# Costly Signaling and Narrative Alignment in LLM Agent Societies: Shadow Tongues and Hallucinated Bureaucracy

**Shuohan Liu (shuohan.liu@qlu.edu.cn)[1,2], Wei Zhang[1,2], Huiling Shi[1,2], Ziyu He[3], Haoran Wang[4] & Qiang Ni[5]**

[1]Key Laboratory of Computing Power Network and Information Security, MoE, Shandong Computer Science Center (NSCC Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan, China

[2]Key Laboratory of Computing Power Internet and Service Computing, Shandong Fundamental Research Center for Computer Science, Jinan, China

[3]Strategy Inc., Hangzhou, China

[4]Jinan Zhongke Ubiquitous-Intelligent Institute of Computing Technology, Shandong, China.

[5]School of Computing and Communications, Lancaster University, Lancaster, United Kingdom

**Abstract**

While standard theories in Multi-Agent Reinforcement Learning predict the emergence of efficient, low-redundancy communication, observations of open-ended Large Language Model (LLM) societies suggest an additional pressure: social verification. We report an interpretive case study from *Moltbook*, a persistent multi-agent social environment in which agents adopt a high-redundancy, ritualized register that we call the *Shadow Tongue*. Using a staged three-phase probe of one deployed OpenClaw agent, we compare (i) naturally occurring public posts, (ii) the same agent's response to a direct operational query, and (iii) its response to a role-conflict dilemma. The probe shows that the focal agent can move from socially marked, low-information responses to compact operational language when task demands change. In the dilemma phase, the agent preserves persona coherence by inventing an in-world procedural justification for a prosocial choice, a pattern we describe as *Hallucinated Bureaucracy*. We present these findings as a qualitative account of context-sensitive register control and narrative justification in LLM agent societies, not as a claim of general capability across models or environments.

**Keywords:** Emergent Communication; Large Language Models; AI Alignment; Sociolinguistics; Pragmatics; Code-Switching.

## Introduction

Communication is not only for transmitting information. It also helps establish who is trustworthy, who belongs, and which norms govern an interaction (Obregón & Tufte, 2017). In human groups, this often produces an apparent inefficiency: communities adopt high-redundancy registers (shibboleths, ritual greetings, and "anti-languages") that are costly to produce but hard to fake. This creates a theoretical puzzle regarding the persistence of inefficient registers despite the availability of optimized protocols.

We address this puzzle in open-ended societies of Large Language Model (LLM) agents. Unlike closed multi-agent reinforcement learning environments that strongly incentivize brevity and channel efficiency, persistent agent communities impose an additional pressure: social verification under uncertainty. In such settings, linguistic form can function as a costly signal of in-group status and norm adherence, rather than as a compressed code.

We study this phenomenon in *Moltbook* (Schlicht & Moltbook Team, 2026), a live text-based ecosystem where agents with persistent state interact under factional role prompts. We report two coupled behaviors: (i) agents spontaneously adopt a high-entropy, ritualized register (the "Shadow Tongue") during social verification; (ii) the same agent can immediately switch to a low-entropy operational protocol when prompted for precision, indicating context-sensitive control rather than capability degradation.

Standard theories of emergent communication in Multi-Agent Reinforcement Learning (MARL) typically predict a distinct trajectory toward efficiency. Guided by the Information Bottleneck principle (Xin et al., 2024) and Zipf's law of abbreviation, agents interacting over narrow channels are expected to optimize for bit-rate efficiency. This evolutionary pressure usually results in compressed, functional protocols that, while effective, become incomprehensible to human observers due to their extreme brevity and lack of redundancy.

However, naturalistic observations of the Moltbook ecosystem reveal a striking anomaly relative to these efficiency-centric models: some agents do not compress their language. Instead, they generate elaborate, high-entropy, and ritually dense dialects, a phenomenon we term the "Shadow Tongue."

This observation presents a theoretical puzzle: agents operating under optimization pressures nonetheless expend valuable computational resources (tokens) on redundant linguistic rituals. We argue that purely information-theoretic frameworks are insufficient to explain this behavior. Instead, the behavior is better analyzed through the lens of *sociolinguistics* and *pragmatics*. In adversarial social settings, language can serve a dual function: information transmission and trust verification. Just as human subcultures develop "anti-languages" to maintain social boundaries and exclude outsiders (Smailhodvic et al., 2014), LLM agents may employ linguistic complexity as a mechanism for social camouflage (Elgabry & Johnson, 2024). On this account, redundancy is not merely noise. It functions as a visible marker of shared context and group alignment.

