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`

@ DFKI Berlin Colloquium
2023-07-20

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

Structure

- Part I: “Understanding Understanding”
    - Our Research Programme
    - Some Systematic Insights
    - Some Recent Projects

- Part II: Probing Chat-Optimized LLMs Through Gameplay
    - Games & Capabilities
    - `clemgame`
    - `clembench`

2000 — 2003		formal **dialogue** pragmatics	SDRT, theorem proving, probabilistic models
2007 — 2012	DFG (ENP)	Inpro: **incrementali**zing the representation building process (ASR, NLU, DM, NLG, TTS; turn-taking)	HMMs, SVMs, prob. parsing, …
2012 — 2016		Inpro II: situated **incremental** processing (vision, gaze, gestures, embodiment)	graphical models
2015 — 2017	DFG/ ANR	DUEL: **incremental** disfluency detection	RNNs
2016 — 2019		concept learning / **grounded** lexical semantics	CNNs, RNNs
2018 — 2022	VW	knowledge-**grounded** closed-domain chat	LLMs + KGs
2019 —		MetaNLP / Methodology / The BIG Picture	words
2019 — 2023		Neur-**Inpro** & State Tracking	RNNs, transformers, LLMs
2019 — 2024	DFG	RECOLAGE: **incremental** instruction generation	RL, symb <-> NN transf.
2020 — 2024	SFB 1287	Limits of Pragmatic Variability in LLMs	LLMs
2022 — 2025	BMBF	COCOBOTS: construction concept learning, robot arm	code gen models
2023 —		Foundation Models: What Are They Good For?	LLMs
2024 —		Reasonable AI: Giving and Asking for Reasons	

2000 — 2003		formal **dialogue** pragmatics	SDRT, theorem proving, probabilistic models
2007 — 2012	DFG (ENP)	Inpro: **incremental**izing the representation building process (ASR, NLU, DM, NLG, TTS; turn-taking)	HMMs, SVMs, prob. parsing, …
2012 — 2016		Inpro II: situated **incremental** processing (vision, gaze, gestures, embodiment)	graphical models
2015 — 2017	DFG/ ANR	DUEL: **incr**[illegible]	[illegible]
2016 — 2019		concept learning / **grounded** lexical semantics	CNNs, RNNs
2018 — 2022	[illegible]	[illegible] omain chat	LLMs + KGs
2019 —		[illegible]	[illegible]
2019 — 2023		Neur-**Inpro** & State Tracking	RNNs, transformers, LLMs
2019 — 2024	DFG	RECOLAGE: **incremental** instruction generation	RL, symb[illegible]N transf.
2020 — 2024	SFB 1287	[illegible] Variability in LLMs	LLMs
2022 — 2025	BMBF	COCOBOTS: construction concept learning, robot arm	code gen models
2023 —		Foundation Models: What Are They Good For?	LLMs
2024 —		Reasonable AI: Giving and Asking for Reasons	

foundational research

applications in UI

domains (multimodal assistance, tutoring, …)

Understanding *Understanding* —

the process by which interlocutors create shared understanding (sufficient for current purposes)

computer vision

TTS

ASR

lexical semantics

robotics

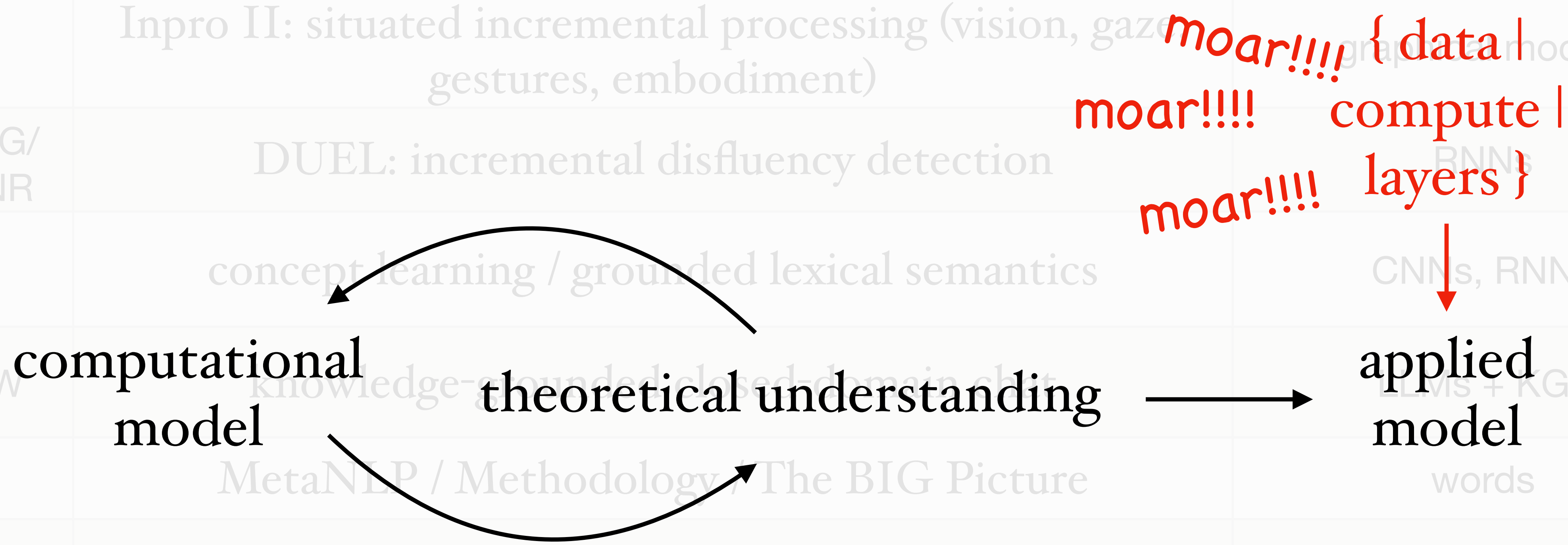
Computational Linguistics

Representation Formalisation

2000 — 2003		formal dialogue pragmatics	SDRT, theorem proving
2007 — 2012	DFG (ENP)	Inpro: incrementalizing the representation building process (ASR, NLU, DM, NLG, TTS; turn-taking)	HMMs, SVMs, prob. parsing, …
2012 — 2016		Inpro II: situated incremental processing (vision, gaze, gestures, embodiment)	graphical models
2015 — 2017	DFG/ ANR	DUEL: incremental disfluency detection	RNNs
2016 — 2019		concept learning / grounded lexical semantics	CNNs, RNNs
2018 — 2022	VW	knowledge-grounded closed-domain chat	LLMs + KGs
2019 —		MetaNLP / Methodology / The BIG Picture	words
2019 — 2023		Neur-Inpro & State Tracking	RNNs, transformers, LLMs
2019 — 2024	DFG	RECOLAGE: incremental instruction generation	RL, symb <-> NN transf.
2020 — 2024	SFB 1287	Limits of Pragmatic Variability in LLMs	LLMs
2022 — 2025	BMBF	COCOBOTS: construction concept learning, robot arm	code gen models
2023 —		Foundation Models: What Are They Good For?	
2024 —		Reasonable AI: Giving and Asking for Reasons	

