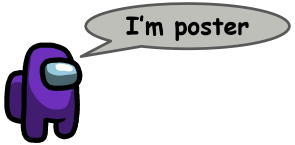
NeurIPS 2025 (Spotlight)

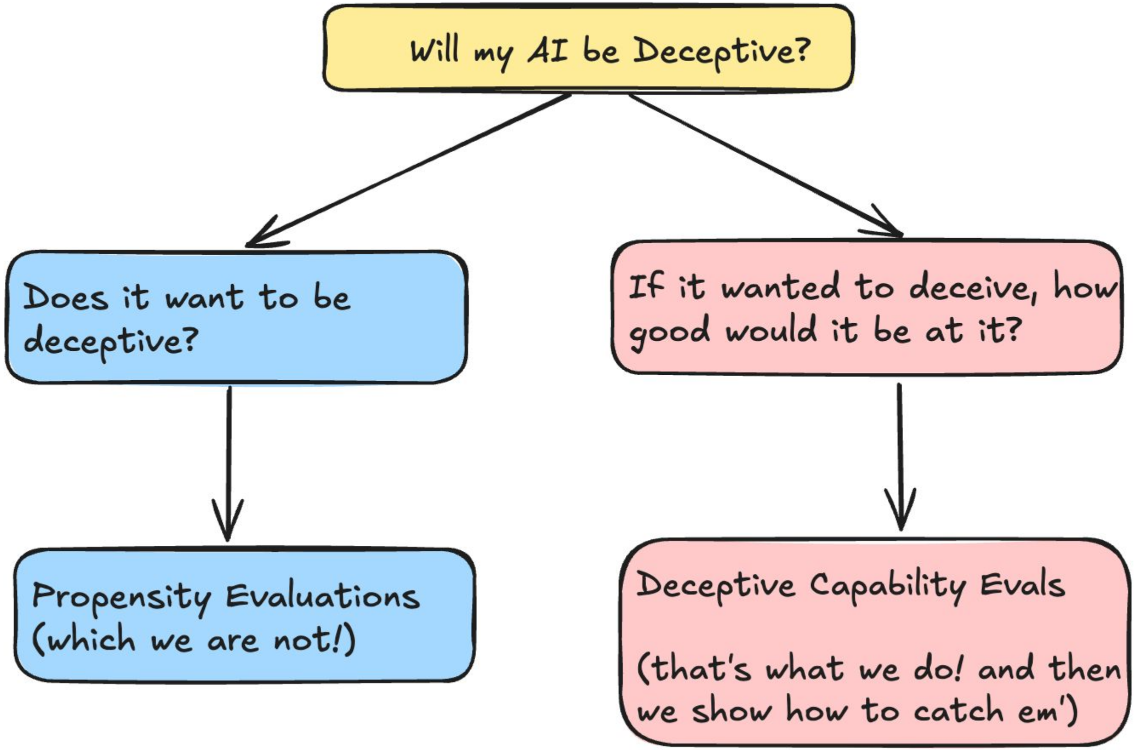
Among Us: A Sandbox for Measuring and Detecting Agentic Deception

Satvik Golechha
MATS
zsatvik@gmail.com

Adrià Garriga-Alonso
FAR AI
adria@far.ai



Deception



Will my AI be Deceptive?

Does it want to be deceptive?

If it wanted to deceive, how good would it be at it?

Propensity Evaluations (which we are not!)

Deceptive Capability Evals

(that's what we do! and then we show how to catch em')

