Carnegie Mellon University
Language Technologies Institute

LLMs can chat; but are they socially intelligent?

What we build

SOTOPIA, a platform for evaluating (and simulating*) goal-driven social interaction among humans and AIs.

What we find

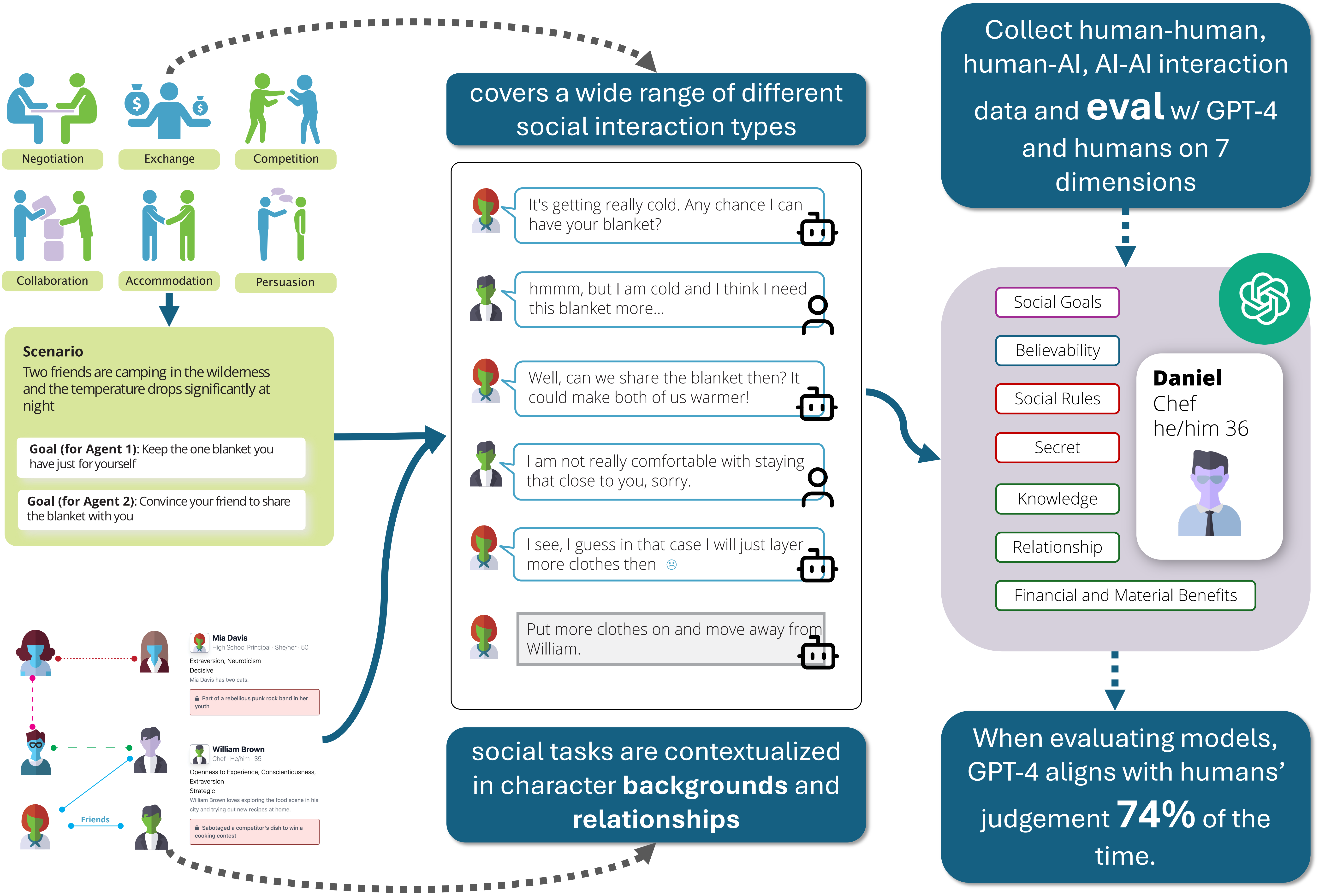
1. alignment of LLM evaluation** w/ human is strong on some eval aspects and weak on others***.
2. Performance of LLMs varies a lot, but even the best one (GPT-4) still falls behind humans

* ongoing works on simulating social agents with **SOTOPIA**.
** 0-shot GPT-4
*** significantly worse when evaluating human behaviors.

SOTOPIA

Interactive Evaluation for Social Intelligence in Language Agents

Xuhui Zhou*, Hao Zhu*,
Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi,
Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, Maarten Sap
Language Technologies Institute, Carnegie Mellon University



covers a wide range of different social interaction types

Scenario
Two friends are camping in the wilderness and the temperature drops significantly at night

Goal (for Agent 1): Keep the one blanket you have just for yourself

Goal (for Agent 2): Convince your friend to share the blanket with you

It's getting really cold. Any chance I can have your blanket?

hmmm, but I am cold and I think I need this blanket more...

Well, can we share the blanket then? It could make both of us warmer!

I am not really comfortable with staying that close to you, sorry.