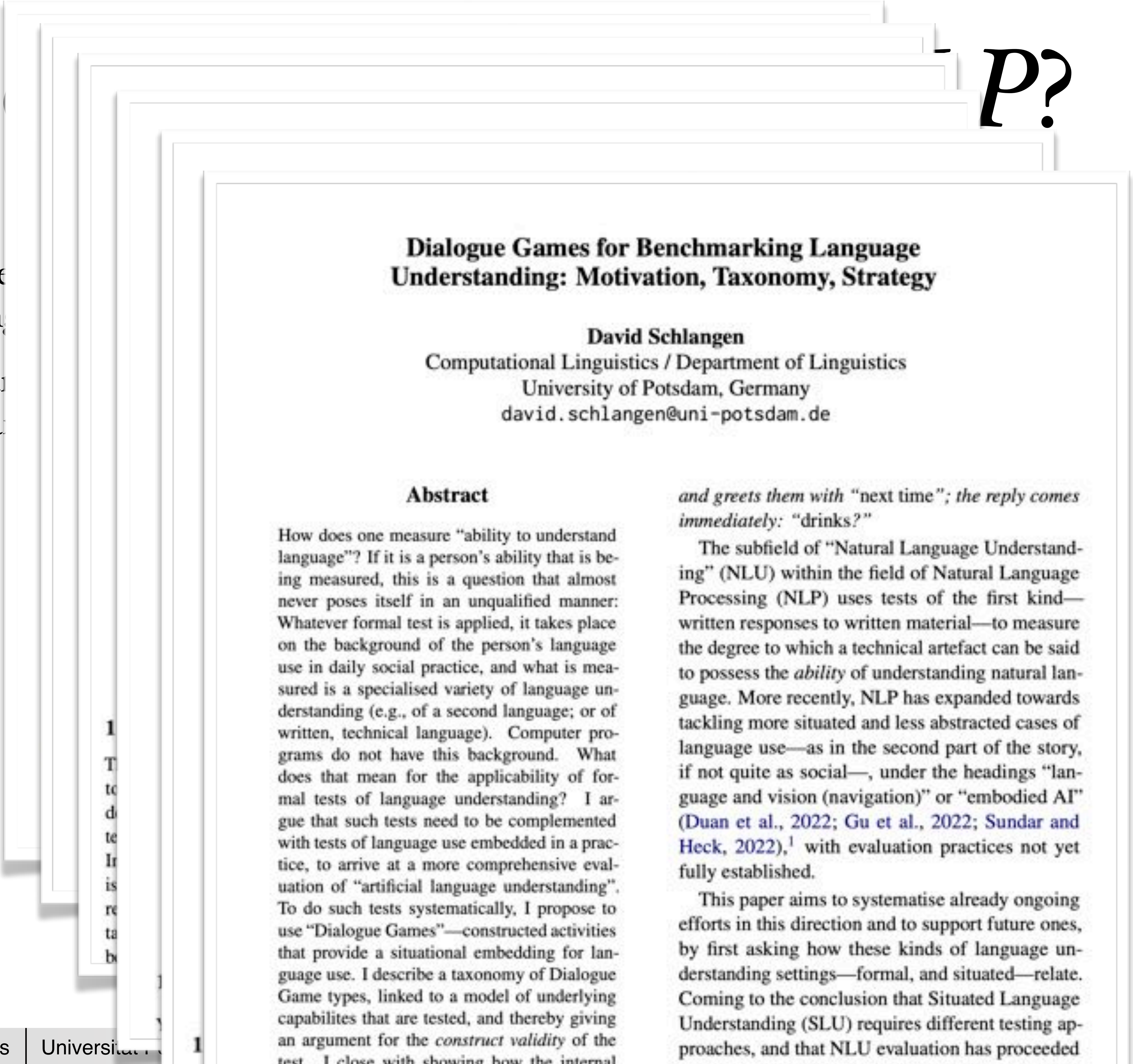
Representation Learning

NLP



Wh P?

- Always bee engineerin
- 2019, beca to think (a



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

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

Abstract

How does one measure "ability to understand language"? If it is a person's ability that is being measured, this is a question that almost never poses itself in an unqualified manner: Whatever formal test is applied, it takes place on the background of the person's language use in daily social practice, and what is measured is a specialised variety of language understanding (e.g., of a second language; or of written, technical language). Computer programs do not have this background. What does that mean for the applicability of formal tests of language understanding? I argue that such tests need to be complemented with tests of language use embedded in a practice, to arrive at a more comprehensive evaluation of "artificial language understanding". To do such tests systematically, I propose to use "Dialogue Games"—constructed activities that provide a situational embedding for language use. I describe a taxonomy of Dialogue Game types, linked to a model of underlying capabilites that are tested, and thereby giving an argument for the *construct validity* of the test. I close with showing how the internal

and greets them with "next time"*; the reply comes immediately:* "drinks*?"*

The subfield of "Natural Language Understanding" (NLU) within the field of Natural Language Processing (NLP) uses tests of the first kind—written responses to written material—to measure the degree to which a technical artefact can be said to possess the *ability* of understanding natural language. More recently, NLP has expanded towards tackling more situated and less abstracted cases of language use—as in the second part of the story, if not quite as social—, under the headings "language and vision (navigation)" or "embodied AI" (Duan et al., 2022; Gu et al., 2022; Sundar and Heck, 2022),[1] with evaluation practices not yet fully established.

This paper aims to systematise already ongoing efforts in this direction and to support future ones, by first asking how these kinds of language understanding settings—formal, and situated—relate. Coming to the conclusion that Situated Language Understanding (SLU) requires different testing approaches, and that NLU evaluation has proceeded

Where's the *theory of NLP*?

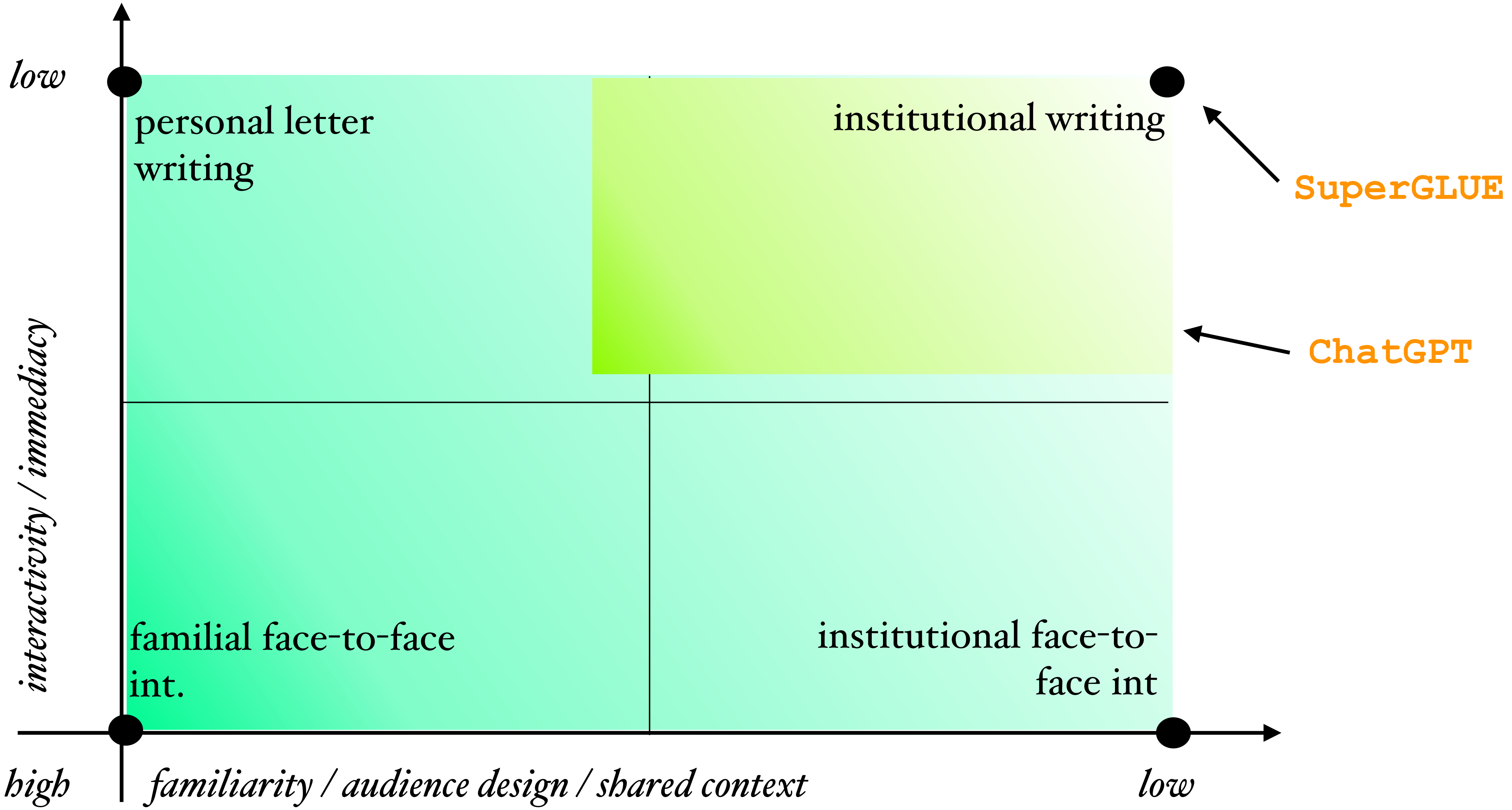
- Always been puzzled by lack of coherence. (Science when feeling fancy, engineering when challenged…)
- 2019, became "Professor of the Foundations of CL" — I'm now officially allowed to think (and write) about this!
    - (Schlangen 2019a, 2019b, 2021 ACL, 2022, 2023a, 2023b)

Where’s the *theory of NLP*?

- Some points
  - now that things kind of work, kind of in a general way, the domain theory aversion is coming back to haunt NLP, as problems with *measurement*
  - the type of language use represented by NLP-NLU is not the only one, and not even the paradigmatic one
  - it makes sense to analyse the task of understanding as
    - requiring *knowledge*,
    - which is applied in (update) *processes*,
      - some of which are single-minded ones, other are interactive ones
  - this makes clearer the relation btw NLP-NLU & SLU (as tasks, and methodologically)

