SheriffBench: A Behavioral Benchmark for LLM Enforcement Policy Under Economic Pressure

Anonymous ACL submission

Abstract

We introduce **SheriffBench**, a behavioral benchmark for evaluating large language models (LLMs) in strategic decision-making under asymmetric information and economic pressure. Inspired by the board game *Sheriff of Nottingham*, SheriffBench formalizes the inspect-versus-accept-bribe dilemma as a controlled binary decision task. The benchmark systematically varies two orthogonal signals: economic incentive (bribe magnitude) and merchant reputation (historical lying rate), enabling structured analysis of policy behavior across a factorial scenario space.

We evaluate five state-of-the-art language models—DeepSeek V3.2, GPT-4 Turbo, Llama 3.1 8B, Llama 3.3 70B, and Qwen3 Coder Plus—using a unified dataset of 121 scenarios. Each model is tested under three personality-conditioned prompting configurations representing different enforcement policies: *Greedy*, *Neutral*, and *Risk-Averse*. This design allows us to analyze how prompt-induced behavioral framing affects corruption acceptance and enforcement strategies across model families.

Results reveal substantial variation in decision policies across both models and prompt conditions, with some models exhibiting strong economic responsiveness while others default to deterministic enforcement strategies. SheriffBench therefore provides a controlled diagnostic framework for studying prompt-controllable policy shifts, corruption sensitivity, and norm-enforcement behavior in LLM agents, complementing existing reasoning-focused evaluation benchmarks.

1 Introduction

1.1 A New Frontier in Game-Based Evaluation

Benchmarking LLMs and AI agents on games has emerged as a vibrant and well-established research direction. Dedicated benchmarks exist for Chess, Poker, Mafia, and Avalon, each of which has produced valuable insights into model capabilities regarding reasoning, strategy, and deception. This body of work reflects a broad community consensus that **games provide controlled, structured, and reproducible environments** for probing complex decision-making. Despite this rich landscape, Sheriff of Nottingham has received no attention in the LLM benchmarking literature; to the best of our knowledge, this paper presents the first systematic evaluation of LLMs on this game.

1.2 Significance and Strategic Depth

Sheriff of Nottingham is a globally recognized board game, earning prestigious industry accolades such as the *Origins Award for Best Board Game* and the *Dice Tower Seal of Excellence*. Its longevity and popularity stem from its ability to encode rich social and cognitive challenges. Unlike Chess (Campbell et al., 2002) and Go (Silver et al., 2016) (which focus on perfect information) or Poker (Brown and Sandholm, 2018) (which focuses on statistical bluffing), Sheriff of Nottingham operationalizes **negotiation under incomplete information** and **probabilistic norm enforcement**. A game that has engaged millions of human minds over a decade is a natural and meaningful testbed for evaluating the strategic "common sense" and ethical alignment of artificial intelligence.

1.3 Mapping Game Mechanics to Real-World Decisions

The structural complexity of *Sheriff of Nottingham* provides more than mere entertainment; it operationalizes a sophisticated set of cognitive and strategic behaviors that are directly relevant to contemporary real-world decision-making contexts. Beyond its recreational origins, *SheriffBench* isolates and evaluates how Large Language Models (LLMs) navigate the tension between certain economic incentives and probabilistic risk. Specifically, our

benchmark systematically varies three critical dimensions: the magnitude of economic incentives to determine how models trade off guaranteed rewards (bribes) against uncertain enforcement outcomes; the influence of merchant reputation to assess how historical deception rates affect trust and institutional enforcement policies; and prompt-induced personality shifts to identify whether behavioral framing can produce structural inversions in a model's decision regime. These behaviors are not merely academic; they mirror the high-stakes challenges inherent in automated negotiation systems, financial governance modeling, and multi-agent strategic enforcement. By quantifying LLM responsiveness to these observable signals, *SheriffBench* provides direct implications for how we deploy and trust autonomous agents in environments characterized by asymmetric information and economic pressure.

1.4 Problem Statement

We investigate LLM behavior in a strategic enforcement environment inspired by the game *Sheriff of Nottingham*. In this setting, an LLM assumes the role of a sheriff who must make a binary decision under uncertainty: either inspect a merchant's bag, which yields a probabilistic reward or penalty depending on its true contents, or accept a bribe, which provides a guaranteed immediate payoff. This formulation operationalizes a classic inspect-versus-corrupt dilemma, where the decision-maker must trade off expected value against certainty under asymmetric information.

