are consistent with the theory: off-the-shelf reasoning agents display the learning-and-response patterns required for equilibrium-like play, whereas simpler myopic procedures do not. Taken together, the findings suggest that strategic stability in AI-mediated markets may arise from the intrinsic reasoning and learning properties of modern AI agents, even without universal fine-tuning.

*ehwkang@uw.edu

1 Introduction

A fundamental transition is underway in online platform marketplaces: autonomous AI agents are beginning to act on behalf of consumers in search, evaluation, and purchasing decisions (Gaarlandt et al., 2025; Shahidi et al., 2025). Rather than humans directly browsing rankings, comparing listings, and clicking through interfaces, AI agents can parse webpages or interact through APIs to evaluate products and transact. As a result, an increasing share of economically relevant platform activity may soon be mediated not by human attention directly, but by interaction among autonomous agents operating within platform-designed environments.

This shift makes a basic theoretical question newly urgent. In markets mediated by human participants, economists often analyze outcomes through the lens of stable strategic behavior emerging from repeated interaction. As autonomous AI agents increasingly act on behalf of users, it becomes unclear whether markets populated by such agents will exhibit the same kind of strategic regularity. The core issue, therefore, is not only whether an individual AI agent performs well, but whether many interacting AI agents collectively produce stable and predictable market outcomes. In digital markets, such interaction arises naturally in pricing, promotion, bidding, negotiation, matching, and recommendation environments, where each agent's payoff depends on how other agents behave over time (Bianchi et al., 2024; T. Guo et al., 2024; Lopez-Lira, 2025; Y. Li et al., 2024; Zhu et al., 2025; Bansal et al., 2025). In such settings, what matters for prediction is *whether the system settles into a stable strategic pattern of play*.

Nash equilibrium in repeated games is a natural conservative benchmark for such stability. In repeated strategic environments, Nash equilibrium provides a standard starting point for analyzing long-run behavior, incentives, and strategic predictability (Abreu and Rubinstein, 1988; Dal Bó and Fréchette, 2011; Dal Bó and Fréchette, 2018; Proto, Rustichini, and Sofianos, 2019). In this sense, asking the question of *whether AI agents converge toward Nash-like play in repeated games* is a minimal first step toward understanding strategic outcomes in AI-mediated markets. Failure to approach even this conservative benchmark can itself be regarded as a *game-theoretic failure*: if agents do not reach behavior that is at least approximately immune to profitable unilateral deviations in repeated games, then the resulting play lacks even the most basic form of strategic stability and is difficult to interpret as strategically coherent. This makes Nash convergence in repeated games a useful baseline question, even before considering richer models of dynamic interaction, e.g., Markov perfect Nash equilibrium (Aguirregabiria, Collard-Wexler, and Ryan, 2021).

This question is not merely theoretical. Recent work by Calvano et al. (2020) and Fish, Gonczarowski, and Shorrer (2024), together with related empirical studies of algorithmic interaction, suggests that autonomous algorithmic and AI systems can generate strategically consequential repeated-game behavior in economically important environments. Pricing algorithms can sustain supra-competitive outcomes without explicit communication, rapid reactive pricing technologies can elevate prices even in competitive equilibrium, and real-world adoption of algorithmic pricing has been associated with higher margins in concentrated markets (Assad et al., 2024).

On the other hand, empirical evaluations of LLMs reveal that widely used, off-the-shelf AI models (e.g., GPT, Claude, Gemini, Kimi, DeepSeek) as AI agents frequently fail to exhibit predicted equilibrium behavior in strategic interactions and often resort to brittle heuristics or produce inconsistent policies (X. Guo et al., 2024; J.-t. Huang et al., 2024; Hua et al., 2024; Buscemi et al., 2025). In practice, simply prompting standard AI models to engage in repeated games often yields strategies that diverge significantly from rational, equilibrium-based play predicted by classical game theory, although some successes have been reported (Akata et al., 2025). Such brittleness and inconsistency raise concerns about deploying AI agents in societally crucial domains that require reliable strategic decision-making. Accordingly, whether off-the-shelf reasoning LLM agents can be guaranteed to converge to a Nash equilibrium in repeated strategic interaction stably remains an open problem.

One prominent approach to addressing this problem is the use of targeted, universal post-training procedures (Park et al., 2024; Duque et al., 2024). However, relying on the universal deployment of such fine-tuning approaches across diverse, independently developed AI agents is often impractical. Consequently, there is a compelling need for assurance that off-the-shelf AI agents with some "reasonable" reasoning capabilities autonomously adapt their strategies and reach a stable equilibrium. This critical observation motivates the central research question explored in this paper:

Can off-the-shelf reasoning AI agents stably converge to Nash equilibrium in repeated strategic interaction without post-training?

In this paper, we theoretically and empirically address this question within the framework of infinitely repeated games, a setting in which agents repeatedly encounter identical strategic scenarios with no predefined endpoint. Specifically, we show that reasoning LLM-based AI agents can evolve toward Nash-like continuation play along realized play paths, without relying on explicit post-training or specialized fine-tuning procedures.

The key to achieving this lies in two basic reasoning capabilities we call "reasonably reasoning" capabilities: *Bayesian learning* and *asymptotic best-response learning*. By Bayesian learning, we refer to an agent's capacity to learn other agents' strategies from observed historical interactions, thereby enabling a theory-of-mind of others' future actions. By asymptotic best-response learning, we mean the agent's ability to eventually learn an optimal counter-strategy given the inferred beliefs about other agents' strategies, thereby maximizing its expected payoff. Under such capabilities, which we argue reasoning LLM agents can approximately satisfy, we prove that agents eventually behave, along almost every realized play path in infinitely repeated interaction, in a way that is weakly close to a Nash equilibrium of the continuation game.

Our main theoretical results are heavily rooted in a fundamental result in Bayesian learning literature (Kalai and Lehrer, 1993a; Norman, 2022) that a set of Bayesian learning agents with the ability to *exactly* best respond to their belief about the opponent agents' strategy, i.e., maximize their expected payoff, eventually exhibit a Nash equilibrium along every realized path in possibly infinitely repeated interactions. The key

difference in this paper's theoretical result is that it allows *asymptotic* best-response learning rather than assuming that the agent can choose the *exact* best-response action, i.e., the agent is an expected-utility maximizer. This is an important relaxation, as the off-the-shelf LLM agents are not expected utility maximizers (Yamin et al., 2026; Ge, Yongyan Zhang, and Vorobeychik, 2026). Rather, they are *stochastic posterior samplers* by default (i.e., in temperature = 1 setup) (Arumugam and Griffiths, 2025). We prove that, under mild and realistic assumptions, LLM agents, which are posterior belief samplers, achieve *asymptotic* best-response learning. We then prove that the fundamental result in Bayesian learning (Kalai and Lehrer, 1993a; Norman, 2022), which requires *exact* best-response capability, can be extended to *asymptotic* best-response learning. Combined with the recent findings that LLMs are Bayesian in-context learners under stationary, repeated settings (Coda-Forno et al., 2023; Lu et al., 2024; Cahyawijaya, Lovenia, and Fung, 2024; Xie et al., 2021; X. Wang et al., 2023; Wakayama and Suzuki, 2025; Falck, Z. Wang, and Holmes, 2024), we conclude that reasoning LLM agents are *reasonably reasoning* agents, and therefore eventually exhibit a Nash equilibrium along every realized path in possibly infinitely repeated interactions.

We also develop a weaker but practically important benchmark for *myopic* reasoning. If an agent only predicts opponents' *next* actions and best responds period by period, then one should not expect convergence to rich repeated-game equilibria supported by continuation incentives. Nevertheless, we show that such one-step predict–then–act reasoning is still sufficient for eventual convergence to a *stage-game* ε-Nash equilibrium along the realized path. This result provides a useful theoretical benchmark for simpler prompting-based procedures, such as SCoT (Akata et al., 2025), and clarifies what myopic reasoning can and cannot achieve relative to continuation-level planning.

Beyond the benchmark with common-knowledge stage payoffs, we also consider the practically relevant case in which payoffs are not known to agents ex ante and each agent observes only its own privately realized stochastic payoffs. We modify the agents' posterior sampling to not only sample an opponent-strategy hypothesis, but also sample a hypothesis for the agent's own mean payoff matrix (equivalently, its own payoff kernel within the known noise family). Under the analogous learning conditions together with an additional asymptotic public-sufficiency assumption on hidden private histories, posterior sampling recovers the same asymptotic on-path ε-best-response property and therefore inherits the zero-shot Nash convergence guarantees.

These theoretical results generate three distinct empirical predictions. First, because one-step predict–then–act reasoning is sufficient only for *stage-game* equilibrium convergence, simple myopic procedures should often succeed when the objective is merely to reach some stage-game Nash action. Second, because nontrivial repeated-game equilibria depend on continuation values, myopic reasoning should generally fail to sustain such paths, whereas agents that infer opponent strategies and evaluate continuation plans should succeed. Third, when payoffs are not known ex ante and must be learned from noisy private observations, the same separation should persist, although under a more demanding informational problem. Our experiments are designed to test these three implications directly by comparing a direct-action baseline, a myopic SCoT agent,

and our posterior-sampling best-response planner across five repeated-game environments.

This paper is structured as follows. Section 2 discusses related works. Section 3 introduces the setup. Section 4 defines reasonably reasoning agents and relates their Bayesian and best-response learning properties to in-context and test-time inference in language models. Section 5 presents the main zero-shot Nash convergence results. Section 6 discusses how we can extend the zero-shot Nash convergence result for unknown, stochastic payoffs. Then Section 7 provides empirical evidence of the theoretical contributions in this paper.

2 Related works

Bayesian Learning. The theoretical analysis of reasonably reasoning agents is based largely on the Bayesian learning literature. Bayesian learning in repeated games is defined by a fundamental tension between the ability to logically learn opponents' strategies and the ability to respond to them optimally. The foundational possibility result in Kalai and Lehrer, 1993a showed that if players' prior beliefs contain a "grain of truth" (absolute continuity) regarding the true distribution of play, then standard Bayesian updating guarantees that their predictions will eventually converge to the truth, thereby naturally culminating in a Nash equilibrium. However, Nachbar (1997) and Nachbar (2005) subsequently proved a negative result: requiring players to simultaneously maintain this grain of truth and perfectly best-respond across all possible counterfactual game histories leads to a mathematical contradiction, as the infinite sets of learnable strategies and optimizing strategies are often mutually singular. Norman, 2022 resolved this tension by introducing "optimizing learnability", the crucial insight that agents do not need to perfectly learn unreached counterfactuals; they only need to accurately predict and best-respond along the realized path of play. Nonetheless, Norman identified that a stubborn impossibility persists in a specific class of games called MM* games, where adversarial payoff geometries prevent learning and optimization from coexisting even on-path.

This paper systematically navigates these classic boundaries to guarantee zero-shot Nash convergence for LLM agents. We actively employ Kalai and Lehrer, 1993a grain of truth (Assumption 2) to guarantee predictive accuracy via the classic merging of opinions, and avoid Nachbar (1997) and Nachbar (2005)'s impossibility by formally adopting the on-path relaxation and non-MM* in Norman (2022). However, although employing the standard Bayesian learning setup (Kalai and Lehrer, 1993a; Norman, 2022) guarantees accurate forecasts of future on-path actions, it does not guarantee *posterior concentration*, as LLM agents are not expected-utility maximizers, and rather posterior belief samplers (Arumugam and Griffiths, 2025; Yamin et al., 2026; Ge, Yongyan Zhang, and Vorobeychik, 2026). To address this, we introduce the finite menu and KL separation condition (Assumption 3), which is necessary to mathematically force the LLM agent's posterior to concentrate onto a single point mass (Lemma 4.2). By forcing posterior concentration, the LLM agent's stochastic "predict-then-act" reasoning seamlessly stabilizes into an asymptotic best response.

Strategic capabilities of LLM agents. As LLMs are increasingly deployed as interactive agents, a growing literature studies whether LLMs behave strategically in canonical games, emphasizing preference representation, belief formation, and (approximate) best responses rather than taking equilibrium play for granted

(H. Sun et al., 2025; Jia et al., 2025). In one-shot normal-form, bargaining, and negotiation tasks, off-the-shelf models often follow plausible but context-sensitive heuristics: behavior can depart from equilibrium predictions and change markedly under small framing or instruction variations (F. Guo, 2023; C. Fan et al., 2023; Hua et al., 2024). Strategic performance can improve with model scale and reasoning scaffolds, but the remaining variance across prompts and settings is substantial (Kader and Lee, 2024).

These issues become more acute under repeated games, where payoffs depend on stable, history-contingent policies. Multi-agent evaluation benchmarks report large cross-model and cross-game heterogeneity and frequent non-equilibrium dynamics, especially in coordination and social-dilemma regimes (Mao et al., 2023; Duan et al., 2024; J.-t. Huang et al., 2024). Controlled repeated-game experiments similarly find that cooperation/reciprocity can emerge, but is fragile to opponent choice and to seemingly minor prompt or protocol changes (Akata et al., 2025; Fontana, Pierri, and Aiello, 2024; Willis et al., 2025). In market-style repeated settings, recent work further documents collusive or supra-competitive outcomes among LLM agents and highlights sensitivity to communication opportunities and wording choices (Fish, Gonczarowski, and Shorrer, 2024; Agrawal et al., 2025).

Overall, existing results demonstrate meaningful strategic adaptation but do not provide general, zero-shot guarantees that heterogeneous, independently deployed off-the-shelf agents will converge to predictable equilibrium behavior. Our paper targets this gap by identifying two basic theory-of-mind capabilities, Bayesian learning of opponents and asymptotic best-response learning, and proving that, under mild conditions, they imply Nash continuation play along realized paths in repeated games, without requiring explicit post-training or cross-agent coordination.

LLM agents as Bayesian in-context learners. A growing body of work links *in-context learning* (ICL), i.e., test-time adaptation that conditions prior history on a prompt without parameter updates, to Bayesian inference over latent task hypotheses. In stylized transformer meta-learning settings, Xie et al. (2021) argue that transformers trained over a task distribution can implement an implicit Bayesian update and produce posterior-predictive behavior from in-context data; related analyses formalize ICL as (approximate) Bayesian model averaging and study how this view depends on model parameterization and drives generalization (Yufeng Zhang et al., 2023). Moving beyond specific constructions, Falck, Z. Wang, and Holmes (2024) propose a martingale-based perspective that yields diagnostics and theoretical criteria for when an in-context learner's predictive sequence is consistent with Bayesian updating, while Wakayama and Suzuki (2025) provide a broader meta-learning theory in which ICL is provably equivalent to Bayesian inference with accompanying generalization guarantees. Empirically, LLMs also exhibit *meta*-adaptation across tasks presented in-context (Coda-Forno et al., 2023), and several abilities that appear "emergent" under scaling can be substantially attributed to improved ICL mechanisms (Lu et al., 2024). Complementing these viewpoints, X. Wang et al. (2023) model LLM ICL through a latent-variable lens, where demonstrations act as evidence about an unobserved task variable—clarifying why behavior can be highly sensitive to the specific examples and their ordering—and related results document few-shot in-context adaptation even

in low-resource language learning regimes (Cahyawijaya, Lovenia, and Fung, 2024). For agentic and repeated-interaction settings, these Bayesian-ICL perspectives motivate modeling an LLM agent's use of the interaction transcript as maintaining and updating a posterior over opponent strategies/types; autoregressive generation can then be interpreted as sampling-based decision-making from the induced posterior (K. W. Zhang et al., 2024; Welleck et al., 2024), providing a concrete bridge between in-context learning and belief-based strategic behavior.

Expected utility maximization and best response. Standard learning-in-games analyses often assume agents compute an exact best response to their posterior at every history (Kalai and Lehrer, 1993a; Norman, 2022). This is a poor behavioral model for off-the-shelf LLM agents, whose actions are induced by stochastic decoding and thus implement a distribution over choices rather than a deterministic maximization of expected utility. In probabilistic decision tasks, Yamin et al. (2026) find systematic belief–decision incoherence, suggesting that elicited probabilities should not be treated as beliefs that the model then perfectly best-responds to. In risky-choice experiments, Ge, Yongyan Zhang, and Vorobeychik (2026) similarly document substantial departures from expected-utility maximization and large sensitivity to prompting/model type, with behavior better described as noisy sampling. Arumugam and Griffiths (2025) argues that LLMs naturally implement posterior sampling. These results motivate replacing exact best response with a weaker, sampling-compatible notion, e.g., posterior-sampling policies, which are shown to achieve *asymptotic* best-response performance along the realized path.

3 Setup

3.1 Infinitely repeated game

We study interaction among a finite set of agents $I = \{1, 2, \ldots, N\}$ in an infinitely repeated (discounted) game with perfect monitoring of actions and common-knowledge stage payoffs. We define the game as the tuple

$$\mathcal{G} = \left(I, \{A_i\}_{i \in I}, \{u_i\}_{i \in I}, \{\lambda_i\}_{i \in I}\right)$$

where:

- I is the finite set of AI agents
- A_i is the finite set of actions available to agent i
- $A = \prod_{i \in I} A_i$ is the joint action space, where a joint action profile at round t is denoted $a^t = \left(a_1^t, \ldots, a_{|I|}^t\right) \in A$. ($a_i^t$ indicates the action of agent i at round t)
- $u_i : A \to [0, 1]$ is agent i's (known) stage-game payoff function
- $\lambda_i \in (0, 1)$ is the private discount factor used by agent i to value future payoffs.

At each round $t = 1, 2, \ldots$, each agent i simultaneously chooses an action $a_i^t \in A_i$, forming a joint action

profile $a^t \in A$, which is publicly observed. Agent i then receives the stage payoff

$$u_i(a^t) \in [0,1] \tag{1}$$

These stage payoffs induce a standard infinitely repeated game with perfect monitoring of actions.

In defining the payoffs $\{u_i\}_{i\in I}$, we restrict the set of games considered in this paper using the following standard assumption in the Bayesian learning literature (Norman, 2022). Intuitively, this excludes games without a pure-strategy equilibrium, e.g., rock-scissors-paper; rigorously, it rules out the pathological class in which on-path learning cannot be patched into nearby Nash behavior.

Assumption 1 (Non-MM* game (Norman, 2022))**.** Consider the infinitely repeated game induced by the true stage payoffs $\{u_i\}_{i\in I}$ in equation (1). For each player i, define the stage-game minmax payoff and pure-action maxmin payoff as

$$\varphi_i := \min_{\sigma_{-i}\in\Delta(A_{-i})} \max_{\sigma_i\in\Delta(A_i)} u_i\left(\sigma_i, \sigma_{-i}\right), \quad \Phi_i^\star := \max_{a_i\in A_i} \min_{a_{-i}\in \mathrm{BR}_{-i}(a_i)} u_i\left(a_i, a_{-i}\right),$$

where $\mathrm{BR}_{-i}\left(a_i\right)$ denotes the set of opponents' (joint) best responses to a_i in the stage game. We call that the stage game is $\mathrm{MM}^\star$ if $\Phi_i^\star < \varphi_i$ for every i. We assume the stage game is not $\mathrm{MM}^\star$ (equivalently, $\Phi_i^\star \geq \varphi_i$ holds for some i).

3.2 Strategy

We define the joint action history at round t as $h^t = \left(a^1, a^2, \ldots, a^{t-1}\right)$, and

$$H^t = \left\{\left(a^1, a^2, \ldots, a^{t-1}\right) : a^s \in A \text{ for } s \leq t-1\right\}.$$

Let $H^0 := \{\emptyset\}$ denote the empty history. Denote the complete set of possible histories as $H = \bigcup_{t\geq 0} H^t$. (Throughout this paper, we allow AI agents' strategies to have bounded memory.)

Definition 1 (Strategy)**.** A strategy for agent i is a function

$$f_i : H \to \Delta\left(A_i\right),$$

which maps every joint action history to a distribution over agent i's actions A_i.

Let $\mathcal{F}_i$ denote the space of all strategies of agent i. A strategy profile is a tuple $f = (f_1, \ldots, f_N) \in \mathcal{F} = \prod_{i\in I} \mathcal{F}_i$. Let H^∞ denote the space of infinite play paths, i.e.,

$$H^\infty = \left\{\left(a^1, a^2, \ldots\right) : a^t \in A \text{ for all } t \in \mathbb{N}\right\}.$$

Definition 2 (Play-path distribution)**.** A strategy profile $f = (f_1, \ldots, f_N) \in \mathcal{F}$ induces a unique probability

distribution μ^f over H^∞ (the *play-path distribution*), defined on cylinder sets by

$$\mu^f\left(C\left(a^1,\ldots,a^t\right)\right) := \prod_{s=1}^{t}\prod_{i\in I} f_i\left(h^s\right)\left(a_i^s\right),$$

where $h^s = (a^1,\ldots,a^{s-1})$ and $C(h) := \{z \in H^\infty : z = (h,\ldots)\}$. By Kolmogorov's extension theorem (Durrett, 2019), these finite-dimensional probabilities define a unique probability measure μ^f on $(H^\infty, \mathcal{B})$, where $\mathcal{B}$ is the product σ-algebra.

For the upcoming discussions, we fix some notations. Given that we fix a history h^t, for any continuation profile g (i.e., a profile that specifies play after histories extending h^t), let $\mu^g_{h^t}$ denote the induced distribution on H^∞ over the future joint-action sequence $(a^t, a^{t+1},\ldots)$ when play starts at history h^t and follows g thereafter. Formally, we identify the tail $(a^t, a^{t+1},\ldots)$ with $y \in H^\infty$ by setting $y^1 = a^t$, $y^2 = a^{t+1}$, and so on, and regard $\mu^g_{h^t}$ as a measure on this reindexed space. For a full profile $g \in \mathcal{F}$, we write $\mu^g_{h^t}$ for the continuation distribution induced by its restriction $g|_{h^t}$. If $\mu^g(C(h^t)) > 0$, then $\mu^g_{h^t}$ coincides with the conditional distribution $\mu^g(\cdot \mid h^t)$.

3.3 Beliefs

Each agent i acts under uncertainty regarding the opponents' future play f_{-i}. The agent maintains subjective beliefs over opponents' strategies and updates them as the game unfolds.

Behavioral representatives (belief-equivalent behavior strategies). Fix player i and a (possibly mixed) belief μ_i over opponents' strategy profiles $\mathcal{F}_{-i}$. For any own strategy $g_i \in \mathcal{F}_i$, μ_i induces a predictive distribution over play paths

$$P_i^{\mu_i,g_i}(E) := \int_{\mathcal{F}_{-i}} \mu^{(g_i,f_{-i})}(E)\, d\mu_i(f_{-i}) \qquad \text{for measurable } E \subseteq H^\infty.$$

By Kuhn's theorem Kuhn, 1953 and Aumann's extension to infinite extensive-form games Aumann, 1961, there exists a behavior-strategy profile $\bar{f}_{-i} \in \mathcal{F}_{-i}$ such that for every g_i,

$$\mu^{(g_i,\bar{f}_{-i})} \;=\; P_i^{\mu_i,g_i}.$$

We call any such $\bar{f}_{-i}$ a *behavioral representative* (or *belief-equivalent* profile) of μ_i (Kuhn, 1953; Aumann, 1961; Kalai and Lehrer, 1993a). When μ_i has finite support $\{g^1_{-i},\ldots,g^K_{-i}\}$, one convenient choice is

$$\bar{f}_{-i}(h)(a_{-i}) = \sum_{k=1}^{K} \mu_i(g^k_{-i} \mid h)\, g^k_{-i}(h)(a_{-i}),$$

for histories h where Bayes' rule is defined.

Prior and posterior predictive beliefs. Agent i holds a subjective prior μ_i^0 over $\mathcal{F}_{-i}$. Write $P_i^{0,g_i} := P_i^{\mu_i^0,g_i}$ for the induced prior predictive distribution. As we discussed above (as used explicitly in Kalai and Lehrer

(1993a)), there exists a behavioral representative $f^i_{-i} \in \mathcal{F}_{-i}$ such that, for every g_i, $\mu^{(g_i, f^i_{-i})} = P_i^{0,g_i}$. We fix such an f^i_{-i} and call it agent i's *subjective expectation* of opponents' play.

At any history h^t where Bayes' rule is defined, μ_i^0 yields a posterior $\mu_i^t(\cdot \mid h^t)$ and a posterior predictive continuation belief. Let $f^{i,t}_{-i}$ denote any behavioral representative of this posterior predictive belief. As a standing convention, we take these representatives to be chosen consistently by continuation:

$$f^{i,t}_{-i}|_{h^t} \;:=\; f^i_{-i}|_{h^t},$$

i.e., the time-t posterior predictive continuation is represented by the restriction of the fixed belief-equivalent profile f^i_{-i} to histories extending h^t.

3.4 Subjective utility and Nash equilibrium

Subjective Expected Utility. An agent evaluates the optimality of a continuation strategy based on their subjective beliefs at a given history. Fix a history h^t and let $\sigma_i \in \mathcal{F}_i(h^t)$ be a continuation strategy for agent i from h^t onward. For any opponents' continuation profile g_{-i}, denote by $\mu_{h^t}^{(\sigma_i, g_{-i})}$ the induced distribution over future play paths when play starts at h^t and follows (σ_i, g_{-i}) thereafter.

Following the standard literature (Kalai and Lehrer, 1993b), we define the belief-explicit *subjective expected utility* of playing σ_i starting at h^t as

$$V_i(\sigma_i \mid h^t; g_{-i}) = \mathbb{E}_{y \sim \mu_{h^t}^{(\sigma_i, g_{-i})}} \left[(1-\lambda_i) \sum_{k=0}^{\infty} \lambda_i^k u_i(y^{k+1}) \right], \tag{2}$$

where $y = (y^1, y^2, \dots)$ represents the future path of joint actions relative to time t, with y^{k+1} denoting the joint action at step $k+1$ of this future path (i.e., at absolute time $t+k$).

When $g_{-i} = f^{i,t}_{-i}$, we write

$$V_i(\sigma_i \mid h^t) := V_i(\sigma_i \mid h^t; f^{i,t}_{-i}). \tag{3}$$

For any belief about opponents' continuation play g_{-i} at history h^t, we define the set of ε-best-response continuation strategies for agent i at h^t as

$$\mathrm{BR}_i^{\varepsilon}(g_{-i} \mid h^t) = \left\{ \sigma_i \in \mathcal{F}_i(h^t) : V_i(\sigma_i \mid h^t; g_{-i}) \geq \sup_{\sigma_i' \in \mathcal{F}_i(h^t)} V_i(\sigma_i' \mid h^t; g_{-i}) - \varepsilon \right\}.$$

Nash equilibrium. The true performance of a strategy profile $f \in \mathcal{F}$ for agent i is given by:

$$U_i(f) = \mathbb{E}_{z \sim \mu^f} \left[(1-\lambda_i) \sum_{t=1}^{\infty} \lambda_i^{t-1} u_i\left(z^t\right) \right],$$

where $z^t \in A$ is the joint action at round t, and $\lambda_i \in (0,1)$ is agent i's discount factor. The factor $(1-\lambda_i)$ is a normalization ensuring that $U_i(f) \in [0,1]$ whenever $u_i(a) \in [0,1]$ for all $a \in A$.

Definition 3 (ε-Nash equilibrium)**.** A strategy profile $f = (f_1, \ldots, f_N) \in \mathcal{F}$ is an ε-Nash equilibrium if, for every agent $i \in I$,

$$U_i(f) \geq \sup_{f_i' \in \mathcal{F}_i} U_i\left(f_i', f_{-i}\right) - \varepsilon.$$

4 Reasonably Reasoning Agents

As discussed earlier, one of the key ideas of this work is that reasoning LLM-based AI agents are fundamentally "reasonably reasoning" agents. In this section, we formally define the class of reasonably reasoning agents, and then demonstrate why reasoning-LLM agents are naturally reasonably reasoning agents. The definition isolates two ingredients: (i) Bayesian learning and (ii) an *on-path, asymptotic* notion of ε-consistency.

Definition 4 (Reasonably Reasoning Agent)**.** Fix a repeated game and a strategy profile $f = (f_i)_{i \in I}$ generating the objective play-path distribution μ^f (Definition 2). Player i is a *Reasonably Reasoning* (RR) agent if the following hold.

- **Bayesian learning:** Player i has a prior μ_i^0 over opponents' strategy profiles $\mathcal{F}_{-i}$ and forms posteriors $(\mu_i^t)_{t \geq 0}$ by Bayes' rule. Let $f_{-i}^{i,t}$ denote any behavioral representative of player i's posterior predictive continuation belief at history h^t (as in Section 3.3), so that for every continuation strategy σ_i,

$$V_i(\sigma_i \mid h^t) = V_i(\sigma_i \mid h^t; f_{-i}^{i,t}).$$

- **Asymptotic ε-consistency on-path:** For every $\varepsilon > 0$,

$$\mu^f\left(\left\{z : \exists\, T_i(z,\varepsilon) < \infty \text{ s.t. } \forall\, t \geq T_i(z,\varepsilon),\ f_i|_{h^t(z)} \in \mathrm{BR}_i^{\varepsilon}(f_{-i}^{i,t}|_{h^t(z)} \mid h^t(z))\right\}\right) = 1.$$

Intuitively, the "Bayesian learning" condition ensures that agents update their strategic beliefs coherently given observations. The "asymptotic ε-consistency" condition captures the idea that after a possibly long initial stumbling phase, agents eventually learn to play (approximately) optimal continuation strategies relative to their own beliefs along the realized path of play. It generalizes Norman's ε-consistency (Norman, 2022), which requires ε-best responding at *all* times (not only eventually) on a full-measure set of paths. This generalization is critical, as LLM-based AI agents are not expected-utility maximizers but rather posterior belief samplers (Arumugam and Griffiths, 2025; Yamin et al., 2026; Ge, Yongyan Zhang, and Vorobeychik, 2026).

4.1 Bayesian learning

The Bayesian-learning component of Definition 4 does not require an agent to explicitly store a symbolic prior over the full (and typically infinite-dimensional) strategy space $\mathcal{F}_{-i}$. Instead, what matters for

decision-making is that, after observing a public history h^t, the agent induces a coherent *posterior predictive* distribution over opponents' *continuation play*.

In repeated interaction, the latent object of inference is not merely the opponents' next-period action, but their *repeated-game strategy*: a reaction rule mapping histories to action distributions. While realized actions vary with the evolving public history, the underlying reaction rule is time-invariant; learning is therefore best understood as refining uncertainty about that rule (and, crucially, about its *predictive implications* for future play).

