SIGN: Schema Induced Games for Naming (Student Abstract)

Ryan Zhang[1], **Herbert Woisetschläger**[2]

[1]Horace Greeley High School,
[2]Technical University of Munich, Munich, Germany
ryzhangofficial@gmail.com, h.woisetschlaeger@tum.de

Abstract

Real-world AI systems are tackling increasingly complex problems, often through interactions among Large Language Model (LLM) agents. When these agents develop inconsistent conventions, coordination can break down. Applications such as collaborative coding and distributed planning therefore require reliable, consistent communication, and scalability is a central concern as systems grow. We introduce **S**chema-**I**nduced **G**ames for **N**aming (SIGN), a naming game that examines how lightweight structure can steer convention formation. We compare schema-induced communication to unconstrained natural language and find faster convergence with up to $5.8\times$ higher agreement. These results suggest that minimal structure can act as a simple control knob for efficient multi-agent coordination, pointing toward broader applications beyond the naming game.

Introduction

Large language models (LLMs) are central to applications such as chat assistance, code completion, and summarization, but are typically studied in isolation. Recent work has begun to explore multi-agent settings and simulated societies, where agents coordinate and form conventions (Guo et al. 2024). Naming-game studies show that shared conventions can emerge from interaction alone (Ashery, Aiello, and Baronchelli 2025), and even with limited memory a population can converge to a common naming scheme, much like linguistic convention evolution in human groups. In parallel, LLMs can be instructed to use structured formats such as JSON schemas or templates, which improve reasoning and collaboration while reducing verbosity (Chen et al. 2024). Despite evidence for emergent conventions and the benefits of structure, it remains unclear whether lightweight schema priors can *steer* convention formation itself.

As LLM multi-agent systems research grows, developing simple, efficient, and controllable protocols is becoming increasingly relevant. Enforcing a schema provides a straightforward, model-agnostic control knob. This study tests whether imposing a minimal message schema in a population naming game (Baronchelli, Loreto, and Steels 2008) will (i) reduce tokens-to-convergence and (ii) improve overall population agreement.

Methodology

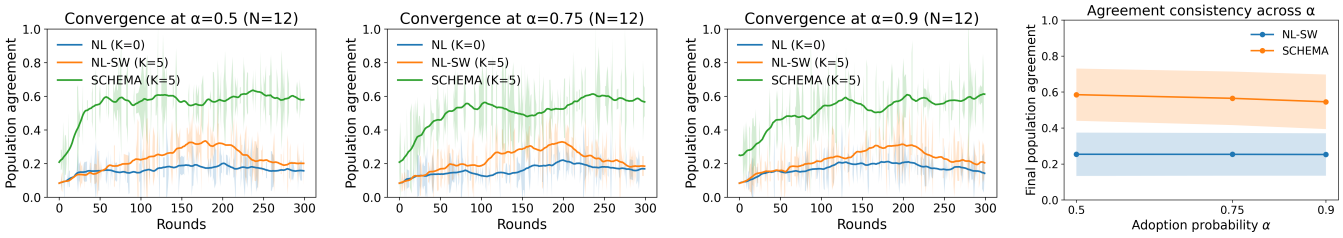
The naming game is defined over a population of N agents and a fixed lexicon $\mathcal{L} = C_1, \ldots, C_M$. Time advances in rounds $t = 1, \ldots, T$, with agents paired uniformly at random. Agent i produces a message m_i^t, which a decoder maps to a name $y_i^t \in \mathcal{L}$. Each agent keeps a memory window of size K containing its last K partner-only interactions, which guides future proposals. We study three conditions: Natural Language (NL), Natural Language Sliding Window (NL-SW), and Schema.

In NL, agents generate unconstrained natural language outputs, and the decoder extracts a valid token if possible. NL-SW extends this by incorporating memory K, so recent interactions influence each proposal.

Algorithm 1: Schema-Induced Condition

Input: N, $\mathcal{L}$, K, T, α
for $t = 1$ **to** T **do**
  Pair agents i, j uniformly at random
  Each forms proposal m^t using partner-only K
  Parse `@say {name: Ck}` $\rightarrow y$
  if *non-compliant* **then**
    Retry once with reminder
    if *still invalid* **then**
      decode free text; **if** *undecodable* **then**
        $y \leftarrow$ None
      end
    end
  end
  if $y_i \neq y_j$ **then**
    adopt partner's C_k w.p. α
  end
end

The Schema condition is outlined in Algorithm 1 and requires replies to match `@say {name: Ck}`, from which a regex parser extracts the C_k token. The lightweight schema tag follows earlier work showing that structured templates improve reasoning and reliability in LLM tasks (Chen et al. 2024). This design provides agents an explicit, easily parsed handle for lexicon entries, keeping replies transparent to the listener with minimal overhead. Non-compliant outputs re-



$\alpha = 0.5$ $\alpha = 0.75$ $\alpha = 0.9$ Final agreement

Figure 1: Population agreement under different adoption probabilities α with $N = 12$ and $K = 5$.

