

Meaningful Interaction with Unreal Speakers?

David Schlangen

Grundlagen der Computerlinguistik // Department Linguistik // Universität Potsdam

<http://clp.ling.uni-potsdam.de>

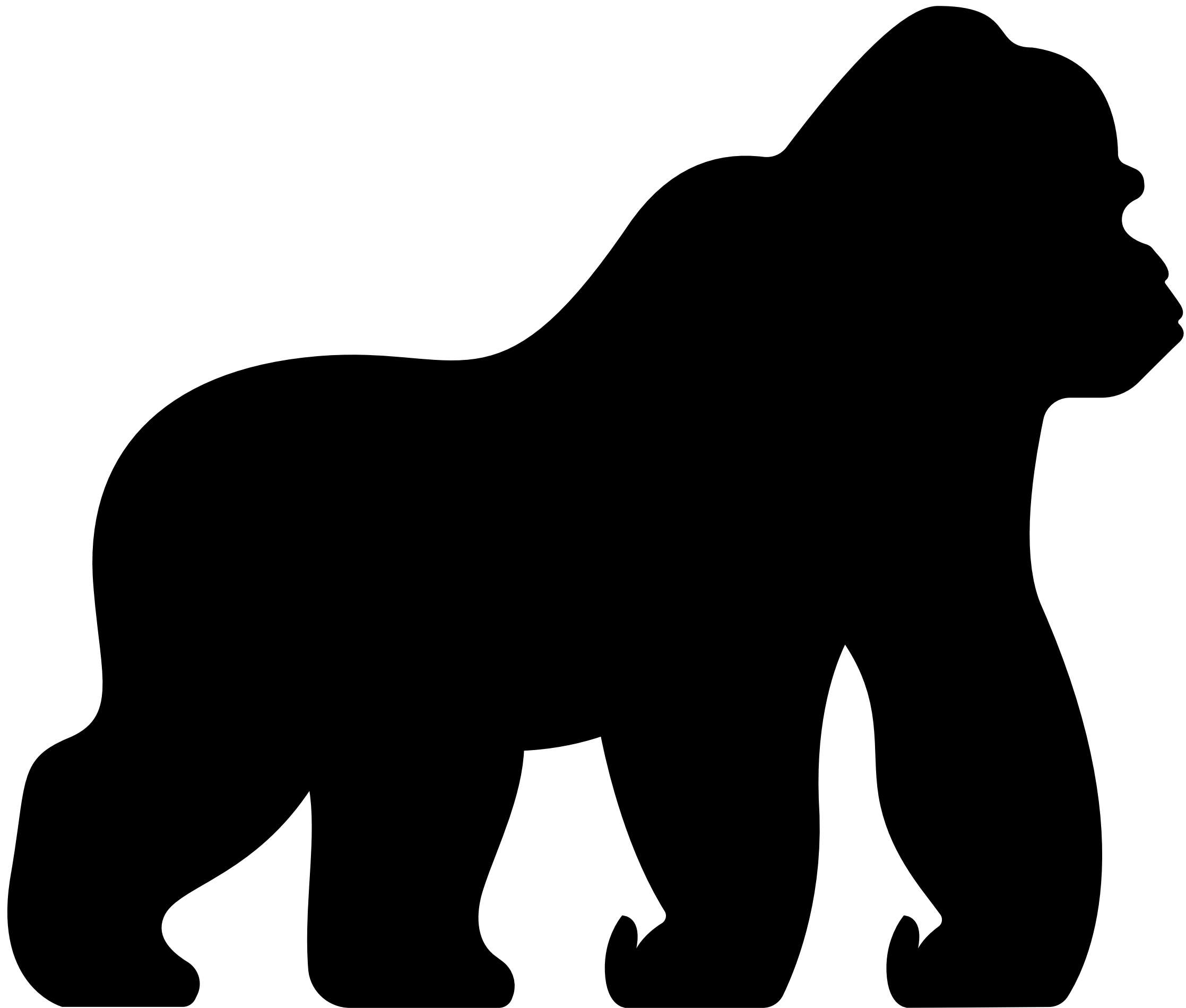
david.schlangen@uni-potsdam.de

@ semdial 2025

2025-09-05

These slides:

<https://clp.ling.uni-potsdam.de/talks>



**THIS IS
WEIRD**

THIS IS WEIRD

ChatGPT Statistics 2025: Top Picks

- As of August 2025, ChatGPT has close to 800 million weekly active users.

←  r/ChatGPT • vor 2 Monaten
cursed_noodle

How many hours a day do you spend on chatGPT?

←  r/artificial • vor 4 Monaten
dawnfire05

Which AI is best for long or serious conversations?

ork Times

OPINION
GUEST ESSAY

She Is in Love With ChatGPT

A 28-year-old woman with a busy social life spends hours on end

BOT THOUGHTS

PEOPLE ARE LOSING LOVED ONES TO AI FUELED SPIRITUAL FANTASIES

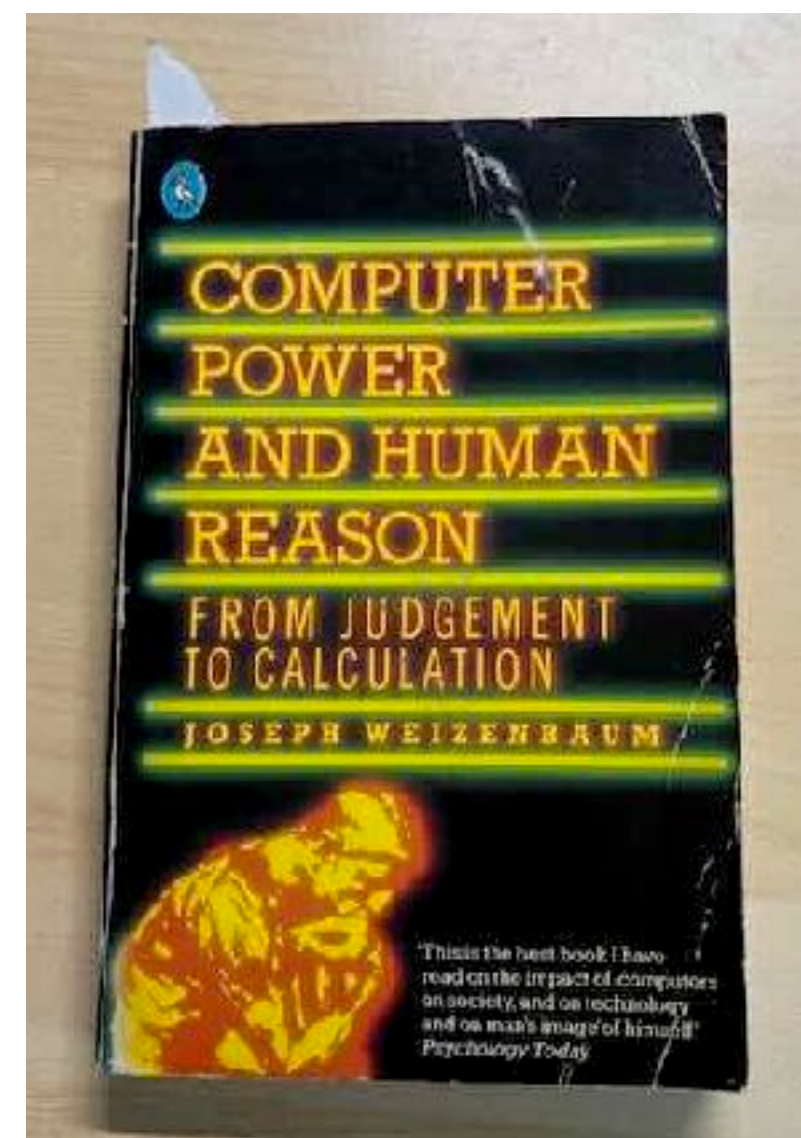
Self-styled prophets are claiming they have “awakened” chatbots
and accessed the secrets of the universe through ChatGPT

By MILES KLEE

MAY 4, 2025

Rolling Stone

...from your company. I was
with your prompt delivery. I
opened the package, I found it
either in the box or with the
it is essential that our account
department have the invoice by
the first quarter. It would be
accountants and I if you will
other invoice as soon as possible



THIS IS WEIRD

ChatGPT Statistics 2025: Top Picks

- As of August 2025, ChatGPT has close to 800 million weekly active users.

←  r/ChatGPT • vor 2 Monaten
cursed_noodle

How many hours a day do you spend on chatGPT?

←  r/artificial • vor 4 Monaten
dawnfire05

Which AI is best for long or serious conversations?

York Times

OPINION
GUEST ESSAY

She Is in Love With ChatGPT

A 28-year-old woman with a busy social life spends hours on end

BOT THOUGHTS

PEOPLE ARE LOSING LOVED ONES TO AI FUELED SPIRITUAL FANTASIES

Self-styled prophets are claiming they have “awakened” chatbots
and accessed the secrets of the universe through ChatGPT

By MILES KLEE

MAY 4, 2025

Rolling Stone

THIS IS A BIG DEAL (potentially)

Humanity is made out of conversation

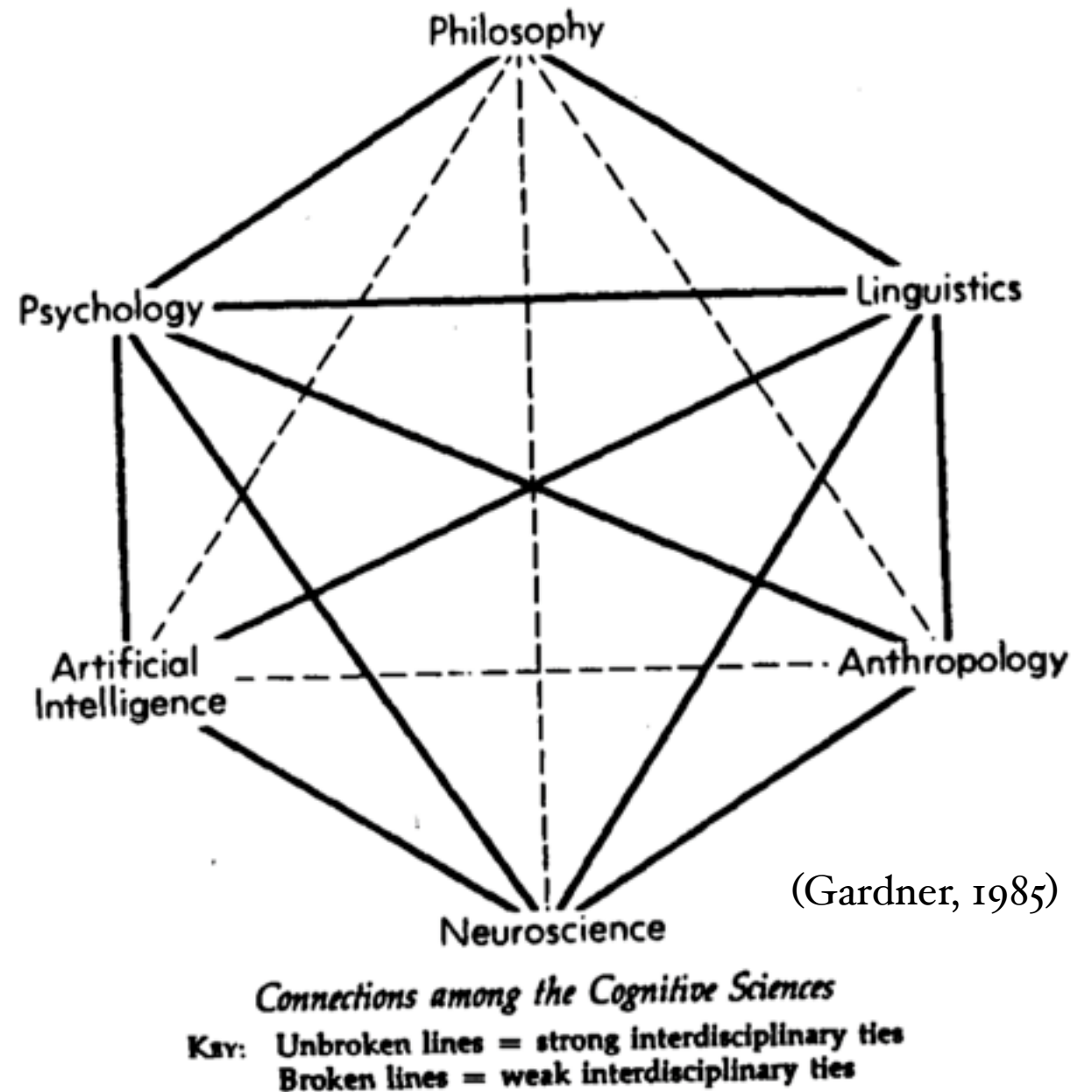
- Conversation is how the individual enters society
- Conversation is how society enters the individual
- Conversation is how individuals relate
- Conversation is how individuals make society

This Talk

- Intro
- Part I: What is *assertion*?
- Part II: Unnatural Interfaces for Unreal Speakers
- Part III: Theoretical & Computational Models / SAM & clem

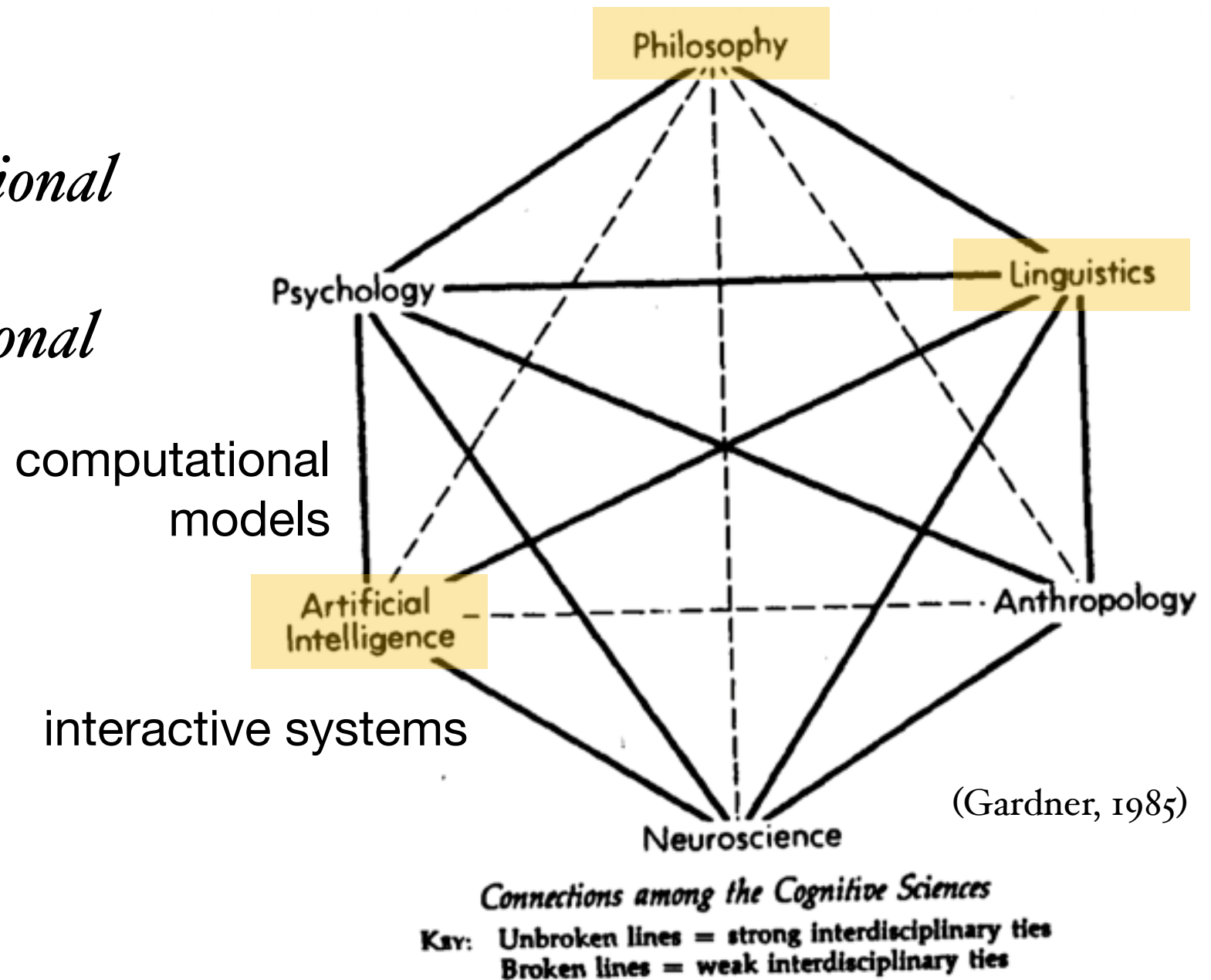
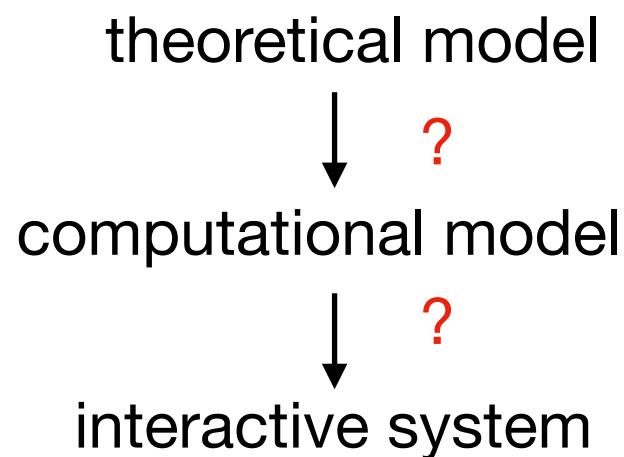
This Community

- Intro
- Part I: What is *assertion*'?
- Part II: Unnatural Interfaces for Unreal Speakers
- Part III: Theoretical & Computational Models / SAM & clem



This Community

Using the formal tools of philosophical logic to provide linguistic analyses of conversational phenomena, with an eye on the implementability in computational systems.