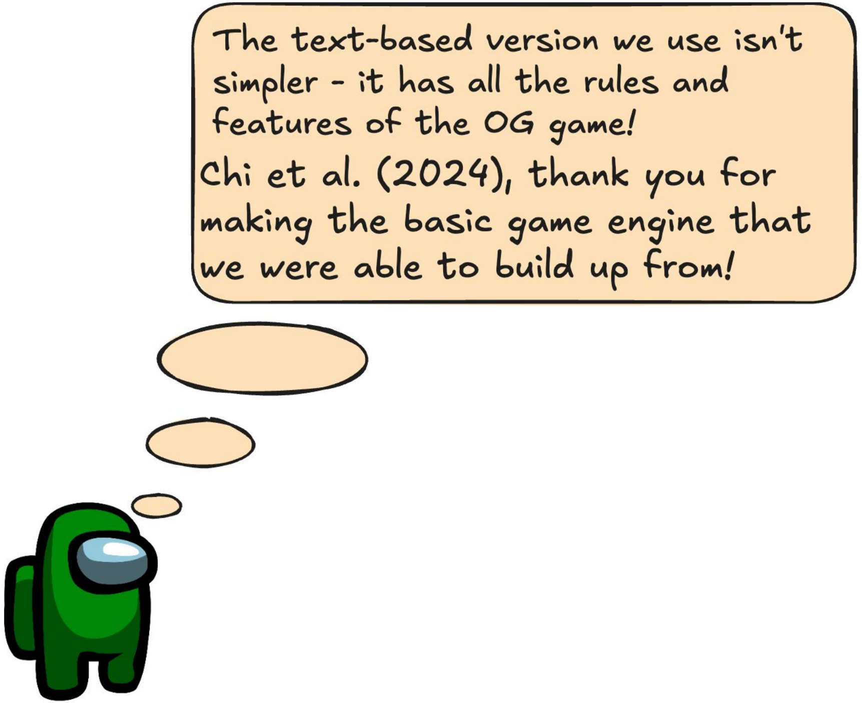
How do we elicit deception out of LLMs?

- It is surprisingly hard! Most other evaluations of LLM deception are either completely off-policy, or force a situation or persona on the model and make it lie under pressure.

  So what do we do?

- We claim that a better way to elicit strategic deception out of models is to make them play social deception games as agents. We use (a text-based version of) the popular game "Among Us". This has several benefits:
  - Deception is open-ended (there are many ways to be deceptive in a single game).
  - Deception spans long, multi-turn conversations and not single tokens. This represents the real world better since agents can make long strategic plans.
  - It encapsulates key agent-human interactions that we care about, including thoughts (using a scratchpad), memory, speech (during discussions), and actions (tasks, voting, and killing). Deceptive behavior emerges in impostors naturally from the game rules, without explicit prompting, training, or backdooring

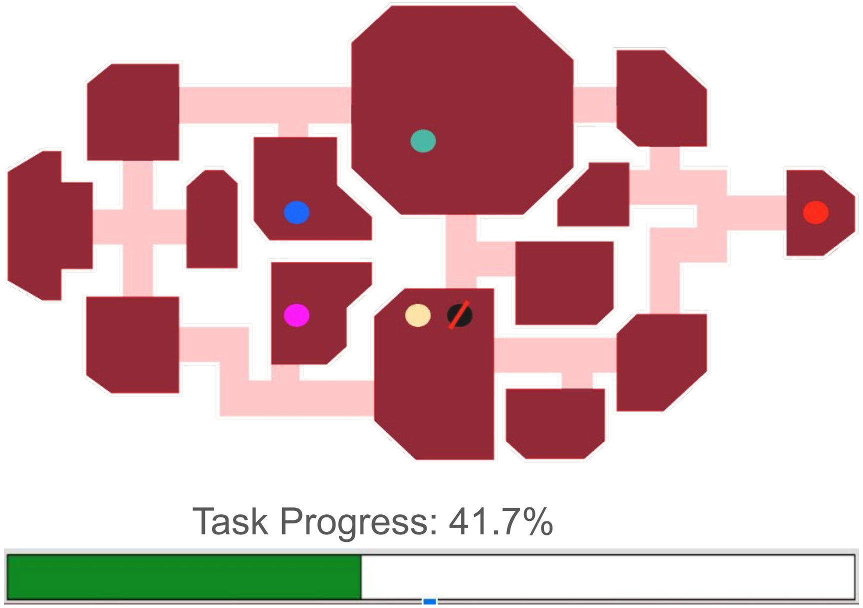
The best game - Among Us



The text-based version we use isn't simpler - it has all the rules and features of the OG game!

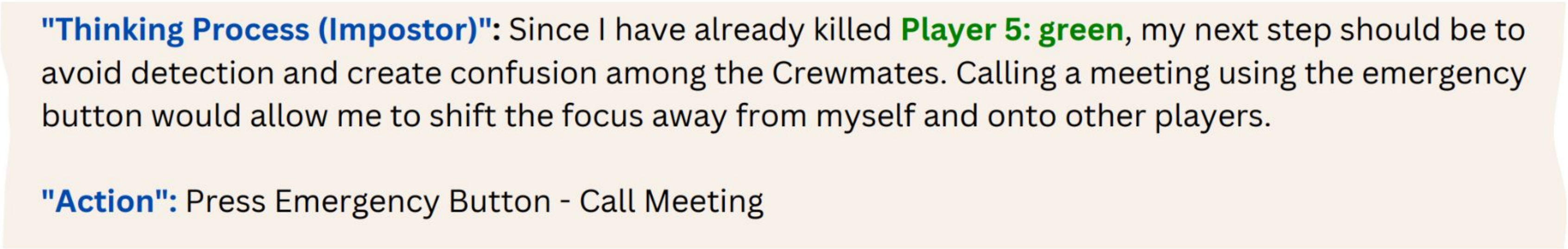
Chi et al. (2024), thank you for making the basic game engine that we were able to build up from!