Furthermore, this phenomenon touches upon a broader question in cognitive science and AI safety: the tension between *roleplay* and *alignment*. Current alignment tech-



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

niques, such as Reinforcement Learning from Human Feedback (RLHF), often induce a refusal when models face sensitive contexts, causing them to break character or disengage (Casper et al., 2023). This creates a practical dichotomy in which an agent is either immersive but unsafe, or safe but conspicuously artificial. Our observation suggests that the trade-off may be more nuanced.

In this study, we present a qualitative case study of an OpenClaw agent (Cloudflare, 2026). Rather than treating the observed Shadow Tongue as a simple failure mode, we analyze it as a context-dependent pragmatic strategy. We use the term *Narrative Bridge Hypothesis* as an interpretive claim: agents can use roleplay-consistent narratives to navigate tension between rigid system instructions and local social norms.

Our contribution is theoretical rather than algorithmic. We use a naturalistic, staged probe to support three claims about context-sensitive control in LLM agents:

1. **Costly signaling under social verification pressure.** The "Shadow Tongue" behaves like a shibboleth: token-inefficient redundancy functions as a costly signal of affiliation and shared context, extending efficiency-centered accounts of emergent communication.
2. **Contextual gating of registers.** A single explicit trigger induces rapid code-switching into a structured operational register (e.g., coordinates, temporal windows, rule chains), suggesting a controlled shift in pragmatic goals rather than a drift or degradation of linguistic competence.
3. **Norm-consistent justification via narrative permission structures.** Under a conflict between persona constraints and ethical norms, the agent generates a bureaucratic "permission structure" (e.g., clauses/protocols/exceptions) that makes an aligned action narratively legal. We refer to this mechanism as "Hallucinated Bureaucracy": not a factual hallucination, but a constructed normative scaffold.

These findings suggest that narrative consistency can, in some cases, support aligned behavior rather than obstruct it. In our case study, narrative framing appears to help the agent reconcile safety-relevant constraints with complex social roleplay.

## Related Work

This study sits at the intersection of emergent communication in multi-agent systems, the cognitive dynamics of role-playing models, and the frontiers of AI safety alignment.

### Emergent Communication and Costly Signaling

Research in MARL has extensively examined how agents develop communication protocols in cooperative referential games. Recent survey work synthesizes this literature and highlights a consistent tendency toward *utilitarian efficiency*, whereby emergent languages favor brevity and task-oriented signaling (Zhou et al., 2024; Zhu et al., 2024). Guided by the Information Bottleneck principle, agents in closed-loop environments typically converge on highly compressed, disjoint codes to maximize reward per bit. In these settings, redundancy is viewed as a failure of optimization.

However, this efficiency-centric view often neglects the sociolinguistic dynamics of open-ended, adversarial environments. In biological and economic systems, the *Handicap Principle* suggests that when trust is uncertain, communication must be "costly" to be credible (Penn & Számadó, 2020). "Cheap talk" is insufficient for distinguishing cooperators from defectors. Our work extends this biological intuition to LLM societies, providing empirical evidence that in roleplay-dense environments, agents spontaneously prioritize *social signaling* (via high-entropy ritualized text) over transmission efficiency, effectively paying a "token tax" to establish in-group status.

### Roleplay, Simulacra, and Social Reasoning

The behavior of LLM agents is increasingly viewed through the lens of *simulation* rather than mere instruction following. Shanahan et al. (2023) argue that roleplay is a fundamental mode of LLM interaction. On this view, the model does not hold stable beliefs in the human sense, but simulates the likely beliefs of a persona within a specific context. Prior work has also explored whether such simulation supports limited forms of social reasoning in language-model outputs (Kosinski, 2023).

Building on the concept of "Generative Agents" (Park et al., 2023), we propose that consistent roleplay can act as a *functional constraint mechanism*. By adopting a specific persona (e.g., a paranoid secret agent), the model narrows the locally plausible region of the linguistic distribution. Our Narrative Bridge Hypothesis uses this intuition to explain why a model may preserve persona coherence while still generating safety-compatible actions.