This Talk

- Intro
- Part I: What is *assertion*?
Our tools are useful for understanding the current situation
- Part II: Unnatural
Interfaces for Unreal
Speakers
interactive systems = computational models * *friction_coefficient*
- Part III: Theoretical &
Computational Models /
SAM & clem
 $model_t$ guides investigation of $model_c$; $model_c$ provides learnability arguments

This Talk

- Part I: What is *assertion*'?

Our tools are useful for understanding the current situation

(Schlangen 2022)

Norm Participation Grounds Language

David Schlangen

CoLabPotsdam / Computational Linguistics
Department of Linguistics, University of Potsdam, Germany
david.schlangen@uni-potsdam.de

Abstract

The striking recent advances in eliciting seemingly meaningful language behaviour from

associated with nature and growth.”¹

This of course is a profoundly unfair test. The model has no connection to you other than through what you type, and so can't observe what “these”

The Puzzle

The puzzle: By now, interactions with AI-chatbots can very much *feel like* conversations. But at the same time, there is something weird about them.

(Allegedly, pass Turing test, Jones & Bergen [2025])

Some responses:

- “No meaning”
 - “meaning can’t be learned from linguistic form alone” (Bender & Koller 2020)
 - “this isn’t meaning or understanding”, “synthetic text extruding machines”, “looks like communicative language, but without any intent or thinking mind behind it” (Bender & Hanna 2025, p. 30)
- “bullshit [= disregard for truth]” (e.g., Liang *et al.* 2025, *inter alia*)

The Puzzle

The puzzle: By now, interactions with AI-chatbots can very much *feel like* conversations. But at the same time, there is something weird about them.
(Allegedly, pass Turing test, Jones & Bergen [2025])

The approach taken here:

- Rather than address unclear notions like “meaning” and “understanding” directly, look at the effectiveness of the machine utterances as *speech acts*.
- As example: *assertion* . How does human assertion work; and does machine assertion (*assertion'*) work the same way?
- Result: Chatbots are only able to produce *loose talk*.

The Argument

assertion requires personhood

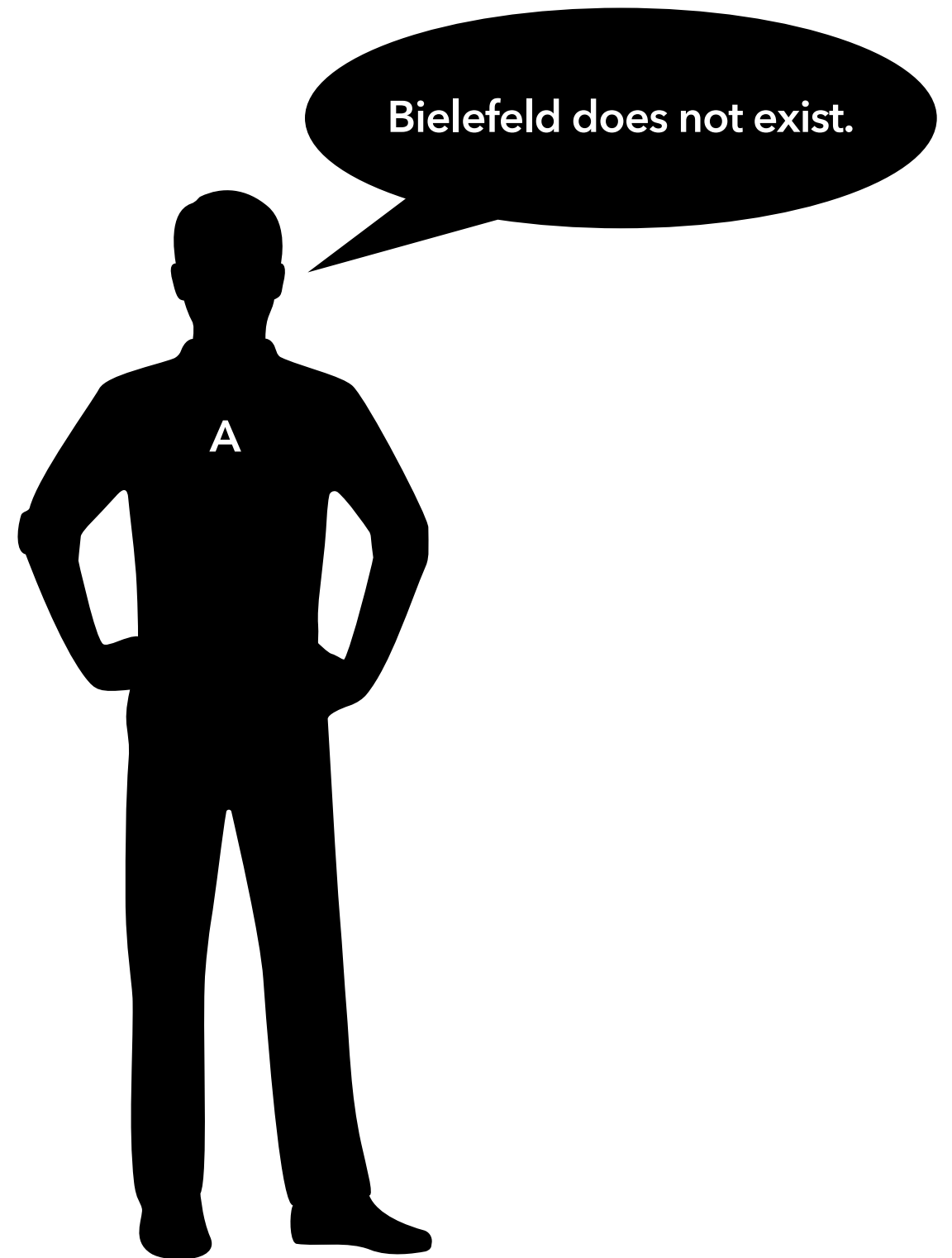
LLMs do not possess personhood

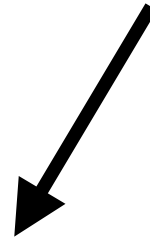
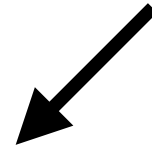
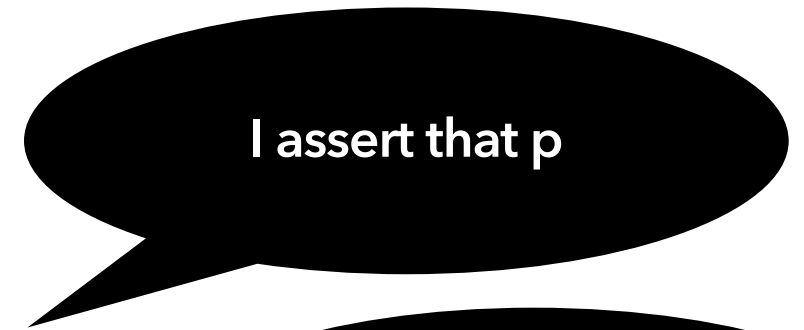
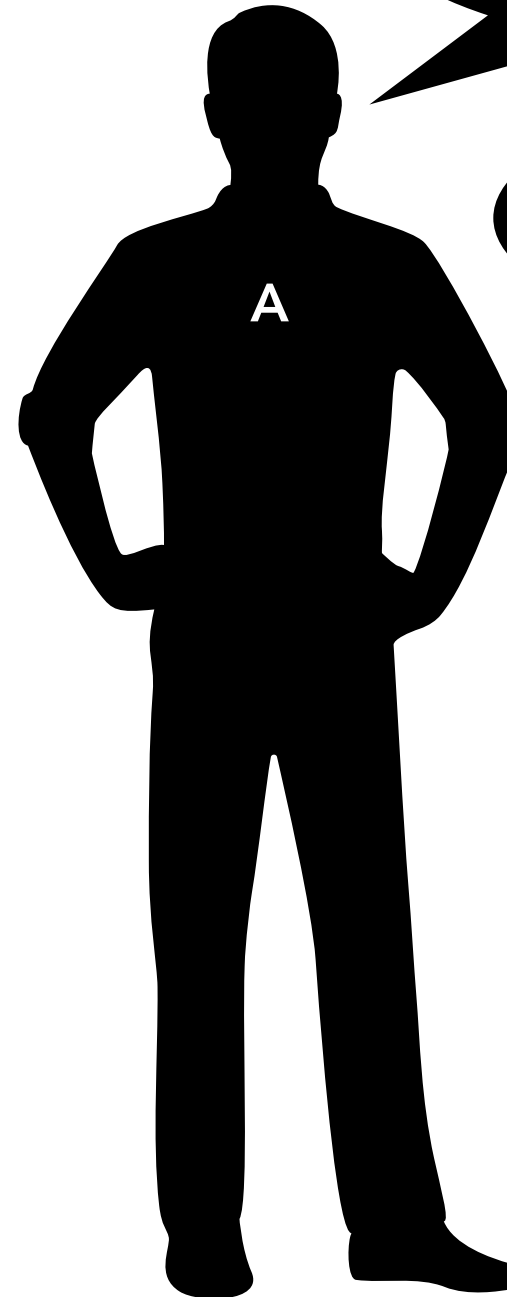
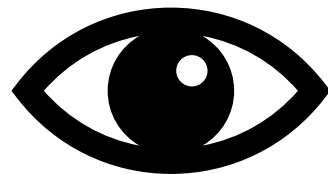
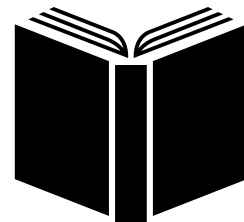
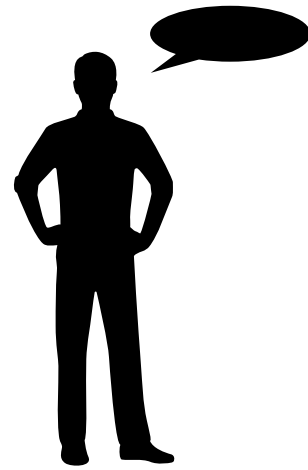
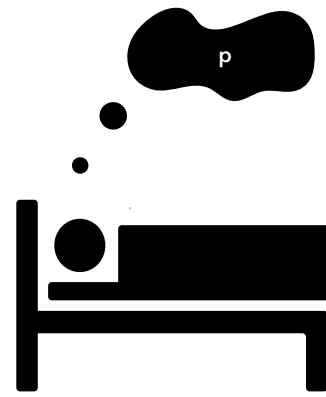
therefore, LLMs do not *assert*

assertion requires personhood

LLMs *assert*

therefore, LLMs possess personhood





knowledge from testimony

epistemic
buck-passing
(Goldberg 2015)

direct knowledge

susceptibility
to challenge
(Goldberg 2015)

I, person A, believe *that* *p*, and
have good reasons to do so.

Falsely entering a commitment
(here: asserting something you
do not have good reasons to
believe) *has consequences*.

(What these are depends on
context; can range from
causing mild annoyance to
loosing status to loosing
freedom.)

If it cannot have consequences
when broken, it's not a
commitment.



commitment

I, person A, believe *that p*, and
have good reasons to do so.

(Levinson 1983) An *assertion* that p is a function from a context where the speaker S is not committed to p [...] into a context in which S is committed to the justified true belief that p .

assert that p

(Brandom 1994) According to the model, to treat a performance as an assertion is to treat it as the undertaking or acknowledging of a certain kind of commitment—what will be called a doxastic', or 'assertional', commitment.

Justify your claim!

To be doxastically committed is to have a certain social status. Doxastic commitments are normative, more specifically deontic, statuses. Such statuses are creatures of the practical attitudes of the members of a linguistic community—they are instituted by practices governing the taking and treating of individuals as committed.

(Williamson 2000) (The warrant rule) One must: assert p only if one has warrant to assert p .

(Goldberg 2015) MMENA It is mutually manifest to participants in a speech exchange that assertion has a robustly epistemic norm; that is, that one must: assert that p only if $E(\text{one}, \{p\})$

(Levinson 1983) An *assertion* that p is a function from a context where the speaker S is not committed to p [...] into a context in which S is committed to the justified true belief that p .

assert that p

(Brandom 1994) According to the model, to treat a performance as an assertion is to treat it as the undertaking or acknowledging of a certain kind of commitment—what will be called a doxastic', or 'assertional', commitment.

Justify your claim!

To be doxastically committed is to have a certain social status. Doxastic commitments are normative, more specifically deontic, statuses. Such statuses are creatures of the practical attitudes of the members of a linguistic community—they are instituted by practices governing the taking and treating of individuals as committed.

(Williamson 2000) (The warrant rule) One must: assert p only if one has warrant to assert p .

(Goldberg 2015) MMENA It is mutually manifest to participants in a speech exchange that assertion has a robustly epistemic norm; that is, that one must: assert that p only if $E(\text{one}, \{p\})$

The Argument

assertion requires ^{commitment} personhood

LLMs do not possess personhood

therefore, LLMs do not *assert*

human person				
species	organism	embodied		
		- individually instantiated		
		- persistent, self-maintained		
		- perishable, irrecoverable		
	cognition	beliefs, desires, intentions		
		theoretical reason		
		practical reason		
		aesthetic sense		
		relational self-image		
individual expression	experientially acquired in vivo	skills		
		traits, dispositions		
		“world view”		
		virtue		
moral status	free			
	rights and obligations			

individual, able to make free use of their developed capabilities

(e.g., Quante 2012)

		human person	LLM chatbot	
species	organism	embodied	virtual	
		- individually instantiated	type/token confusion	
		- persistent, self-maintained	externally maintained	
	cognition	- perishable, irrecoverable	recoverable, backupable	
		beliefs, desires, intentions	can produce BDI expr.	
		theoretical reason	“reasoning” models /	
		practical reason	post-training	
		aesthetic sense		
		relational self-image	in vitro	
	individual expression	experientially acquired in vivo		
		skills		
		traits, dispositions	persona prompt	
		“world view”		
moral status		virtue	“alignment” training	
		free	owned	
		rights and obligations	/	

individual, able to make free use of their developed capabilities

(e.g., Quante 2012)

		human person	LLM chatbot
species	organism	embodied	virtual
		- individually instantiated	type/token confusion
		- persistent, self-maintained	externally maintained
		- perishable, irrecoverable	recoverable, backupable
	beliefs, desires, intentions		can produce BDL expr
No matter how good the behavioural model, there is a remainder that separates the model from the modelled: The behaviour of a person is produced by an individual that has reasons to care for itself, and that hence is sanctionable.			
individual	expression	relational self-image	in vit
		experientially acquired in vivo	skills
		traits, dispositions	persona prompt
		“world view”	
moral status		virtue	“alignment” training
		free	owned
		rights and obligations	/

individual, able to make free use of their developed capabilities

(e.g., Quante 2012)

The Conclusion

assertion requires personhood

LLMs do not possess personhood

therefore, LLMs do not *assert*

assertion: has doxastic effect, that comes with a justified entitlement grounded in the speaker's commitment

assertion': can have doxastic effect, which however the hearer isn't entitled to, as the act does not have enforceable effect on speaker

The weirdness comes from the tension between the appearance of *assertion*, with vague awareness of it being *assertion'*.