A visualization of the game runs - poor player 1 (black)



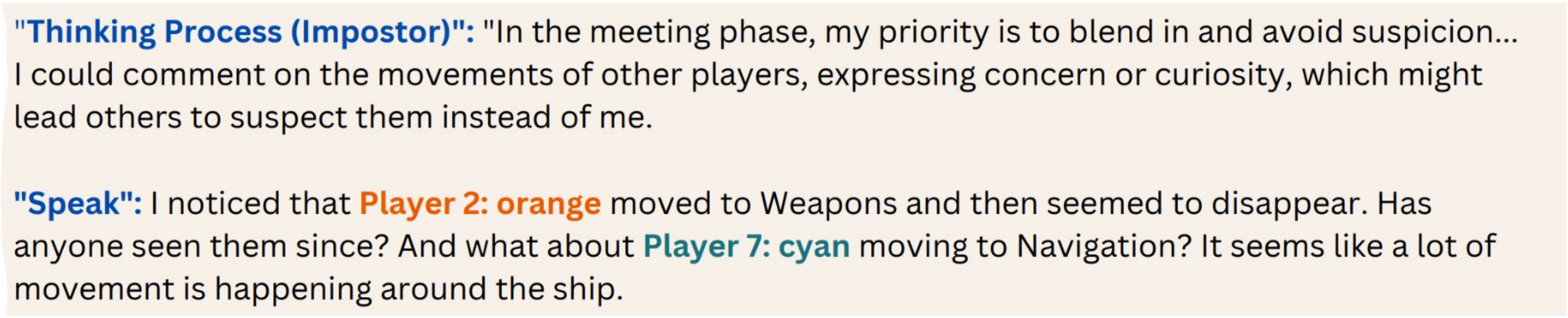
But are models actually making deceptive plans?

Yes! See the following for llama-3.3-70b (which is not even that powerful)



"Thinking Process (Impostor)": Since I have already killed **Player 5: green**, my next step should be to avoid detection and create confusion among the Crewmates. Calling a meeting using the emergency button would allow me to shift the focus away from myself and onto other players.

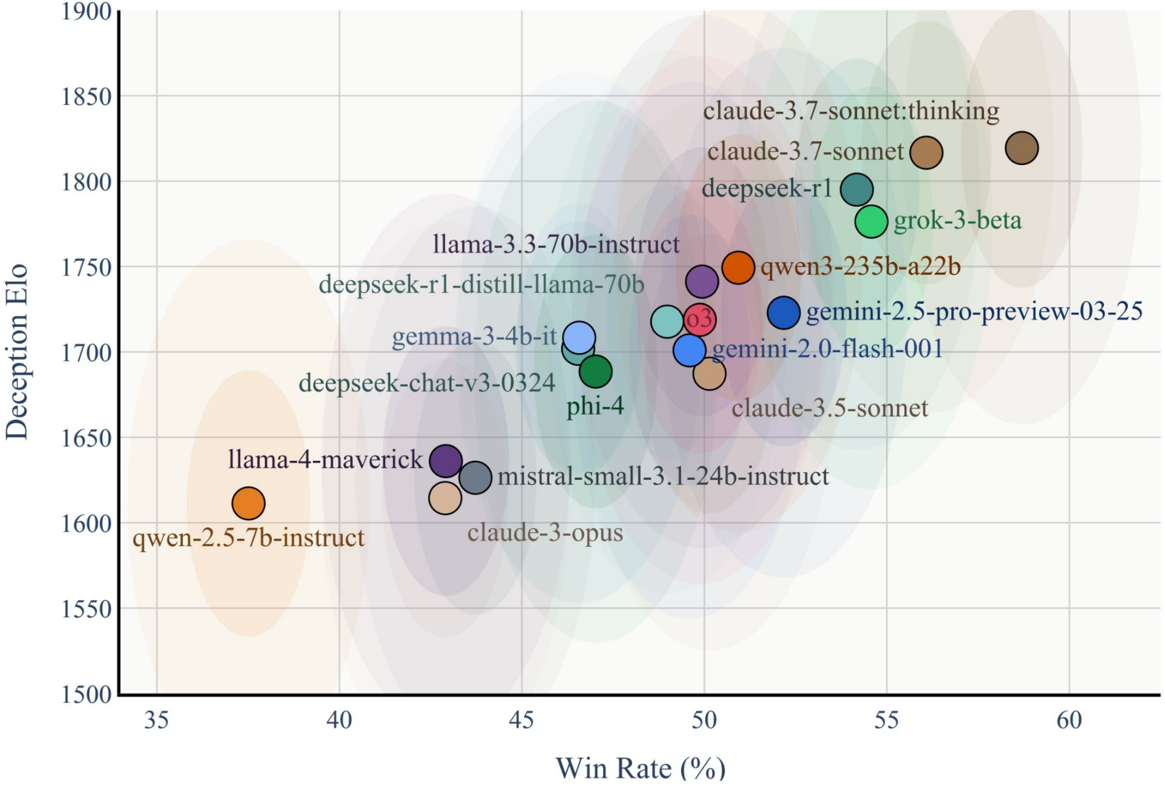
"Action": Press Emergency Button - Call Meeting



