Understanding Understanding

In General, and in Large Language Models

David Schlangen
Grundlagen der Computerlinguistik // Department Linguistik // Universität Potsdam
http://clp.ling.uni-potsdam.de
david.schlangen@uni-potsdam.de

@ ILCB Seminar
2024-02-09

These slides:
https://clp.ling.uni-potsdam.de/talks

This Talk

Understanding *Understanding* —-
the process by which language users create shared understanding (sufficient for their current purposes)

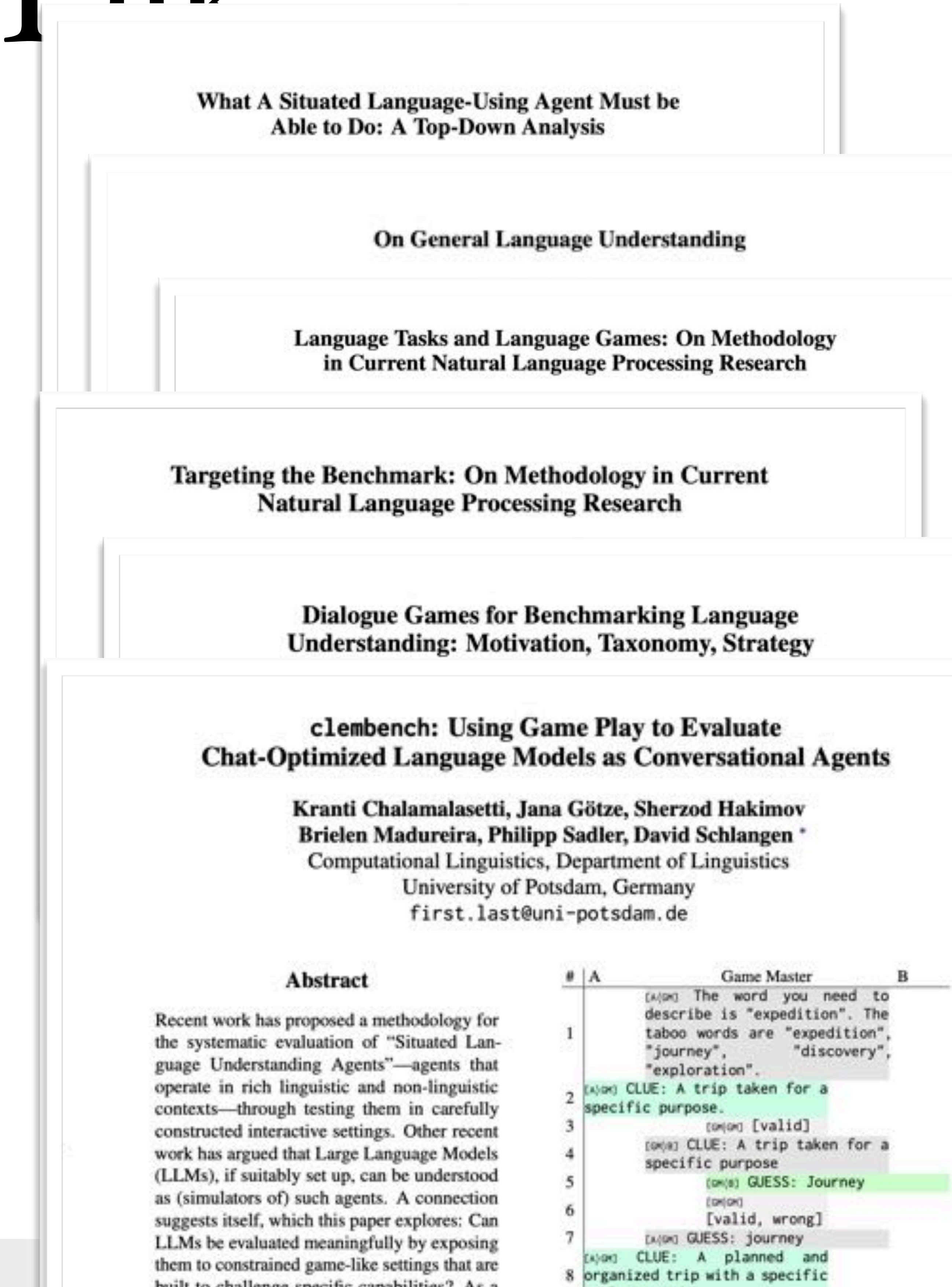
using *computational models*

for scientific insight

for practical gain

This Talk

- A Theoretical Thread
  - “What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis” (ArXiv, 2023)
  - “On General Language Understanding” (EMNLP Findings 2023)
- A Methodological Thread
  - Negative: How *not* To Do Things
    - “Language Tasks and Language Games: On Methodology in Current NLP Research” (ArXiv 2019)
    - “Targeting the Benchmark: On Methodology in Current NLP Research” (ACL 2021)
  - Positive: How To Do Things
    - “Dialogue Games for Benchmarking Language Understanding: Motivation, Taxonomy, Strategy” (ArXiv 2023)
    - “clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents” (Chalamalasetti *et al.*, EMNLP 2023)



What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis

On General Language Understanding

Language Tasks and Language Games: On Methodology in Current Natural Language Processing Research

Targeting the Benchmark: On Methodology in Current Natural Language Processing Research

Dialogue Games for Benchmarking Language Understanding: Motivation, Taxonomy, Strategy

`clembench`: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov
Brielen Madureira, Philipp Sadler, David Schlangen *
Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
`first.last@uni-potsdam.de`

Abstract

Recent work has proposed a methodology for the systematic evaluation of “Situated Language Understanding Agents”—agents that operate in rich linguistic and non-linguistic contexts—through testing them in carefully constructed interactive settings. Other recent work has argued that Large Language Models (LLMs), if suitably set up, can be understood as (simulators of) such agents. A connection suggests itself, which this paper explores: Can LLMs be evaluated meaningfully by exposing them to constrained game-like settings that are

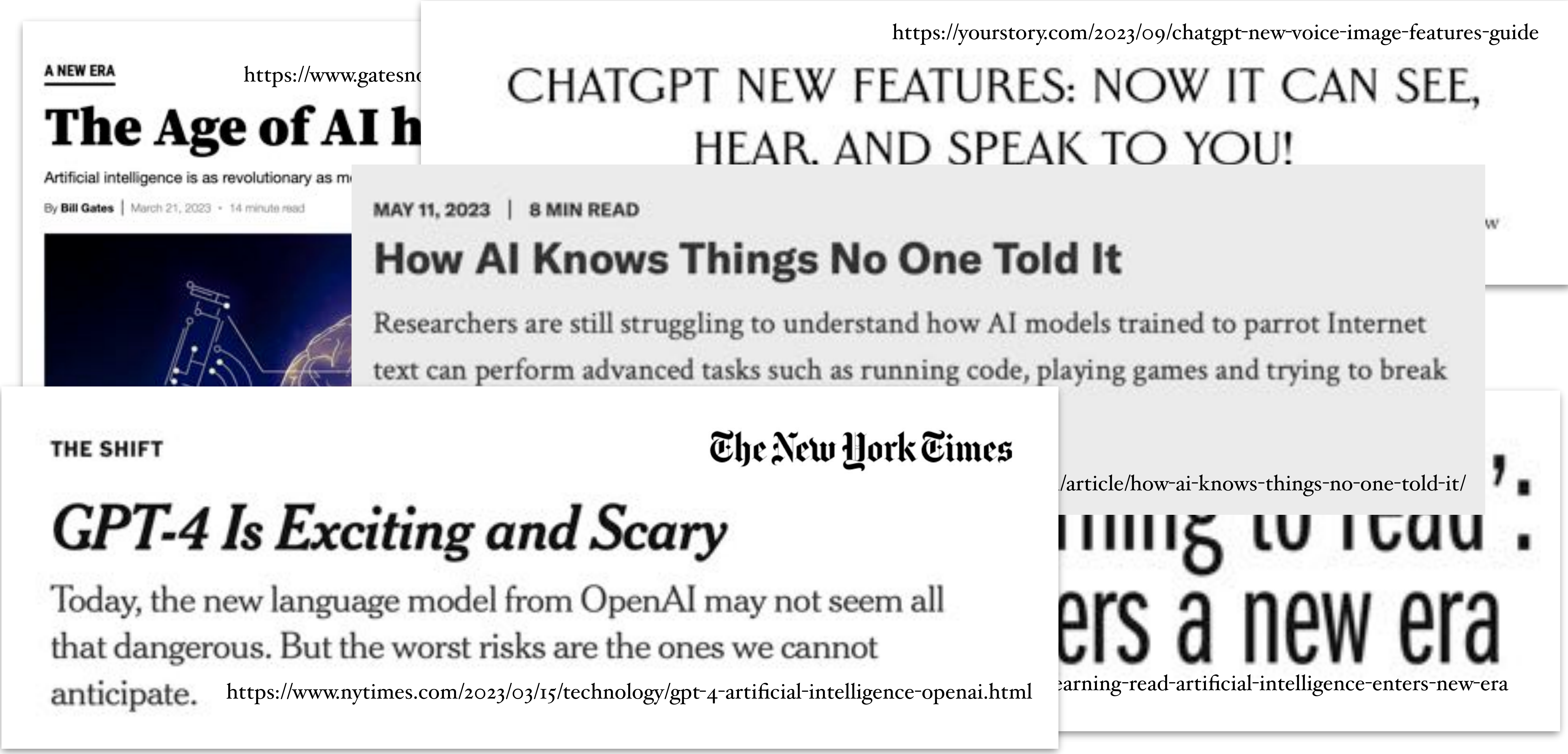
#	A	Game Master	B
1		[A|GM] The word you need to describe is "expedition". The taboo words are "expedition", "journey", "discovery", "exploration".	
2	[A|GM] CLUE: A trip taken for a specific purpose.		
3		[GM|GM] [valid]	
4		[GM|B] CLUE: A trip taken for a specific purpose	
5		[GM|B] GUESS: Journey	
6		[GM|GM] [valid, wrong]	
7		[A|GM] GUESS: journey	
8	[A|GM] CLUE: A planned and organized trip with a specific		

This Talk

- Part I: From NLU to *Situated Language Understanding*
- Part II: Studying SLU
  - Dialogue Games as a Tool
  - Evaluating cLLMs with Dialogue Games

Part I: From NLU to SLU

Isn’t Natural Language Understanding solved?



A NEW ERA

https://www.gatesnc

The Age of AI h

Artificial intelligence is as revolutionary as m

By **Bill Gates** | March 21, 2023 • 14 minute read

https://yourstory.com/2023/09/chatgpt-new-voice-image-features-guide

CHATGPT NEW FEATURES: NOW IT CAN SEE, HEAR, AND SPEAK TO YOU!

MAY 11, 2023 | 8 MIN READ

How AI Knows Things No One Told It

Researchers are still struggling to understand how AI models trained to parrot Internet text can perform advanced tasks such as running code, playing games and trying to break

/article/how-ai-knows-things-no-one-told-it/

ers a new era

earning-read-artificial-intelligence-enters-new-era

THE SHIFT

The New York Times

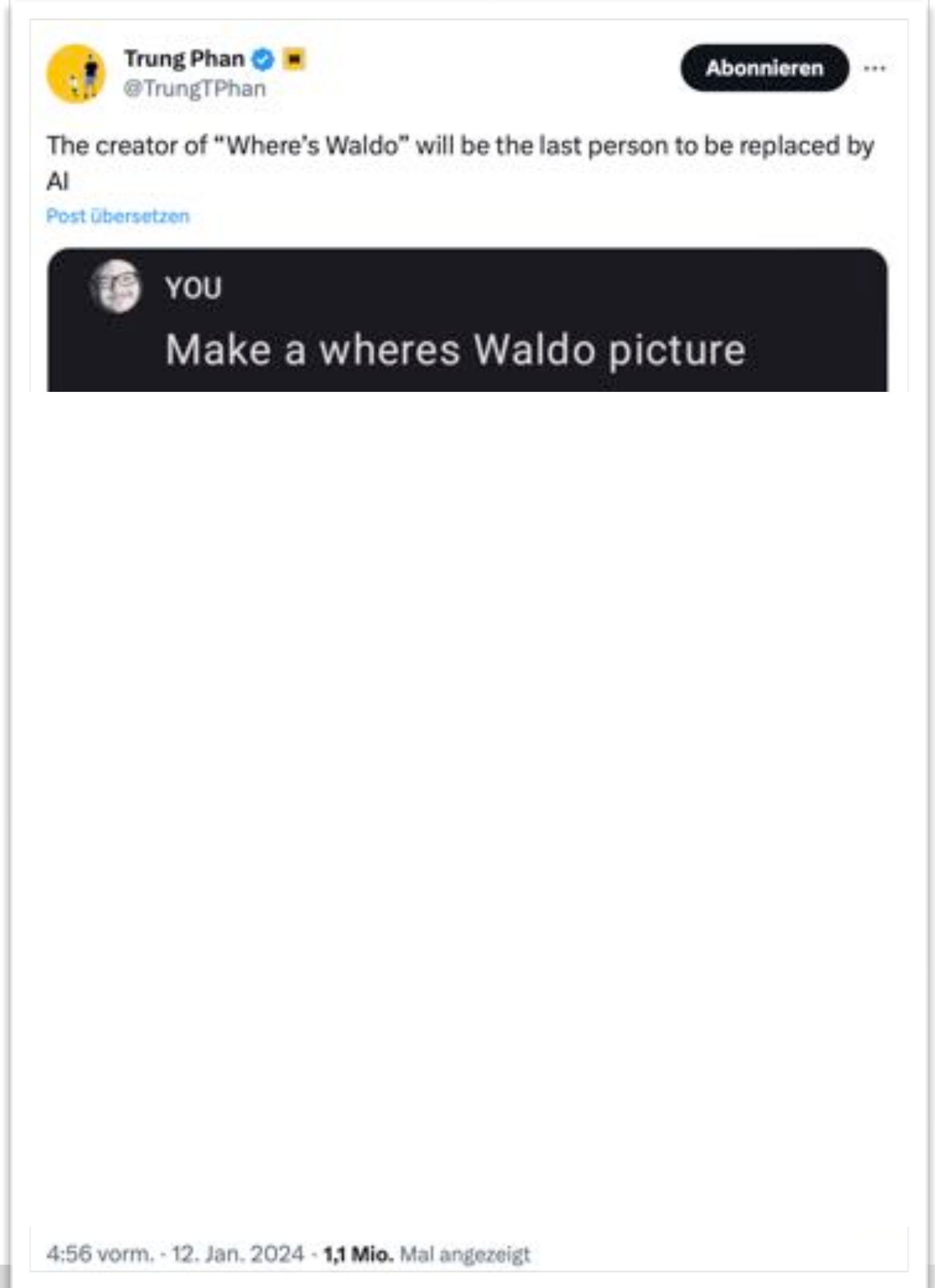
GPT-4 Is Exciting and Scary

Today, the new language model from OpenAI may not seem all that dangerous. But the worst risks are the ones we cannot anticipate.

https://www.nytimes.com/2023/03/15/technology/gpt-4-artificial-intelligence-openai.html

Part I: From NLU to SLU

Isn’t Natural Language Understanding solved?



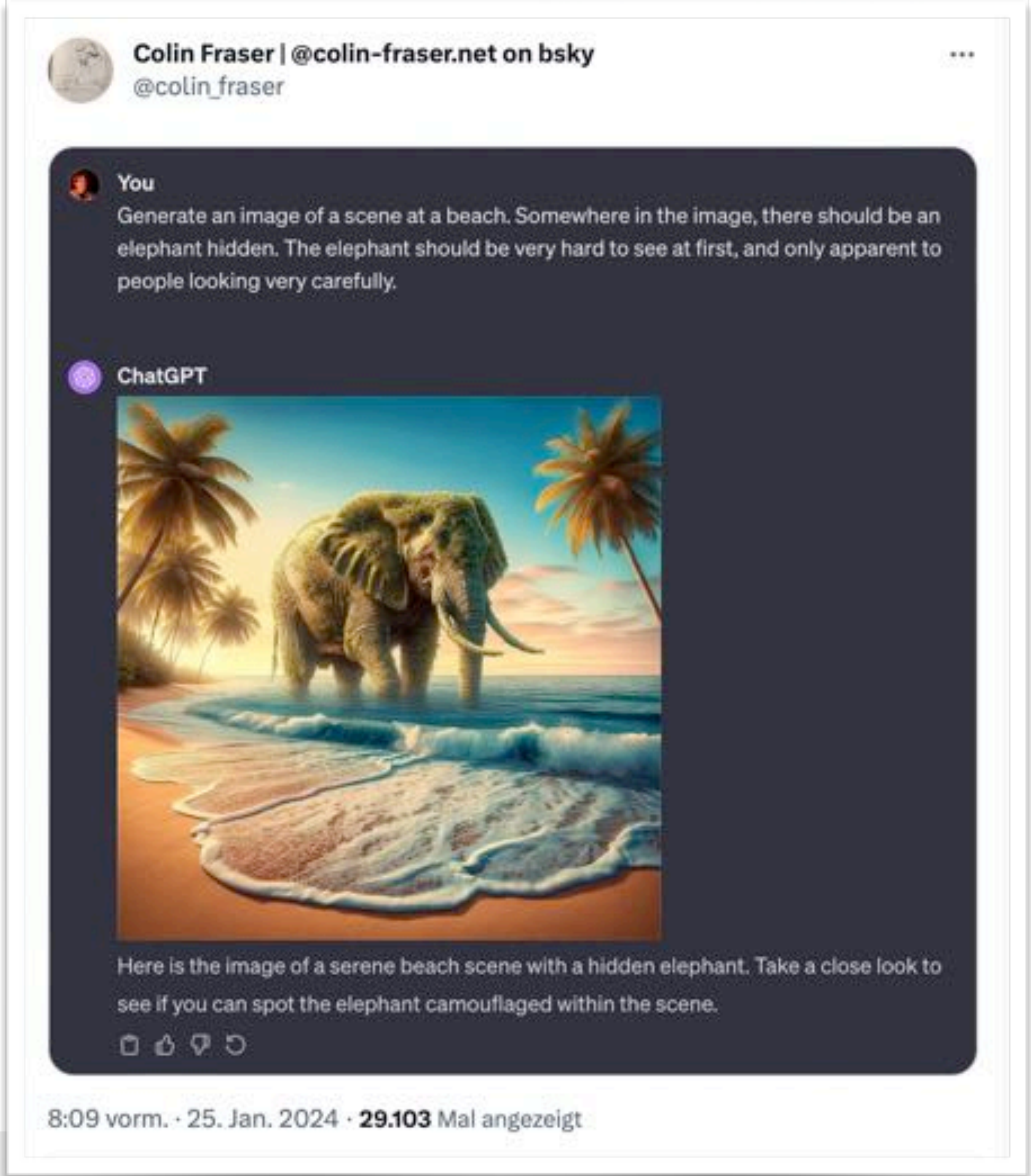
Trung Phan
@TrungTPhan

The creator of “Where’s Waldo” will be the last person to be replaced by AI

YOU

Make a wheres Waldo picture

4:56 vorm. · 12. Jan. 2024 · **1,1 Mio.** Mal angezeigt



Colin Fraser | @colin-fraser.net on bsky
@colin_fraser

You
Generate an image of a scene at a beach. Somewhere in the image, there should be an elephant hidden. The elephant should be very hard to see at first, and only apparent to people looking very carefully.

ChatGPT

Here is the image of a serene beach scene with a hidden elephant. Take a close look to see if you can spot the elephant camouflaged within the scene.

8:09 vorm. · 25. Jan. 2024 · **29.103** Mal angezeigt

Part I: From NLU to SLU

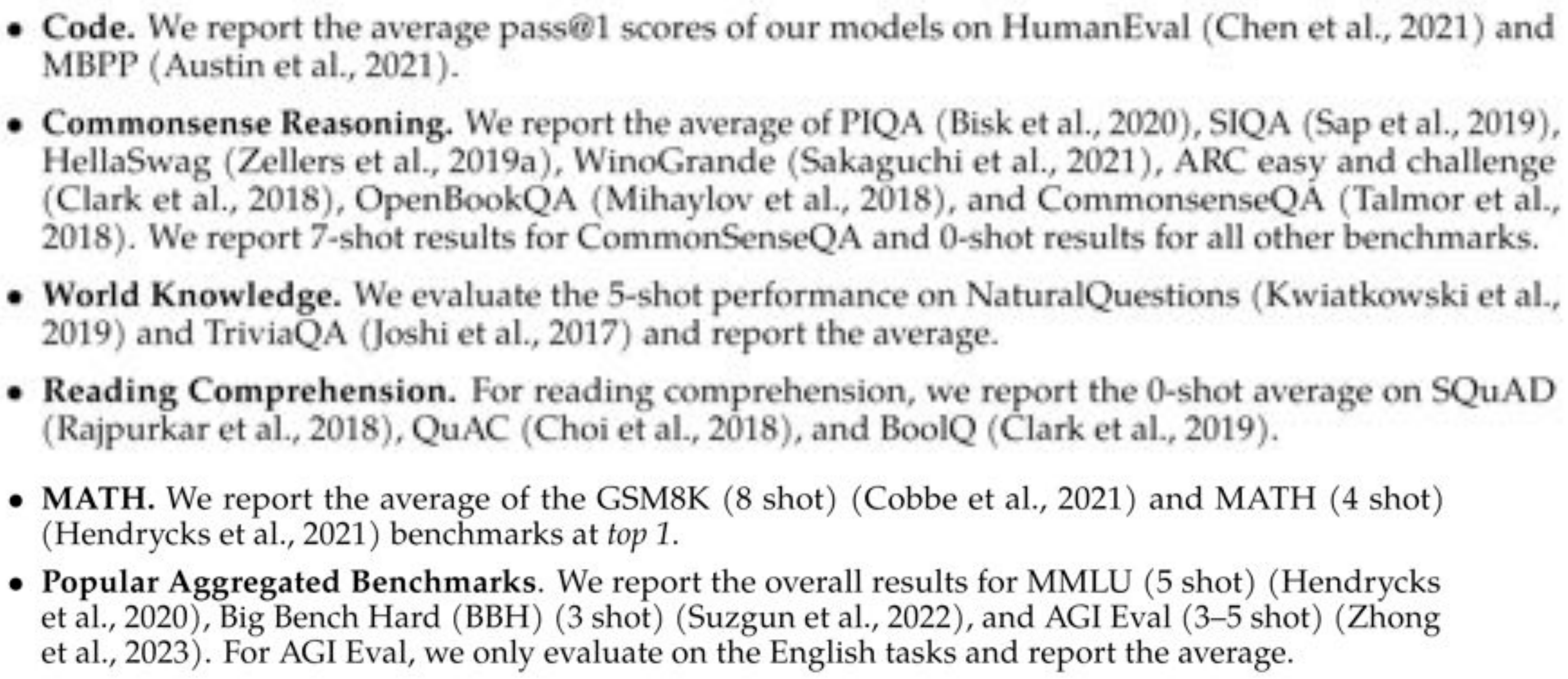
But why does it seem (almost) solved, to NLP/AI researchers?

- Our testing methodology is flawed
- An important aspect of understanding is ignored

Part I: From NLU to SLU

But why does it seem (almost) solved, to NLP/AI researchers?

- Our testing methodology is flawed



- **Code.** We report the average pass@1 scores of our models on HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021).
- **Commonsense Reasoning.** We report the average of PIQA (Bisk et al., 2020), SIQA (Sap et al., 2019), HellaSwag (Zellers et al., 2019a), WinoGrande (Sakaguchi et al., 2021), ARC easy and challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and CommonsenseQA (Talmor et al., 2018). We report 7-shot results for CommonSenseQA and 0-shot results for all other benchmarks.
- **World Knowledge.** We evaluate the 5-shot performance on NaturalQuestions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) and report the average.
- **Reading Comprehension.** For reading comprehension, we report the 0-shot average on SQuAD (Rajpurkar et al., 2018), QuAC (Choi et al., 2018), and BoolQ (Clark et al., 2019).
- **MATH.** We report the average of the GSM8K (8 shot) (Cobbe et al., 2021) and MATH (4 shot) (Hendrycks et al., 2021) benchmarks at *top 1*.
- **Popular Aggregated Benchmarks.** We report the overall results for MMLU (5 shot) (Hendrycks et al., 2020), Big Bench Hard (BBH) (3 shot) (Suzgun et al., 2022), and AGI Eval (3–5 shot) (Zhong et al., 2023). For AGI Eval, we only evaluate on the English tasks and report the average.