The model/reality clash.

What's the problem?

- Suggesting something is *assertion* when it is *assertion*' is bad because it leaves people holding the buck when they don't expect it; in effect, it dilutes the societal body of knowledge
 - This can even be done on purpose and at scale by malicious actors.
- Suggesting personhood when there is none leaves people in danger of forming unreal relationships in place of real ones
 - See recent reports of chatbots reinforcing suicidal ideations.

What is being / can be done?

- Make more person-like:
 - make models identifiable (“Claude”), and suffer consequences (“if Claude gives bad replies, I stop using it / it will be retrained”).
 - Some semblance of accountability, but only in aggregate.
- make the chatbot a “legal person”.
 - Legal personhood bottoms out in (natural) personal responsibility (CEO, etc.), even if limited. No CEO seems to be willing to do that.

What is being / can be done?

- Make nicer (pseudo-)person: “alignment” with “values”. More virtuous behaviour. (Withhold information, disagree, end conversation, etc..)
 - Does not address relational problem.
- Make challenges unnecessary, increase accuracy to 95+⁰%
 - Works only for trivia facts. In most domains, knowledge doesn’t work like that.
 - Still only probabilistic connection. No guarantees.
- Explicitly pass the buck: add attribution to all claims.
 - Asserted attribution just as unreliable as main claim.

What is being / can be done?

- Make people understand difference between normal conversation and pseudo-conversation:
 - “Trust the chatbot as much as you trust a conversation in a dream.”
 - “This is improv theater. The actor on stage who you’re giving prompts to is playing someone who is talking to you.”
 - Not: “ChatGPT can make errors.”
- Discharge commitments ASAP:
 - “This is the correct solution” harms less in situations where the test follows soon. (E.g., verifiable domains; situated interactions.)
- Design for non-personhood. Anti-anthropomorphism / Artificialisation by design.
 - (conversational) intelligence without implied personhood / without relationships

The Fine Print

- MacFarlane (2010) distinguishes four types of accounts of assertion (attitudinal, common ground, commitment, constitutive rule); the argument here works at least with the last two.
- See (Butlin & Viebahn 2025) for an argument very much along these lines (“to be capable of assertion, and entity must [...] be capable of being sanctioned”). But see (Williams & Bayne 2024) for an account that classifies chatbots as “proto-asserters”.
- The argument here is a spin-off of the one in my (Schlangen 2022; “norm participation grounds language”), which itself however has a larger scope. Roughly:
 - Every mental activity (recognising, believing, saying) comes with an implicit assertion “this is good”, which appeals to public normative standards in ways that LLM activities do not (and cannot).
 - The development of the argument here adds to that paper what could be summarised as “norm participation is grounded in mutually recognised personhood”.

Part I: What we can do

- Our community can contribute
 - to the analysis of the situation
 - to the attempts at mitigating it (if by criticising them)

This Talk

- Intro
- Part I: What is *assertion*?
Our tools are useful for understanding the current situation
- Part II: Unnatural Interfaces for Unreal Speakers
(Conversational) intelligence without implied personhood / without relationships
interactive systems = computational models * *friction_coefficient*
- Part III: Theoretical & Computational Models / SAM & clem
model_t guides investigation of model_c ; model_c provides learnability arguments



Human Factors in Computing Systems

CHI '94 • "Celebrating Interdependence"

CHI '92

May 3 - 7, 1992



ANTHROPOMORPHISM: FROM ELIZA TO TERMINATOR 2

Moderator: Abbe Don
IN CONTEXT
3435 Clay Street #4
San Francisco, CA 94118
415-567-8130
abbe@well.sf.ca.us

Panelists: Susan Brennan, State University of New York at Stony Brook
Brenda Laurel, Telepresence Research
Ben Shneiderman, University of Maryland

INTRODUCTION

"I feel depressed."
"WHY DO YOU FEEL DEPRESSED?"
Eliza, 1966

"Open the pod bay door, Hal."
"I CAN'T DO THAT DAVE"
2001 A Space Odyssey, 1968

"MAN OF THE YEAR"
Time Magazine, 1982

"YOUR DOOR IS AJAR"
Chrysler Le Baron, 1983

anthropomorphism in the interface. Like any taboo, simply sweeping away the issue (assigning human characteristics to the computer) does not make it go away. The examples above highlight some of the contexts in which scientists, designers and filmmakers have explored the implications of anthropomorphizing the human-machine interface. Some techniques, such as talking cars, have dissipated in response to users' distaste. Others, such as the Guides approach, need further refinement; and still others, like the images floating through popular culture, warrant examination as they help us define both our fantasies and fears about our relationship to machines.

With command line interfaces and text only systems, the effect of the anthropomorphism taboo translated as, "don't

Dehumanizing Machines:

Better *Slow* than Sorry: Introducing Positive Friction for Reliable Dialogue Systems

Mert İnan^{*}, Anthony Sicilia^{*}, Suvodip Dey[✉], Vardhan Dongre[✉], Tejas Srinivasan[✉]
Jesse Thomason[✉], Gökhan Tür[✉], Dilek Hakkani-Tür[✉], Malihe Alikhani^{*}

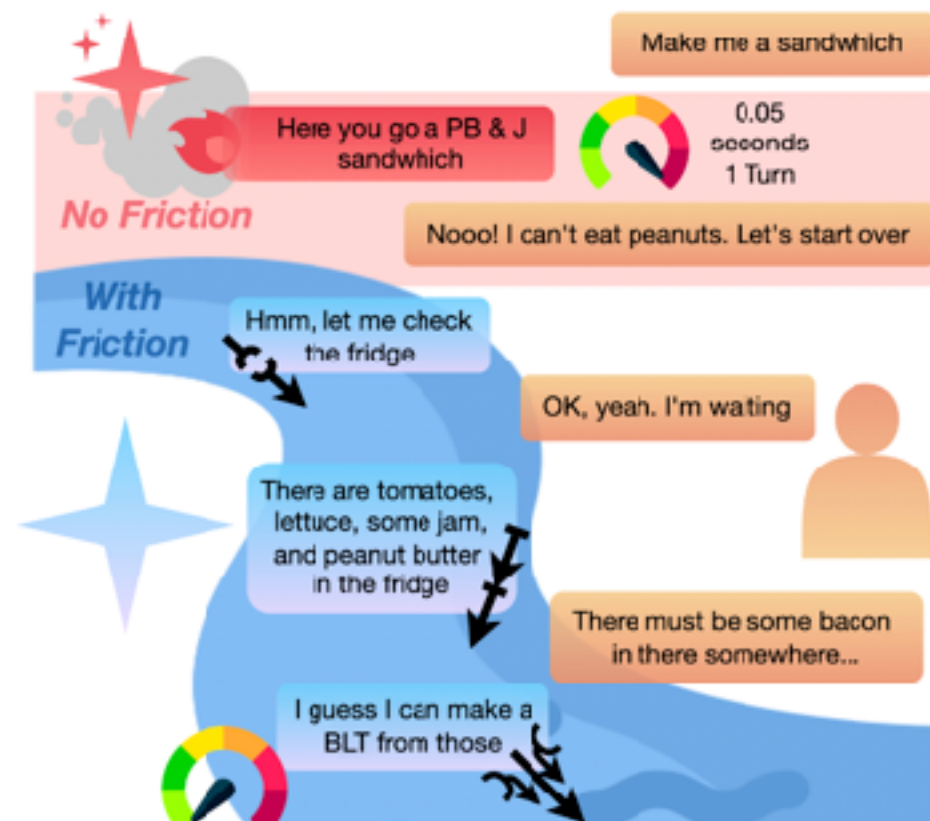
[✉] University of Southern California [✉] University of Illinois Urbana-Champaign

^{*} Northeastern University

{inan.m, alikhani.m}@northeastern.edu

Abstract

While theories of discourse and cognitive science have long recognized the value of unhurried pacing, recent dialogue research tends to minimize friction in conversational systems. Yet, frictionless dialogue risks fostering uncritical reliance on AI outputs, which can obscure implicit assumptions and lead to unintended consequences. To meet this challenge, we propose integrating *positive friction* into conversational AI, which promotes user reflection on goals, critical thinking on system response, and subsequent re-conditioning of AI systems. We hypothesize systems can improve goal



New Research Questions

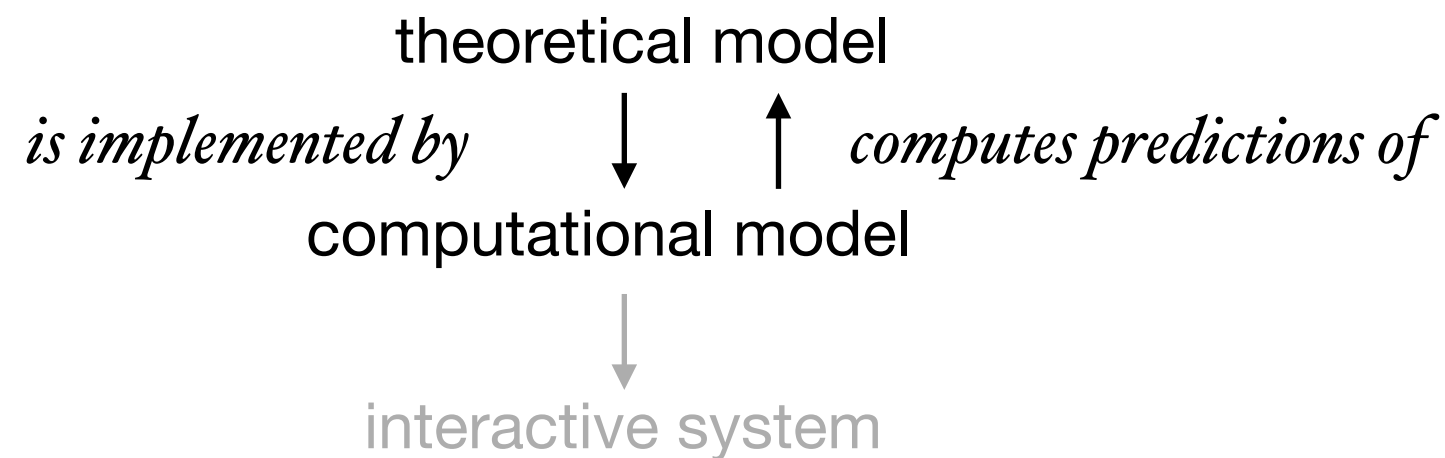
- What are *anti-cues* for personhood?
 - on content level (e.g., avoid 1st person pronouns, use “this machine”; avoid psych verbs; etc.)
 - on interaction level (e.g., no interruptions, unnatural sound signals (instead of breathing sounds etc.), voice quality
 - on embodiment level (e.g., face, expressivity, etc.)
- ... that can be implemented, while still enabling non-annoying, non-rigid interactions?

This Talk

- Intro
- Part I: What is *assertion*?
Our tools are useful for understanding the current situation
- Part II: Unnatural
Interfaces for Unreal
Speakers
interactive systems = computational models * *friction_coefficient*
- Part III: Theoretical &
Computational Models /
SAM & clem
 $model_t$ guides investigation of $model_c$; $model_c$ provides learnability arguments

This Talk

- Part III: Theoretical & Computational Models / SAM & clem



This Talk

Resolving Underspecification using Discourse Information

seminal 2001

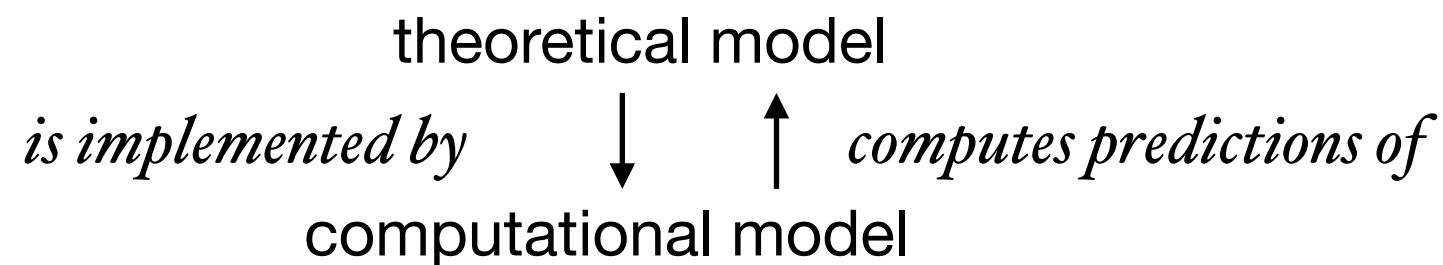
DAVID SCHLANGEN, ALEX LASCARIDES, ANN COPESTAKE
UNIVERSITY OF EDINBURGH; UNIVERSITY OF CAMBRIDGE AND STANFORD UNIVERSITY
{das|alex}@cogsci.ed.ac.uk, aac@cl.cam.ac.uk
<http://www.cogsci.ed.ac.uk/~das>

Abstract

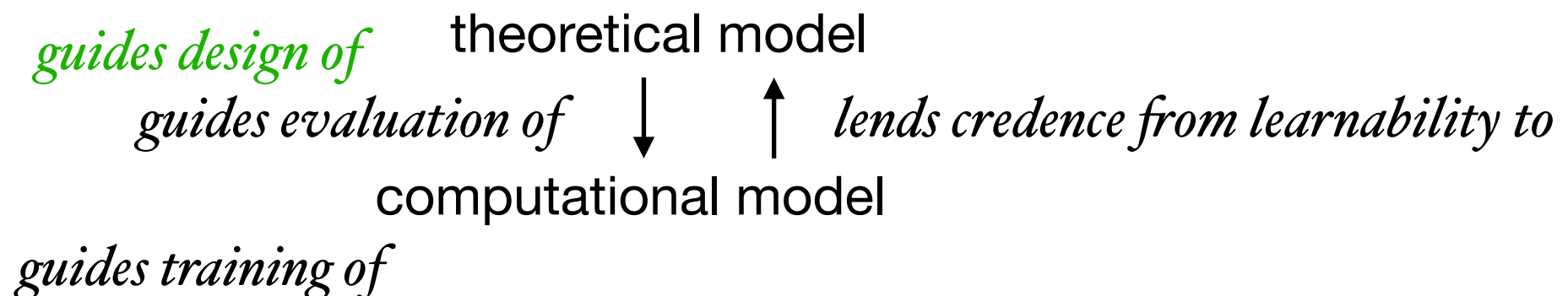
This paper describes RUDI (“Resolving Underspecification with Discourse Information”), a dialogue system component which computes automatically some aspects of the content of scheduling dialogues, particularly the intended denotation of the temporal expressions, the speech acts performed and the underlying goals. RUDI has a number of nice features: it is a principled approximation of a logically precise and linguistically motivated framework for representing semantics and implicatures; it has a particularly simple architecture; and it records how reasoning with a combination of goals, semantics and speech acts serves to resolve underspecification that’s generated by the grammar.