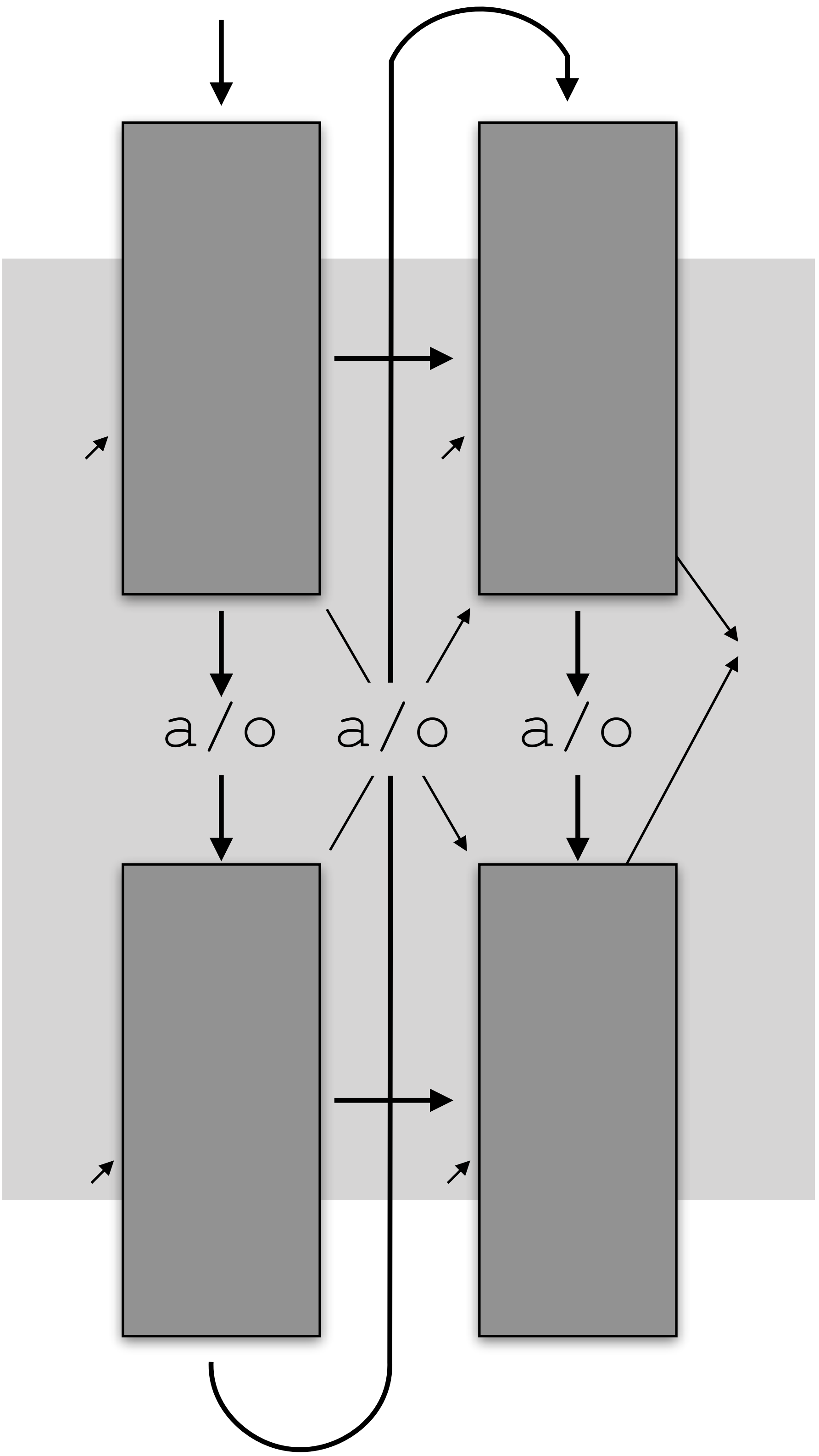
I. Types of Language Use

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

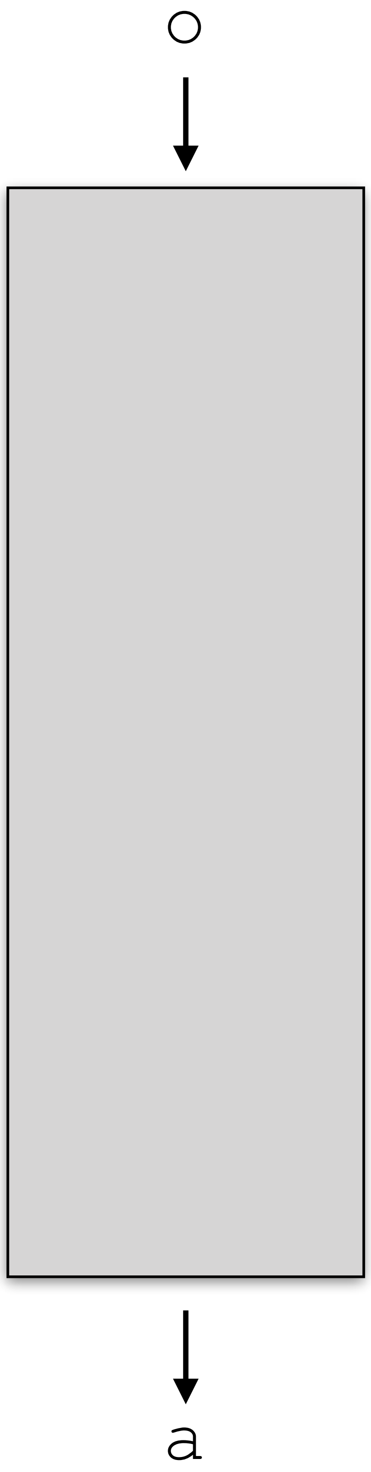

low
personal letter writing
institutional writing
SuperGLUE
ChatGPT
interactivity / immediacy
familial face-to-face int.
institutional face-to-face int
high
familiarity / audience design / shared context
low

II. Task of LU has Structure

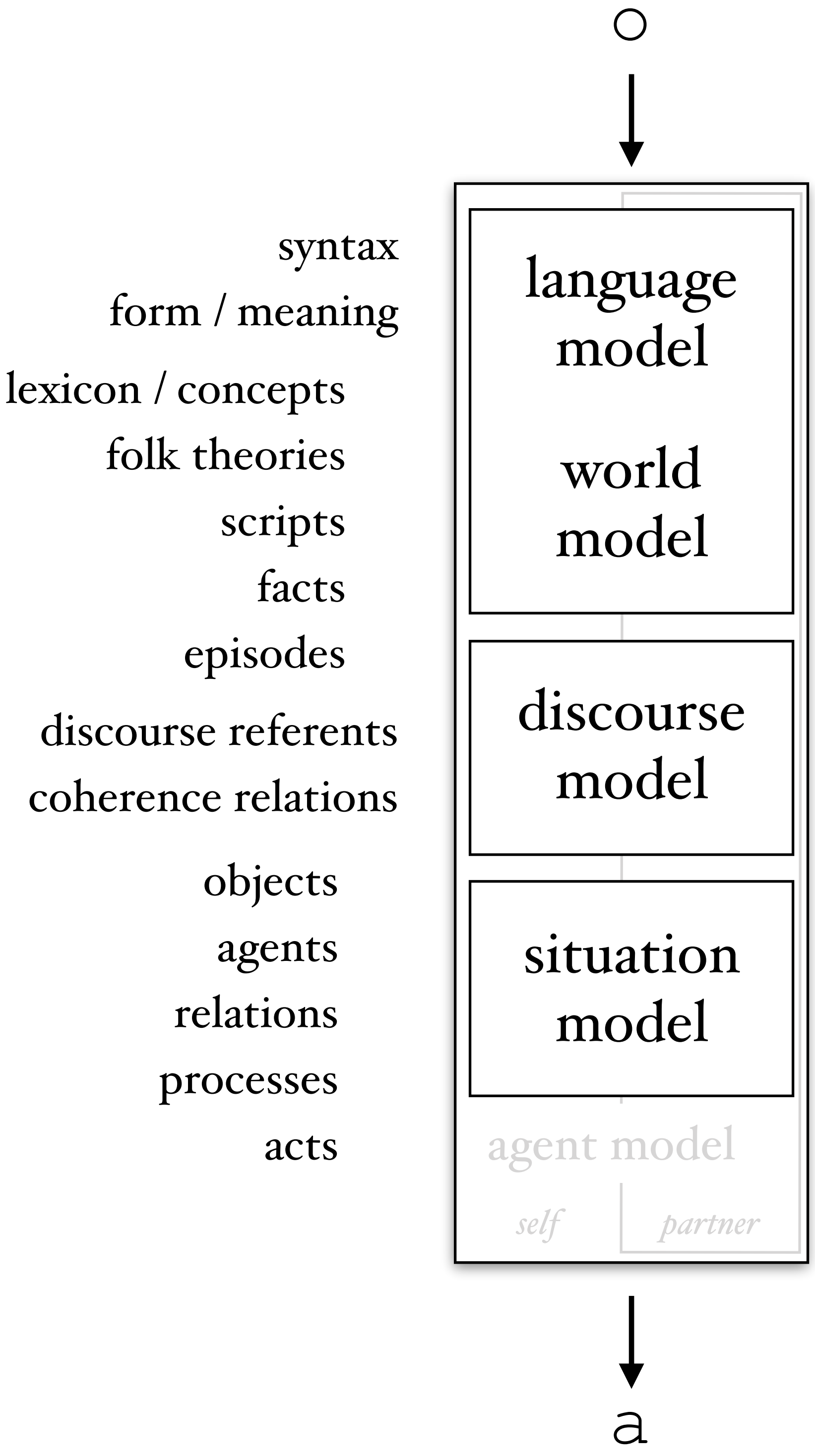
- it makes sense to analyse the task of understanding as
  - requiring *knowledge*,
  - which is applied in (update) *processes*,
    - some of which are single-minded ones, other are interactive ones

