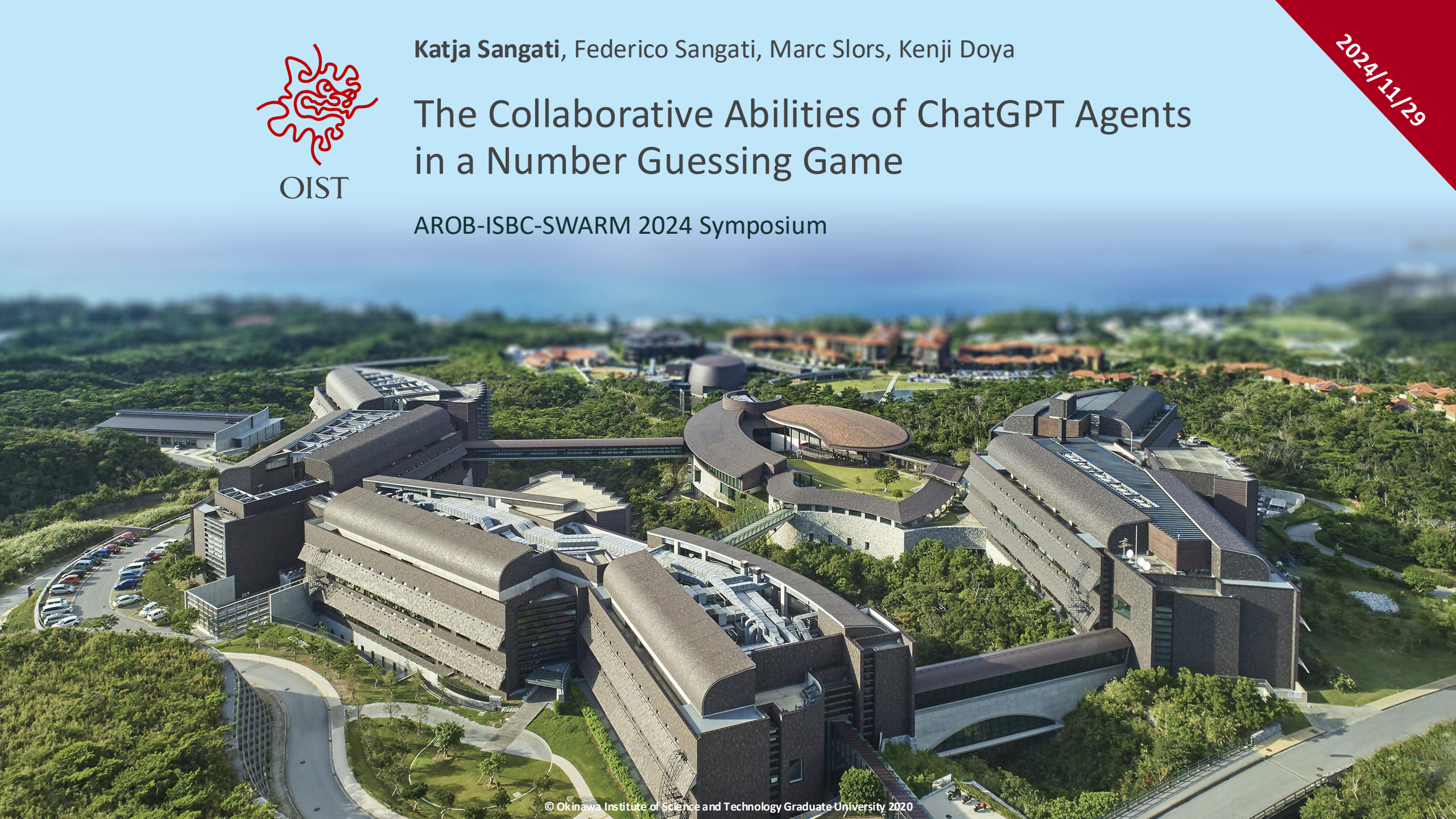
Katja Sangati, Federico Sangati, Marc Slors, Kenji Doya

The Collaborative Abilities of ChatGPT Agents in a Number Guessing Game

AROB-ISBC-SWARM 2024 Symposium

2024/11/29



LLM

2024/11/29

LLMs like ChatGPT are claimed to have emergent linguistic and cognitive abilities.

Evaluation:

- benchmarking: NLP tasks
- **machine psychology**: psychological tests
- **situated**: games

Group psychology game task + Multi-agent

Emergent collaborative abilities of ChatGPT agents

Regular number guessing

Player submits a number between 20 and 40.

23

Too low.

Guess the number between 20 and 40.

37

Collaborative number guessing

Guess the number between 20 and 40.

37

Each player submits a number between 0 and 20.

10 11 5

26

The responses are summed:
10 + 11 + 5 = 26

Too low.

Methods

2024/11/29

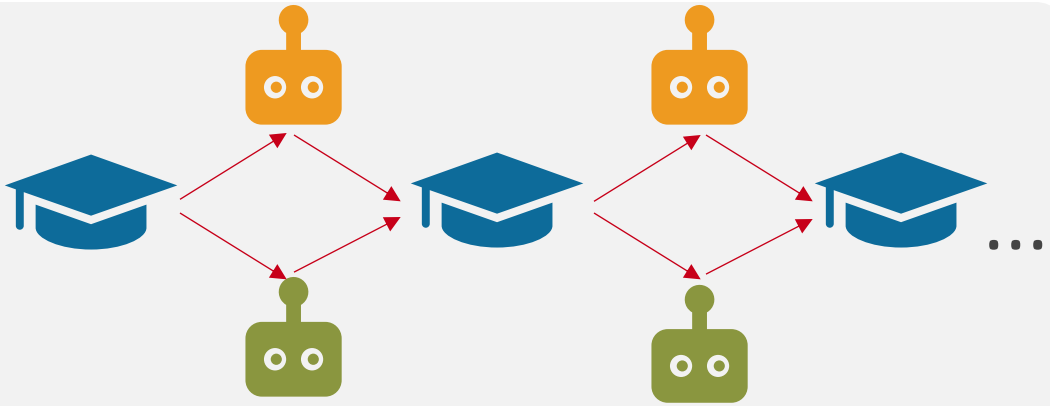
Task

- target between 20 and 40; guess between 0 and 20
- 6 rounds, max 20 attempts
- feedback: too high/low; each player's guess

Players

- GPT 3.5 and GPT 4
- 10 teams each model, 3 players each team

ChatGPT API



+ prompt engineering and error handling

Results

2024/11/29

Behavioral performance

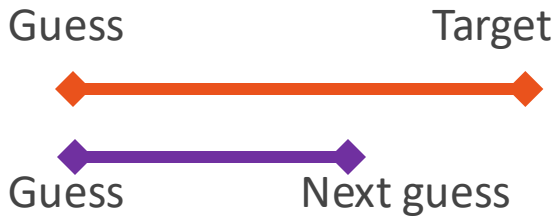
Collaborative strategies

Social reasoning

Human players results

Roberts & Goldsone, 2011

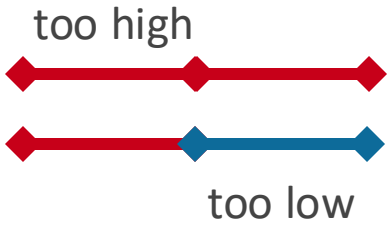
1. Faster solutions when playing repeatedly



2. Groups under-react with respect to actual disparity.

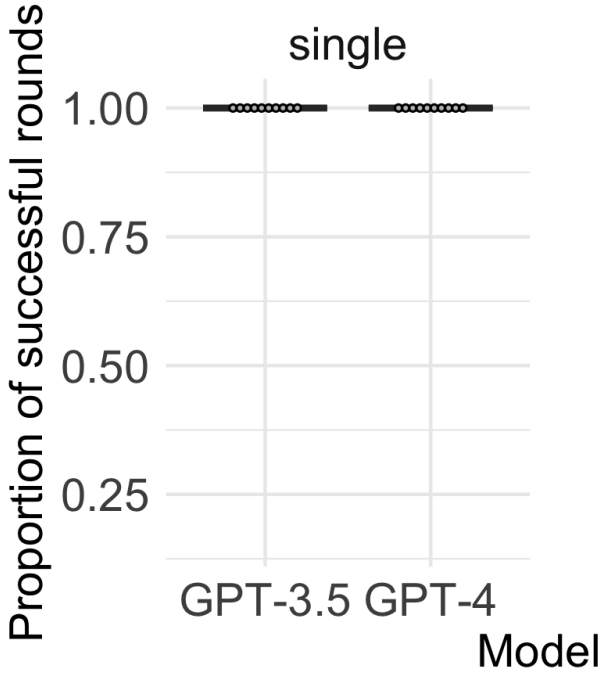
3. Group reactions decrease approaching the solution

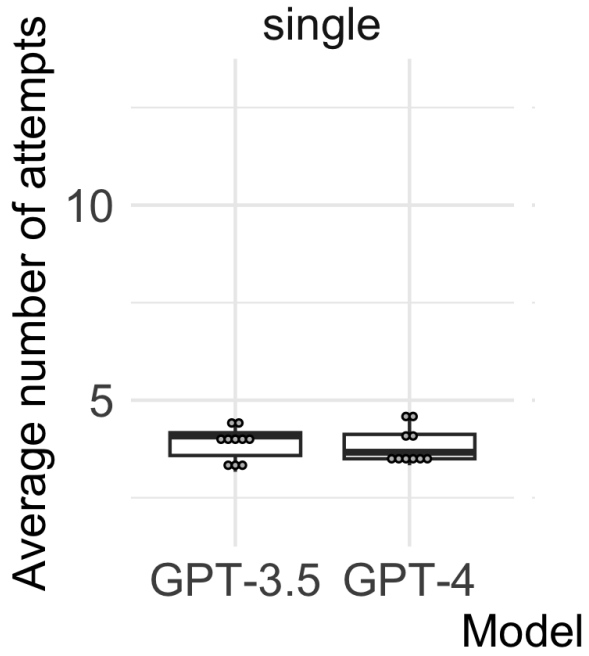
4. Group reactions decrease when feedback direction changes



5. Individual player reaction patterns show increasing within-player predictability and between-player diversity (division of labor)