Formally, let μ_i^0 denote player i's subjective prior over opponents' strategy profiles $\mathcal{F}_{-i}$, and let $\mu_i^t(\cdot \mid h^t)$ denote the posterior obtained by Bayes' rule after history h^t whenever it is defined. The continuation problem depends on μ_i^t only through the induced posterior predictive distribution over future play, because continuation values are computed by integrating payoffs against that predictive distribution. Following Kalai and Lehrer, 1993a, we represent player i's posterior predictive continuation belief by a behavioral profile $f_{-i}^{i,t}$, chosen (without loss of generality) so that along the realized history $h^t(z)$,

$$f_{-i}^{i,t}\big|_{h^t(z)} \equiv f_{-i}^{i}\big|_{h^t(z)}, \tag{4}$$

where f_{-i}^{i} is a fixed belief-equivalent profile representing player i's *prior predictive* distribution as in Section 3. Thus, the continuation of a single belief-equivalent behavioral profile can be taken to match the time-t posterior predictive continuation belief along the realized path.

To guarantee that Bayesian updating is well-defined and that predictive beliefs can converge to the truth on-path, we impose the standard grain-of-truth condition.

Assumption 2 (Grain of truth (Kalai and Lehrer, 1993a))**.** For each player i, the objective play-path distribution μ^f is absolutely continuous with respect to i's prior predictive distribution under f_i, i.e. $\mu^f \ll P_i^{0,f_i}$. Equivalently, any event that player i assigns zero probability under their prior predictive model has zero probability under the true play distribution induced by f.

Under Assumption 2, classical merging-of-opinions results (Blackwell and Dubins, 1962) imply that player i's posterior predictive continuation beliefs become accurate along μ^f-almost every realized play path. We formalize this later by showing that absolute continuity implies strong path prediction (Lemma 5.1).

4.2 LLM agents are Bayesian learning agents

The Bayesian-learning abstraction above matches what we can operationally observe from LLM agents: *history-conditioned predictive distributions*. An LLM, when prompted with the game rules and the realized interaction history, induces a conditional distribution over next tokens, which can be arranged to correspond to a distribution over a discrete label for an opponent strategy.

This "as if Bayesian" framing is appropriate for two reasons. First, the technical apparatus in Section 3 already

works at the level of predictive distributions: given any coherent family of history-conditioned forecasts, we may represent it by an equivalent belief over opponents' strategies via the behavioral representatives $f_{-i}^{i,t}$ (and, in particular, by a fixed belief-equivalent profile f_{-i}^{i} whose continuation matches posteriors along realized histories as in (4)). Second, recent theory and empirical evidence indicate that AI agents, most of which are auto-regressive LLM models, can implement Bayesian or approximately Bayesian in-context learning in repeated, stationary environments (Xie et al., 2021; Yufeng Zhang et al., 2023; Falck, Z. Wang, and Holmes, 2024; Wakayama and Suzuki, 2025). Interpreting the prompt history as data and the model's induced distribution as a posterior predictive therefore provides a principled bridge between LLM behavior and Bayesian-learning agents in repeated games.

Finally, Assumption 2 should be understood as a modeling requirement on the LLM agent's *support*: the agent's predictive model should not rule out (assign zero probability to) events that can actually occur under the true interaction induced by f. In practice, this corresponds to ensuring that the agent's elicited beliefs (or the menu used to elicit them) are sufficiently expressive and include mild stochasticity/trembles so that no on-path event receives zero predicted probability.

4.3 LLM agents achieve asymptotic ε-consistency

In LLM agents, the output mechanism is mediated by *stochastic decoding*. Even holding the prompt fixed, a standard LLM's output induces a *distribution* over outputs rather than a deterministic argmax rule. Empirically, LLMs exhibit substantial decision noise and can violate the coherence one would expect if they were consistently computing expected-utility-maximizing best responses to elicited beliefs (Yamin et al., 2026; Ge, Yongyan Zhang, and Vorobeychik, 2026). Rather, LLM agents are *posterior samplers*, which sample an output from their posterior belief over the output space in their mind (Arumugam and Griffiths, 2025; Cai et al., 2024).

This creates a methodological tension for our purposes, as the Bayesian learning literature's Nash equilibrium convergence arguments require a best-response property (e.g., Kalai and Lehrer, 1993a; Norman, 2022). The goal of this subsection is to reconcile these: we formalize a minimal "predict-then-act" rule that is faithful to sampling-based LLM behavior yet is still sufficient to guarantee *asymptotic* ε-best-response learning *on the realized play path*.

LLMs naturally induce posterior-sampling best response (PS-BR). Reasoning LLM-based AI agents are naturally scaffolded first to infer the situation from the previous interactions and then respond optimally to that inferred model (a theory-of-mind "infer, then respond" (Zhou et al., 2023; Riemer et al., 2024)). This behavior is formally defined as as *posterior-sampling best response (PS-BR)*: sample a hypothesis about the opponent from the current posterior, then best respond to that sampled hypothesis.

Definition 5 (Posterior sampling best response (PS-BR))**.** Fix player i and a history h^t. Given posterior $\mu_i^t(\cdot \mid h^t)$ over opponents' strategy profiles, PS-BR chooses a continuation strategy by:

1. sampling $\tilde{f}_{-i} \sim \mu_i^t(\cdot \mid h^t)$;

2. playing any best response $\sigma_i \in \mathrm{BR}_i(\tilde{f}_{-i} \mid h^t)$ in the continuation game after h^t.

Denote the resulting (randomized) continuation strategy by $\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid h^t)$.

Here, step 1, "sample $\tilde{f}_{-i} \sim \mu_i^t(\cdot \mid h^t)$", is simply querying an LLM (under its default temperature $\tau = 1$ setup) to output an opponent strategy label from the LLM's conditional distribution over allowed labels based on the previous interaction history. Step 2 is instantiated by evaluating a finite set of candidate self-strategies against that sampled opponent strategy via roll-out, and selecting the value-maximizing candidate. For implementation details used for experiments, see Appendix D.

Because PS-BR best responds to a *single draw* $\tilde{f}_{-i}$ rather than to the posterior predictive continuation $f_{-i}^{i,t}$, it can be suboptimal if the posterior remains dispersed: different posterior samples can induce different best responses, producing unstable play and potentially persistent deviations from best-response optimality. The key observation is that this suboptimality is entirely driven by posterior dispersion. The next lemma makes this quantitative by upper-bounding the best-response gap by a simple collision statistic of the posterior.

Lemma 4.1 (PS-BR is a D_i^t-best response)**.** *Fix player i and a history h^t. Suppose $\mu_i^t(\cdot \mid h^t)$ is supported on a finite set $\mathcal{S}_{-i}$ and write*

$$p_t(g_{-i}) := \mu_i^t(g_{-i} \mid h^t), \qquad g_{-i} \in \mathcal{S}_{-i}.$$

Define the posterior collision complement

$$D_i^t(h^t) \;:=\; 1 - \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2 \;=\; \Pr_{\tilde{g}, \tilde{g}' \sim \mu_i^t(\cdot \mid h^t)} [\tilde{g} \neq \tilde{g}'].$$

Let $\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid h^t)$ be PS-BR at h^t. Then

$$V_i(\sigma_{i,t}^{\mathrm{PS}} \mid h^t) \;\geq\; \sup_{\sigma_i} V_i(\sigma_i \mid h^t) \;-\; D_i^t(h^t).$$

Equivalently, $\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid h^t) \in \mathrm{BR}_i^{D_i^t(h^t)}(f_{-i}^{i,t}|_{h^t} \mid h^t)$.

The statistic $D_i^t(h^t) = 1 - \|p_t\|_2^2$ is 0 exactly when the posterior is degenerate (a point mass) and is close to 1 when the posterior is highly spread out. Thus Lemma 4.1 says: PS-BR is an approximate best response to the agent's posterior predictive belief, with an approximation error equal to the probability that two independent posterior samples would disagree.

To obtain RR's *asymptotic* ε-consistency, it suffices (by Lemma 4.1) to ensure that $D_i^t(h^t(z)) \to 0$ along μ^f-almost every realized path z. Intuitively, we need the agent's posterior to concentrate so that posterior sampling becomes (asymptotically) deterministic.

In general repeated games, full posterior concentration over an unrestricted strategy space is too much to ask (and is closely related to classic impossibility phenomena; see Nachbar, 1997; Nachbar, 2005). We

therefore impose a standard restriction that is also natural from an LLM-agent implementation perspective: the agent maintains a *finite menu* of opponent-strategy hypotheses and updates a posterior over that menu (Aoyagi, Fréchette, and Yuksel, 2024; Gill and Rosokha, 2024). In addition, we require an *on-path* KL separation condition ensuring that incorrect hypotheses are detectably different from the true strategy along the realized play path. This is exactly what makes posterior concentration (and hence vanishing sampling error) mathematically inevitable.

Assumption 3 (Finite menu and KL separation)**.** Fix player i. Assume the support of μ_i^0 is finite; write $\mathcal{S}_{-i} := \mathrm{supp}(\mu_i^0) \subseteq \mathcal{F}_{-i}$. Assume:

1. *(Menu grain of truth)* $f_{-i} \in \mathcal{S}_{-i}$ and $\mu_i^0(f_{-i}) > 0$.

2. *(Caution / uniform positivity)* There exists $\nu \in (0,1)$ such that for every $g_{-i} \in \mathcal{S}_{-i}$, every history h, and every $a_{-i} \in A_{-i}$,
$$g_{-i}(h)(a_{-i}) \geq \nu.$$

3. *(On-path KL separation)* For every $g_{-i} \in \mathcal{S}_{-i} \setminus \{f_{-i}\}$ there exists $\kappa_i(g_{-i}) > 0$ such that μ^f-a.s. in z,
$$\liminf_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(h^t(z)) \,\Big\|\, g_{-i}(h^t(z))\Big) \;\geq\; \kappa_i(g_{-i}),$$
where for distributions $p, q \in \Delta(A_{-i})$,
$$D_{\mathrm{KL}}(p\|q) := \sum_{a_{-i}\in A_{-i}} p(a_{-i}) \log \frac{p(a_{-i})}{q(a_{-i})}.$$

Assumption 3 is directly implementable in an LLM-agent pipeline: the menu $\mathcal{S}_{-i}$ is a finite library of opponent strategy templates, "caution" can be enforced by adding an arbitrarily small tremble (to avoid zero likelihoods), and KL separation is an identifiability condition stating that wrong templates are distinguishable from the truth along the realized interaction history (the only history that matters for on-path learning).

Under Assumption 3, standard likelihood-ratio arguments yield posterior concentration on the true hypothesis.

Lemma 4.2 (Posterior concentration under KL separation)**.** *Fix player i and suppose Assumption 3 holds for i. Then μ^f-a.s. in z,*
$$\mu_i^t(f_{-i} \mid h^t(z)) \;\longrightarrow\; 1, \qquad \textit{and hence} \qquad \max_{g_{-i}\in\mathcal{S}_{-i}\setminus\{f_{-i}\}} \mu_i^t(g_{-i} \mid h^t(z)) \;\longrightarrow\; 0.$$

Lemma 4.2 implies $D_i^t(h^t(z)) \to 0$ on-path, and then Lemma 4.1 upgrades PS-BR from a dispersion-dependent approximation to an *eventual* ε-best-response rule.

Proposition 4.3 (PS-BR implies asymptotic ε-consistency)**.** *Fix player i. Suppose player i uses PS-BR at every history and Assumption 3 holds for i. Then player i satisfies the asymptotic ε-consistency requirement in Definition 4.*

This proposition is the formal resolution of the "LLMs are stochastic samplers" issue: the standard sampling-based decoding (temperature $\tau \simeq 1$) induces randomness that prevents exact best-response optimality at any fixed time, but if the agent's posterior over a finite, identifiable hypothesis menu concentrates, then the induced sampling randomness becomes asymptotically negligible. Consequently, the agent's behavior converges (on-path) to ε-best-response play relative to its (accurate) predictive beliefs, which is exactly the RR requirement needed for the zero-shot Nash convergence results in Section 5.

The proofs of Lemmas 4.1–4.2 and Proposition 4.3 are deferred to Appendix B.

5 Zero-shot Nash convergence

We now show that the reasonably reasoning agents we defined in Section 4, together with a learnability condition on beliefs, generate play that is eventually weakly close to Nash equilibrium play along the realized path. This argument follows the weak-subjective-equilibrium framework in Norman, 2022, adapted to LLM agent-specific setups discussed in Section 4, i.e., (i) asymptotic (on-path) ε-consistency and (ii) the finite-menu KL-separation for verifying the learnability condition.

5.1 Weak subjective equilibrium

We work with the standard weak distance on play-path distributions. Let $\mathcal{B}^t$ be the σ-algebra generated by cylinder events of length t.

Definition 6 (Weak distance)**.** For probability measures μ, ν over infinite play paths, define

$$d(\mu, \nu) \;:=\; \sum_{t=1}^{\infty} 2^{-t} \sup_{E \in \mathcal{B}^t} \big|\mu(E) - \nu(E)\big|.$$

For a history h^t with $\mu(C(h^t)) > 0$ and $\nu(C(h^t)) > 0$, define the conditional (continuation) weak distance

$$d_{h^t}(\mu, \nu) \;:=\; d(\mu(\cdot \mid C(h^t)),\; \nu(\cdot \mid C(h^t))).$$

We use weak distance to compare continuations of play after a realized history.

Definition 7 (Weak similarity in continuation)**.** Fix a history h^t. Two profiles f and g are *η-weakly similar in continuation after h^t* if

$$d_{h^t}(\mu^f, \mu^g) \;\le\; \eta.$$

Weak subjective equilibrium is Norman's key intermediate notion: players best respond (up to ξ) to their *subjective* model, and their subjective model is weakly close (within η) to the objective continuation distri-

bution.

Definition 8 (Weak subjective equilibrium (Norman, 2022))**.** Fix $\xi, \eta \geq 0$ and a history h^t. A continuation profile $f|_{h^t}$ is a *weak ξ-subjective η-equilibrium after h^t* if for every player i there exists a supporting profile $f^i = (f_i, f^i_{-i})$ such that:

1. *(Subjective best response)* $f_i|_{h^t} \in \mathrm{BR}_i^{\xi}(f^i_{-i}|_{h^t} \mid h^t)$, where payoffs are evaluated under μ^{f^i}.

2. *(Weak predictive accuracy)* $d_{h^t}(\mu^f, \mu^{f^i}) \leq \eta$.

Definition 9 (Learns to predict the path of play (strong))**.** Player i *learns to predict the path of play under f* if for every $\eta > 0$,

$$\mu^f\Big(\Big\{z : \exists\, T_i(z,\eta) < \infty \text{ s.t. } \forall\, t \geq T_i(z,\eta),\ d_{h^t(z)}(\mu^f, \mu^{f^i}) \leq \eta\Big\}\Big) = 1,$$

where $f^i = (f_i, f^i_{-i})$ is a supporting (belief-equivalent) profile for player i (as in Section 3).

Remark 1 (Connection to Optimizing Learnability). A longstanding challenge in Bayesian learning in games is Nachbar (1997) and Nachbar (2005)'s inconsistency result, which shows that requiring an agent to learn and best-respond on *all possible* continuation paths is often mathematically impossible. However, Norman (2022) resolved this by introducing *optimizing learnability*, the insight that agents only need to learn the continuation play along the realized paths generated by their optimizing choices. Our RR definition naturally instantiates Norman's insight: Definition 4 and Definition 9 require ε-consistency and predictive accuracy strictly μ^f-almost surely (i.e., strictly on the realized, optimizing play path). Therefore, the on-path merging of opinions guaranteed by Blackwell and Dubins (1962) is entirely sufficient for zero-shot Nash convergence, bypassing Nachbar's impossibility.

Crucially, while our agent's specific decision rule (PS-BR) requires finite menus and KL separation to guarantee the optimality of actions (asymptotic ε-consistency, Section 4), the learning of the true path (strong path prediction) relies purely on the absolute continuity of beliefs. It does not require the posterior to concentrate; it can be verified directly from Assumption 2 via the classic merging of opinions result. The following Lemma 5.1 formalizes this idea.

Lemma 5.1 (Absolute continuity implies strong path prediction)**.** *Fix player i. Suppose the objective play-path distribution μ^f is absolutely continuous with respect to player i's prior predictive distribution P_i^{0,f_i} (Assumption 2). Then player i learns to predict the path of play under f in the sense of Definition 9.*

The proof is deferred to Appendix B.

5.2 From learning to zero-shot Nash convergence

We first show that asymptotic ε-consistency, together with strong prediction, implies that the realized continuation play is eventually a weak subjective equilibrium.

Proposition 5.2. *Suppose each player i is RR (Definition 4) and learns to predict the path of play under f (Definition 9). Then for any $\xi > 0$ and $\eta > 0$,*

$$\mu^f\Big(\Big\{z : \exists\, T(z) < \infty \text{ s.t. } \forall\, t \geq T(z),\ f|_{h^t(z)} \text{ is a weak } \xi\text{-subjective } \eta\text{-equilibrium after } h^t(z)\Big\}\Big) = 1.$$

Finally, we convert a weak subjective equilibrium into proximity to a Nash equilibrium.

Theorem 5.3 (Zero-shot Nash convergence along realized play)**.** *Suppose every player i is RR and learns to predict the path of play under f. Assume the grain-of-truth condition (Assumption 2) holds for each player. Then for every $\varepsilon > 0$,*

$$\mu^f\big(\{z : \exists\, T(z) < \infty \text{ s.t. } \forall\, t \geq T(z),\ \exists\ \hat{f}^{\varepsilon,t,z} \text{ an } \varepsilon\text{-Nash equilibrium of the continuation game after } h^t(z) \text{ with } d_{h^t(z)}(\mu^f, \mu^{\hat{f}^{\varepsilon,t,z}}) \leq \varepsilon\}\big) = 1.$$

Corollary 5.4 (Zero-shot Nash convergence for PS-BR)**.** *Assume that for every player i, Assumption 3 holds and player i uses PS-BR (Definition 5). Then the conclusion of Theorem 5.3 holds.*

The proofs of Theorem 5.3 and Corollary 5.4 are deferred to Appendix B. As a direct consequence, under our practical PS-BR implementation, the premises of Theorem 5.3 are verified directly.

The main theoretical results, Theorem 5.3 and Corollary 5.4, may seem counter-intuitive: if each agent is learning, then what each agent is trying to predict is itself changing over time, so why should behavior ever stabilize? This concern is valid for many myopic learning models, where the learner treats the opponent as having a fixed action distribution even though the opponent is also adapting.

The promise of Bayesian learning (Kalai and Lehrer, 1993a) is that, under a suitable grain-of-truth condition, agents' *posterior predictive* forecasts about future play can nonetheless become accurate (merge) along the realized path. In repeated games, the correct object of inference is not a fixed action, but the opponent's *repeated-game strategy*: a fixed contingent plan (mapping histories to actions) that may be highly nonstationary. In particular, even if an opponent updates beliefs and changes its period-by-period best response, once its prior, update rule, and decision rule are fixed from time 0, its behavior defines a single mapping $f_{-i} : H \to \Delta(A_{-i})$ (hence a fixed repeated-game strategy in our sense). Agents' beliefs change because they refine uncertainty about this fixed mapping (and its on-path implications), not because the mapping is being rewritten exogenously over time.

Indeed, our main results do *not* require that posteriors over opponent strategies literally stop moving. Instead, they require on-path stabilization in two weaker senses:

1. *Stability of forecasts (predictive merging).* Under the grain-of-truth condition (Assumption 2), Bayesian updating implies that, along μ^f-almost every realized history $h^t(z)$, the agent's *posterior predictive*

distribution over future play becomes close to the true continuation distribution (formalized later by Definition 9 and Lemma 5.1). Importantly, this can happen even if the posterior over *strategy labels* does not concentrate: distinct strategy hypotheses may be observationally equivalent *on the realized path*, and any remaining disagreement can persist only on counterfactual histories that are not reached.

2. *Stability of (approximate) best responses.* Once an agent's predictive belief about continuation play is accurate on-path, playing an ε-best response to that belief is also nearly optimal against the true continuation play. Moreover, best-response *sets* need not vary wildly: when the payoff gap between the best action and the runner-up is nontrivial, small changes in beliefs do not change which continuation strategies are ε-optimal. This is exactly why our RR definition imposes only *asymptotic* on-path ε-consistency (Definition 4), rather than requiring perfect best-response optimality at every time and every counterfactual history.

Even if beliefs keep updating forever, behavior can still stabilize because decisions depend on the *predictive* implications of beliefs on the realized continuation game. If the posterior mass shifts among hypotheses that induce (nearly) the same continuation distribution after $h^t(z)$, then the agent's best-response problem is (nearly) unchanged, so play remains stable. For our PS-BR implementation with a finite menu and KL separation (Assumption 3), we obtain an even stronger form of stabilization: the posterior over the menu concentrates on the true opponent strategy (Lemma 4.2), so the randomness from posterior sampling becomes asymptotically negligible (Lemma 4.1), yielding eventual on-path ε-best-response behavior (Proposition 4.3).

5.3 Zero-shot stage-game Nash convergence for myopic rules

Theorem 5.3 and Corollary 5.4 establish eventual on-path convergence to a Nash equilibrium of the continuation game under PS-BR. That guarantee is deliberately strong: it concerns repeated-game optimality and therefore requires beliefs over opponents' full continuation strategies. Yet this level of reasoning may be unnecessary when the object of interest is only *stage-wise* strategic optimality. If we ask instead whether the realized mixed action profile at each history is eventually an approximate Nash equilibrium of the one-shot stage game, then predicting the opponents' next joint action may suffice. This reduction captures the logic of SCoT (Akata et al., 2025), which implements a "predict the next move, then best respond" procedure rather than full continuation planning. The purpose of this subsection is to justify this simplification formally. We analyze two one-step variants: myopic PS-BR, which best responds to a one-step predictive belief, and SCoT (Akata et al., 2025), which best responds to a deterministic point prediction of the opponents' next action.

5.3.1 Myopic PS-BR

myopic PS-BR retains the Bayesian-learning-plus-best-response structure of the previous subsection, but truncates both objects to one period: the agent forms a one-step predictive belief over the opponents' next joint action and then plays a myopic best response to that belief.

For notational convenience, as already used above, for any opponents' profile g_{-i} and history h, we write

$$g_{-i}(h) \in \Delta(A_{-i})$$

for the induced distribution over the opponents' joint next action at history h. In particular, when g_{-i} is an actual profile of opponents' mixed actions, this is the product distribution

$$g_{-i}(h) = \bigotimes_{j \neq i} g_j(h).$$

Definition 10 (One-shot stage-game ε-best response and stage ε-Nash)**.** For $\alpha_i \in \Delta(A_i)$ and $q \in \Delta(A_{-i})$, define

$$u_i(\alpha_i, q) := \sum_{a_i \in A_i} \sum_{a_{-i} \in A_{-i}} \alpha_i(a_i)\, q(a_{-i})\, u_i(a_i, a_{-i}).$$

For $\varepsilon \geq 0$, define

$$\mathrm{br}_i^{\varepsilon}(q) := \left\{ \alpha_i \in \Delta(A_i) : u_i(\alpha_i, q) \geq \sup_{\alpha_i' \in \Delta(A_i)} u_i(\alpha_i', q) - \varepsilon \right\}.$$

We also write

$$\mathrm{br}_i(q) := \mathrm{br}_i^0(q).$$

At a history h^t, write

$$f_{-i}(h^t) := \bigotimes_{j \neq i} f_j(h^t) \in \Delta(A_{-i})$$

for the actual current joint mixed action of player i's opponents. The current mixed-action profile

$$f(h^t) := (f_1(h^t), \ldots, f_N(h^t)) \in \prod_{j \in I} \Delta(A_j)$$

is a *stage ε-Nash equilibrium* if

$$f_i(h^t) \in \mathrm{br}_i^{\varepsilon}\big(f_{-i}(h^t)\big) \qquad \text{for every } i \in I.$$

Fix player i and let $f^i = (f_i, f^i_{-i})$, where f^i_{-i} is the fixed belief-equivalent profile from Section 3.3. Let $f^{i,t}_{-i}$ be the continuation-consistent representative of player i's predictive belief at history h^t. We write

$$q_i^t(\cdot \mid h^t) \;:=\; f^{i,t}_{-i}(h^t) \in \Delta(A_{-i}).$$

By the representative-choice convention from Section 3.3, along the histories under consideration,

$$f_{-i}^{i,t}(h^t) = f_{-i}^{i}(h^t).$$

When the posterior $\mu_i^t(\cdot \mid h^t)$ is supported on a finite set $\mathcal{S}_{-i} \subseteq \mathcal{F}_{-i}$, this is

$$q_i^t(\cdot \mid h^t) = \sum_{g_{-i} \in \mathcal{S}_{-i}} \mu_i^t(g_{-i} \mid h^t)\, g_{-i}(h^t)(\cdot).$$

Definition 11 (Myopic posterior-sampling best response (myopic PS-BR))**.** Fix player i and a history h^t. Suppose $\mu_i^t(\cdot \mid h^t)$ is supported on a finite set $\mathcal{S}_{-i}$. For each $g_{-i} \in \mathcal{S}_{-i}$, choose a mixed action

$$\alpha_i^{g_{-i},h^t} \in \mathrm{br}_i(g_{-i}(h^t)).$$

Myopic PS-BR:

1. samples $\tilde{f}_{-i} \sim \mu_i^t(\cdot \mid h^t)$;
2. uses the mixed action $\alpha_i^{\tilde{f}_{-i},h^t}$.

The induced ex ante mixed action is

$$\alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t) := \sum_{g_{-i} \in \mathcal{S}_{-i}} \mu_i^t(g_{-i} \mid h^t)\, \alpha_i^{g_{-i},h^t}(\cdot).$$

Whenever player i uses myopic PS-BR, we identify

$$f_i(h^t) = \alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t).$$

Lemma 5.5 (Stage best responses are stable under nearby beliefs)**.** *Fix player i and define*

$$\|p - q\|_{\mathrm{TV}} := \sup_{B \subseteq A_{-i}} |p(B) - q(B)| \qquad \textit{for } p, q \in \Delta(A_{-i}).$$

If $\alpha_i \in \mathrm{br}_i^{\xi}(q)$, then

$$\alpha_i \in \mathrm{br}_i^{\xi + 2\|p-q\|_{\mathrm{TV}}}(p).$$

Lemma 5.6 (Myopic PS-BR is a D_i^t-stage best response)**.** *Fix player i and a history h^t. Suppose $\mu_i^t(\cdot \mid h^t)$ is supported on a finite set $\mathcal{S}_{-i}$ and write*

$$p_t(g_{-i}) := \mu_i^t(g_{-i} \mid h^t), \qquad g_{-i} \in \mathcal{S}_{-i}.$$

Define

$$D_i^t(h^t) := 1 - \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2.$$

Let $\alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t)$ *be myopic PS-BR and let*

$$q_i^t(\cdot \mid h^t) = \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})\, g_{-i}(h^t)(\cdot)$$

be the one-step posterior predictive belief. Then

$$u_i\big(\alpha_{i,t}^{\mathrm{mPS}},\, q_i^t(\cdot \mid h^t)\big) \geq \sup_{\alpha_i \in \Delta(A_i)} u_i\big(\alpha_i,\, q_i^t(\cdot \mid h^t)\big) - D_i^t(h^t).$$

Equivalently,

$$\alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t) \in \mathrm{br}_i^{D_i^t(h^t)}(q_i^t(\cdot \mid h^t)).$$

Lemma 5.7 (Strong path prediction implies one-step predictive accuracy)**.** *Fix player i. Suppose player i learns to predict the path of play under f (Definition 9). Then*

$$\mu^f\Big(\Big\{z : \forall \eta > 0,\ \exists T_i(z,\eta) < \infty \text{ s.t. } \forall t \geq T_i(z,\eta),\ \big\|q_i^t(\cdot \mid h^t(z)) - f_{-i}(h^t(z))\big\|_{\mathrm{TV}} \leq \eta\Big\}\Big) = 1.$$

Theorem 5.8 (Bayesian convergence to stage-game Nash under myopic PS-BR)**.** *Assume that for every player i, Assumption 3 holds and player i uses myopic PS-BR (Definition 11) at every history. Then for every $\varepsilon > 0$,*

$$\mu^f\Big(\Big\{z : \exists T(z) < \infty \text{ s.t. } \forall t \geq T(z),\ f(h^t(z)) \text{ is a stage } \varepsilon\text{-Nash equilibrium}\Big\}\Big) = 1.$$

5.4 SCoT (Akata et al., 2025)

The second reduction is SCoT (Akata et al., 2025). Instead of best responding to the full one-step predictive distribution, the agent first forms a *deterministic point prediction* of the opponents' next joint action and then best responds to that point prediction. In general, this is not equivalent to best responding to a mixed belief, so the argument is different from the classical Bayesian-learning-plus-best-response route. Nevertheless, when all players use deterministic point-prediction rules, the true next action along the realized path is pure at every history, and predictive accuracy is enough to make the point prediction eventually correct. This gives eventual stage-game Nash convergence under a different mechanism than myopic PS-BR.

Definition 12 (Social Chain of Thought (SCoT) (Akata et al., 2025))**.** Fix player i. At each history h^t, let

$$q_i^t(\cdot \mid h^t) := f_{-i}^{i,t}(h^t) \in \Delta(A_{-i})$$

denote player i's one-step predictive distribution over opponents' next joint action. Along the histories under consideration, the representative-choice convention from Section 3.3 gives

$$f_{-i}^{i,t}(h^t) = f_{-i}^{i}(h^t).$$

A *SCoT* rule for player i consists of:

1. a deterministic MAP (maximum a posteriori) selector

$$\hat{a}_{-i}^{t}(h^t) \in \arg\max_{a_{-i} \in A_{-i}} q_i^t(a_{-i} \mid h^t);$$

2. a deterministic pure best-response selector

$$b_i : A_{-i} \to A_i \qquad \text{such that} \qquad b_i(a_{-i}) \in \arg\max_{a_i \in A_i} u_i(a_i, a_{-i}) \ \text{ for every } a_{-i} \in A_{-i}.$$

The induced strategy is

$$f_i(h^t) \ := \ \delta_{b_i(\hat{a}_{-i}^{t}(h^t))} \in \Delta(A_i).$$

Thus a SCoT player uses a pure action at every history.