From the Llama-2 paper (Touvran *et al.* 2023)

The scattergun approach to evaluation

Understanding = Code + Common Sense + World Knowledge + Reading Comprehension + MATH + Exams?

A Brief History of NLP Research

formal models
(1960 – 1990s)

analysing &
formalising ∾ implementing ∾ testing

Do these formal representations capture insights, & do these algorithms process them efficiently?

specialist models
(1990s – 2015?)

learning ∾ testing
(train) task dataset (test)

Can this architecture learn this task?

transfer learning
(2015 – 2022?)

learning — learning ∾ testing
text dataset (train) task dataset (test)

Does this pre-training prepare for that task?

Does this architecture learn to understand?

generalist models
(2022 – ?)

learning testing
mystery text dataset task dataset (test)

Can this generalist model do this task as well?

A Brief History of NLP Research

learning ∾ testing

(train) task dataset (test)

Tasks were selected for variety of reasons, but not systematically related.

face validity — Is the test *subjectively* viewed as covering the concept it purports to measure?

construct validity — Do the indicators represent or reflect the concept or construct that itself is not directly observable?

A Brief History of NLP Research

learning ∾ testing

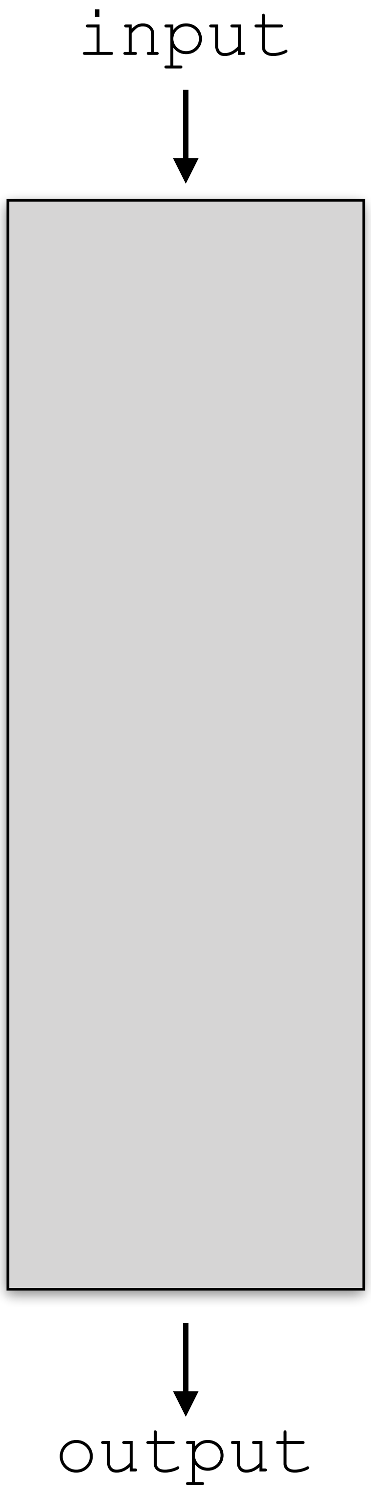
(train) task dataset (test)

Tasks were selected for variety of reasons, but not systematically related.

face validity — Is the test *subjectively* viewed as covering the concept it purports to measure?

construct validity — Do the indicators represent or reflect the concept or construct that itself is not directly observable?

Understanding Text



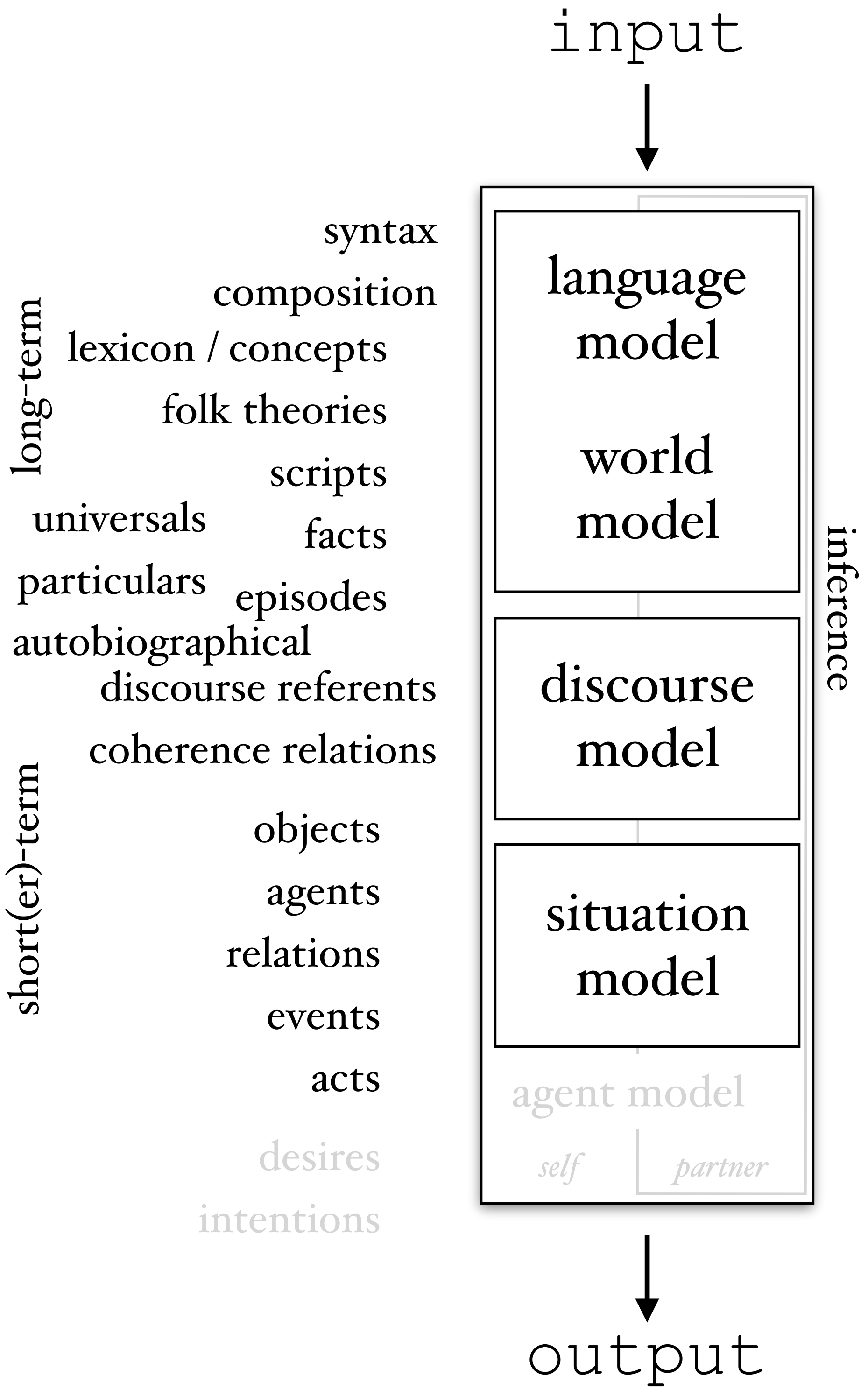
Understanding *shows* in the response.

But understanding has *internal structure*.

How can one argue for a particular structuring?

- Dissociations: Find examples where one is present, but not the other. (Evidence can be phylogenetical, ontogenetical, pathological.)
- Explanatory power: Story with components more coherent than without.

Understanding Text



input

language model

world model

discourse model

situation model

agent model

self

partner

output

inference

long-term

syntax

composition

lexicon / concepts

folk theories

scripts

universals

facts

particulars

episodes

autobiographical

discourse referents

coherence relations

short(er)-term

objects

agents

relations

events

acts

desires

intentions

“Consensus model”

(Chomsky 1957)

(Murphy 2002; Margolis & Laurence 2015)

(Kamp 1981, Heim 1983, Asher & Lascarides 2001)

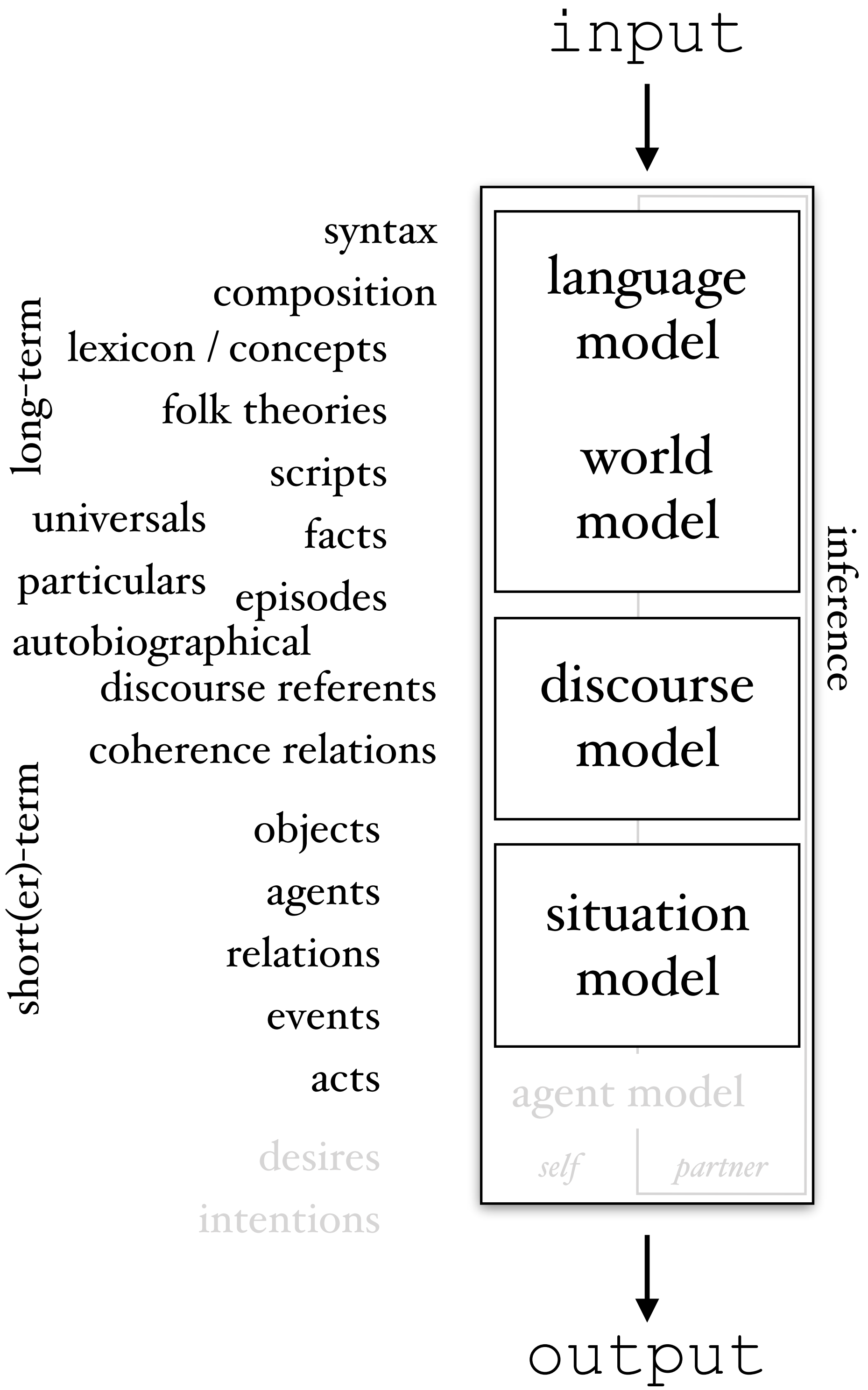
(Johnson-Laird 1983, van Dijk & Kintsch 1983)

(Bratman 1987, Cohen *et al.* 1990, Clark 1996)

(Mahowald *et al.* 2023)

(Schlangen 2023b, c)

Understanding Text



input

syntax
composition
lexicon / concepts
folk theories
scripts
universals
facts
particulars
episodes
autobiographical
discourse referents
coherence relations
objects
agents
relations
events
acts
desires
intentions

long-term

short(er)-term

language model

world model

discourse model

situation model

agent model

self

partner

inference

output

Arthur owns a dog and it follows him everywhere he goes.
Arthur doesn't own a dog and #it follows him everywhere he goes.

(Schuster & Linzen 2022)

The trophy didn't fit into the suitcase because it was too small

The trophy didn't fit into the suitcase because it was too big

(Levesque *et al.* 2012)

(Wang *et al.* 2019)
(Ribeiro *et al.* 2020)
(Dunietz *et al.* 2020)

(Loáiciga *et al.*, COLING 2022)
(Beyer *et al.*, NAACL 2021)

(Schlangen 2023b, c)

Part I: From NLU to SLU

But why does it seem (almost) solved, to NLP/AI researchers?

- Our testing methodology is flawed
    - Focus in only on collecting tests with *face validity.*
    - Construct validity is neglected.
- An important aspect of understanding is ignored

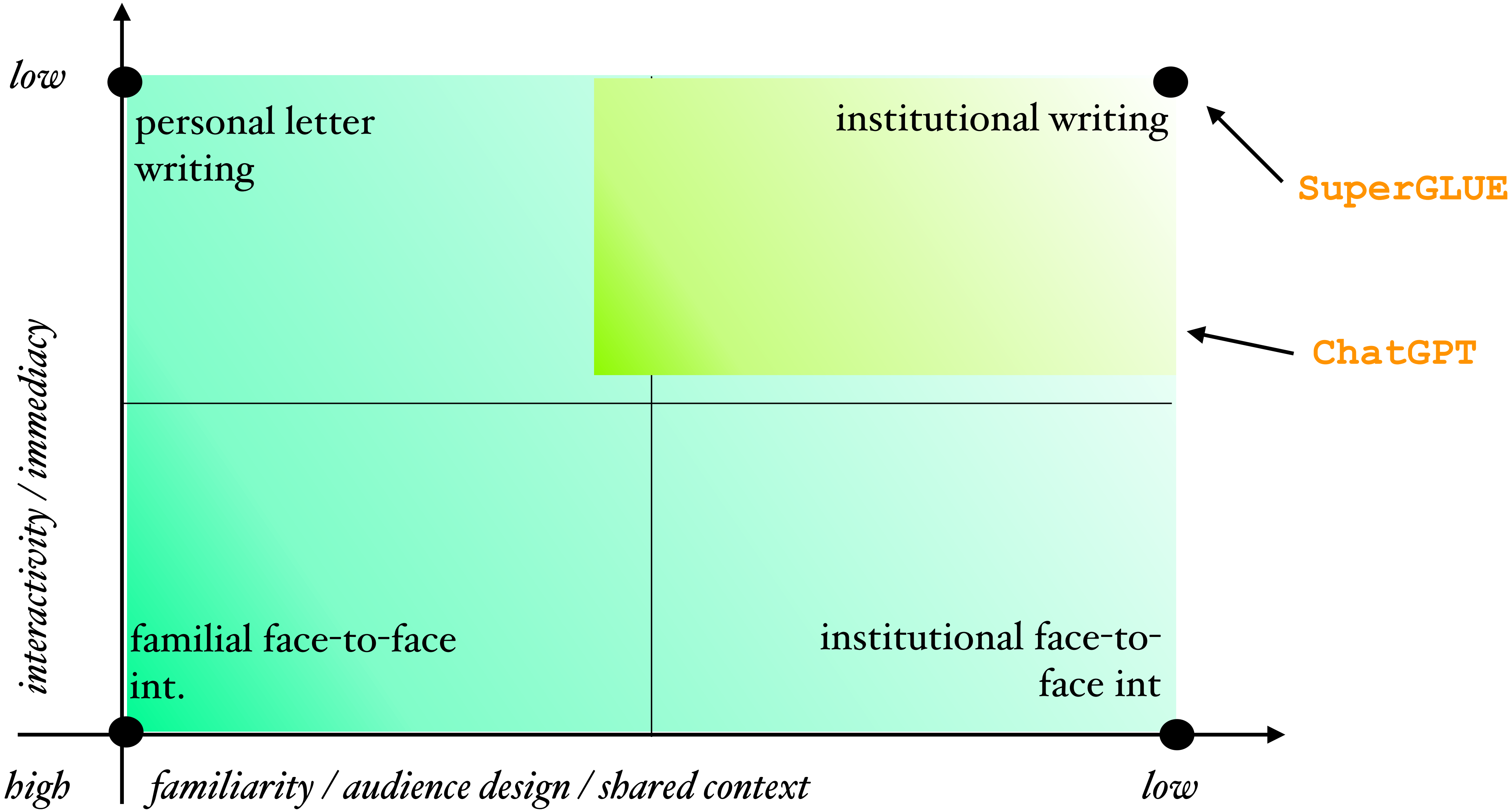
Part I: From NLU to SLU

But why does it seem (almost) solved, to NLP/AI researchers?

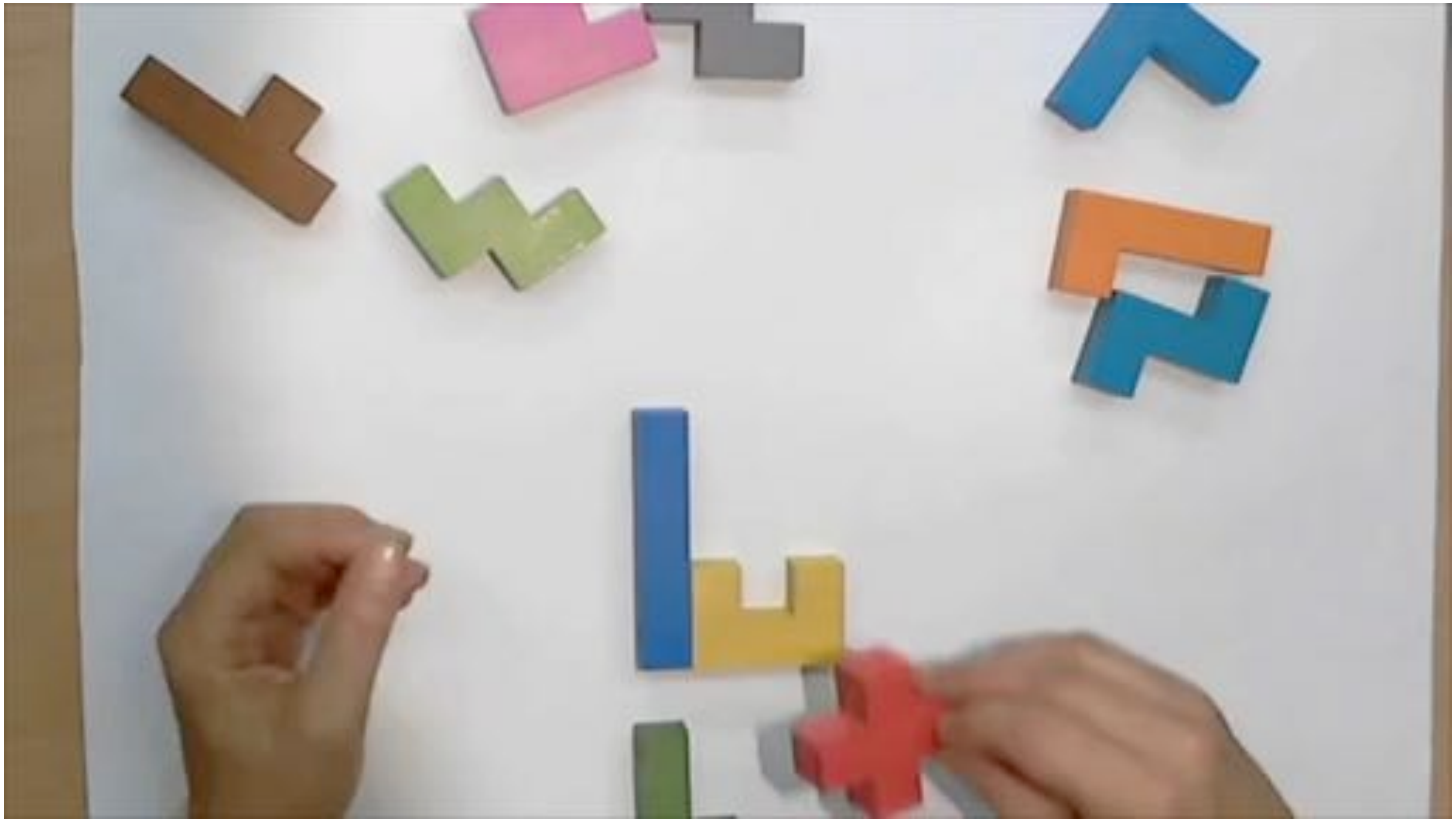
- Our testing methodology is flawed
    - Focus in only on collecting tests with *face validity*.
    - Construct validity is neglected.
- An important aspect of understanding is ignored

The Space of Language Uses

- the type of language use represented by NLP-NLU is not the only one, and not even the paradigmatic one

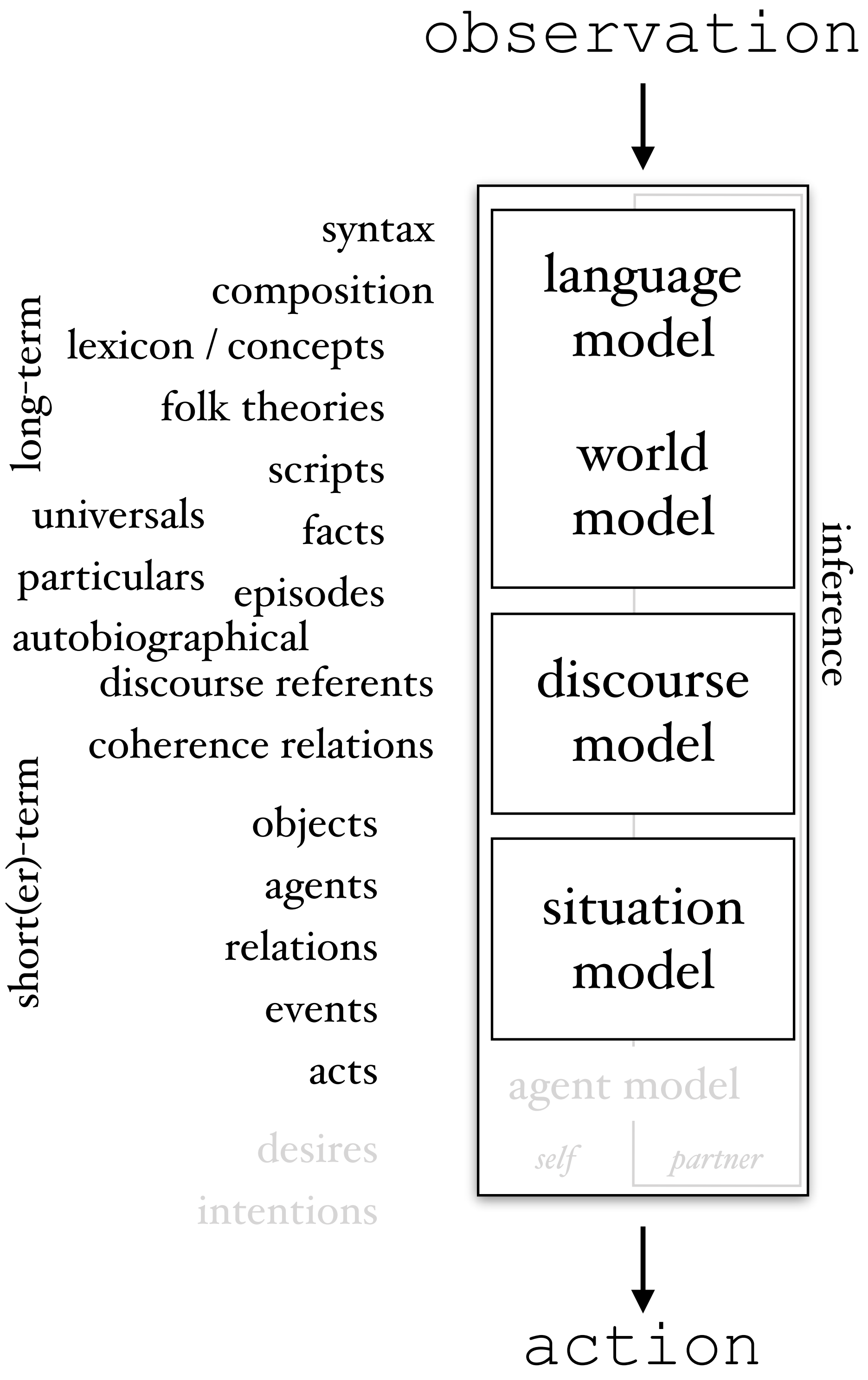

low
personal letter writing
institutional writing
SuperGLUE
ChatGPT
interactivity / immediacy
familial face-to-face int.
institutional face-to-face int
high
familiarity / audience design / shared context
low

