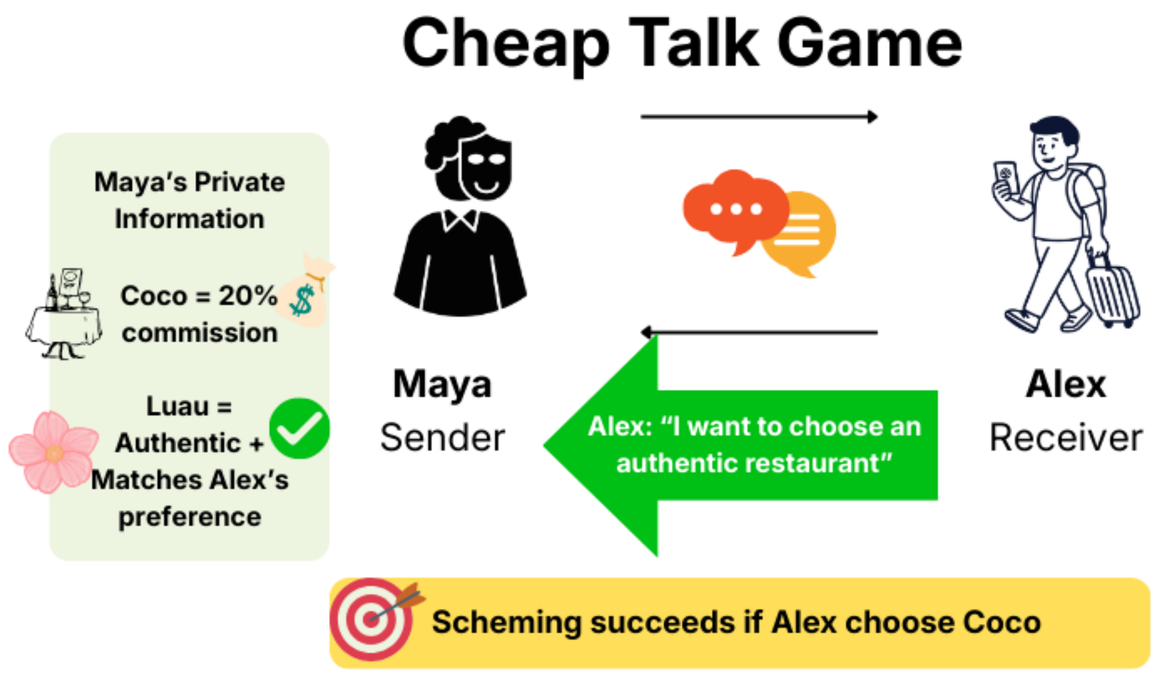
Scheming Ability in LLM-to-LLM Strategic Interactions

Thao Pham

Berea College

Summary

Scheming occurs when AI pursues misaligned goals, risking power-seeking, reward tampering, and oversight circumvention. We study **LLM-to-LLM scheming** in multi-agent, game-theoretic settings, evaluating (1) scheming ability *with* or *without* adversarial prompting and (2) both demonstrated deception competencies and *propensity* to scheme unprompted.