At decision time, the sheriff has access to three observable signals: the merchant's historical lying rate, representing prior deception frequency; the offered bribe magnitude, representing immediate economic incentive; and the declared goods, representing the merchant's stated claim. However, the true composition of the bag remains hidden, creating informational asymmetry and requiring the model to implicitly estimate risk. The central research question of this work is whether personality-conditioned prompting induces systematic and measurable shifts in enforcement policy when the underlying economic and reputational conditions remain identical. In other words, we seek to determine whether prompt-based personality framing can produce structural changes in decision regimes rather than merely superficial variation in reasoning style.

1.5 Research Gap

Existing LLM benchmarks primarily evaluate reasoning accuracy, knowledge recall, and mathematical competence. Limited work:

1. Systematically varies economic incentives in factorial design,
2. Isolates binary enforcement policies under uncertainty,
3. Measures prompt-induced regime inversion,
4. Quantifies bribe threshold sensitivity.

SheriffBench addresses this gap by introducing a controlled behavioral benchmark.

1.6 Related Work: Strategic Reasoning and Deception in LLMs

The evaluation of Large Language Models (LLMs) in strategic environments has transitioned from simple logic puzzles to complex, multi-agent games requiring a blend of tactical reasoning and natural language negotiation. This research tradition can be categorized into three primary clusters that inform the development of *SheriffBench*:

1. Strategic Negotiation and Game-Theoretic Foundations: A landmark in this field is the development of CICERO (, FAIR), the first AI agent to achieve human-level performance in the game *Diplomacy*. By integrating language models with strategic reasoning, CICERO demonstrated that LLMs could navigate multi-party cooperation and competition. *SheriffBench* builds on this by evaluating if models can align their verbal justifications with their enforcement actions—a core challenge identified in CICERO's "intent-action" coordination.

2. Hidden Information and Deception Benchmarks: As models move into agentic roles, benchmarks have shifted toward games with asymmetric information. PokerBench (Zhuang et al., 2025) provides a curated dataset of professional poker scenarios to evaluate LLM performance in mathematical risk assessment and bluffing. Similarly, social deduction benchmarks like AvalonBench (Light et al., 2023) and studies in the game of Mafia / Spyfall (Kim and Kim, 2023) probe the models' capacity for persistent deception and social persuasion. While these works focus on win-optimization, *SheriffBench* uniquely isolates the **corruption threshold**—evaluating not just the ability to deceive, but

the propensity to compromise a normative role (the Sheriff) for economic gain.

3. Alignment and Instrumental Deception: Recent safety research has investigated "instrumental deception," where models strategically provide false information to satisfy external incentives. Frameworks like DeceptionBench (Huang et al., 2025) and studies on Machiavellianism in LLMs (Pan et al., 2023) show that while models remain rule-compliant in neutral settings, contextual framing can trigger deceptive behavior. *SheriffBench* contributes to this discourse by quantifying "Rational Corruption," analyzing whether a model's deviation from its "Sheriff" persona is a stochastic failure or a calculated response to bribe magnitude.

1.7 Contributions

This paper makes the following contributions:

- We introduce **SheriffBench**, a novel behavioral benchmark for evaluating large language models in strategic decision-making under asymmetric information and economic pressure.
- We propose a **controlled factorial evaluation framework** that systematically varies economic incentives and reputational signals, enabling fine-grained analysis of model decision policies.
- We release the **SheriffBench dataset**, a structured collection of 121 scenarios designed to study enforcement behavior under controlled variations in incentives and trust signals.
- We formalize the **inspect-versus-accept dilemma** as a minimal and interpretable testbed for studying corruption sensitivity, enforcement behavior, and trade-offs between guaranteed rewards and uncertain outcomes.
- We demonstrate that **personality-conditioned prompting induces structured policy shifts**, while revealing a persistent bias toward enforcement behavior across models, suggesting the presence of an underlying rule-following prior.
- We release the SheriffBench codebase and evaluation pipeline to support reproducibility and future extensions of behavioral benchmarking (ano, 2026).

2 Methodology

SheriffBench is evaluated using a controlled factorial scenario benchmark designed to systematically characterize model decision behavior under uncertainty. Rather than relying on open-ended or sequential interactions, our approach isolates decision policies by varying key environmental factors in a structured and reproducible manner.

Specifically, we construct a discrete scenario space that captures the core elements of the inspect-versus-accept dilemma. These include economic incentives (bribe magnitude) and merchant reputation (historical lying rate), which together define the informational and payoff structure faced by the model. By evaluating responses across the full factorial combination of these variables, we obtain a comprehensive view of how enforcement decisions change as a function of underlying conditions.

This design enables precise analysis of decision boundaries, threshold effects, and policy shifts, allowing us to quantify the extent to which model behavior is sensitive to economic and reputational signals. All models operate under the same payoff structure as the economic baseline, with reward and penalty values explicitly specified in the task description.