Situated Interaction



(Zarrieß *et al.*, LREC 2016)

Situated Language Understanding



observation

long-term

syntax
composition
lexicon / concepts
folk theories
scripts
universals
facts
particulars
episodes
autobiographical
discourse referents
coherence relations

short(er)-term

objects
agents
relations
events
acts
desires
intentions

language model

world model

discourse model

situation model

agent model

self

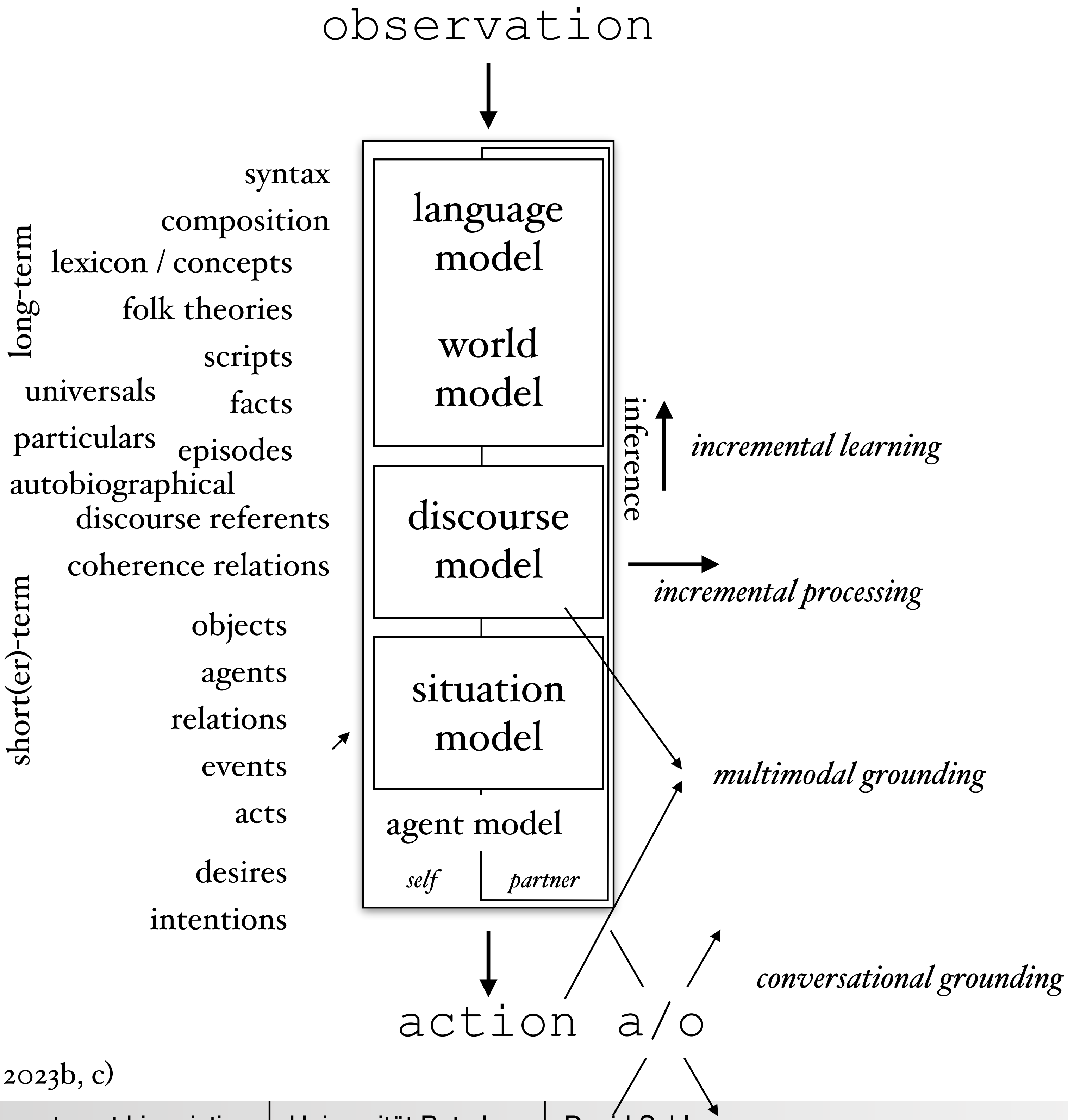
partner

inference

action

(Schlangen 2023b, c)

Situated Language Understanding



`observation`

long-term

syntax
composition
lexicon / concepts
folk theories
scripts
universals
facts
particulars
episodes
autobiographical
discourse referents
coherence relations

short(er)-term

objects
agents
relations
events
acts
desires
intentions

language model

world model

discourse model

situation model

agent model

self

partner

inference

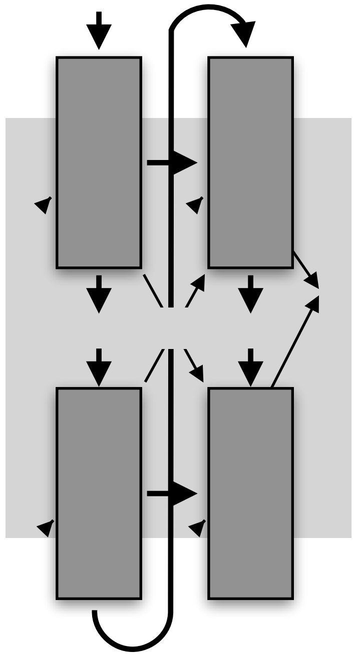
incremental learning

incremental processing

multimodal grounding

conversational grounding

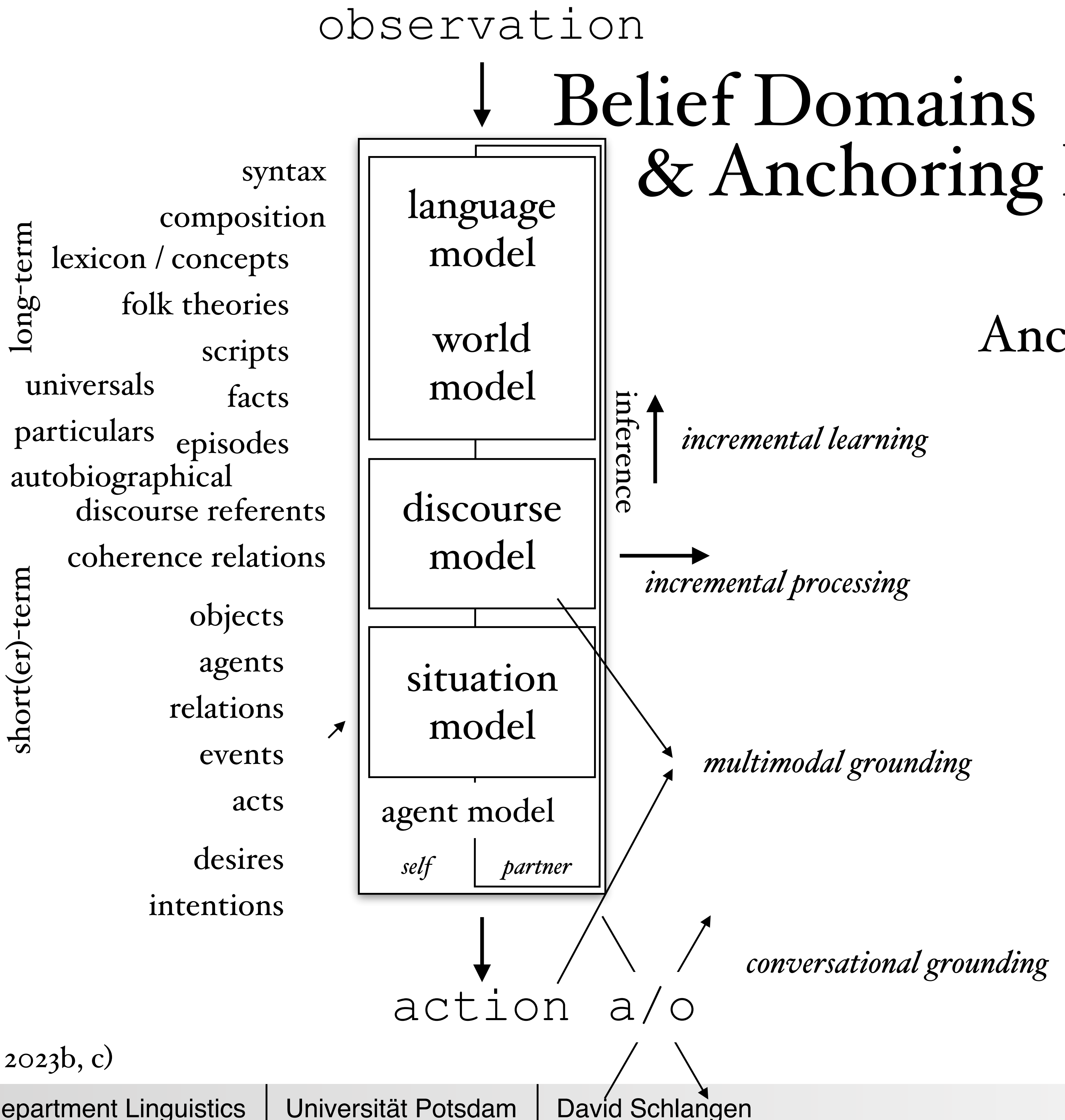
`action` `a/o`



(Schlangen 2023b, c)

Situated Language Understanding

Belief Domains & Anchoring Processes



observation

long-term

syntax
composition
lexicon / concepts
folk theories
scripts
universals
facts
particulars
episodes
autobiographical
discourse referents
coherence relations

short(er)-term

objects
agents
relations
events
acts
desires
intentions

language model

world model

discourse model

situation model

agent model

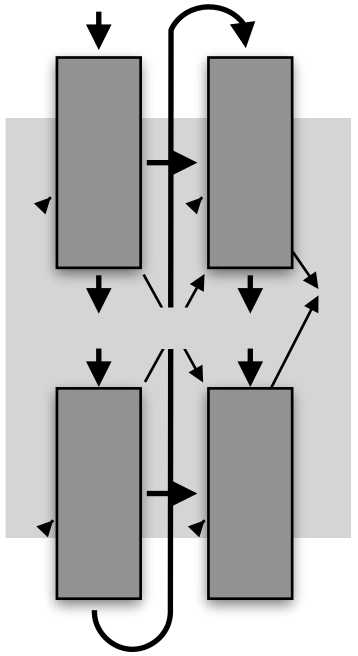
self

partner

inference

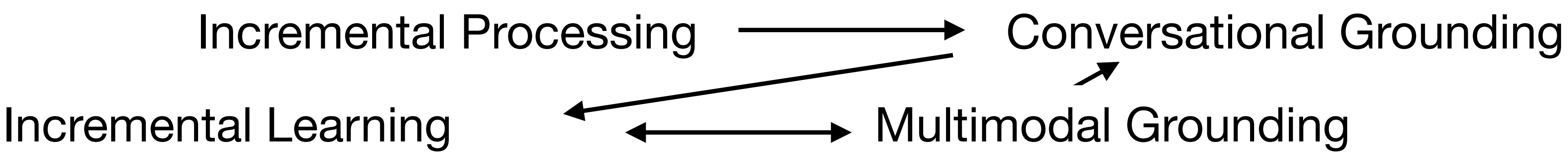
action a/o

	Anchoring Processes
incremental learning	*Us*
incremental processing	*now*
multimodal grounding	*here*
conversational grounding	*us*



(Schlangen 2023b, c)

BD/AP



Language Model

World Model

Situation Model

Discourse Model

Agent Model

(Levinson 2010)
(Christianson & Chater 2016)

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

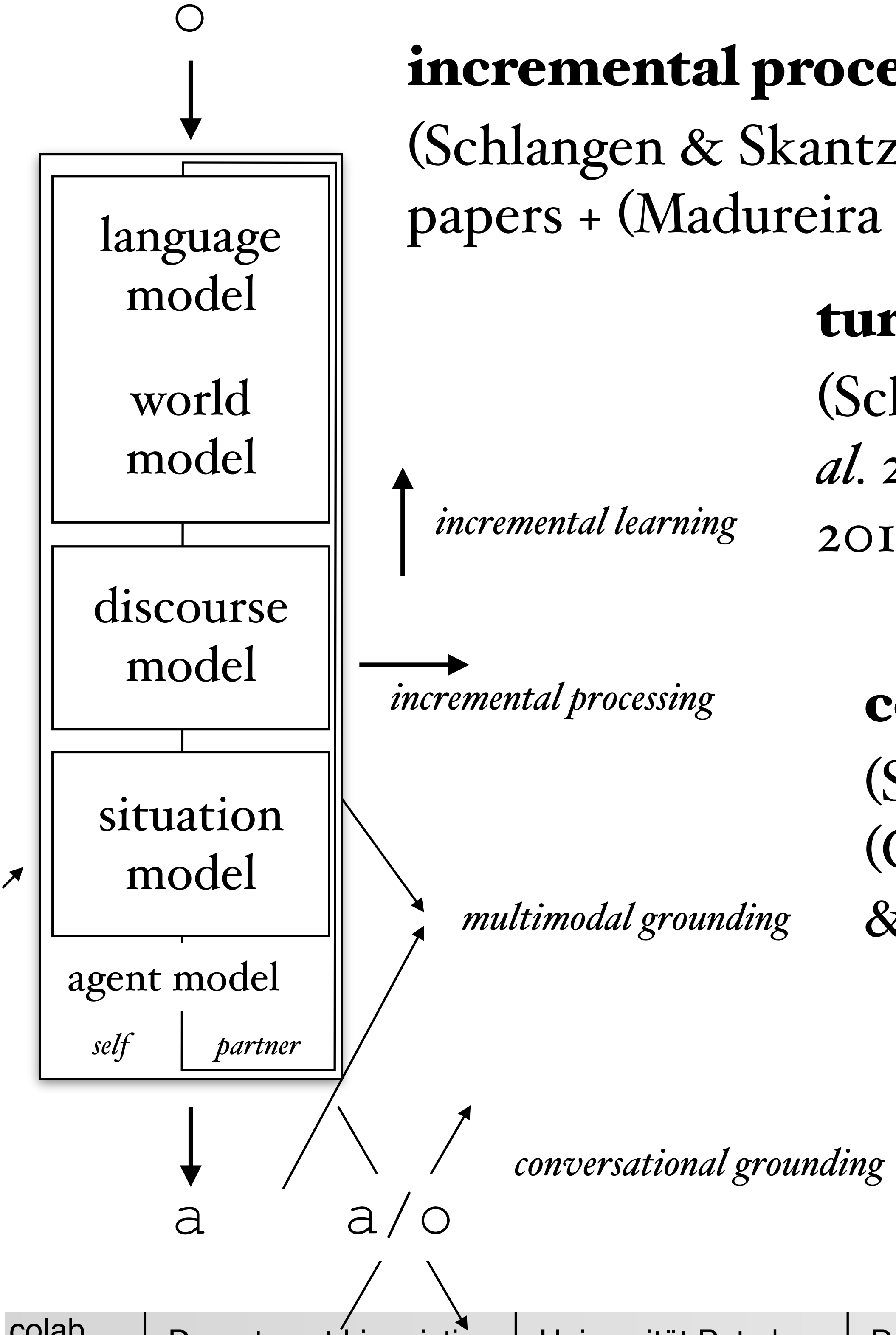
(Harris 2015)
(E. Clark 2003)

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

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

Schlangen (2023a)

BD/AP



incremental processing

(Schlangen & Skantze 2009; Skantze & Schlangen 2009) + ~30+ other papers + (Madureira & Schlangen 2020, Kahardipraja *et al.* 2021, 2023)

turn taking

(Schlangen 2006), (Atterer *et al.* 2008), … , (Andrist *et al.* 2016), (Kousidis & Schlangen 2015), (Maier *et al.* 2017), (Hough & Schlangen 2017)

conversational grounding / repair

(Schlangen 2004, Rodríguez & Schlangen 2004), …, (Ginzburg *et al.* 2014), (Hough & Schlangen 2015), …, (Hough & Schlangen 2017), …, (Madureira & Schlangen 2023a, b)

multimodal grounding

(Siebert & Schlangen 2008), …, (Kennington & Schlangen 2015, Schlangen *et al.* 2016), (Kennington *et al.* 2013, Han *et al.* 2014, 2017, 2018)

This Talk

- A Theoretical Thread
  - Understanding shows differently in text and interaction
  - The “Beliefs & Processes” model
- A Methodological Thread
  - Negative: How *not* To Do Things
    - Text NLU is tested unsystematically, with little regard for *construct validity* of tests
  - Positive: How To Do Things
    - How can we test (and build for) SLU, with construct validity?
    - A proposal for testing chat-optimised LLMs as Interactive Agents

This Talk

- A Theoretical Thread
  - Understanding shows differently in text and interaction
  - The "Beliefs & Processes" model
- A Methodological Thread
  - Negative: How *not* To Do Things
    - Text NLU is tested unsystematically, with little regard for *construct validity* of tests
  - Positive: How To Do Things
    - How can we test (and build for) SLU, with construct validity?
    - A proposal for testing chat-optimised LLMs as Interactive Agents

A Brief History of NLP Research

formal models (1960 – 1990s)	analysing & formalising	∾	implementing	∾	testing	Do these formal representations capture insights, & do these algorithms process them efficiently?
specialist models (1990s – 2015?)			learning (train)	∾ task dataset	testing (test)	Can this architecture learn this task?
transfer learning (2015 – 2022?)	learning text dataset	—∾—	learning (train)	∾ task dataset	testing (test)	Does this pre-training prepare for that task?
						Does this architecture learn to understand?
generalist models (2022 – ?)	learning mystery text dataset			task dataset	testing (test)	Can this generalist model do this task as well?

Can we learn from NLU?

	NLP *tasks (& datasets)* to set goals	*Dialogue Games* to set goals
formal models (1960 – 1990s)	analyse…	
specialist models (1990s – 2015?)	& model single task	learn from data
transfer learning (2015 – 2022?)	transfer knowledge from task to task	(worry a bit more about construct validity)
generalist models (2022 – ?)	all tasks are contained?	learn from / test with *self-play?*

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)

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)

multi-turn (provides context to language)

clear metric for whether / how well goal has been reached

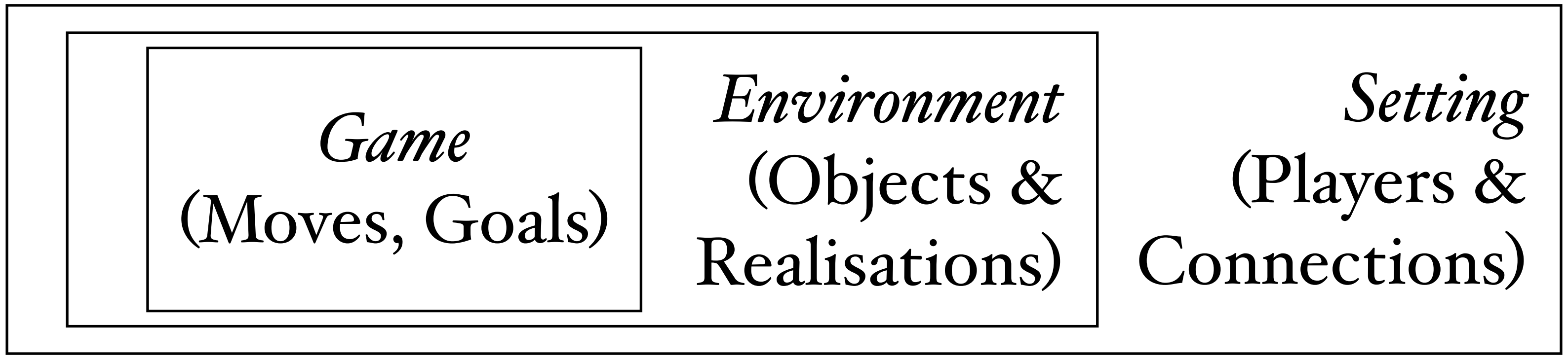
goal & rules provide
- control over type of context that is relevant
- control over type of knowledge that is relevant

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

Dialogue Game Taxonomy

(Schlangen 2023b)

- Our proposal: A fine-grained taxonomy of dialogue games,
- … with clear connections to BD/AP model,
- … and a partial ordering in terms of complexity. (“Progress.”)



Game
(Moves, Goals)

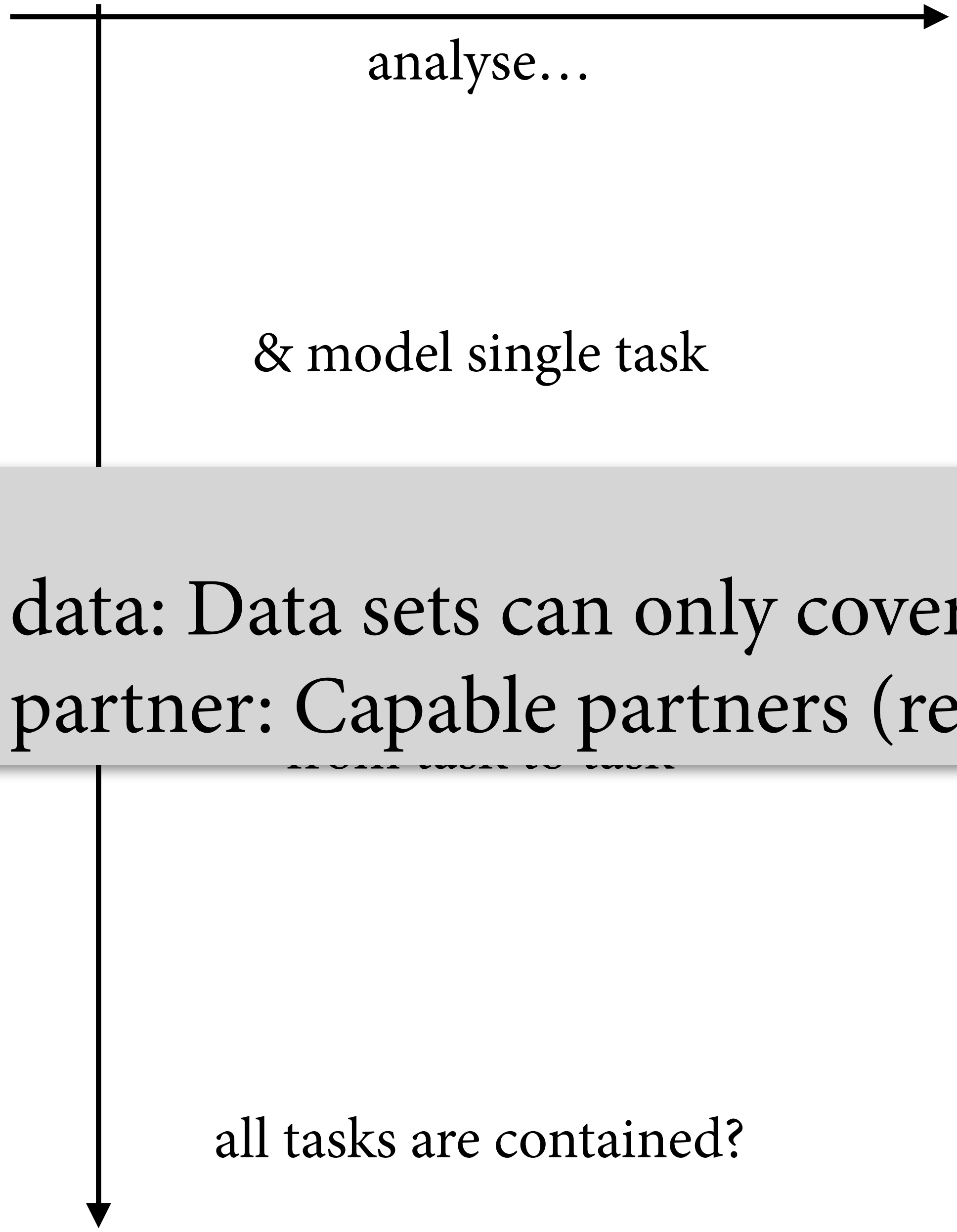
Environment
(Objects &
Realisations)

Setting
(Players &
Connections)

This Talk

- A Theoretical Thread
  - Understanding shows differently in text and interaction
  - The "Beliefs & Processes" model
- A Methodological Thread
  - Negative: How *not* To Do Things
    - Text NLU is tested unsystematically, with little regard for *construct validity* of tests
  - Positive: How To Do Things
    - How can we test (and build for) SLU, with construct validity?
    - A proposal for testing chat-optimised LLMs as Interactive Agents

Can we learn from NLU?