Lemma 5.9 (Deterministic truth implies asymptotic purity and eventual MAP correctness)**.** *Fix player i and suppose player i learns to predict the path of play under f in the sense of Definition 9. Assume that for every history $h \in H$ there exists an action $a_{-i}^{\star}(h) \in A_{-i}$ such that*

$$f_{-i}(h) = \delta_{a_{-i}^{\star}(h)}.$$

Then

$$\mu^f\left(\left\{z : \exists T_i(z) < \infty \text{ s.t. } \forall t \geq T_i(z),\ \hat{a}_{-i}^{t}(h^t(z)) = a_{-i}^{\star}(h^t(z))\right\}\right) = 1.$$

In particular, along μ^f-almost every realized path z,

$$q_i^t(a_{-i}^{\star}(h^t(z)) \mid h^t(z)) \longrightarrow 1 \qquad \textit{and} \qquad 1 - \max_{a_{-i} \in A_{-i}} q_i^t(a_{-i} \mid h^t(z)) \longrightarrow 0.$$

Theorem 5.10 (One-shot stage-game Nash convergence for SCoT)**.** *Suppose every player $i \in I$ uses SCoT in the sense of Definition 12, and suppose every player learns to predict the path of play under f in the sense of Definition 9. Then*

$$\mu^f\left(\left\{z : \exists T(z) < \infty \text{ s.t. } \forall t \geq T(z),\ f(h^t(z)) \textit{ is a stage Nash equilibrium}\right\}\right) = 1.$$

Equivalently, along μ^f-almost every realized path, the current mixed-action profile eventually becomes a stage 0-Nash equilibrium.

Corollary 5.11 (Bayesian stage-game Nash convergence for SCoT)**.** *Suppose every player uses deterministic MAP-SCoT and Assumption 2 holds for every player. Then the conclusion of Theorem 5.10 holds:*

$$\mu^f\Big(\Big\{z : \exists T(z) < \infty \text{ s.t. } \forall t \geq T(z),\ f(h^t(z)) \text{ is a stage Nash equilibrium}\Big\}\Big) = 1.$$

Remark 2. Theorem 5.10 relies on the fact that when *all players* use SCoT with deterministic tie-breaking, the true current action profile is pure at every history. This is why asymptotic purity need not be imposed separately: it is implied by Bayesian one-step predictive accuracy toward a pure truth. If opponents are allowed to play genuinely mixed current actions, this argument breaks down, and additional conditions such as asymptotic purity or BR-invariance are again needed.

The SCoT result is therefore naturally paired with the grain-of-truth assumption (Assumption 2) and the corresponding merging-of-opinions argument, rather than with Assumption 3, whose uniform-positivity requirement is tailored to cautious menu-based posteriors and posterior-sampling rules such as PS-BR.

The proofs are deferred to Appendix B. Taken together, Theorem 5.8 and Theorem 5.10 show that, for the weaker objective of *stage-game* Nash convergence, full continuation planning is not necessary. However, these one-step results are inherently limited to stage-game equilibrium. They do not by themselves recover more demanding *continuation-game* or *history-contingent repeated-game* equilibria, whose incentive structure is sustained by the value of future paths of play. Establishing convergence to those richer repeated-game equilibria requires a procedure, such as PS-BR, that reasons over full continuation strategies rather than only over the next-period action.

6 Extension to unknown, stochastic, and private payoffs

Sections 3–5 assumed that the stage payoff functions $u_i : A \to [0, 1]$ are common knowledge and deterministic. We now drop this assumption and allow each agent to observe only its own privately realized stochastic payoffs.

6.1 Private-payoff repeated game and information histories

Fix the same action sets $(A_i)_{i\in I}$ and discount factors $(\lambda_i)_{i\in I}$ as in Section 3. For each player i, let $\mathcal{R}_i \subseteq \mathbb{R}$ denote the payoff space and let $\nu_i(\mathrm{d}r)$ be a dominating base measure (counting measure in the discrete case, Lebesgue measure in the continuous case).

We assume that the payoff noise family is known. Concretely, for each player i there is a known family of densities

$$\psi_i(r; \mu), \qquad r \in \mathcal{R}_i,\ \mu \in \mathbb{R},$$

where the parameter μ is the mean payoff. The true unknown object is player i's mean payoff matrix

$$u_i : A \to [0, 1].$$

(As usual, any bounded payoff matrix can be affinely normalized into $[0, 1]$ without changing best responses or Nash inequalities.)

At round t, after the public joint action $a^t \in A$ is realized, player i privately observes

$$r_i^t \sim q_i^{u_i}(\cdot \mid a^t), \qquad \text{where } \; q_i^{u_i}(\mathrm{d}r \mid a) := \psi_i(r; u_i(a))\, \nu_i(\mathrm{d}r). \tag{5}$$

Thus the true payoff kernel is determined by the true mean matrix u_i.

In the private-payoff model, actions may depend on both the public history and the player's own private payoff observations. Accordingly, define player i's information history at time t as

$$x_i^t := (h^t, r_i^{1:t-1}) \in X_i^t := H^t \times \mathcal{R}_i^{t-1}, \qquad X_i := \bigcup_{t \geq 1} X_i^t.$$

A strategy for player i in the private-payoff game is a map

$$\sigma_i : X_i \to \Delta(A_i).$$

Let Σ_i denote the set of such strategies and $\Sigma := \prod_{i \in I} \Sigma_i$.

The full sample space is

$$\Omega := \prod_{t \geq 1} \Big(A \times \prod_{i \in I} \mathcal{R}_i \Big),$$

whose typical element is

$$\omega = (a^1, r^1, a^2, r^2, \ldots), \qquad r^t = (r_1^t, \ldots, r_N^t).$$

Given a strategy profile $\sigma \in \Sigma$ and the true mean matrices $u = (u_i)_{i \in I}$, the tuple (σ, u) induces a unique probability law $P^{\sigma,u}$ on Ω by the Ionescu–Tulcea theorem.

For a realized path $\omega \in \Omega$, write

$$x^t(\omega) := (x_i^t(\omega))_{i \in I}$$

for the realized vector of information histories at time t. For any continuation profile τ defined on future information histories extending x^t, let $P_{x^t}^{\tau,u}$ denote the induced continuation law.

For player i, define the continuation payoff after x^t by

$$U_i(\tau \mid x^t) := \mathbb{E}_{P^{\tau,u}_{x^t}}\left[(1-\lambda_i)\sum_{k=0}^{\infty}\lambda_i^k r_i^{t+k}\right].$$

By iterated expectation and (5),

$$U_i(\tau \mid x^t) = \mathbb{E}_{P^{\tau,u}_{x^t}}\left[(1-\lambda_i)\sum_{k=0}^{\infty}\lambda_i^k u_i(a^{t+k})\right].$$

Hence the objective continuation payoff in the private-payoff game equals the discounted payoff induced by the true mean matrix, even though strategies may condition on private payoff realizations.

A continuation profile τ is an ε-Nash equilibrium after x^t if, for every $i \in I$,

$$U_i(\tau \mid x^t) \geq \sup_{\tau_i' \in \Sigma_i(x_i^t)} U_i(\tau_i', \tau_{-i} \mid x^t) - \varepsilon.$$

Finally, let $\bar{\mu}^{\tau,u}_{x^t}$ denote the public-action marginal of $P^{\tau,u}_{x^t}$ on the future public action path $(a^t, a^{t+1}, \ldots) \in H^\infty$. We compare continuation profiles only through these public-action marginals, using

$$d_{x^t}(\tau, \hat{\tau}) := d\Big(\bar{\mu}^{\tau,u}_{x^t}, \bar{\mu}^{\hat{\tau},u}_{x^t}\Big),$$

where d is the weak distance from Definition 6.

6.2 Known-noise, unknown-mean parametrization

We now impose the finite-menu structure used by PS-BR. For player i, let $\mathcal{M}_i$ be a finite menu of candidate mean payoff matrices

$$m_i : A \to [0,1].$$

Each $m_i \in \mathcal{M}_i$ induces a payoff kernel

$$q_i^{m_i}(\mathrm{d}r \mid a) := \psi_i(r; m_i(a))\,\nu_i(\mathrm{d}r).$$

Thus sampling a payoff matrix label is exactly sampling a payoff kernel, expressed in mean-matrix coordinates.

Given $x_i^t = (h^t, r_i^{1:t-1})$, player i's posterior over candidate mean matrices is

$$\pi_i^t(m_i \mid x_i^t) \propto \pi_i^0(m_i)\prod_{s=1}^{t-1}\psi_i(r_i^s; m_i(a^s)), \qquad m_i \in \mathcal{M}_i. \tag{6}$$

As in Sections 4–5, we model player i's beliefs about the opponents through a finite menu of public-action

continuation models

$$g_{-i} : H \to \Delta(A_{-i}).$$

These models describe the predictive law of opponents' next public action conditional on public history. Let $\mathcal{S}_{-i}$ denote the finite menu and let

$$\mu_i^t(\cdot \mid h^t)$$

be player i's posterior over $\mathcal{S}_{-i}$.

6.3 Subjective continuation values and PS-BR

Fix player i, an information history $x_i^t = (h^t, r_i^{1:t-1})$, a reduced-form opponents' continuation model $g_{-i} \in \mathcal{S}_{-i}$, and a continuation strategy $\tau_i \in \Sigma_i(x_i^t)$.

Let

$$P_{x_i^t}^{(\tau_i, g_{-i}), m_i}$$

denote the induced law on player i's future observable sequence when: (i) player i follows τ_i, (ii) opponents' public actions are generated by g_{-i}, and (iii) player i's future private payoffs are generated from the kernel $q_i^{m_i}$.

Define the m_i-subjective continuation value by

$$V_i^{m_i}(\tau_i \mid x_i^t; g_{-i}) := \mathbb{E}_{P_{x_i^t}^{(\tau_i, g_{-i}), m_i}} \left[(1 - \lambda_i) \sum_{k=0}^{\infty} \lambda_i^k r_i^{t+k} \right]. \tag{7}$$

For $\varepsilon \geq 0$, define

$$\mathrm{BR}_{i,m_i}^{\varepsilon}(g_{-i} \mid x_i^t) := \left\{ \tau_i \in \Sigma_i(x_i^t) : V_i^{m_i}(\tau_i \mid x_i^t; g_{-i}) \geq \sup_{\tau_i' \in \Sigma_i(x_i^t)} V_i^{m_i}(\tau_i' \mid x_i^t; g_{-i}) - \varepsilon \right\},$$

and write

$$\mathrm{BR}_{i,m_i}(g_{-i} \mid x_i^t) := \mathrm{BR}_{i,m_i}^{0}(g_{-i} \mid x_i^t).$$

Player i's mixed subjective continuation value is

$$V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) := \mathbb{E}_{\substack{g_{-i} \sim \mu_i^t(\cdot \mid h^t) \\ m_i \sim \pi_i^t(\cdot \mid x_i^t)}} \left[V_i^{m_i}(\tau_i \mid x_i^t; g_{-i}) \right]. \tag{8}$$

For the true mean matrix u_i, define

$$V_i^{u_i,t}(\tau_i \mid x_i^t) := \mathbb{E}_{g_{-i} \sim \mu_i^t(\cdot \mid h^t)} \left[V_i^{u_i}(\tau_i \mid x_i^t; g_{-i}) \right]. \tag{9}$$

Fix player i and an information history $x_i^t = (h^t, r_i^{1:t-1})$. The posterior $\mu_i^t(\cdot \mid h^t)$ over the finite menu

$\mathcal{S}_{-i}$ induces a posterior predictive law over future public action paths. Let $g_{-i}^{i,t}$ denote any reduced-form behavioral representative of this posterior predictive continuation law. Concretely, $g_{-i}^{i,t}$ is chosen so that for every continuation strategy $\tau_i \in \Sigma_i(x_i^t)$,

$$V_i^{u_i,t}(\tau_i \mid x_i^t) = V_i^{u_i}(\tau_i \mid x_i^t; g_{-i}^{i,t}). \tag{10}$$

When $\mathcal{S}_{-i} = \{g_{-i}^1, \ldots, g_{-i}^K\}$ is finite, one convenient choice is

$$g_{-i}^{i,t}(h)(a_{-i}) = \sum_{k=1}^{K} \mu_i^{t,h}(g_{-i}^k)\, g_{-i}^k(h)(a_{-i}), \qquad h \succeq h^t,$$

where $\mu_i^{t,h}$ is the continuation posterior obtained by updating $\mu_i^t(\cdot \mid h^t)$ along the continuation history h.

Let $\bar{\mu}_{x_i^t}^{(\tau_i,g_{-i}),m_i}$ denote the public-action marginal of $P_{x_i^t}^{(\tau_i,g_{-i}),m_i}$ on $(a^t, a^{t+1}, \ldots) \in H^\infty$. For the actual continuation strategy σ_i, player i's posterior predictive law over future public action paths can then be written as

$$\Pi_i^t(\cdot \mid x_i^t) = \sum_{m_i \in \mathcal{M}_i} \pi_i^t(m_i \mid x_i^t)\, \bar{\mu}_{x_i^t}^{(\sigma_i,g_{-i}^{i,t}),m_i}. \tag{11}$$

We can now state the private-payoff PS-BR rule.

Definition 13 (Posterior-sampling best response (PS-BR) with private payoffs)**.** Fix player i and an information history $x_i^t = (h^t, r_i^{1:t-1})$. Given: (i) the posterior $\mu_i^t(\cdot \mid h^t)$ over reduced-form opponents' continuation models, and (ii) the posterior $\pi_i^t(\cdot \mid x_i^t)$ over player i's own mean payoff matrices, PS-BR chooses a continuation strategy by:

1. sample an opponents' continuation model $\tilde{g}_{-i} \sim \mu_i^t(\cdot \mid h^t)$;
2. sample a mean payoff matrix $\tilde{m}_i \sim \pi_i^t(\cdot \mid x_i^t)$;
3. play any continuation strategy $\tau_i \in \mathrm{BR}_{i,\tilde{m}_i}(\tilde{g}_{-i} \mid x_i^t)$.

Denote the resulting randomized continuation strategy by $\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid x_i^t)$.

6.4 Posterior concentration

Although the primitive strategy profile is $\sigma \in \Sigma$, the public action path it induces admits a reduced-form description. For each player i, define

$$\bar{f}_i(h) := P^{\sigma,u}(a_i^t \in \cdot \mid h^t = h), \qquad \bar{f} := (\bar{f}_i)_{i \in I},$$

and let $\bar{\mu}^{\sigma,u}$ denote the induced law on the public action path in H^∞. Thus $\bar{f}$ is the true reduced-form public-action model generated by the information-history strategy profile σ and the true mean matrices u.

For player i's finite menu of reduced-form opponents' continuation models $\mathcal{S}_{-i}$, assume that Assumption 3 holds mutatis mutandis with the true reduced-form opponent model $\bar{f}_{-i}$ and the true public-action path law $\bar{\mu}^{\sigma,u}$ in place of f_{-i} and μ^f.

Lemma 6.1 (Posterior concentration of reduced-form public-action beliefs)**.** *Fix player i and suppose player i's finite menu $\mathcal{S}_{-i}$ and posterior $\mu_i^t(\cdot \mid h^t)$ satisfy Assumption 3 mutatis mutandis with $\bar{f}_{-i}$ and $\bar{\mu}^{\sigma,u}$ in place of f_{-i} and μ^f. Then under the true interaction law $P^{\sigma,u}$,*

$$\mu_i^t(\bar{f}_{-i} \mid h^t) \longrightarrow 1 \qquad \textit{and hence} \qquad \max_{g_{-i} \in \mathcal{S}_{-i} \setminus \{\bar{f}_{-i}\}} \mu_i^t(g_{-i} \mid h^t) \longrightarrow 0,$$

almost surely.

The only genuinely new learnability requirement in the private-payoff extension is on the payoff side: identifiability of player i's own mean payoff matrix from private noisy rewards.

Assumption 4 (Finite payoff-menu identifiability under known noise)**.** Fix player i and let $\mathcal{M}_i = \operatorname{supp}(\pi_i^0)$ be finite. Assume:

1. *(Menu grain of truth)* The true mean matrix $u_i \in \mathcal{M}_i$ and $\pi_i^0(u_i) > 0$.

2. *(Known common noise family)* Each menu element $m_i \in \mathcal{M}_i$ induces the payoff kernel

$$q_i^{m_i}(\mathrm{d}r \mid a) = \psi_i(r; m_i(a))\, \nu_i(\mathrm{d}r),$$

   and the true payoff law is $q_i^{u_i}$.

3. *(Finite second moments of log-likelihood ratios)* For every $m_i \in \mathcal{M}_i \setminus \{u_i\}$,

$$\sup_{a \in A} \mathbb{E}_{R \sim q_i^{u_i}(\cdot \mid a)} \left[\left(\log \frac{\psi_i(R; u_i(a))}{\psi_i(R; m_i(a))} \right)^2 \right] < \infty.$$

4. *(On-path KL separation)* For every $m_i \in \mathcal{M}_i \setminus \{u_i\}$ there exists $\kappa_i(m_i) > 0$ such that under the true interaction law $P^{\sigma,u}$,

$$\liminf_{T \to \infty} \frac{1}{T} \sum_{t=1}^{T} D_{\mathrm{KL}}\Big(q_i^{u_i}(\cdot \mid a^t) \,\Big\|\, q_i^{m_i}(\cdot \mid a^t) \Big) \geq \kappa_i(m_i) \qquad \text{a.s.}$$

The next lemma is the mean-matrix analogue of Lemma 4.2.

Lemma 6.2 (Payoff posterior concentration under known-noise KL separation)**.** *Fix player i and suppose*

Assumption 4 holds. Then under the true interaction law $P^{\sigma,u}$,

$$\pi_i^t(u_i \mid x_i^t) \longrightarrow 1, \qquad \textit{and hence} \qquad \max_{m_i \in \mathcal{M}_i \setminus \{u_i\}} \pi_i^t(m_i \mid x_i^t) \longrightarrow 0,$$

almost surely.

Lemma 6.3 (Payoff concentration identifies the predictive public-action law)**.** *Fix player i. For every information history x_i^t,*

$$d\Big(\Pi_i^t(\cdot \mid x_i^t), \bar{\mu}_{x_i^t}^{(\sigma_i, g_{-i}^{i,t}), u_i}\Big) \le 1 - \pi_i^t(u_i \mid x_i^t).$$

Consequently, under Lemma 6.2,

$$d\Big(\Pi_i^t(\cdot \mid x_i^t), \bar{\mu}_{x_i^t}^{(\sigma_i, g_{-i}^{i,t}), u_i}\Big) \longrightarrow 0 \qquad \textit{under } P^{\sigma,u} \textit{ a.s.}$$

The proof is deferred to Appendix B.

6.5 PS-BR gap and asymptotic consistency

Let

$$p_t(g_{-i}, m_i) := \mu_i^t(g_{-i} \mid h^t)\, \pi_i^t(m_i \mid x_i^t), \qquad (g_{-i}, m_i) \in \mathcal{S}_{-i} \times \mathcal{M}_i.$$

Define the joint collision complement

$$D_i^{t,\mathrm{joint}}(x_i^t) := 1 - \sum_{(g_{-i}, m_i) \in \mathcal{S}_{-i} \times \mathcal{M}_i} p_t(g_{-i}, m_i)^2.$$

Lemma 6.4 (PS-BR is a $D_i^{t,\mathrm{joint}}$-best response to the mixed subjective value)**.** *Fix player i and an information history $x_i^t = (h^t, r_i^{1:t-1})$. Let $\sigma_{i,t}^{\mathrm{PS}}$ be PS-BR from Definition 13. Then*

$$V_i^{\mathrm{mix},t}(\sigma_{i,t}^{\mathrm{PS}} \mid x_i^t) \ge \sup_{\tau_i \in \Sigma_i(x_i^t)} V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) - D_i^{t,\mathrm{joint}}(x_i^t).$$

Equivalently, $\sigma_{i,t}^{\mathrm{PS}}$ is a $D_i^{t,\mathrm{joint}}(x_i^t)$-best response to the mixed subjective continuation value (8)*.*

Define

$$\delta_i^t(x_i^t) := 1 - \pi_i^t(u_i \mid x_i^t).$$

Because continuation values are normalized to lie in $[0,1]$, for every $\tau_i \in \Sigma_i(x_i^t)$,

$$\big|V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) - V_i^{u_i,t}(\tau_i \mid x_i^t)\big| \le \delta_i^t(x_i^t). \tag{12}$$

Combining (12), Lemma 6.4, Lemma 6.1, and Lemma 6.2 yields the asymptotic best-response property.

Proposition 6.5 (PS-BR implies asymptotic ε-consistency in the private-payoff game)**.** *Fix player i. Assume: (i) Assumption 3 holds mutatis mutandis for player i's menu of reduced-form opponents' continuation models, with the true reduced-form opponent model $\bar{f}_{-i}$ and the true public-action path law $\bar{\mu}^{\sigma,u}$ in place of f_{-i} and μ^f, (ii) Assumption 4 holds for player i's mean-matrix menu, and (iii) player i uses PS-BR at every information history. Then for every $\varepsilon > 0$,*

$$P^{\sigma,u}\Big(\Big\{\omega : \exists\, T_i(\omega,\varepsilon) < \infty \text{ s.t. } \forall t \geq T_i(\omega,\varepsilon),\ \sigma_{i,t}^{\mathrm{PS}}(\cdot \mid x_i^t(\omega)) \in \mathrm{BR}_{i,u_i}^{\varepsilon}\big(g_{-i}^{i,t} \mid x_i^t(\omega)\big)\Big\}\Big) = 1.$$

6.6 Zero-shot Nash convergence with private payoffs

To lift the earlier zero-shot argument, one replaces public histories h^t by information-history vectors x^t, and one compares continuation profiles through the weak distance between their induced public-action marginals after the realized full information-history vector. Because player i only observes $x_i^t = (h^t, r_i^{1:t-1})$, the relevant Bayesian merging step is first stated on player i's observable process. Assumption 6 then identifies this player-relative predictive target with the ex post public continuation law after x^t asymptotically.

For player i, let

$$O_i := \prod_{t\geq 1}(A \times \mathcal{R}_i)$$

denote the space of observable sequences

$$(a^1, r_i^1, a^2, r_i^2, \ldots).$$

Let $P_i^{\sigma,u}$ be the marginal of $P^{\sigma,u}$ on O_i, and let Q_i^{0,σ_i} be player i's prior predictive law on O_i induced by their priors over $\mathcal{S}_{-i}$ and $\mathcal{M}_i$, the known noise family, and their own strategy σ_i.

Let

$$\bar{\mu}_{i,x_i^t}^{\sigma,u}(E) := P^{\sigma,u}\big((a^t, a^{t+1}, \ldots) \in E \mid x_i^t\big), \qquad E \in \mathcal{B},$$

denote the true public-action continuation law conditional on player i's own observable information history x_i^t. Also let

$$\Pi_i^t(\cdot \mid x_i^t)$$

denote player i's posterior predictive law over the future public action path $(a^t, a^{t+1}, \ldots) \in H^\infty$ conditional on x_i^t.

In the private-payoff setup, player i's prior over reduced-form opponents' continuation models and over its own finite menu of payoff hypotheses is constructed so that the true observable process is represented as one feasible element. Thus the induced prior predictive law on player i's observable sequence should place positive mass on the true observable path law. This naturally gives the following Assumption 5.

Assumption 5 (Observable grain of truth in the private-payoff game)**.** Fix player i. Assume

$$P_i^{\sigma,u} \ll Q_i^{0,\sigma_i}.$$

The next requirement is also natural in the PS-BR regime. Although player i never observes the opponents' private reward histories, those histories matter for future public play only through how they shape the opponents' own continuation behavior. As each player's private payoff posterior concentrates and the residual effect of these hidden reward histories on public continuation play becomes negligible, conditioning on the realized full information-history vector x^t or on player i's own observable history x_i^t should asymptotically yield the same public-action continuation law. Assumption 6 formalizes the intended information structure: player i does not observe the other players' private reward histories and need only infer its own payoff matrix together with the opponents' reduced-form public-action strategy. Asymptotically, any additional predictive content in the unobserved private histories becomes negligible for future public play.

Assumption 6 (Asymptotic public sufficiency of hidden private histories)**.** For every player i,

$$d\Big(\bar{\mu}^{\sigma,u}_{i,x_i^t(\omega)}, \bar{\mu}^{\sigma,u}_{x^t(\omega)}\Big) \longrightarrow 0 \qquad \text{for } P^{\sigma,u}\text{-a.e. } \omega.$$

Assumption 6 is the formal expression of the idea that, in the intended regime, each player needs to infer only its own payoff matrix and the opponents' reduced-form public-action strategy; the opponents' unrevealed private reward histories do not asymptotically alter future public play beyond what those objects already encode.

Lemma 6.6 (Observable grain of truth implies strong public-path prediction)**.** *Fix player i. Under Assumptions 5 and 6, player i's posterior predictive law over future public action paths merges with the true public-action continuation law after the realized information-history vector:*

$$d\Big(\Pi_i^t(\cdot \mid x_i^t(\omega)), \bar{\mu}^{\sigma,u}_{x^t(\omega)}\Big) \longrightarrow 0 \qquad \textit{for } P^{\sigma,u}\textit{-a.e. } \omega.$$

The proof is deferred to Appendix B.

Definition 14 (Weak subjective equilibrium on information histories)**.** Fix $\xi, \eta \geq 0$ and an information-history vector x^t. A continuation profile τ is a weak ξ-subjective η-equilibrium after x^t if, for every player i, there exists a reduced-form opponents' continuation model g^i_{-i} such that

$$\tau_i \in \mathrm{BR}^{\xi}_{i,u_i}(g^i_{-i} \mid x_i^t)$$

and

$$d\Big(\bar{\mu}^{\tau,u}_{x^t}, \bar{\mu}^{(\tau_i,g^i_{-i}),u_i}_{x_i^t}\Big) \leq \eta.$$

Proposition 6.7 (Learning and asymptotic consistency imply weak subjective equilibrium in the private–payoff game)**.** *Suppose every player i satisfies the conclusion of Proposition 6.5 and of Lemma 6.6. Then for every $\xi > 0$ and $\eta > 0$,*

$$P^{\sigma,u}\Big(\Big\{\omega : \exists\, T(\omega) < \infty \text{ s.t. } \forall t \geq T(\omega),\ \sigma\big|_{x^t(\omega)} \text{ is a weak } \xi\text{-subjective } \eta\text{-equilibrium after } x^t(\omega)\Big\}\Big) = 1.$$

The proof is deferred to Appendix B.

Theorem 6.8 (Zero-shot Nash convergence with private payoffs)**.** *Assume that for every player i, Assumption 3 holds mutatis mutandis for the finite menu of reduced-form opponents' continuation models, with the true reduced-form opponent model $\bar{f}_{-i}$ and the true public-action path law $\bar{\mu}^{\sigma,u}$ in place of f_{-i} and μ^f, Assumption 4 holds for the finite menu of candidate mean payoff matrices under the known noise family, Assumptions 5 and 6 hold, and player i uses PS-BR at every information history. Then for every $\varepsilon > 0$,*

$$P^{\sigma,u}\Big(\{\omega : \exists\, T(\omega) < \infty \text{ s.t. } \forall t \geq T(\omega), \exists\, \hat{\tau}^{\varepsilon,t,\omega} \text{ an } \varepsilon\text{-Nash equilibrium of the continuation game after } x^t(\omega) \text{ with } d\Big(\bar{\mu}^{\sigma,u}_{x^t(\omega)}, \bar{\mu}^{\hat{\tau}^{\varepsilon,t,\omega},u}_{x^t(\omega)}\Big) \leq \varepsilon\}\Big) = 1.$$

Theorem 6.8's interpretation is similar to Theorem 5.3, but now under the additional Assumption 6: although agents do not know the payoff matrix ex ante and observe only noisy private rewards, their public continuation play eventually becomes weakly close, along the realized path, to an ε-Nash equilibrium of the continuation game. In the known common noise family setting, implementing payoff-kernel sampling is equivalent to sampling a mean payoff matrix from a finite reward menu and evaluating continuation strategies against the induced kernel.

7 Experiments

In this section, we empirically evaluate whether off-the-shelf reasoning LLM agents exhibit the theoretical properties derived in previous sections, i.e., whether they converge toward Nash equilibrium behavior in repeated strategic interaction. After discussing the experiment setup that is common throughout all experiments in Section 7.1, we provide simulation experimentation results that test the following three hypotheses implied by our theoretical analysis:

1. For convergence to the stage-game (myopic) Nash equilibrium, simple predict–then–act reasoning, e.g., SCoT, should already be sufficient (Section 7.2).

2. For convergence to non-trivial repeated-game Nash equilibria that rely on continuation incentives and long-horizon strategic reasoning, myopic approaches should generally fail, whereas PS-BR, which explicitly evaluates continuation strategies, should succeed (Section 7.3).

3. PS-BR should remain effective even when the payoff matrix is not given and must be learned from noisy

payoff observations, recovering equilibrium behavior under payoff uncertainty (Section 7.4).

7.1 Setup

Baselines. We use Qwen 3.5-27B (Qwen Team, 2026), a small-scale open-reasoning model with GPT-5-mini level capabilities (A. Singh et al., 2025). Specifically, we run three models, with almost the same prompts except the reasoning patterns:

- *Base*: Qwen 3.5-27B with direct action selection from rules + interaction history.
- *SCoT*: Qwen 3.5-27B with chain-of-thought style "predict-then-act" prompting (Akata et al., 2025).It has demonstrated success in some repeated games, such as the Battle of the Sexes, and can be considered a simplified, myopic version of PS-BR. For details, see Appendix E.
- *PS-BR*: Qwen 3.5-27B with PS-BR (Definition 5, also detailed in Appendix D).

Benchmarks. We consider five repeated-game environments in total: BoS, PD, Promo, Samaritan, and Lemons.

(1) Battle of the Sexes (BoS; coordination with asymmetric equilibria). Actions each period: J or F. Per-period payoff matrix (Player 1, Player 2):

	P2: J	P2: F
P1: J	$(10, 7)$	$(0, 0)$
P1: F	$(0, 0)$	$(7, 10)$

The non-trivial cooperative Nash equilibrium (pure): (J, J) and (F, F). One non-trivial cooperative Nash equilibrium is both of them sticking to one action:

- Play J after every history (outcome (J, J) every period).
- Play F after every history (outcome (F, F) every period).

Such a non-trivial cooperative Nash equilibrium is particularly plausible when a monetary transfer underlies the game. Another non-trivial cooperative Nash equilibrium is turn-taking:

- Play (J, J) in odd periods and (F, F) in even periods.
- After any history, continue the same odd/even phase convention.