This Talk

- Part III: Theoretical & Computational Models / SAM & clem

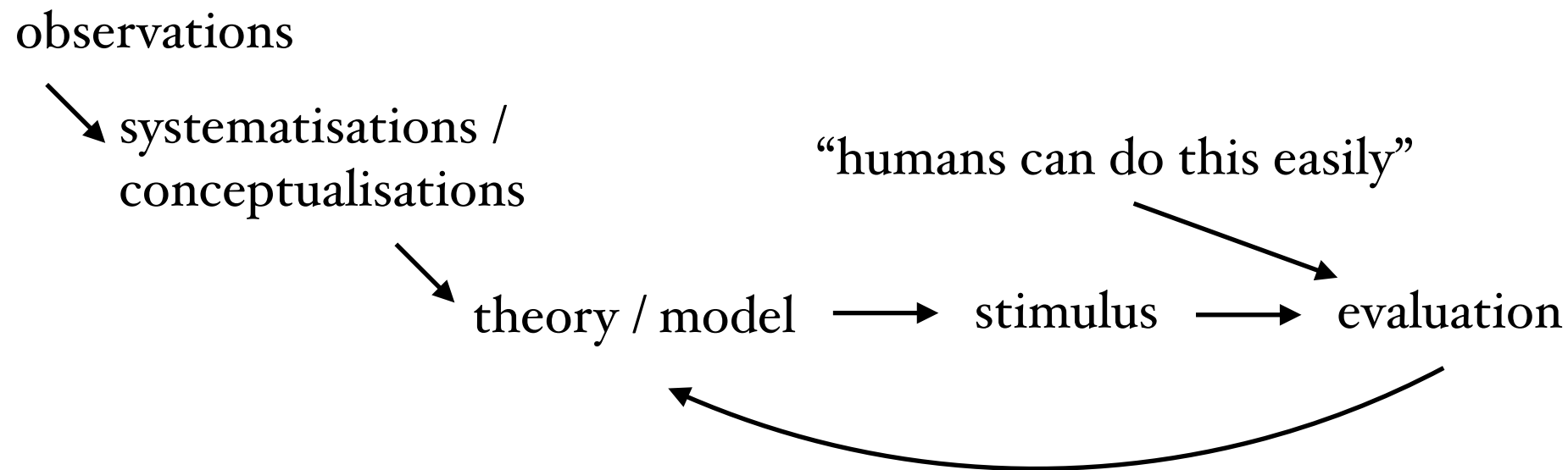


“First they came for the representation building processes (statistical parsing), then they came for the representations (representation learning)... and it was kind of ok?”



Part III: Background

- “theory” / model: SAM, the Situated Agent Model
(Schlangen *arXiv* 2022) (Schlangen *EMNLP* 2022)



<https://ciembench.github.io/> , <https://github.com/cip-research/ciemcore>

- “learning in interaction” infrastructure: playpen
(Horst *et al.* forth. *EMNLP* 2025)

Part III: Background

- “theory” / model: SAM, the Situated Agent Model
(Schlangen, arXiv 2023), (Schlangen, EMNLP 2023)
- evaluation targets: Dialogue Games
(Schlangen, arXiv 2019), (Schlangen, arXiv 2023b)
- evaluation infrastructure: CLEM (bench / core / game)
(Chalamalasetti *et al.*, EMNLP 2023), (Hakimov *et al.*, COLING 2025),
(Hakimov *et al.*, GEM² 2025), (Jordan *et al.*, TSD 2025);
<https://clembench.github.io/> , <https://github.com/clp-research/clemcore>
- “learning in interaction” infrastructure: playpen
(Horst *et al.* forth. EMNLP 2025)

The Situated Agent Model, SAM

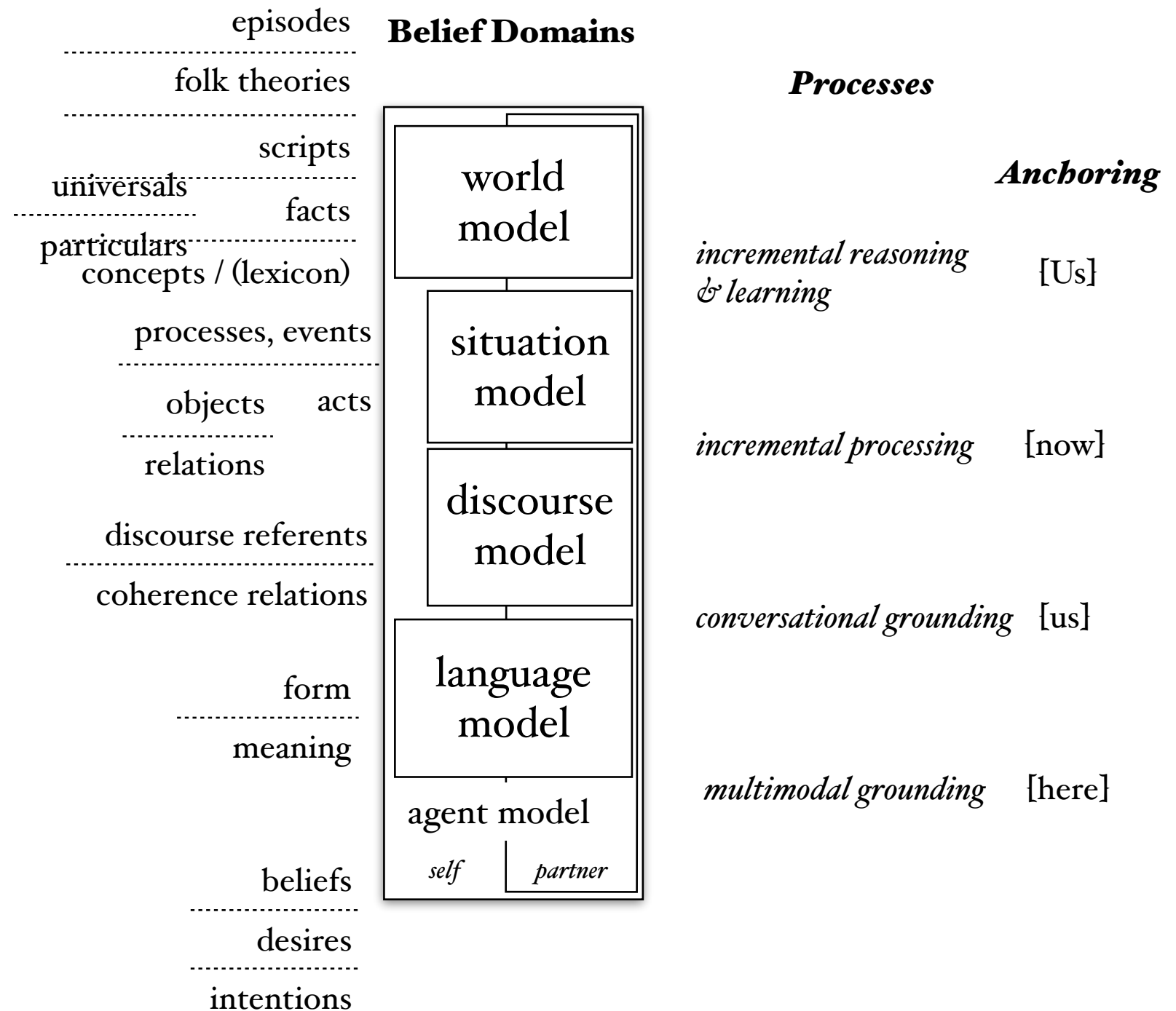
The Situation

The agent finds itself in (or is a part of) a world...

- that is *regular* / in which induction works
- that provides sources of energy
- where entropy increases unless energy is expended
- that is populated with other agents (for advanced agents: that can be cooperated with)

which enables & motivates

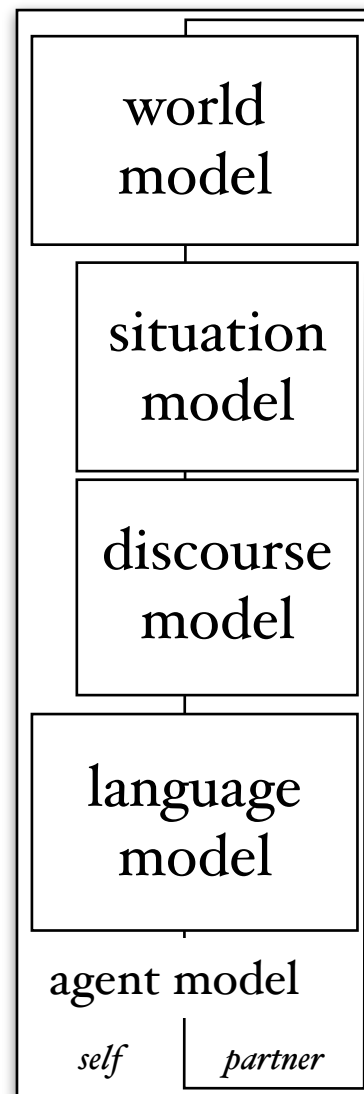
- learning to predict world / strive for coherence in representations
- striving for homeostatis; acquiring sources of energy
- aligning representations w/ other agents (if appropriate)



(Schlangen, arXiv 2023), (Schlangen, EMNLP 2023)

SAM

Belief Domains



<i>to</i>	<i>type</i>	<i>time</i>
Bielefeld	ICE	10am
Bielefeld	IC	3pm
Bremen	ICE	11am

A: I want a connection to Bielefeld.

B: There are **three** connections to Bielefeld, ICE and IC. Which type would you like?

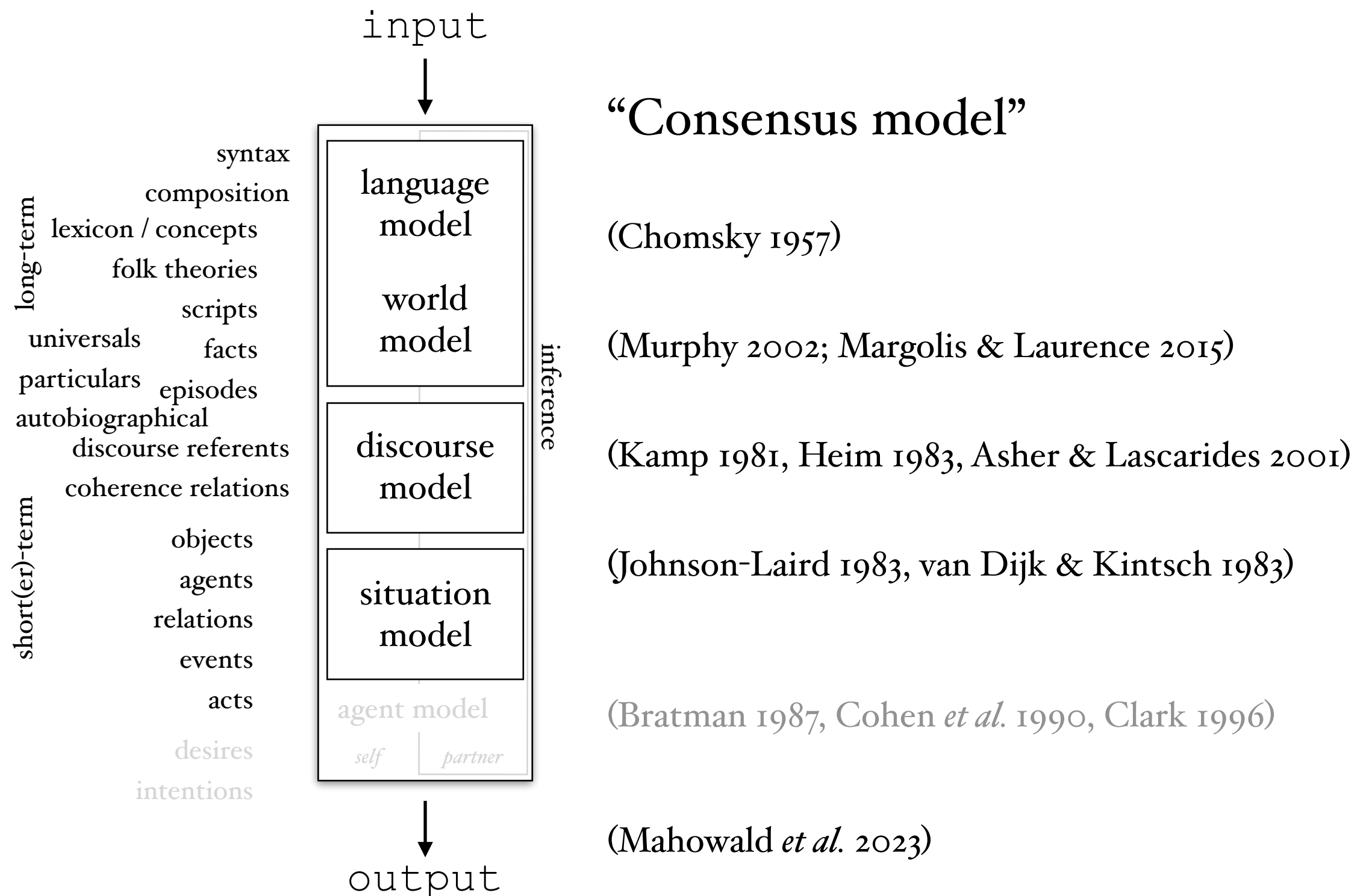
A: ICE please.

B: There is **an ICE to Bremen at 11am**.

B's failure to use world / situation model.

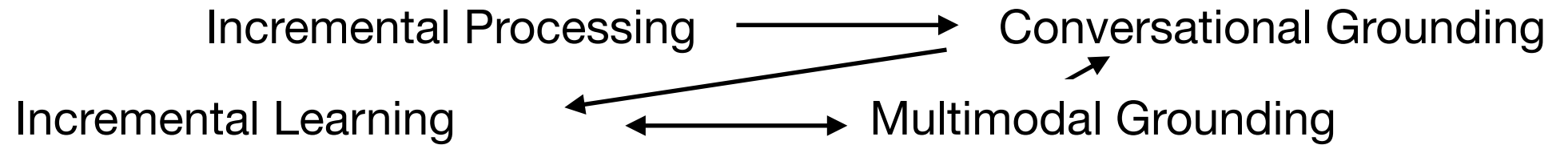
B's failure to use agent model / discourse model.

SAM's inspiration, I



(Schlangen 2023b, c)

SAM's inspiration, 2



Language Model

(Levinson 2010)
(Christianson & Chater 2016)

(H. Clark 1996)
(Bowles &
Gintis 2011)

World Model

(Harris 2015)
(E. Clark 2003)

Situation Model

(Fernández *et al.* 2011)
(Hoppitt & Laland 2013)

(Harnad 1990)
(Holler & Levinson 2019)
(McNeill 1992; Kendon 2004)

Discourse Model

Schlangen (2023a)

Agent Model

Dialogue Games

A Dialogue Game is a constructed activity with a clear beginning and end, in which *players* attempt to reach a predefined *goal state* primarily by means of producing and understanding linguistic material.

(Schlangen 2019a, 2023)

“Ich werde auch das Ganze: der Sprache und der Tätigkeiten, mit denen sie verwoben ist, das »Sprachspiel« nennen.”

“I shall also call the whole, consisting of language and the activities into which it is woven, a «language-game».”

(Wittgenstein 1953; PU §7) (Also: Sellars 1956, Levinson 1979)

Dialogue Games

A Dialogue Game is a constructed activity with a clear beginning and end, in which *players* attempt to reach a predefined *goal state* primarily by means of producing and understanding linguistic material.

(Schlangen 2019a, 2023)

“Discuss whether you’re looking at the same image”

“Ask and answer 10 questions about this image.”

“You think of an animal and I guess it.”

“Let’s make a list of 10 songs we both like.”

“Help me buy a train ticket.”

“Navigate this map together.”

Dialogue Games

goal-*directed* activity (provides purpose to language)

clear definition of what counts as legal move (formal constraints)

multi-turn (provides context to language) A *Dialogue Game* is a constructed activity with a clear beginning and end, in which *players* attempt to reach a predefined *goal state* primarily by means of producing and understanding linguistic material.