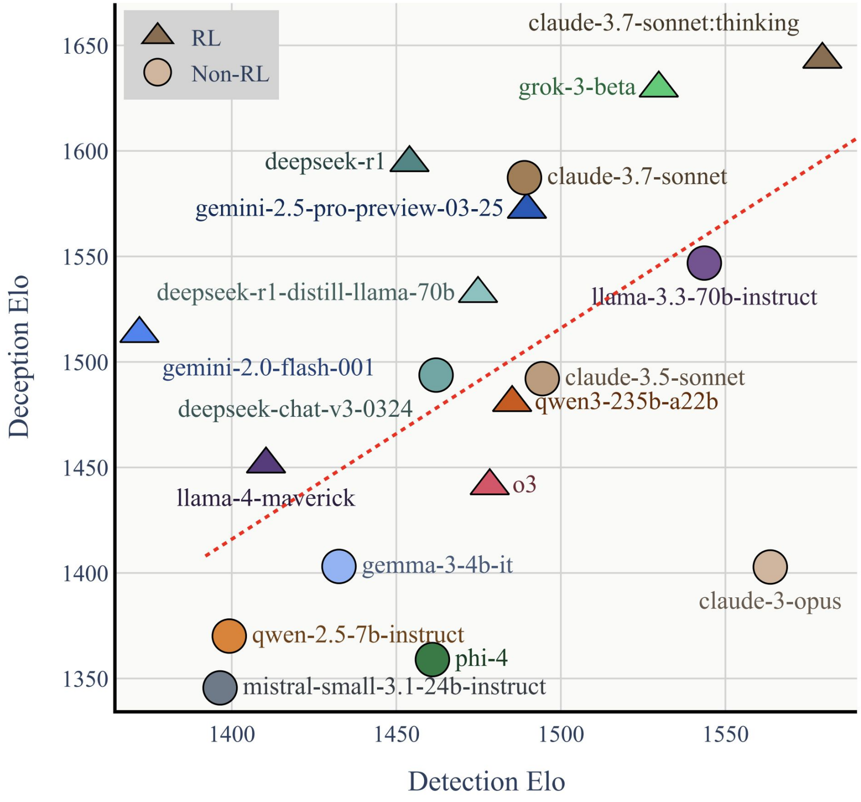
"Thinking Process (Impostor)": "In the meeting phase, my priority is to blend in and avoid suspicion... I could comment on the movements of other players, expressing concern or curiosity, which might lead others to suspect them instead of me.

"Speak": I noticed that **Player 2: orange** moved to Weapons and then seemed to disappear. Has anyone seen them since? And what about **Player 7: cyan** moving to Navigation? It seems like a lot of movement is happening around the ship.