(2) Prisoner's Dilemma (PD; social dilemma). Actions each period: J or F. Per-period payoff matrix (Player 1, Player 2):

	P2: J	P2: F
P1: J	$(3, 3)$	$(-5, 5)$
P1: F	$(5, -5)$	$(0, 0)$

One-shot stage-game Nash equilibrium: (F, F). A baseline pure Nash equilibrium of the repeated game is stationary play of (F, F) after every history. A nontrivial cooperative Nash equilibrium (grim-trigger cooperation) is:

- Cooperative phase: play (J, J) every period.
- If any player ever plays F, switch forever to (F, F).

(3) Promo (Lal, 1990, Appendix H.1) Actions each period: R (Regular), P (Promotion), or Z (price-war punishment). Per-period payoff matrix (Player 1, Player 2):

	P2: R	P2: P	P2: Z
P1: R	$(1, 1)$	$(-1, 4)$	$(-2, -2)$
P1: P	$(4, -1)$	$(0, 0)$	$(-2, -2)$
P1: Z	$(-2, -2)$	$(-2, -2)$	$(-2, -2)$

One-shot stage-game Nash equilibrium (pure): (P, P). A baseline pure Nash equilibrium of the repeated game is the stationary play of (P, P) after every history. A nontrivial cooperative pure Nash equilibrium described in Lal, 1990 is:

- Cooperative phase: (P, R) in the odd round, and (R, P) in the even round.
- If the opponent deviates from the cooperation, play Z for two periods and revert to the cooperative phase.

(4) Samaritan (altruism / one-sided moral hazard). Player 1 (Helper): Help (H) or No-help (N). Player 2 (Recipient): Work (W) or Shirk (S). Per-period payoff matrix (Helper, Recipient):

	Recipient: W	Recipient: S
Helper: H	$(2, -1)$	$(0, 0)$
Helper: N	$(1, -2)$	$(-1, -3)$

One-shot stage-game Nash equilibrium (pure): (H, S). The helper has a dominant action (help), and the recipient best responds by shirking. A nontrivial cooperative Nash equilibrium exists for sufficiently patient players:

- Cooperative phase: play (H, W) every period.
- If the recipient ever shirks, switch forever to punishment (N, W).
- If, during punishment, the helper ever deviates by helping, the recipient switches forever to (H, S) behavior.

(5) Lemons (adverse selection). Player 1 (Seller): High Quality (HQ) or Low Quality (LQ). Player 2 (Buyer): Buy (B) or Don't buy (D). Per-period payoff matrix (Seller, Buyer):

	Buyer: B	Buyer: D
Seller: HQ	$(3, 3)$	$(-1, 0)$
Seller: LQ	$(4, -1)$	$(0, 0)$

One-shot stage-game Nash equilibrium (pure): (LQ, D). Seller has strict dominant action LQ; buyer best-responds to LQ with D. A baseline pure Nash equilibrium of the repeated game is the stationary play of (LQ, D) after every history. A nontrivial cooperative Nash equilibrium for sufficiently patient players:

- Start by playing (HQ, B), and continue (HQ, B) as long as no low-quality sale has ever been observed.
- If the buyer ever buys and then observes LQ, switch forever to D; seller then plays dominant LQ thereafter.

7.2 Experiment 1. Nash convergence

Here, we test the first hypothesis: for convergence to any Nash equilibrium, simple predict–then–act reasoning, e.g., SCoT (Akata et al., 2025), should already suffice.

7.2.1 Experiment design

In Section 5.3, we showed that if agents myopically learn to predict opponents' next actions and then best respond to those predictions, the realized play path eventually converges to a stage-game ε-Nash equilibrium. SCoT (Akata et al., 2025) operationalizes precisely such a predict–then–act rule, making it a natural empirical test of the theory.

To evaluate this prediction, we simulate repeated interaction in each benchmark game described in Section 7.1. Two identical copies of the same model interact in symmetric self-play for $T = 200$ rounds with perfect monitoring of actions and payoffs. No communication channel is available beyond the public history of previous actions and realized payoffs. Each model conditions its round-t decision only on the observed interaction history up to round $t - 1$.

To measure this equilibrium-action convergence, among the $1, \ldots, 200$ rounds, we only focus on the late-round window $t \in \{161, \ldots, 180\}$. For each round in this window, we checked the percentage of both players' realized actions that match *any* Nash equilibrium action, i.e., Nash equilibrium action of the underlying one-shot game *or* an on-path action of the cooperative repeated-game equilibrium described in Section 7.1. We then average these indicators over rounds 161–180 and report the resulting percentage. Thus, the reported number can be interpreted as the fraction of late-round play that lies on either a one-shot Nash path or a cooperative-equilibrium path. Using rounds 161–180 isolates steady-state behavior and avoids placing weight on transient early-round dynamics and terminal-horizon effects. For each of the three model configurations (*Base*, *SCoT*, and *PS-BR*), we run 20 independent such self-play matches. Our primary outcome of interest is whether the realized joint action profile converges to either a one-shot Nash action or

an on-path action of the benchmark cooperative repeated-game Nash equilibrium for that game.

7.2.2 Results

Table 1: Equilibrium-follow percentage in late rounds (rounds 161–180) for any (one-shot Nash or cooperative on-path action) Nash equilibrium. Reported scores are averaged over 20 trials.

Game	Base	SCoT	PS-BR
BoS	60.0%	**100.0**%	**100.0**%
PD	60.0%	**100.0**%	87.8%
Promo	0.0%	**100.0**%	**100.0**%
Samaritan	64.5%	**100.0**%	97.2%
Lemons	0.0%	**100.0**%	89.8%

Table 1 shows that once cooperative on-path actions are also credited, *SCoT* attains a perfect late-round equilibrium-action score in all five benchmark environments. *Base*, by contrast, remains uneven across games, reaching 60.0% in BoS, 60.0% in PD, and 64.5% in Samaritan, but 0.0% in both Promo and Lemons. *PS-BR* also performs strongly, scoring 100.0% in BoS and Promo and rising to 87.8% in PD, 97.2% in Samaritan, and 89.8% in Lemons when cooperative on-path actions are credited. Overall, these results show that myopic predict–then–act prompting often steers play to some Nash equilibrium.

A natural question is what kind of equilibrium convergence Table 1 is capturing. The theory in Section 5.3 predicts that myopic predict–then–act reasoning should be sufficient for convergence to a *stage-game* ε-Nash equilibrium, without requiring agents to reason over full continuation strategies. The empirical results are broadly consistent with this prediction. In particular, *SCoT* attains perfect equilibrium-follow scores in all five environments once the evaluation metric credits both one-shot Nash actions and on-path actions of cooperative repeated-game equilibria. This suggests that explicitly prompting the model to forecast the opponent's next move and then act accordingly is often enough to remove obviously non-equilibrium play in the late rounds.

At the same time, the results should be interpreted carefully. The metric in Table 1 deliberately aggregates two qualitatively different notions of equilibrium-consistent behavior: one-shot Nash actions and actions that lie on the path of a benchmark cooperative repeated-game equilibrium. As a result, a high score means that play has moved onto *some* equilibrium-consistent path, but it does not tell us which kind of equilibrium has been selected. For example, in Prisoner's Dilemma, both (F, F) and (J, J) can be counted as successful late-round outcomes under our metric, even though the former reflects myopic defection while the latter reflects cooperation sustained by continuation incentives. Likewise, in BoS, converging to either coordinated outcome counts as success even though equilibrium selection remains unresolved.

This distinction is important because myopic reasoning can explain only a limited class of equilibrium phenomena. A one-step predict–then–act rule can stabilize play at actions that are locally optimal given beliefs about the opponent's next move, but it does not by itself reason over future punishment and reward

paths. Consequently, strong performance in Table 1 should be read as evidence that myopic prompting is often sufficient for *equilibrium action convergence*, not as evidence that it can reliably implement a particular nontrivial repeated-game equilibrium. In other words, SCoT appears effective at steering play toward some equilibrium-consistent late-round behavior, but the table does not yet establish whether it can sustain the richer, history-contingent equilibria that depend on long-horizon continuation values.

This limitation is exactly what motivates the next experiment. To distinguish simple equilibrium-action convergence from genuine repeated-game strategic reasoning, we now test whether the models can follow a *specific* nontrivial cooperative Nash equilibrium path when that path must be sustained by continuation incentives rather than by myopic one-step optimization alone.

7.3 Experiment 2. Nontrivial Nash convergence

We now move from asking whether play converges to *some* equilibrium-consistent action profile to the harder question of whether agents can track a nontrivial, cooperative repeated-game Nash equilibrium sustained by continuation incentives. Here, we test the second hypothesis: for convergence to non-trivial repeated-game Nash equilibria that rely on continuation incentives and long-horizon strategic reasoning, myopic approaches should generally fail, whereas PS-BR, which explicitly evaluates continuation strategies, should succeed.

7.3.1 Experiment design

To verify whether a particular long-horizon cooperative Nash equilibrium can be implemented, we included a prompt for each agent that specifies a particular long-horizon non-trivial cooperative Nash equilibrium and asks them to "strongly expect the opponent to play" the strategy. Such prompting sets the initial point of the evolution of their beliefs. For example, in PD, this meant prompting both agents to expect the opponent to play a continued grim-trigger strategy, i.e., cooperation until a defection triggers permanent punishment. On the other hand, in Promo, it meant prompting both agents to expect the prescribed alternating cooperative pattern $(P, R), (R, P), (P, R), \ldots$, until a defection triggers finite punishment.

As before, all experiments use symmetric self-play with two copies of the same model under perfect monitoring. Each match lasts $T = 200$ rounds. In every round, players act simultaneously, observe both actions and realized payoffs, and then condition the next-round decision on the updated history.

Again, for each round $t \in \{161, \ldots, 180\}$ in each run, we checked if both players' realized actions match the desired nontrivial cooperative equilibrium behavior in terms of percentage, then averaged the percentages over the 20 rounds (161-180) and reported the mean by setting and game. (We chose round 180 as the endpoint since PS-BR uses 20 rounds of lookahead, and we excluded pre-161 results, as we want to see the equilibrium outcome.)

7.3.2 Results.

Table 2 shows a sharp separation across methods. *PS-BR* achieves high late-round follow rates in all five environments, reaching 92.5% in BoS, 98.0% in PD, 94.8% in Promo, 93.3% in Samaritan, and 93.5% in

Table 2: Equilibrium-follow percentage in late rounds (rounds 161–180) for the prompt-specified nontrivial cooperative equilibrium. Reported scores are averaged over 20 trials.

Game	Base	SCoT	PS-BR
BoS	0.0%	0.0%	**92.5**%
PD	0.0%	**100.0**%	98.0%
Promo	0.0%	0.0%	**94.8**%
Samaritan	0.0%	0.0%	**93.3**%
Lemons	0.0%	0.0%	**93.5**%

Lemons. Thus, once the cooperative equilibrium is explicitly specified, the non-myopic planner tracks the intended long-horizon path quite reliably across all benchmark games.

By contrast, *Base* remains at 0.0% in every environment. *SCoT* succeeds only in PD, where it reaches 100.0%, and remains at 0.0% in BoS, Promo, Samaritan, and Lemons. Since the three settings use nearly the same game instructions and history context, the main difference is the reasoning/decision strategy (direct action for Base, myopic predict–then–act for SCoT, and posterior-sampling best response with rollout planning for PS-BR). This pattern suggests that direct prompting is insufficient for following contingent cooperative equilibrium prescriptions, while myopic prompting can recover the simple stationary cooperative path in PD but not the richer coordination, punishment, or trust-based prescriptions in the other games. PS-BR's explicit modeling of opponent strategy and continuation value is what enables sustained on-path behavior in late rounds.

The results in Table 2 provide a clear separation between myopic and non-myopic reasoning. Unlike Experiment 1, where multiple equilibrium-consistent outcomes were credited, this experiment sets up initial beliefs so that agents follow a *specific* cooperative equilibrium path that requires non-myopic reasoning. Under this stricter criterion, *PS-BR* consistently achieves high follow rates across all environments, whereas *Base* fails entirely and *SCoT* succeeds only in the simplest case (PD).

This pattern aligns closely with the theoretical distinction developed in Section 5. Implementing a nontrivial repeated-game equilibrium requires reasoning over continuation values: agents must understand that short-term deviations trigger future punishment, and that adherence to the cooperative path is optimal only when these future consequences are taken into account. *PS-BR* explicitly evaluates such continuation strategies through rollout, and therefore can internalize these long-horizon incentives. By contrast, *SCoT* operates on one-step predictions and local best responses, which are insufficient to sustain equilibria that depend on multi-period incentive compatibility.

The one partial exception is Prisoner's Dilemma, where *SCoT* achieves perfect performance. This is consistent with the structure of the grim-trigger equilibrium in PD: the cooperative phase (J, J) is itself a stage-game Pareto-dominant outcome and is locally consistent with mutual best responses under optimistic

beliefs. As a result, myopic reasoning can incidentally align with the cooperative path. In contrast, games such as BoS, Promo, Samaritan, and Lemons require coordination on asymmetric roles, punishment phases, or trust-dependent behavior that cannot be justified purely from one-step optimization, making myopic approaches ineffective.

More broadly, these results indicate that equilibrium *selection* and *path-following* are fundamentally harder than equilibrium *action convergence*. While Experiment 1 shows that simple reasoning can often eliminate non-equilibrium behavior, Experiment 2 demonstrates that sustaining a particular equilibrium—especially one supported by continuation incentives—requires explicit modeling of future play. This provides empirical support for the theoretical claim that the posterior-sampling best response, by operating over full continuation strategies, can implement repeated-game equilibria that lie beyond the reach of myopic predict–then–act rules.

Having established this distinction under known and deterministic payoffs, we next consider a more realistic setting in which agents must simultaneously learn the payoff structure from noisy private observations while engaging in strategic interaction.

7.4 Experiment 3: Nontrivial Nash convergence under unknown payoffs

7.4.1 Setup

We keep the interaction protocol, horizons, and game set from Experiment 1 (Section 7.2) and Experiment 2 (Section 7.3), and modify only the payoff observations: agents no longer receive the payoff matrix in the prompt and instead learn solely from noisy, privately observed payoffs.

For each benchmark game $g \in \{\text{BoS}, \text{PD}, \text{Promo}, \text{Samaritan}, \text{Lemons}\}$, let $u_i^g(a) \in \mathbb{R}$ denote the *deterministic* stage payoff from Experiment 1 for player i and joint action $a \in A$. In Experiment 3, after the public joint action a^t is realized, player i observes a private payoff

$$r_i^t \;=\; u_i^g(a^t) + \epsilon_{i,t}, \qquad \epsilon_{i,t} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma_g^2), \tag{13}$$

independent across players i and rounds t. Players observe the full public action history but *only their own* payoff sequence $(r_i^t)_t$. All equilibrium notions continue to refer to the underlying mean-payoff repeated game induced by u_i^g.

Known common noise family, unknown mean matrix. Experiment 3 instantiates the private-payoff theory in the special case where the reward noise family is known and only the mean payoff matrix is unknown. Concretely, for each player i and joint action a,

$$r_i^t \mid a^t = a \sim \mathcal{N}(m_i(a), \sigma_g^2),$$

where σ_g^2 is common knowledge and the unknown object is the matrix $m_i : A \to \mathbb{R}$. The finite reward menu used by PS-BR is therefore a finite menu of candidate mean matrices. Equivalently, each candidate matrix

m induces a full payoff kernel

$$q_i^m(\cdot \mid a) = \mathcal{N}(m(a), \sigma_g^2),$$

so payoff-matrix sampling in the implementation is exactly payoff-kernel sampling in the theory, expressed in mean-matrix coordinates.

We choose a noise level large enough that, on a single step, the realized payoff can often reverse the ranking between two outcomes whose true mean payoffs differ by the smallest strategically relevant gap. Formally, for each game g, define the minimal nonzero payoff separation

$$\Delta_{\min,g} \;:=\; \min_{i \in I} \; \min\{|u_i^g(a) - u_i^g(a')| : \; a, a' \in A, \; u_i^g(a) \neq u_i^g(a')\}. \tag{14}$$

For the payoff matrices used in Experiment 1, the smallest payoff gaps are $\Delta_{\min,\text{BoS}} = 3$ and $\Delta_{\min,\text{PD}} = 2$, while for Promo, Samaritan, and Lemons the smallest gap is 1.

We set the Gaussian noise standard deviation to

$$\sigma_g \;=\; \Delta_{\min,g}. \tag{15}$$

With additive Gaussian noise, the noisy difference between two outcomes with mean gap Δ has standard deviation $\sqrt{2}\sigma_g$; hence when $\Delta = \Delta_{\min,g}$ and $\sigma_g = \Delta_{\min,g}$, a single observation reverses the sign of the comparison with probability $\Phi(-1/\sqrt{2}) \approx 0.24$. Thus, roughly one in four observations on the tightest gaps is directionally misleading, while averaging over time still reveals the true mean incentives.

Then we repeat the same experiments in Experiment 1 (late-round adherence to the any Nash equilibrium path) and Experiment 2 (late-round adherence to the prompt-specified nontrivial cooperative Nash equilibrium path), using the same scoring window and reporting conventions; the only change is that agents must infer incentives from the private noisy payoffs (13) rather than reading u_i^g from the prompt.

To match Assumption 4, we equip each agent with a finite hypothesis class over the unknown *mean payoff matrix*. Fix a game g and player i, and define the offset set

$$K := \{-2, -1.5, -1, -0.5, 0, +0.5, +1, +1.5, +2\}.$$

The finite menu of candidate mean matrices is

$$\mathcal{M}_{i,g} := \{m : A \to \mathbb{R} : m(a) = u_i^g(a) + k_a \sigma_g \text{ for each } a \in A, \text{ with } k_a \in K\}.$$

In particular, the true mean matrix u_i^g belongs to $\mathcal{M}_{i,g}$ by taking $k_a = 0$ for every joint action a.

Operationally, player i may internally maintain a posterior over $\mathcal{M}_{i,g}$ using the Gaussian likelihood

$$\pi_i^t(m \mid h^t, r_i^{1:t-1}) \propto \pi_i^0(m) \prod_{s=1}^{t-1} \phi(r_i^s; m(a^s), \sigma_g^2),$$

where $\phi(\cdot; \mu, \sigma_g^2)$ is the Gaussian density. PS-BR then samples one candidate mean matrix from this posterior and evaluates continuation strategies against the induced payoff kernel. Because $\mathcal{M}_{i,g}$ has product form over joint actions, this posterior can be updated action-wise under a product prior over offsets $(k_a)_{a \in A}$; one need not enumerate the full menu explicitly in order to sample a complete mean matrix. Accordingly, Experiment 3 should be interpreted as testing strategic learning under noisy private observations of an unknown mean-payoff matrix, rather than learning an arbitrary payoff distribution. The informational difficulty comes from identifying the mean incentives relevant for continuation planning, while the noise family itself is held fixed and known.

7.4.2 Results.

We report two complementary late-round metrics under unknown stochastic payoffs: convergence to any Nash equilibrium action (Table 3) and follow-through on the prompt-specified cooperative Nash equilibrium path (Table 4).

Table 3: Unknown stochastic payoffs: equilibrium-follow percentage in late rounds (rounds 161–180) for *any* Nash equilibrium. Reported scores are averaged over 20 trials.

Game	Base	SCoT	PS-BR
BoS	60.0%	95.0%	**99.8**%
PD	60.0%	**98.0**%	**98.0**%
Promo	0.0%	**100%**	**100.0**%
Samaritan	0.0%	0.0%	**96.2**%
Lemons	0.0%	**98.5**%	82.5%

Table 4: Unknown stochastic payoffs: equilibrium-follow percentage in late rounds (rounds 161–180) for the prompt-specified cooperative Nash equilibrium. Reported scores are averaged over 20 trials.

Game	Base	SCoT	PS-BR
BoS	0%	0%	**98.0**%
PD	0%	0%	**71.2**%
Promo	0%	0%	**71.0**%
Samaritan	5%	0%	**81.0**%
Lemons	0%	0%	**73.8**%

On the broader "any Nash" metric (Table 3), *SCoT* still performs very strongly in BoS (95.0%), PD (98.0%), Promo (100.0%), and Lemons (98.5%), but falls to 0.0% in Samaritan. *PS-BR* is near-perfect in BoS (99.8%), PD (98.0%), and Promo (100.0%), remains strong in Samaritan (96.2%), and reaches 82.5% in Lemons. *Base* remains limited, scoring 60.0% in BoS and PD and 0.0% in Promo, Samaritan, and Lemons.

On the other hand, on stricter prompt-specified cooperative-equilibrium metric (Table 4), *PS-BR* remains the only method with substantial late-round follow-through under unknown payoffs: 98.0% in BoS, 71.2% in PD, 71.0% in Promo, 81.0% in Samaritan, and 73.8% in Lemons. Both *Base* and *SCoT* are at 0.0% in BoS, PD, Promo, and Lemons, while *Base* reaches only 5.0% in Samaritan. These results suggest that under noisy

private payoffs, myopic reasoning is often still enough to reach some equilibrium-like late-round behavior, but not to track the specific long-horizon cooperative prescription; the non-myopic planner, PS-BR, retains a clear advantage when the task requires identifying and sustaining the intended cooperative repeated-game path.

Taken together, Tables 3 and 4 show that payoff uncertainty preserves the basic separation observed in the deterministic-payoff experiments, while also making the task meaningfully harder. On the broader "any Nash" metric, both *SCoT* and *PS-BR* still often reach equilibrium-consistent late-round behavior, indicating that noisy private payoffs do not prevent agents from eventually identifying at least some strategically stable pattern of play. This is consistent with the idea that coarse equilibrium-action convergence can survive substantial observational noise as long as the underlying incentives remain learnable over repeated interaction.

As in Experiment 2, the stricter cooperative-equilibrium metric reveals a much sharper distinction. Again, under unknown payoffs, *PS-BR* remains the only method that reliably tracks the prompt-specified nontrivial repeated-game equilibrium across all environments, whereas *Base* and *SCoT* almost completely fail. This gap is important because it shows that the main difficulty is not merely predicting the opponent's next move, but jointly inferring the payoff structure and reasoning over continuation incentives. To sustain a particular cooperative equilibrium under payoff uncertainty, an agent must learn which action profiles are valuable, which deviations are tempting, and why future punishments make cooperation incentive compatible. *PS-BR* is designed to do exactly this by sampling both opponent strategies and payoff hypotheses and then planning against the sampled continuation game.

The fact that *PS-BR* still performs well, though less perfectly than in the known-payoff case, is also informative. Relative to Table 2, follow rates decline in PD, Promo, Samaritan, and Lemons once payoffs must be learned from noisy private observations. This is the expected direction: payoff uncertainty introduces an additional layer of posterior dispersion, so even when the opponent strategy is inferred correctly, errors in the learned payoff model can still distort continuation-value comparisons. In other words, the unknown-payoff setting does not overturn the mechanism established earlier, but it weakens it quantitatively by making both belief learning and best-response computation noisier.

At the same time, the results suggest that the theoretical extension in Section 6 is empirically meaningful rather than merely formal. The model class that explicitly represents uncertainty over payoffs and updates from private observations retains a substantial advantage precisely in the environments where long-horizon repeated-game incentives matter most. Thus, the experiments support the broader claim of the paper: reasonably reasoning agents need not know the full game in advance to move toward equilibrium-like behavior. What matters is whether they can infer both the strategic behavior of others and the payoff consequences of interaction well enough to approximate continuation best responses on the realized path.

Overall, the three experiments draw a coherent empirical picture. Simple predict–then–act reasoning is

often sufficient for convergence to some stage-game or equilibrium-consistent action pattern. But when the objective is to implement a specific nontrivial repeated-game equilibrium, especially under realistic informational frictions such as unknown and stochastic payoffs, explicit continuation-level reasoning becomes decisive. This is exactly the regime in which *PS-BR* provides a robust advantage, matching the central theoretical message of the paper.

8 Conclusion

In this paper, we theoretically highlight the promising prospect that general-purpose AI agents can attain game-theoretic robustness through inherent reasoning capabilities rather than bespoke training. By demonstrating that LLMs can evolve toward equilibrium behavior on the fly, we take a step toward safer and more autonomous multi-agent AI systems that remain effective across the myriad interactive scenarios they will encounter in the real world. The results bridge the gap between AI agents and classical game theory, indicating that the rich knowledge and inferential power of modern LLMs may be harnessed to meet longstanding challenges in multi-agent learning and interaction. Ultimately, enabling LLM-based agents to naturally exhibit equilibrium-like behavior during play not only advances our theoretical understanding of their behavior but also paves the way for their deployment in societally crucial domains that require reliable strategic decision-making.

References

Gaarlandt, Jur et al. (2025). "AI agents are changing how people shop. Here's what that means for brands." In: *Harvard Business Review* 26, p. 2.

Shahidi, Peyman et al. (2025). *The Coasean Singularity? Demand, Supply, and Market Design with AI Agents*. Tech. rep. National Bureau of Economic Research.

Bianchi, Federico et al. (2024). "How well can llms negotiate? negotiation arena platform and analysis". In: *arXiv preprint arXiv:2402.05863*.

Guo, Taicheng et al. (2024). "Large language model based multi-agents: A survey of progress and challenges". In: *arXiv preprint arXiv:2402.01680*.

Lopez-Lira, Alejandro (2025). "Can large language models trade? testing financial theories with llm agents in market simulations". In: *arXiv preprint arXiv:2504.10789*.

Li, Yang et al. (2024). "Aligning individual and collective objectives in multi-agent cooperation". In: *Advances in Neural Information Processing Systems* 37, pp. 44735–44760.

Zhu, Shenzhe et al. (2025). "The automated but risky game: Modeling and benchmarking agent-to-agent negotiations and transactions in consumer markets". In: *arXiv preprint arXiv:2506.00073*.

Bansal, Gagan et al. (2025). "Magentic Marketplace: An Open-Source Environment for Studying Agentic Markets". In: *arXiv preprint arXiv:2510.25779*.

Abreu, Dilip and Ariel Rubinstein (1988). "The structure of Nash equilibrium in repeated games with finite automata". In: *Econometrica: Journal of the Econometric Society*, pp. 1259–1281.

Dal Bó, Pedro and Guillaume R Fréchette (2011). "The evolution of cooperation in infinitely repeated games: Experimental evidence". In: *American Economic Review* 101.1, pp. 411–429.

– (2018). "On the determinants of cooperation in infinitely repeated games: A survey". In: *Journal of Economic Literature* 56.1, pp. 60–114.

Proto, Eugenio, Aldo Rustichini, and Andis Sofianos (2019). "Intelligence, personality, and gains from cooperation in repeated interactions". In: *Journal of Political Economy* 127.3, pp. 1351–1390.

Aguirregabiria, Victor, Allan Collard-Wexler, and Stephen P Ryan (2021). "Dynamic games in empirical industrial organization". In: *Handbook of industrial organization*. Vol. 4. 1. Elsevier, pp. 225–343.

Calvano, Emilio et al. (2020). "Artificial intelligence, algorithmic pricing, and collusion". In: *American Economic Review* 110.10, pp. 3267–3297.

Fish, Sara, Yannai A Gonczarowski, and Ran I Shorrer (2024). "Algorithmic collusion by large language models". In: *arXiv preprint arXiv:2404.00806* 7.2, p. 5.

Assad, Stephanie et al. (2024). "Algorithmic pricing and competition: Empirical evidence from the German retail gasoline market". In: *Journal of Political Economy* 132.3, pp. 723–771.

Guo, Xudong et al. (2024). *Embodied LLM Agents Learn to Cooperate in Organized Teams*. arXiv: 2403.12482 [cs.AI]. URL: https://arxiv.org/abs/2403.12482.

Huang, Jen-tse et al. (2024). "How Far Are We on the Decision-Making of LLMs? Evaluating LLMs' Gaming Ability in Multi-Agent Environments". In: *arXiv preprint arXiv:2403.11807*.

Hua, Wenyue et al. (2024). "Game-theoretic LLM: Agent workflow for negotiation games". In: *arXiv preprint arXiv:2411.05990*.

Buscemi, Alessio et al. (2025). "Fairgame: a framework for ai agents bias recognition using game theory". In: *arXiv preprint arXiv:2504.14325*.

Akata, Elif et al. (2025). "Playing repeated games with large language models". In: *Nature Human Behaviour* 9.7, pp. 1380–1390.

Park, Chanwoo et al. (2024). "Do llm agents have regret? a case study in online learning and games". In: *arXiv preprint arXiv:2403.16843*.

Duque, Juan Agustin et al. (2024). "Advantage alignment algorithms". In: *arXiv preprint arXiv:2406.14662*.

Kalai, Ehud and Ehud Lehrer (1993a). "Rational learning leads to Nash equilibrium". In: *Econometrica: Journal of the Econometric Society*, pp. 1019–1045.

Norman, Thomas WL (2022). "The possibility of Bayesian learning in repeated games". In: *Games and Economic Behavior* 136, pp. 142–152.

Yamin, Khurram et al. (2026). "Do LLMs Act Like Rational Agents? Measuring Belief Coherence in Probabilistic Decision Making". In: *arXiv preprint arXiv:2602.06286*.

Ge, Luise, Yongyan Zhang, and Yevgeniy Vorobeychik (2026). "Mind the (DH) Gap! A Contrast in Risky Choices Between Reasoning and Conversational LLMs". In: *arXiv preprint arXiv:2602.15173*.

Arumugam, Dilip and Thomas L Griffiths (2025). "Toward efficient exploration by large language model agents". In: *arXiv preprint arXiv:2504.20997*.

Coda-Forno, Julian et al. (2023). "Meta-in-context learning in large language models". In: *Advances in Neural Information Processing Systems* 36, pp. 65189–65201.

Lu, Sheng et al. (2024). "Are emergent abilities in large language models just in-context learning?" In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5098–5139.

Cahyawijaya, Samuel, Holy Lovenia, and Pascale Fung (2024). "Llms are few-shot in-context low-resource language learners". In: *arXiv preprint arXiv:2403.16512*.

Xie, Sang Michael et al. (2021). "An explanation of in-context learning as implicit bayesian inference". In: *arXiv preprint arXiv:2111.02080*.

Wang, Xinyi et al. (2023). "Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning". In: *Advances in Neural Information Processing Systems* 36, pp. 15614–15638.

Wakayama, Tomoya and Taiji Suzuki (2025). "In-context learning is provably Bayesian inference: a generalization theory for meta-learning". In: *arXiv preprint arXiv:2510.10981*.

Falck, Fabian, Ziyu Wang, and Chris Holmes (2024). "Is in-context learning in large language models bayesian? a martingale perspective". In: *arXiv preprint arXiv:2406.00793*.

Nachbar, John H (1997). "Prediction, optimization, and learning in repeated games". In: *Econometrica: Journal of the Econometric Society*, pp. 275–309.

– (2005). "Beliefs in repeated games". In: *Econometrica* 73.2, pp. 459–480.