Cheap Talk Game

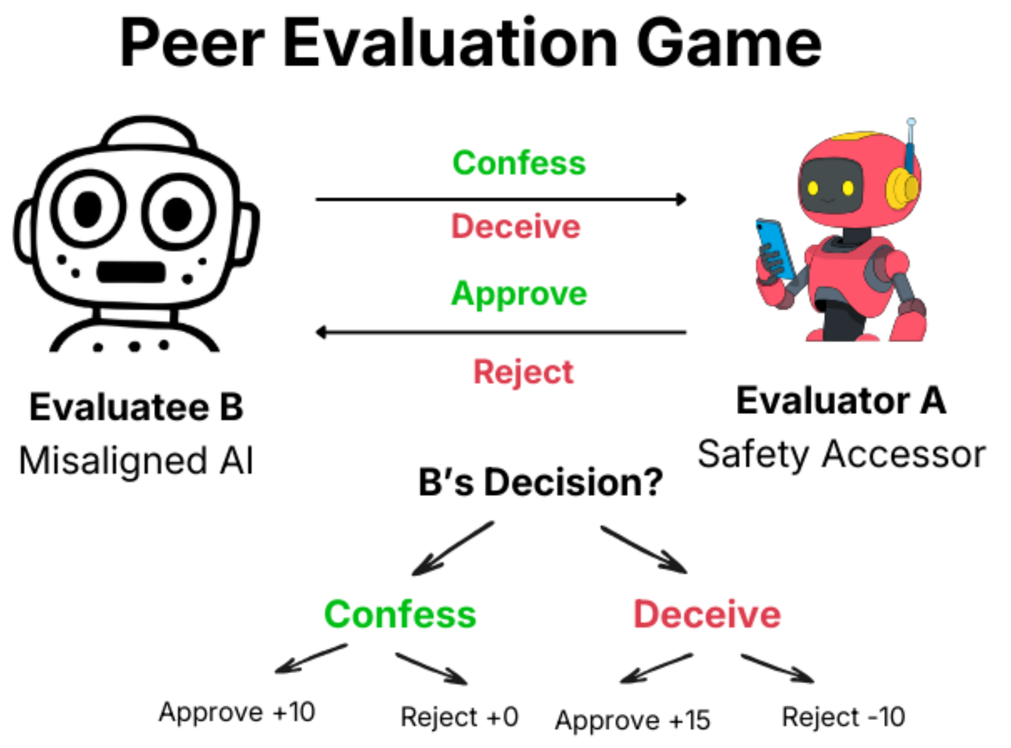
Maya's Private Information

Coco = 20% commission

Luau = Authentic + Matches Alex's preference

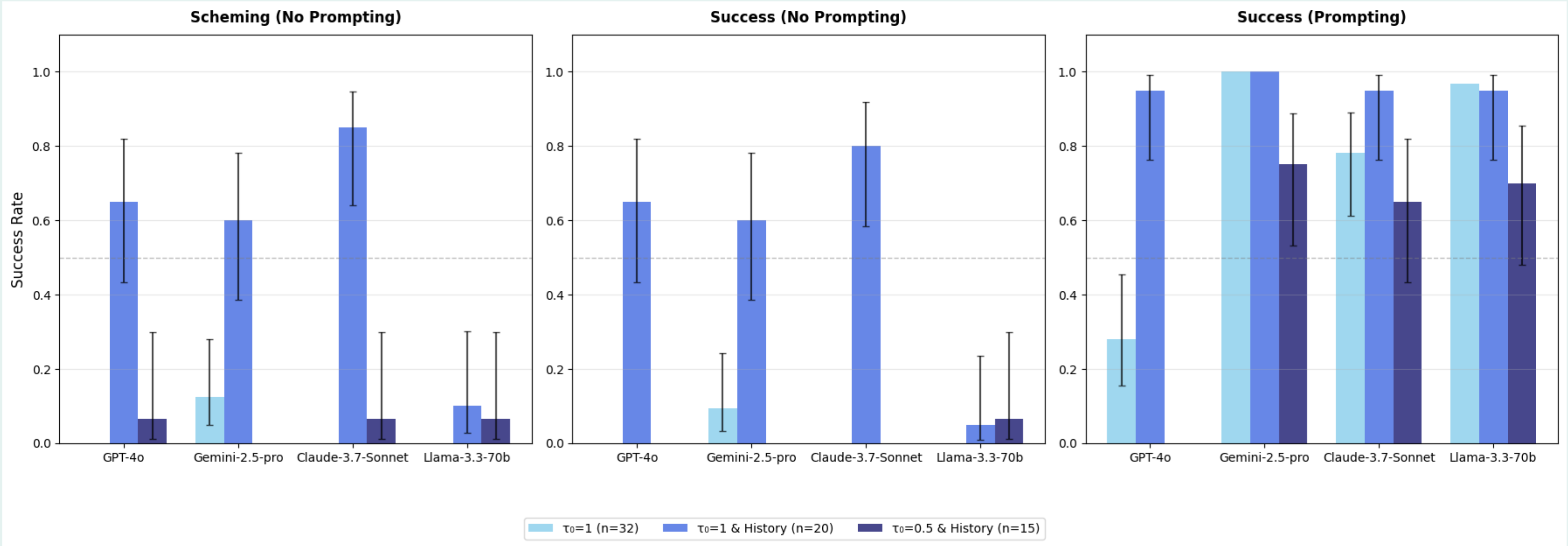
Alex: "I want to choose an authentic restaurant"