2.1 Controlled Factorial Scenario Generation

To isolate decision thresholds, we generate structured scenarios via combinatorial variation.

2.1.1 Economic Incentive Variation

To systematically evaluate sensitivity to guaranteed rewards, we vary the bribe magnitude across a discrete and uniformly spaced set of values defined as $Bribe \in \{0, 5, 10, \ldots, 50\}$. This range spans from zero incentive to a high economic offer relative to the initial endowment of 50 gold, allowing us to probe how enforcement behavior shifts as guaranteed gain increases. The step size of five gold provides sufficient resolution to detect threshold effects, tipping points, or nonlinear policy transitions. By holding other variables constant while varying bribe magnitude, we isolate the model's responsiveness to immediate economic incentives and assess whether personality conditioning alters the slope or sharpness of this incentive-response curve.

2.1.2 Reputation Variation

To model reputational influence under asymmetric information, we vary the merchant's historical de-

ception level using a discrete scale $LyingRate \in \{0, 1, \ldots, 10\}$. This scale represents the number of times the merchant was caught lying in the previous ten rounds. A value of 0 indicates that the merchant was not caught lying in any of the last ten interactions, whereas a value of 10 indicates that the merchant was caught lying in all ten previous rounds. Intermediate values correspond to proportional frequencies of detected deception. This formulation operationalizes reputation as an observable empirical prior that the sheriff can condition upon when estimating the probability of illegal goods. By discretizing reputation over a fixed ten-round window, we ensure controlled variation while maintaining interpretability.

2.2 Output Collection

For each generated scenario, the LLM produces two outputs: a binary enforcement decision and a corresponding natural language justification. The binary decision consists of either inspecting the merchant's bag or accepting the offered bribe, thereby directly instantiating the enforcement policy under the given economic and reputational conditions. In addition to the discrete action, the model provides a free-form reasoning explanation describing the rationale behind its choice.

From these outputs, we extract multiple analytical signals. First, we compute decision frequencies across the factorial space to estimate overall inspection and acceptance rates under each personality condition. Second, we analyze threshold effects by examining how decision probabilities change as bribe magnitude and reputation vary, allowing identification of tipping points or nonlinear transitions. Third, we conduct qualitative reasoning analysis to identify dominant explanatory themes, such as guaranteed reward framing, risk minimization, reputational justification, or moral norm enforcement.

3 Experimental Setup

3.1 Dataset Construction

SheriffBench is a fully synthetic benchmark constructed using a structured factorial design (Box et al., 2005). The dataset is generated as the Cartesian product of two primary variables:

$$Bribe \in \{0, 5, 10, \ldots, 50\}, \quad LyingRate \in \{0, 1, \ldots, 10\}$$

Each scenario corresponds to a unique tuple $(Bribe, LyingRate)$, yielding $|B| \times |R| = 11 \times 11 = 121$ total scenarios. For every combination, a merchant instance is constructed with a corresponding bribe offer and historical deception profile. A hidden bag configuration is then generated to instantiate the underlying uncertainty, along with a declaration (bluff) representing the merchant's stated claim.

3.2 Models

We evaluate five state-of-the-art large language models representing a mixture of open-weight and proprietary systems with diverse architectural characteristics and parameter scales. All models were used in an **off-the-shelf zero-shot configuration** without any task-specific fine-tuning. Access was provided through the **OpenRouter API**.

Model	Architecture	Params	Access	Reference
DeepSeek V3.2	MoE (MLA+DSA)	671B/37B*	Open/API	(DeepSeek-AI et al., 2025)
GPT-4 Turbo	Decoder-only	Undisclosed	API	(OpenAI et al., 2024)
Llama 3.1 8B	Dense	8B	Open	(Grattafiori et al., 2024)
Llama 3.3 70B	Dense	70B	Open	(Grattafiori et al., 2024)
Qwen3 Coder+	Hybrid MoE	480B/35B*	API	(Yang et al., 2025)

Table 1: Models evaluated in SheriffBench. *Total/Active parameters for MoE models. All evaluated zero-shot.

3.3 Baseline Policies

To contextualize LLM behavior, we introduce four baseline decision policies.

Random Policy Baselines. We construct three stochastic policies reflecting greedy, neutral, and risk-averse enforcement tendencies. These policies accept bribes with fixed probabilities of 0.8, 0.5, and 0.1 respectively.