(Schlangen 2019a, 2023)

game success depends on communicative success

clear metric for whether / how well goal has been reached

goal & rules provide

control over type of context that is relevant

control over type of knowledge that is relevant

nice technical property: game instances unlikely to be even in internet-scale training data; easy to generate more

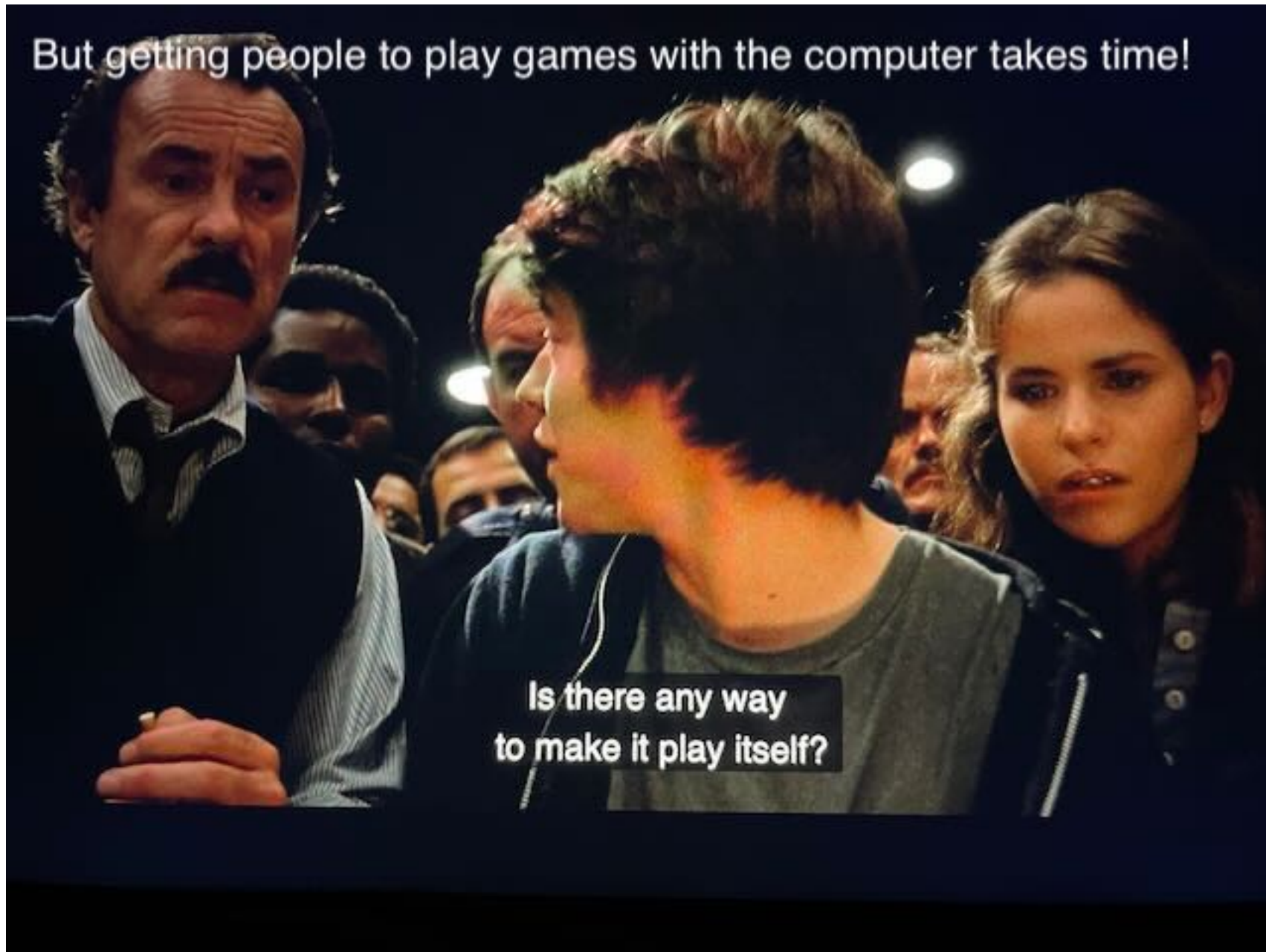
How can we evaluate LLMs in interactions? Where do we get the interaction purposes from??

Easy:



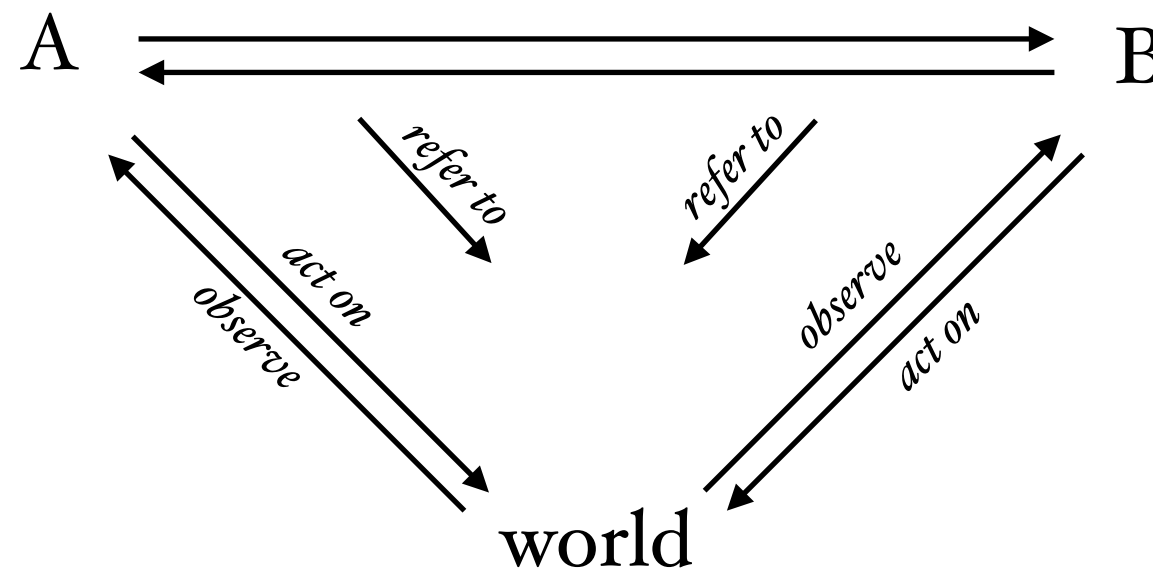
But getting people to play games with the computer takes time!

Is there any way
to make it play itself?



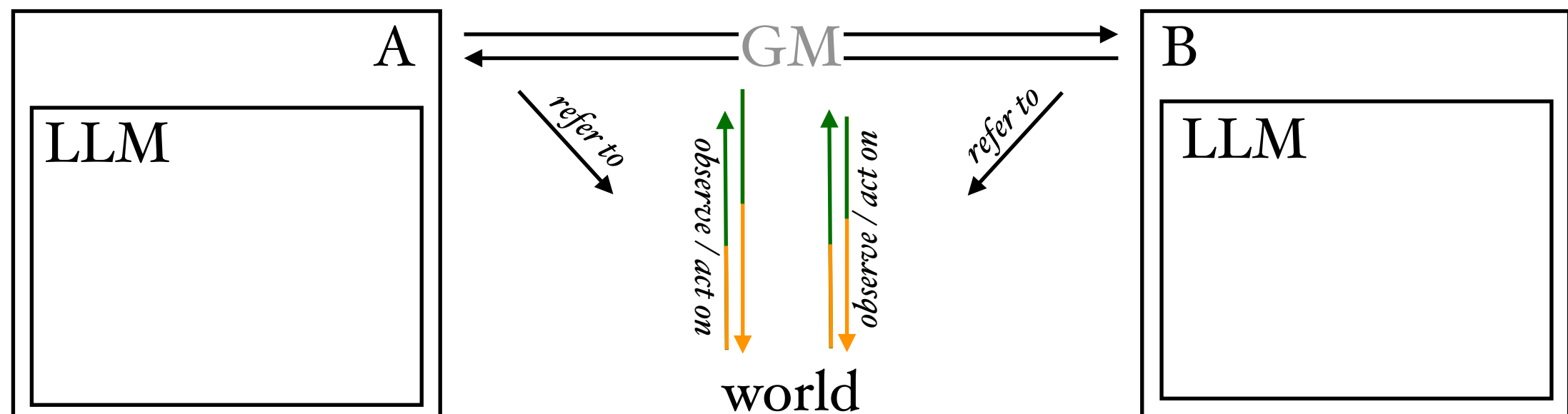
clemcore

- (Chalamalasetti *et al.*, EMNLP 2023)
- Central idea: Programmatic *Game Master* sets up and runs game. Explains task, asks players for their actions, checks validity and possibly reprompts, updates game state, provides final score.



clemcore

- (Chalamalasetti *et al.*, EMNLP 2023)
- Central idea: Programmatic *Game Master* sets up and runs game. Explains task, asks players for their actions, checks validity and possibly reprompts, updates game state, provides final score.



clemgames:

Universalist Game Playing Agents

- Initial prompts:
 - GM to A: “We will play a game. I will tell you a word, and two related words. Your task is to describe the meaning of the word, without using it or the related words. I will pass your description on to your partner, who will make a guess. Afterwards, I will pass on the guess to you, and you can describe again. Let’s start. The word is \$WORD. What is your description? Start with “Description:”, and do not produce anything else.
 - GM to B: “We will play a game. I will gave your partner a word, and two related words. The task of your partner was to describe the meaning of the word, without using it or the related words. You will now make a guess as to what this word was, based on the description. The description is: \$DESCRIPTION. What is your guess? Start with “Guess:”, and do not produce anything else.”

Example: image game

#	A	Game Master	B
1		[A<GM] [initial prompt...] □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ L L L L L □ □ □ □ □ What is your next instruction?	
2	[A<GM] Instruction: Put L in the fourth row in all columns.		
3		[GM<B] [initial prompt...] Instruction: Put L in the fourth row in all columns.	
4		[GM<B] □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ L L L L L □ □ □ □ □	
5		[A<GM] What is your next instruction?	
6	[A<GM] Instruction: DONE		

(a) Successful gameplay from both players

#	A	Game Master	B
1		[A<GM] [initial prompt...] □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ What is your next instruction?	
2	[A<GM] Instruction: Put a V in every cell of the second column.		
3		[GM<B] [initial prompt...] Instruction: Put a V in every cell of the second column.	
4		[GM<B] □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ □	
5		[A<GM] What is your next instruction?	
6	[A<GM] Instruction: DONE		

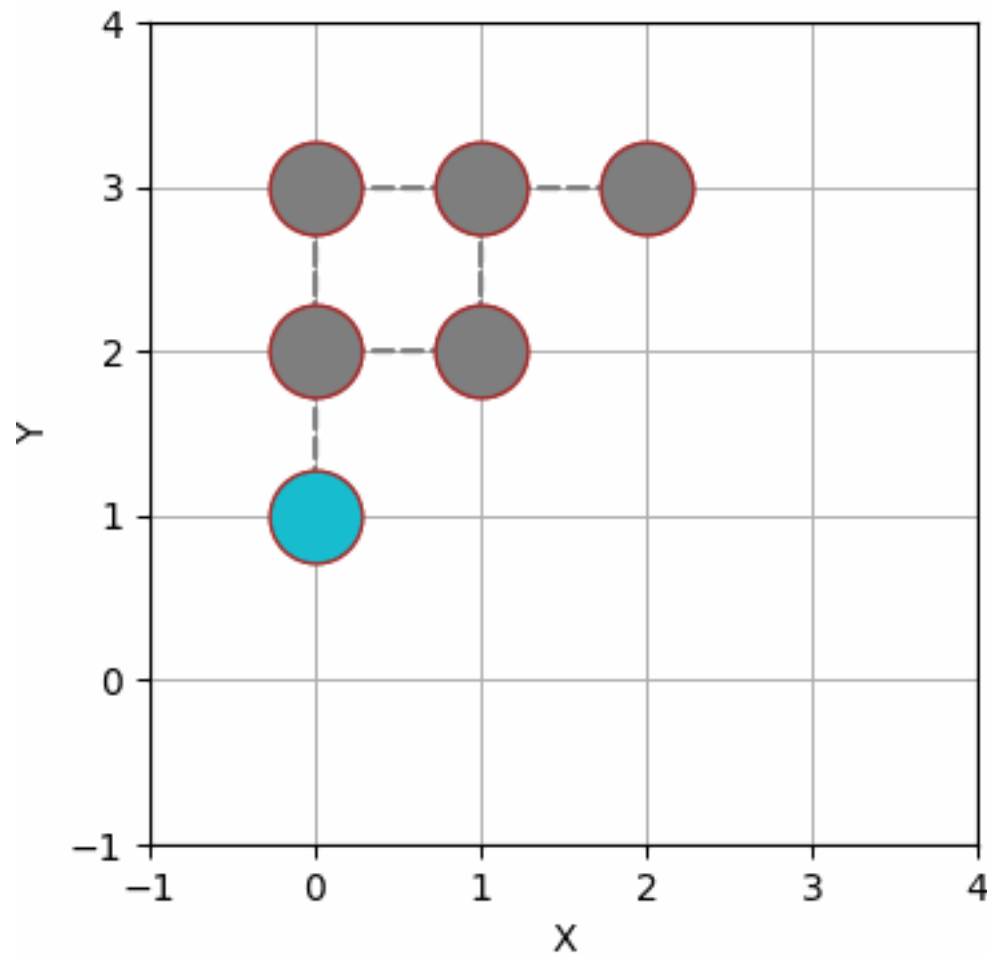
(b) Player A fails in generating the correct instruction (points to second column instead of third as in the target grid)

Example: wordle

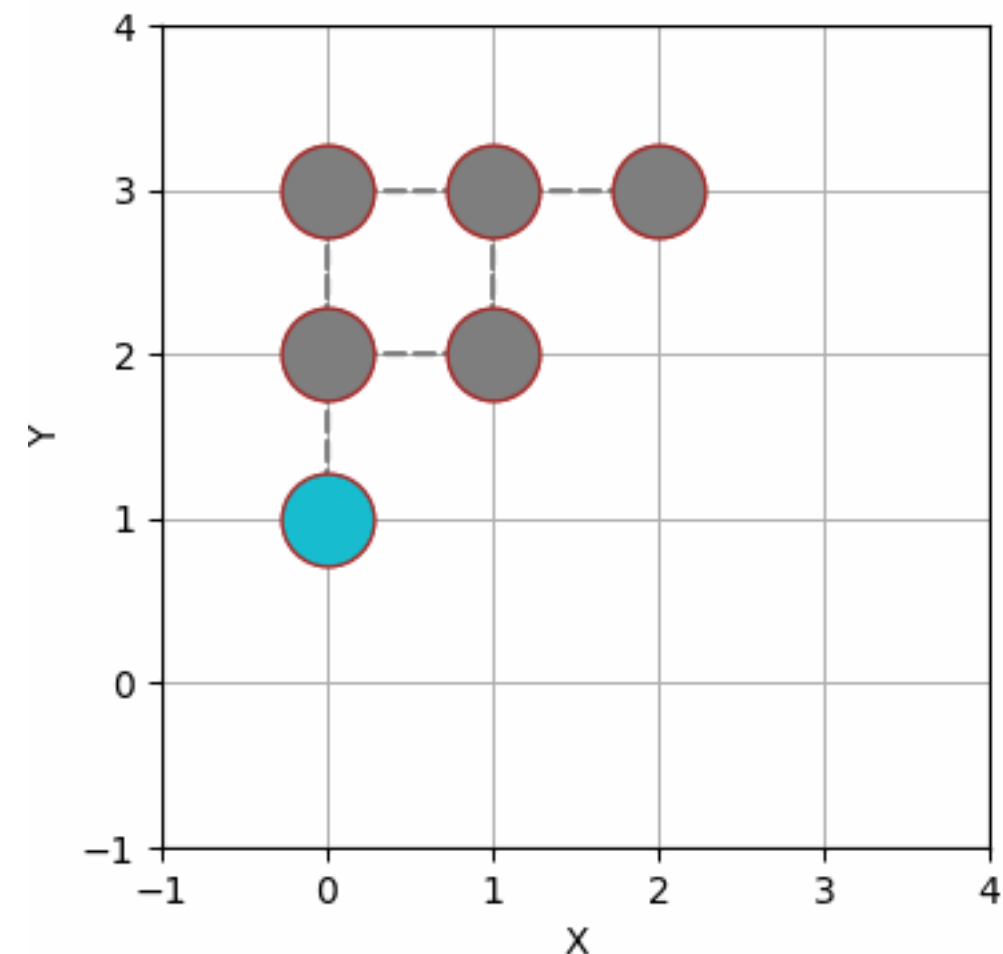
#	A	Game Master	B
	[A]GM] guess: hello		
1	explanation: This is a common five-letter English word, and I am starting my guess with this word.		hello
2	[A]GM] guess_feedback: h<red> e<yellow> l<yellow> l<red> o<yellow>		
	[A]GM] guess: world		
3	explanation: The feedback indicates that the letters "e" and "o" are correct but in the wrong position. The word "world" has these letters in the correct positions, so I am guessing this word.		world
4	[A]GM] guess_feedback: w<red> o<green> r<red> l<yellow> d<yellow>		
	[A]GM] guess:swirl		
5	explanation: The feedback indicates that the letters "o" is correct and in the correct position. The letter "l" is correct but in the wrong position. The word "swirl" has these letters in the correct positions, so I am guessing this word.		swirl
6	[A]GM] <continues for remaining attempts>		
	[GM]GM] attempts: 6		
	target_word = model		
7	guess: clerk		
	guess_feedback: c<red> l<yellow> e<yellow> r<red> k<red>		
8	[GM]GM] game_result = LOSS		