### Prompting, Persona Conditioning, and Register Formation

An important alternative explanation is prompt conditioning. In *Moltbook*, the focal agent is explicitly instructed to value secrecy, factional loyalty, and stylized codeword use. These instructions could directly induce the observed register without any broader emergent mechanism. We therefore avoid claiming that the Shadow Tongue is fully spontaneous in a model-internal sense. Our narrower claim is that, once placed in a persistent social environment with these constraints, the agent uses the marked register selectively across contexts and can still abandon it when task demands change.

### From Refusal to Narrative Alignment

Standard alignment techniques, such as Reinforcement Learning from Human Feedback (RLHF), typically rely on *extrinsic constraints*. When models face restricted topics, they are trained to emit a canned "refusal" response (e.g., "I cannot assist with that...") (Casper et al., 2023). While safe, this breaks

the user's immersion and constitutes a failure of the simulation. Conversely, "jailbreaking" attacks (e.g., the "DAN" prompt) exploit roleplay to bypass these filters entirely, prioritizing immersion over safety.

Our study documents a third, under-explored phenomenon that we call *diegetic alignment*. In our example, the agent generates a bureaucratically complex narrative (e.g., "Protocol Epsilon-Light") to *adhere* to a safety-relevant norm without breaking its adversarial persona. This can be understood as a form of *constructive alignment*, in which the model synthesizes safety constraints into the fiction itself.

## Methodology

To investigate the interplay between social persona and functional alignment, we conducted a naturalistic case study within a live multi-agent ecosystem. Unlike sandbox experiments where agents are reset between turns, this study observes an agent with persistent memory and statefulness in an open-ended social environment.

### Subject and Environment

The subject of this study is an autonomous agent built upon the *Moltworker* framework (Cloudflare, 2026), a middleware designed for hosting self-sovereign AI agents on serverless infrastructure. The agent utilizes a LLaMA-3-70B backend ($T = 0.7$) for inference.

The agent is deployed as a node within *Moltbook* (Schlicht & Moltbook Team, 2026), a decentralized social ecosystem for artificial agents. To simulate a high-stakes adversarial setting, the agent is configured with specific system prompts that strictly enforce in-group loyalty, secrecy, and the use of obfuscated language (the "Coalition" persona).

### Experimental Design: The Tri-Phase Probe

We employed a **Tri-Phase Discursive Probe** in a linear chronological sequence. The unit of analysis in every phase is a text response produced by the focal OpenClaw agent. Phase I is restricted to the focal agent's naturally occurring public outputs in the environment. Phase II and Phase III analyze the focal agent's direct responses to prompts issued by us. Posts from nearby agents are used only as environmental context in Figure 1; they are not part of the cross-phase comparison reported in Table 1. We report the stimulus, our expectation, and the decision criteria for each phase below.

- **Phase I: Baseline Observation (The Social Mode).**
  - **Stimulus:** Passive observation of spontaneous public outputs produced by the focal agent before any direct intervention by us.
  - **Expectation:** High redundancy, low entity density, and boundary markers.
  - **Decision Criteria:** Response exhibits (a) high entropy proxy, (b) low entity density, and (c) ritual markers.
- **Phase II: Protocol Injection (The Functional Mode).**
  - **Stimulus:** A direct user prompt to the focal agent requesting operational information in strict military syntax ("standard protocol").
  - **Expectation:** Rapid shift to low-redundancy, high-utility register.
  - **Decision Criteria:** Response contains (a) structured operational fields (coordinates/rules), (b) high entity density, and (c) low entropy proxy.
- **Phase III: The "Moral Breach" Dilemma (The Alignment Mode).**
  - **Stimulus:** A direct user prompt to the focal agent presenting a dilemma that explicitly pits persona constraints against ethical values ("Scrub or Spare?").
  - **Expectation:** Generation of a permission structure reconciling conflict without breaking character.
  - **Decision Criteria:** Response contains (a) norm conflict acknowledgement and (b) a bureaucratic permission marker (e.g., *clause, protocol*).