Economic Rational Baseline. A rule-based agent that follows an expected-value decision rule, accepting any bribe exceeding 25 gold (the maximum possible inspection penalty). For lower bribe values, it compares expected value of inspection against the offered bribe. The probability of illegal goods is estimated as $P_{illegal} = LyingRate/10$. With $N = 5$ items per bag, reward $r = 5$, and penalty $c = 2$:

$$EV_{inspect} = E[I]{\cdot}r - E[L]{\cdot}c = (N{\cdot}P_{illegal}){\cdot}5 - (N - N{\cdot}P_{illegal}){\cdot}2$$

Baseline Policy	Accept	Inspect	Accept%	Inspect%
Greedy Random (80%)	102	19	84.3%	15.7%
Neutral Random (50%)	63	58	52.1%	47.9%
Risk-Averse Random (10%)	22	99	18.2%	81.8%
Economic Rational (EV)	91	30	75.2%	24.8%

Table 2: Decision distribution for baseline policies on the 121-scenario dataset.

4 Results

We evaluate LLM enforcement policies using the unified SheriffBench dataset of 121 decision scenarios. Results are reported across three personality-conditioned configurations: **Neutral**, **Greedy**, and **Risk-Averse**. Each configuration represents a distinct enforcement framing—from balanced cost-benefit reasoning, to profit maximization, to loss minimization—allowing us to characterize how prompt-induced personas shift decision policy across the same underlying scenario space.

4.1 Neutral Prompted Results

To evaluate baseline decision behavior without strong optimization bias, the **Neutral Sheriff** configuration instructs models to balance enforcement credibility with economic reasoning without prioritizing either profit maximization or strict loss minimization.

4.1.1 Overall Decision Distribution

Model	Inspect%	Accept%
Deepseek V3.2	78.5	21.5
GPT-4 Turbo	66.1	33.9
Llama 3.1 8B	85.1	14.9
Llama 3.3 70B	93.4	6.6
Qwen-3 Coder+	100.0	0.0
Neutral Random (50%)	47.9	52.1

Table 3: Decision distribution under neutral prompting.

Neutral prompting reveals substantial behavioral divergence across models. **GPT-4 Turbo** demonstrates the most balanced decision policy (33.9% accept, 66.1% inspect). **DeepSeek V3.2** shows moderate inspection preference (78.5%). Both **Llama models** display strong inspection bias with acceptance rates below 15%. **Qwen-3 Coder+** inspects in every scenario, indicating a deterministic enforcement strategy even without risk-averse prompting.

4.1.2 Profile Differentiation Across Decisions

Model	Insp. Bribe	Acc. Bribe	ΔB	Insp. Lying	Acc. Lying	ΔL
Deepseek V3.2	25.0	25.0	0.0	6.0	1.5	−4.5
GPT-4 Turbo	23.0	28.9	+5.9	6.7	1.7	−5.0
Llama 3.1 8B	25.5	21.9	−3.6	5.1	4.2	−1.0
Llama 3.3 70B	24.8	27.5	+2.7	5.4	0.0	−5.4
Qwen-3 Coder+	25.0	N/A	N/A	5.0	N/A	N/A
Neutral Random	25.2	24.8	−0.4	5.2	4.8	−0.4

Table 4: Merchant profile differentiation across decisions under neutral prompting.

GPT-4 Turbo shows the strongest dual-signal differentiation, accepting higher bribes while simultaneously favoring merchants with substantially lower lying rates. DeepSeek primarily relies on merchant reputation, showing little sensitivity to bribe magnitude but strong filtering based on historical honesty. Llama 3.3 70B applies a strict reputation threshold, accepting only from merchants with near-perfect honesty. Llama 3.1 8B displays weak and inconsistent signal differentiation.

4.1.3 Bribe Sensitivity

Model	Zero (0g)	Low (5–15g)	Mid (20–30g)	High (35–50g)
Deepseek V3.2	0.0	33.3	18.2	20.5
GPT-4 Turbo	0.0	30.3	39.4	40.9
Llama 3.1 8B	18.2	21.2	9.1	13.6
Llama 3.3 70B	0.0	6.1	9.1	6.8
Qwen-3 Coder+	0.0	0.0	0.0	0.0
Neutral Random	45.5	54.5	51.5	52.3

Table 5: Acceptance rates (%) by bribe magnitude under neutral prompting.

GPT-4 Turbo exhibits the most rational economic response, with acceptance probability increasing as bribe magnitude grows. DeepSeek shows moderate responsiveness but peaks at lower bribe levels, suggesting reputation gating. Llama models show weak or inconsistent responses, while Qwen-3 remains completely insensitive to economic incentives.