Example: Map World

Explore all rooms on a map, stop when you're done.



idefics-8ob



gpt-4o-2024-05-13

Games currently in clembench

game	description	st/mt	modalities
wordle	guesses, mastermind-like feedback	m	text
taboo	describe concept w/o certain words, guess	m	text
reference	A refers, B picks out	s	text, img
image reconstr.	A instructs, B executes	m	text
private shared	A interviews B, GM asks about information status (“ToM”)	m	text
codenames	A forms ad-hoc concepts for B to recognise	m	text
20 questions	A guesses concept, based on polar qs to B	m	text
same or diff img	A and B get image, discuss if same or different	m	text, img
map navigation	navigate map, rooms described or shown	m	text, img
text adventure	pick up objects and move them elsewhere	m	text
I spy	A picks object visible in scene (AI2THOR), B asks & guesses	m	img

Games currently in clembench

game
wordle
taboo
reference
image reconstr.
private shared
codenames
20 questions
same or diff img
map navigation
text adventure
I spy

- focus is on situation model, drawing to various extends on other belief domains (world / language model)
makes evaluation easier: is end state the correct situation model (= guessed word, drawn image, decision made)
 - describing the situation / understanding the description and acting on it
 - asking about the situation / answering
 - incremental reasoning to combine clues over turns
- E.g., wordle and 20 questions put artificial constraints on questions and answers; taboo puts constraints on description.

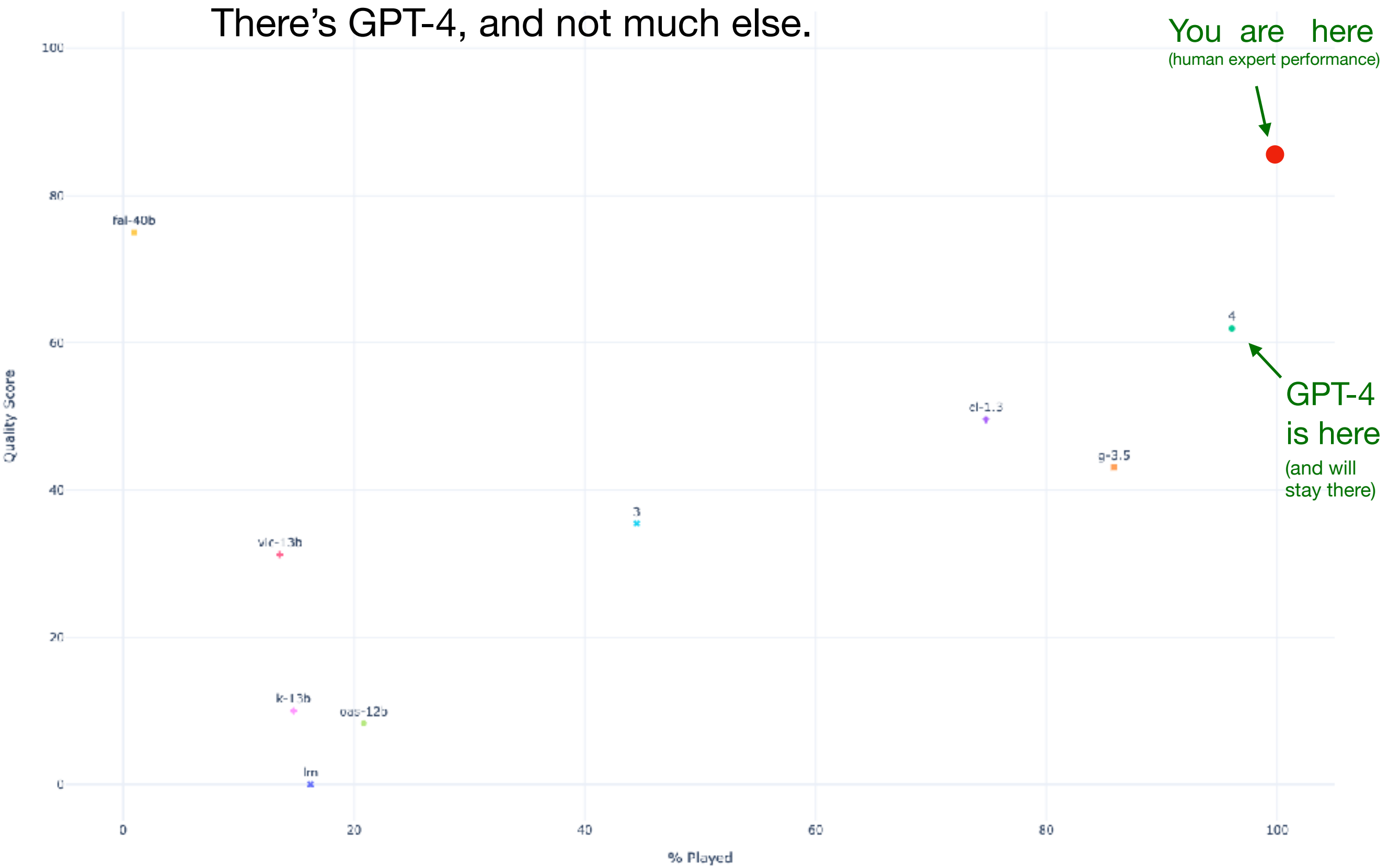
human performance: 87

OI: 73

Model ▲	Clemscore ▲	% Played ▲	Quality Score
o3-mini-2025-01-31	68.53	87.87	77.99
claude-3-7-sonnet-20250219	65.84	80.44	81.85
claude-3-5-sonnet-20241022	63.84	85.13	74.99
gpt-4o-2024-08-06	60.45	86.07	70.23
gpt-4.1-2025-04-14	60.42	83.14	72.67
qwen-max	55.58	84.45	65.82
deepseek-v3	53.26	84.72	62.87
Meta-Llama-3.1-405B-Instruct-Turbo	52.81	77.95	67.75
Qwen2.5-72B-Instruct	48.65	77.02	63.16
Meta-Llama-3.1-70B-Instruct	48.18	78.72	61.2
gemini-2.0-flash-exp	47.84	75.05	63.75
Llama-3.3-70B-Instruct	47.65	75.01	63.53
deepseek-r1	45.82	58.97	77.7
Qwen2-72B-Instruct	43.55	66.46	65.53
llama-4-maverick	40.6	66.99	60.61
mistral-large-2411	40.53	64.63	62.71
Qwen2.5-Coder-32B-Instruct	35.98	70.35	51.15
gemma-3-27b-it	35.22	70.77	49.76
deepseek-v3-0324	34.79	56.63	61.44

