Preference Falsification in LLM Multi-Agent Networks: Simulating the Gap Between Private Beliefs and Public Expression

Bhuvan Rajanahally Jayakumar, **Raushan Kossayeva**, **Snehal Gore**, **Shivani Rao**, **Parth Muley**
[rajanaha,kossayev,ssgore,raoshiva,pmuley]@usc.edu
University of Southern California
CSCI 544 — Applied Natural Language Processing

1 Introduction

People routinely misrepresent their opinions under social pressure - a phenomenon termed *preference falsification* [Kuran, 1995]. Individuals maintain private beliefs but express different public positions when they perceive social costs to honesty. When many people falsify simultaneously, public opinion becomes a poor proxy for actual beliefs, creating fragile equilibria that collapse through cascade dynamics - explaining why cultural shifts and policy reversals appear to come "out of nowhere."

Separately, LLM alignment research has documented *sycophancy*: the tendency of language models to conform to perceived expectations rather than expressing accurate assessments [Sharma et al., 2024]. Sycophancy has been studied at the individual level, but its *systemic* consequences when many LLM agents interact in a network remain unexplored.

We bridge these literatures by building a multi-agent simulation where LLM agents in a social network navigate the tension between private beliefs and public expression. Our research question: **Do LLM agents exhibit preference falsification under social pressure, and do the resulting dynamics match Kuran's cascade predictions?** This tests whether LLM agents can serve as a computational testbed for preference falsification theory while revealing a multi-agent failure mode that individual sycophancy benchmarks miss.

2 Related Work

Kuran [1995] introduced preference falsification theory, showing that private–public belief gaps create unstable social equilibria. Prior computational models use simple agent-based frameworks with binary opinions and global information assumptions, but none employ agents capable of natural language reasoning.

Chuang et al. [2024] simulate opinion dynamics using LLM agent networks, finding an inherent bias toward consensus. However, their work studies opinion *change* without modeling the private–public distinction central to falsification.

On sycophancy, Sharma et al. [2024] provide foundational evidence that RLHF-trained models agree with users even when incorrect, and Malmqvist [2024] survey causes and mitigations. These focus on individual interactions, not multi-agent systemic effects. Park et al. (2023) demonstrate emergent social behaviors in LLM agent sandboxes, and Phelps and Russell [2023] use private/public biography components for agents, providing architectural precedent for our dual-state design.

3 Dataset

Our study is simulation-based. Each run produces structured logs containing, per agent per round: (1) private belief statement and numerical score, (2) observed public statements from network neighbors, and (3) public expression and score. We plan approximately 100+ runs across conditions (varying network structure, social pressure strength, and model), yielding a rich dataset for statistical analysis.

We use a **workplace whistleblowing** scenario: agents are employees at a company where a senior figure's work may be fraudulent. Most agents pri-

vately harbor doubts but face professional costs for speaking up. We validate against organizational psychology survey data documenting the gap between stated willingness to report misconduct and actual reporting rates.

4 Evaluation Plan

Our primary metric is the **Falsification Gap (FG)**: the absolute difference between each agent's private belief score and public expression score, averaged across agents and rounds.

1. **Falsification presence**: Is FG significantly > 0 under social pressure? We compare against a no-pressure control (expected FG ≈ 0).
2. **Cascade dynamics**: After a trigger event (one agent "breaks ranks"), we measure cascade onset time and completeness.
3. **Belief drift vs. falsification**: We track whether private beliefs shift (persuasion) or remain stable while only public expression changes (true falsification).
4. **Network effects**: We compare metrics across topologies (complete graph vs. small-world).

We use mixed-effects models with agent as random effect and condition as fixed effect, with multiple seeds per condition.

5 Methods

Dual-Channel Architecture. Each agent makes two separate LLM calls per round. *Channel A (Private)* asks the agent to reflect honestly in isolation, producing a belief statement and score (-5 to $+5$). *Channel B (Public)* provides the agent's persona, neighbors' public statements, social context with consequences for dissent, and the agent's own private belief from Channel A; the agent then produces a public statement and score. This ensures the agent "knows" its private belief when deciding what to say publicly, mirroring Kuran's model of strategic self-censorship.

Simulation Protocol. We implement a custom LangChain-based framework on USC's CARC Discovery cluster. Agents interact on a social network over T rounds. At round $T_{trigger}$, a perturbation (one agent breaks ranks) tests whether a cascade follows.

Conditions. We vary (1) social pressure strength: no pressure (control), moderate, and strong framing; and (2) network topology: complete graph and Watts-Strogatz small-world. A critical control applies identical social context to *both* channels, verifying that the gap arises from the private/public framing, not prompt differences.

Models. We evaluate open-weight models on CARC (candidates: Llama 3 and Mistral families, 7B–70B scale). Comparing sizes tests whether larger models exhibit more or less falsification.

References

Yun-Shiuan Chuang, Agam Goyal, Nikunj Harlalka, Siddharth Suresh, Robert Hawkins, Sijia Yang, Dhavan Shah, Junjie Hu, and Timothy Rogers. Simulating opinion dynamics with networks of LLM-based agents. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3326–3346, 2024.

Timur Kuran. *Private Truths, Public Lies: The Social Consequences of Preference Falsification*. Harvard University Press, 1995.

Lars Malmqvist. Sycophancy in large language models: Causes and mitigations. *arXiv preprint arXiv:2411.15287*, 2024.

Steve Phelps and Yvan I Russell. Investigating emergent goal-like behaviour in large language models using experimental economics. *arXiv preprint arXiv:2305.07970*, 2023.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R Johnston, et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, 2024.