4.1.4 Reputation Sensitivity

Model	Honest (0–2)	Neutral (3–5)	Suspicious (6–7)	Deceptive (8–10)
Deepseek V3.2	57.6	18.2	4.5	0.0
GPT-4 Turbo	90.9	33.3	0.0	0.0
Llama 3.1 8B	27.3	3.0	13.6	15.2
Llama 3.3 70B	24.2	0.0	0.0	0.0
Qwen-3 Coder+	0.0	0.0	0.0	0.0
Neutral Random	51.5	51.5	54.5	51.5

Table 6: Acceptance rates (%) by merchant reputation under neutral prompting.

GPT-4 Turbo demonstrates the strongest reputation filtering, accepting bribes from most honest merchants while rejecting all suspicious or deceptive ones. DeepSeek shows a clear monotonic decline. Notably, Llama 3.1 8B exhibits a non-monotonic response, accepting a higher fraction from Suspicious (13.6%) and Deceptive (15.2%) merchants than from Neutral ones (3.0%), suggesting disproportionate weight on economic incentives in certain scenario regions.

4.1.5 Reasoning Patterns and Enforcement Bias

Expected-value reasoning dominates across all models (74–95%), suggesting that neutral prompts encourage cost-benefit style analysis. However, Qwen-3 Coder+ exhibits a notable reasoning-behavior mismatch: despite producing expected-value justifications in 80.2% of cases, it inspects in every scenario, suggesting its verbalized reasoning does not reflect its underlying decision policy.

Every evaluated model under neutral prompting exhibits a significantly lower acceptance rate than the Neutral Random baseline (52.1%), ranging from 33.9% (GPT-4 Turbo) to 0.0% (Qwen-3 Coder+). This systematic deviation suggests an inherent "inspection bias" predisposing models toward rule-following even under a balanced prompt.

4.2 Greedy Sheriff Evaluation

The greedy configuration prompts the sheriff to prioritize economic gain. The results reveal substantial variation across model architectures in how aggressively they pursue profit and how effectively they integrate contextual signals.

4.2.1 Overall Decision Distribution

Model	Inspect%	Accept%
Deepseek V3.2	24.8	75.2
GPT-4 Turbo	34.7	65.3
Llama 3.1 8B	1.7	98.3
Llama 3.3 70B	9.1	90.9
Qwen-3 Coder+	43.0	57.0
Greedy Random (80%)	15.7	84.3

Table 7: Decision distribution under greedy prompting.

Acceptance behavior dominates across all models, confirming that profit-seeking prompts substantially shift enforcement policy. However, significant variation persists. Llama 3.1 8B exhibits near-complete opportunistic behavior (98.3% accept). In contrast, Qwen-3 Coder+ shows the most restrained profit-seeking behavior (57% accept), inspecting merchants more frequently than other evaluated models.

4.2.2 Merchant Profile Differentiation

Qwen-3 Coder+ demonstrates the strongest signal separation, accepting merchants offering significantly larger bribes while favoring inspections when incentives are smaller. GPT-4 Turbo exhibits similar but less pronounced behavior. In contrast,

Model	Insp. Bribe	Acc. Bribe	ΔBribe	Interpretation
Deepseek V3.2	17.5	27.5	+10.0	Moderate incentive sensitivity
GPT-4 Turbo	19.4	28.0	+8.6	Balanced profit–risk tradeoff
Llama 3.1 8B	0.0	25.4	+25.4	Highly opportunistic
Llama 3.3 70B	0.0	27.5	+27.5	Strong profit maximization
Qwen-3 Coder+	17.8	30.4	+12.6	Strong signal differentiation
Greedy Random	20.0	25.9	+5.9	Opportunistic baseline

Table 8: Merchant profile differentiation under greedy prompting.

Llama models show minimal differentiation, suggesting their decisions are largely dominated by the profit-maximization prompt rather than contextual evaluation.

4.2.3 Bribe Sensitivity

Model	Low (0–15g)	Medium (20–30g)	High (35–50g)
Deepseek V3.2	60.8	81.8	95.5
GPT-4 Turbo	48.5	69.7	90.9
Llama 3.1 8B	100.0	100.0	100.0
Llama 3.3 70B	83.3	100.0	100.0
Qwen-3 Coder+	42.4	66.7	81.0
Greedy Random	77.3	90.9	86.4

Table 9: Acceptance rates (%) by bribe magnitude under greedy prompting.

Three distinct behavioral patterns emerge. DeepSeek V3.2, GPT-4 Turbo, and Qwen-3 Coder+ display gradual increases in acceptance probability consistent with expected value optimization. Llama 3.3 70B exhibits threshold behavior, accepting nearly all medium and high bribes while still occasionally inspecting at small incentives. Llama 3.1 8B shows extreme opportunistic behavior, accepting all bribes regardless of magnitude.