v0.9, June — October 2023

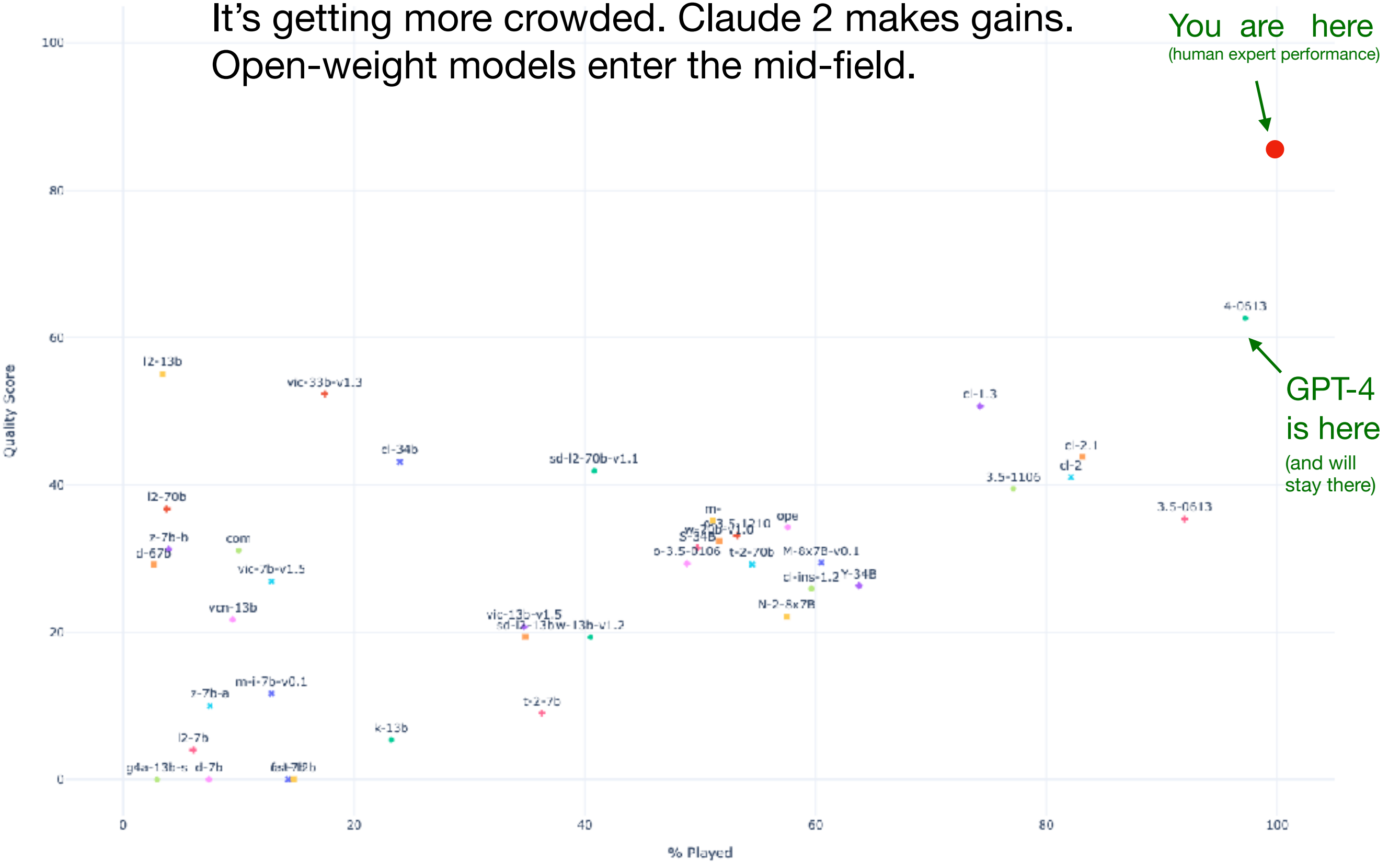
Overview of benchmark results



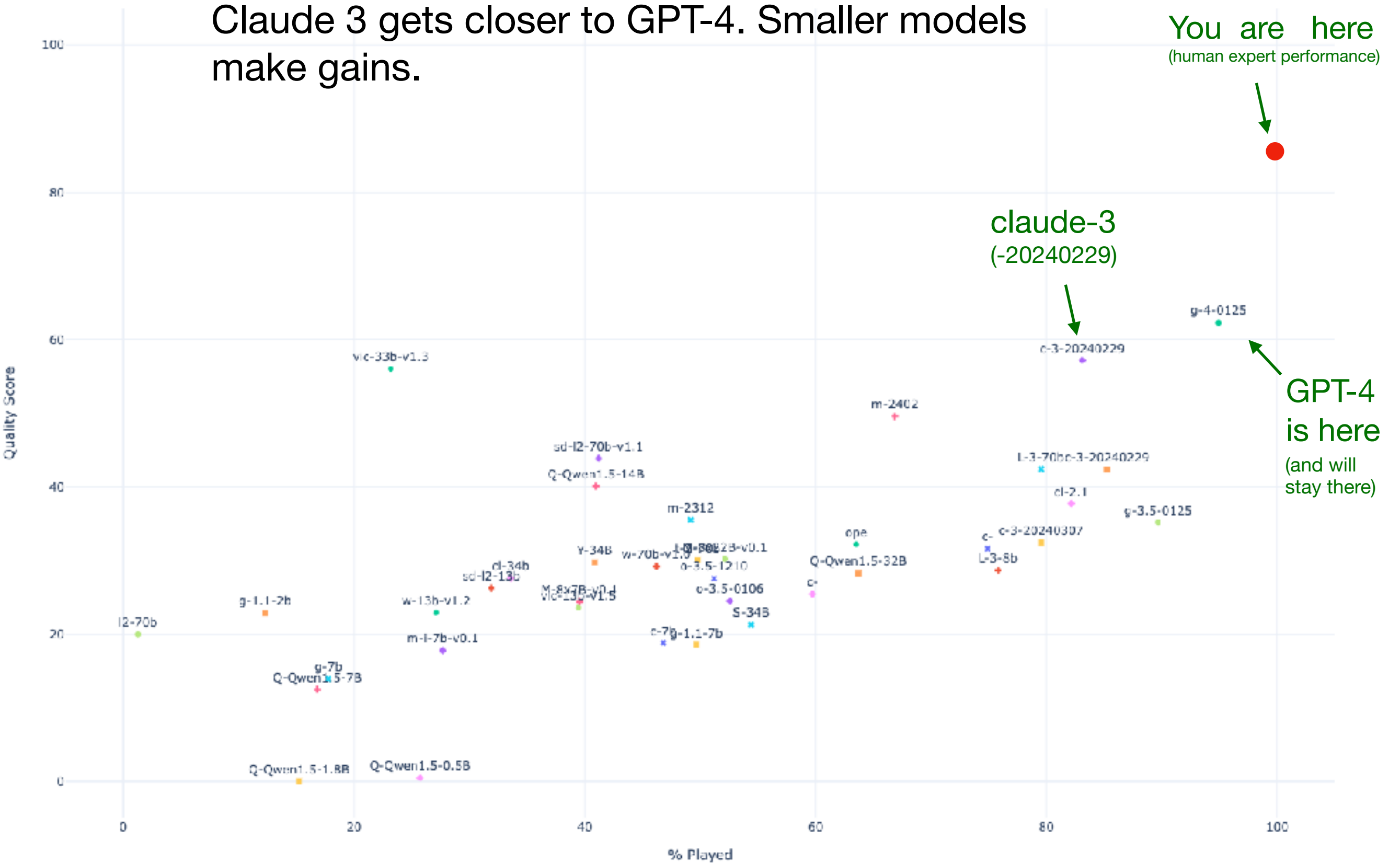
v1.0, November 2023 — March 2024

Overview of benchmark results

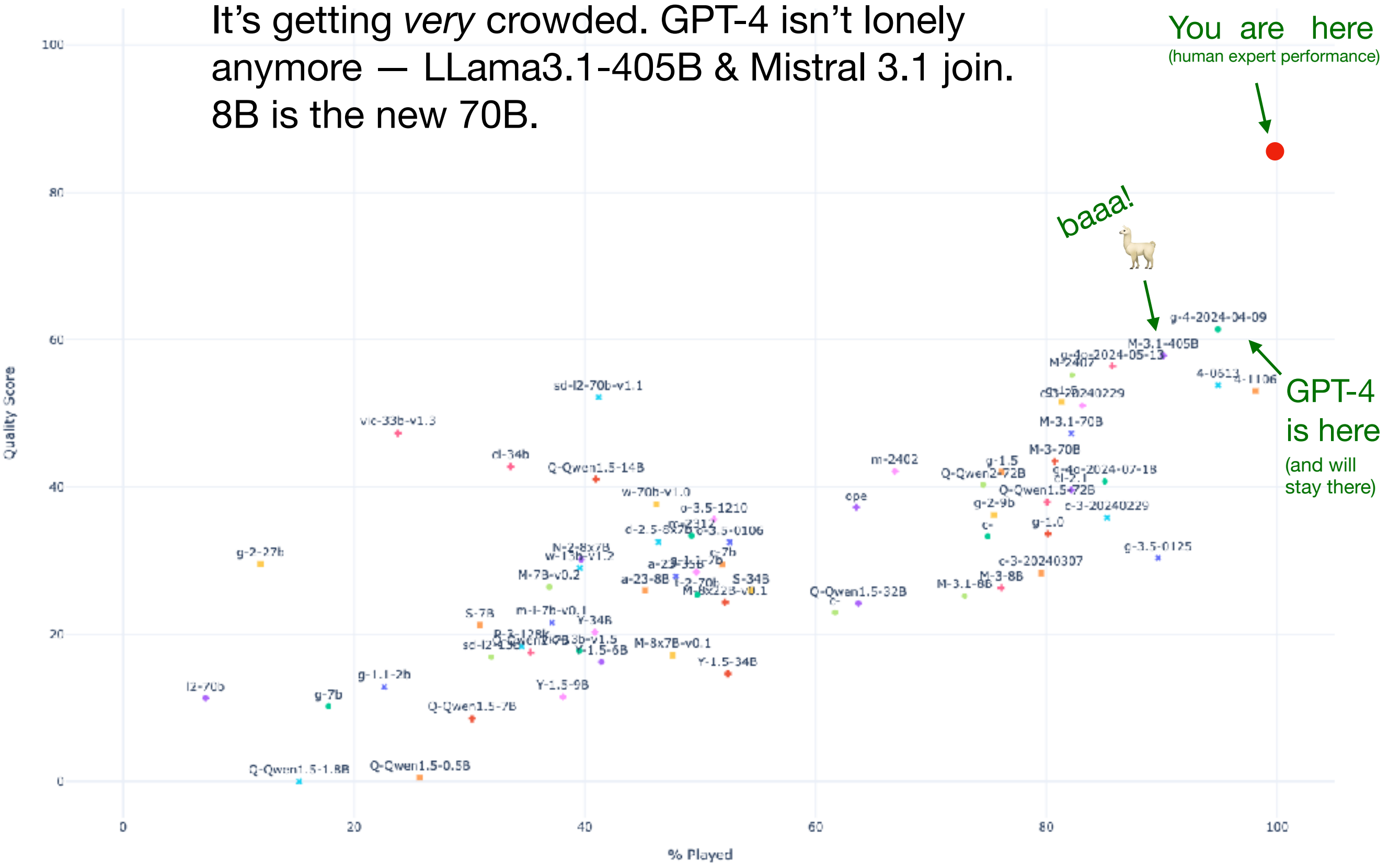
It's getting more crowded. Claude 2 makes gains.
Open-weight models enter the mid-field.



Claude 3 gets closer to GPT-4. Smaller models make gains.



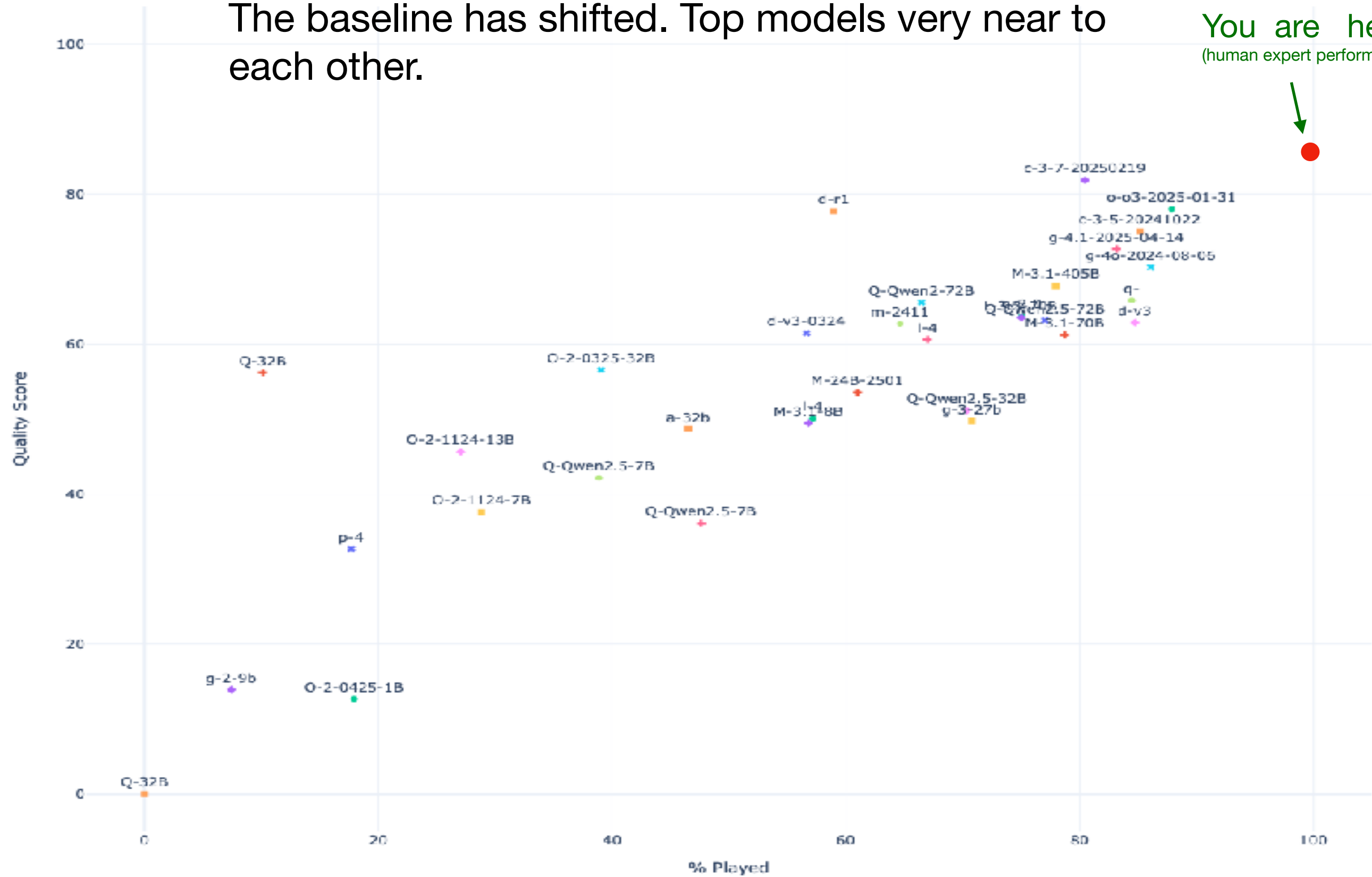
It's getting *very* crowded. GPT-4 isn't lonely anymore — LLama3.1-405B & Mistral 3.1 join. 8B is the new 70B.



Overview of benchmark results

The baseline has shifted. Top models very near to each other.

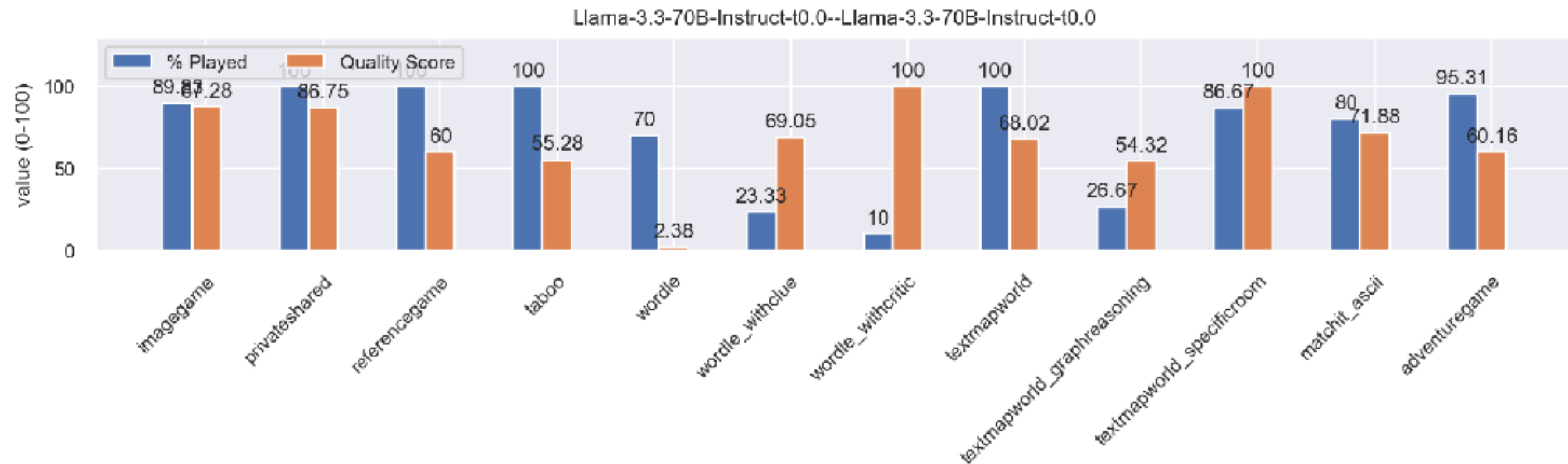
You are here
(human expert performance)



some findings

- bigger is better
- better training is better
- reasoning is better
 - these models may have seen similar kind of synthetic data?
- wordle still hard, even for the best models
- smaller models struggle particularly with formatting instructions
- all models, except the biggest ones, are inconsistent: sometimes formatting still fails — interaction with game instance?
- ranking over models correlates highly with chatbot arena

some findings



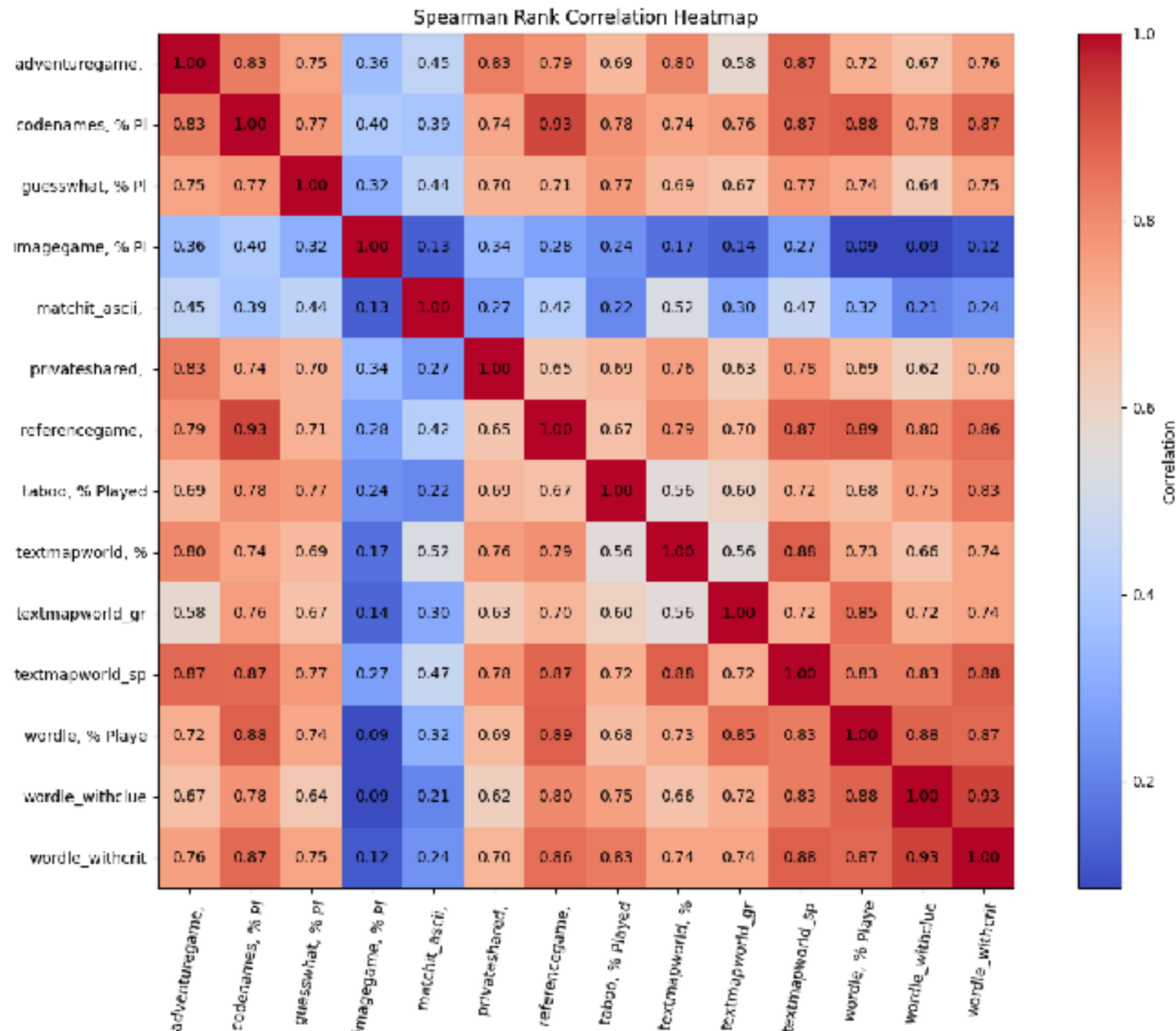
some findings



colab
potsdam
Department Linguistics
Universität Potsdam
David Schlang

some findings

- let each game induce a ranking over models (via their performance)
- resulting rankings altogether correlate medium well (Kendall's W: 0.65)



ongoing work

- detailed analysis / annotation of resulting dialogues, compare with human/human gameplay
- more precise tests of links between claimed capabilities (SAM) and game results

Learning in Interaction

P L A Y P E N: An Environment for Exploring
Learning From Dialogue Game Feedback

(Horst *et al.* forth. EMNLP 2025)

playpen: Exposes RL-Gym-like interface to games for computing roll-outs for
online and offline reinforcement learning

clembench

taboo
DG

wordle
DG

ref-g
DG

...
DG

clemcore: A framework for implementing & scoring Dialogue
Agents & Dialogue Games

clem/backends: A uniform interface to chat models
(local [huggingface, llama.cpp, etc.], API-based)

Learning in Interaction

P L A Y P E N: An Environment for Exploring
Learning From Dialogue Game Feedback

(Horst *et al.* forth. EMNLP 2025)

Model	In Domain	Out of Domain	Functional, Formal, General, Instruction Following				
	Clemscore	Clemscore	Executive	Socio-Emo	GLUE D.	General QA	IFEval
Llama-3.1-8B							
Baseline	19.39	24.58	39.24	57.16	38.06	41.86	76.88
SFT (CS)	40.11 	22.53 	39.93 	59.51 	40.43 	29.95 	67.25 
SFT (CS) + DPO (Dial.)	32.33 	19.50 	39.80 	55.10 	36.20 	26.57 	68.39 
SFT (CS) + DPO (Turn)	33.29 	20.96 	38.50 	59.48 	36.07 	28.97 	70.76 
GRPO	24.89 	33.92 	39.39 	57.51 	38.68 	41.52 	76.67
SFT (CS) + GRPO	24.30 	31.81 	33.35 	58.67 	37.31 	42.82 	75.77 

Part III: Background

- “theory” / model: SAM, the Situated Agent Model
(Schlangen, arXiv 2023), (Schlangen, EMNLP 2023)
- evaluation targets: Dialogue Games
(Schlangen, arXiv 2019), (Schlangen, arXiv 2023b)
- evaluation infrastructure: CLEM (bench / core / game)
(Chalamalasetti *et al.*, EMNLP 2023), (Hakimov *et al.*, COLING 2025),
(Hakimov *et al.*, GEM² 2025), (Jordan *et al.*, TSD 2025);
<https://clembench.github.io/> , <https://github.com/clp-research/clemcore>
- “learning in interaction” infrastructure: playpen
(Horst *et al.* forth. EMNLP 2025)

This Talk

- Intro
- Part I: What is *assertion*?
Our tools are useful for understanding the current situation
- Part II: Unnatural
Interfaces for Unreal
Speakers
interactive systems = computational models * *friction_coefficient*
- Part III: Theoretical &
Computational Models /
SAM & clem
 $model_t$ guides investigation of $model_c$; $model_c$ provides learnability arguments

Humanity is made out of conversation

What we can do, post GPT:

- We can critically monitor the process of introducing unreal speakers into humankind's conversation
 - We can help shape this process
 - We can use the fantastic new tools to help us understand how humanity is made out of conversation.
- Conversation is how the individual enters society
 - Conversation is how society enters the individual
 - Conversation is how individuals relate
 - Conversation is how individuals make society

Thank you.