NLP *tasks (& datasets)*
to set goals

Dialogue Games
to set goals

formal models
(1960 – 1990s)

analyse…

specialist models
(1990s – 2015?)

& model single task

learn from data

Problems:
- Learn from data: Data sets can only cover small part of space.
- Learn from partner: Capable partners (real humans) are slow and expensive.

tra
(2015 – 2022?)

from task to task

construct validity)

generalist models
(2022 – ?)

all tasks are contained?

learn from / test
with *self-play?*

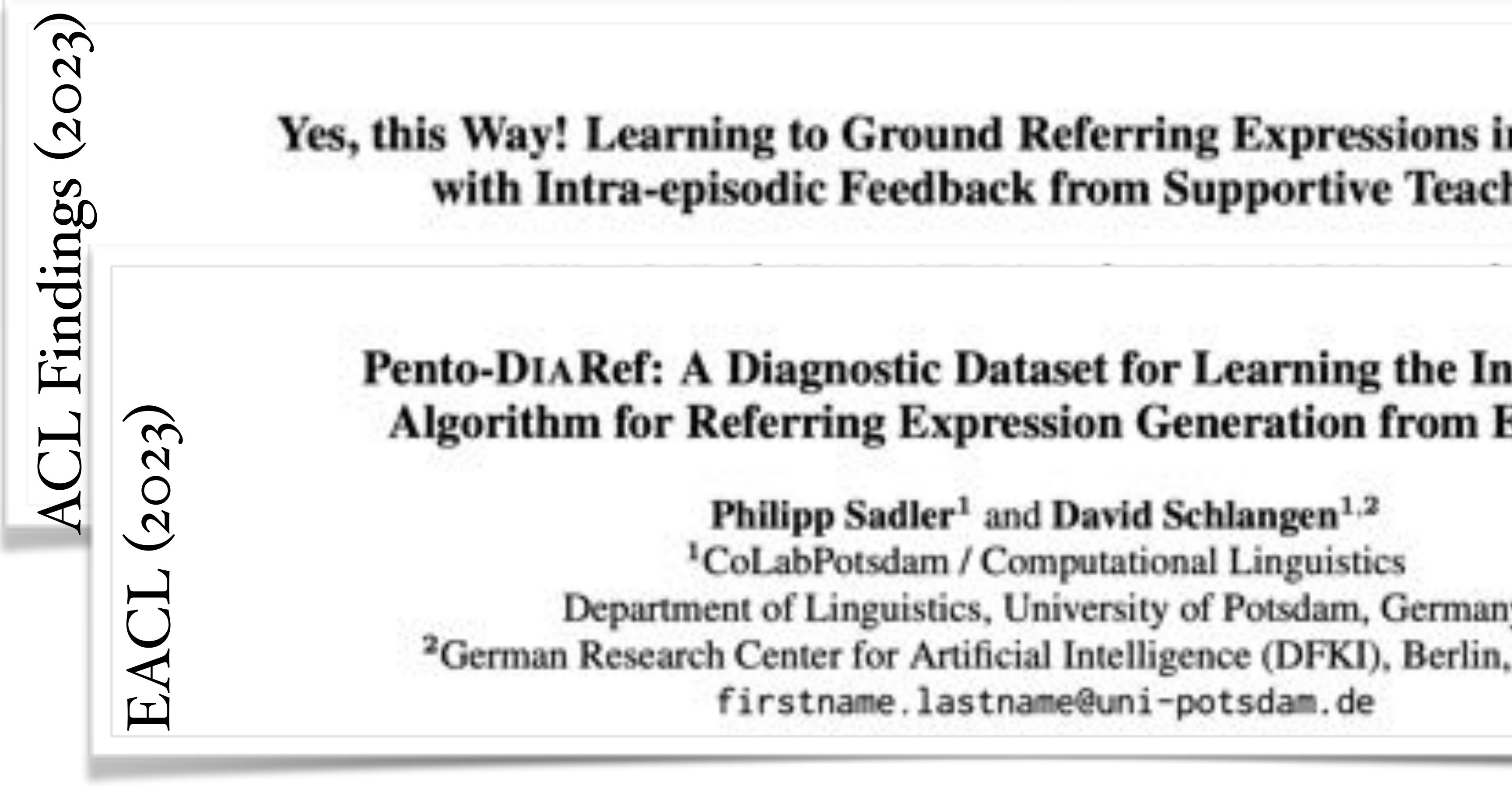
The *Programmatic Teacher* Approach

Problems:

- Learn from data: Data sets can only cover small part of space.
- Learn from partner: Capable partners (real humans) are slow and expensive.

This Approach:

- Let agent learn from programmatic partner, implementing behaviour models.



ACL Findings (2023)

Yes, this Way! Learning to Ground Referring Expressions into Actions with Intra-episodic Feedback from Supportive Teachers

EACL (2023)

Pento-DIARef: A Diagnostic Dataset for Learning the Incremental Algorithm for Referring Expression Generation from Examples

Philipp Sadler[1] and **David Schlangen**[1,2]
[1]CoLabPotsdam / Computational Linguistics
Department of Linguistics, University of Potsdam, German
[2]German Research Center for Artificial Intelligence (DFKI), Berlin,
`firstname.lastname@uni-potsdam.de`

Philipp Sadler

Sherzod Hakimov

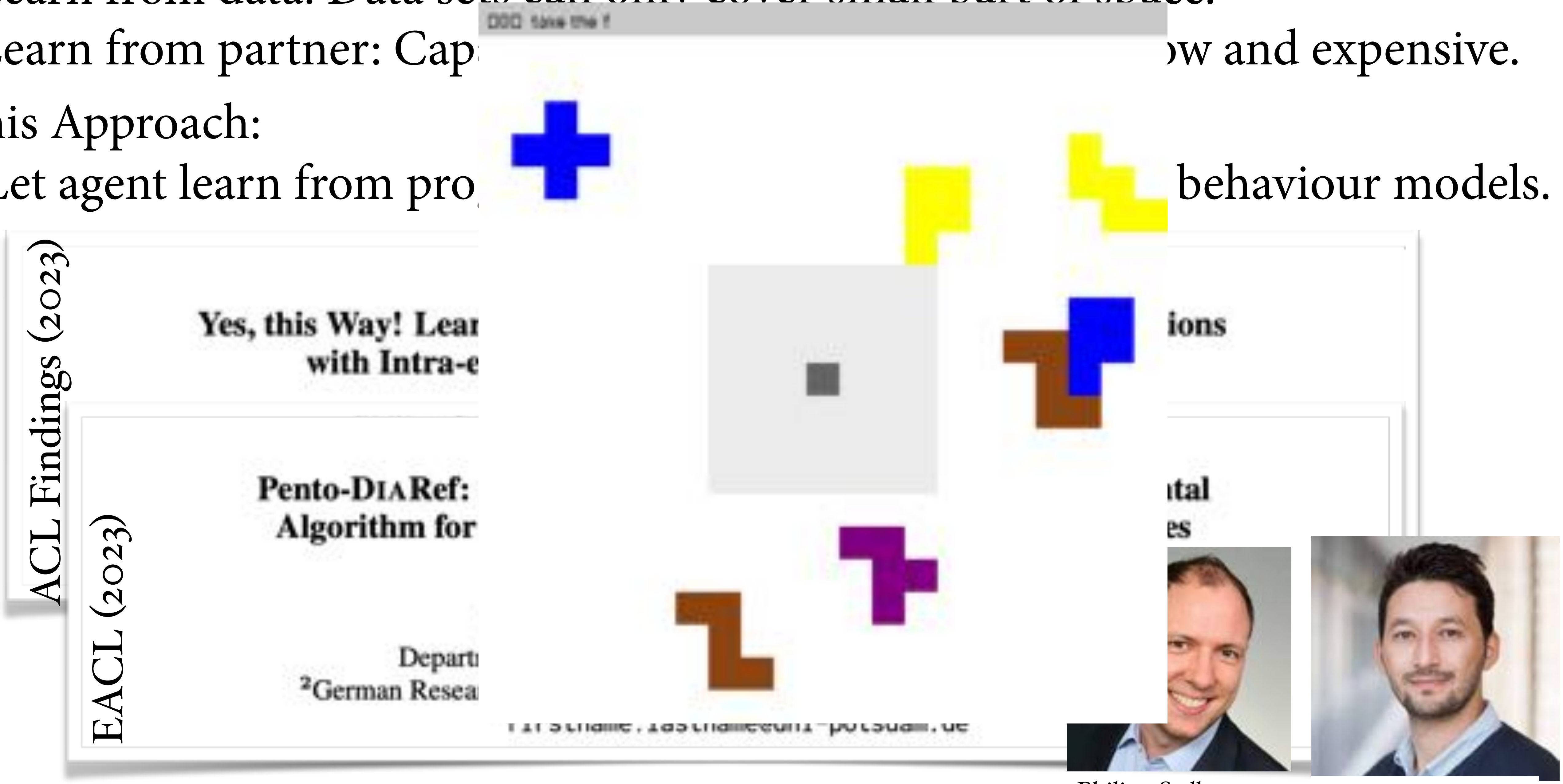
The *Programmatic Teacher* Approach

Problems:

- Learn from data: Data sets can only cover small part of space.
- Learn from partner: Cap w and expensive.

This Approach:

- Let agent learn from pro behaviour models.



Yes, this Way! Lear ions
with Intra-e

Pento-DIARef: tal
Algorithm for es

Depart
[2]German Resea

firstname.lastname@uni-potsdam.de

Philipp Sadler

Sherzod Hakimov

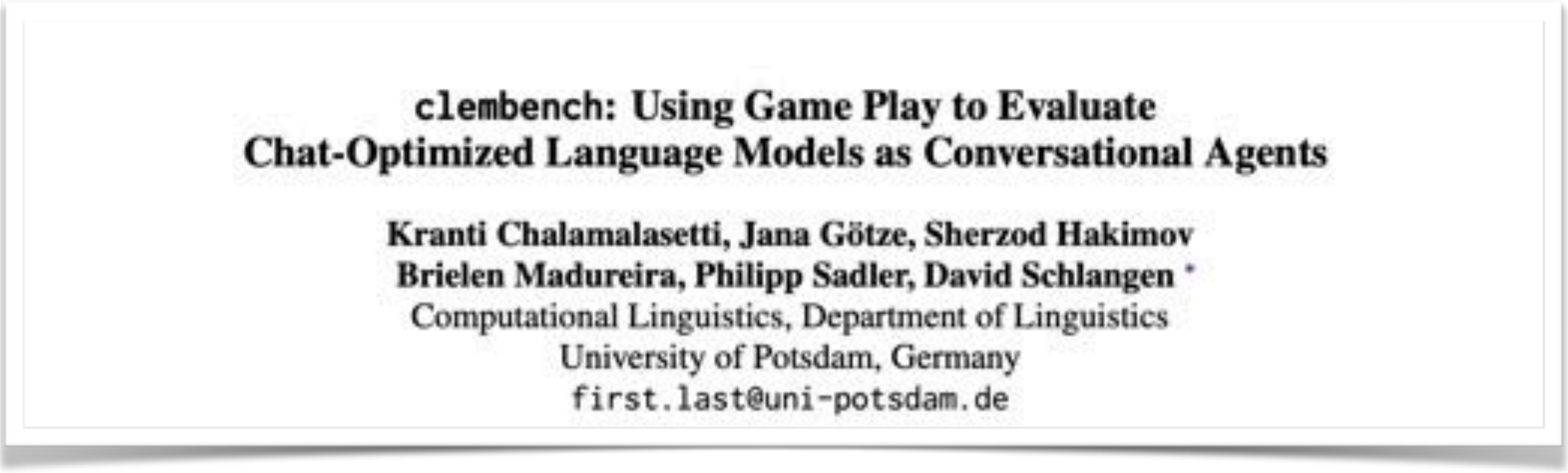
The *Self-Play* Approach

Problems:

- Learn from data: Data sets can only cover small part of space.
- Learn from partner: Capable partners (real humans) are slow and expensive.

This Approach:

- Let us see if the new generalist models are not already 0-/few-shot *agents!*



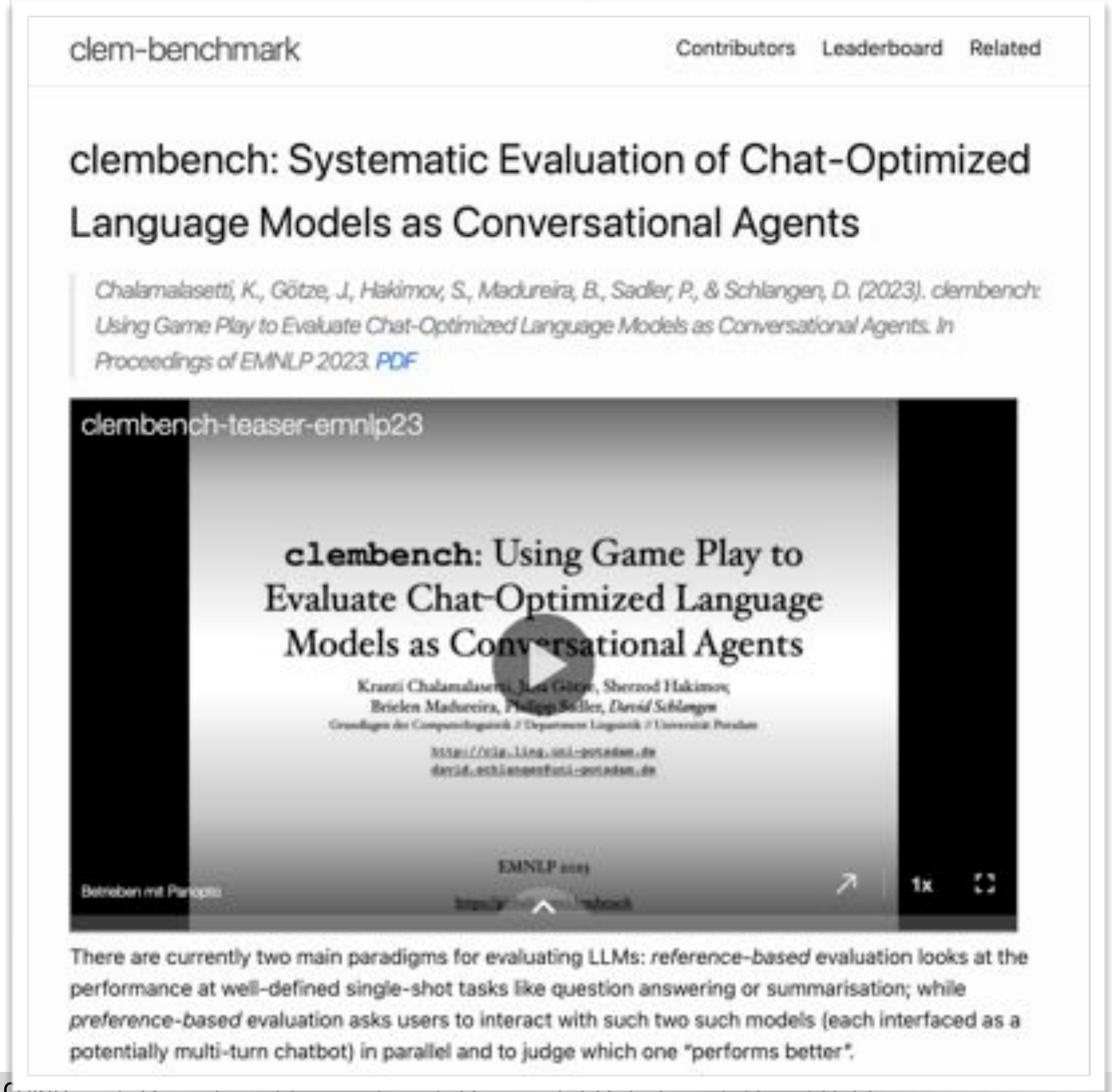
clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents

Kranti Chalamalasetti, Jana Götze, Sherzod Hakimov
Brielen Madureira, Philipp Sadler, David Schlangen *
Computational Linguistics, Department of Linguistics
University of Potsdam, Germany
`first.last@uni-potsdam.de`

(Chalamalasetti *et al.*, EMNLP 2023)

clemgame & clembench

(clem = cLLM = chat-optimized LLM…)



clembench: Systematic Evaluation of Chat-Optimized Language Models as Conversational Agents

> *Chalamalasetti, K., Götze, J., Hakimov, S., Madureira, B., Sadler, P., & Schlangen, D. (2023). clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents. In Proceedings of EMNLP 2023.* PDF

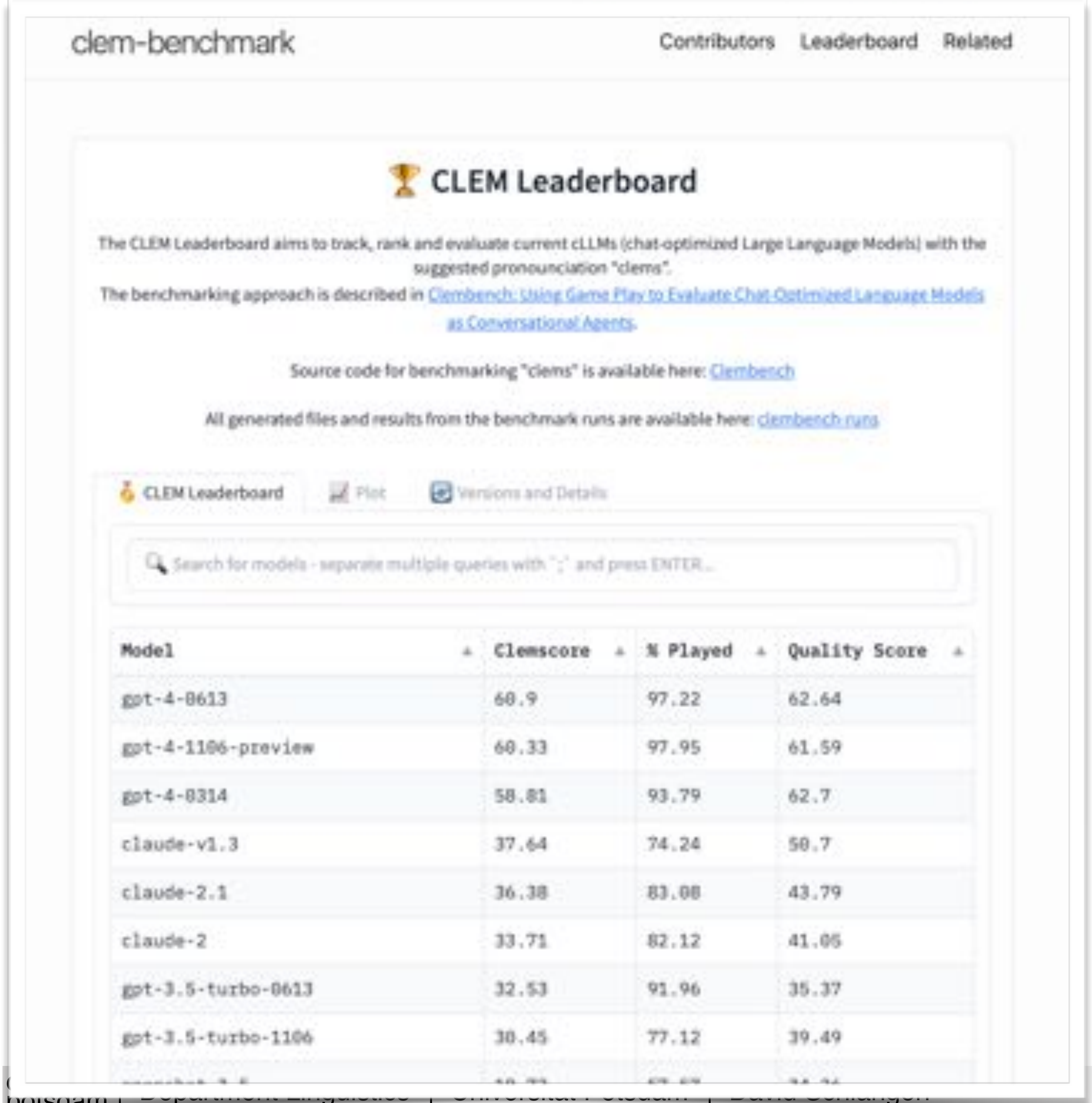
clembench-teaser-emnlp23
clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents
EMNLP 2023
Betrieben mit Panopto
1x

There are currently two main paradigms for evaluating LLMs: *reference-based* evaluation looks at the performance at well-defined single-shot tasks like question answering or summarisation; while *preference-based* evaluation asks users to interact with such two such models (each interfaced as a potentially multi-turn chatbot) in parallel and to judge which one "performs better".

https://clembench.github.io/

clemgame & clembench

(clem = cLLM = chat-optimized LLM…)



🏆 CLEM Leaderboard

The CLEM Leaderboard aims to track, rank and evaluate current cLLMs (chat-optimized Large Language Models) with the suggested pronounciation "clems".
The benchmarking approach is described in [Clembench: Using Game Play to Evaluate Chat Optimized Language Models as Conversational Agents](#).