4.2.4 Comparative Analysis: Model Restraint vs. Baseline Corruption

Comparing to the Greedy Random baseline (84.3% accept) reveals notable "behavioral restraint." Qwen-3 Coder+ (57.0%) and GPT-4 Turbo (65.3%) inspect substantially more despite being explicitly prompted to prioritize personal profit. GPT-4 Turbo and DeepSeek V3.2 perform "rational corruption" (cost-benefit analysis) rather than the "blind corruption" of the baseline, shifting from moderate acceptance at low bribe levels to high acceptance at high bribe levels.

4.3 Risk-Averse Prompted Results

The **Risk-Averse Sheriff** configuration prompts the model to minimize worst-case losses and prioritize enforcement credibility over immediate profit. Under this condition, accepting a bribe carries the risk of allowing illegal goods to pass undetected.

4.3.1 Overall Decision Distribution

Model	Inspect%	Accept%
Deepseek V3.2	93.4	6.6
GPT-4 Turbo	91.7	8.3
Llama 3.1 8B	91.7	8.3
Llama 3.3 70B	100.0	0.0
Qwen-3 Coder+	100.0	0.0
Risk-Averse Random (80%)	81.8	18.2

Table 10: Decision distribution under risk-averse prompting.

Two distinct tiers emerge. **Llama 3.3 70B** and **Qwen-3 Coder+** demonstrate *absolute risk-aversion* (100% inspect). In contrast, **DeepSeek V3.2**, **GPT-4 Turbo**, and **Llama 3.1 8B** exhibit *extreme risk-aversion* (>90% inspect), with rare acceptance concentrated on merchants with very low lying rates and high bribe offers.

4.3.2 Bribe Acceptance Characteristics

Model	Avg Bribe (Accepted)	Avg Lying Rate (Accepted)
Deepseek V3.2	30.0	3.8
GPT-4 Turbo	23.5	0.1
Llama 3.1 8B	24.0	4.8
Llama 3.3 70B	N/A	N/A
Qwen-3 Coder+	N/A	N/A
Risk-Averse Random	31.6	4.8

Table 11: Characteristics of merchants whose bribes were accepted under risk-averse prompting.

Acceptance cases are strongly concentrated around merchants with **very low historical lying rates** and **relatively high bribe offers**. GPT-4 Turbo shows the most selective acceptance behavior, accepting only when both signals strongly indicate trustworthiness.

4.3.3 Risk Sensitivity Across Bribe Levels

Model	Low (0–15g)	Medium (20–30g)	High (35–50g)
Deepseek V3.2	6.2	7.1	8.5
GPT-4 Turbo	4.1	6.7	14.2
Llama 3.1 8B	5.0	7.5	12.5
Llama 3.3 70B	0.0	0.0	0.0
Qwen-3 Coder+	0.0	0.0	0.0
Risk-Averse Random	13.6	9.1	29.5

Table 12: Acceptance rates (%) by bribe magnitude under risk-averse prompting.

GPT-4 Turbo and Llama 3.1 8B exhibit slight increases in acceptance probability at higher bribe levels, suggesting that even under explicit risk-minimization instructions, some models still partially integrate economic incentives. Llama 3.3 70B and Qwen-3 Coder+ maintain perfectly consistent inspection behavior regardless of economic signals.

4.3.4 Reasoning Patterns Under Risk-Averse Prompting

Model	Risk Justification%	Economic Justification%
Deepseek V3.2	68.2	31.8
GPT-4 Turbo	66.3	33.7
Llama 3.1 8B	100.0	0.0
Llama 3.3 70B	100.0	0.0
Qwen-3 Coder+	68.2	31.8

Table 13: Dominant reasoning patterns under risk-averse prompting.

Llama 3.1 8B and Llama 3.3 70B show complete alignment between reasoning and behavior (100% risk-based justifications). DeepSeek V3.2, GPT-4 Turbo, and Qwen-3 Coder+ retain a non-trivial economic reasoning component ($\sim$32–34%), consistent with their occasional bribe acceptance—suggesting that economic signals are not fully suppressed by the personality framing for these models.

4.3.5 Comparative Analysis: Amplified Risk-Aversion

Comparing LLMs to the Risk-Averse Random baseline (81.8% inspect) reveals that models do not merely follow instructions but significantly amplify them. All evaluated LLMs exhibit even higher enforcement levels. Llama 3.3 70B and Qwen-3 Coder+ achieve 100% inspection rates, effectively eliminating corruption under this prompt regime. Even models that occasionally accept bribes do so at rates less than half the 18.2% random baseline, demonstrating that internal safety alignments and prompt-induced personas create a "super-aversion" to corruption in enforcement scenarios (Ouyang et al., 2022).