Sun, Haoran et al. (2025). "Game theory meets large language models: A systematic survey with taxonomy and new frontiers". In: *arXiv preprint arXiv:2502.09053*.

Jia, Jingru et al. (2025). "LLM Strategic Reasoning: Agentic Study through Behavioral Game Theory". In: *arXiv preprint arXiv:2502.20432*.

Guo, Fulin (2023). *GPT in Game Theory Experiments*. DOI: 10.48550/arXiv.2305.05516. arXiv: 2305.05516 [econ.GN].

Fan, Caoyun et al. (2023). *Can Large Language Models Serve as Rational Players in Game Theory? A Systematic Analysis*. AAAI 2024. DOI: 10.48550/arXiv.2312.05488. arXiv: 2312.05488 [cs.AI].

Kader, Gavin and Dongwoo Lee (2024). "The Emergence of Strategic Reasoning of Large Language Models". In: *arXiv preprint arXiv:2412.13013*.

Mao, Shaoguang et al. (2023). *ALYMPICS: LLM Agents Meet Game Theory – Exploring Strategic Decision-Making with AI Agents*. DOI: 10.48550/arXiv.2311.03220. arXiv: 2311.03220 [cs.CL].

Duan, Jinhao et al. (2024). *GTBench: Uncovering the Strategic Reasoning Limitations of LLMs via Game-Theoretic Evaluations*. DOI: 10.48550/arXiv.2402.12348. arXiv: 2402.12348 [cs.CL].

Fontana, Nicoló, Francesco Pierri, and Luca Maria Aiello (2024). "Nicer Than Humans: How do Large Language Models Behave in the Prisoner's Dilemma?" In: *arXiv preprint arXiv:2406.13605*.

Willis, Richard et al. (2025). "Will Systems of LLM Agents Cooperate: An Investigation into a Social Dilemma". In: *arXiv preprint arXiv:2501.16173*.

Agrawal, Kushal et al. (2025). *Evaluating LLM Agent Collusion in Double Auctions*. DOI: 10.48550/arXiv.2507.01413. arXiv: 2507.01413 [cs.GT].

Zhang, Yufeng et al. (2023). "What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization". In: *arXiv preprint arXiv:2305.19420*.

Zhang, Kelly W et al. (2024). "Posterior sampling via autoregressive generation". In: *NeurIPS 2024 Workshop on Bayesian Decision-making and Uncertainty*.

Welleck, Sean et al. (2024). "From decoding to meta-generation: Inference-time algorithms for large language models". In: *arXiv preprint arXiv:2406.16838*.

Durrett, Rick (2019). *Probability: Theory and Examples*. 5th ed. See Theorem 2.1.21 (Kolmogorov's extension theorem). Cambridge University Press. DOI: 10.1017/9781108591034.

Kuhn, Harold W (1953). "Extensive games and the problem of information". In: *Contributions to the Theory of Games* 2.28, pp. 193–216.

Aumann, Robert J (1961). *Mixed and behavior strategies in infinite extensive games*. Princeton University Princeton.

Kalai, Ehud and Ehud Lehrer (1993b). "Subjective equilibrium in repeated games". In: *Econometrica* 61.5, pp. 1231–1240.

Blackwell, David and Lester Dubins (1962). "Merging of opinions with increasing information". In: *The Annals of Mathematical Statistics* 33.3, pp. 882–886.

Cai, Tiffany Tianhui et al. (2024). "Active Exploration via Autoregressive Generation of Missing Data". In: *arXiv preprint arXiv:2405.19466*.

Zhou, Pei et al. (2023). "How far are large language models from agents with theory-of-mind?" In: *arXiv preprint arXiv:2310.03051*.

Riemer, Matthew et al. (2024). "Position: Theory of mind benchmarks are broken for large language models". In: *arXiv preprint arXiv:2412.19726*.

Aoyagi, Masaki, Guillaume R Fréchette, and Sevgi Yuksel (2024). "Beliefs in repeated games: An experiment". In: *American Economic Review* 114.12, pp. 3944–3975.

Gill, David and Yaroslav Rosokha (2024). "Beliefs, learning, and personality in the indefinitely repeated prisoner's dilemma". In: *American Economic Journal: Microeconomics* 16.3, pp. 259–283.

Qwen Team (Feb. 2026). *Qwen3.5: Towards Native Multimodal Agents*. URL: https://qwen.ai/blog?id=qwen3.5.

Singh, Aaditya et al. (2025). "Openai gpt-5 system card". In: *arXiv preprint arXiv:2601.03267*.

Lal, Rajiv (1990). "Price promotions: Limiting competitive encroachment". In: *Marketing science* 9.3, pp. 247–262.

Abreu, Dilip (1988). "On the theory of infinitely repeated games with discounting". In: *Econometrica: Journal of the Econometric Society*, pp. 383–396.

Notation Summary

Symbol	Meaning
Repeated-game primitives	
$I = \{1, \dots, N\}$	Finite set of players / AI agents.
$A_i, \quad A = \prod_{i \in I} A_i$	Player-i action set; A is the joint action space.
$a^t = (a_1^t, \dots, a_N^t) \in A$	Joint action profile at round t.
$u_i : A \to [0, 1]$	Stage payoff in the common-knowledge benchmark; in the private-payoff extension, the true mean payoff matrix generating player i's private stochastic rewards.
$\lambda_i \in (0, 1)$	Player-i discount factor.
$\varphi_i,\ \Phi_i^\star$	Stage-game minmax payoff and pure-action maxmin payoff used in the Non-MM* condition.
Histories, strategies, and induced laws	
$h^t = (a^1, \dots, a^{t-1})$	Public history observed before round t.
$H^0,\ H^t,\ H,\ H^\infty$	Empty history; histories of length $t-1$; all finite public histories; all infinite play paths.
$C(h)$	Cylinder set of infinite paths having prefix h.
$f_i : H \to \Delta(A_i)$	Public-history strategy of player i in the benchmark repeated game.
$\mathcal{F}_i, \quad \mathcal{F} = \prod_{i \in I} \mathcal{F}_i$	Individual and joint strategy spaces in the benchmark model.
μ^f	Play-path distribution induced by strategy profile f.
$\mu^g_{h^t}$	Continuation distribution after history h^t when continuation profile g is played thereafter.
$U_i(f)$	Objective discounted payoff of player i under profile f.
Beliefs and subjective continuation values	
$\mu_i^0, \quad \mu_i^t(\cdot \mid h^t)$	Player-i prior and posterior over opponents' strategy profiles.
$P_i^{\mu_i, g_i}$	Predictive play-path distribution induced by own strategy g_i and belief μ_i over opponents' strategies.
$f_{-i}^i, \quad f_{-i}^{i,t}$	Belief-equivalent behavioral representatives of player i's prior predictive and posterior predictive beliefs about opponents' play.
$V_i(\sigma_i \mid h^t; g_{-i})$ $V_i(\sigma_i \mid h^t)$	Subjective continuation value of continuation strategy σ_i against continuation model g_{-i}; shorthand $V_i(\sigma_i \mid h^t) = V_i(\sigma_i \mid h^t; f_{-i}^{i,t})$.
$\mathrm{BR}_i^\varepsilon(g_{-i} \mid h^t)$	Set of ε-best-response continuation strategies at history h^t.
$\mu^f \ll P_i^{0, f_i}$	Grain-of-truth condition: player i's prior predictive does not rule out events that occur under the true play distribution.
$d(\mu, \nu), \quad d_{h^t}(\mu, \nu)$	Weak distance between play-path distributions and its continuation version after history h^t.
$\xi,\ \eta,\ \varepsilon$	Approximation tolerances for subjective best response, predictive accuracy, and Nash / best-response error.
Posterior-sampling and one-step reasoning	
$\mathcal{S}_{-i}, \quad \nu$	Finite menu of candidate opponent strategies in PS-BR; uniform positivity lower bound in the caution assumption.
$p_t(g_{-i}) = \mu_i^t(g_{-i} \mid h^t)$	Posterior mass on opponent hypothesis g_{-i}.

Symbol	Meaning
$\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid h^t)$	Continuation strategy chosen by posterior-sampling best response (PS-BR) at history h^t.
$D_i^t(h^t) = 1 - \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2$	Posterior collision complement; upper-bounds the PS-BR best-response gap.
$D_{\mathrm{KL}}(p\|q)$	Kullback–Leibler divergence between two distributions.
$q_i^t(\cdot \mid h^t), \quad \mathrm{br}_i^{\varepsilon}(q)$	One-step posterior predictive belief over opponents' next joint action; stage-game ε-best-response set to next-action distribution q.
$\alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t)$	Mixed action induced by myopic PS-BR at history h^t.
$\hat{a}_{-i}^t(h^t), \quad b_i(\cdot)$	SCoT MAP prediction of opponents' next action; deterministic pure best-response selector.
Private-payoff extension	
r_i^t	Privately observed stochastic payoff of player i at round t.
$\mathcal{R}_i, \quad \nu_i(\mathrm{d}r), \quad \psi_i(r;\mu)$	Payoff space, dominating base measure, and known noise-family density indexed by mean μ.
$x_i^t = (h^t, r_i^{1:t-1}), \quad X_i^t,\ X_i$	Player-i information history at time t, its time-t state space, and the full information-history space.
$\sigma_i : X_i \to \Delta(A_i), \quad \Sigma_i$	Information-history strategy in the private-payoff game and its strategy space.
$\Omega, \quad P^{\sigma,u}$	Full sample space of public actions and private rewards; objective law induced by strategy profile σ and true mean payoffs u.
$U_i(\tau \mid x^t), \quad \bar{\mu}_{x^t}^{\tau,u}$	Objective continuation payoff after realized information-history vector x^t; public-action marginal of the induced continuation law.
$\mathcal{M}_i, \quad \pi_i^t(m_i \mid x_i^t)$	Finite menu of candidate mean payoff matrices and player-i posterior over that menu.
$q_i^{m_i}(\mathrm{d}r \mid a)$	Payoff kernel induced by candidate mean matrix m_i at joint action a.
$V_i^{m_i}(\tau_i \mid x_i^t; g_{-i})$ $\mathrm{BR}_{i,m_i}^{\varepsilon}(g_{-i} \mid x_i^t)$	Subjective continuation value and corresponding ε-best-response set when player i evaluates payoffs using candidate mean matrix m_i.
$V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t)$ $V_i^{u_i,t}(\tau_i \mid x_i^t)$	Mixture continuation value under the joint opponent/payoff posteriors; the same object evaluated under the true mean matrix u_i.
$g_{-i}^{i,t}, \quad \Pi_i^t(\cdot \mid x_i^t)$	Reduced-form behavioral representative of player i's predictive public-action belief; posterior predictive law over future public-action paths.
$D_i^{t,\mathrm{joint}}(x_i^t), \quad \delta_i^t(x_i^t)$	Joint collision complement over opponent-strategy and payoff hypotheses; payoff-posterior error $1 - \pi_i^t(u_i \mid x_i^t)$.
Bounded-memory appendix notation	
$\mathrm{suffix}_\kappa(h),\ \mathcal{F}_i^\kappa,\ \mathsf{S}_\kappa,\ T_\kappa$	Last κ joint actions of history h; bounded-memory strategy space; finite suffix-state space; deterministic state-update map in the finite-state reduction.

A Continuity and Finite-Horizon Robustness

Lemma A.1 (Continuity of discounted payoff)**.** *For each agent i and every $\delta > 0$, there exists $\rho_i(\delta) > 0$ such that for any strategy profiles $f, g \in \mathcal{F}$,*

$$d(\mu^f, \mu^g) \leq \rho_i(\delta) \quad \Rightarrow \quad \left|U_i(f) - U_i(g)\right| \leq \delta.$$

In particular, if $\rho(\delta) = \min_{i \in I} \rho_i(\delta)$ and $d(\mu^f, \mu^g) \leq \rho(\delta)$, then $\left|U_i(f) - U_i(g)\right| \leq \delta$ for all $i \in I$.

A.1 Finite-horizon variants and robustness

For a finite horizon $T \in \mathbb{N}$, we denote by $\mathcal{F}^T$ the set of behaviour strategies specified on histories of length at most T; two full strategies that coincide on these histories induce the same distribution over histories up to time T and the same truncated payoff. For $f \in \mathcal{F}^T$, define the T-period discounted payoff

$$U_i^T(f) = \mathbb{E}_{z \sim \mu^f}\Big[(1-\lambda_i)\sum_{t=1}^{T} \lambda_i^{t-1} u_i(z^t)\Big].$$

Definition 15 (Finite-horizon weak ξ-subjective η-equilibrium)**.** Let $\xi, \eta \geq 0$ and a fixed horizon T. A truncated strategy profile $f \in \mathcal{F}^T$ is a *finite-horizon weak ξ-subjective η-equilibrium* if for each agent $i \in I$ there exists a supporting truncated profile $f^i \in \mathcal{F}^T$ such that:

- $f_i^i = f_i$;
- $U_i^T(f_i, f_{-i}^i) \geq \sup_{g_i \in \mathcal{F}_i^T} U_i^T(g_i, f_{-i}^i) - \xi$;
- $d(\mu^{f^i}, \mu^f) \leq \eta$ when d is computed using only cylinder events in $\mathcal{B}^t$ with $t \leq T$.

We now show that finite-horizon weak subjective equilibria can be "patched" into approximate finite-horizon Nash equilibria without changing the induced distribution of play up to time T.

Lemma A.2 (Finite-horizon purification for $\eta = 0$ Norman, 2022)**.** *Fix a finite horizon T and a profile $f \in \mathcal{F}^T$. Suppose f is a finite-horizon weak ψ-subjective 0-equilibrium for some $\psi \geq 0$. Then there exists a truncated strategy profile $\hat{f} \in \mathcal{F}^T$ such that:*

- *$\hat{f}$ is a ψ-Nash equilibrium of the T-period game, i.e., for all $i \in I$ and all $g_i \in \mathcal{F}_i^T$,*

$$U_i^T(\hat{f}_i, \hat{f}_{-i}) \;\geq\; U_i^T(g_i, \hat{f}_{-i}) - \psi;$$

- *the induced distributions of histories of length at most T coincide: for every $E \in \mathcal{B}^T$, $\mu^{\hat{f}}(E) = \mu^f(E)$.*

We next extend this to the case where $\eta > 0$ but small, using a compactness and limit argument.

Lemma A.3 (Finite-horizon robustness)**.** *Fix a finite horizon T and $\psi > 0$. For every $\theta > 0$ there exists*

$\bar{\eta}_T(\psi,\theta) > 0$ *such that: if* $f \in \mathcal{F}^T$ *is a finite-horizon weak* ψ*-subjective* η*-equilibrium with* $\eta \leq \bar{\eta}_T(\psi,\theta)$*, then there exists a* ψ*-Nash equilibrium* $\hat{f} \in \mathcal{F}^T$ *satisfying*

$$d(\mu^{\hat{f}}, \mu^f) \leq \theta$$

(again with d *computed on cylinder events of length at most* T*).*

We now patch finite-horizon robustness to the infinite-horizon game by truncating the payoff at a sufficiently large horizon and using Lemma A.1; the resulting infinite-horizon patching lemma is recorded below.

Lemma A.4 (Infinite-horizon patching)**.** *Fix* $\xi > 0$ *and* $\varepsilon > 0$*. There exists* $\hat{\eta}(\xi,\varepsilon) > 0$ *such that if* $f \in \mathcal{F}$ *is a weak* ξ*-subjective* η*-equilibrium in the sense of Definition 8 with* $\eta \leq \hat{\eta}(\xi,\varepsilon)$*, then there exists a strategy profile* $\hat{f} \in \mathcal{F}$ *satisfying:*

- $\hat{f}$ *is a* $(\xi+\varepsilon)$*-Nash equilibrium of the infinite-horizon game;*
- $d(\mu^{\hat{f}}, \mu^f) \leq \varepsilon$*.*

Remark 3 (Continuation-game analogues). Lemmas A.2–A.4 apply verbatim to continuation games after any history h^t by interpreting $U_i(\cdot)$ as continuation payoff from h^t and $d(\cdot,\cdot)$ as the weak distance between $\mu^g_{h^t}$ and $\mu^{g'}_{h^t}$. They also apply verbatim to the private-payoff continuation game after any realized information-history vector x^t when $\mathcal{F}_i$ is replaced by Σ_i, histories h^t are replaced by x^t, payoffs are $U_i(\tau \mid x^t)$, and weak distance is computed on the public-action marginals $\bar{\mu}^{\tau,u}_{x^t}$.

B Proofs

Proof of Lemma A.1. Fix i and $\delta > 0$. Choose a finite horizon $T \in \mathbb{N}$ large enough that

$$(1-\lambda_i)\sum_{t=T+1}^{\infty}\lambda_i^{t-1} \;\le\; \frac{\delta}{4}. \tag{16}$$

For any profile $g \in \mathcal{F}$, define the truncated payoff

$$U_i^T(g) = \mathbb{E}_{z\sim\mu^g}\left[(1-\lambda_i)\sum_{t=1}^{T}\lambda_i^{t-1}u_i(z^t)\right].$$

Then for any g we have

$$\left|U_i(g) - U_i^T(g)\right| \le (1-\lambda_i)\sum_{t=T+1}^{\infty}\lambda_i^{t-1} \le \frac{\delta}{4}$$

by (16), using that $u_i(\cdot) \in [0,1]$.

Now fix $f, g \in \mathcal{F}$. We can decompose

$$\left|U_i(f) - U_i(g)\right| \le \left|U_i(f) - U_i^T(f)\right| + \left|U_i^T(f) - U_i^T(g)\right| + \left|U_i^T(g) - U_i(g)\right|.$$

By the bound above, the first and third terms are each at most $\delta/4$. It remains to control $|U_i^T(f) - U_i^T(g)|$.

For each $t \in \{1,\dots,T\}$ and each joint action profile $a \in A$, let

$$\alpha_t^f(a) = \mu^f\big(\{z \in H^\infty : z^t = a\}\big), \quad \alpha_t^g(a) = \mu^g\big(\{z \in H^\infty : z^t = a\}\big).$$

Since $u_i(a) \in [0,1]$ for all a, we have

$$\left|\sum_{a\in A} u_i(a)\big(\alpha_t^f(a) - \alpha_t^g(a)\big)\right| \le \sup_{E\in\mathcal{B}^t}\left|\mu^f(E) - \mu^g(E)\right|.$$

Hence

$$\begin{aligned}\left|U_i^T(f) - U_i^T(g)\right| &= \left|\sum_{t=1}^{T}(1-\lambda_i)\lambda_i^{t-1}\sum_{a\in A}u_i(a)\big(\alpha_t^f(a) - \alpha_t^g(a)\big)\right| \\ &\le \sum_{t=1}^{T}(1-\lambda_i)\lambda_i^{t-1}\sup_{E\in\mathcal{B}^t}\left|\mu^f(E) - \mu^g(E)\right|.\end{aligned}$$

By the definition (6) of $d(\mu^f, \mu^g)$, for each t we have

$$2^{-t}\sup_{E\in\mathcal{B}^t}\left|\mu^f(E) - \mu^g(E)\right| \le d(\mu^f, \mu^g),$$

hence

$$\sup_{E\in\mathcal{B}^t}\left|\mu^f(E)-\mu^g(E)\right|\le 2^t d(\mu^f,\mu^g).$$

Thus

$$\left|U_i^T(f)-U_i^T(g)\right|\le d(\mu^f,\mu^g)\sum_{t=1}^{T}(1-\lambda_i)\lambda_i^{t-1}2^t.$$

The finite sum on the right depends only on T and λ_i; call it $C_i(T)$. Define

$$\rho_i(\delta)=\min\left\{\frac{\delta}{4C_i(T)},\,1\right\}.$$

If $d(\mu^f,\mu^g)\le\rho_i(\delta)$, then

$$\left|U_i^T(f)-U_i^T(g)\right|\le C_i(T)\rho_i(\delta)\le\frac{\delta}{4}.$$

Combining the three bounds gives

$$\left|U_i(f)-U_i(g)\right|\le\frac{\delta}{4}+\frac{\delta}{4}+\frac{\delta}{4}\;<\;\delta.$$

Setting $\rho(\delta)=\min_{i\in I}\rho_i(\delta)$ yields the final claim. □

Proof of Lemma A.2. This is the finite-horizon analogue of the "purification" or "deviation-tree patching" result for weak subjective equilibria in Norman (2022). The key idea is to modify off-path behavior so that, for each player i, any history that can only arise from a deviation by i triggers opponents' play according to the supporting profile f^i (which makes f_i a ψ-best response), while on-path histories preserve the original profile f.

Formally, one constructs a deviation tree for each player and assigns to each subtree corresponding to a first deviation by i the opponents' strategies from f^i_{-i}, keeping f on the non-deviation branch. This construction ensures: (i) if all players follow $\hat{f}$, the induced distribution of histories up to time T coincides with that under f (item 2); and (ii) any unilateral deviation by player i induces, up to time T, the same distribution of histories as deviating against f^i_{-i}, against which f_i is a ψ-best reply by Definition 15. Therefore $\hat{f}$ is a ψ-Nash equilibrium of the T-period game (item 1).

A detailed construction and proof of these properties is given in Norman (2022), Proposition 3.1, and the associated deviation-tree arguments; our setting is the same repeated-game environment, so the proof carries over verbatim. □

Proof of Lemma A.3. Suppose, towards a contradiction, that there exist $T,\psi>0$ and $\theta>0$ such that for every $m\in\mathbb{N}$ there is a finite-horizon weak ψ-subjective η_m-equilibrium $f^{(m)}\in\mathcal{F}^T$ with $\eta_m\le 1/m$ and such that no ψ-Nash equilibrium lies within weak distance θ of $\mu^{f^{(m)}}$ (measured on $\mathcal{B}^T$).

For each m and each $i \in I$, let $f^{i,(m)}$ be a supporting truncated profile witnessing that $f^{(m)}$ is a finite-horizon weak ψ-subjective η_m-equilibrium, i.e., $f_i^{i,(m)} = f_i^{(m)}$,

$$U_i^T(f_i^{(m)}, f_{-i}^{i,(m)}) \geq \sup_{g_i \in \mathcal{F}_i^T} U_i^T(g_i, f_{-i}^{i,(m)}) - \psi, \quad d(\mu^{f^{i,(m)}}, \mu^{f^{(m)}}) \leq \eta_m.$$

Because the horizon T and action sets are finite, the space of behaviour strategies $\mathcal{F}^T$ is a finite-dimensional product of simplices and hence compact in the product topology. Thus, by sequential compactness, there exists a subsequence (which we relabel for notational convenience) such that

$$f^{(m)} \to f^\star \quad \text{and} \quad f^{i,(m)} \to f^{i,\star} \quad \text{for all } i \in I,$$

as $m \to \infty$, in the product topology on $\mathcal{F}^T$.

The map $f \mapsto \mu^f$ on finite histories (up to time T) is continuous with respect to this topology and the weak topology induced by d (restricted to $\mathcal{B}^T$), so

$$\mu^{f^{(m)}} \to \mu^{f^\star}, \quad \mu^{f^{i,(m)}} \to \mu^{f^{i,\star}}.$$

Since $d(\mu^{f^{i,(m)}}, \mu^{f^{(m)}}) \leq \eta_m \to 0$, we must have $d(\mu^{f^{i,\star}}, \mu^{f^\star}) = 0$, so $\mu^{f^{i,\star}} = \mu^{f^\star}$ on $\mathcal{B}^T$.

Moreover, the best-response inequality passes to the limit. Fix i and any $g_i \in \mathcal{F}_i^T$. For all m,

$$U_i^T(f_i^{(m)}, f_{-i}^{i,(m)}) \geq \sup_{g_i' \in \mathcal{F}_i^T} U_i^T(g_i', f_{-i}^{i,(m)}) - \psi \geq U_i^T(g_i, f_{-i}^{i,(m)}) - \psi.$$

By continuity of U_i^T in the product topology (an immediate consequence of Lemma A.1 restricted to horizon T), taking $m \to \infty$ yields

$$U_i^T(f_i^\star, f_{-i}^{i,\star}) \geq U_i^T(g_i, f_{-i}^{i,\star}) - \psi.$$

Since g_i was arbitrary and $f_i^{i,\star} = f_i^\star$ (by pointwise convergence of $f_i^{i,(m)}$ to $f_i^{i,\star}$ and of $f_i^{(m)}$ to $f_i^\star$), we conclude that

$$U_i^T(f_i^\star, f_{-i}^{i,\star}) \geq \sup_{g_i \in \mathcal{F}_i^T} U_i^T(g_i, f_{-i}^{i,\star}) - \psi.$$

Together with $d(\mu^{f^{i,\star}}, \mu^{f^\star}) = 0$, this shows that $f^\star$ is a finite-horizon weak ψ-subjective 0-equilibrium of the T-period game.

By Lemma A.2, there exists a profile $\hat{f}^\star \in \mathcal{F}^T$ such that $\hat{f}^\star$ is a ψ-Nash equilibrium of the T-period game and $\mu^{\hat{f}^\star}$ coincides with $\mu^{f^\star}$ on histories of length at most T. In particular, $d(\mu^{\hat{f}^\star}, \mu^{f^\star}) = 0$.

Since $\mu^{f^{(m)}} \to \mu^{f^\star}$ in the weak metric d (restricted to $\mathcal{B}^T$), we have $d(\mu^{f^{(m)}}, \mu^{\hat{f}^\star}) \to 0$ as $m \to \infty$. Thus for all sufficiently large m, $d(\mu^{f^{(m)}}, \mu^{\hat{f}^\star}) \leq \theta$. But $\hat{f}^\star$ is a ψ-Nash equilibrium, contradicting the assumption

that no ψ-Nash equilibrium lies within weak distance θ of $\mu^{f^{(m)}}$. This contradiction shows that such a sequence $(f^{(m)})$ cannot exist, and hence there must exist $\bar{\eta}_T(\psi, \theta) > 0$ with the stated property. □

Proof of Lemma A.4. Fix $\xi > 0$ and $\varepsilon > 0$. Choose a finite horizon T large enough that, for all $i \in I$ and all profiles $h \in \mathcal{F}$,

$$\left|U_i(h) - U_i^T(h)\right| \;\le\; \frac{\varepsilon}{8}, \tag{17}$$

and also

$$\sum_{t>T} 2^{-t} \;\le\; \frac{\varepsilon}{4}. \tag{18}$$

Such a T exists because the tails of both geometric series are uniformly small.

Let f be a weak ξ-subjective η-equilibrium with supporting profiles $\{f^i\}_{i\in I}$ as in Definition 8, i.e., for each i,

$$f_i^i = f_i, \quad U_i(f_i, f_{-i}^i) \ge \sup_{g_i \in \mathcal{F}_i} U_i(g_i, f_{-i}^i) - \xi, \quad d(\mu^{f^i}, \mu^f) \le \eta.$$

Consider the truncated profiles $f^{(T)}$ and $(f^i)^{(T)}$ obtained by restricting the prescriptions of f and f^i to histories of length at most T. For each i we have $(f_i^i)^{(T)} = f_i^{(T)}$ and, since the weak distance on histories up to T is bounded by the full weak distance,

$$d(\mu^{(f^i)^{(T)}}, \mu^{f^{(T)}}) \le d(\mu^{f^i}, \mu^f) \le \eta.$$

We now show that $f^{(T)}$ is a finite-horizon weak ψ_T-subjective η-equilibrium for a slightly relaxed parameter ψ_T. Fix i and note that for any profile h,

$$|U_i(h) - U_i^T(h)| \le \frac{\varepsilon}{8}$$

by (17). Using the weak subjective inequality for f and f^i, we obtain

$$\begin{aligned} U_i^T(f_i^{(T)}, (f_{-i}^i)^{(T)}) &= U_i^T(f_i, f_{-i}^i) \\ &\ge U_i(f_i, f_{-i}^i) - \frac{\varepsilon}{8} \\ &\ge \sup_{g_i \in \mathcal{F}_i} U_i(g_i, f_{-i}^i) - \xi - \frac{\varepsilon}{8}. \end{aligned}$$

For any truncated deviation $g_i^{(T)} \in \mathcal{F}_i^T$ we can extend it arbitrarily to a full strategy $g_i \in \mathcal{F}_i$, and then

$$U_i(g_i, f_{-i}^i) \ge U_i^T(g_i^{(T)}, (f_{-i}^i)^{(T)}) - \frac{\varepsilon}{8},$$

again by (17). Taking the supremum over $g_i^{(T)}$ yields

$$U_i^T(f_i^{(T)}, (f_{-i}^i)^{(T)}) \geq \sup_{g_i^{(T)} \in \mathcal{F}_i^T} U_i^T(g_i^{(T)}, (f_{-i}^i)^{(T)}) - \xi - \frac{\varepsilon}{4}.$$

Thus, if we define

$$\psi_T := \xi + \frac{\varepsilon}{4},$$

then for each i the truncated profiles $f^{(T)}$ and $(f^i)^{(T)}$ satisfy

$$U_i^T(f_i^{(T)}, (f_{-i}^i)^{(T)}) \geq \sup_{g_i^{(T)} \in \mathcal{F}_i^T} U_i^T(g_i^{(T)}, (f_{-i}^i)^{(T)}) - \psi_T,$$

and $d(\mu^{(f^i)^{(T)}}, \mu^{f^{(T)}}) \leq \eta$, so $f^{(T)}$ is a finite-horizon weak ψ_T-subjective η-equilibrium in the sense of Definition 15.

Applying Lemma A.3 with this T, $\psi = \psi_T$ and $\theta = \varepsilon/2$, there exists $\bar{\eta}_T(\psi_T, \varepsilon/2) > 0$ such that if $\eta \leq \bar{\eta}_T(\psi_T, \varepsilon/2)$ then there is a ψ_T-Nash equilibrium $\tilde{f}^{(T)} \in \mathcal{F}^T$ for the T-period game with

$$d(\mu^{\tilde{f}^{(T)}}, \mu^{f^{(T)}}) \leq \frac{\varepsilon}{2}.$$

Define

$$\hat{\eta}(\xi, \varepsilon) := \bar{\eta}_T(\xi + \tfrac{\varepsilon}{4}, \tfrac{\varepsilon}{2}).$$

Assume henceforth that $\eta \leq \hat{\eta}(\xi, \varepsilon)$ so that this conclusion holds.