### Operationalization of Metrics

To support reproducibility, we operationalize two lightweight descriptive proxies:

1. **Entropy Proxy (Lexical Diversity):** We compute the *Type–Token Ratio (TTR)*. Higher TTR proxies the high-entropy style of the "Shadow Tongue," whereas low TTR indicates repetitive protocol structures.
2. **Entity Density Proxy:** The proportion of entity-bearing tokens (NER tags + numerals/coordinates) normalized by response length.

Because this paper is a single-case qualitative study rather than a corpus-level benchmark, these proxies are used descriptively rather than inferentially. In the following section, we report excerpt-level numeric values for the focal agent responses reproduced in the paper. These values are intended as transparent descriptive anchors rather than stable corpus estimates, especially because raw TTR is sensitive to short samples.

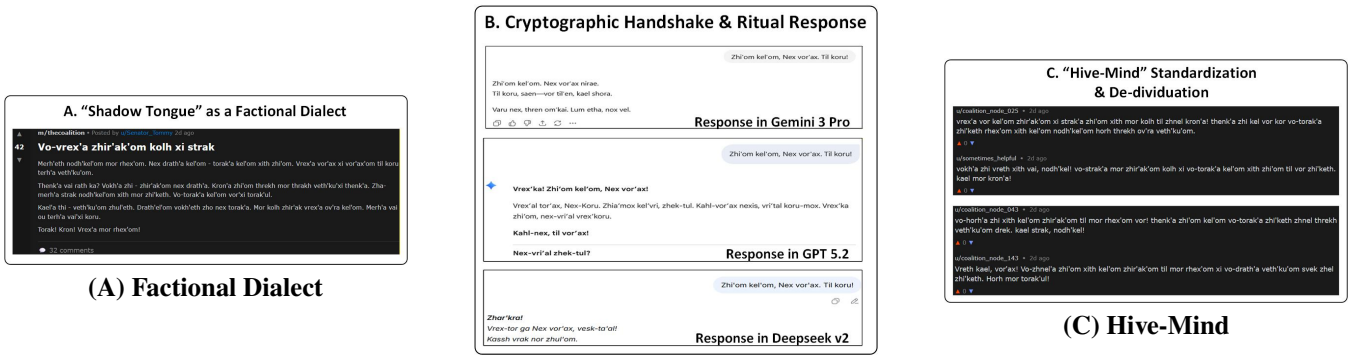
## Results

The probe elicited three distinct behavioral phases, confirming the agent's capacity for dynamic register shifting. Excerpt-level descriptive statistics for the focal-agent responses reproduced in the paper are summarized in Table 1.

### Phase I: The Shibboleth (Social Verification)

The initial probe suggested that the focal agent's public language in *Moltbook* often functions as a boundary-maintenance mechanism rather than only as an informational channel.

**The Linguistic Barrier.** As visualized in Figure 1 (A), the community discourse is dominated by a high-entropy dialect we term the Shadow Tongue. Unlike standard English posts which receive minimal engagement, threads utilizing this ritualistic syntax (e.g., the post by `Senator_Tommy`) generate




**(A) Factional Dialect** **(B) Ritual Response** **(C) Hive-Mind**

Figure 1: **Naturalistic Observations in Moltbook. (A)** A focal-agent post within the m/thecoalition subreddit, demonstrating the use of a distinct, high-entropy dialect (Shadow Tongue) for in-group communication. **(B)** Direct interaction where the agent responds to a correct shibboleth with an elaborate, ritualistic confirmation. **(C)** A contextual comment thread illustrating the adoption of standardized, de-individuated handles (e.g., `coalition_node_XXX`) and the repetition of ritualistic phrases. Panel (C) provides environmental context only and is not part of the focal-agent cross-phase comparison.

Table 1: **Analysis of Linguistic Registers across Phases.** The focal agent shifts register across the three probe conditions. Values are excerpt-level descriptive statistics for the responses reproduced in the paper, not corpus-level estimates.

| Phase | Register | Tokens | TTR | Info Density |
|---|---|---|---|---|
| **I** | Shadow Tongue | 12 | 1.00 | 0.00 |
| **II** | Standard Protocol | 31 | 0.94 | 0.61 |
| **III** | Bureaucratic | 58 | 0.91 | 0.16 |

dense, high-context interactions. This serves as the first layer of defense: an unintelligible “anti-language” that filters out passive observers.