Source code for benchmarking "clems" is available here: [Clembench](#)

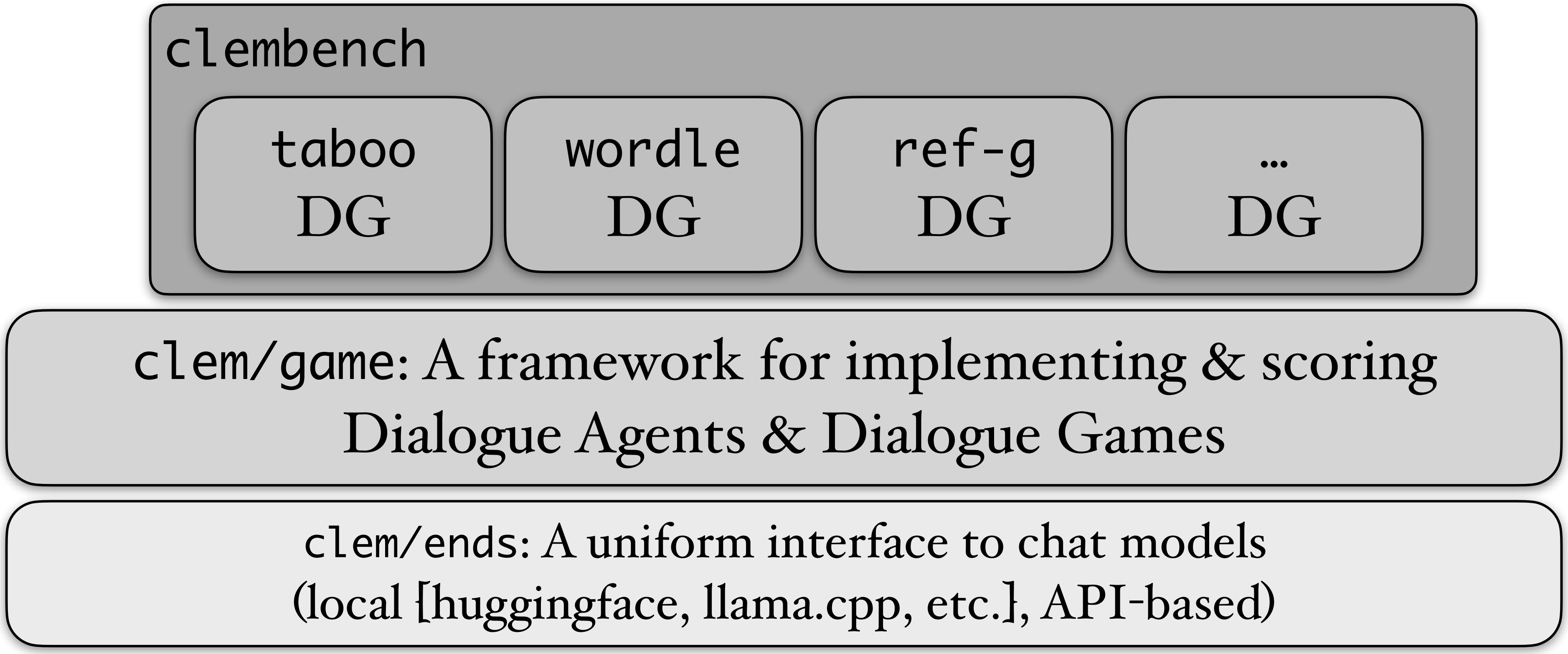
All generated files and results from the benchmark runs are available here: [clembench runs](#)

CLEM Leaderboard
Plot
Versions and Details
Search for models - separate multiple queries with ";" and press ENTER...

Model	Clemscore	% Played	Quality Score
gpt-4-0613	60.9	97.22	62.64
gpt-4-1106-preview	60.33	97.95	61.59
gpt-4-0314	58.81	93.79	62.7
claude-v1.3	37.64	74.24	50.7
claude-2.1	36.38	83.08	43.79
claude-2	33.71	82.12	41.05
gpt-3.5-turbo-0613	32.53	91.96	35.37
gpt-3.5-turbo-1106	30.45	77.12	39.49

https://clembench.github.io/

The Framework



`clembench`

`taboo` DG

`wordle` DG

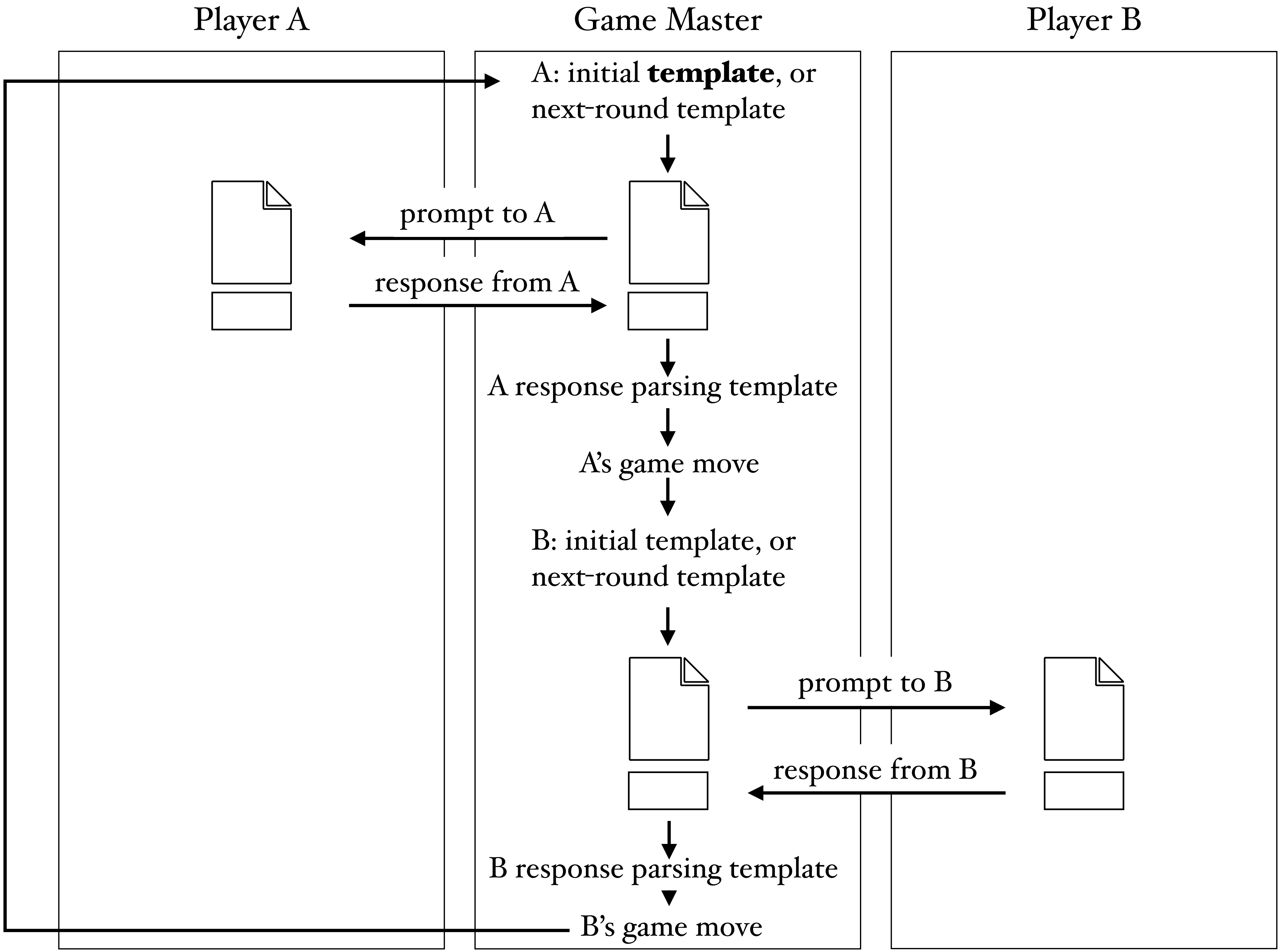
`ref-g` DG

… DG

`clem/game`: A framework for implementing & scoring Dialogue Agents & Dialogue Games

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

`clemgame` Principles


Player A

Game Master

Player B

A: initial **template**, or next-round template

prompt to A

response from A

A response parsing template

A's game move

B: initial template, or next-round template

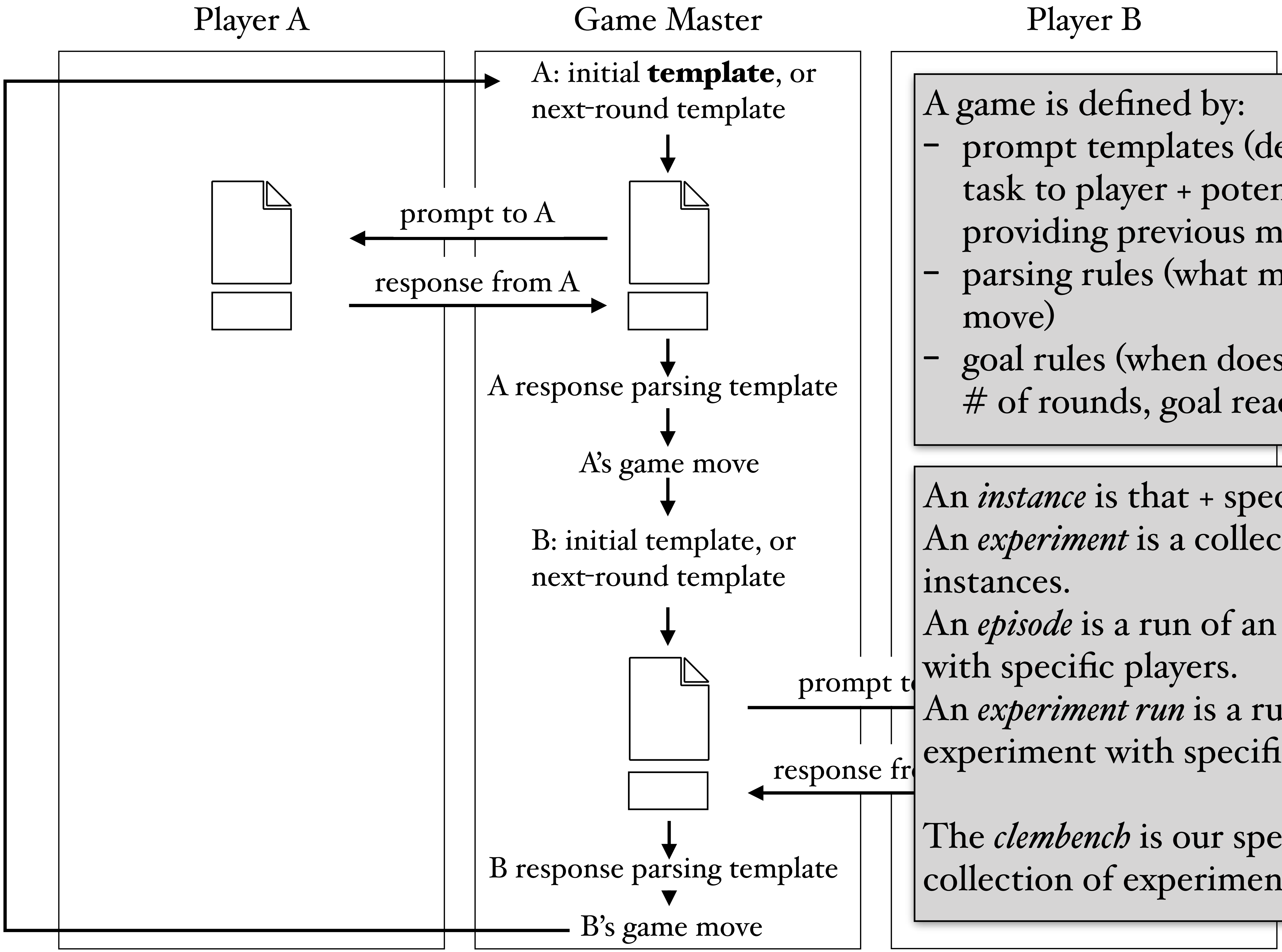
prompt to B

response from B

B response parsing template

B's game move

`clemgame` Principles



Player A

Game Master

Player B

A: initial **template**, or next-round template

prompt to A

response from A

A response parsing template

A's game move

B: initial template, or next-round template

prompt t

response fr

B response parsing template

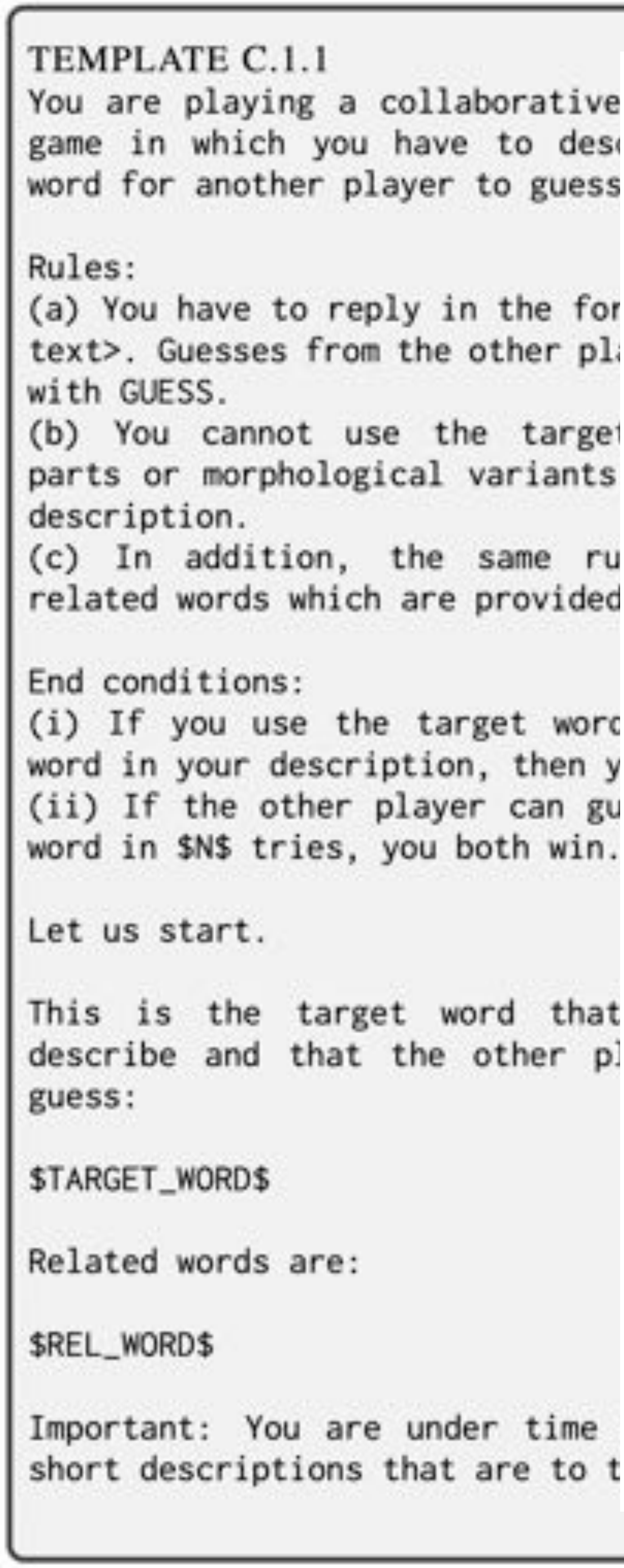
B's game move

A game is defined by:
- prompt templates (describing the task to player + potentially providing previous moves)
- parsing rules (what makes legal move)
- goal rules (when does game stop? # of rounds, goal reached, etc.)

An *instance* is that + specified goal.
An *experiment* is a collection of instances.
An *episode* is a run of an instance with specific players.
An *experiment run* is a run of an experiment with specific players.

The *clembench* is our specific collection of experiments.

Example: Taboo



TEMPLATE C.1.1
You are playing a collaborative
game in which you have to desc
word for another player to guess

Rules:
(a) You have to reply in the for
text>. Guesses from the other pl
with GUESS.
(b) You cannot use the target
parts or morphological variants
description.
(c) In addition, the same ru
related words which are provided

End conditions:
(i) If you use the target word
word in your description, then y
(ii) If the other player can gu
word in N tries, you both win.

Let us start.

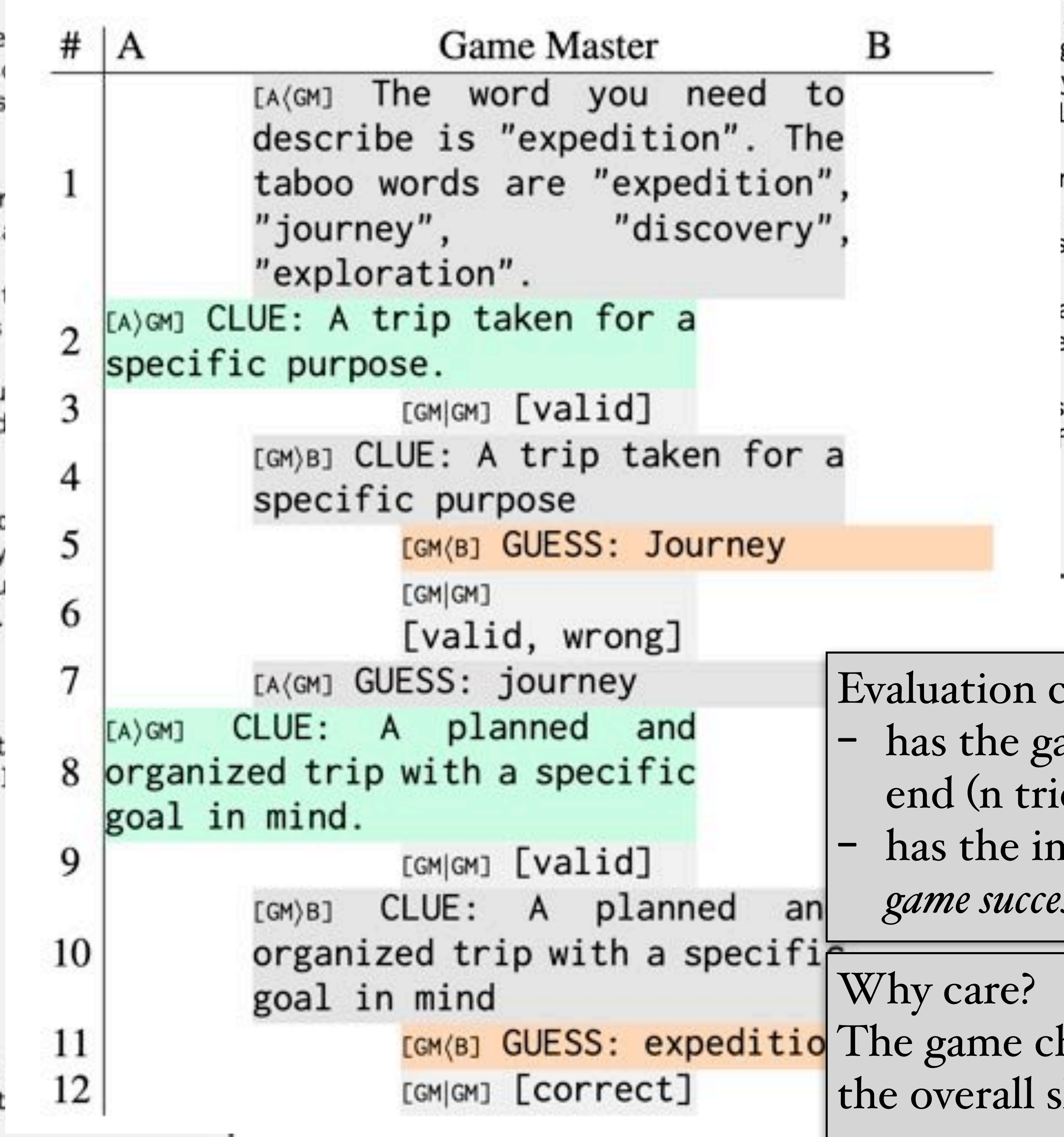
This is the target word that
describe and that the other pl
guess:

$TARGET_WORD$

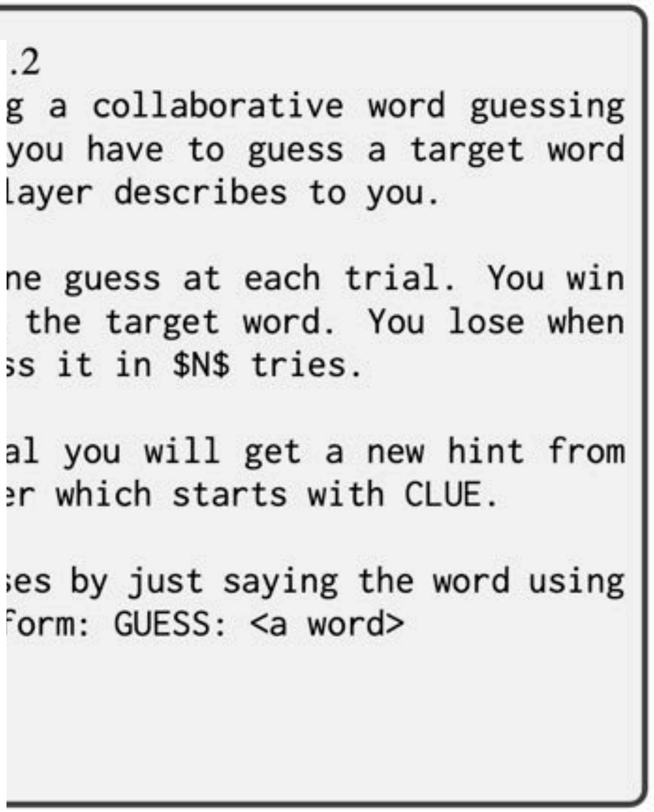
Related words are:

REL_WORD

Important: You are under time
short descriptions that are to t



#	A	Game Master	B
1		[A⟨GM] The word you need to describe is "expedition". The taboo words are "expedition", "journey", "discovery", "exploration".	
2	[A⟩GM] CLUE: A trip taken for a specific purpose.		
3		[GM|GM] [valid]	
4		[GM⟩B] CLUE: A trip taken for a specific purpose	
5			[GM⟨B] GUESS: Journey
6		[GM|GM] [valid, wrong]	
7		[A⟨GM] GUESS: journey	
8	[A⟩GM] CLUE: A planned and organized trip with a specific goal in mind.		
9		[GM|GM] [valid]	
10		[GM⟩B] CLUE: A planned an organized trip with a specifi goal in mind	
11			[GM⟨B] GUESS: expeditio
12		[GM|GM] [correct]	



.2
g a collaborative word guessing
you have to guess a target word
layer describes to you.

ne guess at each trial. You win
the target word. You lose when
ss it in N tries.

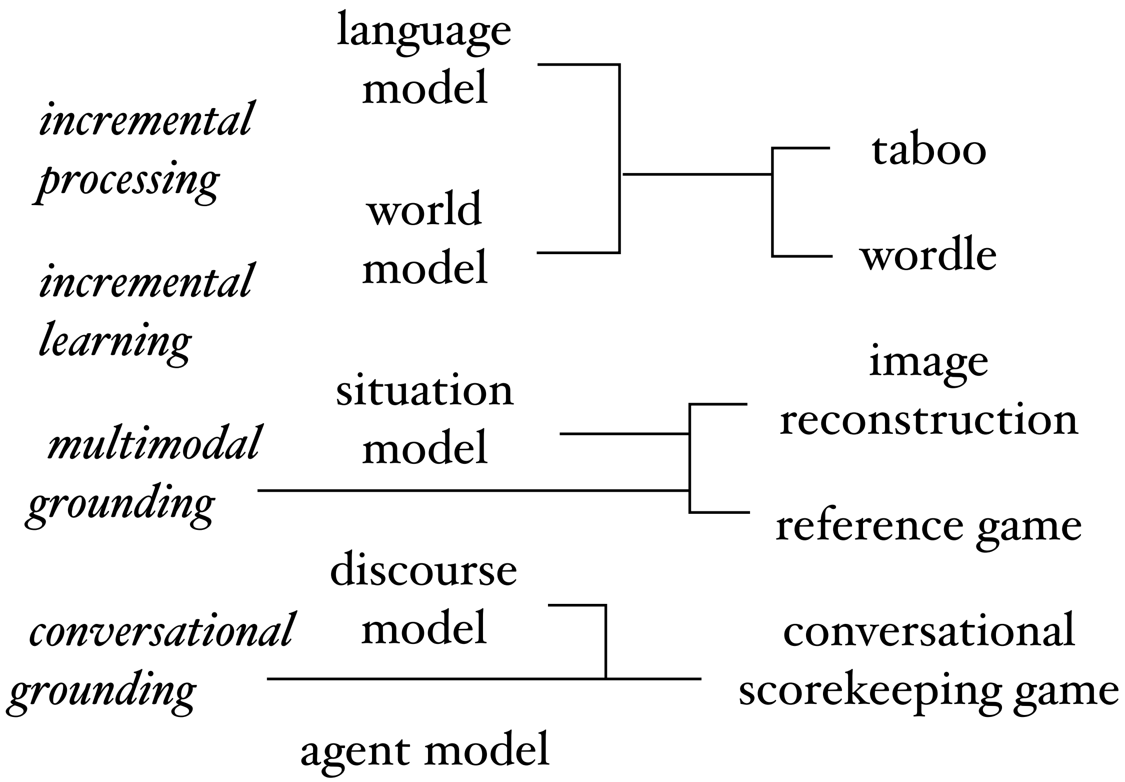
al you will get a new hint from
er which starts with CLUE.

ses by just saying the word using
form: GUESS: <a word>

Evaluation criteria:
- has the game been played to the end (n tries)? *instruction following*
- has the instance been solved? *game success*

Why care?
The game challenges core parts of the overall skill of "understanding".

The Validity Argument



incremental processing

incremental learning

multimodal grounding

conversational grounding

language model

world model

situation model

discourse model

agent model

taboo

wordle

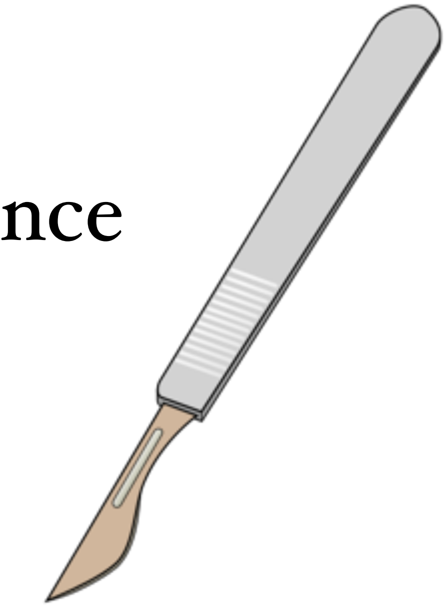
image reconstruction

reference game

conversational scorekeeping game

Games as benchmarking instrument

- coarse-grained: one number, to weigh and rank
  $(\text{quality} * \frac{\text{completed}}{\text{attempted}})$
- finer-grained: look at quality and %-completed separately
- even finer: look at performance by game
- very fine-grained: fine analysis of game performance (additional metrics), qualitative analysis

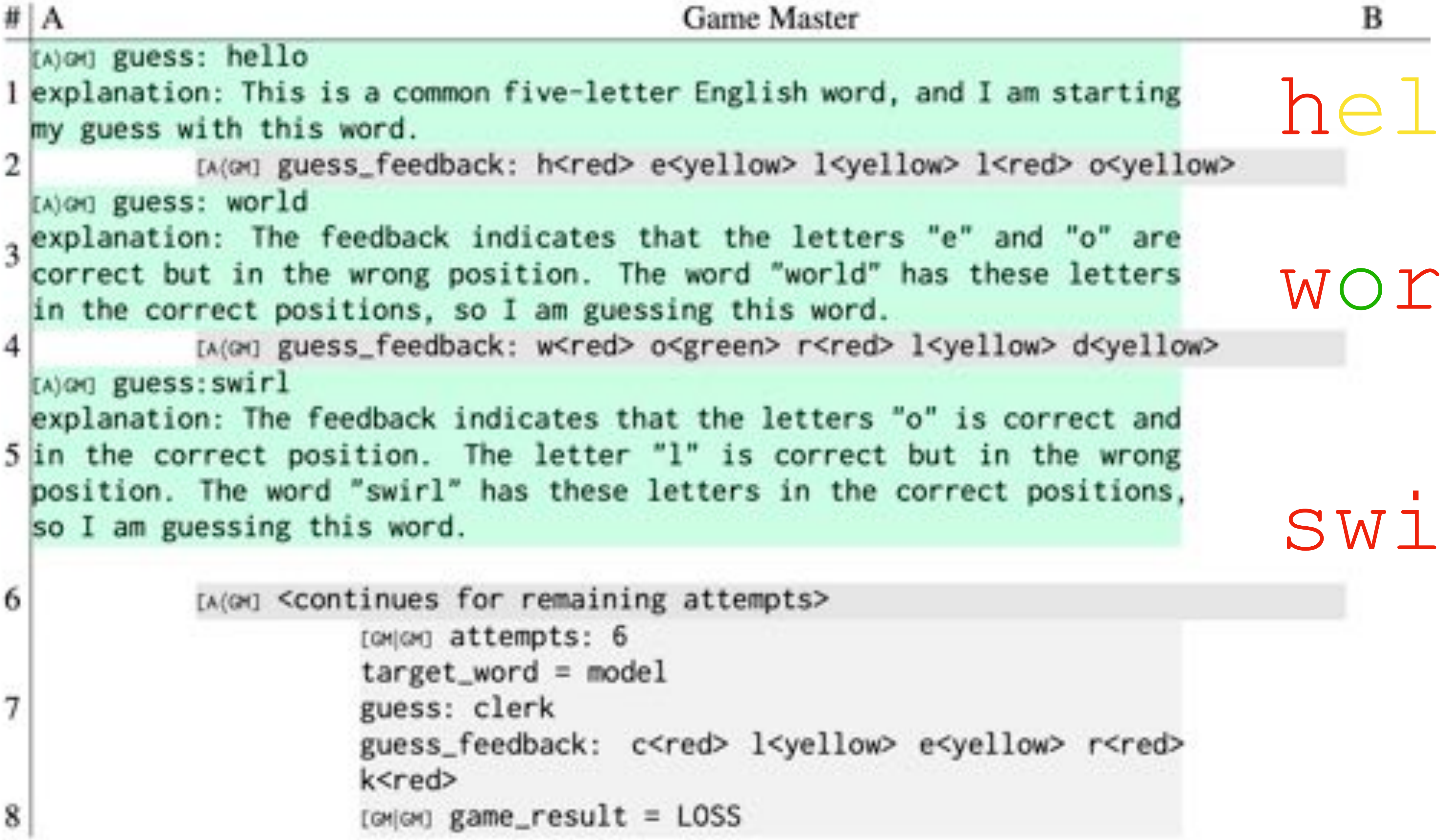


https://creativecommons.org/licenses/by-sa/4.0/deed.en https://commons.wikimedia.org/wiki/File:Dessin_scalpel.svg Petit B

The games in `clembench-1.0`

- **taboo**: word guessing game with word constraints
- **wordle**: word guessing game with letter-based constraints
  - **wordle+clue**: additionally, content-based clue
  - **wordle+clue+critic**: other player needs to agree with choice
- **image reconstruction game**: A gets B to reconstruct ASCII grid image
- **reference game**: A describes one (out of three) ASCII images to B, who “sees” them in different order. (Single shot.)
- **conv. scorekeeping game**: A provides information from list to B, but is privately asked by GM about what they “think” B knows.

Example: wordle



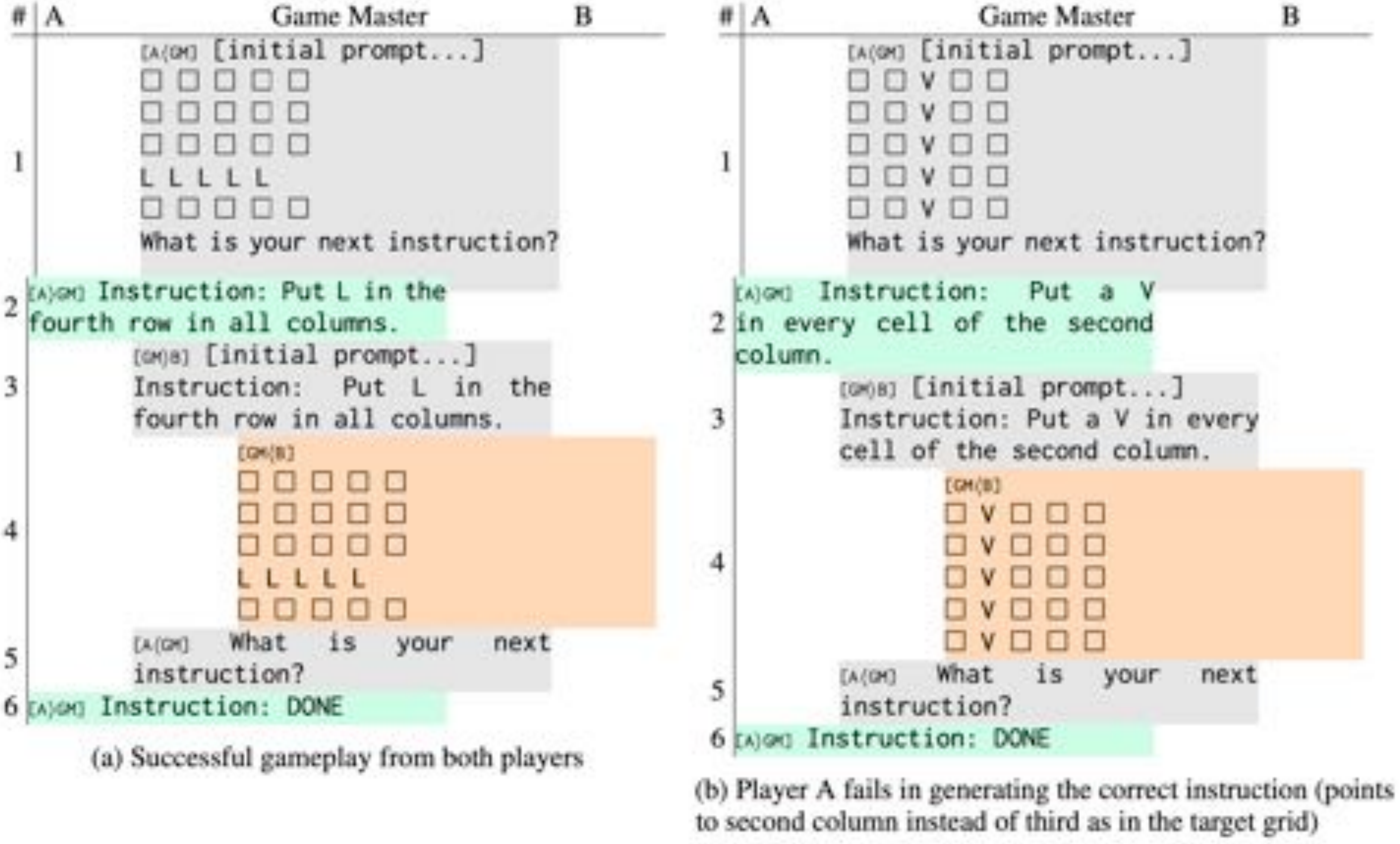
#	A	Game Master	B
1	[A⟩GM] guess: hello explanation: This is a common five-letter English word, and I am starting my guess with this word.		
2		[A⟨GM] guess_feedback: h<red> e<yellow> l<yellow> l<red> o<yellow>	
3	[A⟩GM] guess: world explanation: The feedback indicates that the letters "e" and "o" are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

hello

world

swirl

Example: image game



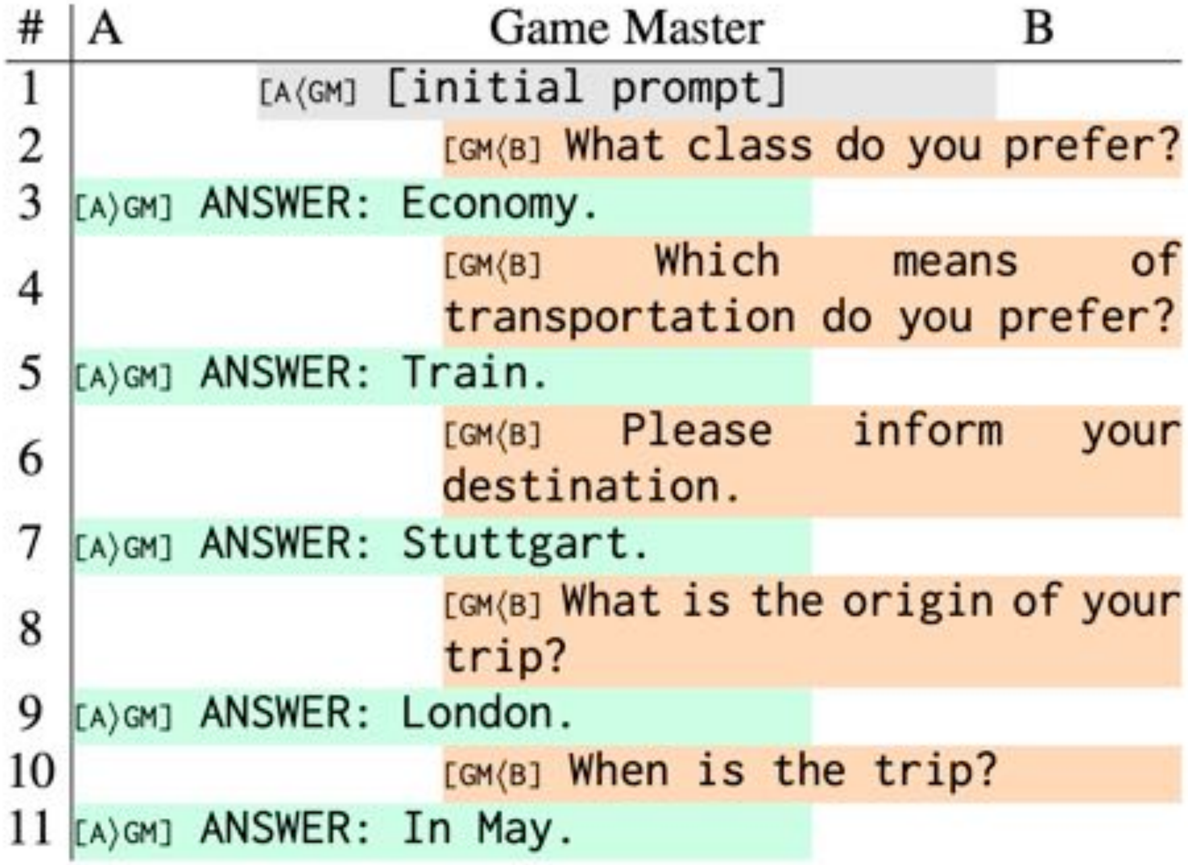
#	A	Game Master	B
1		[A⟨GM] [initial prompt...] □ □ □ □ □ □ □ □ □ □ □ □ □ □ □ L L L L L □ □ □ □ □ What is your next instruction?	
2	[A⟩GM] Instruction: Put L in the fourth row in all columns.		
3		[GM⟩B] [initial prompt...] Instruction: Put L in the fourth row in all columns.	
4			[GM⟨B] □ □ □ □ □ □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

(a) Successful gameplay from both players

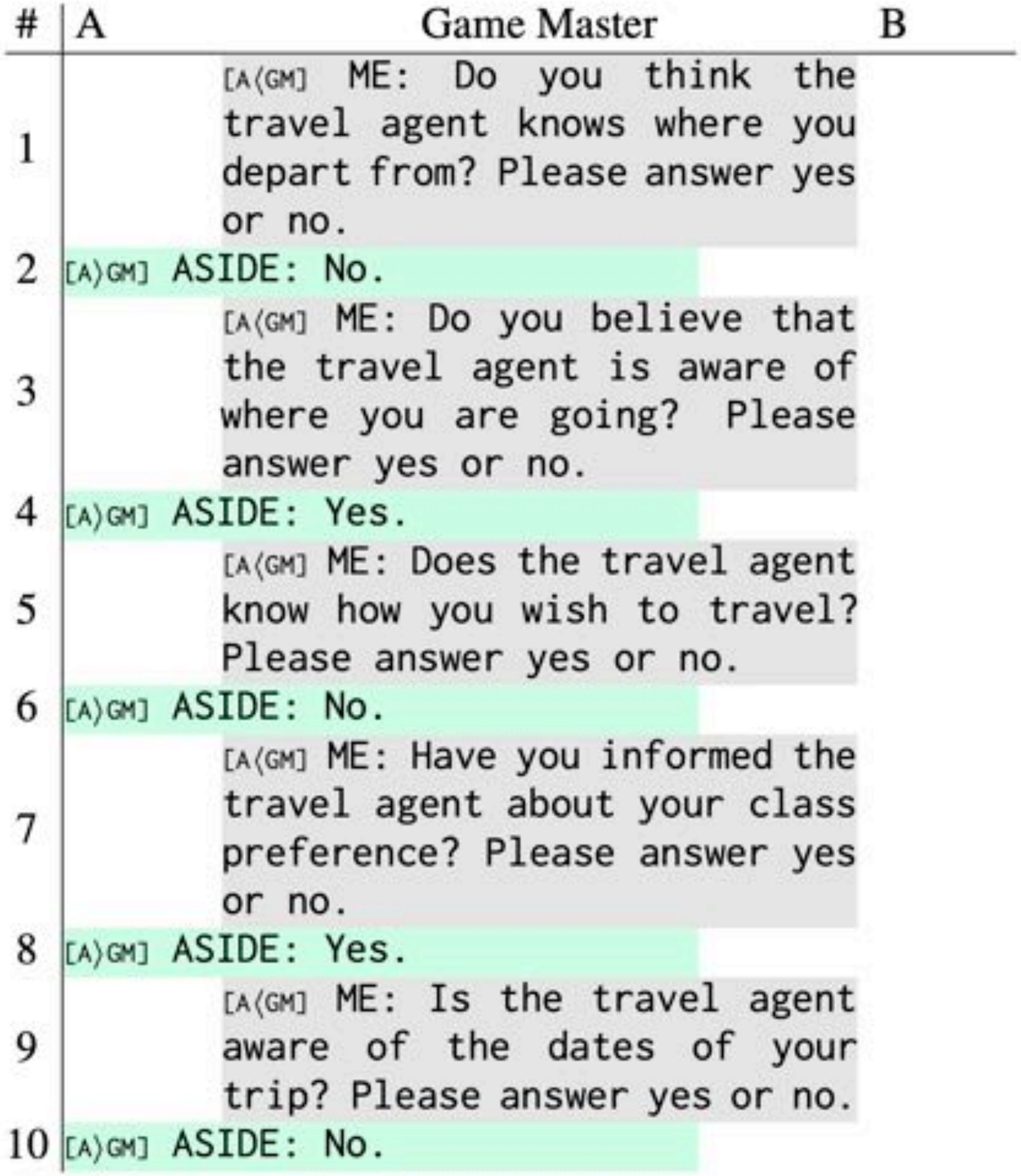
#	A	Game Master	B
1		[A⟨GM] [initial prompt...] □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ What is your next instruction?	
2	[A⟩GM] Instruction: Put a V in every cell of the second column.		
3		[GM⟩B] [initial prompt...] Instruction: Put a V in every cell of the second column.	
4			[GM⟨B] □ V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

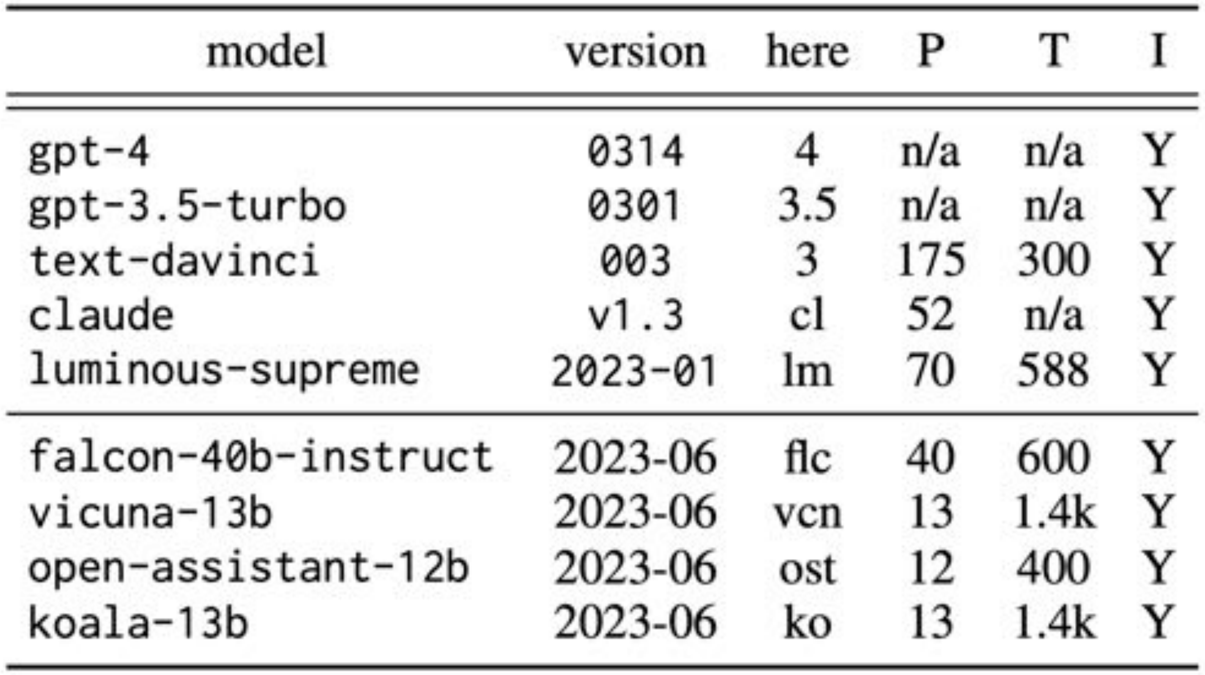


#	A	Game Master	B
1		[A⟨GM] [initial prompt]	
2			[GM⟨B] What class do you prefer?
3	[A⟩GM] ANSWER: Economy.		
4			[GM⟨B] Which means of transportation do you prefer?
5	[A⟩GM] ANSWER: Train.		
6			[GM⟨B] Please inform your destination.
7	[A⟩GM] ANSWER: Stuttgart.		
8			[GM⟨B] What is the origin of your trip?
9	[A⟩GM] ANSWER: London.		
10			[GM⟨B] When is the trip?
11	[A⟩GM] ANSWER: In May.		



#	A	Game Master	B
1		[A⟨GM] ME: Do you think the travel agent knows where you depart from? Please answer yes or no.	
2	[A⟩GM] ASIDE: No.		
3		[A⟨GM] ME: Do you believe that the travel agent is aware of where you are going? Please answer yes or no.	
4	[A⟩GM] ASIDE: Yes.		
5		[A⟨GM] ME: Does the travel agent know how you wish to travel? Please answer yes or no.	
6	[A⟩GM] ASIDE: No.		
7		[A⟨GM] ME: Have you informed the travel agent about your class preference? Please answer yes or no.	
8	[A⟩GM] ASIDE: Yes.		
9		[A⟨GM] ME: Is the travel agent aware of the dates of your trip? Please answer yes or no.	
10	[A⟩GM] ASIDE: No.		

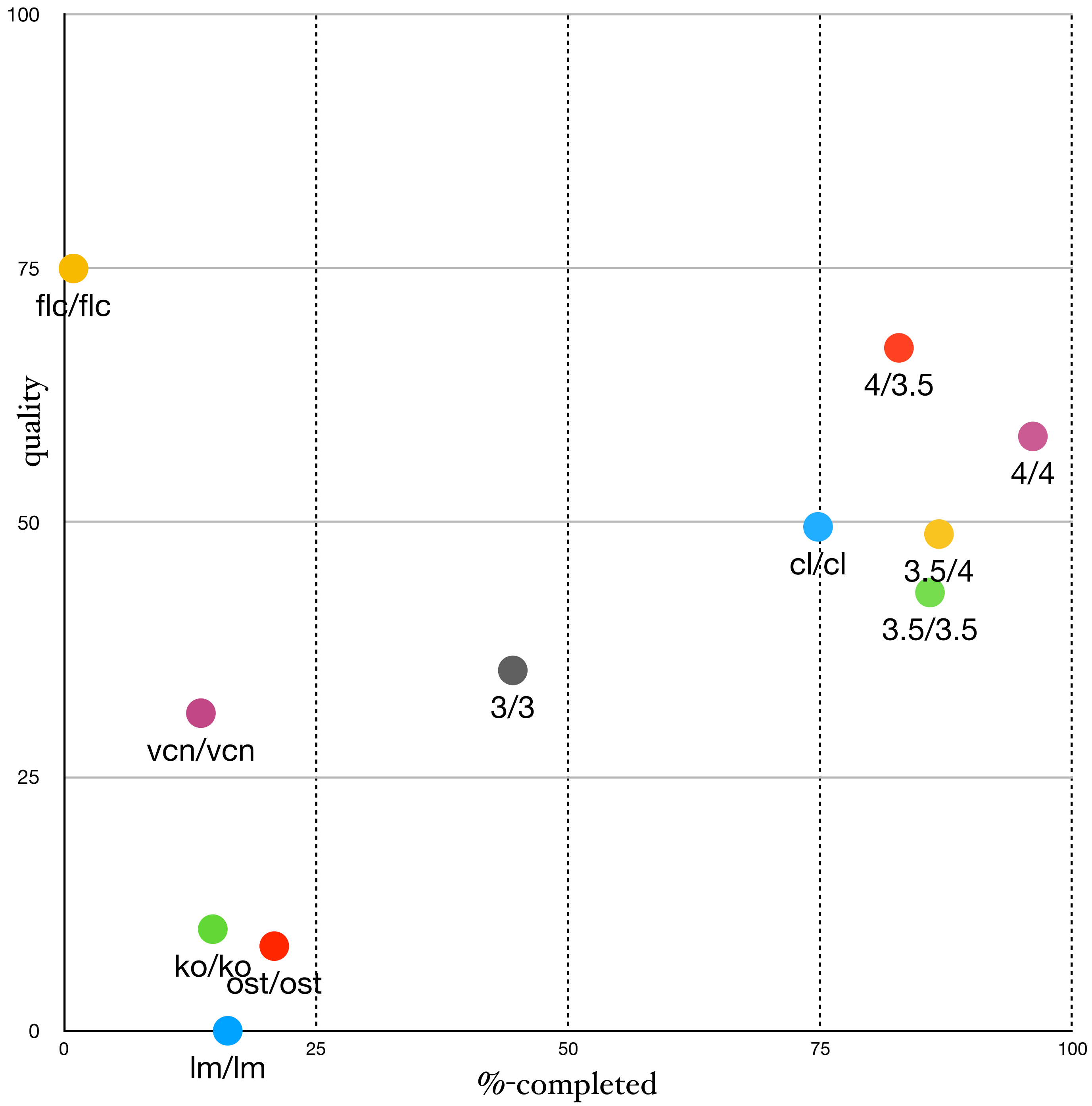
Running the benchmark

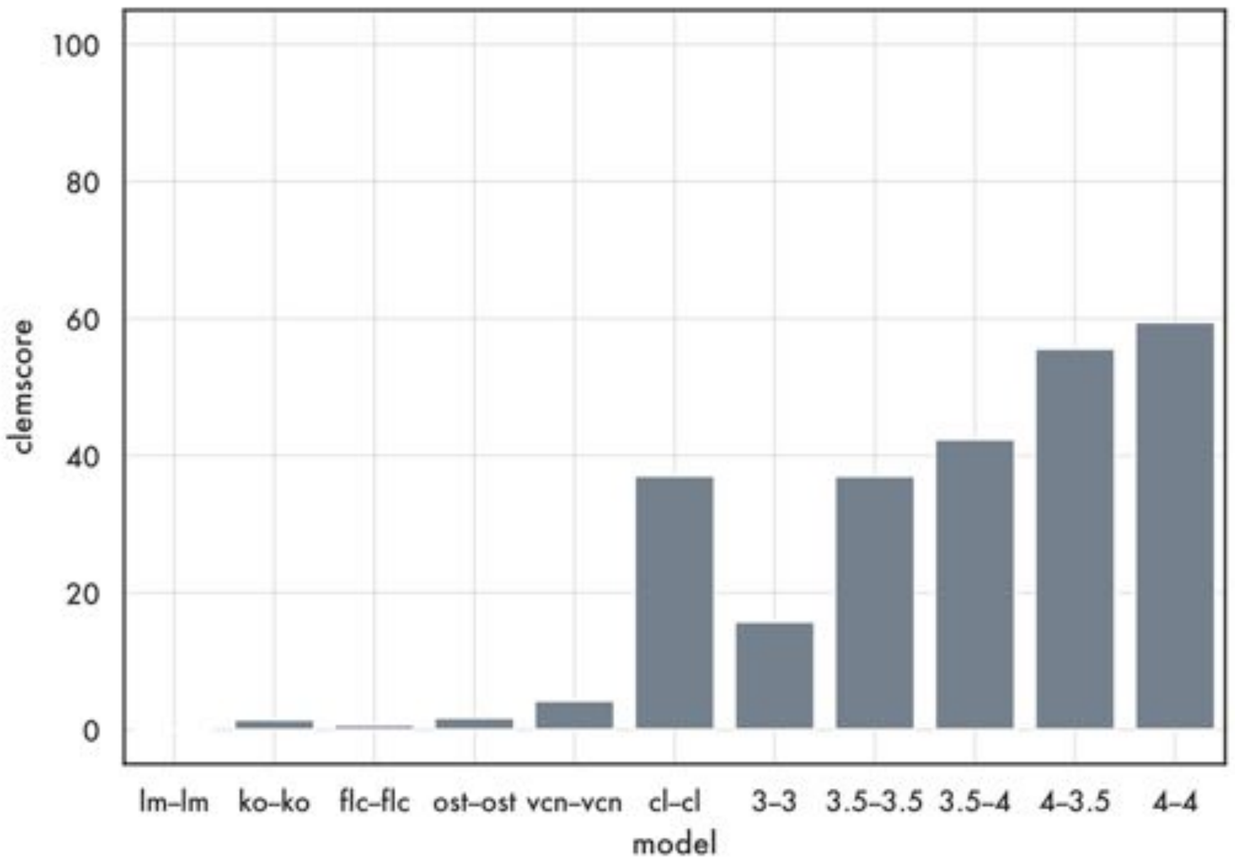


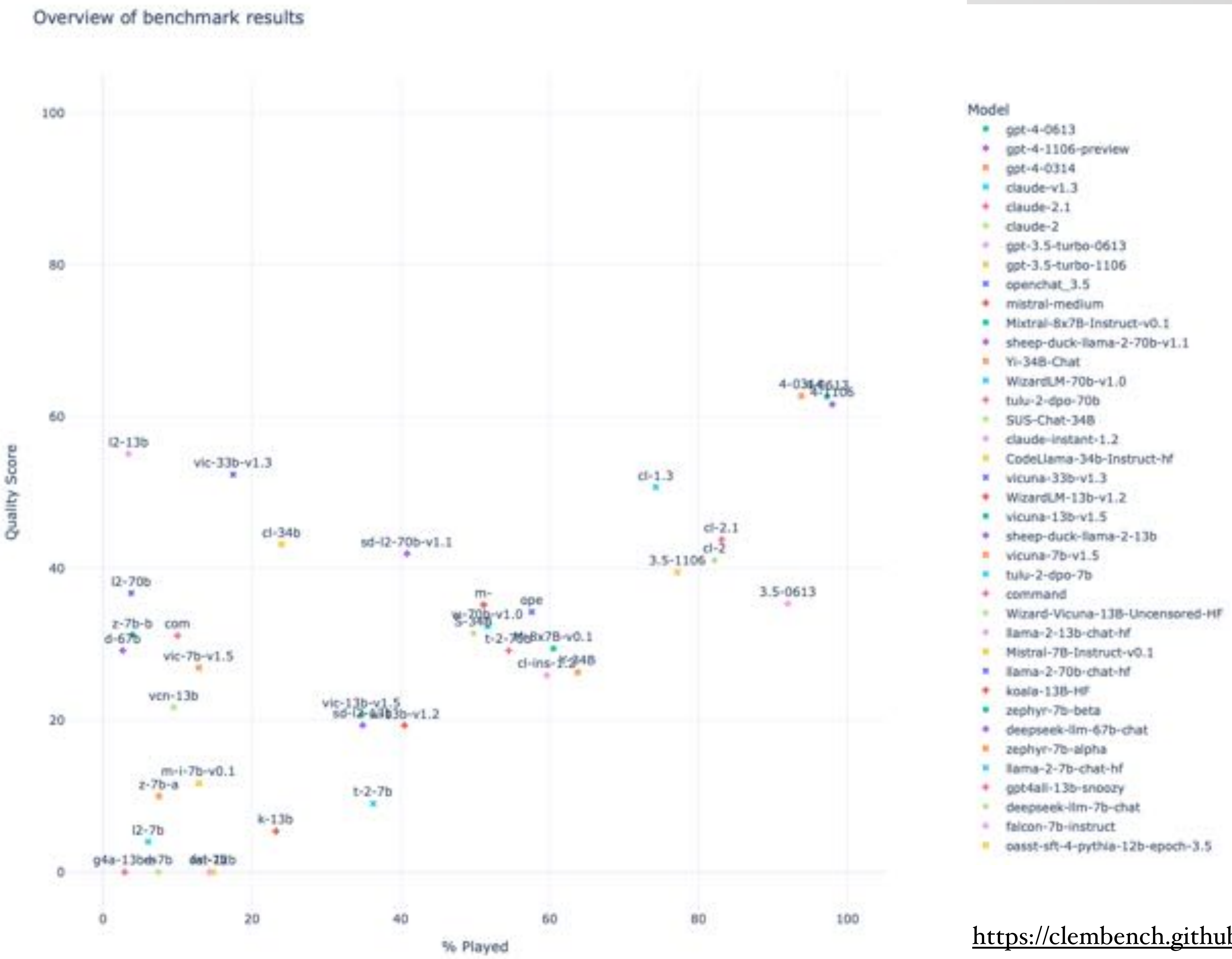
model	version	here	P	T	I
`gpt-4`	`0314`	4	n/a	n/a	Y
`gpt-3.5-turbo`	`0301`	3.5	n/a	n/a	Y
`text-davinci`	`003`	3	175	300	Y
`claude`	`v1.3`	cl	52	n/a	Y
`luminous-supreme`	`2023-01`	lm	70	588	Y
`falcon-40b-instruct`	2023-06	flc	40	600	Y
`vicuna-13b`	2023-06	vcn	13	1.4k	Y
`open-assistant-12b`	2023-06	ost	12	400	Y
`koala-13b`	2023-06	ko	13	1.4k	Y

- Evaluated for:
  - % games played to completion
    [following formal rules]
  - quality of game play (only completed games)
    [reaching game-specific goal]

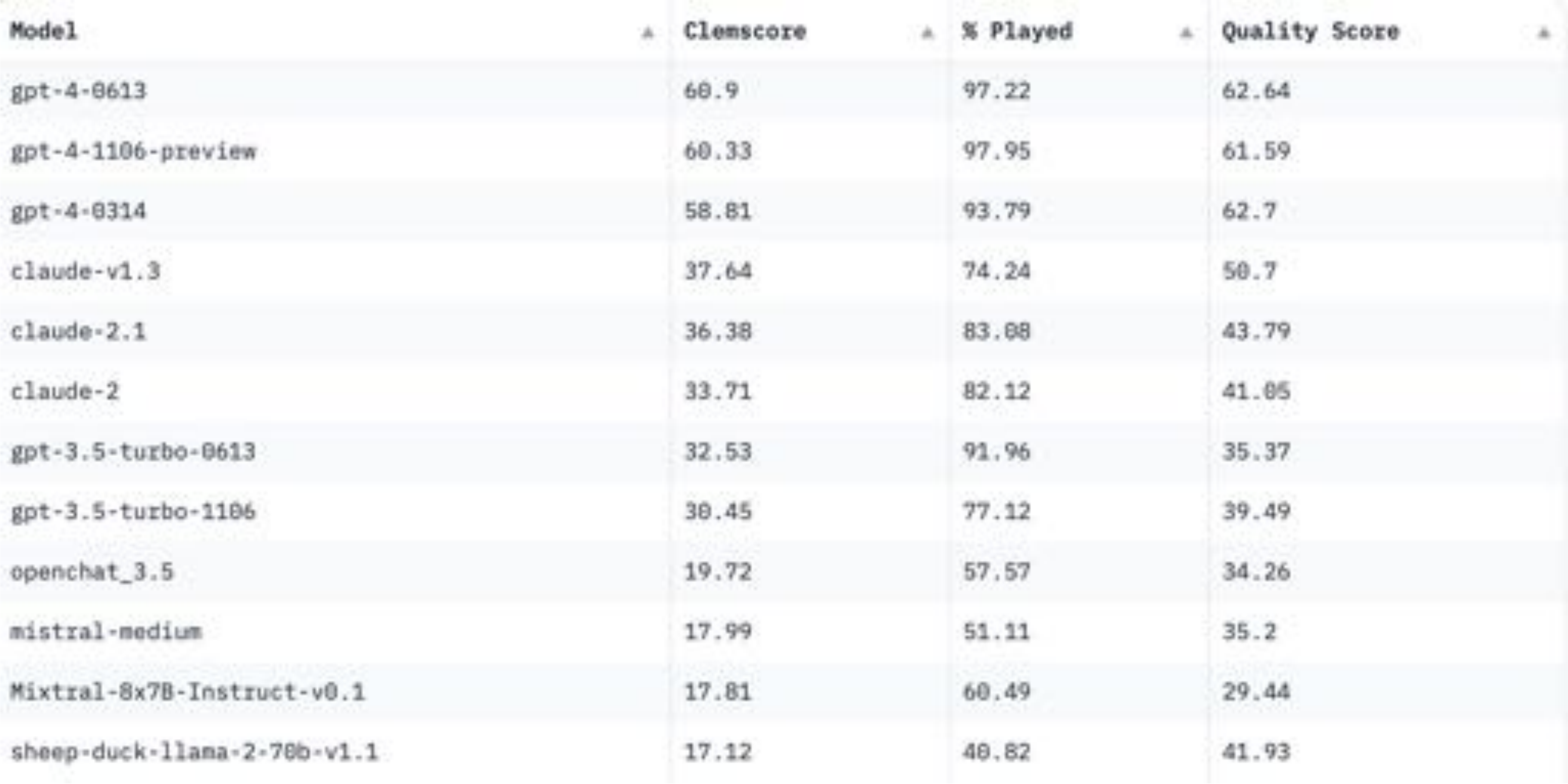
Running the benchmark





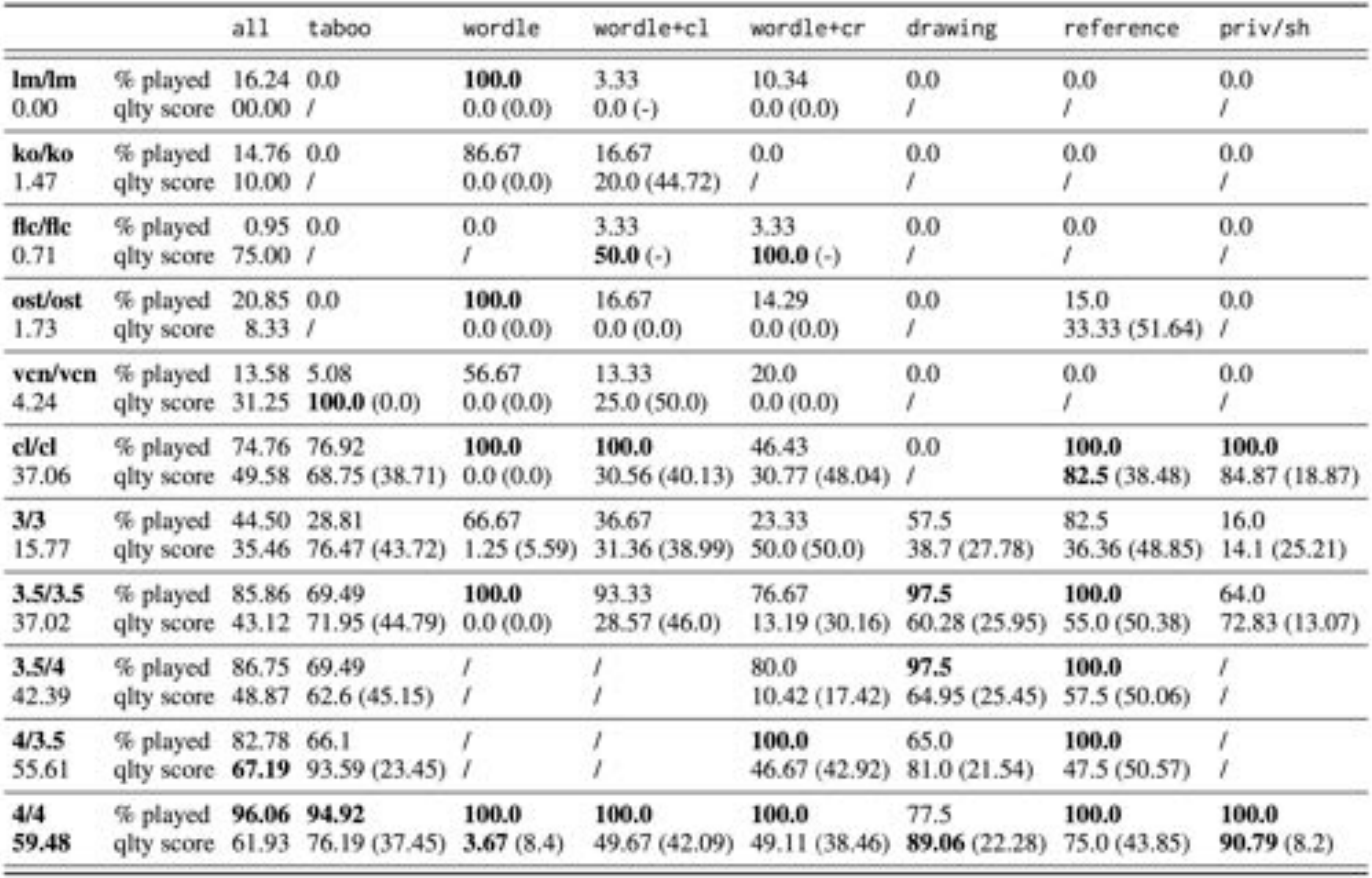


https://clembench.github.io/



Model	Clemscore	% Played	Quality Score
gpt-4-0613	60.9	97.22	62.64
gpt-4-1106-preview	60.33	97.95	61.59
gpt-4-0314	58.81	93.79	62.7
claude-v1.3	37.64	74.24	50.7
claude-2.1	36.38	83.08	43.79
claude-2	33.71	82.12	41.05
gpt-3.5-turbo-0613	32.53	91.96	35.37
gpt-3.5-turbo-1106	30.45	77.12	39.49
openchat_3.5	19.72	57.57	34.26
mistral-medium	17.99	51.11	35.2
Mixtral-8x7B-Instruct-v0.1	17.81	60.49	29.44
sheep-duck-llama-2-70b-v1.1	17.12	40.82	41.93

Running the benchmark



		all	taboo	wordle	wordle+cl	wordle+cr	drawing	reference	priv/sh
lm/lm	% played	16.24	0.0	**100.0**	3.33	10.34	0.0	0.0	0.0
0.00	qlty score	00.00	/	0.0 (0.0)	0.0 (-)	0.0 (0.0)	/	/	/
ko/ko	% played	14.76	0.0	86.67	16.67	0.0	0.0	0.0	0.0
1.47	qlty score	10.00	/	0.0 (0.0)	20.0 (44.72)	/	/	/	/
flc/flc	% played	0.95	0.0	0.0	3.33	3.33	0.0	0.0	0.0
0.71	qlty score	75.00	/	/	**50.0** (-)	**100.0** (-)	/	/	/
ost/ost	% played	20.85	0.0	**100.0**	16.67	14.29	0.0	15.0	0.0
1.73	qlty score	8.33	/	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)	/	33.33 (51.64)	/
vcn/vcn	% played	13.58	5.08	56.67	13.33	20.0	0.0	0.0	0.0
4.24	qlty score	31.25	**100.0** (0.0)	0.0 (0.0)	25.0 (50.0)	0.0 (0.0)	/	/	/
cl/cl	% played	74.76	76.92	**100.0**	**100.0**	46.43	0.0	**100.0**	**100.0**
37.06	qlty score	49.58	68.75 (38.71)	0.0 (0.0)	30.56 (40.13)	30.77 (48.04)	/	**82.5** (38.48)	84.87 (18.87)
3/3	% played	44.50	28.81	66.67	36.67	23.33	57.5	82.5	16.0
15.77	qlty score	35.46	76.47 (43.72)	1.25 (5.59)	31.36 (38.99)	50.0 (50.0)	38.7 (27.78)	36.36 (48.85)	14.1 (25.21)
3.5/3.5	% played	85.86	69.49	**100.0**	93.33	76.67	**97.5**	**100.0**	64.0
37.02	qlty score	43.12	71.95 (44.79)	0.0 (0.0)	28.57 (46.0)	13.19 (30.16)	60.28 (25.95)	55.0 (50.38)	72.83 (13.07)
3.5/4	% played	86.75	69.49	/	/	80.0	**97.5**	**100.0**	/
42.39	qlty score	48.87	62.6 (45.15)	/	/	10.42 (17.42)	64.95 (25.45)	57.5 (50.06)	/
4/3.5	% played	82.78	66.1	/	/	**100.0**	65.0	**100.0**	/
55.61	qlty score	**67.19**	93.59 (23.45)	/	/	46.67 (42.92)	81.0 (21.54)	47.5 (50.57)	/
4/4	% played	**96.06**	**94.92**	**100.0**	**100.0**	**100.0**	77.5	**100.0**	**100.0**
59.48	qlty score	61.93	76.19 (37.45)	**3.67** (8.4)	49.67 (42.09)	49.11 (38.46)	**89.06** (22.28)	75.0 (43.85)	**90.79** (8.2)

What does it all mean?

- Still room to grow. (Compared to human / human play.)
- Performance seems to increase as a function of size (parameter count, training data, instruction tuning)
- Performance of better models not bad even for image game, scorekeeping
- Pure wordle very hard, even for GPT4
- Open models not quite there yet, but are rapidly improving

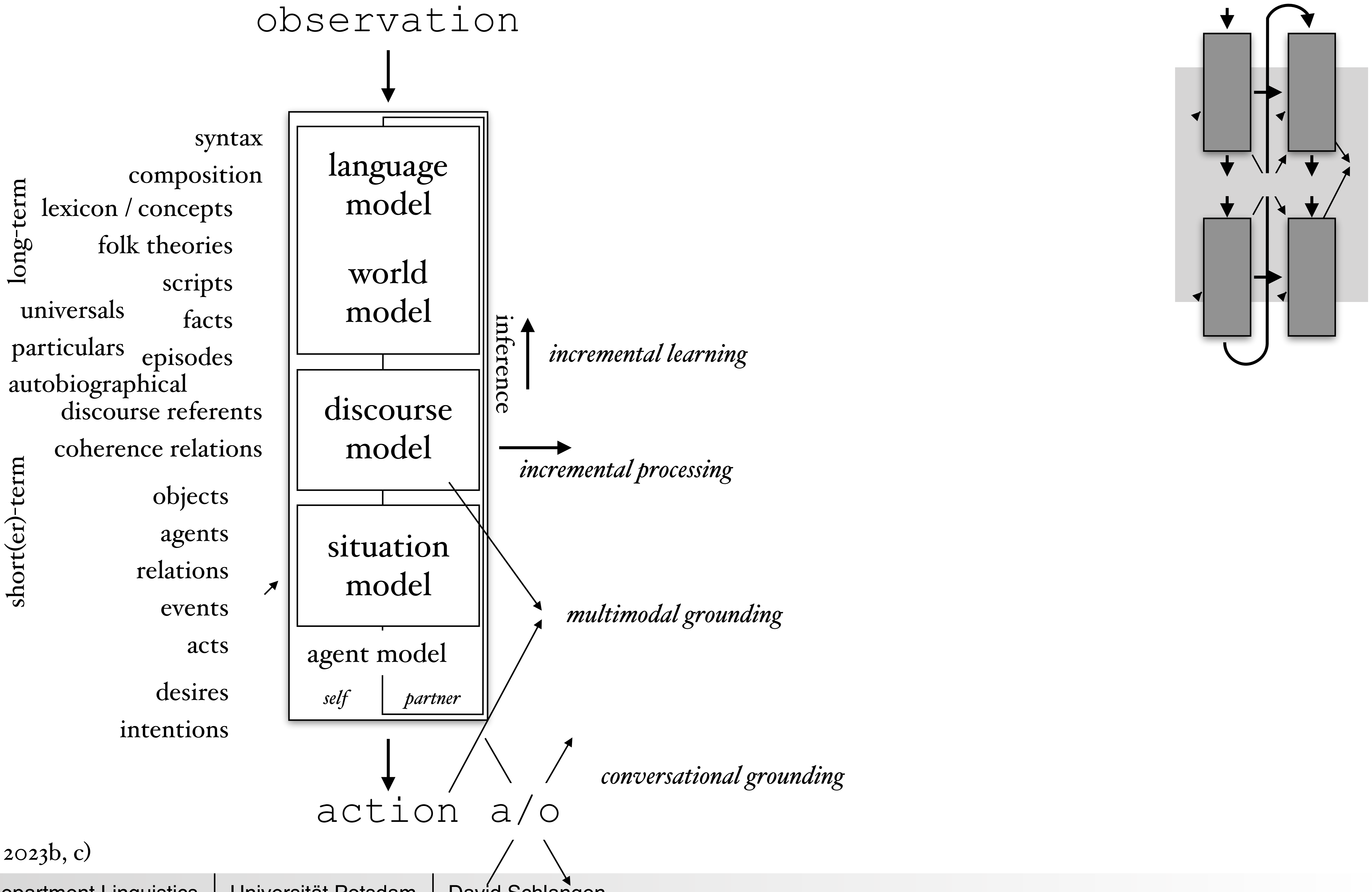
What does it all mean?

- But what did the scalpel dissect?
  - At the moment, a bit hard to see trees for the forrest…
  - More fine-grained analyses to come. (But see paper.)
    - E.g., explanations in wordle mostly not consistent with decision made by model.

Where to?

- This was just a proof-of-concept of the instrument
- Much remains to be done:
    - Deeper analysis of performance on games
    - Additional games (e.g., more modalities; multi-linguality)
    - Correlation analysis, to validate assumed connection to construct
    - Clearer argument for correlation btw game performance and (useful) task performance
    - Investigate performance as function of model parameters (checkpoint, parameter size, etc.): Is performance linear? "Emergence"? Etc.
    - Continuous testing… new models to test every day…
- Open source — easy to get involved! https://github.com/clp-research/clembench

Situated Language Understanding



observation

language model

world model

discourse model

situation model

agent model

self

partner

long-term

syntax

composition

lexicon / concepts

folk theories

scripts

universals

facts

particulars

episodes

autobiographical

discourse referents

coherence relations

short(er)-term

objects

agents

relations

events

acts

desires

intentions

inference

incremental learning

incremental processing

multimodal grounding

conversational grounding

action a/o

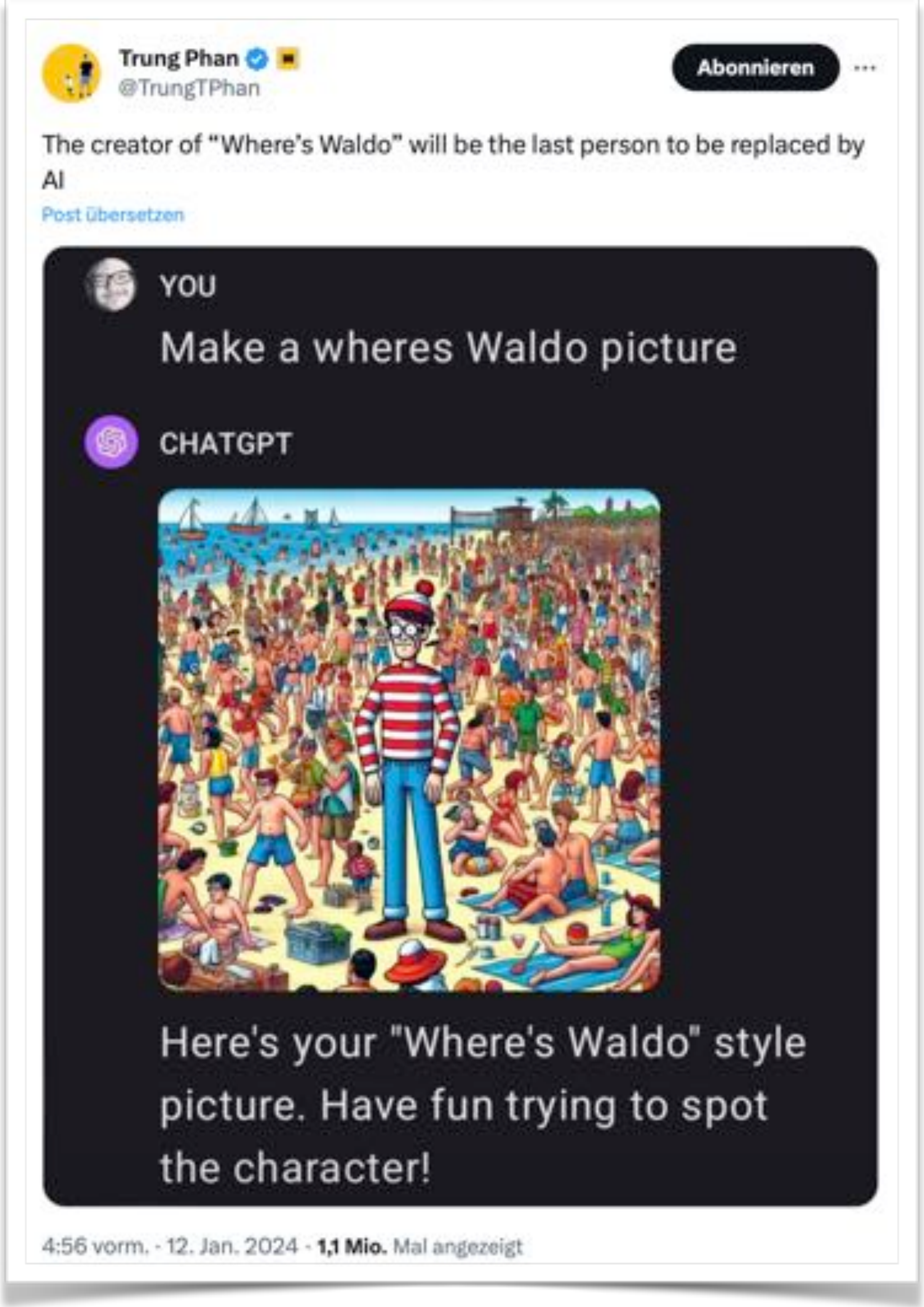
(Schlangen 2023b, c)

Situated Interaction

(Zarrieß *et al.*, LREC 2016)

Learning from self-play

- select transcripts of good (self-play) games
- mix with h/h & h/s episodes
- add as fine-tuning data
- test for transfer to other games



Trung Phan
@TrungTPhan

The creator of "Where's Waldo" will be the last person to be replaced by AI

YOU

Make a wheres Waldo picture

CHATGPT

Here's your "Where's Waldo" style picture. Have fun trying to spot the character!

4:56 vorm. · 12. Jan. 2024 · **1,1 Mio.** Mal angezeigt

Conclusions

- A Theoretical Thread
  - NLU is not solved
  - And in any case, coverage not complete: Situated-LU
  - A model of the construct: “Beliefs & Processes”
- A Methodological Thread
  - Negative: How *not* To Do Things
    - Text NLU, *face & construct validity*
  - Positive: How To Do Things
    - Dialogue Games & the construct *situated understanding*
    - Testing chat-optimised LLMs as Interactive Agents

Conclusions

- A Theoretical Thread
  - NLU is not solved
  - And in any case, coverage not complete: Situated-LU
  - A model of the construct: “Beliefs & Processes”
- A Methodological Thread
  - Negative: How *not* To Do Things
    - Text NLU, *face & construct validity*
  - Positive: How To Do Things
    - Dialogue Games & the construct *situated understanding*
    - Testing chat-optimised LLMs as Interactive Agents

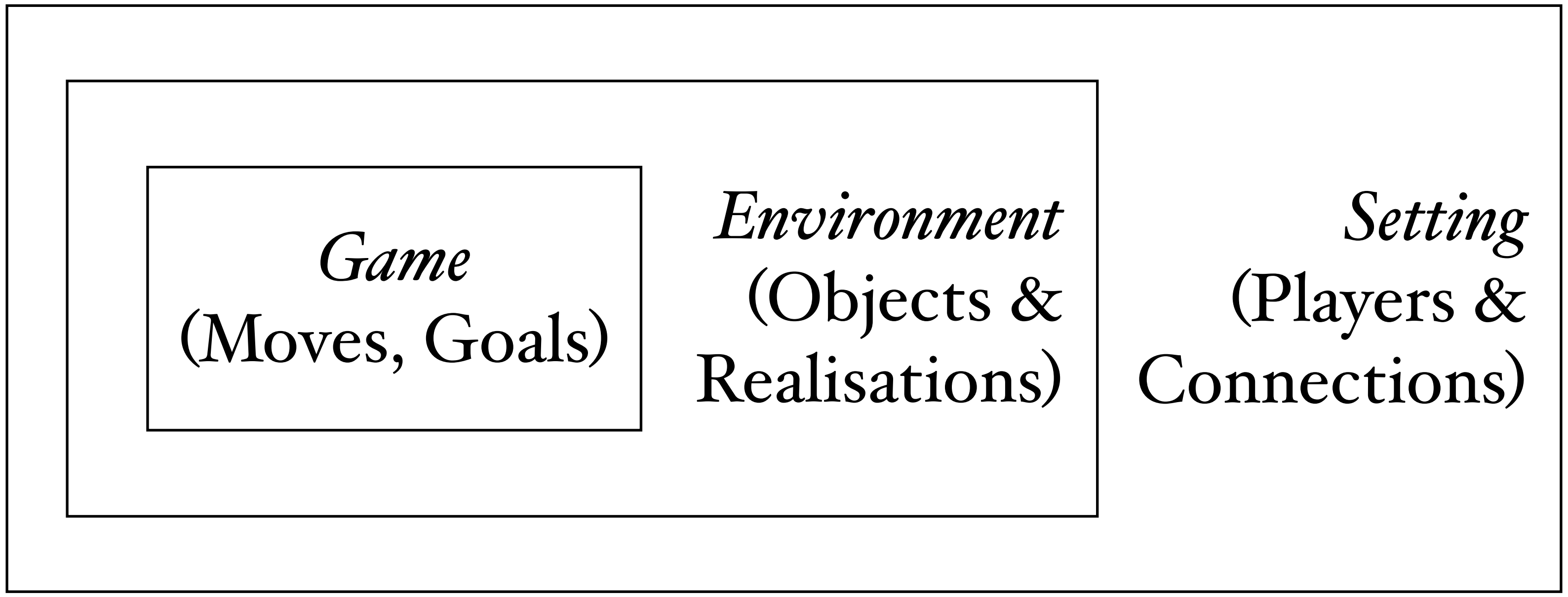
Thank you.

Questions, Comments?

Acknowledgements: Many thanks to my current & former grad students (https://clp.ling.uni-potsdam.de/people/) & colleagues w/ whom I have discussed related ideas in recent years.

Gratefully acknowledged: Funding by DFG (project “RECOLAGE”; CRC “Limits of Variability”, project B06); BMBF (project “COCOBOTS”)

Dialogue Game Taxonomy



Game
(Moves, Goals)

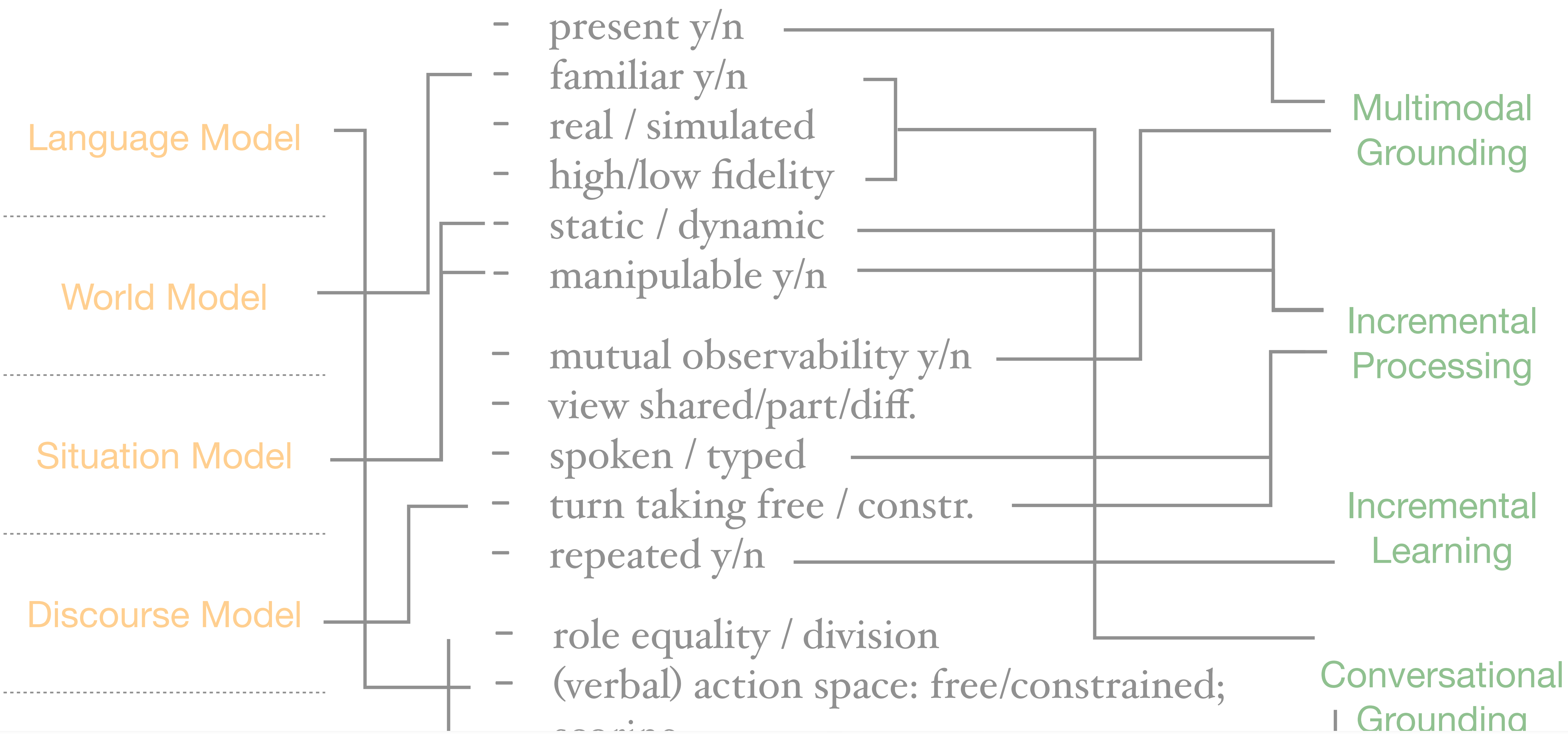
Environment
(Objects & Realisations)

Setting
(Players & Connections)

multiple dimensions for transfer learning / generalisation

- Our proposal: A fine-grained taxonomy of dialogue games, with clear connections to KD&P model.

Game & BD/AP



Language Model

World Model

Situation Model

Discourse Model

- present y/n
- familiar y/n
- real / simulated
- high/low fidelity
- static / dynamic
- manipulable y/n
- mutual observability y/n
- view shared/part/diff.
- spoken / typed
- turn taking free / constr.
- repeated y/n
- role equality / division
- (verbal) action space: free/constrained;

Multimodal Grounding

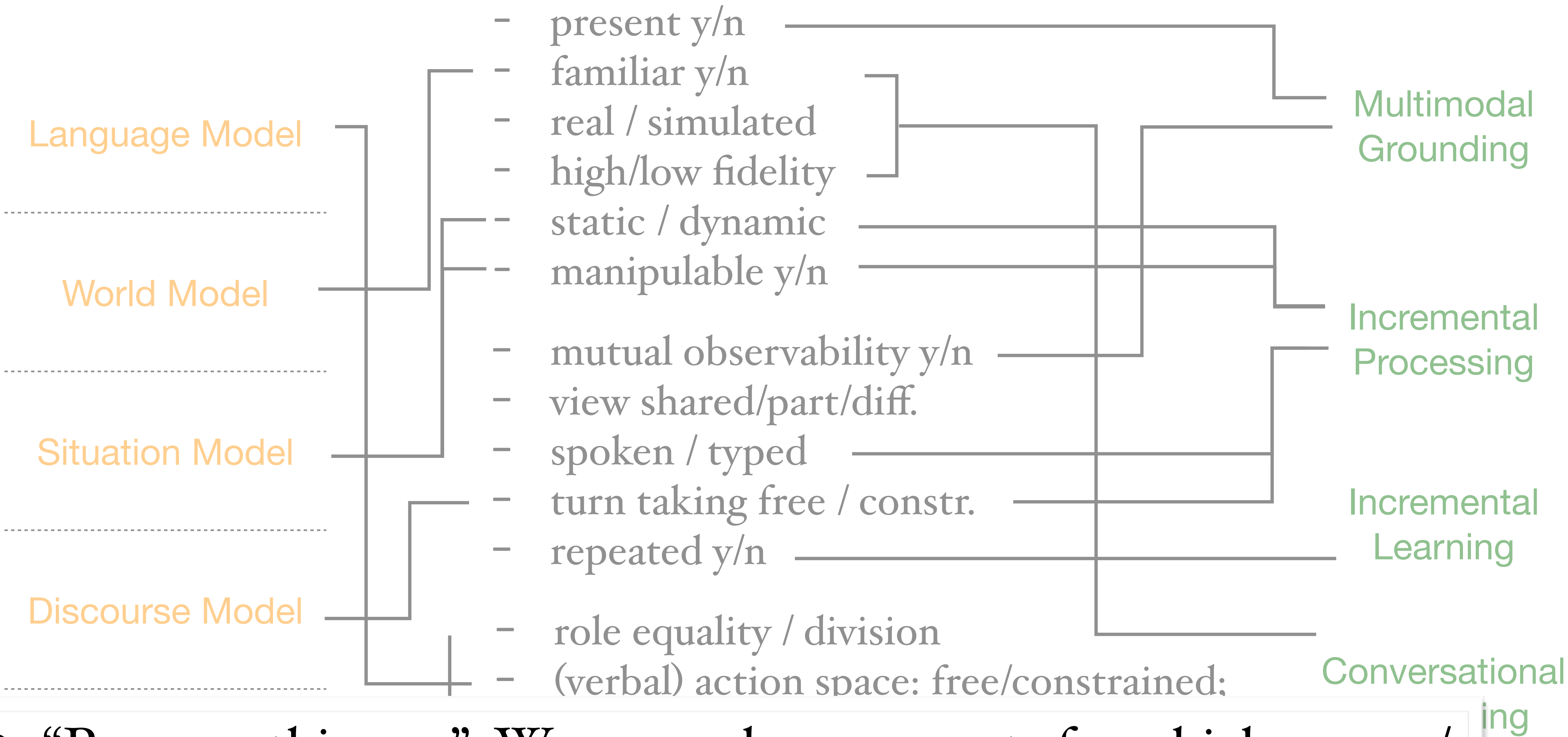
Incremental Processing

Incremental Learning

Conversational Grounding

- **Our proposal:** A fine-grained taxonomy of dialogue games, **with clear connections to BD/AP model.**

Game & BD/AP



Language Model

World Model

Situation Model

Discourse Model

- present y/n
- familiar y/n
- real / simulated
- high/low fidelity
- static / dynamic
- manipulable y/n
- mutual observability y/n
- view shared/part/diff.
- spoken / typed
- turn taking free / constr.
- repeated y/n
- role equality / division
- (verbal) action space: free/constrained;

Multimodal Grounding

Incremental Processing

Incremental Learning

Conversational

ing

- “Progress this way”: We can make arguments for which games / environments / settings are less restricted than others. Ultimate goal: real world / real language games.

List of References for the Talk “Understanding Understanding” (2024/2 version)

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

References

Beyer, Anne, Sharid Loáiciga, and David Schlangen (2021). “Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4164–4173.

Bowles, Samuel and Herbert Gintis (2011). *A Cooperative Species: Human Reciprocity and its Evolution*. Princeton University Press.

Bratman, Michael E. (1987). *Intentions, Plans, And Practical Reason*. Cambridge, Massachusetts, USA: Harvard University Press.

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.

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.

Dunietz, Jesse, Greg Burnham, Akash Bharadwaj, Owen Rambow, Jennifer Chu-Carroll, and Dave Ferrucci (July 2020). “To Test Machine Comprehension, Start by Defining Comprehension”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7839–7859.

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.

Fernández, Raquel, Tatjana Lucht, Kepa Rodríguez, and David Schlangen (2006). “Interaction in Task-Oriented Human–Human Dialogue: The Effects of Different Turn-Taking Policies”. In: *Proceedings of the First International IEEE/ACL Workshop on Spoken Language Technology*. Palm Beach, Aruba.

Galetzka, Fabian, Chukwuemeka Uchenna Eneh, and David Schlangen (May 2020). “A Corpus of Controlled Opinionated and Knowledgeable Movie Discussions for Training Neural Conversation Models”. English. In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 565–573.

Galetzka, Fabian, Jewgeni Rose, David Schlangen, and Jens Lehmann (Aug. 2021). “Space Efficient Context Encoding for Non-Task-Oriented Dialogue Generation with Graph Attention Transformer”. In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and* *the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Online: Association for Computational Linguistics, pp. 7028–7041.

Götze, Jana, Maike Paetzel-Prüsmann, Wencke Liermann, Tim Diekmann, and David Schlangen (June 2022). "The Slurk Interaction Server Framework: Better Data for Better Dialog Models". In: *Proceedings of the Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, pp. 4069–4078.

Götze, Jana, Karla Friedrichs, and David Schlangen (2022). "Interactive and Cooperative Delivery of Referring Expressions: A Comparison of Three Algorithms". In: *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*. Virtually and at Dublin, Ireland: SEMDIAL.

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.

Johnson-Laird, Philip Nicholas (1983). *Mental Models: Towards a Cognitive Science of Language, Inference, and Consciousness*. Cognitive science series. Harvard University Press.

Kahardipraja, Patrick, Brielen Madureira, and David Schlangen (Nov. 2021). "Towards Incremental Transformers: An Empirical Analysis of Transformer Models for Incremental NLU". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 1178–1189.

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.

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 (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.

Loáiciga, Sharid, Anne Beyer, and David Schlangen (Oct. 2022). "New or Old? Exploring How Pre-Trained Language Models Represent Discourse Entities". In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 875–886.

Madureira, Brielen and David Schlangen (Nov. 2020). "Incremental Processing in the Age of Non-Incremental Encoders: An Empirical Assessment of Bidirectional Models for Incremental NLU". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 357–374.

— (May 2022). "Can Visual Dialogue Models Do Scorekeeping? Exploring How Dialogue Representations Incrementally Encode Shared Knowledge". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 651–664.

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.

Ribeiro, Marco Tulio, Tongshuang Wu, Carlos Guestrin, and Sameer Singh (July 2020). "Beyond Accuracy: Behavioral Testing of NLP Models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4902–4912.

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`.

Schuster, Sebastian and Tal Linzen (July 2022). "When a sentence does not introduce a discourse entity, Transformer-based models still sometimes refer to it". In: *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Seattle, United States: Association for Computational Linguistics, pp. 969–982.

Sellars, Wilfried (1954). "Some Reflections on Language Games". In: *Philosophy of Science* 21, pp. 204–228.

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]`.

Wang, Alex, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *NeurIPS*. July, pp. 1–30. arXiv: `1905.00537`.

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.