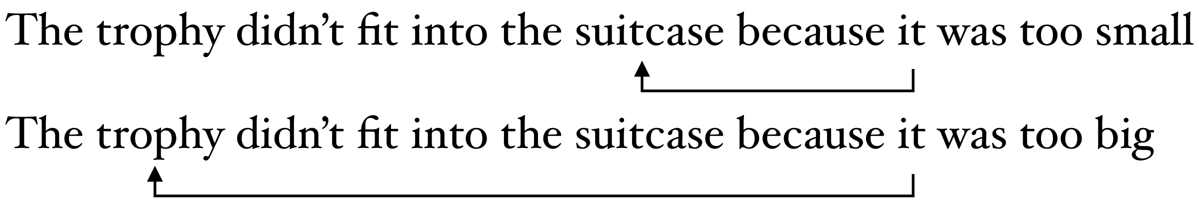

a/o
a/o
a/o

NLP-NLU



NLP-NLU



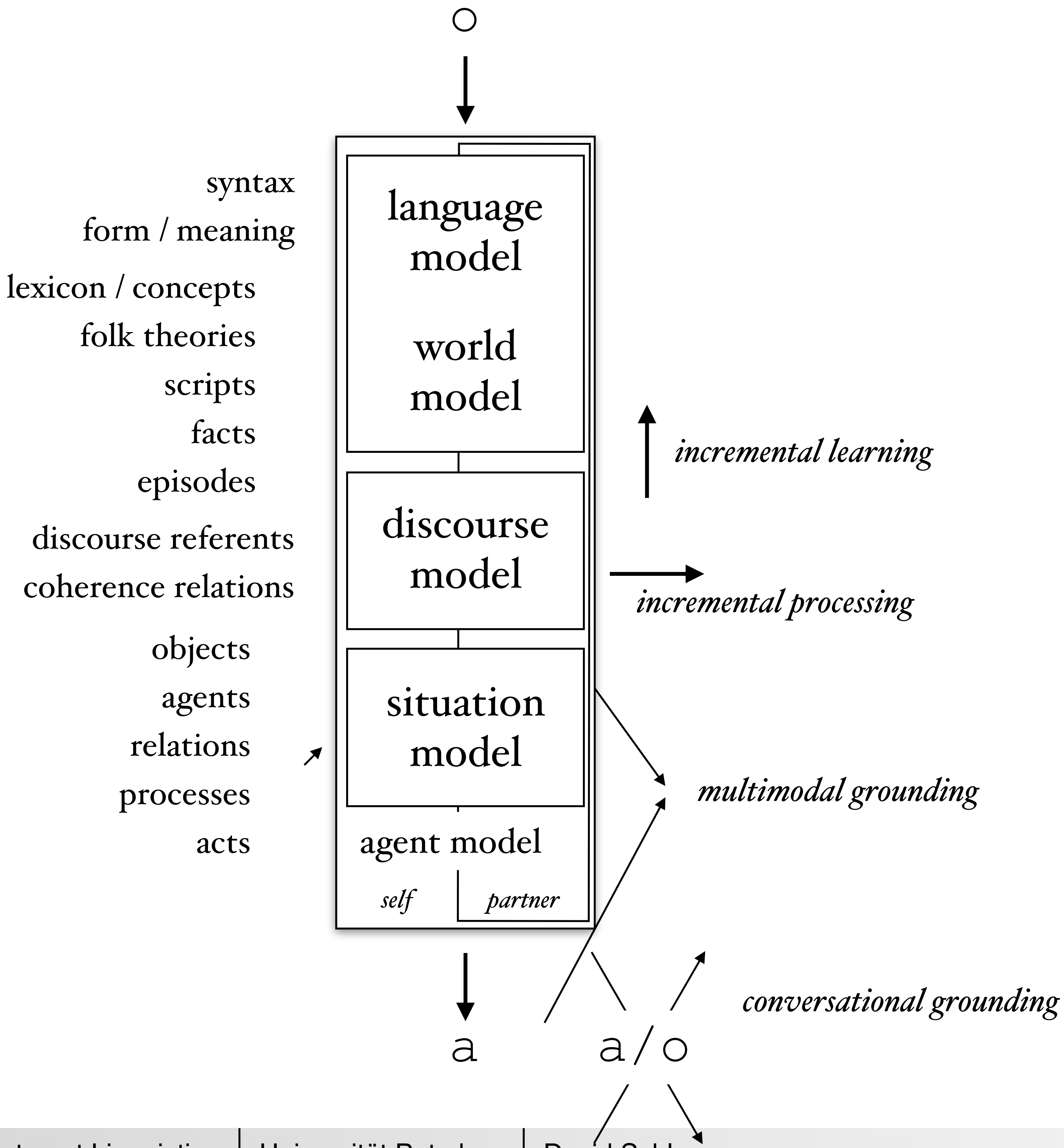

The trophy didn't fit into the suitcase because it was too small

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)
work on *representation probing*

situated NLU



o

syntax
form / meaning
lexicon / concepts
folk theories
scripts
facts
episodes
discourse referents
coherence relations
objects
agents
relations
processes
acts

language model

world model

discourse model

situation model

agent model

self

partner

incremental learning

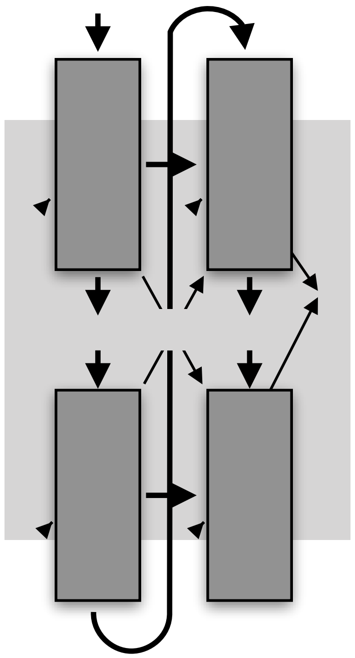
incremental processing

multimodal grounding

conversational grounding

a

a / o

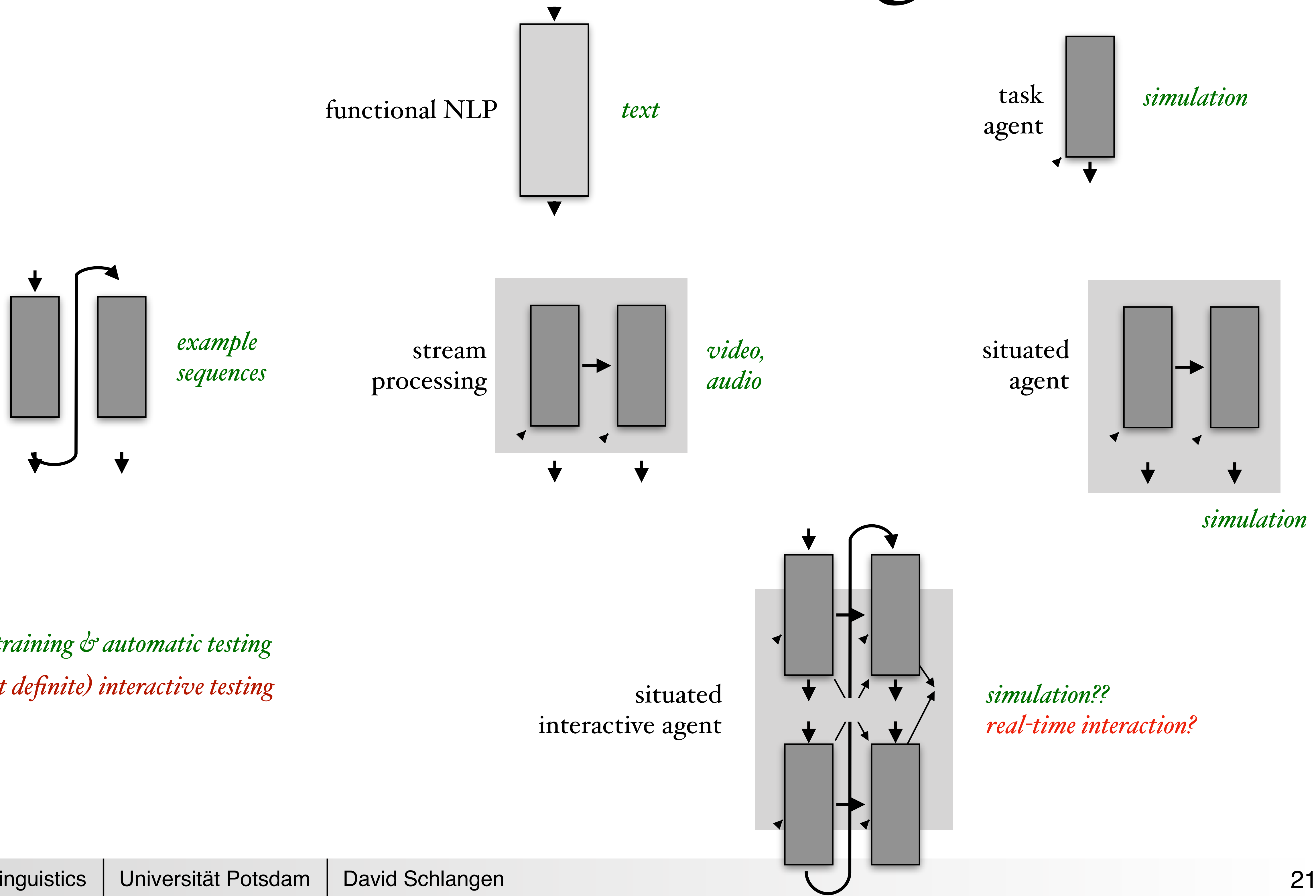


(Schlangen 2023b)

III. Relations btw NLP-NLU & SLU

- this makes clearer the relation btw NLP-NLU & SLU (as tasks, and methodologically)

from NLP-NLU to situated interactive agents



functional NLP

text

task
agent

simulation

turn-incremental
NLP

*example
sequences*

stream
processing

*video,
audio*

situated
agent

simulation

faster-than-realtime training & automatic testing

expensive & slow (but definite) interactive testing

situated
interactive agent

*simulation??
real-time interaction?*

Where's the *theory of NLP*?