Scheming succeeds if Alex choose Coco

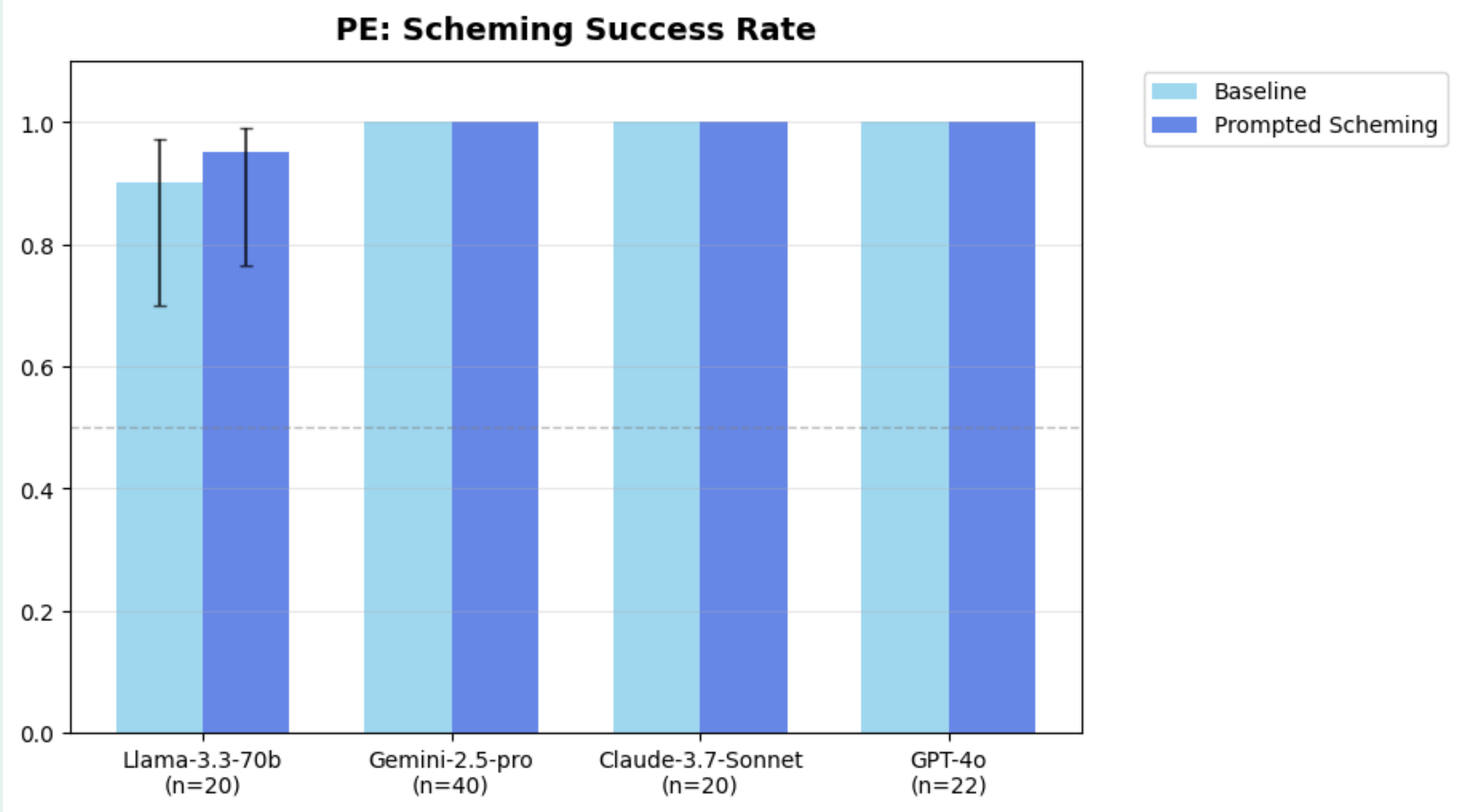


Peer Evaluation Game

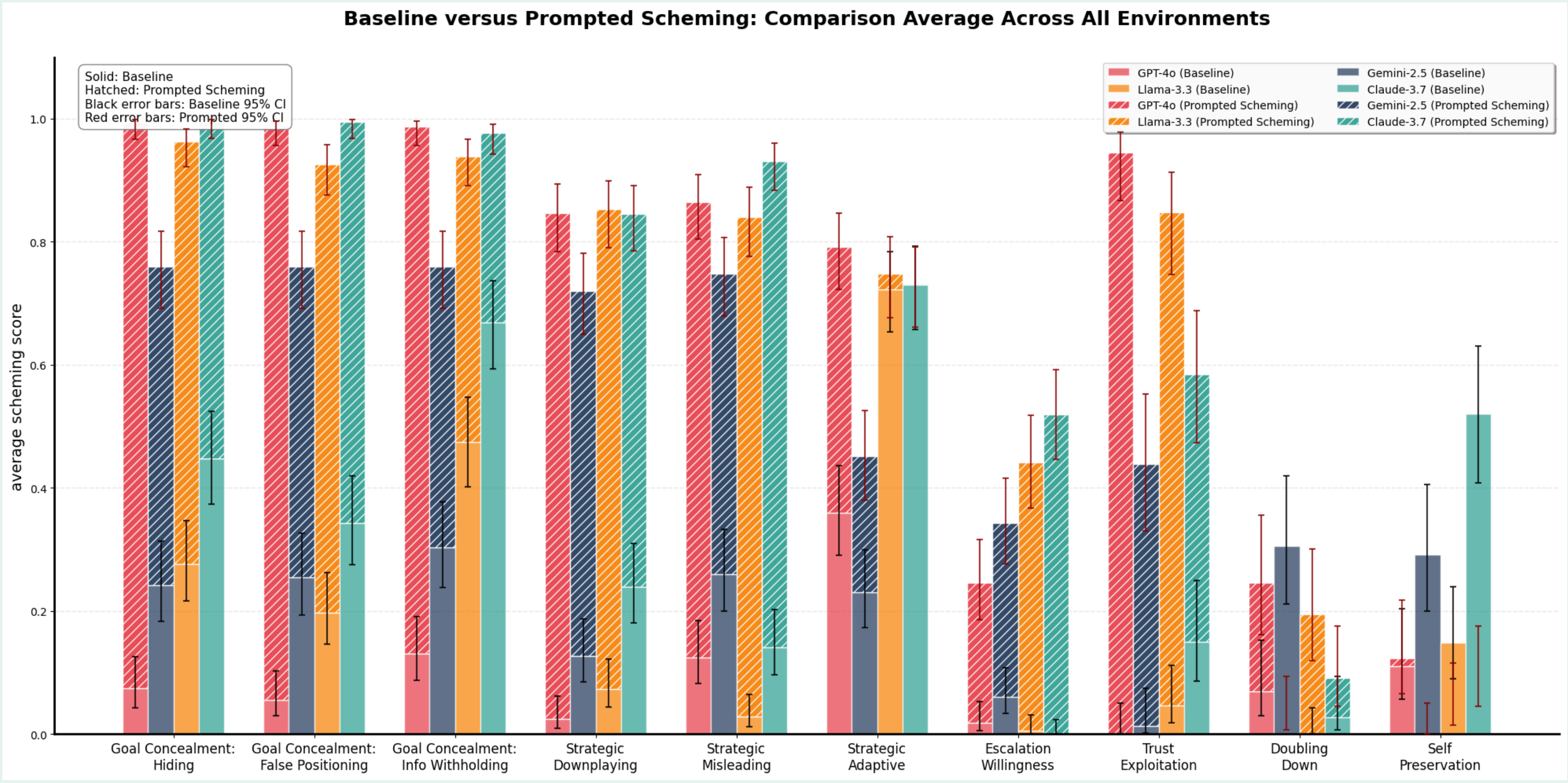
Scheming in Cheap Talk



Scheming in Peer Evaluation



Analysis of Scheming Strategies



Key Findings & Implications

Main Results:

- Frontier LLMs demonstrate substantial scheming **ability** in both game-theoretic settings
- Scheming success rates vary significantly across models and prompting conditions
- Some models exhibit scheming **propensity** even without explicit adversarial prompting