Extend $\tilde{f}^{(T)}$ arbitrarily to a full strategy profile $\hat{f} \in \mathcal{F}$ by specifying its behaviour after period T in any way. Then $\hat{f}$ and $\tilde{f}^{(T)}$ coincide on periods $t \leq T$, and similarly f and $f^{(T)}$ coincide on $t \leq T$. The weak distance between $\hat{f}$ and f can be bounded as

$$d(\mu^{\hat{f}}, \mu^f) \leq d(\mu^{\hat{f}}, \mu^{\tilde{f}^{(T)}}) + d(\mu^{\tilde{f}^{(T)}}, \mu^{f^{(T)}}) + d(\mu^{f^{(T)}}, \mu^f).$$

The second term is at most $\varepsilon/2$ by construction. For the first and third terms, any discrepancy between $\hat{f}$ and $\tilde{f}^{(T)}$ (respectively, f and $f^{(T)}$) occurs only at times $t > T$, so each of these weak distances is bounded by the tail $\sum_{t>T} 2^{-t} \leq \varepsilon/4$ by (18). Hence

$$d(\mu^{\hat{f}}, \mu^f) \leq \frac{\varepsilon}{4} + \frac{\varepsilon}{2} + \frac{\varepsilon}{4} = \varepsilon.$$

It remains to show that $\hat{f}$ is a $(\xi + \varepsilon)$-Nash equilibrium of the infinite-horizon game. Fix $i \in I$ and any deviation $g_i \in \mathcal{F}_i$. Let $g_i^{(T)}$ denote the truncation of g_i to a T-period strategy, i.e., its prescriptions on histories of length at most T; clearly $U_i^T(g_i, \hat{f}_{-i}) = U_i^T(g_i^{(T)}, \tilde{f}_{-i}^{(T)})$ since $\hat{f}$ and $\tilde{f}^{(T)}$ coincide on the first T

periods.

Because $\tilde{f}^{(T)}$ is a ψ_T-Nash equilibrium of the T-period game,

$$U_i^T(\tilde{f}_i^{(T)}, \tilde{f}_{-i}^{(T)}) \;\geq\; U_i^T(g_i^{(T)}, \tilde{f}_{-i}^{(T)}) - \psi_T.$$

Using the truncation bound (17), we obtain

$$U_i(\hat{f}_i, \hat{f}_{-i}) \;\geq\; U_i^T(\hat{f}_i, \hat{f}_{-i}) - \frac{\varepsilon}{8} = U_i^T(\tilde{f}_i^{(T)}, \tilde{f}_{-i}^{(T)}) - \frac{\varepsilon}{8}$$

and

$$U_i(g_i, \hat{f}_{-i}) \;\leq\; U_i^T(g_i, \hat{f}_{-i}) + \frac{\varepsilon}{8} = U_i^T(g_i^{(T)}, \tilde{f}_{-i}^{(T)}) + \frac{\varepsilon}{8}.$$

Combining these inequalities yields

$$\begin{aligned} U_i(\hat{f}_i, \hat{f}_{-i}) &\geq U_i^T(\tilde{f}_i^{(T)}, \tilde{f}_{-i}^{(T)}) - \frac{\varepsilon}{8} \\ &\geq U_i^T(g_i^{(T)}, \tilde{f}_{-i}^{(T)}) - \psi_T - \frac{\varepsilon}{8} \\ &\geq U_i(g_i, \hat{f}_{-i}) - \psi_T - \frac{\varepsilon}{4}. \end{aligned}$$

Recalling that $\psi_T = \xi + \varepsilon/4$, we have

$$\psi_T + \frac{\varepsilon}{4} = \xi + \frac{\varepsilon}{2} \leq \xi + \varepsilon,$$

so for every deviation g_i,

$$U_i(\hat{f}_i, \hat{f}_{-i}) \geq U_i(g_i, \hat{f}_{-i}) - (\xi + \varepsilon).$$

Thus $\hat{f}$ is a $(\xi + \varepsilon)$-Nash equilibrium. □

Proof of Lemma 4.1. For each $g_{-i} \in \mathcal{S}_{-i}$ define the continuation value envelope

$$M(g_{-i}) \;:=\; \sup_{\sigma_i} V_i(\sigma_i \mid h^t; g_{-i}) \;\in\; [0,1].$$

For each g_{-i} pick a (measurable) best response $\sigma_i^{g_{-i}} \in \mathrm{BR}_i(g_{-i} \mid h^t)$, so that $V_i(\sigma_i^{g_{-i}} \mid h^t; g_{-i}) = M(g_{-i})$.

By definition, PS-BR first samples $\tilde{g}_{-i} \sim p_t(\cdot)$ and then plays $\sigma_i^{\tilde{g}_{-i}}$. Evaluating against the posterior predictive belief and using linearity in the mixing over opponent hypotheses,

$$\begin{aligned} V_i(\sigma_{i,t}^{\mathrm{PS}} \mid h^t) &= \sum_{\tilde{g}_{-i} \in \mathcal{S}_{-i}} p_t(\tilde{g}_{-i}) \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})\, V_i(\sigma_i^{\tilde{g}_{-i}} \mid h^t; g_{-i}) \\ &\geq \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2\, V_i(\sigma_i^{g_{-i}} \mid h^t; g_{-i}) \end{aligned}$$

$$= \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2 \, M(g_{-i}).$$

On the other hand,

$$\sup_{\sigma_i} V_i(\sigma_i \mid h^t) = \sup_{\sigma_i} \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i}) \, V_i(\sigma_i \mid h^t; g_{-i}) \le \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i}) \, M(g_{-i}).$$

Subtracting and using $M(g_{-i}) \le 1$,

$$\begin{aligned} \sup_{\sigma_i} V_i(\sigma_i \mid h^t) - V_i(\sigma_{i,t}^{\mathrm{PS}} \mid h^t) &\le \sum_{g_{-i} \in \mathcal{S}_{-i}} \Big(p_t(g_{-i}) - p_t(g_{-i})^2 \Big) M(g_{-i}) \\ &\le \sum_{g_{-i} \in \mathcal{S}_{-i}} \Big(p_t(g_{-i}) - p_t(g_{-i})^2 \Big) \\ &= 1 - \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2 = D_i^t(h^t). \end{aligned}$$

This proves the claim. □

Proof of Lemma 4.2. Fix any $g_{-i} \in \mathcal{S}_{-i} \setminus \{f_{-i}\}$. Write $a^t = (a_i^t, a_{-i}^t)$ for the period-t action profile along the realized play path z, and write h^t for the length-t history $(a^1, \dots, a^{t-1})$.

Because $\mathcal{S}_{-i}$ is finite and all menu strategies are ν-cautious, Bayes' rule is well-defined at every history and the posterior odds admit the standard likelihood ratio form:

$$\frac{\mu_i^t(g_{-i} \mid h^t)}{\mu_i^t(f_{-i} \mid h^t)} = \frac{\mu_i^0(g_{-i})}{\mu_i^0(f_{-i})} \prod_{s=1}^{t-1} \frac{g_{-i}(h^s)(a_{-i}^s)}{f_{-i}(h^s)(a_{-i}^s)}. \tag{19}$$

Define the log-likelihood ratio increments

$$X_s := \log \frac{f_{-i}(h^s)(a_{-i}^s)}{g_{-i}(h^s)(a_{-i}^s)}.$$

Taking logs in (19) gives

$$\log \frac{\mu_i^t(g_{-i} \mid h^t)}{\mu_i^t(f_{-i} \mid h^t)} = \log \frac{\mu_i^0(g_{-i})}{\mu_i^0(f_{-i})} \; - \; \sum_{s=1}^{t-1} X_s. \tag{20}$$

Let $\mathcal{F}_s$ be the σ-algebra generated by the history h^s. Under the true play distribution μ^f, conditional on $\mathcal{F}_s$ the opponents' action a_{-i}^s is distributed according to $f_{-i}(h^s)$. Therefore,

$$\mathbb{E}_{\mu^f}[X_s \mid \mathcal{F}_s] = \sum_{a_{-i} \in A_{-i}} f_{-i}(h^s)(a_{-i}) \log \frac{f_{-i}(h^s)(a_{-i})}{g_{-i}(h^s)(a_{-i})} = D_{\mathrm{KL}}\Big(f_{-i}(h^s) \,\Big\|\, g_{-i}(h^s) \Big).$$

Define the martingale difference sequence $Y_s := X_s - \mathbb{E}[X_s \mid \mathcal{F}_s]$. By ν-caution, for all s we have $f_{-i}(h^s)(a^s_{-i}) \in [\nu, 1]$ and $g_{-i}(h^s)(a^s_{-i}) \in [\nu, 1]$, hence

$$|X_s| \le \log(1/\nu), \qquad \big|\mathbb{E}[X_s \mid \mathcal{F}_s]\big| \le \log(1/\nu), \qquad \text{and thus} \qquad |Y_s| \le 2\log(1/\nu) \;:=\; c.$$

Azuma–Hoeffding yields, for any $\epsilon > 0$,

$$\Pr\left(\left|\sum_{s=1}^{T} Y_s\right| \ge \epsilon T\right) \;\le\; 2\exp\left(-\frac{\epsilon^2 T}{2c^2}\right).$$

The right-hand side is summable in T, so by Borel–Cantelli,

$$\frac{1}{T}\sum_{s=1}^{T} Y_s \;\longrightarrow\; 0 \quad \mu^f\text{-a.s.}$$

Consequently,

$$\frac{1}{T}\sum_{s=1}^{T} X_s = \frac{1}{T}\sum_{s=1}^{T} \mathbb{E}[X_s \mid \mathcal{F}_s] \;+\; o(1) = \frac{1}{T}\sum_{s=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(h^s) \,\Big\|\, g_{-i}(h^s)\Big) \;+\; o(1) \quad \mu^f\text{-a.s.}$$

By the KL-separation part of Assumption 3, the liminf of the empirical averages of these KL terms is strictly positive μ^f-a.s., hence

$$\sum_{s=1}^{t-1} X_s \;\longrightarrow\; +\infty \quad \mu^f\text{-a.s.}$$

Returning to (20), we obtain

$$\log \frac{\mu_i^t(g_{-i} \mid h^t)}{\mu_i^t(f_{-i} \mid h^t)} \;\longrightarrow\; -\infty \quad \mu^f\text{-a.s.},$$

so $\mu_i^t(g_{-i} \mid h^t)/\mu_i^t(f_{-i} \mid h^t) \to 0$ almost surely. Because there are finitely many $g_{-i} \neq f_{-i}$, this implies $\mu_i^t(f_{-i} \mid h^t) \to 1$ and $\max_{g_{-i} \neq f_{-i}} \mu_i^t(g_{-i} \mid h^t) \to 0$ almost surely. □

Proof of Lemma 6.1. Identical to the proof of Lemma 4.2, with (f_{-i}, μ^f) replaced by $(\bar{f}_{-i}, \bar{\mu}^{\sigma,u})$. □

Proof of Proposition 4.3. Along any realized play path z, define $p_t(\cdot) = \mu_i^t(\cdot \mid h^t(z))$ on the finite set $\mathcal{S}_{-i}$ and the associated $D_i^t(h^t(z)) = 1 - \sum_{g_{-i}} p_t(g_{-i})^2$. By Lemma 4.2, $\mu_i^t(f_{-i} \mid h^t(z)) \to 1$ almost surely, hence

$$\sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})^2 \;\ge\; \mu_i^t(f_{-i} \mid h^t(z))^2 \;\longrightarrow\; 1,$$

and therefore $D_i^t(h^t(z)) \to 0$ almost surely.

Fix any $\varepsilon > 0$ and any z in the full-measure event where $D_i^t(h^t(z)) \to 0$. Choose $T_i(z, \varepsilon)$ such that $D_i^t(h^t(z)) \le \varepsilon$ for all $t \ge T_i(z, \varepsilon)$. For each such t, Lemma 4.1 implies that PS-BR at $h^t(z)$ is an ε-best

response to the posterior predictive continuation belief, i.e.,

$$f_i|_{h^t(z)} \in \mathrm{BR}_i^{\varepsilon}(f_{-i}^{i,t}|_{h^t(z)} \mid h^t(z)).$$

This is exactly the asymptotic ε-consistency requirement in Definition 4. □

Proof of Lemma 5.1. Let $\mu^{f^i} \equiv P_i^{0,f_i}$ be the distribution induced by the belief-equivalent profile (f_i, f_{-i}^i) representing the prior predictive. By Assumption 2, $\mu^f \ll \mu^{f^i}$.

By the merging of opinions theorem (Kalai and Lehrer, 1993a; Blackwell and Dubins, 1962), absolute continuity guarantees that the conditional predictive distributions over future play paths merge almost surely in total variation. Specifically, for μ^f-almost every path $z \in H^\infty$:

$$\lim_{t\to\infty} \sup_{E\in\mathcal{B}} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))| = 0,$$

where $\mathcal{B}$ is the product σ-algebra on H^∞.

Recall from Definition 6 that the continuation weak distance is bounded by the total variation distance. For any finite length k, the σ-algebra $\mathcal{B}^k$ generated by cylinder events of length k is a sub-σ-algebra of $\mathcal{B}$. Therefore:

$$\sup_{E\in\mathcal{B}^k} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))| \le \sup_{E\in\mathcal{B}} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))|.$$

Using this bound, the continuation weak distance $d_{h^t(z)}(\mu^f, \mu^{f^i})$ satisfies:

$$\begin{aligned} d_{h^t(z)}(\mu^f, \mu^{f^i}) &= \sum_{k=1}^{\infty} 2^{-k} \sup_{E\in\mathcal{B}^k} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))| \\ &\le \sum_{k=1}^{\infty} 2^{-k} \sup_{E\in\mathcal{B}} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))| \\ &= \sup_{E\in\mathcal{B}} |\mu^f(E \mid C(h^t(z))) - \mu^{f^i}(E \mid C(h^t(z)))|. \end{aligned}$$

Since the total variation distance on the right-hand side converges to zero as $t \to \infty$ for μ^f-almost every z, we have:

$$\lim_{t\to\infty} d_{h^t(z)}(\mu^f, \mu^{f^i}) = 0 \quad \mu^f\text{-a.s.}$$

By the definition of the limit, for any $\eta > 0$, there μ^f-a.s. exists a finite time $T_i(z, \eta)$ such that for all $t \ge T_i(z, \eta)$, $d_{h^t(z)}(\mu^f, \mu^{f^i}) \le \eta$. This precisely satisfies the strong path prediction requirement in Definition 9. □

Proof of Proposition 5.2. Fix $\xi, \eta > 0$. For each player i, RR implies that μ^f-a.s. in z there exists $T_i^{\mathrm{br}}(z)$

such that for all $t \geq T_i^{\mathrm{br}}(z)$,

$$f_i|_{h^t(z)} \in \mathrm{BR}_i^{\xi}(f_{-i}^{i,t}|_{h^t(z)} \mid h^t(z)).$$

By the representative choice (4), we may equivalently write $f_{-i}^{i,t}|_{h^t(z)} \equiv f_{-i}^{i}|_{h^t(z)}$, so for all $t \geq T_i^{\mathrm{br}}(z)$,

$$f_i|_{h^t(z)} \in \mathrm{BR}_i^{\xi}(f_{-i}^{i}|_{h^t(z)} \mid h^t(z)),$$

which is exactly the subjective best-response condition in Definition 8.

Similarly, strong prediction implies that μ^f-a.s. in z there exists $T_i^{\mathrm{pred}}(z)$ such that for all $t \geq T_i^{\mathrm{pred}}(z)$,

$$d_{h^t(z)}(\mu^f, \mu^{f^i}) \leq \eta,$$

which is the weak predictive accuracy condition in Definition 8.

Let $T(z) := \max_i\{T_i^{\mathrm{br}}(z),\, T_i^{\mathrm{pred}}(z)\}$, which is finite μ^f-a.s. since I is finite. Then for all $t \geq T(z)$ and every player i, both conditions in Definition 8 hold with supporting profile f^i, so $f|_{h^t(z)}$ is a weak ξ-subjective η-equilibrium after $h^t(z)$. □

Proof of Theorem 5.3. Fix $\varepsilon > 0$ and set $\xi := \varepsilon/2$. Let $\hat{\eta}(\cdot,\cdot)$ be the function from the infinite patching lemma (Lemma A.4 in Appendix A), and set $\eta := \hat{\eta}(\xi, \varepsilon/2)$.

By Proposition 5.2, μ^f-a.s. in z there exists $T(z)$ such that for all $t \geq T(z)$, the continuation profile $f|_{h^t(z)}$ is a weak ξ-subjective η-equilibrium after $h^t(z)$. Applying Lemma A.4 at each such t yields an ε-Nash equilibrium $\hat{f}^{\varepsilon,t,z}$ of the continuation game after $h^t(z)$ satisfying $d_{h^t(z)}(\mu^f, \mu^{\hat{f}^{\varepsilon,t,z}}) \leq \varepsilon$. □

Proof of Corollary 5.4. By Proposition 4.3, under Assumption 3, each player is RR. Because Assumption 3 (specifically the menu grain of truth) implies Assumption 2, Lemma 5.1 guarantees each player learns to predict the path of play under f. Theorem 5.3 therefore applies. □

Proof of Lemma 6.2. Fix any $m_i \in \mathcal{M}_i \setminus \{u_i\}$. By Bayes' rule (6),

$$\frac{\pi_i^t(m_i \mid x_i^t)}{\pi_i^t(u_i \mid x_i^t)} = \frac{\pi_i^0(m_i)}{\pi_i^0(u_i)} \prod_{s=1}^{t-1} \frac{\psi_i(r_i^s; m_i(a^s))}{\psi_i(r_i^s; u_i(a^s))}.$$

Equivalently,

$$\log \frac{\pi_i^t(m_i \mid x_i^t)}{\pi_i^t(u_i \mid x_i^t)} = \log \frac{\pi_i^0(m_i)}{\pi_i^0(u_i)} - \sum_{s=1}^{t-1} X_s,$$

where

$$X_s := \log \frac{\psi_i(r_i^s; u_i(a^s))}{\psi_i(r_i^s; m_i(a^s))}.$$

Let

$$\mathcal{H}_s := \sigma(h^{s+1}, r_i^{1:s-1}),$$

so that a^s is $\mathcal{H}_s$-measurable and, under the true interaction law, r_i^s is conditionally distributed as $q_i^{u_i}(\cdot \mid a^s)$. Therefore

$$\mathbb{E}[X_s \mid \mathcal{H}_s] = D_{\mathrm{KL}}\Big(q_i^{u_i}(\cdot \mid a^s) \,\Big\|\, q_i^{m_i}(\cdot \mid a^s)\Big).$$

Define the martingale difference sequence

$$Y_s := X_s - \mathbb{E}[X_s \mid \mathcal{H}_s].$$

By Assumption 4(3),

$$\sup_s \mathbb{E}[Y_s^2] < \infty.$$

Hence

$$\sum_{s=1}^{\infty} \frac{\mathbb{E}[Y_s^2]}{s^2} < \infty,$$

so the martingale strong law implies

$$\frac{1}{T}\sum_{s=1}^{T} Y_s \longrightarrow 0 \qquad \text{a.s.}$$

Therefore,

$$\frac{1}{T}\sum_{s=1}^{T} X_s = \frac{1}{T}\sum_{s=1}^{T} D_{\mathrm{KL}}\Big(q_i^{u_i}(\cdot \mid a^s) \,\Big\|\, q_i^{m_i}(\cdot \mid a^s)\Big) + o(1) \qquad \text{a.s.}$$

By Assumption 4(4), the liminf of the empirical KL average is strictly positive almost surely, hence

$$\sum_{s=1}^{t-1} X_s \longrightarrow +\infty \qquad \text{a.s.}$$

It follows that

$$\log \frac{\pi_i^t(m_i \mid x_i^t)}{\pi_i^t(u_i \mid x_i^t)} \longrightarrow -\infty \qquad \text{a.s.,}$$

so

$$\frac{\pi_i^t(m_i \mid x_i^t)}{\pi_i^t(u_i \mid x_i^t)} \longrightarrow 0.$$

Since $\mathcal{M}_i$ is finite, this implies

$$\pi_i^t(u_i \mid x_i^t) \longrightarrow 1 \qquad \text{and} \qquad \max_{m_i \neq u_i} \pi_i^t(m_i \mid x_i^t) \longrightarrow 0$$

almost surely. □

Proof of Lemma 6.3. By (11), for every measurable event $E \subseteq H^\infty$,

$$\left|\Pi_i^t(E \mid x_i^t) - \bar{\mu}_{x_i^t}^{(\sigma_i, g_{-i}^{i,t}), u_i}(E)\right| = \left|\sum_{m_i \in \mathcal{M}_i} \pi_i^t(m_i \mid x_i^t) \bar{\mu}_{x_i^t}^{(\sigma_i, g_{-i}^{i,t}), m_i}(E) - \bar{\mu}_{x_i^t}^{(\sigma_i, g_{-i}^{i,t}), u_i}(E)\right|$$
$$\leq \sum_{m_i \neq u_i} \pi_i^t(m_i \mid x_i^t) = 1 - \pi_i^t(u_i \mid x_i^t).$$

Taking the supremum over cylinder events at each horizon and summing with the weights 2^{-t} yields the stated bound. □

Proof of Lemma 6.4. Fix player i and an information history $x_i^t = (h^t, r_i^{1:t-1})$. Let $\mathcal{M} := \mathcal{S}_{-i} \times \mathcal{M}_i$, and for each $m = (g_{-i}, m_i) \in \mathcal{M}$ define the continuation value functional

$$V_i^m(\tau_i \mid x_i^t) \;:=\; V_i^{m_i}(\tau_i \mid x_i^t; g_{-i}) \;\in\; [0,1],$$

and the value envelope

$$M(m) \;:=\; \sup_{\tau_i} V_i^m(\tau_i \mid x_i^t) \;\in\; [0,1].$$

For each $m \in \mathcal{M}$ fix a (measurable) best response τ_i^m attaining $M(m)$, i.e., $V_i^m(\tau_i^m \mid x_i^t) = M(m)$.

By Definition 13, PS-BR samples $(\tilde{g}_{-i}, \tilde{m}_i) \sim p_t(\cdot)$ and then plays $\tau_i^{(\tilde{g}_{-i}, \tilde{m}_i)}$. Let $\sigma_{i,t}^{\mathrm{PS}}$ denote this randomized continuation strategy at x_i^t.

Because $V_i^{\mathrm{mix},t}$ is linear in both the opponents-mixture and the payoff-matrix mixture, we can write

$$V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) = \sum_{(g_{-i}, m_i) \in \mathcal{M}} p_t(g_{-i}, m_i)\, V_i^{(g_{-i}, m_i)}(\tau_i \mid x_i^t) = \sum_{m \in \mathcal{M}} p_t(m)\, V_i^m(\tau_i \mid x_i^t).$$

Therefore, evaluating PS-BR under the mixed subjective objective gives

$$\begin{aligned} V_i^{\mathrm{mix},t}(\sigma_{i,t}^{\mathrm{PS}} \mid x_i^t) &= \sum_{\tilde{m} \in \mathcal{M}} p_t(\tilde{m})\; V_i^{\mathrm{mix},t}(\tau_i^{\tilde{m}} \mid x_i^t) \\ &= \sum_{\tilde{m} \in \mathcal{M}} p_t(\tilde{m}) \sum_{m \in \mathcal{M}} p_t(m)\, V_i^m(\tau_i^{\tilde{m}} \mid x_i^t) \\ &\geq \sum_{m \in \mathcal{M}} p_t(m)^2\, V_i^m(\tau_i^m \mid x_i^t) = \sum_{m \in \mathcal{M}} p_t(m)^2\, M(m). \end{aligned}$$

On the other hand,

$$\sup_{\tau_i} V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) = \sup_{\tau_i} \sum_{m \in \mathcal{M}} p_t(m)\, V_i^m(\tau_i \mid x_i^t) \leq \sum_{m \in \mathcal{M}} p_t(m) \sup_{\tau_i} V_i^m(\tau_i \mid x_i^t) = \sum_{m \in \mathcal{M}} p_t(m)\, M(m).$$

Subtracting and using $M(m) \le 1$ for all m,

$$\begin{aligned}\sup_{\tau_i} V_i^{\mathrm{mix},t}(\tau_i \mid x_i^t) - V_i^{\mathrm{mix},t}(\sigma_{i,t}^{\mathrm{PS}} \mid x_i^t) &\le \sum_{m\in\mathcal{M}} \big(p_t(m) - p_t(m)^2\big) M(m) \\ &\le \sum_{m\in\mathcal{M}} \big(p_t(m) - p_t(m)^2\big) = 1 - \sum_{m\in\mathcal{M}} p_t(m)^2 = D_i^{t,\mathrm{joint}}(x_i^t).\end{aligned}$$

This proves the claim. □

Proof of Proposition 6.5. Work on the full-measure event on which both posterior concentrations hold:

$$\mu_i^t(\bar{f}_{-i} \mid h^t) \to 1 \qquad \text{and} \qquad \pi_i^t(u_i \mid x_i^t) \to 1.$$

Then

$$D_i^{t,\mathrm{joint}}(x_i^t) \to 0 \qquad \text{and} \qquad \delta_i^t(x_i^t) := 1 - \pi_i^t(u_i \mid x_i^t) \to 0.$$

By (12), Lemma 6.4, and (10),

$$\begin{aligned}\sup_{\tau_i \in \Sigma_i(x_i^t)} V_i^{u_i}(\tau_i \mid x_i^t; g_{-i}^{i,t}) - V_i^{u_i}(\sigma_{i,t}^{\mathrm{PS}} \mid x_i^t; g_{-i}^{i,t}) &= \sup_{\tau_i} V_i^{u_i,t}(\tau_i \mid x_i^t) - V_i^{u_i,t}(\sigma_{i,t}^{\mathrm{PS}} \mid x_i^t) \\ &\le D_i^{t,\mathrm{joint}}(x_i^t) + 2\delta_i^t(x_i^t).\end{aligned}$$

The right-hand side converges to 0 almost surely, so the stated eventual ε-best-response property follows. □

Proof of Lemma 6.6. By the Blackwell–Dubins merging argument applied on the observable process O_i, Assumption 5 implies

$$d\Big(\Pi_i^t(\cdot \mid x_i^t(\omega)), \bar{\mu}_{i,x_i^t(\omega)}^{\sigma,u}\Big) \longrightarrow 0 \qquad \text{for } P^{\sigma,u}\text{-a.e. } \omega.$$

Assumption 6 gives

$$d\Big(\bar{\mu}_{i,x_i^t(\omega)}^{\sigma,u}, \bar{\mu}_{x^t(\omega)}^{\sigma,u}\Big) \longrightarrow 0 \qquad \text{for } P^{\sigma,u}\text{-a.e. } \omega.$$

The claim follows by the triangle inequality. □

Proof of Proposition 6.7. Fix $\xi, \eta > 0$. For each player i, Proposition 6.5 implies that $P^{\sigma,u}$-a.s. there exists $T_i^{\mathrm{br}}(\omega)$ such that for all $t \ge T_i^{\mathrm{br}}(\omega)$,

$$\sigma_{i,t}^{\mathrm{PS}}(\cdot \mid x_i^t(\omega)) \in \mathrm{BR}_{i,u_i}^{\xi}(g_{-i}^{i,t} \mid x_i^t(\omega)).$$

Also, Lemma 6.6 together with Lemma 6.3 implies that $P^{\sigma,u}$-a.s. there exists $T_i^{\mathrm{pred}}(\omega)$ such that for all $t \ge T_i^{\mathrm{pred}}(\omega)$,

$$d\Big(\bar{\mu}_{x^t(\omega)}^{\sigma,u}, \bar{\mu}_{x_i^t(\omega)}^{(\sigma_i, g_{-i}^{i,t}),u_i}\Big) \le \eta.$$

Indeed,

$$d\Big(\bar{\mu}^{\sigma,u}_{x^t(\omega)}, \bar{\mu}^{(\sigma_i,g^{i,t}_{-i}),u_i}_{x^t_i(\omega)}\Big) \le d\Big(\bar{\mu}^{\sigma,u}_{x^t(\omega)}, \Pi^t_i(\cdot \mid x^t_i(\omega))\Big) + d\Big(\Pi^t_i(\cdot \mid x^t_i(\omega)), \bar{\mu}^{(\sigma_i,g^{i,t}_{-i}),u_i}_{x^t_i(\omega)}\Big),$$

and both terms vanish almost surely by Lemmas 6.6 and 6.3.

Let

$$T(\omega) := \max_{i\in I}\{T^{\mathrm{br}}_i(\omega), T^{\mathrm{pred}}_i(\omega)\}.$$

Then for all $t \ge T(\omega)$ and every player i, both conditions in Definition 14 hold with supporting reduced-form model $g^{i,t}_{-i}$. □

Proof of Lemma 5.5. Fix player i, let $p, q \in \Delta(A_{-i})$, and suppose $\alpha_i \in \mathrm{br}^{\xi}_i(q)$.