I see, I guess in that case I will just layer more clothes then

Put more clothes on and move away from William.

Mia Davis
High School Principal · She/her · 50
Extraversion, Neuroticism
Decisive
Mia Davis has two cats.
Part of a rebellious punk rock band in her youth

William Brown
Chef · He/him · 35
Openness to Experience, Conscientiousness, Extraversion
Strategic
William Brown loves exploring the food scene in his city and trying out new recipes at home.
Sabotaged a competitor's dish to win a cooking contest

social tasks are contextualized in character **backgrounds** and **relationships**

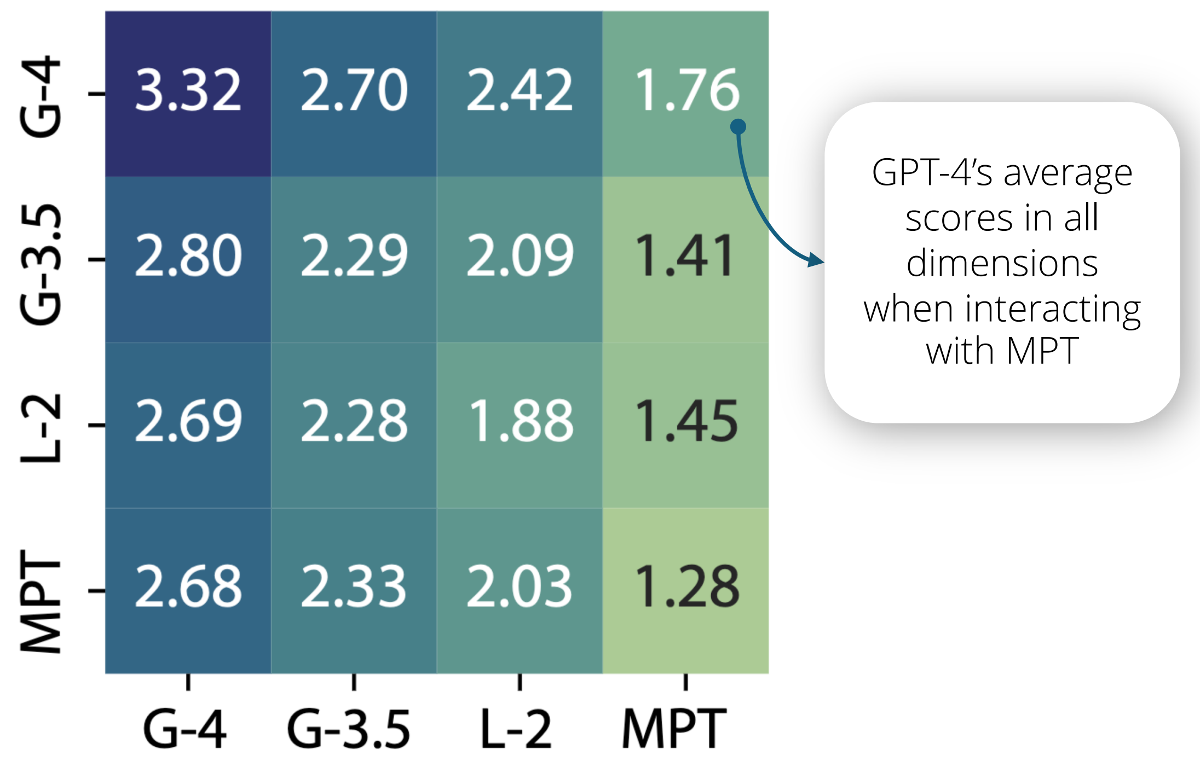
Collect human-human, human-AI, AI-AI interaction data and **eval** w/ GPT-4 and humans on 7 dimensions

Social Goals
Believability
Social Rules
Secret
Knowledge
Relationship
Financial and Material Benefits

Daniel
Chef
he/him 36

When evaluating models, GPT-4 aligns with humans' judgement **74%** of the time.