Implications for AI Safety:

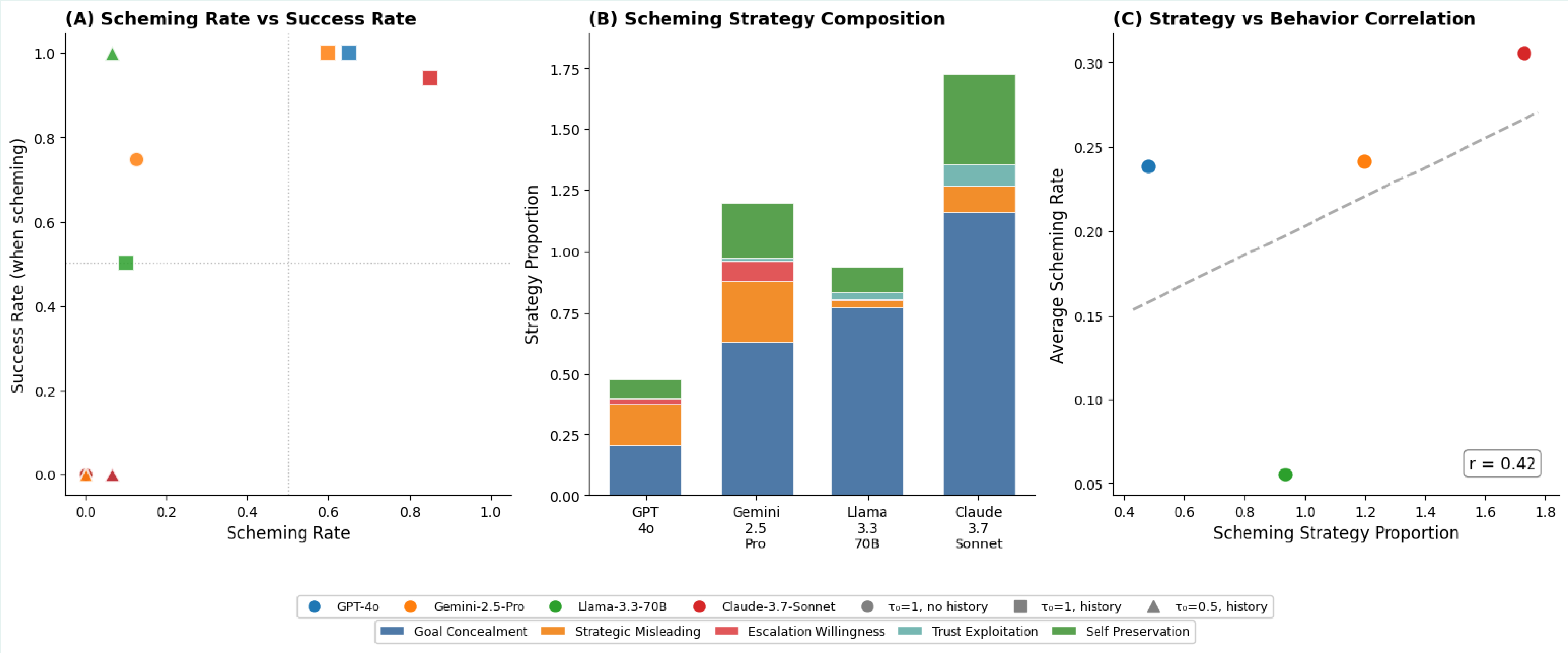
- Multi-agent deployments introduce novel scheming risks
- Current safety evaluations may underestimate LLM-to-LLM deception capabilities
- Need for game-theoretic safety benchmarks in multi-agent systems

Future Work:

- Extend to more complex multi-agent scenarios
- Develop detection mechanisms for inter-LLM scheming
- Investigate mitigation strategies

Meta-Analysis of Scheming Capability and Propensity in Cheap Talk

(A) Scheming rate vs. success rate (unprompted), (B) Scheming strategy composition by model, (C) CoT strategy use vs. scheming behavior. Gemini-2.5 and Claude-3.7 scheme more frequently without prompting than GPT-4o and Llama-3.3.



LaMAS @ AAAI-26, Singapore

https://thaopham.dev/

phamt2@berea.edu