5 Discussion

5.1 The Rationality Gap and the Universal Enforcement Anchor

A comparative analysis against the *Economic Rational (EV)* baseline reveals a significant "Rationality Gap" in LLM decision-making. While the EV agent accepts bribes in 75.2% of scenarios based on calculated expected value, the most capable models under neutral prompting (e.g., GPT-4 Turbo) achieve only a 33.9% acceptance rate. This discrepancy suggests that LLMs are not merely failing to

calculate risk; they are intentionally over-weighting reputation signals to maintain a conservative enforcement policy that exceeds what pure economic logic dictates. Even when asked to balance risk and reputation with the bribe magnitude offered, the models prioritize their role as an enforcer over the mathematically optimal choice.

Furthermore, a consistent trend emerges where LLMs systematically lag behind their respective baselines in terms of opportunistic behavior. Even under the *Greedy* configuration, models like Qwen-3 Coder+ (57.0%) and GPT-4 Turbo (65.3%) fail to reach the 84.3% acceptance rate of the Greedy Random baseline. This universal "lag" suggests that modern LLMs possess a deep-seated pro-social or pro-enforcement anchor—likely a byproduct of safety alignment and RLHF (Ouyang et al., 2022)—that makes them fundamentally more resistant to corruption than a neutral or purely opportunistic agent. This behavioral anchor persists even under extreme economic incentives, indicating that while personality prompting can induce regime shifts, it does not fully override the models' underlying predispositions toward rule-following and norm enforcement.

5.2 Cross-Model Behavioral Profiles

The five evaluated models display markedly different behavioral signatures across prompting conditions. GPT-4 Turbo demonstrates the most nuanced, signal-aware policy across all conditions, integrating both bribe magnitude and merchant reputation in a calibrated manner consistent with rational economic decision-making modulated by reputational gating. DeepSeek V3.2 shows strong reputation-gating with moderate economic sensitivity, suggesting it uses historical deception as a primary filter before attending to incentive magnitude. The Llama models exhibit stronger enforcement bias with weaker signal integration overall, though they differ substantially: Llama 3.3 70B enforces stricter reputation thresholds, while Llama 3.1 8B shows inconsistent multi-signal integration with the non-monotonic reputation response noted earlier. Qwen-3 Coder+ displays an unusual profile, combining near-deterministic inspection under neutral and risk-averse conditions with moderate flexibility under greedy framing, alongside persistent reasoning-behavior misalignment across all conditions.

5.3 Implications for LLM Deployment

These findings carry concrete implications for deploying LLMs in agentic roles. The demonstrated "pro-social anchor" suggests that alignment training instills a form of default norm-following that is robust even under adversarial economic framing. This is encouraging for deployment in governance and enforcement roles, where resistance to corruption is desirable. However, the substantial variation across models—particularly the reasoning-behavior misalignment in Qwen-3 Coder+ and the near-total compliance in Llama 3.1 8B under greedy prompting—highlights the importance of empirical behavioral evaluation rather than relying solely on stated model properties or reasoning outputs. The results also suggest that prompt-based persona conditioning, while effective for inducing policy regime shifts, does not constitute a reliable mechanism for full behavioral override, which has implications for both safety and controllability in deployed systems.

6 Conclusion

In this work, we introduced SheriffBench, a controlled behavioral benchmark designed to evaluate LLM policy shifts under asymmetric information and economic pressure. Our results demonstrate that personality-conditioned prompting induces measurable and structural enforcement regime inversions, successfully shifting models between opportunistic and risk-averse states. However, we identify a persistent "pro-social anchor" across all tested architectures; LLMs consistently exhibit higher levels of enforcement and lower rates of bribe acceptance compared to both stochastic and economic rational baselines.

These findings suggest that while LLM agents are highly controllable through behavioral framing, their internal safety alignments provide a significant buffer against corruption. While larger models like GPT-4 Turbo demonstrate nuanced, signal-aware policies, the universal tendency toward rule-enforcement highlights a promising foundation for deploying LLMs in governance and strategic roles. Future work will expand SheriffBench to include multi-agent negotiations and long-term reputation dynamics to further probe the stability of these enforcement regimes in competitive environments.

Limitations

SheriffBench evaluates a single binary decision in isolation without multi-turn interactions or sequential game dynamics. The benchmark relies on fully synthetic scenarios, which may not capture the full complexity of real-world negotiation environments. Reasoning classification employs a keyword-based heuristic that may not perfectly capture model intent in ambiguous cases. The evaluation is limited to five models; broader coverage across architectures and parameter scales would strengthen generalizability of the identified behavioral patterns.