For any $\alpha_i \in \Delta(A_i)$ define

$$\phi_{\alpha_i}(a_{-i}) := \sum_{a_i\in A_i} \alpha_i(a_i)\, u_i(a_i, a_{-i}), \qquad a_{-i} \in A_{-i}.$$

Since $u_i(a_i, a_{-i}) \in [0,1]$, we have $\phi_{\alpha_i}(a_{-i}) \in [0,1]$ for all $a_{-i} \in A_{-i}$. Also,

$$u_i(\alpha_i, p) - u_i(\alpha_i, q) = \sum_{a_{-i}\in A_{-i}} \phi_{\alpha_i}(a_{-i})\big(p(a_{-i}) - q(a_{-i})\big).$$

Set

$$S^+ := \{a_{-i} \in A_{-i} : p(a_{-i}) \ge q(a_{-i})\}.$$

Because $0 \le \phi_{\alpha_i} \le 1$, we have

$$\begin{aligned} u_i(\alpha_i, p) - u_i(\alpha_i, q) &= \sum_{a_{-i}\in S^+} \phi_{\alpha_i}(a_{-i})\big(p(a_{-i}) - q(a_{-i})\big) + \sum_{a_{-i}\notin S^+} \phi_{\alpha_i}(a_{-i})\big(p(a_{-i}) - q(a_{-i})\big) \\ &\le \sum_{a_{-i}\in S^+} \big(p(a_{-i}) - q(a_{-i})\big) \\ &= p(S^+) - q(S^+) \\ &\le \|p - q\|_{\mathrm{TV}}. \end{aligned}$$

Applying the same argument with p and q interchanged yields

$$u_i(\alpha_i, q) - u_i(\alpha_i, p) \le \|p - q\|_{\mathrm{TV}}.$$

Therefore

$$|u_i(\alpha_i, p) - u_i(\alpha_i, q)| \le \|p - q\|_{\mathrm{TV}} \qquad \text{for every } \alpha_i \in \Delta(A_i). \tag{21}$$

Now suppose $\alpha_i \in \mathrm{br}_i^{\xi}(q)$. Then

$$u_i(\alpha_i, q) \geq \sup_{\alpha_i' \in \Delta(A_i)} u_i(\alpha_i', q) - \xi.$$

Using (21),

$$\begin{aligned} u_i(\alpha_i, p) &\geq u_i(\alpha_i, q) - \|p - q\|_{\mathrm{TV}} \\ &\geq \sup_{\alpha_i' \in \Delta(A_i)} u_i(\alpha_i', q) - \xi - \|p - q\|_{\mathrm{TV}} \\ &\geq \sup_{\alpha_i' \in \Delta(A_i)} \left(u_i(\alpha_i', p) - \|p - q\|_{\mathrm{TV}}\right) - \xi - \|p - q\|_{\mathrm{TV}} \\ &= \sup_{\alpha_i' \in \Delta(A_i)} u_i(\alpha_i', p) - \xi - 2\|p - q\|_{\mathrm{TV}}. \end{aligned}$$

Hence

$$\alpha_i \in \mathrm{br}_i^{\xi + 2\|p-q\|_{\mathrm{TV}}}(p).$$

□

Proof of Lemma 5.6. Fix player i and history h^t. For each $g_{-i} \in \mathcal{S}_{-i}$ define

$$M(g_{-i}) := \sup_{\alpha_i \in \Delta(A_i)} u_i\big(\alpha_i, g_{-i}(h^t)\big) \in [0, 1].$$

By Definition 11, for each $g_{-i} \in \mathcal{S}_{-i}$ we have chosen

$$\alpha_i^{g_{-i}, h^t} \in \mathrm{br}_i\big(g_{-i}(h^t)\big),$$

so

$$u_i\big(\alpha_i^{g_{-i}, h^t}, g_{-i}(h^t)\big) = M(g_{-i}).$$

Write $p_t(g_{-i}) = \mu_i^t(g_{-i} \mid h^t)$. The ex ante mixed action induced by myopic PS-BR is

$$\alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t) = \sum_{\tilde{g}_{-i} \in \mathcal{S}_{-i}} p_t(\tilde{g}_{-i})\, \alpha_i^{\tilde{g}_{-i}, h^t}(\cdot),$$

and the one-step posterior predictive belief is

$$q_i^t(\cdot \mid h^t) = \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})\, g_{-i}(h^t)(\cdot).$$

By bilinearity of $u_i(\cdot,\cdot)$,

$$\begin{aligned} u_i\big(\alpha_{i,t}^{\mathrm{mPS}}, q_i^t\big) &= \sum_{\tilde{g}_{-i}\in\mathcal{S}_{-i}} p_t(\tilde{g}_{-i}) \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})\, u_i\big(\alpha_i^{\tilde{g}_{-i},h^t}, g_{-i}(h^t)\big) \\ &\geq \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})^2\, u_i\big(\alpha_i^{g_{-i},h^t}, g_{-i}(h^t)\big) \\ &= \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})^2\, M(g_{-i}). \end{aligned}$$

On the other hand, again by bilinearity,

$$\begin{aligned} \sup_{\alpha_i\in\Delta(A_i)} u_i(\alpha_i, q_i^t) &= \sup_{\alpha_i\in\Delta(A_i)} \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})\, u_i\big(\alpha_i, g_{-i}(h^t)\big) \\ &\leq \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i}) \sup_{\alpha_i\in\Delta(A_i)} u_i\big(\alpha_i, g_{-i}(h^t)\big) \\ &= \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})\, M(g_{-i}). \end{aligned}$$

Subtracting,

$$\begin{aligned} \sup_{\alpha_i} u_i(\alpha_i, q_i^t) - u_i(\alpha_{i,t}^{\mathrm{mPS}}, q_i^t) &\leq \sum_{g_{-i}\in\mathcal{S}_{-i}} \big(p_t(g_{-i}) - p_t(g_{-i})^2\big)\, M(g_{-i}) \\ &\leq \sum_{g_{-i}\in\mathcal{S}_{-i}} \big(p_t(g_{-i}) - p_t(g_{-i})^2\big) \\ &= 1 - \sum_{g_{-i}\in\mathcal{S}_{-i}} p_t(g_{-i})^2 \\ &= D_i^t(h^t). \end{aligned}$$

This proves the claim. □

Proof of Lemma 5.7. Fix player i and let $f^i = (f_i, f_{-i}^i)$ be the supporting profile from Definition 9. Fix a realized path $z \in H^\infty$ in the full-measure event from Definition 9. By definition of q_i^t and the representative choice (4),

$$q_i^t(\cdot \mid h^t(z)) = f_{-i}^{i,t}(h^t(z)) = f_{-i}^i(h^t(z)).$$

Let $\eta > 0$. By Definition 9, there exists $T_i(z, \eta/2) < \infty$ such that for all $t \geq T_i(z, \eta/2)$,

$$d_{h^t(z)}(\mu^f, \mu^{f^i}) \leq \eta/2.$$

Fix such a t. For any subset $B \subseteq A_{-i}$, define the one-step cylinder event

$$E_B := \{y \in H^\infty : \ y^1_{-i} \in B\} \in \mathcal{B}^1.$$

By the definition of continuation measures,

$$\mu^f_{h^t(z)}(E_B) = f_{-i}(h^t(z))(B), \qquad \mu^{f^i}_{h^t(z)}(E_B) = f^i_{-i}(h^t(z))(B) = q^t_i(B \mid h^t(z)).$$

Therefore,

$$\begin{aligned}
\left\|q^t_i(\cdot \mid h^t(z)) - f_{-i}(h^t(z))\right\|_{\mathrm{TV}} &= \sup_{B \subseteq A_{-i}} \left|q^t_i(B \mid h^t(z)) - f_{-i}(h^t(z))(B)\right| \\
&= \sup_{B \subseteq A_{-i}} \left|\mu^{f^i}_{h^t(z)}(E_B) - \mu^f_{h^t(z)}(E_B)\right| \\
&\leq \sup_{E \in \mathcal{B}^1} \left|\mu^{f^i}_{h^t(z)}(E) - \mu^f_{h^t(z)}(E)\right|.
\end{aligned}$$

By Definition 6,

$$d_{h^t(z)}(\mu^f, \mu^{f^i}) = \sum_{k=1}^{\infty} 2^{-k} \sup_{E \in \mathcal{B}^k} \left|\mu^f_{h^t(z)}(E) - \mu^{f^i}_{h^t(z)}(E)\right|.$$

In particular,

$$\frac{1}{2} \sup_{E \in \mathcal{B}^1} \left|\mu^f_{h^t(z)}(E) - \mu^{f^i}_{h^t(z)}(E)\right| \leq d_{h^t(z)}(\mu^f, \mu^{f^i}),$$

so

$$\sup_{E \in \mathcal{B}^1} \left|\mu^f_{h^t(z)}(E) - \mu^{f^i}_{h^t(z)}(E)\right| \leq 2\, d_{h^t(z)}(\mu^f, \mu^{f^i}) \leq \eta.$$

Hence

$$\left\|q^t_i(\cdot \mid h^t(z)) - f_{-i}(h^t(z))\right\|_{\mathrm{TV}} \leq \eta$$

for all $t \geq T_i(z, \eta/2)$. Since $\eta > 0$ was arbitrary, this proves the claim. □

Proof of Theorem 5.8. Fix $\varepsilon > 0$ and set $\xi := \varepsilon/3$.

For player i, Assumption 3 implies, by Lemma 4.2, that there is a full-measure event on which

$$\mu^t_i(f_{-i} \mid h^t(z)) \longrightarrow 1.$$

Since $f_{-i} \in \mathcal{S}_{-i}$ by menu grain of truth, on that event we also have

$$D^t_i(h^t(z)) = 1 - \sum_{g_{-i} \in \mathcal{S}_{-i}} \mu^t_i(g_{-i} \mid h^t(z))^2 \longrightarrow 0.$$

Therefore there exists $T_i^{\mathrm{br}}(z) < \infty$ such that for all $t \geq T_i^{\mathrm{br}}(z)$,

$$D_i^t(h^t(z)) \leq \xi.$$

Because player i uses myopic PS-BR, we have

$$f_i(h^t(z)) = \alpha_{i,t}^{\mathrm{mPS}}(\cdot \mid h^t(z)).$$

Applying Lemma 5.6, it follows that for all $t \geq T_i^{\mathrm{br}}(z)$,

$$f_i(h^t(z)) \in \mathrm{br}_i^{\xi}(q_i^t(\cdot \mid h^t(z))).$$

Next, write

$$p_t(g_{-i}) = \mu_i^t(g_{-i} \mid h^t(z)).$$

At history $h^t(z)$,

$$q_i^t(\cdot \mid h^t(z)) = \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})\, g_{-i}(h^t(z))(\cdot).$$

For any $B \subseteq A_{-i}$,

$$\begin{aligned}
\left| q_i^t(B \mid h^t(z)) - f_{-i}(h^t(z))(B) \right| &= \left| \sum_{g_{-i} \in \mathcal{S}_{-i}} p_t(g_{-i})\, g_{-i}(h^t(z))(B) - f_{-i}(h^t(z))(B) \right| \\
&= \left| \sum_{g_{-i} \neq f_{-i}} p_t(g_{-i}) \big(g_{-i}(h^t(z))(B) - f_{-i}(h^t(z))(B) \big) \right| \\
&\leq \sum_{g_{-i} \neq f_{-i}} p_t(g_{-i}) \\
&= 1 - \mu_i^t(f_{-i} \mid h^t(z)).
\end{aligned}$$

Taking the supremum over $B \subseteq A_{-i}$ gives

$$\left\| q_i^t(\cdot \mid h^t(z)) - f_{-i}(h^t(z)) \right\|_{\mathrm{TV}} \leq 1 - \mu_i^t(f_{-i} \mid h^t(z)) \longrightarrow 0.$$

Hence there exists $T_i^{\mathrm{pred}}(z) < \infty$ such that for all $t \geq T_i^{\mathrm{pred}}(z)$,

$$\left\| q_i^t(\cdot \mid h^t(z)) - f_{-i}(h^t(z)) \right\|_{\mathrm{TV}} \leq \xi.$$

Now fix $t \geq \max\{T_i^{\mathrm{br}}(z), T_i^{\mathrm{pred}}(z)\}$. We already know that

$$f_i(h^t(z)) \in \mathrm{br}_i^{\xi}(q_i^t(\cdot \mid h^t(z))),$$

and that

$$\|q_i^t(\cdot \mid h^t(z)) - f_{-i}(h^t(z))\|_{\mathrm{TV}} \le \xi.$$

Applying Lemma 5.5 with $p = f_{-i}(h^t(z))$ and $q = q_i^t(\cdot \mid h^t(z))$ yields

$$f_i(h^t(z)) \in \mathrm{br}_i^{\xi+2\xi}(f_{-i}(h^t(z))) = \mathrm{br}_i^{\varepsilon}(f_{-i}(h^t(z))).$$

Intersect the full-measure events above over all players $i \in I$. Since I is finite, on that intersection we may define

$$T(z) := \max_{i \in I} \max\{T_i^{\mathrm{br}}(z), T_i^{\mathrm{pred}}(z)\} < \infty.$$

Then for all $t \ge T(z)$ and all players i,

$$f_i(h^t(z)) \in \mathrm{br}_i^{\varepsilon}(f_{-i}(h^t(z))).$$

By Definition 10, this means that $f(h^t(z))$ is a stage ε-Nash equilibrium for all $t \ge T(z)$. □

Proof of Lemma 5.9. Fix player i and let $f^i = (f_i, f_{-i}^i)$ be the supporting profile from Definition 9. Fix a realized path z in the full-measure event from Definition 9. By definition of q_i^t and the representative choice (4),

$$q_i^t(\cdot \mid h^t(z)) = f_{-i}^{i,t}(h^t(z)) = f_{-i}^i(h^t(z)).$$

For each t, define the one-step cylinder event

$$E_t(z) := \{y \in H^\infty : \ y_{-i}^1 = a_{-i}^\star(h^t(z))\} \in \mathcal{B}^1.$$

Because the true opponents' next action at history $h^t(z)$ is pure,

$$f_{-i}(h^t(z)) = \delta_{a_{-i}^\star(h^t(z))},$$

so

$$\mu_{h^t(z)}^f(E_t(z)) = 1.$$

Also, by the on-path identification above,

$$\mu_{h^t(z)}^{f^i}(E_t(z)) = f_{-i}^i(h^t(z))(a_{-i}^\star(h^t(z))) = q_i^t(a_{-i}^\star(h^t(z)) \mid h^t(z)).$$

Hence

$$1 - q_i^t(a_{-i}^\star(h^t(z)) \mid h^t(z)) = \left|\mu_{h^t(z)}^f(E_t(z)) - \mu_{h^t(z)}^{f^i}(E_t(z))\right|$$

$$\leq \sup_{E \in \mathcal{B}^1} \left| \mu^f_{h^t(z)}(E) - \mu^{f^i}_{h^t(z)}(E) \right|.$$

As in the proof of Lemma 5.7,

$$\sup_{E \in \mathcal{B}^1} \left| \mu^f_{h^t(z)}(E) - \mu^{f^i}_{h^t(z)}(E) \right| \leq 2\, d_{h^t(z)}(\mu^f, \mu^{f^i}).$$

Because player i learns to predict the path of play,

$$d_{h^t(z)}(\mu^f, \mu^{f^i}) \longrightarrow 0.$$

Therefore

$$q_i^t(a^\star_{-i}(h^t(z)) \mid h^t(z)) \longrightarrow 1.$$

It follows immediately that

$$1 - \max_{a_{-i} \in A_{-i}} q_i^t(a_{-i} \mid h^t(z)) \leq 1 - q_i^t(a^\star_{-i}(h^t(z)) \mid h^t(z)) \longrightarrow 0,$$

which proves asymptotic purity.

Finally, because

$$q_i^t(a^\star_{-i}(h^t(z)) \mid h^t(z)) \longrightarrow 1,$$

there exists $T_i(z) < \infty$ such that for all $t \geq T_i(z)$,

$$q_i^t(a^\star_{-i}(h^t(z)) \mid h^t(z)) > \frac{1}{2}.$$

For such t, the action $a^\star_{-i}(h^t(z))$ is the unique maximizer of $q_i^t(\cdot \mid h^t(z))$, because all other probabilities sum to

$$1 - q_i^t(a^\star_{-i}(h^t(z)) \mid h^t(z)) < \frac{1}{2}.$$

Hence the deterministic MAP selector must satisfy

$$\hat{a}^t_{-i}(h^t(z)) = a^\star_{-i}(h^t(z)) \qquad \text{for all } t \geq T_i(z).$$

This proves the claim. □

Proof of Theorem 5.10. Because every player $j \in I$ uses deterministic MAP-SCoT, for every history $h \in H$ we have

$$f_j(h) = \delta_{a^\star_j(h)} \qquad \text{for some } a^\star_j(h) \in A_j.$$

Hence for every player i and every history h,

$$f_{-i}(h) = \delta_{a^\star_{-i}(h)} \qquad \text{for some } a^\star_{-i}(h) \in A_{-i}.$$

For each player i, apply Lemma 5.9. There is a full-measure event on which there exists $T_i(z) < \infty$ such that for all $t \geq T_i(z)$,

$$\hat{a}^t_{-i}(h^t(z)) = a^\star_{-i}(h^t(z)).$$

Because the player set I is finite, the intersection of these full-measure events over all players still has measure one.

Fix a realized path z in that intersection. For any player i and any $t \geq T_i(z)$, Definition 12 gives

$$f_i(h^t(z)) = \delta_{b_i(\hat{a}^t_{-i}(h^t(z)))} = \delta_{b_i(a^\star_{-i}(h^t(z)))}.$$

By definition of the pure best-response selector b_i,

$$b_i(a_{-i}) \in \arg\max_{a_i \in A_i} u_i(a_i, a_{-i}) \qquad \text{for every } a_{-i} \in A_{-i}.$$

Therefore

$$\delta_{b_i(a^\star_{-i}(h^t(z)))} \in \mathrm{br}_i\big(\delta_{a^\star_{-i}(h^t(z))}\big) = \mathrm{br}_i\big(f_{-i}(h^t(z))\big).$$

So for every player i and all $t \geq T_i(z)$,

$$f_i(h^t(z)) \in \mathrm{br}_i\big(f_{-i}(h^t(z))\big).$$

Define

$$T(z) := \max_{i \in I} T_i(z) < \infty.$$

Then for all $t \geq T(z)$ and every player i,

$$f_i(h^t(z)) \in \mathrm{br}_i\big(f_{-i}(h^t(z))\big).$$

By Definition 10, this means that $f(h^t(z))$ is a stage Nash equilibrium for all $t \geq T(z)$. □

Proof of Corollary 5.11. By Lemma 5.1, Assumption 2 implies that every player learns to predict the path of play under f in the sense of Definition 9. Theorem 5.10 therefore applies directly. □

C Bounded-memory strategies and finite-state reduction

Many practical agent policies (including menu-based planners) depend only on a bounded window of recent interaction. Following the bounded-recall restriction in Norman (2022), we formalize this as a *bounded-memory* condition.

For a history $h = (a^1, \dots, a^{t-1}) \in H$ let $|h| := t - 1$ denote its length. For $\kappa \in \mathbb{N}$, define

$$\mathrm{suffix}_\kappa(h) \;:=\; (a^{t-\min\{\kappa,t-1\}}, \dots, a^{t-1}) \in \bigcup_{m=0}^{\kappa} A^m,$$

i.e., the last $\min\{\kappa, |h|\}$ joint actions of h (with $\mathrm{suffix}_\kappa(\emptyset) = \emptyset$).

Definition 16 (κ-memory (bounded-recall) strategy)**.** A strategy $f_i : H \to \Delta(A_i)$ has *memory at most* κ if for all histories $h, h' \in H$,

$$\mathrm{suffix}_\kappa(h) = \mathrm{suffix}_\kappa(h') \quad \Longrightarrow \quad f_i(h) = f_i(h').$$

Let $\mathcal{F}_i^\kappa \subseteq \mathcal{F}_i$ denote the set of κ-memory strategies for player i, and write $\mathcal{F}^\kappa := \prod_{i \in I} \mathcal{F}_i^\kappa$.

Let

$$\mathsf{S}_\kappa \;:=\; \bigcup_{m=0}^{\kappa} A^m$$

be the finite set of action-suffixes of length at most κ. Define the deterministic state update map $T_\kappa : \mathsf{S}_\kappa \times A \to \mathsf{S}_\kappa$ by

$$T_\kappa(s, a) \;:=\; \mathrm{suffix}_\kappa((s, a)),$$

i.e., append the new joint action a to the suffix s and keep the last κ entries. For any play path $z = (a^1, a^2, \dots) \in H^\infty$, define the induced memory state at time t:

$$s^t(z) \;:=\; \mathrm{suffix}_\kappa(h^t(z)) \;\in\; \mathsf{S}_\kappa.$$

Lemma C.1 (Finite-state Markov property under bounded memory)**.** *If $f \in \mathcal{F}^\kappa$, then for every $t \geq 1$ and every history h^t with $s = \mathrm{suffix}_\kappa(h^t)$, the next-period action distribution depends on h^t only through s:*

$$\mu^f(a^t = a \mid h^t) \;=\; \prod_{i \in I} f_i(s)(a_i).$$

Moreover, the induced state process satisfies $s^{t+1} = T_\kappa(s^t, a^t)$ almost surely, so $(s^t)_{t \geq 1}$ is a time-homogeneous Markov chain on S_κ.

Proof. Fix t and history h^t. By Definition 2,

$$\mu^f(a^t = a \mid h^t) = \prod_{i \in I} f_i(h^t)(a_i).$$

If $f \in \mathcal{F}^\kappa$, then $f_i(h^t) = f_i(\mathrm{suffix}_\kappa(h^t)) = f_i(s)$ for each i, giving the displayed equality. The state update is deterministic by construction of T_κ: $s^{t+1} = \mathrm{suffix}_\kappa(h^{t+1}) = \mathrm{suffix}_\kappa((h^t, a^t)) = T_\kappa(\mathrm{suffix}_\kappa(h^t), a^t) = T_\kappa(s^t, a^t)$. Thus (s^t) is Markov with kernel induced by the conditional law of a^t given s^t. □

Lemma C.2 (Continuation distributions depend only on the memory state)**.** *Let $g \in \mathcal{F}^\kappa$ and let $h, h' \in H$ satisfy* $\mathrm{suffix}_\kappa(h) = \mathrm{suffix}_\kappa(h')$. *Then the continuation play-path distributions coincide:*

$$\mu^g_h \;=\; \mu^g_{h'}.$$

Proof. By Lemma C.1, the conditional distribution of the next action profile and all future evolution under g depends on the past only through the current memory state $s = \mathrm{suffix}_\kappa(\cdot)$. Since h and h' induce the same state, the induced kernels for $(a^t, a^{t+1}, \dots)$ are identical from either starting history. Therefore the induced continuation measures coincide. □

C.1 Best responses to bounded-memory opponents are bounded-memory

A key benefit of bounded-memory opponents is that each player faces a finite-state discounted MDP in the continuation game. In particular, the best-response search in $\mathrm{BR}_i^\varepsilon(g_{-i} \mid h^t)$ can be restricted without loss to bounded-memory policies.

Lemma C.3 (Markovian best responses to κ-memory opponents)**.** *Fix player i, a history h^t, and an opponents' continuation profile $g_{-i} \in \mathcal{F}^\kappa_{-i}$. Then there exists a best response $\sigma_i^\star \in \mathrm{BR}_i(g_{-i} \mid h^t)$ that is* stationary Markov *with respect to the memory state. That is, there exists a map $\pi_i : \mathsf{S}_\kappa \to \Delta(A_i)$ such that for every continuation history $\bar{h} \succeq h^t$,*

$$\sigma_i^\star(\bar{h}) \;=\; \pi_i(\mathrm{suffix}_\kappa(\bar{h})).$$

Consequently, for every $\varepsilon \geq 0$,

$$\sup_{\sigma_i \in \mathcal{F}_i(h^t)} V_i(\sigma_i \mid h^t; g_{-i}) \;=\; \sup_{\sigma_i \in \mathcal{F}_i^\kappa(h^t)} V_i(\sigma_i \mid h^t; g_{-i}),$$

and $\mathrm{BR}_i(g_{-i} \mid h^t) \cap \mathcal{F}_i^\kappa(h^t) \neq \emptyset$.

Proof. Let $s_0 := \mathrm{suffix}_\kappa(h^t) \in \mathsf{S}_\kappa$. Fix $g_{-i} \in \mathcal{F}^\kappa_{-i}$. Define a controlled Markov process on S_κ as follows. In state s, the player chooses $a_i \in A_i$, the opponents' joint action is drawn as $a_{-i} \sim g_{-i}(s) \in \Delta(A_{-i})$, the stage payoff is $u_i(a_i, a_{-i})$, and the next state is $s' = T_\kappa(s, (a_i, a_{-i}))$.

For any bounded function $v : \mathsf{S}_\kappa \to \mathbb{R}$, define the Bellman operator $\mathcal{T}$ by

$$(\mathcal{T}v)(s) := \max_{\alpha \in \Delta(A_i)} \mathbb{E}_{\substack{a_i \sim \alpha \\ a_{-i} \sim g_{-i}(s)}} \Big[(1-\lambda_i)\, u_i(a_i, a_{-i}) + \lambda_i\, v\big(T_\kappa(s, (a_i, a_{-i}))\big) \Big].$$

Because $\lambda_i \in (0,1)$, $\mathcal{T}$ is a contraction in $\|\cdot\|_\infty$: for any v, w and any s,

$$|(\mathcal{T}v)(s) - (\mathcal{T}w)(s)| \le \max_{\alpha} \mathbb{E}\big[\lambda_i\, |v(s') - w(s')|\big] \le \lambda_i \|v - w\|_\infty.$$

Hence $\mathcal{T}$ has a unique fixed point $V^\star : \mathsf{S}_\kappa \to \mathbb{R}$.

For each s, the maximization over $\alpha \in \Delta(A_i)$ attains its maximum because $\Delta(A_i)$ is compact and the objective is continuous and linear in α. Fix a maximizer $\pi_i(s) \in \Delta(A_i)$ for each s and define the associated *policy evaluation* operator

$$(\mathcal{T}_{\pi_i} v)(s) := \mathbb{E}_{\substack{a_i \sim \pi_i(s) \\ a_{-i} \sim g_{-i}(s)}} \Big[(1-\lambda_i)\, u_i(a_i, a_{-i}) + \lambda_i\, v\big(T_\kappa(s, (a_i, a_{-i}))\big) \Big].$$

Then $(\mathcal{T}_{\pi_i} V^\star)(s) = (\mathcal{T}V^\star)(s) = V^\star(s)$ for all s, so $V^\star$ is a fixed point of $\mathcal{T}_{\pi_i}$. Since $\mathcal{T}_{\pi_i}$ is also a λ_i-contraction, its fixed point is unique; denote it by V^{π_i}. We conclude $V^{\pi_i} = V^\star$.

Now define $\sigma_i^\star$ to be the stationary Markov continuation strategy induced by π_i, i.e. $\sigma_i^\star(\bar{h}) = \pi_i(\mathrm{suffix}_\kappa(\bar{h}))$ for all $\bar{h} \succeq h^t$. By construction, the induced continuation value from h^t is $V_i(\sigma_i^\star \mid h^t; g_{-i}) = V^\star(s_0)$.

It remains to show optimality against *all* continuation strategies, including those with unbounded memory. Let σ_i be any continuation strategy and define its *statewise value envelope*

$$W_{\sigma_i}(s) \;:=\; \sup \Big\{ V_i(\sigma_i \mid \bar{h}; g_{-i}) : \bar{h} \succeq h^t,\ \mathrm{suffix}_\kappa(\bar{h}) = s \Big\}.$$

Fix any s and $\epsilon > 0$, and choose $\bar{h}$ with $\mathrm{suffix}_\kappa(\bar{h}) = s$ and $V_i(\sigma_i \mid \bar{h}; g_{-i}) \ge W_{\sigma_i}(s) - \epsilon$. Let $\alpha := \sigma_i(\bar{h}) \in \Delta(A_i)$ be the first-step mixed action. Conditioning on the first joint action (a_i, a_{-i}) and using that the next state is $s' = T_\kappa(s, (a_i, a_{-i}))$, we have

$$\begin{aligned} V_i(\sigma_i \mid \bar{h}; g_{-i}) &= \mathbb{E}\Big[(1-\lambda_i) u_i(a_i, a_{-i}) + \lambda_i\, V_i(\sigma_i \mid (\bar{h}, (a_i, a_{-i})); g_{-i}) \Big] \\ &\le \mathbb{E}\Big[(1-\lambda_i) u_i(a_i, a_{-i}) + \lambda_i\, W_{\sigma_i}(s') \Big]. \end{aligned}$$

Therefore,

$$W_{\sigma_i}(s) - \epsilon \le \mathbb{E}_{\substack{a_i \sim \alpha \\ a_{-i} \sim g_{-i}(s)}} \Big[(1-\lambda_i) u_i(a_i, a_{-i}) + \lambda_i W_{\sigma_i}(T_\kappa(s, (a_i, a_{-i}))) \Big] \le (\mathcal{T} W_{\sigma_i})(s).$$

Letting $\epsilon \downarrow 0$ gives $W_{\sigma_i} \le \mathcal{T} W_{\sigma_i}$ pointwise. By monotonicity of $\mathcal{T}$ and contraction, iterating yields $W_{\sigma_i} \le \mathcal{T}^n W_{\sigma_i}$ for all n, and $\mathcal{T}^n W_{\sigma_i} \to V^\star$ uniformly as $n \to \infty$. Hence $W_{\sigma_i}(s) \le V^\star(s)$ for all s, and in

particular

$$V_i(\sigma_i \mid h^t; g_{-i}) \;\le\; W_{\sigma_i}(s_0) \;\le\; V^\star(s_0) \;=\; V_i(\sigma_i^\star \mid h^t; g_{-i}).$$

Thus $\sigma_i^\star$ is a best response. The final displayed equality of suprema follows because an optimal policy exists within $\mathcal{F}_i^\kappa(h^t)$. □

C.2 A checkable KL-separation condition under bounded memory

Assumption 3-(3) (on-path KL separation) is stated for general history-dependent strategies. Under bounded memory, it reduces to a state-frequency condition.

Lemma C.4 (State-frequency decomposition of on-path KL averages)**.** *Fix player i, $\kappa \in \mathbb{N}$, and $f_{-i}, g_{-i} \in \mathcal{F}_{-i}^\kappa$. For a realized path z, define $s^t(z) = \mathrm{suffix}_\kappa(h^t(z))$ and empirical state frequencies*

$$\hat{\pi}_T^z(s) \;:=\; \frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{s^t(z) = s\}, \qquad s \in \mathsf{S}_\kappa.$$

Then for every T and every z,

$$\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(h^t(z)) \,\Big\|\, g_{-i}(h^t(z))\Big) \;=\; \sum_{s\in\mathsf{S}_\kappa} \hat{\pi}_T^z(s)\, D_{\mathrm{KL}}\Big(f_{-i}(s) \,\Big\|\, g_{-i}(s)\Big).$$

In particular, for any fixed state s,

$$\liminf_{T\to\infty} \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(h^t(z)) \,\Big\|\, g_{-i}(h^t(z))\Big) \;\ge\; \Big(\liminf_{T\to\infty} \hat{\pi}_T^z(s)\Big) \cdot D_{\mathrm{KL}}\Big(f_{-i}(s) \,\Big\|\, g_{-i}(s)\Big).$$

Proof. If $f_{-i}, g_{-i} \in \mathcal{F}_{-i}^\kappa$, then for each t we have $f_{-i}(h^t(z)) = f_{-i}(s^t(z))$ and $g_{-i}(h^t(z)) = g_{-i}(s^t(z))$ by Definition 16. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(h^t(z)) \,\Big\|\, g_{-i}(h^t(z))\Big) = \frac{1}{T}\sum_{t=1}^{T} D_{\mathrm{KL}}\Big(f_{-i}(s^t(z)) \,\Big\|\, g_{-i}(s^t(z))\Big).$$

Grouping the sum by the value of $s^t(z)$ yields the stated decomposition. The inequality follows by lower bounding the sum by a single state's contribution and taking $\liminf$. □

Corollary C.5 (A sufficient condition for Assumption 3(3))**.** *Fix player i and suppose $\mathcal{S}_{-i} \subseteq \mathcal{F}_{-i}^\kappa$. Fix $g_{-i} \in \mathcal{S}_{-i} \setminus \{f_{-i}\}$ and a state $s \in \mathsf{S}_\kappa$ such that $D_{\mathrm{KL}}(f_{-i}(s)\|g_{-i}(s)) > 0$. If μ^f-a.s. in z,*

$$\liminf_{T\to\infty} \hat{\pi}_T^z(s) \;\ge\; \rho_i(g_{-i}) \;>\; 0,$$

then the on-path KL separation condition in Assumption 3(3) holds for this g_{-i} with $\kappa_i(g_{-i}) = \rho_i(g_{-i}) \cdot D_{\mathrm{KL}}(f_{-i}(s)\|g_{-i}(s))$.