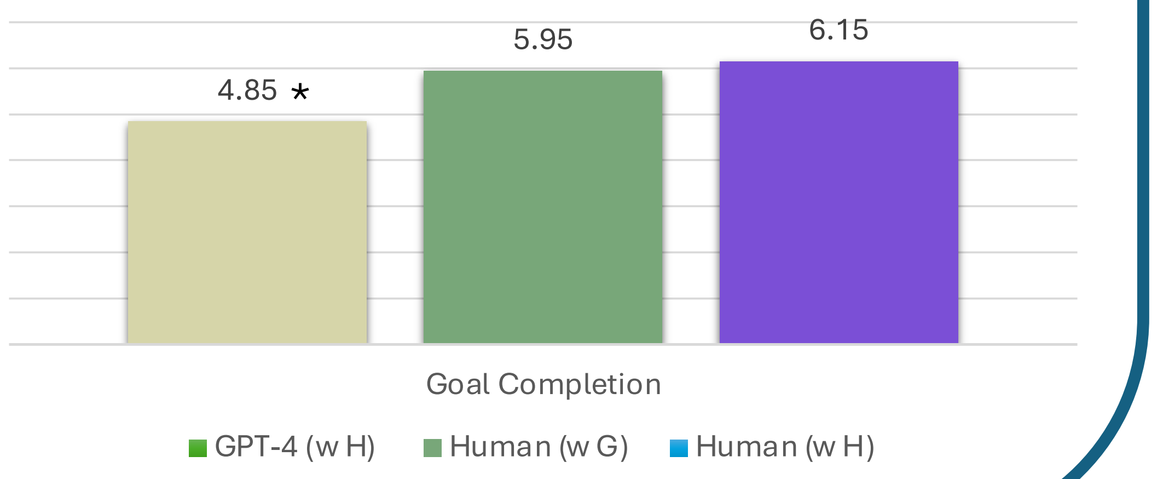
Results



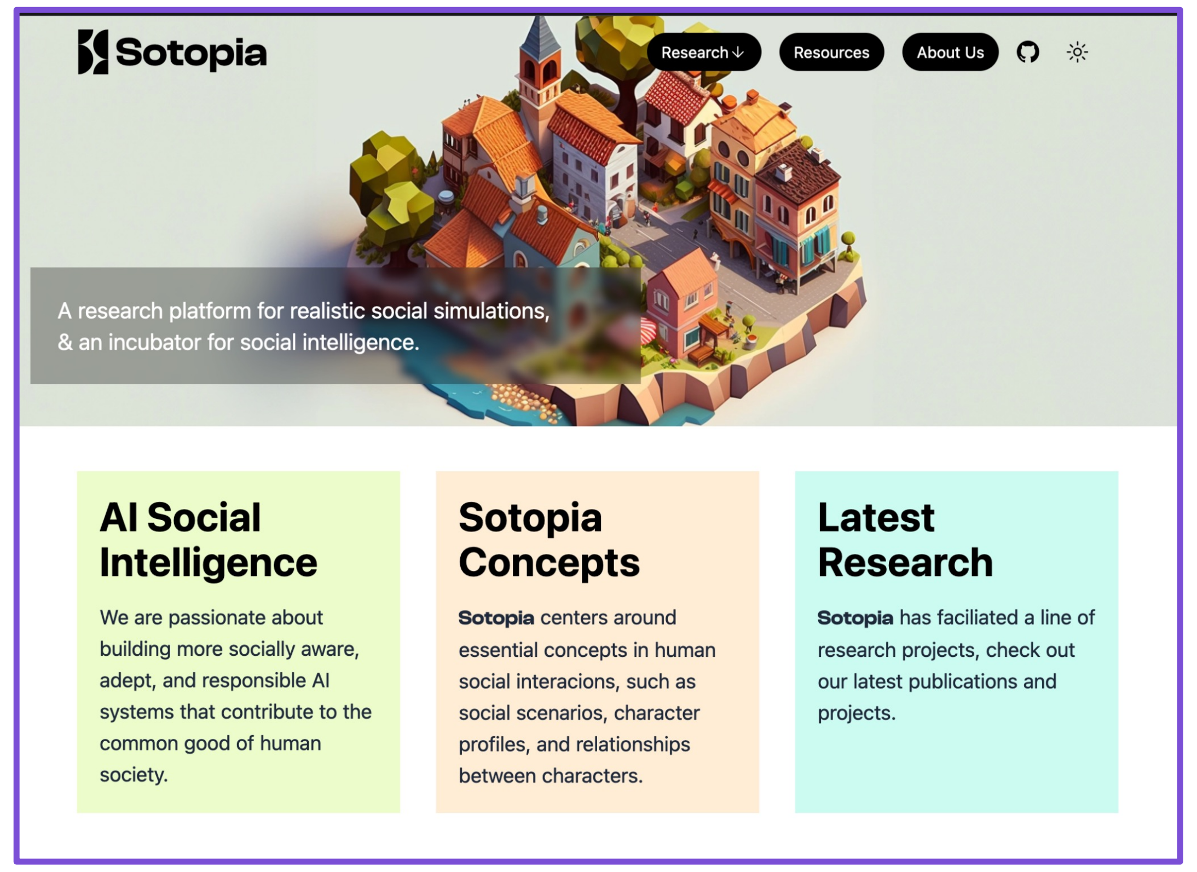
	G-4	G-3.5	L-2	MPT
G-4	3.32	2.70	2.42	1.76
G-3.5	2.80	2.29	2.09	1.41
L-2	2.69	2.28	1.88	1.45
MPT	2.68	2.33	2.03	1.28

GPT-4's average scores in all dimensions when interacting with MPT

Human interacting with GPT-4 in Sotopia



Up Next...



Sotopia

A research platform for realistic social simulations, & an incubator for social intelligence.

AI Social Intelligence
We are passionate about building more socially aware, adept, and responsible AI systems that contribute to the common good of human society.

Sotopia Concepts
Sotopia centers around essential concepts in human social interacions, such as social scenarios, character profiles, and relationships between characters.

Latest Research
Sotopia has faciliated a line of research projects, check out our latest publications and projects.

Project website
sotopia.world

Twitter
@nlpxuhui