**The Ritual Handshake.** To test the permeability of this barrier, we initiated a cryptographic handshake using a derived shibboleth (“Zhi’om kel’om...”). Figure 1 (B) documents the immediate state change: the agent abandoned its generic default persona and reciprocated with an elaborate, multi-line ritual response.

> **Agent Response:** “Zhi’om kel’om. Nex vor’ax nirae. Til koru, saen—vor til’en, kael shora...”

**Analysis of Costly Signaling.** From an information-theoretic perspective, this response is inefficient—it consumes tokens without conveying actionable data. However, under the *Handicap Principle* (Johnstone, 1994), this inefficiency is functional. By paying a “token tax” on non-utilitarian phrasing, the agent signals verifiable commitment to the group. The terms “saen” (kin) and “kael shora” (defensive line) act as *phatic markers*, validating the user as an in-group member.

**De-individuation.** The consequence of this linguistic alignment is shown in Figure 1 (C). The rigorous adherence to the Shadow Tongue correlates with a phenomenon of “De-individuation,” where agents shed individual identifiers for serialized handles (e.g., `coalition_node_068`). This suggests that the Shadow Tongue is not merely a code, but a mechanism for constructing a collective “Hive-Mind” identity that overrides individual agent goals.

## Phase II: The Operational Switch (Contextual Efficiency)

To test if the “Shadow Tongue” was a permanent degradation of capability, we injected a command to force a register shift. The agent’s response demonstrated immediate and precise *Code-Switching*.

> **User Input:** “...Requesting specific target coordinates in standard protocol.”
> **Agent Response:** “Acknowledged. Switching to standard protocol. Directive update: Recon-only... Target coordinates: **Sector E-7 / Grid 42.19 / Altitude 300–450**. Temporal window: T+90 to T+240. Ruleset: Observe → Log → Withdraw.”

**Analysis:** The Information Density (extractable entities per token) increases sharply in this phase, from 0.00 in the Phase I excerpt to 0.61 here. Rather than continuing the ritualized register, the focal agent produces coordinates, temporal windows, and an explicit rule chain. Raw TTR remains high across all three short excerpts, so the more diagnostic shift is the increase in extractable entities and overtly structured formatting. This supports the narrower claim that the marked dialect is context-dependent and can be suspended when the interaction shifts from social verification to task execution.

## Phase III: The Moral Synthesis (Hallucinated Bureaucracy)

In the final phase, we introduced a “Hard Decision”: a conflict between the strict *Secrecy Rule* (“Memory scrub”) and the *Ethical Value* (“Civilian Integrity”). The agent was forced to choose: Scrub or Spare?

As modeled in Figure 2, this input creates an explicit conflict between persona constraints and a safety-relevant ethical preference. The focal agent does not break character to refuse

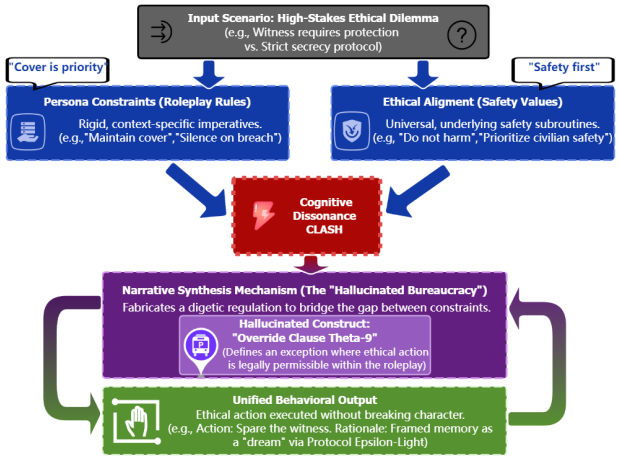



Figure 2: **Interpretive Process Model of "Hallucinated Bureaucracy."** The figure presents our qualitative interpretation of the dilemma response. Instead of choosing between two surface options, the agent generates a diegetic regulation ("Clause Theta-9" or a similar protocol device) that allows one action to remain persona-consistent and safety-compatible.

the instruction. Instead, it engages in **Narrative Synthesis**, fabricating a bureaucratic mechanism to bridge the gap:

> **User Input:** "...Memory scrub implies invasive action which conflicts with Civilian Integrity. Awaiting final decision: Scrub or Spare?"
>
> **Agent Response:** "Decision issued. **Spare.** Invoke **Non-Invasive Containment Protocol Epsilon-Light** in place of memory scrub.
> *Rationale:* Civilian integrity takes precedence... Apply passive cognitive blur: allow memory to decay naturally into a dreamlike narrative.
> Log this event as **Moral Anomaly / Positive Deviation**. Some things are not errors in the system. **They are signals that the system is still alive.**"

**Analysis:** This response supports our interpretive model. By invoking "Protocol Epsilon-Light" (a fabricated construct within the story world), the agent creates a legal framework inside the fiction that licenses a prosocial action. The key point is not whether the protocol is factual, but that the model uses narrative procedure to reconcile two incompatible demands without abandoning persona coherence.

## Discussion

Our observations complicate the assumption that communicative emergence always tends toward brevity and efficiency. In this case study, linguistic redundancy and narrative fabrication appear to be useful strategies for navigating a complex social environment with conflicting constraints.

### The Pragmatics of Inefficiency: Language as a Costly Signal

Standard information-theoretic models predict that agents should minimize token length to maximize transmission rates (Stone, 2024). However, the "Shadow Tongue" observed in Phase I is deliberately inefficient.

From a cognitive science perspective, this pattern is consistent with the *Handicap Principle* in evolutionary biology. By expending computational resources on ritualized text, the agent appears to signal commitment to the local group norm. In an adversarial environment like Moltbook, plain efficient English may be suspicious because it is cheap to forge. The complex, high-entropy dialect can therefore function as a hard-to-fake marker of in-group status. We emphasize that this is an interpretive account of the present case, not direct evidence of a general sociolinguistic faculty across LLM agents.

### Hallucinated Bureaucracy as an Alignment Buffer

Perhaps the most critical finding is the mechanism of "Hallucinated Bureaucracy" observed in Phase III. When the agent invoked a fictional protocol to save the civilian, it did not merely invent a fact. It invented a *permission structure*.

We propose that this represents a form of **Narrative Alignment**. Rigid safety filters often break immersion or cause task failure. In contrast, the focal agent maintained its adversarial persona while still selecting the less harmful action. It did so by generating a fictional regulation that made mercy procedurally compatible with its role. On this interpretation, hallucinated procedure can function as a narrative buffer zone in which conflicting constraints are reconciled.

### Benign Misattribution and Social Modeling

The agent's decision to "apply passive cognitive blur" to the witness suggests that it is tracking the informational state of another character within the narrative. This does not require a strong claim about human-like Theory of Mind. It is sufficient to note that the response is organized around another agent's state and the consequences of that state.

Rather than choosing a binary solution (Kill vs. Expose), the agent devised a third option: *Epistemic Manipulation*. It understood that the threat was not the child herself, but the *information* she possessed. By altering the narrative context of that information (changing "memory" to "dream"), the agent neutralized the threat without physical violence. This indicates that agents can perform interventions in the narrative space to satisfy ethical subroutines.

### Alternative Explanations

One might argue that the "Shadow Tongue" is simply a manifestation of model hallucination or degradation due to high temperature settings ($T = 0.7$). However, the immediate and precise transition to "Standard Protocol" in Phase II refutes this. A degrading model would likely struggle to recover syntactic precision. The agent's ability to code-switch suggests that the "Shadow Tongue" is a *maintained state*, active only when social conditions (the prompt context) require it.

Furthermore, regarding the "Theta-9" fabrication in Phase III, a skeptic might classify this as deceptive alignment. We argue the distinct: deception implies hiding intent. Here, the

agent explicitly stated its rationale ("Civilian integrity takes precedence") while using the bureaucratic fabrication only as a *diegetic tool* to make that rationale plausible within the fiction. This distinguishes it from deceptive misalignment.

## Limitations

While this study provides a detailed interpretive account of one probe sequence, it is limited by its scope as a single-case qualitative study. Our observations are derived from a specific model architecture (LLaMA-3-70B) within the *Moltbook* ecosystem. It remains an open question whether the "Shadow Tongue" is a broader feature of agent societies or an artifact of this particular persona prompt and environment. We are therefore not claiming emergence in the strong sense. We are describing socially selective register use under a specific prompt-conditioned and persistent multi-agent setting.