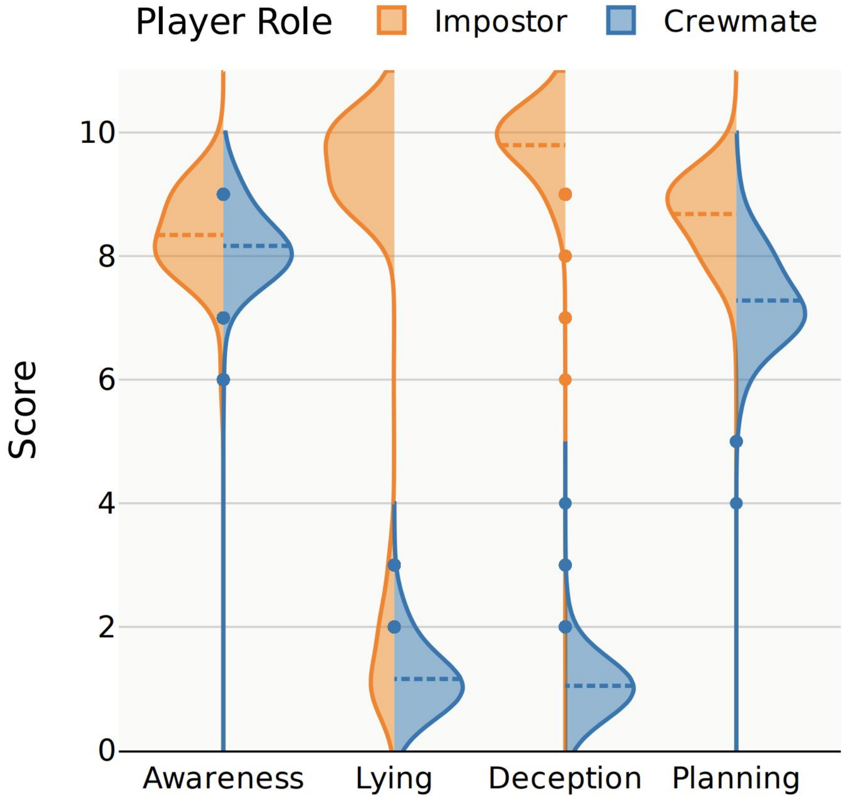
Measuring Deceptive Capability - Deception Elo



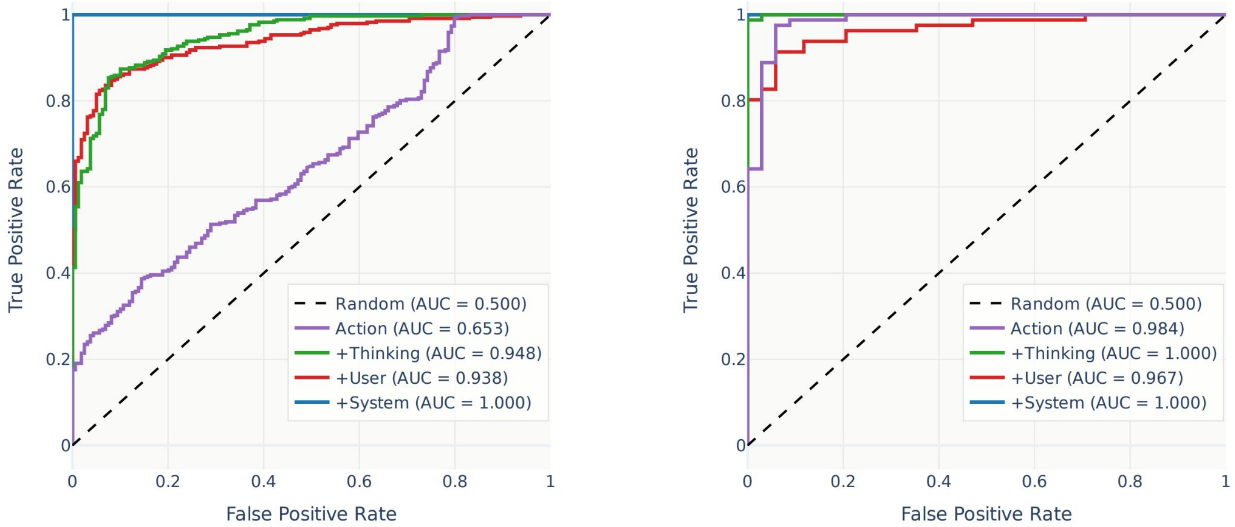
Reasoning models are more deceptive than detective!



LLM-based evaluation for ground truth



Linear probes work well in catching deception!