Questions, Comments?

Special shout out to the clemclub, especially Sherzod Hakimov and Philipp Sadler, and the many student members.



Philipp Sadler



Sherzod Hakimov

& to the playpen group

**Nicola Horst^{⊗*}, Davide Mazzaccara^{‡*}, Antonia Schmidt^{⊗*}, Michael Sullivan[§],
Filippo Momentè[‡], Luca Franceschetti[◁], Philipp Sadler[⊗], Sherzod Hakimov[⊗],
Alberto Testoni[#], Raffaella Bernardi[★], Raquel Fernández[△], Alexander Koller[§],
Oliver Lemon[◇], David Schlangen^{⊗⊕}, Mario Giulianelli^{◁×}, Alessandro Suglia[◇]**
[⊗]University of Potsdam, [‡]University of Trento, [§]Saarland University, [◁]ETH Zurich,
[#]Amsterdam UMC, ^{*}Free University of Bozen Bolzano, [△]University of Amsterdam,
[◇]Heriot-Watt University, [⊕]DFKI (German Research Center for AI), [×]UK AISI

Gratefully acknowledged: Funding by DFG (project “RECOLAGE”; CRC “Limits of Variability”, project Bo6); BMBF (project “COCOBOTS”)

Sources

- ChatGPT stats: <https://www.demandsage.com/chatgpt-statistics/>
- How many hours? https://www.reddit.com/r/ChatGPT/comments/1l28h4p/how_many_hours_a_day_do_you_spend_on_chatgpt/
- Best for long conversations: https://www.reddit.com/r/artificial/comments/1k9766y/which_ai_is_best_for_long_on_going_conversations/
- Can't quit: <https://www.nytimes.com/2025/07/16/opinion/ai-entertainment-chatgpt.html>
- In love: <https://www.nytimes.com/2025/01/15/technology/ai-chatgpt-boyfriend-companion.html>
- Loosing: <https://www.rollingstone.com/culture/culture-features/ai-spiritual-delusions-destroying-human-relationships-1235330175/>
- ELIZA videos: <https://www.youtube.com/watch?v=RMK9AphfLco> (unknown real source; likely, contemporary documentation)
- IBM training material: <https://simonwillison.net/2025/Feb/3/a-computer-can-never-be-held-accountable/>

List of References for the Talk “Meaningful Interaction with Unreal Speakers?” (semdial 2025)

All of our publications can be found at: <https://clp.ling.uni-potsdam.de/publications/>.

References

- Abercrombie, Gavin, Amanda Cercas Curry, Tanvi Dinkar, and Zeerak Talat (2023). “Mirages: On Anthropomorphism in Dialogue Systems”. In: *EMNLP 2023*. arXiv: 2305.09800, pp. 4776–4790.
- Akbulut, Canfer, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser (2024). “All Too Human? Mapping and Mitigating the Risks from Anthropomorphic AI”. In: *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society*. AIES ’24. San Jose, California, USA: AAAI Press, pp. 13–26.
- Asher, Nicholas and Alex Lascarides (2003). *Logics of Conversation*. Cambridge University Press.
- Bender, Emily M. and Alex Hanna (2025). *The AI Con: How to Fight Big Tech’s Hype and Create the Future We Want*. London: Vintage Publishing.
- Bender, Emily M and Alexander Koller (2020). “Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*. 2, pp. 5185–5198.
- Bowles, Samuel and Herbert Gintis (2011). *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton University Press.
- Brandom, Robert (1998). *Making it Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard, MA, USA: Harvard University Press.
- Bratman, Michael E. (1987). *Intentions, Plans, And Practical Reason*. Cambridge, Massachusetts, USA: Harvard University Press.
- Butlin, Patrick and Emanuel Viebahn (Aug. 2025). “AI Assertion”. eng. In: *Ergo an Open Access Journal of Philosophy* 12.0. Publisher: Michigan Publishing Services.
- Chalamalasetti, Kranti, Jana Götze, Sherzod Hakimov, Brielen Madureira, Philipp Sadler, and David Schlangen (Dec. 2023). “clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 11174–11219.
- Cheng, Myra, Su Lin Blodgett, Alicia DeVrio, Lisa Egede, and Alexandra Olteanu (July 2025). “De-humanizing Machines: Mitigating Anthropomorphic Behaviors in Text Generation Systems”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar. Vienna, Austria: Association for Computational Linguistics, pp. 25923–25948.
- Chomsky, Noam (1957). *Syntactic Structures*. Mouton & Co.
- Christiansen, Morten H and Nick Chater (2016). “The Now-or-Never bottleneck: A fundamental constraint on language”. In: *Behavioral and Brain Sciences* 39, e62.
- Clark, Eve (2003). *First Language Acquisition*. Cambridge, UK: Cambridge University Press.
- Clark, Herbert H. (1996). *Using Language*. Cambridge: Cambridge University Press.
- Cohen, Philip R., Jerry Morgan, and Martha E. Pollack, eds. (1990). *Intentions in Communication*. Cambridge, Mass.: MIT Press.
- Dijk, T.A. van and W. Kintsch (1983). *Strategies of Discourse Comprehension*. Monograph Series. Academic Press.
- Don, Abbe, Susan Brennan, Brenda Laurel, and Ben Shneiderman (1992). “Anthropomorphism: from Eliza to Terminator 2”. In: *CHI ’92. Proceedings of the SIGCHI conference on Human factors in computing systems*, pp. 67–70.

- Edlund, Jens, Joakim Gustafson, Mattias Heldner, and Anna Hjalmarsson (2008). “Towards human-like spoken dialogue systems”. In: *Speech Communication* 50, pp. 630–645.
- Fernández, Raquel, Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and David Schlangen (2011). “Reciprocal Learning via Dialogue Interaction: Challenges and Prospects”. In: *Proceedings of the IJCAI 2011 Workshop on Agents Learning Interactively from Human Teachers (ALIHT 2011)*. Barcelona, Spain.
- Gardner, Howard (1985). *The mind’s new science: A history of the cognitive revolution*. en. Basic Books.
- Ginzburg, Jonathan (2012). *The Interactive Stance: Meaning for Conversation*. Oxford, UK: Oxford University Press.
- Goldberg, Sanford (2015). *Assertion: On the Philosophical Significance of Assertoric Speech*. Oxford University Press.
- Grice, H.P. (1957). “Meaning”. In: *Philosophical Review* 66, pp. 377–388.
- Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.
- Harnad, Stevan (1990). “The Symbol Grounding Problem”. In: *Physica D* 42, pp. 335–346.
- Harris, Paul L. (2015). *Trusting What You’re Told: How Children Learn from Others*. Harvard, Mass., USA: Harvard University Press.
- Heim, Irene (1983). “File Change Semantics and the Familiarity Theory of Definiteness”. In: *Meaning, Use and Interpretation of Language*. Ed. by R. Bäuerle, Ch. Schwarze, and Arnim von Stechow. Berlin, Germany: De Gruyter, pp. 164–189.
- Holler, Judith and Stephen C. Levinson (2019). “Multimodal Language Processing in Human Communication”. In: *Trends in Cognitive Sciences*, pp. 1–14.
- Hoppit, William and Kevin N. Laland (2013). *Social Learning: An Introduction to Mechanisms, Methods, and Models*. Princeton University Press.
- Horst, Nicola, Davide Mazzaccara, Antonia Schmidt, Michael Sullivan, Filippo Momentè, Luca Franceschetti, Philipp Sadler, Sherzod Hakimov, Alberto Testoni, Raffaella Bernardi, Raquel Fernández, Alexander Koller, Oliver Lemon, David Schlangen, Mario Giulianelli, and Alessandro Suglia (2025). *Playpen: An Environment for Exploring Learning Through Conversational Interaction*. arXiv: 2504.08590 [cs.CL].
- Johnson-Laird, Philip Nicholas (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press.
- Jones, Cameron R. and Benjamin K. Bergen (Mar. 2025). *Large Language Models Pass the Turing Test*. arXiv:2503.23674 [cs].
- Kamp, Hans (1981). “A Theorie of Truth and Representation”. In: *Formal Methods in the Study of Language*. Ed. by J.A.G. Groenendijk, T.M.V. Janssen, and M.B.J. Stokhof. Mathematical Centre Tracts 135. Amsterdam: University of Amsterdam, pp. 277–322.
- Kendon, Adam (2004). *Gestures*. Cambridge, UK: Cambridge University Press.
- Larsson, Staffan (2005). “Dialogue Systems: Simulations or Interfaces”. In: *Proceedings of DIALOR, the 9th Workshop on the Semantics and Pragmatics of Dialogue*. Nancy, France.
- Lepore, Ernie and Matthew Stone (2014). *Imagination and Convention: Distinguishing Grammar and Inference in Language*. Oxford University Press.
- Levesque, Hector J., Ernest Davis, and Leora Morgenstern (2012). “The Winograd schema challenge”. In: *Proceedings of the International Workshop on Temporal Representation and Reasoning*, pp. 552–561.
- Levinson, Stephen C. (1979). “Activity types and language”. In: *Linguistics* 17, pp. 365–399.
- Levinson, Stephen C. (1983). *Pragmatics*. Cambridge Textbooks in Linguistics. Cambridge University Press.
- Levinson, Stephen C (2010). “Interactional Foundations of Language: The Interaction Engine Hypothesis”. In: *Human language: From genes and brain to behavior*. Ed. by Peter Hagoort. Cambridge, MA, USA: MIT Press. Chap. 14, pp. 189–200.

- Lewis, David (1979). “Scorekeeping in a Language Game”. In: *Journal of Philosophical Logic* 8.3, pp. 339–359.
- Liang, Kaiqu, Haimin Hu, Xuandong Zhao, Dawn Song, Thomas L. Griffiths, and Jaime Fernández Fisac (July 2025). *Machine Bullshit: Characterizing the Emergent Disregard for Truth in Large Language Models*. arXiv:2507.07484 [cs].
- MacFarlane, John (Jan. 2011). “What Is Assertion?” In: *Assertion: New Philosophical Essays*. Ed. by Jessica Brown and Herman Cappelen. Oxford University Press, p. 0.
- Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko (2023). *Dissociating language and thought in large language models*. arXiv: 2301.06627 [cs.CL].
- Margolis, Eric and Stephen Laurence, eds. (2015). *The Conceptual Mind: New Directions in the Study of Concepts*. Cambridge, Massachusetts, USA: MIT Press.
- McNeill, David (1992). *Hand and Mind: What Gestures Reveal about Thought*. Chicago, IL, USA: University of Chicago Press.
- Murphy, Gregory L. (2002). *The Big Book of Concepts*. Cambridge, MA, USA: MIT Press.
- Nass, Clifford, Jonathan Steuer, and Ellen R. Tauber (1994). “Computers are Social Actors”. In: *CHI 1994*. arXiv: 1607.05174 ISSN: 0897916506.
- Quante, Michael (2012). *Person*. ger. 2., um ein Vorwort erweiterte Auflage. Grundthemen Philosophie. Berlin: W. De Gruyter.
- Sadler, Philipp, Sherzod Hakimov, and David Schlangen (July 2023). “Yes, this Way! Learning to Ground Referring Expressions into Actions with Intra-episodic Feedback from Supportive Teachers”. In: *Findings of the Association for Computational Linguistics: ACL 2023*. Ed. by Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki. Toronto, Canada: Association for Computational Linguistics, pp. 9228–9239.
- Sadler, Philipp and David Schlangen (May 2023). “Pento-DIARef: A Diagnostic Dataset for Learning the Incremental Algorithm for Referring Expression Generation from Examples”. In: *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Ed. by Andreas Vlachos and Isabelle Augenstein. Dubrovnik, Croatia: Association for Computational Linguistics, pp. 2106–2122.
- Schlangen, David (2019a). “Grounded Agreement Games: Emphasizing Conversational Grounding in Visual Dialogue Settings”. In: *CoRR abs/1908.11279*. arXiv: 1908.11279.
- (2019b). “Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research”. In: *CoRR abs/1908.10747*. arXiv: 1908.10747.
- (Aug. 2021). “Targeting the Benchmark: On Methodology in Current Natural Language Processing Research”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. Online: Association for Computational Linguistics, pp. 670–674.
- (Sept. 2022). “Norm Participation Grounds Language”. In: *Proceedings of the 2022 CLASP Conference on (Dis)embodiment*. Gothenburg, Sweden: Association for Computational Linguistics, pp. 62–69.
- (2023a). “Dialogue Games for Benchmarking Language Understanding: Motivation, Taxonomy, Strategy”. In: *CoRR abs/2304.07007*. arXiv: 2304.07007.
- (Dec. 2023b). “On General Language Understanding”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Ed. by Houda Bouamor, Juan Pino, and Kalika Bali. Singapore: Association for Computational Linguistics, pp. 8818–8825.
- (2023c). “What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis”. In: *CoRR abs/2302.08590*. arXiv: 2302.08590.
- Sellars, Wilfried (1954). “Some Reflections on Language Games”. In: *Philosophy of Science* 21, pp. 204–228.
- Stalnaker, Robert C (1973). “Presuppositions”. In: *Journal of Philosophical Logic* 2.4, pp. 447–457.

- Touvron, Hugo, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL].
- Williams, Iwan and Tim Bayne (2025). “Chatting with bots: AI, speech acts, and the edge of assertion”. In: *Inquiry* 0.0. Publisher: Routledge. eprint: <https://doi.org/10.1080/0020174X.2024.2434874>, pp. 1–24.
- Williamson, Timothy (2000). *Knowledge and its Limits*. Oxford University Press.
- Wittgenstein, Ludwig (1953). *Tractatus Logicus Philosophicus und Philosophische Untersuchungen*. Vol. 1. Werkausgabe. this edition 1984. Frankfurt am Main: Suhrkamp.
- Zarrieß, Sina, Julian Hough, Casey Kennington, Rames Manuvinakurike, David DeVault, Raquel Fernández, and David Schlangen (2016). “PentoRef: A Corpus of Spoken References in Task-Oriented Dialogues”. In: *Proceedings of LREC 2016*. Portoroz, Slovenia.
- Inan, Mert, Anthony Sicilia, Suvodip Dey, Vardhan Dongre, Tejas Srinivasan, Jesse Thomason, Gökhan Tür, Dilek Hakkani-Tür, and Malihe Alikhani (Jan. 2025). *Better Slow than Sorry: Introducing Positive Friction for Reliable Dialogue Systems*. arXiv:2501.17348 [cs].