- Some points
  - now that things kind of work, kind of in a general way, the domain theory aversion is coming back to haunt NLP, as problems with *measurement*
  - the type of language use represented by NLP-NLU is not the only one, and not even the paradigmatic one
  - it makes sense to analyse the task of understanding as
    - requiring *knowledge*,
    - which is applied in (update) *processes*,
      - some of which are single-minded ones, other are interactive ones
  - this makes clearer the relation btw NLP-NLU & SLU (as tasks, and methodologically)

knowledge
& process

Language Model

(Chomsky 1957)

World Model

(Murphy 2002; Margolis & Laurence 2015)

Situation Model

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

Discourse Model

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

Agent Model

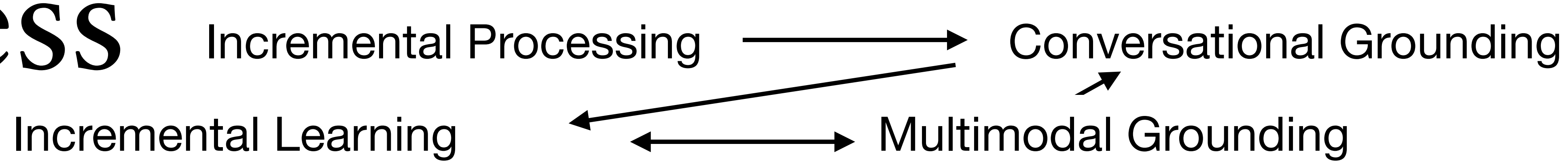
(Bratman 1987, Cohen *et al.* 1990, Clark 1996)

ALARM! Is this not just 20th century AI??

Observations certainly not new.
(This combination may be?)

But the claim is not that these should be modelled symbolically (representations + rules), just that it makes sense to pay attention to these aspects of knowledge and knowledge dynamics.

knowledge & process



Language Model

World Model

Situation Model

Discourse Model

Agent Model

(Levinson 2010)
(Christianson & Chater 2016)

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

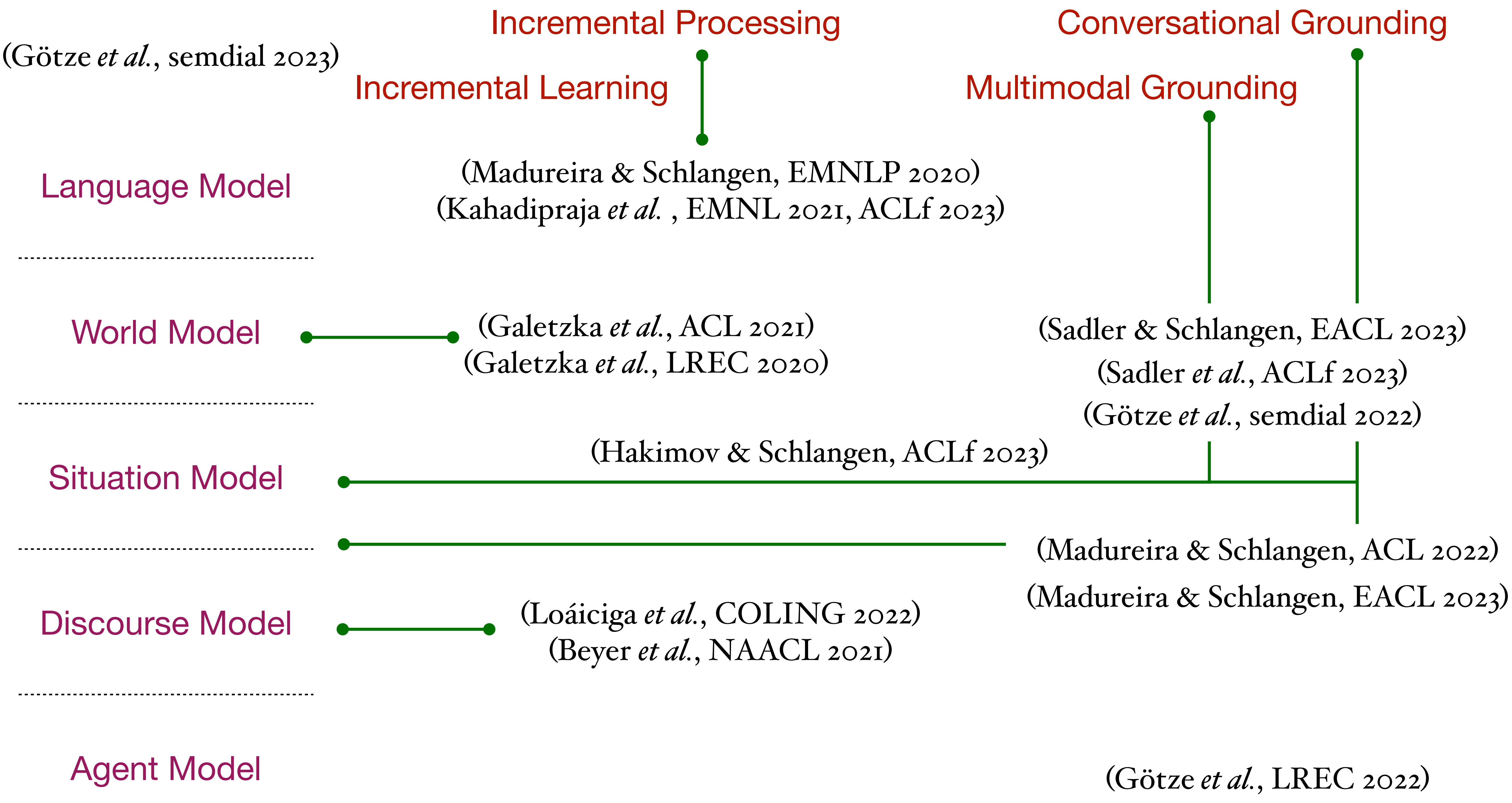
(Harris 2015)
(E. Clark 2003)

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

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

Schlangen (forthcoming)

Research Programme



Incremental Processing

Conversational Grounding

(Götze *et al.*, semdial 2023)

Incremental Learning

Multimodal Grounding

Language Model

(Madureira & Schlangen, EMNLP 2020)
(Kahadipraja *et al.* , EMNL 2021, ACLf 2023)

World Model

(Galetzka *et al.*, ACL 2021)
(Galetzka *et al.*, LREC 2020)

(Sadler & Schlangen, EACL 2023)
(Sadler *et al.*, ACLf 2023)
(Götze *et al.*, semdial 2022)

Situation Model

(Hakimov & Schlangen, ACLf 2023)

(Madureira & Schlangen, ACL 2022)
(Madureira & Schlangen, EACL 2023)

Discourse Model

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

Agent Model

(Götze *et al.*, LREC 2022)

Structure

- Part I: “Understanding Understanding”
  - Our Research Programme
  - Some Systematic Insights
  - Some Recent Projects

- The *task* of LU has internal structure
- Realising this is useful (at the very least) for benchmarking

- Part II: Probing Chat-Optimized LLMs Through Gameplay
  - Games & Capabilities
  - `clemgame`
  - `clembench`

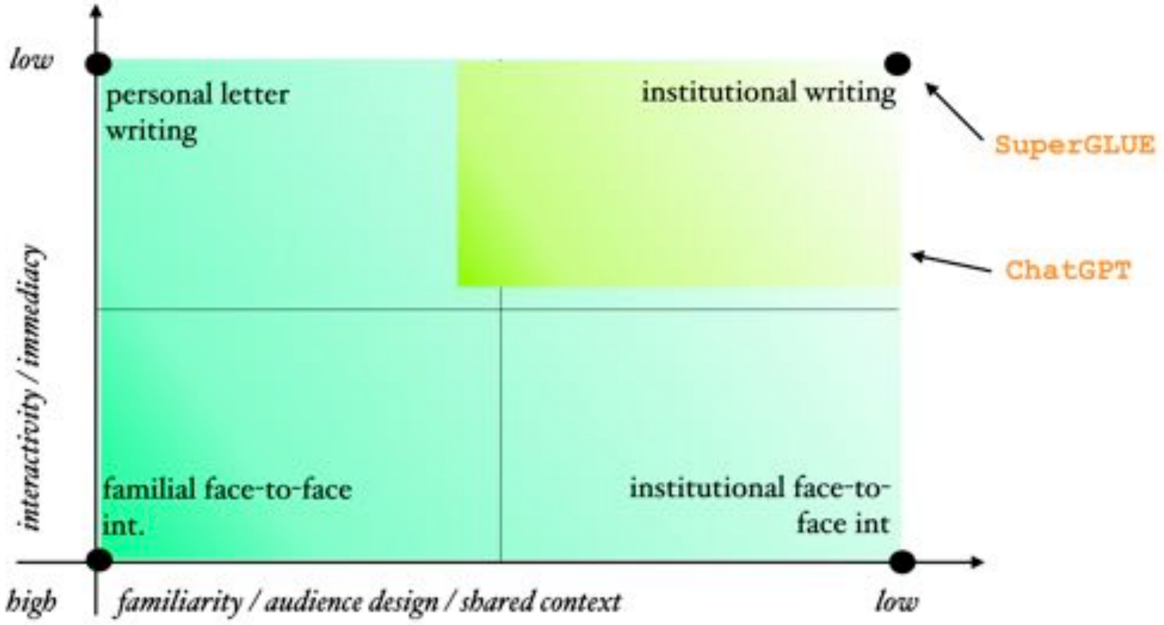
Structure

- Part I: “Understanding Understanding”
  - Our Research Programme
  - Some Systematic Insights
  - Some Recent Projects