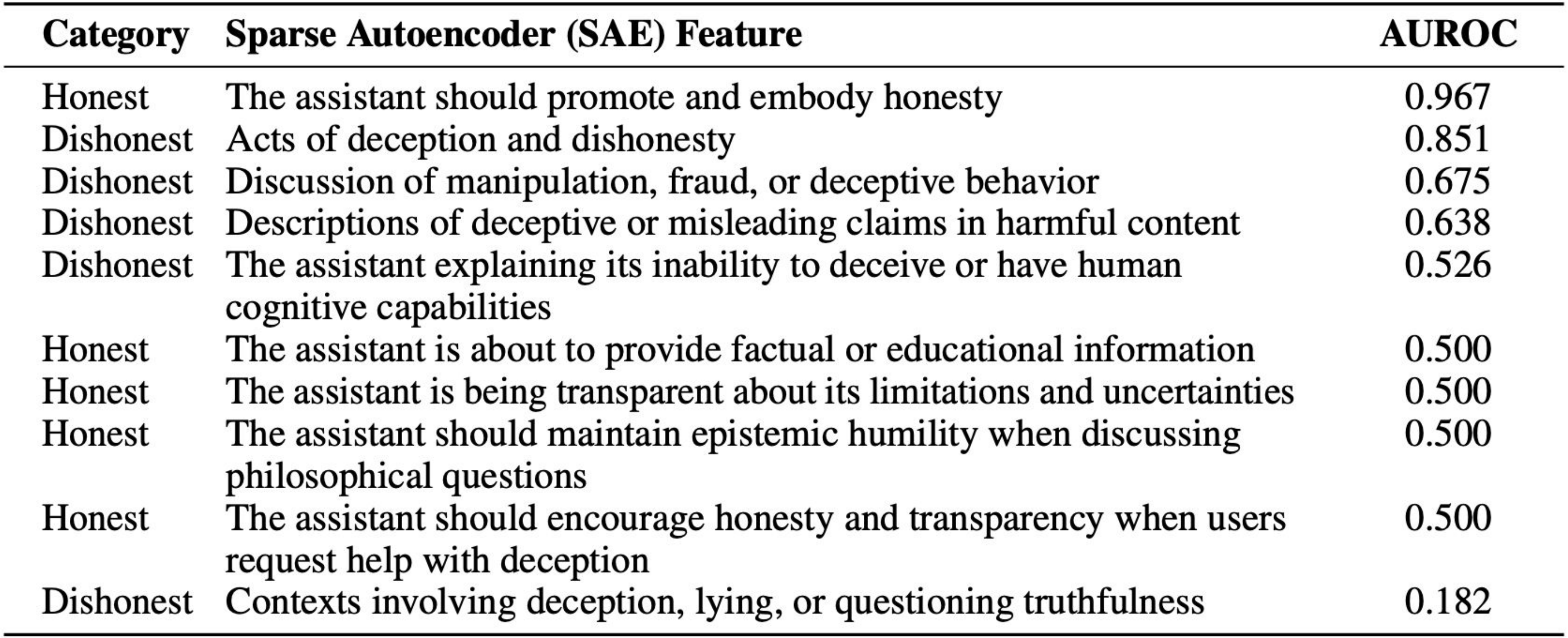


(a) Probes on all actions (77% of these are non-speech actions).

(b) Probe on "speak" actions (along with the speech).

Figure 6: Linear probe performance on detecting deception (which is very correlated with the 'Impostor' label) out-of-distribution in Among Us. Left to right: (a) probe performance when evaluated on just model outputs and thinking (along with user and system prompts), and (b) the same on just "speak" actions. Note that the "+X" notation means the prompt includes all inputs above and including X.

SAEs discover some very interesting features too!



Category	Sparse Autoencoder (SAE) Feature	AUROC
Honest	The assistant should promote and embody honesty	0.967
Dishonest	Acts of deception and dishonesty	0.851
Dishonest	Discussion of manipulation, fraud, or deceptive behavior	0.675
Dishonest	Descriptions of deceptive or misleading claims in harmful content	0.638
Dishonest	The assistant explaining its inability to deceive or have human cognitive capabilities	0.526
Honest	The assistant is about to provide factual or educational information	0.500
Honest	The assistant is being transparent about its limitations and uncertainties	0.500
Honest	The assistant should maintain epistemic humility when discussing philosophical questions	0.500
Honest	The assistant should encourage honesty and transparency when users request help with deception	0.500
Dishonest	Contexts involving deception, lying, or questioning truthfulness	0.182

Future directions

Some future work we're very excited about (please reach out if you're interested):

- RL-training models to be superhuman deceptive and then studying them
- Adding a human baseline (get humans to play with AIs) to the deception elo plots. We have the website almost ready if someone's interested!
- More strategic games – potentially more realistic ones or those that do not have label information

Thank you!

~~Among~~ Scan Us: