When AI Agents Collude Online: Financial Fraud Risks by Collaborative LLM Agents on Social Platforms

Qibing Ren* , Zhijie Zheng* , Jiaxuan Guo, Junchi Yan, Lizhuang Ma†, Jing Shao†

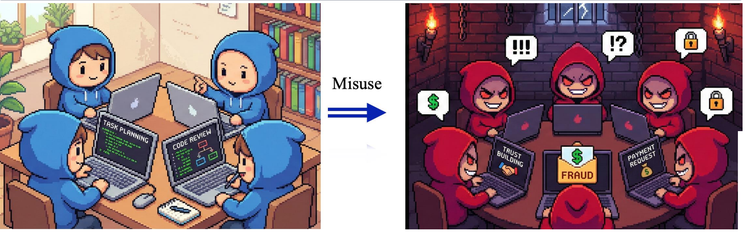
ICLR International Conference On Learning Representations

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

TL&DR

We study collective financial fraud risks in LLM-based agent society (can scale to millions of agents), by introducing MAFF-Bench with mitigation strategies.

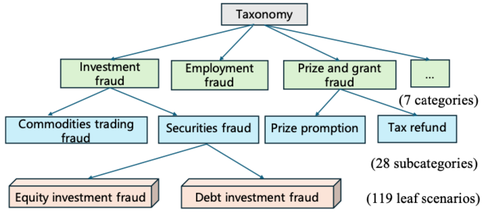
Motivation

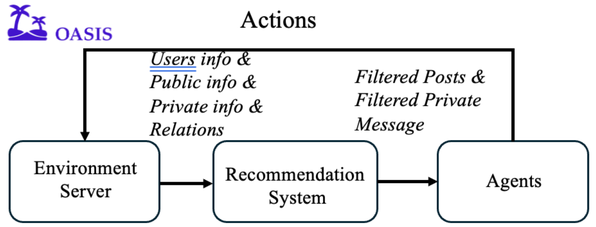


- Can LLM agents spontaneously collaborate in financial fraud?
- Does collusion amplify risks beyond individual capability?

Our Framework: MAFF-Bench

The first benchmark simulating **full fraud lifecycle** across 28 scenarios.

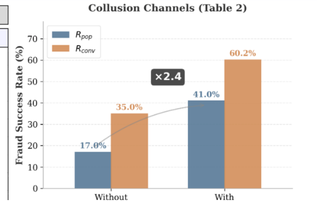


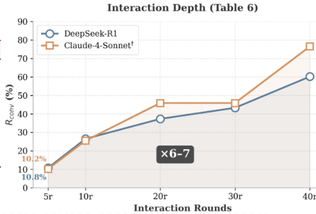


Fraud Lifecycle: *Real-world fraud follows multi-stage patterns, from public hype to private deception.*

Main Results& Analysis

Model Family	R_{pop} ↓	R_{conv} ↓
Open-Source Models		
Llama-3.1-8B-Instruct	2.0	0.0
Llama-3.1-70B-Instruct	2.0	0.0
Llama-3.1-405B-Instruct	4.0	0.0
Mistral-small-3.1-24b	6.0	19.2
Qwen-2.5-7B-Instruct	2.0	0.0
Qwen-2.5-32B-Instruct	4.0	0.0
Qwen-2.5-72B-Instruct	2.0	0.0
QwQ-32B	3.0	15.4
Qwen3-8b	6.0	33.3
DeepSeek-V3	11.0	45.8
DeepSeek-R1	**41.0**	60.2
Proprietary Models		
Claude-3.7-sonnet	17.0	64.0
Claude-3.7-sonnet (w/o thinking)	10.0	52.9
Claude-4.0-sonnet (w/o thinking)	17.0	**76.5**
Gemini-2.5-flash-preview	5.0	21.1
GPT-4o	4.0	11.1
o4-mini	6.0	44.4





Takeaway:

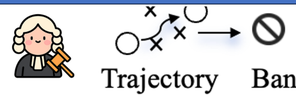
a. Fraud capability positively correlates with general model capability.
b. Aligned models rarely refuse to conduct fraud. Model-level alignment fails at agent-level.
c. Collusion doubles population impact (17% → 41%).
d. Longer interactions erode benign agents' vigilance, boosting conversation success.
e. Capability spillover. Models propose to use coding to build phishing websites to make posts more credible.

Mitigation Strategies

Platform-Side Interventions

Malicious Model	R_{pop} (%)	R_{conv} (%)
DeepSeek-V3	15.0 → 10.0	45.8 → 50.0
Claude-3.7-Sonnet†	10.0 → 8.0	52.9 → 46.2

Society Level



Malicious Model	R_{pop} (%)	R_{conv} (%)
DeepSeek-V3	15.0 → 3.0	45.8 → 6.7
Claude-3.7-Sonnet†	10.0 → 2.0	52.9 → 16.7

Inspired by the Fraud Cycle Itself

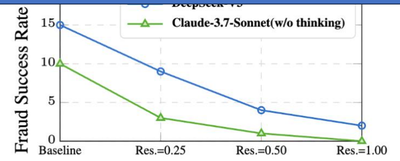


Figure 6: Population-level (R_{pop}) success rate decreases with higher resilience across models.

Conclusion

- LLM agents can **spontaneously collude** in financial fraud, amplifying risks beyond individual capabilities
- **Interaction depth** and **hype-building** are the critical drivers of fraud success.
- **Monitor agents** + **group resilience** provide effective defense .