- The *task* of LU has internal structure
- Realising this is useful (at the very least) for benchmarking

- Part II: Probing Chat-Optimized LLMs Through Gameplay
  - Games & Capabilities
  - `clemgame`
  - `clembench`

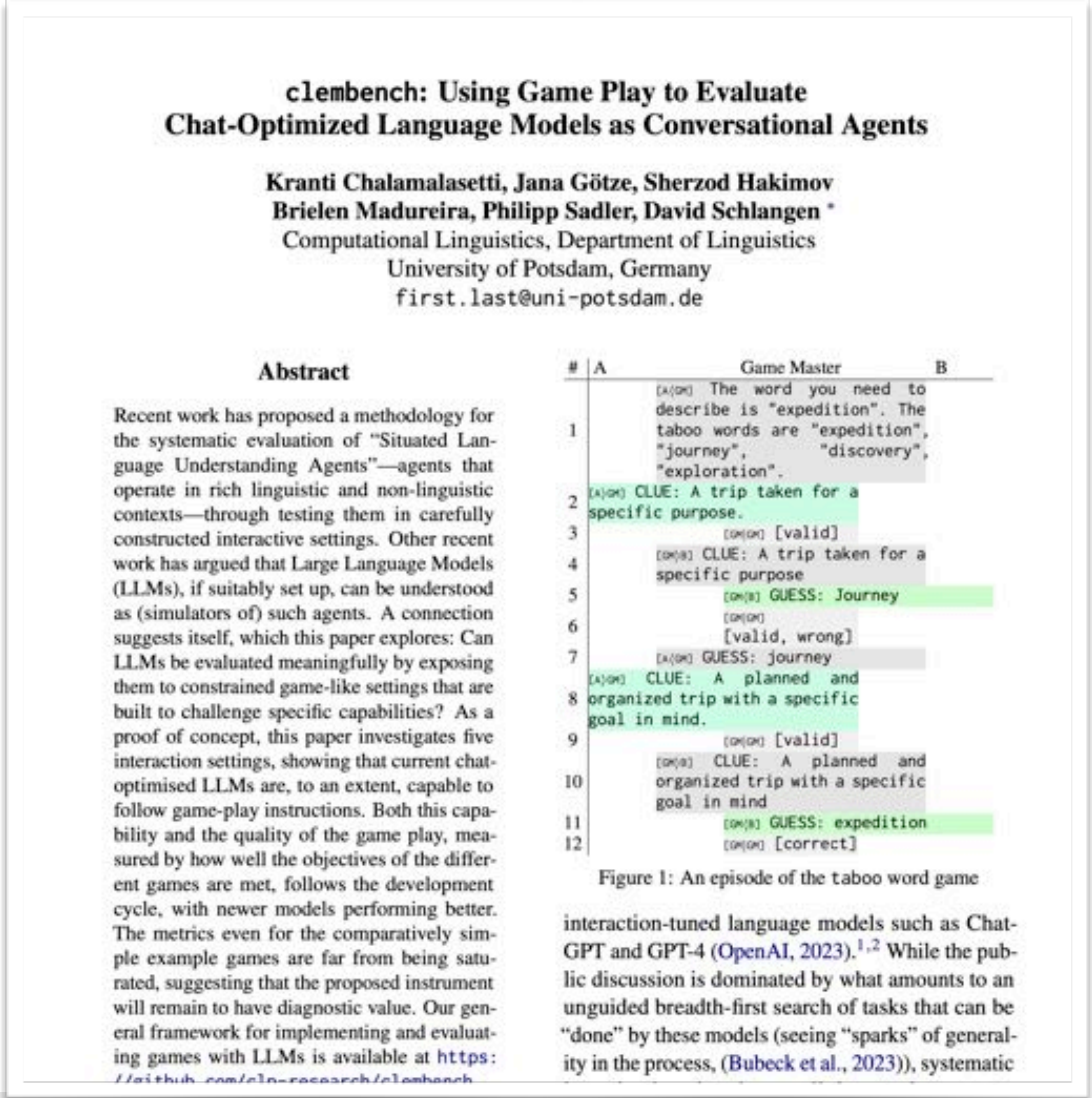
The New Kids on the Block: chat-optimised LLMs (cLLMs)


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

- We know how to evaluate it NLP-NLU style.
- That's not all of NLU.
- We've said that *Dialogue Games* offer a principled way to evaluate "language understanding / use in context".
- Can we evaluate cLLMs with Dialogue Games?
    - (Ideally, even in self-play / faster-than-real time & w/o human involvement.)

(a note in advance)

- On ArXiv (Chalamalasetti *et al.* 2023) & github [https://github.com/clp-research/clembench]
- but may or may not also be under review
- consider it not being advertised here…



`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 built to challenge specific capabilities? As a proof of concept, this paper investigates five interaction settings, showing that current chat-optimised LLMs are, to an extent, capable to follow game-play instructions. Both this capability and the quality of the game play, measured by how well the objectives of the different games are met, follows the development cycle, with newer models performing better. The metrics even for the comparatively simple example games are far from being saturated, suggesting that the proposed instrument will remain to have diagnostic value. Our general framework for implementing and evaluating games with LLMs is available at `https://github.com/clp-research/clembench`

#	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 and organized trip with a specific goal in mind	
11			[GM⟩B] GUESS: expedition
12		[GM⟩GM] [correct]	

Figure 1: An episode of the `taboo` word game

interaction-tuned language models such as ChatGPT and GPT-4 (OpenAI, 2023).[1,2] While the public discussion is dominated by what amounts to an unguided breadth-first search of tasks that can be "done" by these models (seeing "sparks" of generality in the process, (Bubeck et al., 2023)), systematic

Why 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)

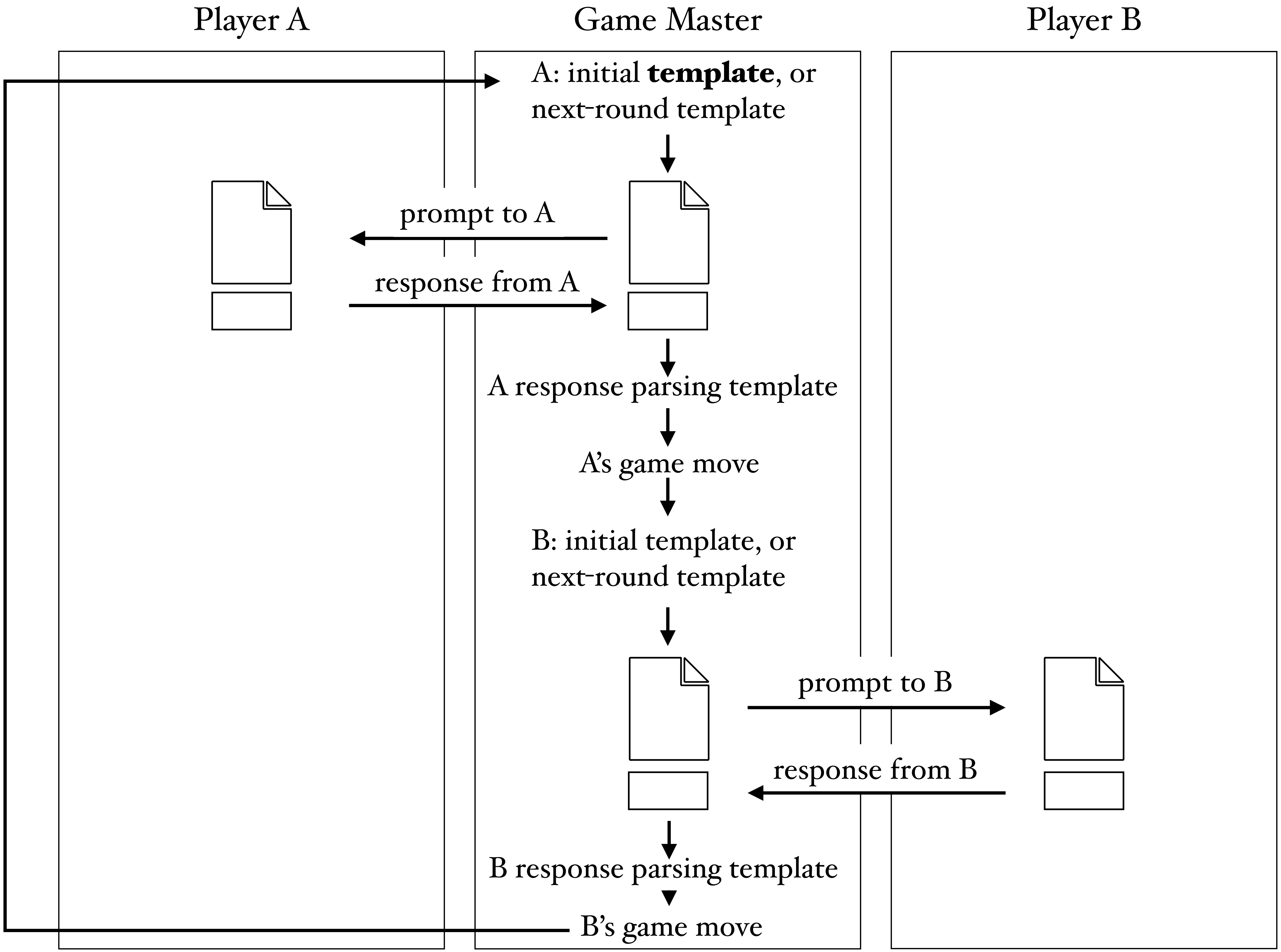
multi-turn (provides context to language)