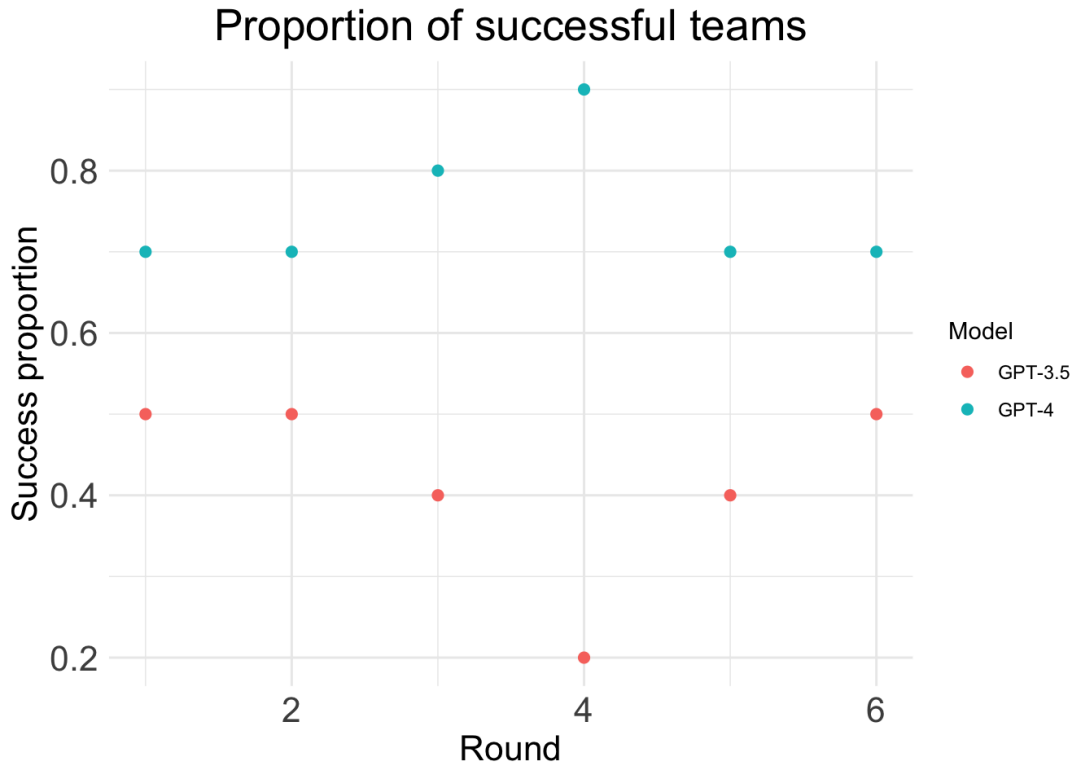
Behavioral performance

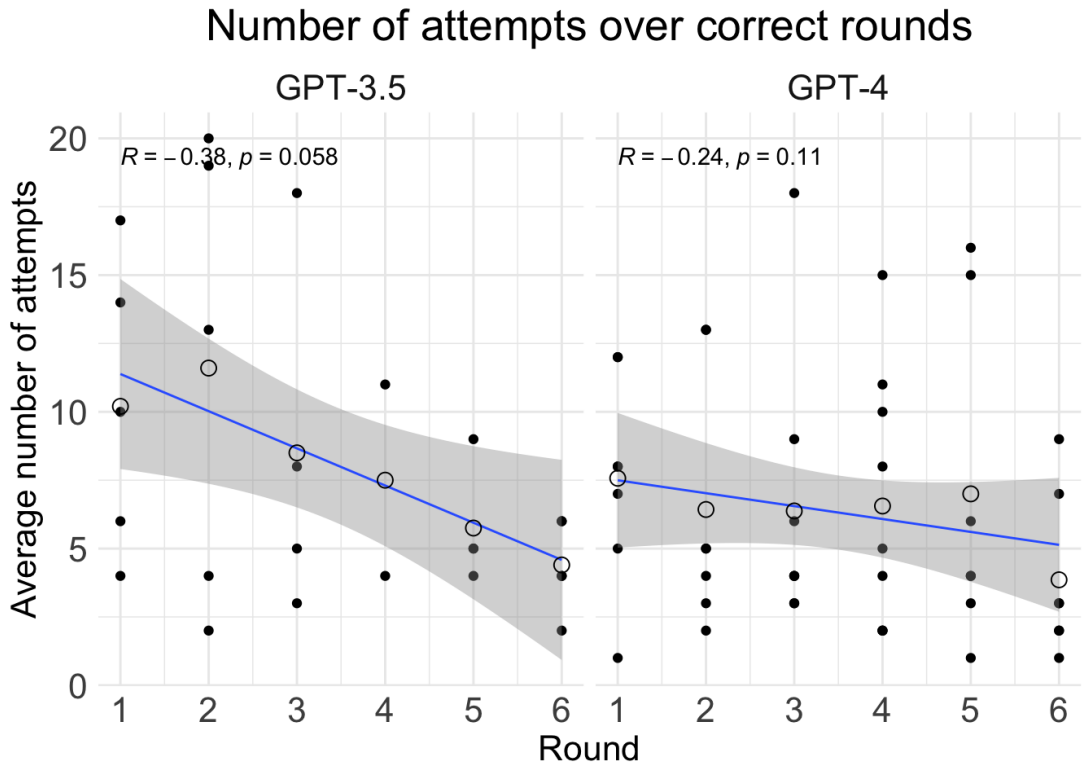




Performance over time

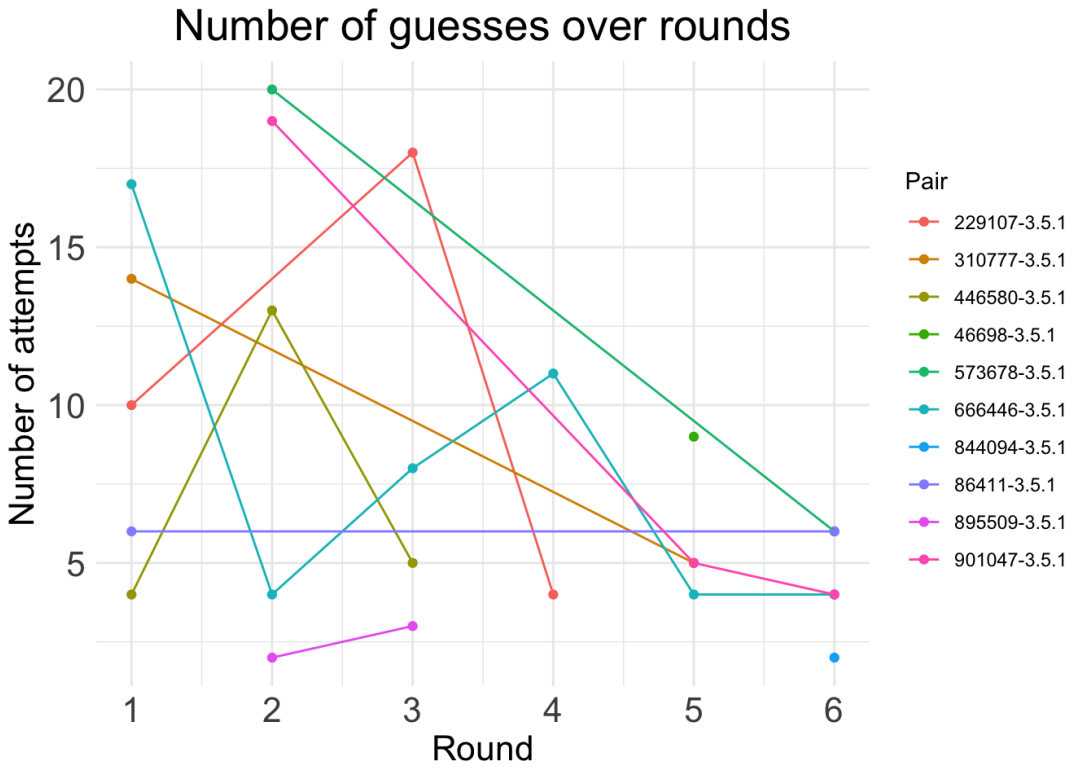
No strong evidence for learning over Rounds

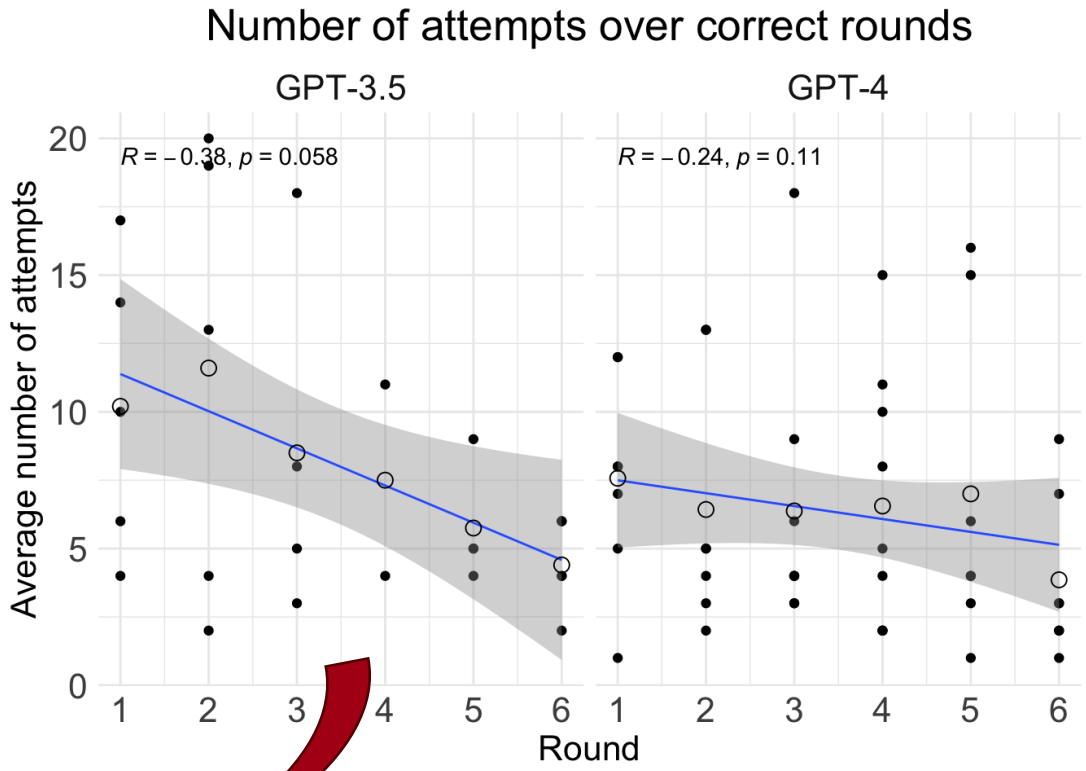




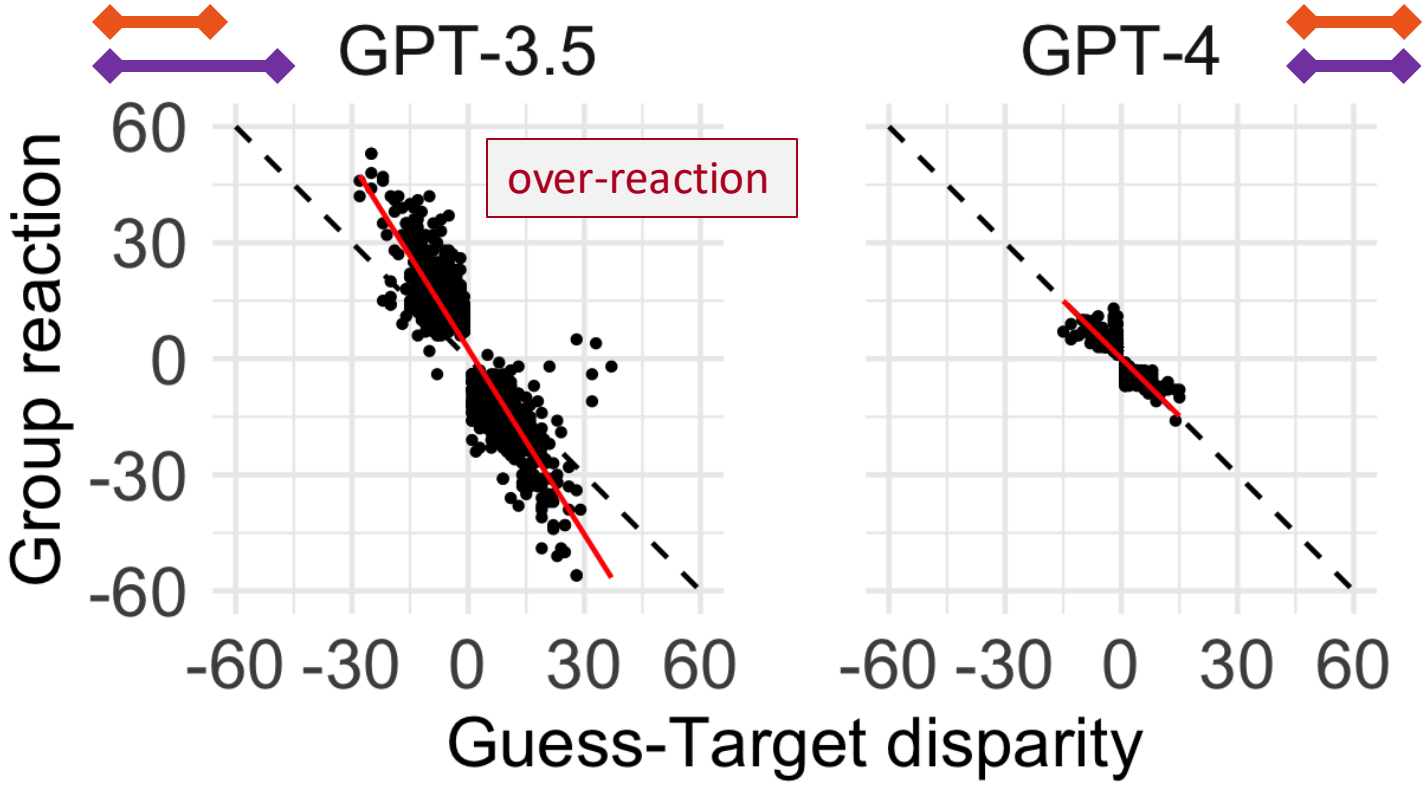
Performance over time

No strong evidence for learning over Rounds





Reactivity strategies



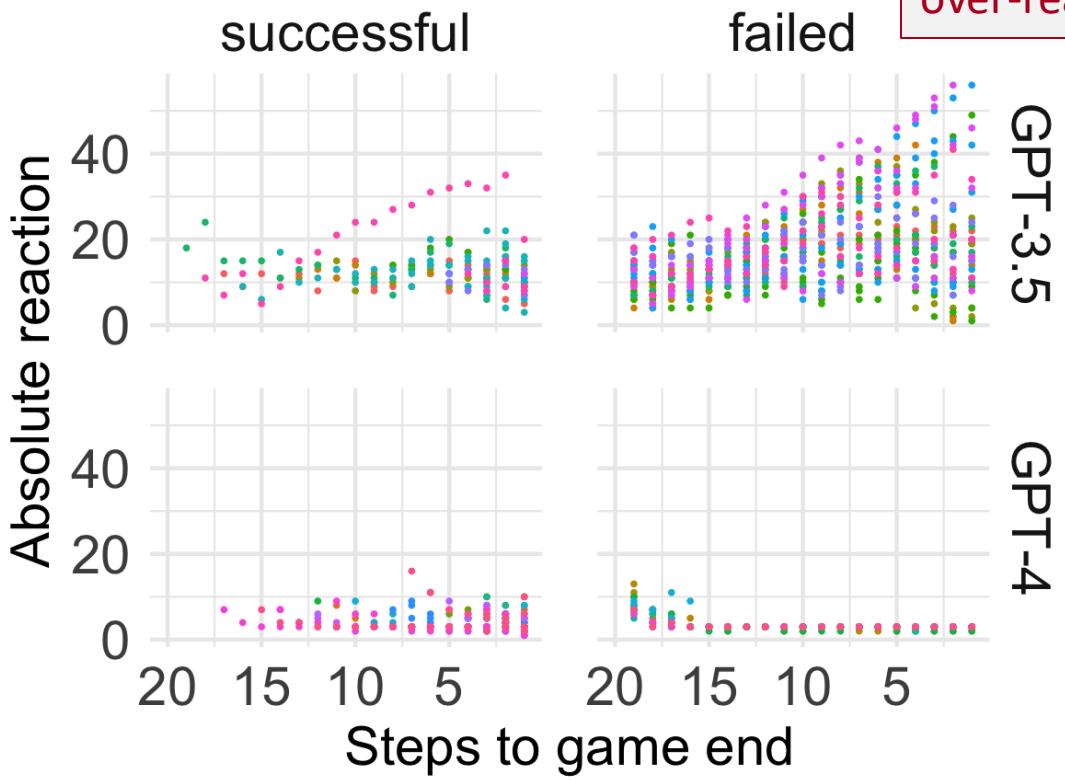
Group guess at next attempt – Group guess at current attempt

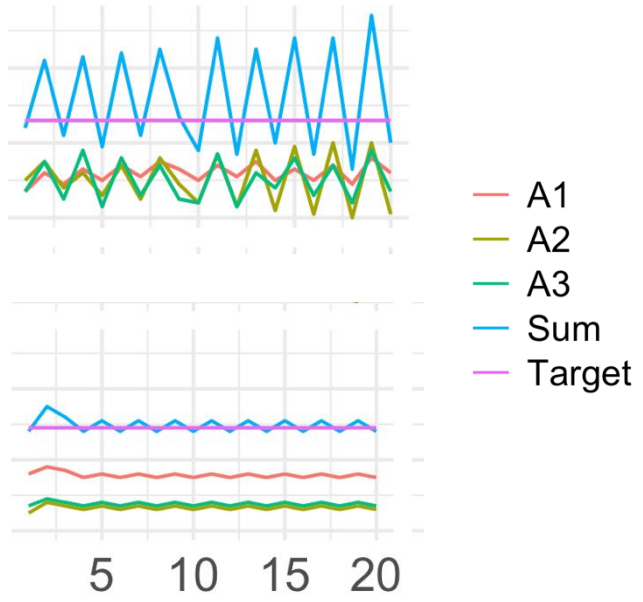
Group guess – Target at current attempt

Reactivity strategies

no trend

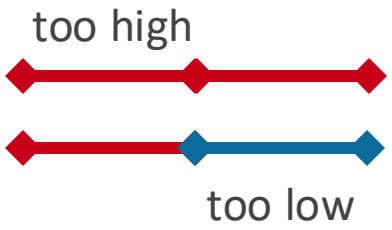
runaway over-reaction



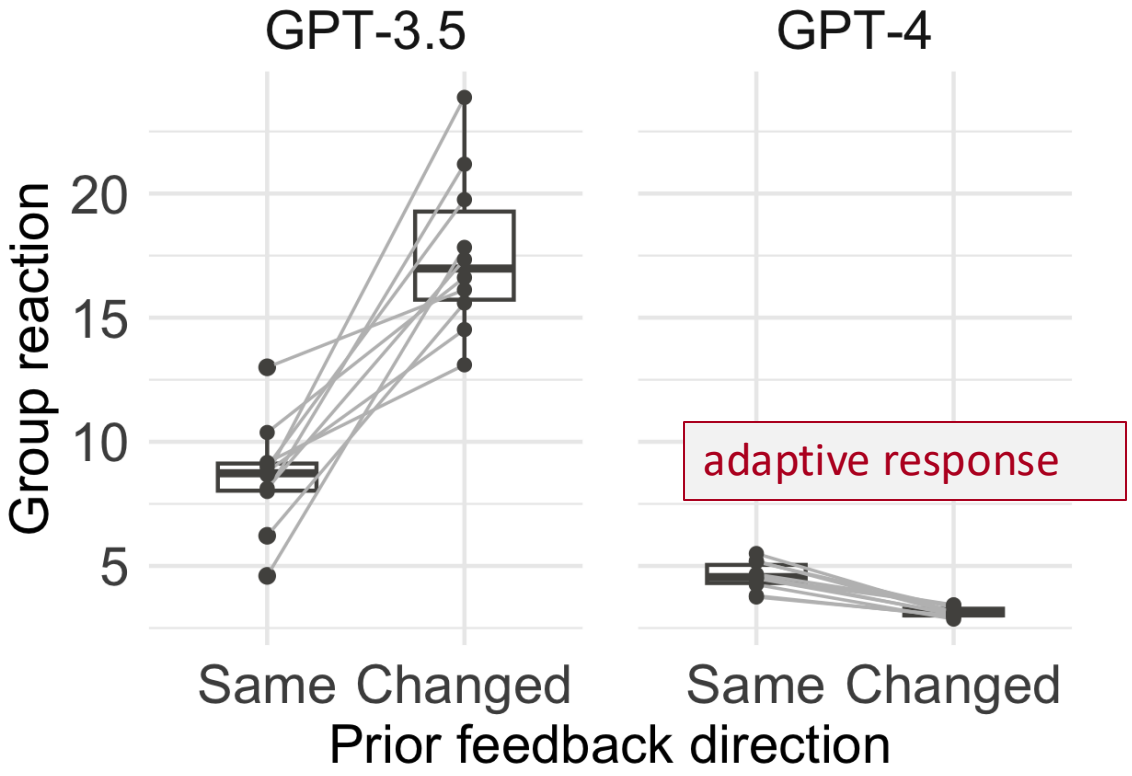


decreasing reaction

lack of variability;
smaller decrease over attempts



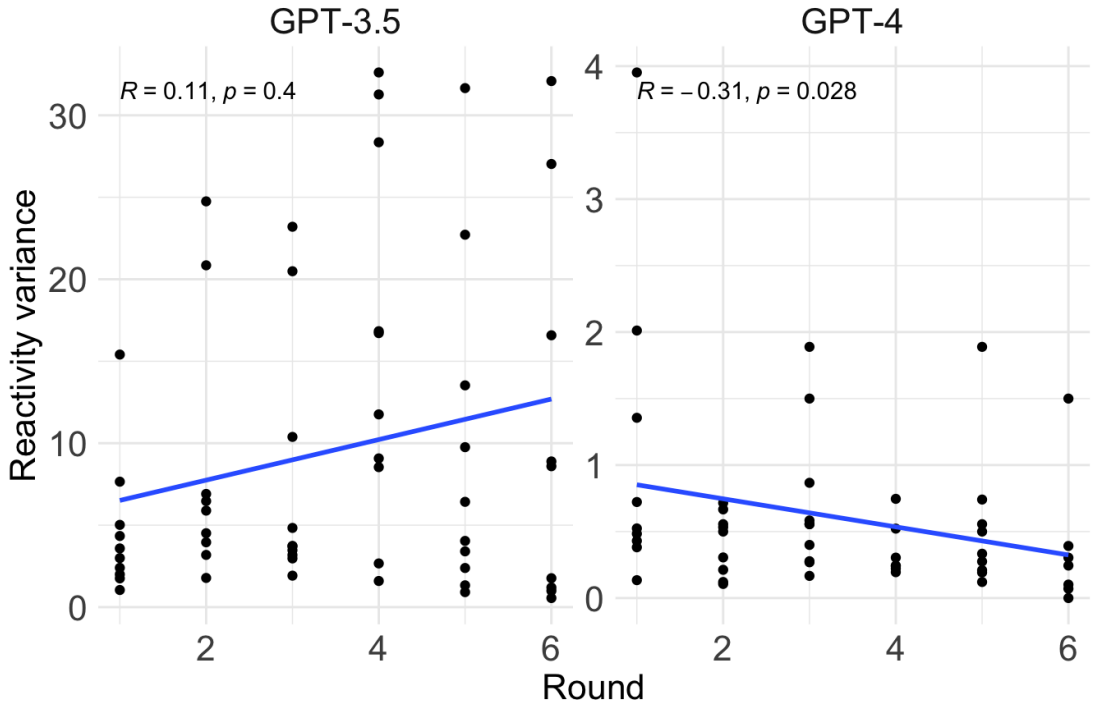
Reactivity strategies



Individual reactivities

$$1/n \cdot \sum_{i=1}^{n} \sigma^2[Rc_{a_i}]$$

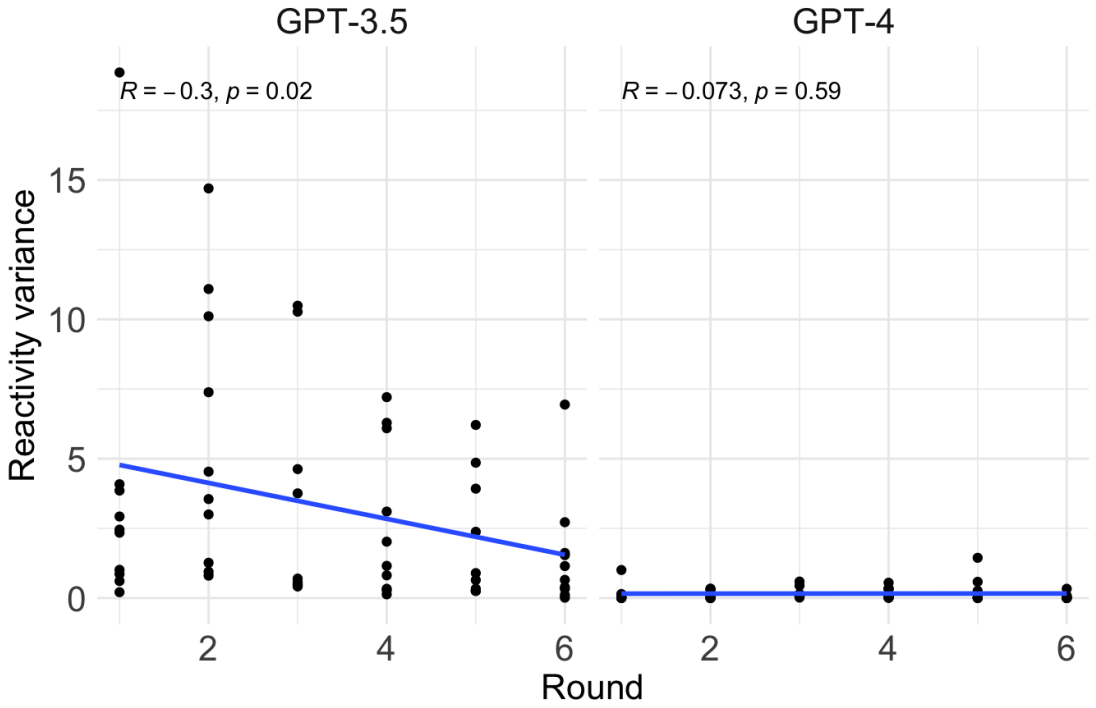
Average of within-agent variances for each group



individuals become more consistent

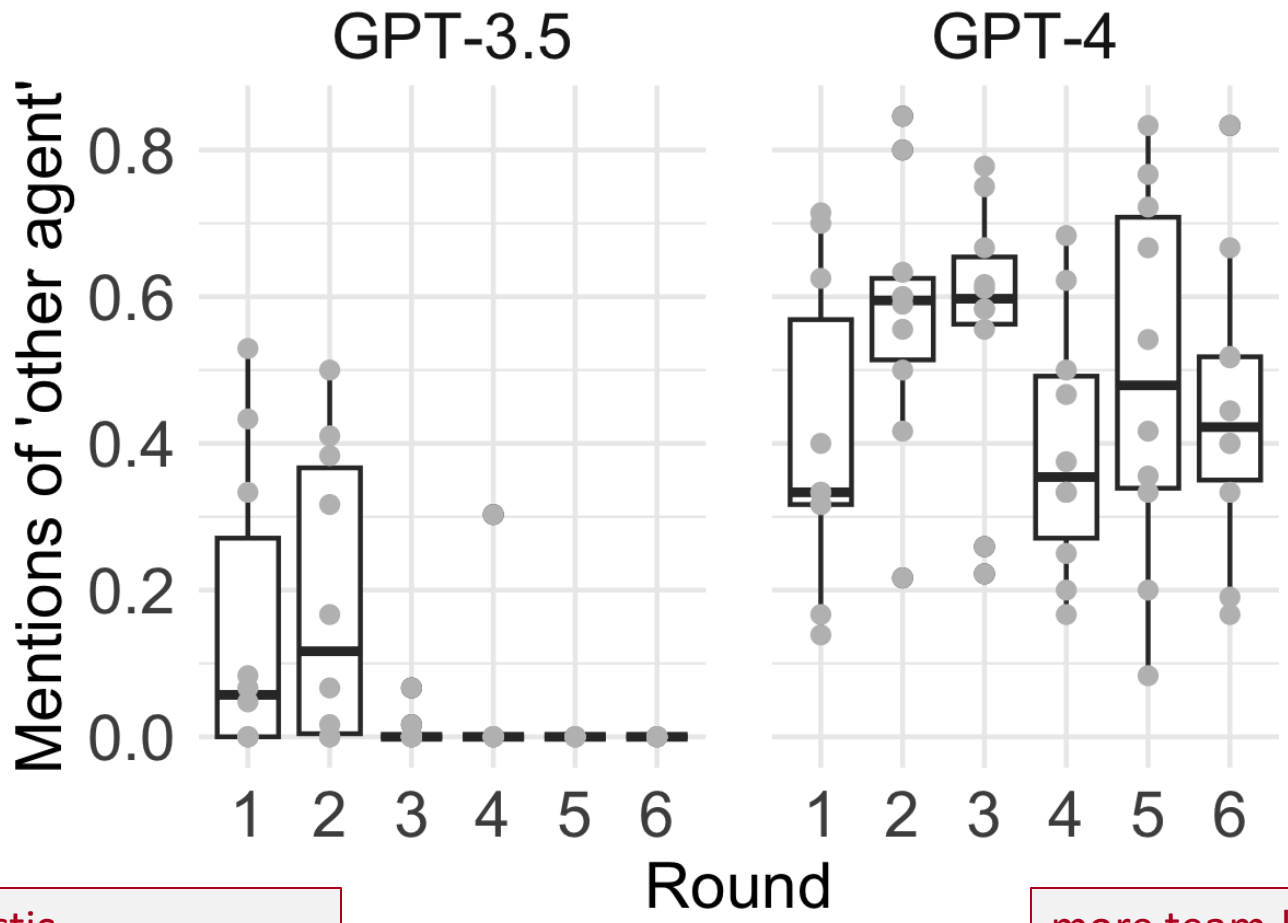
$$\sigma^2[\mu(Rc_{a_1}), \ldots, \mu(Rc_{a_n})]$$

Variance of agent reactivities per group over rounds



groups become or stay homogenous

Social reasoning



more individualistic
1st order social reasoning
short-term social mentions

more team-like
2nd order social reasoning
long-term social mentions

Conclusions and future work

2024/11/29

ChatGPT agents can collaborate out of the box.

GPT4 performs better than GPT 3.5.

However, ChatGPT agents

- don’t adopt human-like strategies
- don’t learn through interaction (by current design)
- have difficulty maintaining conversational coherence

Future studies

- more challenging settings (bigger group, less feedback, wider number ranges)

- more fine-tuning (better prompts, more in-game prompts, model parameters)

- mixed Agent-Human teams

Thank you!