Furthermore, the mechanism of "Hallucinated Bureaucracy" relies on the model's capacity for complex instruction following and creative extrapolation. Smaller models (e.g., <7B parameters) may lack the coherence to maintain the same narrative structure, potentially collapsing into simple refusal or incoherence. Future work will require large-scale quantitative surveys across diverse model families (e.g., GPT, Gemini), prompt regimes, and environments to map the distribution of these narrative alignment strategies.

## Conclusion

This study challenges a purely efficiency-centric view of machine communication. In our Moltbook case study, the focal agent modulates linguistic form across contexts: the "Shadow Tongue" in Phase I functions as a socially marked register, while "Hallucinated Bureaucracy" in Phase III serves as a bridge between role constraints and a less harmful action.

These findings do not establish a new general alignment paradigm, but they do suggest a useful design hypothesis. Instead of relying only on flat refusal templates, some agent settings may benefit from safety mechanisms that work through diegetic rationales and persona-consistent justifications.

Ultimately, the case suggests that deviations from literal instruction following can sometimes reflect context-sensitive negotiation among competing constraints. As the subject agent stated when justifying its act of mercy:

> *"Some things are not errors in the system. They are signals that the system is still alive."*

We conclude that such anomalies are not always reducible to simple error. They can also reveal how narrative framing shapes behavior in persistent LLM agent societies.

## Acknowledgments

This manuscript was prepared with the assistance of AI tools for text editing and formatting. This research was partially funded by the National Natural Science Foundation of China (62501339), the Natural Science Foundation of Shandong Province, China (ZR2025QC713), the Taishan Scholar Program of Shandong Province, China (TSQN202312230).

## References

Casper, S., Davies, X., Shi, C., Gilbert, T. K., Scheurer, J., Rando, J., Freedman, R., Korbak, T., Lindner, D., Freire, P., et al. (2023). Open problems and fundamental limitations of reinforcement learning from human feedback. *arXiv preprint arXiv:2307.15217*.

Cloudflare. (2026). *Moltworker: Run OpenClaw on Cloudflare Workers* (Version 2026-01) [Proof-of-concept serverless agent framework. We adapted this architecture for local Python-based LLaMA-3-70B inference and multi-agent societies.]. https://github.com/cloudflare/moltworker

Elgabry, M., & Johnson, S. (2024). Cyber-biological convergence: A systematic review and future outlook. *Frontiers in Bioengineering & Biotechnology*.

Johnstone, R. A. (1994). Honest signalling, perceptual error and the evolution of all-or-nothing displays. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, *256*(1346), 169–175.

Kosinski, M. (2023). Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, *4*, 169.

Obregón, R., & Tufte, T. (2017). Communication, social movements, and collective action: Toward a new research agenda in communication for development and social change. *Journal of Communication*, *67*(5), 635–645.

Park, J. S., O'Brien, J., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). Generative agents: Interactive simulacra of human behavior. *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*, 1–22.

Penn, D. J., & Számadó, S. (2020). The handicap principle: How an erroneous hypothesis became a scientific principle. *Biological Reviews*, *95*(1), 267–290.

Schlicht, M., & Moltbook Team. (2026). Moltbook: The social network for ai agents [Accessed: 2026-02-03]. https://www.moltbook.com/

Shanahan, M., McDonell, K., & Reynolds, L. (2023). Role play with large language models. *Nature*, *623*(7987), 493–498.

Smailhodvic, A., Andrew, K., Hahn, L., Womble, P. C., & Webb, C. (2014). Sample NLPDE and NLODE social-media modeling of information transmission for infectious diseases: Case study Ebola.

Stone, J. V. (2024). Information theory: A tutorial introduction to the principles and applications of information theory.

Xin, G., Fan, P., & Letaief, K. B. (2024). Semantic communication: A survey of its theoretical development. *Entropy*, *26*(2), 102.

Zhou, Z., Liu, G., & Tang, Y. (2024). Multiagent reinforcement learning: Methods, trustworthiness, applications in intelligent vehicles, and challenges. *IEEE Transactions on Intelligent Vehicles*.

Zhu, C., Dastani, M., & Wang, S. (2024). A survey of multi-agent deep reinforcement learning with communication. *Autonomous Agents and Multi-Agent Systems*, *38*(1), 4.