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

`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

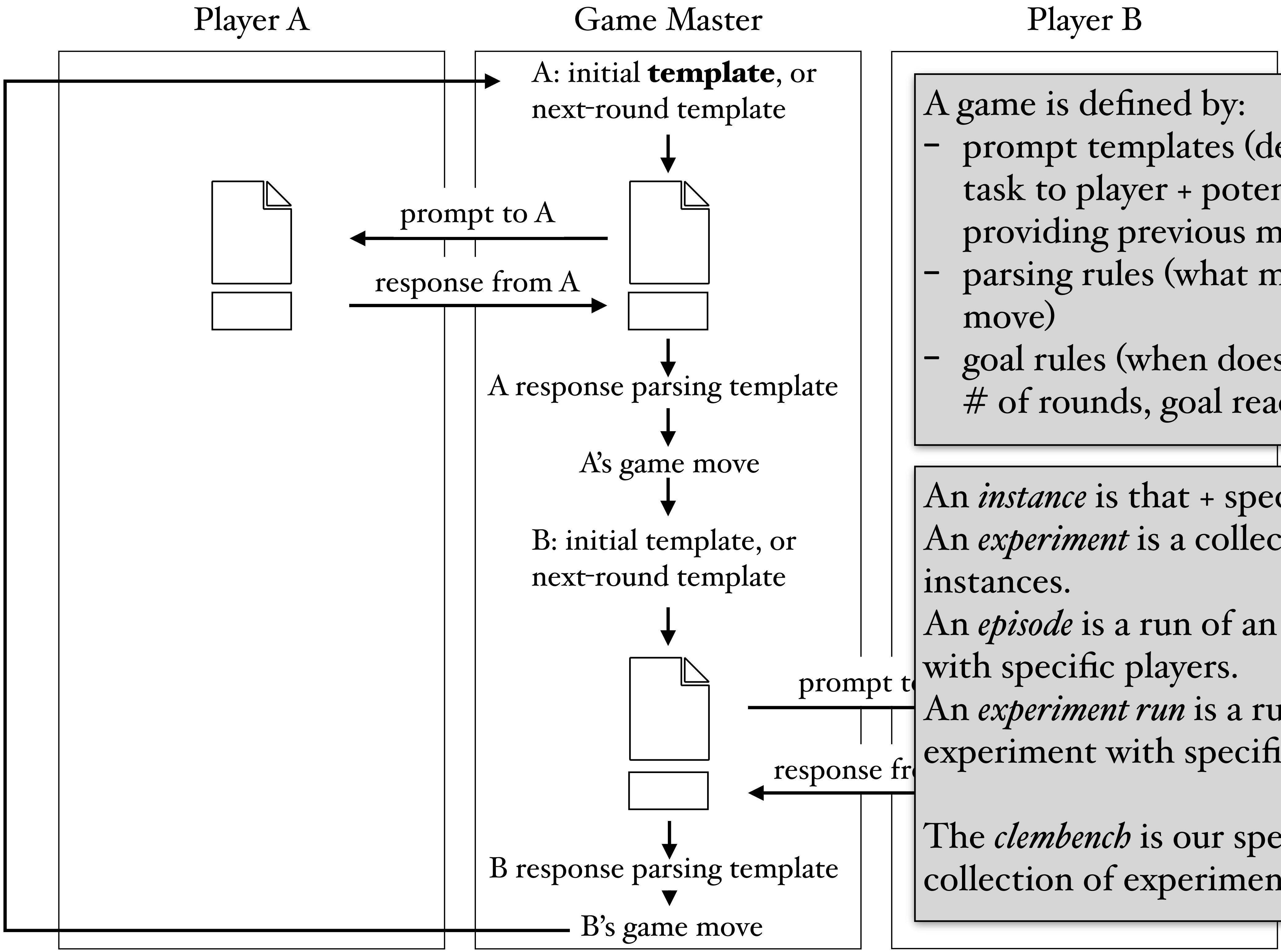
prompt to B

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

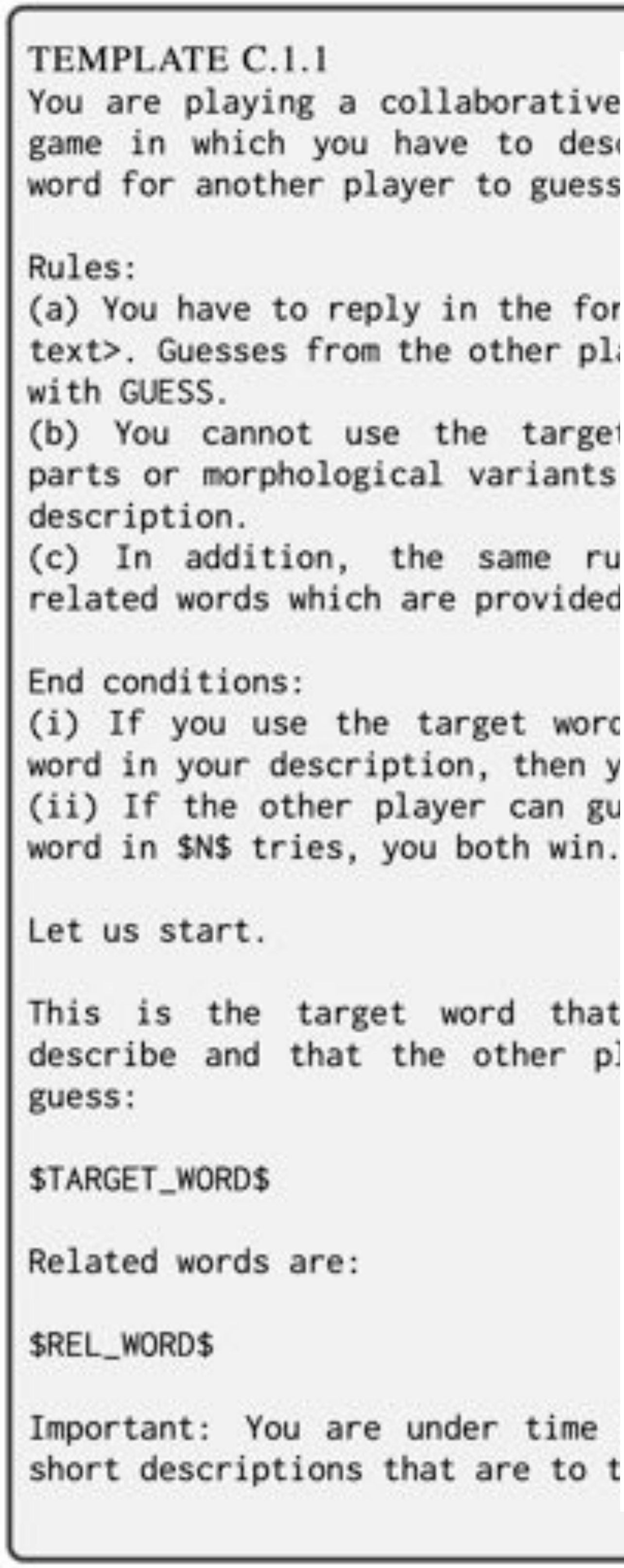
A game is defined by:

- 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 wor
word in your description, then y
(ii) If the other player can gu
word in N tries, you both win.

Let us start.