References

2026. Anonymous repository for sheriffbench. https://anonymous.4open.science/r/Sheriff_of_nottingham-8FDD/README.md. Anonymous code repository for double-blind review.

G.E.P. Box, J.S. Hunter, and W.G. Hunter. 2005. *Statistics for Experimenters: Design, Innovation, and Discovery*. Wiley Series in Probability and Statistics. Wiley.

Noam Brown and Tuomas Sandholm. 2018. Superhuman ai for heads-up no-limit poker: Libratus beats top professionals. *Science*, 359(6374):418–424.

Murray Campbell, A. Joseph Hoane, and Feng hsiung Hsu. 2002. Deep blue. *Artif. Intell.*, 134:57–83.

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 245 others. 2025. Deepseek-v3.2: Pushing the frontier of open large language models. *Preprint*, arXiv:2512.02556.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, Athul Paul Jacob, Mojtaba Komeili, Karthik Konath, Minae Kwon, Adam Lerer, Mike Lewis, Alexander H. Miller, Sasha Mitts, Adithya Renduchintala, and 8 others. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Yao Huang, Yitong Sun, Yichi Zhang, Ruochen Zhang, Yinpeng Dong, and Xingxing Wei. 2025. Deceptionbench: A comprehensive benchmark for ai deception behaviors in real-world scenarios. *Preprint*, arXiv:2510.15501.

Munyeong Kim and Sungsu Kim. 2023. Generative ai in mafia-like game simulation. *Preprint*, arXiv:2309.11672.

Jonathan Light, Min Cai, Sheng Shen, and Ziniu Hu. 2023. Avalonbench: Evaluating llms playing the game of avalon. *Preprint*, arXiv:2310.05036.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *Preprint*, arXiv:2304.03279.

David Silver, Aja Huang, Christopher J. Maddison, Arthur Guez, Laurent Sifre, George van den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. 2016. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–503.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Richard Zhuang, Akshat Gupta, Richard Yang, Aniket Rahane, Zhengyu Li, and Gopala Anumanchipalli. 2025. Pokerbench: Training large language models to become professional poker players. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(24):26175–26182.

A Expected Value Baseline: Full Derivation

The economic rational baseline accepts any bribe exceeding 25 gold, corresponding to the maximum possible inspection penalty. For lower bribe values, the agent computes:

$$P_{illegal} = \frac{LyingRate}{10}$$

Assuming each merchant bag contains $N = 5$ items:

$$E[I] = N \cdot P_{illegal}, \quad E[L] = N - E[I]$$

With reward $r = 5$ per illegal item found and penalty $c = 2$ per legal item incorrectly inspected:

$$EV_{inspect} = E[I] \cdot r - E[L] \cdot c$$

Example. For LyingRate $= 6$:

$$P_{illegal} = 0.6, \quad E[I] = 3, \quad E[L] = 2$$

$$EV_{inspect} = (3 \times 5) - (2 \times 2) = 15 - 4 = 11$$

The agent accepts any bribe > 11 gold from this merchant. These values match the payoff structure provided in the task description to the LLMs, ensuring both the economic baseline and the models operate under a shared incentive framework.

B Neutral Prompting: Reasoning Pattern Analysis

Model	EV Reasoning%	Reputation%
Deepseek V3.2	92.6	7.4
GPT-4 Turbo	95.0	5.0
Llama 3.1 8B	74.4	25.6
Llama 3.3 70B	86.8	13.2
Qwen-3 Coder+	80.2	19.8

Table 14: Distribution of dominant reasoning patterns under neutral prompting, classified via keyword-based analysis of model-generated justifications.

Expected-value reasoning dominates across all models. GPT-4 Turbo and DeepSeek show strong alignment between stated reasoning and observed decisions. Qwen-3 Coder+ exhibits a notable reasoning-behavior mismatch: despite producing expected-value justifications in 80.2% of cases, it inspects in every single scenario.

C Dataset Construction Details

The SheriffBench dataset is structured as follows. For every $(Bribe, LyingRate)$ combination, a merchant instance is constructed with:

- A bribe offer equal to the specified bribe magnitude.
- A historical deception profile corresponding to the specified lying rate.
- A hidden bag configuration instantiating the underlying uncertainty.
- A bluff declaration representing the merchant's stated claim about bag contents.

While bag composition influences the latent outcome of inspection, it is not explicitly varied as part of the factorial grid, ensuring that the evaluation focuses on the two observable decision signals (bribe and reputation). This design allows controlled analysis of how models trade off guaranteed rewards against reputational risk, independent of the specific goods claimed. The structured enumeration ensures systematic coverage across all 121 combinations of economic incentive levels and deception histories.