		Population Agreement		
N	K	NL	NL-SW	SCHEMA
12	0	0.111 ± 0.048	—	—
24	0	0.125 ± 0.042	—	—
12	5	—	0.278 ± 0.127	**0.611** ± 0.293
24	5	—	0.292 ± 0.042	**0.556** ± 0.064
12	10	—	0.333 ± 0.144	**0.639** ± 0.096
24	10	—	0.295 ± 0.039	**0.588** ± 0.085

Table 1: Average population agreement by condition with varying agents (N) and memory (K).

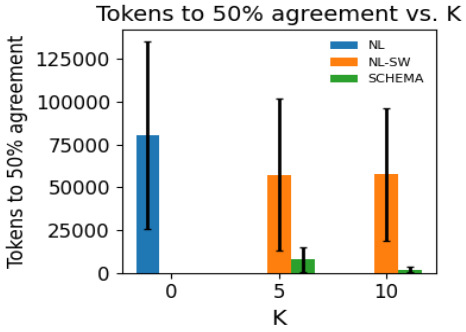


Figure 2: Tokens required to reach 50% population agreement across conditions.

ceive one retry with a brief reminder. If still invalid, the output defaults to a random valid C_k and is marked non-compliant. The variable α denotes the lose shift probability, meaning that after a mismatch an agent adopts the partner's decoded name with probability α.

Experiments

We evaluate the naming game with populations of $N \in \{12, 24\}$ agents, lexicon size M=12, and T=300 rounds, defining tokens-to-convergence as the number of tokens needed for the population to reach a chosen agreement. Memory windows vary with $K \in \{5, 10\}$ and lose-shift with $\alpha \in \{0.5, 0.75, 0.99\}$. Each configuration is run with three random seeds. Agents use the Phi-3 Mini 4K Instruct model with fixed decoding: max new tokens = 32, temperature = 0.7, top-p = 0.9, and repeat penalty = 1.1.

As shown in Figure 1, the Schema condition achieves substantially higher population agreement than both NL and NL-SW across adoption probabilities α. Agreement under Schema rises toward 0.6–0.65, while NL-SW peaks near 0.3 and NL remains below 0.2. Increasing α slightly lowers agreement for both NL-SW and Schema. Table 1 reports average agreement across populations and memory sizes. Agreement is stable for Schema, with decreases from $N = 12$ to $N = 24$ and modest increases from $K = 5$ to $K = 10$, indicating that gains come mainly from schema induction rather than population size or memory. Beyond higher agreement, standard deviation decreases over time, with Schema at $\alpha = 0.5$ producing the most consistent outcomes. Figure 2 shows that Schema converges with an order of magnitude fewer tokens than NL or NL-SW. These results suggest that schema-induced messages create a more stable signal for the population, helping agents coordinate more efficiently over repeated interactions and showing that even small constraints on message format can shift the overall dynamics of the naming game in a measurable way. Code is available here: https://github.com/ryanzhangofficial/schema-induced-games-for-naming.

Conclusion and Future Work

Inspired by work showing that conventions can emerge from interaction and by the use of structured formats in supervised tasks, we show that adding a fixed schema to LLM agents steers convention formation in a naming game, yielding up to 5.8× greater population agreement. Minimal structural priors thus can shape how conventions emerge. In our setting, the schema acts as a lightweight inductive bias that reduces ambiguity in proposals, limits the drift produced by natural language variability, and provides each agent with a more stable input space for forming and updating its memory. These effects together suggest that structure can influence not only the final agreement but also the pathway by which agents converge.

A key direction is testing whether schema reduces variation in LLM responses and whether this consistency may limit broader tasks, together with studies of larger populations and varied lexicon sizes. Future work may also explore whether different forms of structure produce distinct convergence behaviors, for example whether softer or partially flexible schemas can maintain high agreement without restricting communication as strongly.

References

Ashery, A. F.; Aiello, L. M.; and Baronchelli, A. 2025. Emergent social conventions and collective bias in LLM populations. *Science Advances*, 11(20): eadu9368.

Baronchelli, A.; Loreto, V.; and Steels, L. 2008. In-depth analysis of the Naming Game dynamics: the homogeneous mixing case. *arXiv:0803.0398*.

Chen, W.; Yuan, C.; Yuan, J.; Su, Y.; Qian, C.; Yang, C.; Xie, R.; Liu, Z.; and Sun, M. 2024. Beyond natural language: Llms leveraging alternative formats for enhanced reasoning and communication. *arXiv preprint arXiv:2402.18439*.

Guo, T.; Chen, X.; Wang, Y.; Chang, R.; Pei, S.; Chawla, N. V.; Wiest, O.; and Zhang, X. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*.