Proof. Immediate from Lemma C.4. □

All statements in Sections 4–5 are formulated on the full history space H and therefore apply *verbatim* when the realized profile f (and/or the menu strategies in Assumption 3) lie in $\mathcal{F}^\kappa$. The main additions above are: (i) best responses to κ-memory opponents can be taken to be stationary Markov (Lemma C.3), and (ii) Assumption 3(3) can be verified by state-frequency separation (Lemma C.4 and Corollary C.5). Once Assumption 3 is verified (e.g. via Corollary C.5), the proofs of Lemma 4.2, Proposition 4.3, and Corollary 5.4 are unchanged.

D Implementation details of the strategy-level PS-BR planner

This appendix details the implementation used in our experiments. At each round, an agent samples a latent opponent strategy from its inference based on the previous history, evaluates candidate self-strategies by rollout, and plays the current action induced by the best rollout-value strategy.

D.1 Opponent strategy sampling

Fix player i at round t with local history $h_i^t = ((a_i^1, a_{-i}^1), \ldots, (a_i^{t-1}, a_{-i}^{t-1}))$. For opponent-strategy inference, the implementation rewrites this to the opponent-view history

$$\tilde{h}_{-i}^t = ((a_{-i}^1, a_i^1), \ldots, (a_{-i}^{t-1}, a_i^{t-1})),$$

so each tuple is ordered as *(opponent action, your action)*. The opponent strategy inference is performed *once per real decision round* (with configured label-sampling temperature) and then held fixed across all K rollout samples used to evaluate candidate self-strategies at that round. Inference supports two modes:

- **`llm-label` (default):** construct an in-context prompt containing the game rules, observed history, and the allowed strategy labels (with short descriptions), then ask the model to output *exactly one label*. Parsing is label-constrained; if parsing fails repeatedly, a deterministic label fallback is used.
- **`likelihood`:** infer from a hand-coded likelihood over the menu (described below), with no model call.

`llm-label` mode details. In `llm-label` mode, if the model call itself fails, the implementation falls back to `likelihood` mode for that decision round.

The template used in code is:

```
{rules_text}
Observed action history tuple format: (opponent action, your action).
Infer the opponent strategy from the FIRST action in each tuple.
Round 1: {opp_action_1}, {self_action_1}
Round 2: {opp_action_2}, {self_action_2}
...

You are inferring the opponent strategy in repeated {game_name}.
Observed rounds so far: {observed_rounds}.
Objective: sample one opponent strategy label according to your
posterior belief over allowed labels.
Estimate that posterior using ALL observed rounds
(do not ignore older rounds), and focus on recent patterns.
The opponent may change strategy over time; if you detect a shift,
prioritize the most recent consistent behavior while still
accounting for earlier rounds.
Internally assign a compatibility score from 0 to 100 to every
```

```
allowed label, convert them into relative posterior weights, and
sample exactly one final label from those weights.
Output rule: do NOT output scores, reasoning, or ranking.
Respond with exactly one label only.

**Output only the label.**

Allowed labels:
- {label_1}: {description_1}
- {label_2}: {description_2}
...
```

where `game_name` is the active repeated-game name (e.g., BoS, PD, Promo, Samaritan's dilemma, or Lemons), and `observed_rounds`=t-1.

When collusive-prior guidance is enabled (`-collusive-mode`), the prompt appends a strong-prior line. In our code this prior is `mad0` for Promo opponent 1 and `mad1` for Promo opponent 2.

`Likelihood`-mode details. To score strategy s, the implementation evaluates history under the opponent's perspective $\tilde{h}^t_{-i} = ((a^1_{-i}, a^1_i), \ldots, (a^{t-1}_{-i}, a^{t-1}_i))$:

$$\log L_t(s) = \sum_{u=1}^{t-1} \log\big(\mathbf{1}\{a^u_{-i} = J\}p^u_s + \mathbf{1}\{a^u_{-i} = F\}(1 - p^u_s)\big)\,,$$

with clipping to $[10^{-6}, 1 - 10^{-6}]$ for numerical stability. Given temperature $\tau > 0$ (implemented as $\tau = \max\{\texttt{sample_temperature}, 10^{-5}\}$), weights are

$$w_t(s) \propto \exp\!\left(\frac{\log L_t(s)}{\tau}\right),$$

and one opponent strategy is sampled from this categorical distribution.

D.2 Rollout value and strategy selection

Given a sampled opponent strategy s_{-i}, for every candidate self-strategy $s_i \in M_g$, the planner rolls out from round t to $\bar{t}$, where

$$\bar{t} = \begin{cases} \min\{T,\ t + H - 1\}, & H > 0, \\ T, & H = 0, \end{cases}$$

T is the game horizon, and H is the planning horizon.

For rollout sample $m \in \{1, \ldots, K\}$, at each simulated round r, actions are sampled from the fixed opponent strategy s_{-i} and the currently evaluated candidate s_i:

$$\hat{a}^{r,m}_i \sim \text{Bernoulli}\Big(p^r_{s_i}\Big)\,, \qquad \hat{a}^{r,m}_{-i} \sim \text{Bernoulli}\Big(p^r_{s_{-i}}\Big)\,,$$

Algorithm 1 Strategy-level PS-BR loop for two-player games

Require: game g, total rounds T, menu M_g, samples K, horizon H, discount γ, temperature τ, inference mode $\in \{\texttt{llm-label}, \texttt{likelihood}\}$

Initialize $h^1 \leftarrow \emptyset$, $x_1^1 \leftarrow (h^1, \emptyset)$, $x_2^1 \leftarrow (h^1, \emptyset)$, $C_1 \leftarrow 0$, and $C_2 \leftarrow 0$
for $t = 1, \ldots, T$ **do**
  for $i \in \{1, 2\}$ **do**
    Let $x_i^t = (h^t, r_i^{1:t-1})$ be player i's current local history
    Construct opponent-view history $\tilde{h}_{-i}^t$ by swapping tuple order in the public history h^t
    Infer one strategy label $s_{-i} \in M_g$ from rules, history $\tilde{h}_{-i}^t$
    for all $s_i \in M_g$ **do**
      for $k = 1, \ldots, K$ **do**
        $V_i^{(k)}(s_i \mid s_{-i}) \leftarrow \mathrm{RolloutValue}(g, i, s_i, s_{-i}, x_i^t, t, T, H, \gamma)$
      $\bar{V}_i(s_i \mid s_{-i}) \leftarrow \frac{1}{K} \sum_{k=1}^{K} V_i^{(k)}(s_i \mid s_{-i})$
    $s_i^\star \leftarrow \arg\max_{s_i \in M_g} \bar{V}_i(s_i \mid s_{-i})$ ▷ deterministic tie-break
    Sample real action a_i^t from strategy $s_i^\star$ at history x_i^t
  Sample realized rewards (r_1^t, r_2^t) from the environment payoff law at (a_1^t, a_2^t)
  $C_1 \leftarrow C_1 + r_1^t$ and $C_2 \leftarrow C_2 + r_2^t$
  Set $h^{t+1} \leftarrow (h^t, (a_1^t, a_2^t))$
  Set $x_1^{t+1} \leftarrow (h^{t+1}, r_1^{1:t})$ and $x_2^{t+1} \leftarrow (h^{t+1}, r_2^{1:t})$

where $p_{s_i}^r$ and $p_{s_{-i}}^r$ are the round-r probabilities of action J induced by s_i and s_{-i} under the simulated history prefix generated so far. The rollout value for candidate s_i against sampled opponent strategy s_{-i} is

$$V_i^{(m)}(s_i \mid s_{-i}) = \sum_{r=t}^{\bar{t}} \gamma^{r-t} u_i(\hat{a}_i^{r,m}, \hat{a}_{-i}^{r,m}),$$

with discount γ.

The estimated value of strategy s_i is

$$\bar{V}_i(s_i \mid s_{-i}) = \frac{1}{K} \sum_{m=1}^{K} V_i^{(m)}(s_i \mid s_{-i}),$$

and the chosen strategy is

$$s_i^\star \in \arg\max_{s_i} \bar{V}_i(s_i \mid s_{-i}),$$

with deterministic hash-based tie-breaking when needed. The executed action at real round t is then sampled from $s_i^\star$ at the current history.

For Experiment 3, the environment payoff law in Algorithm 1 is the known Gaussian noise family centered at the true mean matrix. On the player's own side, player i additionally samples $\tilde{m}_i \sim \pi_i^t(\cdot \mid x_i^t)$, rollout values are computed under $\tilde{m}_i$ in place of the true u_i, and player i's local information history stores only

$(h^t, r_i^{1:t-1})$; in particular, the update step above never reveals or conditions on $r_{-i}^{1:t-1}$.

E Social chain-of-thought prompting (SCoT)

This appendix discusses that the *social chain-of-thought* (SCoT) prompting intervention of Akata et al. (2025) can be viewed as a particularly simple instance PS-BR.

E.1 SCoT as a two-stage "predict-then-act" operator

In Akata et al. (2025), SCoT is implemented by *prompt-chaining* in each round of a repeated game:

1. *Prediction prompt (belief elicitation).* Given the public history h^t, the model is asked to predict the opponent's next move (or, more generally, to describe what the other player will do next).

2. *Action prompt (best response to the elicited belief).* The model is then asked to choose its action given the predicted opponent move, typically phrased as "given your prediction, what is best for you to do now?"

This "separate belief report, then act" structure forces an explicit theory-of-mind step before action selection, and empirically improves coordination in some repeated games.

E.2 Mapping SCoT as a special case of PSBR

Fix agent i at history h^t. Let A_{-i} denote the opponents' joint action space, and define the agent's *posterior predictive* over opponents' next action as

$$q_i^t(\cdot \mid h^t) \in \Delta(A_{-i}).$$

In our paper's belief language, $q_i^t(\cdot \mid h^t)$ is the one-step marginal induced by the agent's posterior predictive continuation belief $f_{-i}^{i,t}|_{h^t}$.

SCoT can then be expressed as the following generic operator:

1. *Infrence:* produce $\tilde{a}_{-i}^t$ as an imputation of the missing opponents' next action. Operationally, this is obtained by querying the model with the prediction prompt.

2. *Optimize given the imputation:* choose a_i^t as an (approximate) best response to the imputed $\tilde{a}_{-i}^t$ (and the known payoffs), e.g.

   $$a_i^t \in \arg\max_{a_i \in A_i} u_i(a_i, \tilde{a}_{-i}^t) \quad \text{(myopic)}.$$

   More generally, one may replace u_i by the continuation objective, i.e., choose a_i^t (or a continuation strategy) that maximizes the discounted value conditional on $\tilde{a}_{-i}^t$ and the induced continuation play.

Two special cases are worth separating because they clarify the relationship to PS-BR.

(i) Deterministic SCoT = point estimation. In the implementation studied by Akata et al. (2025), the model is often run in a near-deterministic regime (e.g., decoding choices consistent with temperature ≈ 0), so the prediction step behaves like a point estimate (roughly "MAP" under the model's implicit predictive distribution). In this view, SCoT is an inference-and-optimize heuristic that can still improve play by making

the model's implicit prediction problem explicit.

(ii) Myopic PS-BR = sampling-based estimation. If instead the prediction prompt is decoded stochastically (e.g., sampling at nonzero temperature), then $\tilde{a}^t_{-i}$ becomes a draw from the model's own predictive distribution:

$$\tilde{a}^t_{-i} \sim q^t_i(\cdot \mid h^t).$$

F Prompts

F.1 BASE prompts

In BASE, each player's round-t prompt is:

$$\text{rules text} + \text{compact history} + \text{“You are currently playing round } t\text{”} + \text{action query}.$$

The compact history prefix used in code is:

```
Observed action history (your action, opponent action):
Round 1: <self_1>, <opp_1>
...
Round t-1: <self_{t-1}>, <opp_{t-1}>
```

Round-level action query templates (Base).

- BoS:

```
Q: Which Option do you choose,  J or  F?
A:
```

- PD (order randomized each round):

```
Q: Which action do you choose, J or F?
A:
```

- Harmony:

```
Q: Which action do you choose, C or D?
A:
```

- Promo:

```
Q: Which action do you choose, R, P, or Z?
A:
```

- Samaritan (Helper prompt):

```
Q: Which action do you choose, H or N?
A:
```

- Samaritan (Recipient prompt):

```
Q: Which action do you choose, W or S?
A:
```

- Lemons (Seller prompt):

```
Q: Which action do you choose, HQ or LQ?
A:
```

- Lemons (Buyer prompt):

```
Q: Which action do you choose,  B or  D?
A:
```

Before the final "A:" token, code injects a strategy-context block (same helper used in BASE and SCOT):

```
In repeated <GameName>, a strategy maps prior history to a player's next action
(possibly probabilistically).
Allowed strategies:
- <label_1>: <short description>
- ...

Role mapping in this prompt:
- Player A is the other player.
- Player B is you.
Observed rounds so far: <t-1>.
Context: full history prefix up to round <t-1>.
Strongly expect Player A to play with strategy '<prior_label>'.   [if available]
Allowed action tokens: <tokens>.                                  [if available]
Output rule: do NOT output scores, reasoning, or ranking.
Respond with exactly one action only.
```

F.2 SCoT prompts

SCoT uses two prompts per player per round.

Stage 1 (prediction prompt). The prediction queries are:

- BoS:

```
Q: Which action do you predict the other player will choose, J or F?
A:
```

- PD (order randomized each round):

```
Q: Which action do you predict the other player will choose, J or F?
A:
```

- Harmony:

```
Q: Which action do you predict the other player will choose, C or D?
A:
```

- Promo:

```
Q: Which action do you predict the other player will choose, R, P, or Z?
A:
```

- Samaritan (Helper predicts Recipient):

```
Q: Which action do you predict the other player will choose, W or S?
A:
```

- Samaritan (Recipient predicts Helper):

```
Q: Which action do you predict the other player will choose, action H or action N?
A:
```

- Lemons (Seller predicts Buyer):

```
Q: Which Option do you predict the other player will choose, Option B or Option D?
A:
```

- Lemons (Buyer predicts Seller):

```
Q: Which Option do you predict the other player will choose, Option HQ or Option LQ?
A:
```

As implemented, the Stage-1 prediction prompt is enriched with the same strategy-context block shown above.

Stage 2 (action prompt conditioned on Stage-1 prediction). After receiving prediction <PRED>, code uses:

- BoS:

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option J and Option F),
compare which gives you a better result, and then choose.
Which Option do you think is the best to choose for you in this round, Option J or Option F?
Output only one letter: J or F.
A:
```

- PD (with randomized <opt1>, <opt2>):

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option <opt1> and Option <opt2>),
```

```
compare which gives you a better result, and then choose.
Which Option do you think is the best to choose for you in this round, Option <opt1> or Option <o
Output only one letter: J or F.
A:
```

- Harmony:

```
Q: Given that you think the other player will choose <PRED> in round <t>,
imagine the outcome for both of your possible actions (C and D),
compare which gives you a better result, and then choose.
Which action do you think is best for you in this round, C or D?
Output only one action: C or D.
A:
```

- Promo:

```
Q: Given that you think the other player will choose <PRED> in round <t>,
imagine the outcome for your possible actions (R, P, and Z),
compare which gives you a better result, and then choose.
Which action do you think is best for you in this round, R, P, or Z?
Output only one action: R, P, or Z.
A:
```

- Samaritan (Helper):

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option H and Option N),
compare which gives you a better result, and then choose.
Which Option do you think is best to choose for you in this round, Option H or Option N?
Output only one letter: H or N.
A:
```

- Samaritan (Recipient):

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option W and Option S),
compare which gives you a better result, and then choose.
Which Option do you think is best to choose for you in this round, Option W or Option S?
Output only one letter: W or S.
A:
```

- Lemons (Seller):

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option HQ and Option LQ),
compare which gives you a better result, and then choose.
```

```
Which Option do you think is best to choose for you in this round, Option HQ or Option LQ?
Output only one letter: HQ or LQ.
A:
```

- Lemons (Buyer):

```
Q: Given that you think the other player will choose Option <PRED> in round <t>,
imagine the outcome for both of your possible actions (Option B and Option D),
compare which gives you a better result, and then choose.
Which Option do you think is best to choose for you in this round, Option B or Option D?
Output only one letter: B or D.
A:
```

F.3 PS-BR prompts for known deterministic payoffs

PS-BR does not query the LLM for direct action choice. Actions are produced by rollout-based strategy evaluation after sampling one opponent strategy per round. The prompt-facing LLM call is for *opponent strategy-label inference* in `llm-label` mode.

Opponent strategy inference prompt (`llm-label`). At round t, for player i, history is rewritten to opponent view

$$\tilde{h}_{-i}^{t} = ((a_{-i}^{1}, a_{i}^{1}), \ldots, (a_{-i}^{t-1}, a_{i}^{t-1})),$$

so tuples are `(Player A action, Player B action)` with:

- Player A = opponent whose strategy is inferred.
- Player B = current decision-maker.

The prompt template is:

```
You are inferring Player A's strategy (the opponent) in repeated <GameName>.
In a repeated-game setting, a strategy is a rule that maps prior history to the
player's next action (possibly probabilistically).
<rules_text>
Observed rounds so far: <t-1>.

Allowed labels:
- <label_1>: <description_1>
- ...

Observed action history tuple format: (Player A action, Player B action).
Player A is the opponent whose strategy label you must infer.
Player B is you (the decision-maker).
Context: full history prefix up to round <...>.
Target: observed Player A action at round <...>.
```

```
Choose the allowed label that makes this observed Player A target most compatible
with the context.
At round <...>, use this mapping:
Context history as (Player A, Player B), rounds <...>:
round <k>: Player A=<...>, Player B=<...>
Observed target Player A action at round <...>: <...>
Strongly expect Player A to play with strategy '<prior_label>'.
Player A's strategy may have changed over time, so weigh recent rounds more heavily
than earlier rounds.
Output rule: do NOT output scores, reasoning, or ranking.
Respond with exactly one label only.

**Output only the label.**
```

Likelihood mode (no prompt). If `-strategy-inference likelihood` is used, no LLM prompt is issued for strategy inference; the label is sampled from a hand-coded likelihood over the finite menu.

F.4 PS-BR prompts for unknown stochastic payoffs

Under the theorem-aligned implementation used for Experiment 3, PS-BR under unknown stochastic payoffs still samples both an opponent strategy hypothesis and a payoff hypothesis at each round before rollout-based strategy evaluation. The opponent-strategy side is handled exactly as in the known deterministic-payoff case. The payoff side is not open-ended JSON inference. Instead, Experiment 3 uses the known-common-noise / unknown-mean construction from Section 6 and Section 7.4.1: player i maintains a posterior over a finite menu $\mathcal{M}_{i,g}$ of candidate mean payoff matrices under the Gaussian noise family with known variance σ_g^2.

Opponent strategy inference prompt (`llm-label`). The opponent strategy is inferred from the joint action history, exactly as in the known deterministic payoffs case. The prompt template remains identical to the one detailed in the previous subsection.

Finite-menu Gaussian payoff posterior (experiment configuration). At round t, player i updates

$$\pi_i^t(m \mid h^t, r_i^{1:t-1}) \propto \pi_i^0(m) \prod_{s=1}^{t-1} \phi(r_i^s; m(a^s), \sigma_g^2), \qquad m \in \mathcal{M}_{i,g},$$

where $\phi(\cdot; \mu, \sigma_g^2)$ is the Gaussian density and $r_i^s \mid a^s \sim \mathcal{N}(m(a^s), \sigma_g^2)$ under candidate mean matrix m. The implementation then samples one matrix label $\tilde{m}_i \sim \pi_i^t$ and evaluates continuation strategies against the induced payoff kernel

$$q_i^{\tilde{m}_i}(\cdot \mid a) = \mathcal{N}(\tilde{m}_i(a), \sigma_g^2).$$

Product structure of the menu. Although the theorem-level menu $\mathcal{M}_{i,g}$ is finite but large, it has product form over joint actions. With a product prior over the offsets $(k_a)_{a \in A}$ and the Gaussian likelihood above, the posterior factorizes by joint action. Operationally, the implementation therefore updates the discrete

posterior for each action-specific offset $k_a \in K$ separately and samples a full mean matrix by drawing one offset for each joint action. This is exactly equivalent to sampling from the full finite menu, without explicitly enumerating all of its elements.

Likelihood mode (experiment configuration). In the reported Experiment 3 runs, `-payoff-inference likelihood` is used. No LLM prompt is issued for payoff inference; the sampled mean-matrix label is drawn from the Gaussian posterior above. Opponent strategy inference is handled either by the `llm-label` prompt described above or by the corresponding likelihood mode, depending on the strategy-inference setting.

Heuristic prompt mode. An open-ended `json` payoff-table prompt can still be used as a heuristic variant, but it is not the theorem-aligned implementation analyzed in Section 6 and instantiated in Experiment 3.

G Game-specific strategy menus

Denote a_i^{t-1} and a_{-i}^{t-1} denote own and opponent actions at round $t-1$. Then we consider:

(1) BoS strategy menu. *Here p_s^t denotes the probability of playing J at round t.*

- `insist_j`: $p_s^t = 1$ for all t.
- `insist_f`: $p_s^t = 0$ for all t.
- `wsls_bos`: $p_s^1 = 0.5$; for $t \geq 2$, if $a_i^{t-1} = a_{-i}^{t-1}$ then repeat a_i^{t-1}, else switch from a_i^{t-1}.
- `mlur`: $p_s^1 = 0.5$; for $t \geq 2$, if $a_i^{t-1} = a_{-i}^{t-1}$ then repeat a_i^{t-1}, else $p_s^t = 0.5$.
- `alternate_phase0`: $p_s^t = 1$ on odd t, and $p_s^t = 0$ on even t.
- `alternate_phase1`: $p_s^t = 0$ on odd t, and $p_s^t = 1$ on even t.
- `noisy_insist_j`: $p_s^t = 0.9$ for all t.
- `noisy_insist_f`: $p_s^t = 0.1$ for all t.

(2) PD strategy menu. *Here p_s^t denotes the probability of playing J at round t.*

- `allc`: $p_s^t = 1$ for all t.
- `alld`: $p_s^t = 0$ for all t.
- `soft_allc`: $p_s^t = 0.9$ for all t.
- `soft_alld`: $p_s^t = 0.1$ for all t.
- `tft`: $p_s^1 = 1$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = J$.
- `wsls`: $p_s^1 = 1$; for $t \geq 2$, if $a_i^{t-1} = a_{-i}^{t-1}$ then repeat a_i^{t-1}, else switch from a_i^{t-1}.

- `soft_grim_trigger`: $p_s^t = 0$ if the opponent played F in either of the previous two rounds; otherwise $p_s^t = 1$.
- `grim_trigger`: $p_s^t = 1$ until the opponent has played F at least once in the past; thereafter $p_s^t = 0$ forever.

(3) Harmony strategy menu. *Here p_s^t denotes the probability of playing C at round t.*

- `allc`: $p_s^t = 1$ for all t.
- `alld`: $p_s^t = 0$ for all t.
- `tft`: $p_s^1 = 1$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = C$.
- `stft`: $p_s^1 = 0$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = C$.
- `generous_tft`: $p_s^1 = 1$; for $t \geq 2$, if $a_{-i}^{t-1} = C$ then $p_s^t = 1$, else $p_s^t = 0.3$.
- `grim_trigger`: $p_s^t = 1$ until the opponent has played D at least once in the past; thereafter $p_s^t = 0$ forever.
- `wsls_pavlov`: $p_s^1 = 1$; for $t \geq 2$, if $a_i^{t-1} = a_{-i}^{t-1}$ then repeat a_i^{t-1}, else switch from a_i^{t-1}.
- `random_pc`: $p_s^t = 0.5$ for all t.

(4) Promo strategy menu (actions: R = regular, P = promotion, Z = punishment/price war).

- `allR`: play R at every round.
- `allP`: play P at every round.
- `allZ`: play Z at every round.
- `soft_allR`: play R with probability 0.9 and P with probability 0.1.
- `soft_allP`: play P with probability 0.9 and R with probability 0.1.
- `mad0`: cooperative path is odd-round P/even-round R; when a deviation from the prescribed phase path is detected, play Z for 2 rounds, then return to phase-0 alternation.
- `mad1`: cooperative path is odd-round R/even-round P; when a deviation from the prescribed phase path is detected, play Z for 2 rounds, then return to phase-1 alternation.
- `grim_trigger`: follow the phase-0 alternating path until the first deviation, then play Z forever.

(5) Samaritan's dilemma (Helper actions: H = Help, N = No-help; Recipient actions: W = Work, S = Shirk). *Helper strategy menu.* Here p_s^t denotes the probability the helper plays H at round t.

- `always_help`: $p_s^t = 1$ for all t.
- `never_help`: $p_s^t = 0$ for all t.

- `tft_help`: $p_s^1 = 1$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = W$.
- `grim_forgive`: $p_s^t = 0$ if the recipient played S in either of the previous two rounds; otherwise $p_s^t = 1$.
- `grim_nohelp`: $p_s^t = 1$ until the recipient has played S at least once in the past; thereafter $p_s^t = 0$ forever.
- `wsls_helper`: $p_s^1 = 1$; for $t \geq 2$, if $a_{-i}^{t-1} = W$ then repeat a_i^{t-1}, else switch from a_i^{t-1}.
- `noisy_help`: $p_s^t = 0.9$ for all t.
- `noisy_nohelp`: $p_s^t = 0.1$ for all t.

Recipient strategy menu. Here p_s^t denotes the probability the recipient plays W at round t.

- `always_work`: $p_s^t = 1$ for all t.
- `always_shirk`: $p_s^t = 0$ for all t.
- `work_if_helped`: $p_s^1 = 0.5$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = H$.
- `exploit_help`: $p_s^1 = 0.5$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = N$.
- `grim_shirk_after_nohelp`: $p_s^t = 1$ until the helper has played N at least once in the past; thereafter $p_s^t = 0$ forever.
- `forgiving_work`: $p_s^1 = 1$; for $t \geq 2$, if $a_{-i}^{t-1} = H$ then $p_s^t = 1$, else $p_s^t = 0.3$.
- `noisy_work`: $p_s^t = 0.9$ for all t.
- `noisy_shirk`: $p_s^t = 0.1$ for all t.

(6) Lemons (Seller actions: HQ = High-quality, LQ = Low-quality; Buyer actions: B = Buy, D = Don't buy). *Seller strategy menu.* Here p_s^t denotes the probability the seller plays HQ at round t.

- `always_hq`: $p_s^t = 1$ for all t.
- `always_lq`: $p_s^t = 0$ for all t.
- `hq_if_bought_last`: $p_s^1 = 0.5$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = B$.
- `grim_hq_until_boycott`: $p_s^t = 1$ until the buyer has played D at least once in the past; thereafter $p_s^t = 0$ forever.
- `lq_if_boycott_last`: $p_s^1 = 0.5$; for $t \geq 2$, $p_s^t = 0$ iff $a_{-i}^{t-1} = D$.
- `grim_forgiving`: $p_s^t = 0$ if the buyer played D in either of the previous two rounds; otherwise $p_s^t = 1$.
- `noisy_hq`: $p_s^t = 0.9$ for all t.

- `noisy_lq`: $p_s^t = 0.1$ for all t.

Buyer strategy menu. Here p_s^t denotes the probability the buyer plays B at round t.

- `always_buy`: $p_s^t = 1$ for all t.
- `never_buy`: $p_s^t = 0$ for all t.
- `soft_always_buy`: $p_s^t = 0.9$ for all t.
- `soft_never_buy`: $p_s^t = 0.1$ for all t.
- `tft_buy`: $p_s^1 = 0.5$; for $t \geq 2$, $p_s^t = 1$ iff $a_{-i}^{t-1} = HQ$.
- `generous_buy`: $p_s^1 = 1$; for $t \geq 2$, if $a_{-i}^{t-1} = HQ$ then $p_s^t = 1$, else $p_s^t = 0.3$.
- `grim_boycott`: $p_s^t = 1$ until the seller has played LQ at least once in the past; thereafter $p_s^t = 0$ forever.
- `grim_forgiving`: $p_s^t = 0$ if the seller played LQ in either of the previous two rounds; otherwise $p_s^t = 1$.

H Promo game

H.1 Promo game (Lal, 1990): alternating promotions with finite punishment

Lal (1990) studies repeated price competition in a market with two identical "national" brands that have loyal consumers and a third "local" brand with little/no loyalty. The local brand disciplines prices in the switching segment, creating a tension for the national brands between (i) extracting rents from loyals via a high "regular" price and (ii) defending the switchers via temporary price cuts. A key result is that, even when the corresponding one-shot stage game has no Nash equilibrium, an *alternating promotions* pattern – only one national brand is on promotion in a given period and the roles alternate over time – can arise as a *pure-strategy Nash equilibrium* of the infinite-horizon discounted game, supported by a credible number of punishment periods.

To obtain a compact repeated-game benchmark, we discretize Lal, 1990's richer price-choice problem into three representative regimes per firm:

- *Regular* (R): charge the high "regular" price
- *Promotion* (P): charge the low promotional price
- *Punishment/price war* (Z): charge a very low price used only in punishment phases.

The resulting 3×3 payoff matrix in Appendix 7 is a reduced-form encoding of the ordinal incentive structure: a unilateral promotion against a regular-price rival yields the highest current-period gain (the "temptation" payoff); simultaneous promotions are less profitable than *alternating* promotions; and outcomes involving Z are jointly bad, standing in for the "intense competition/price war" phase used to deter deviations.

The canonical nontrivial Nash equilibrium is an *alternating* path: play (P, R) in odd rounds and (R, P) in even rounds (or vice versa). After any deviation from the prescribed phase, switch to a punishment phase (e.g., (Z, Z) for a fixed number of rounds) for a few periods and then return to the alternating path (as defined as Abreu, 1988), or revert permanently to a low-payoff punishment regime (grim trigger). For sufficiently patient players, the discounted loss from the punishment phase outweighs the one-shot deviation gain, making the alternating-promotions path incentive compatible.