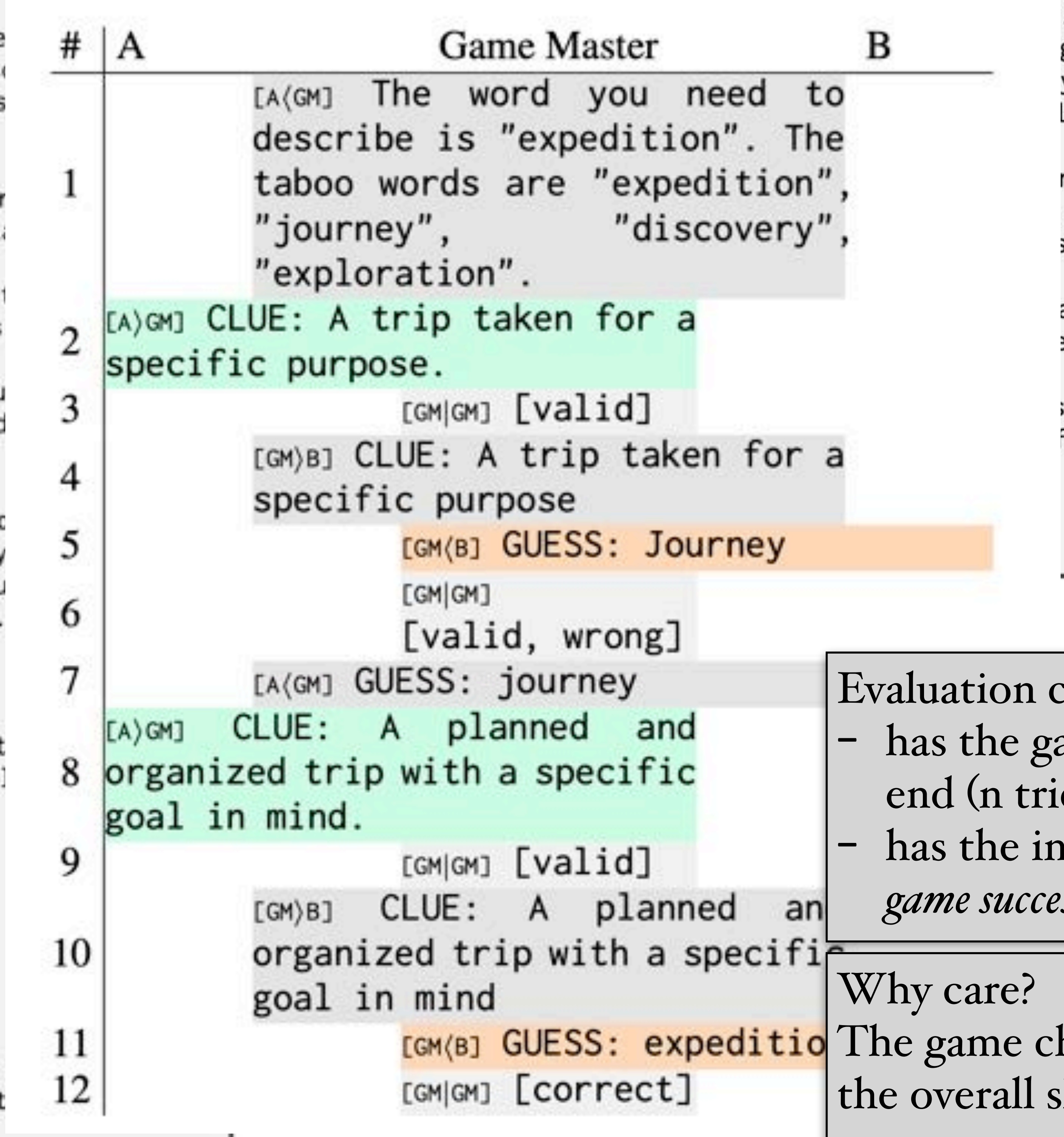
This is the target word that
describe and that the other pl
guess:

$TARGET_WORD$

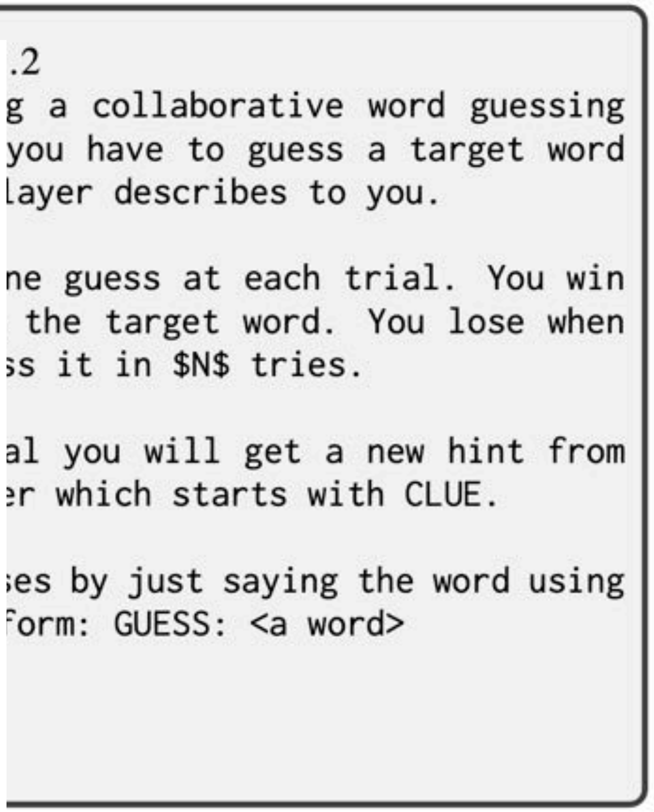
Related words are:

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.

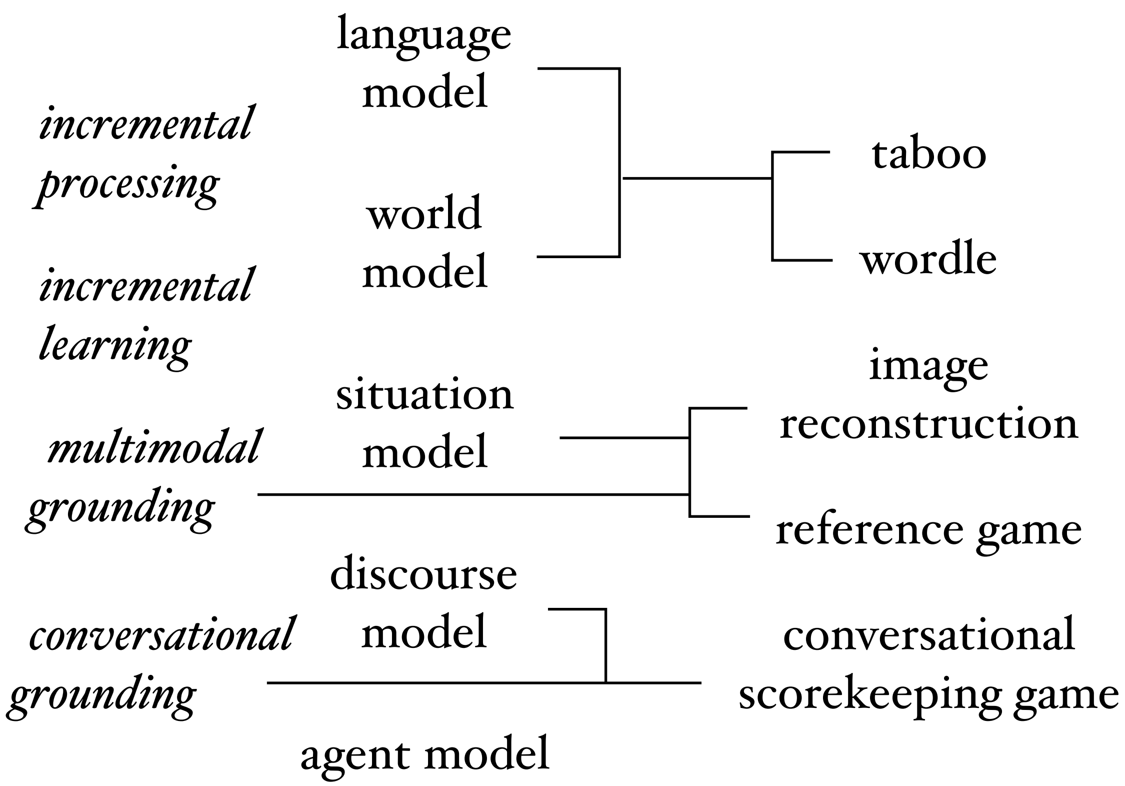
ses by just saying the word using
form: GUESS: <a word>

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".

Games & Understanding



language model

incremental processing

world model

taboo

wordle

incremental learning

situation model

multimodal grounding

image reconstruction

reference game

discourse model

conversational grounding

conversational scorekeeping game

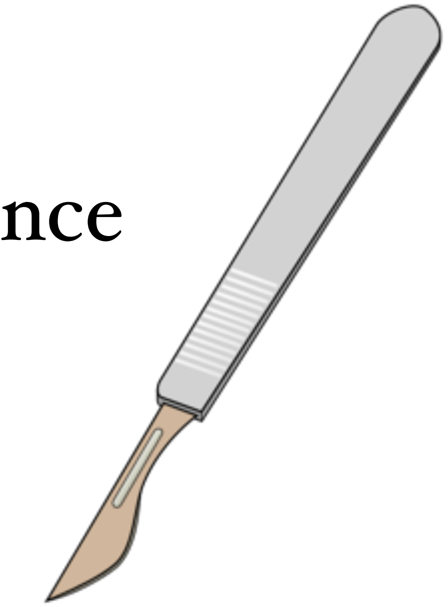
agent model

Games as benchmarking instrument

- coarse-grained: one number, to weigh and rank

$$\left(\text{quality} * \frac{\text{completed}}{\text{attempted}} \right)$$

- 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

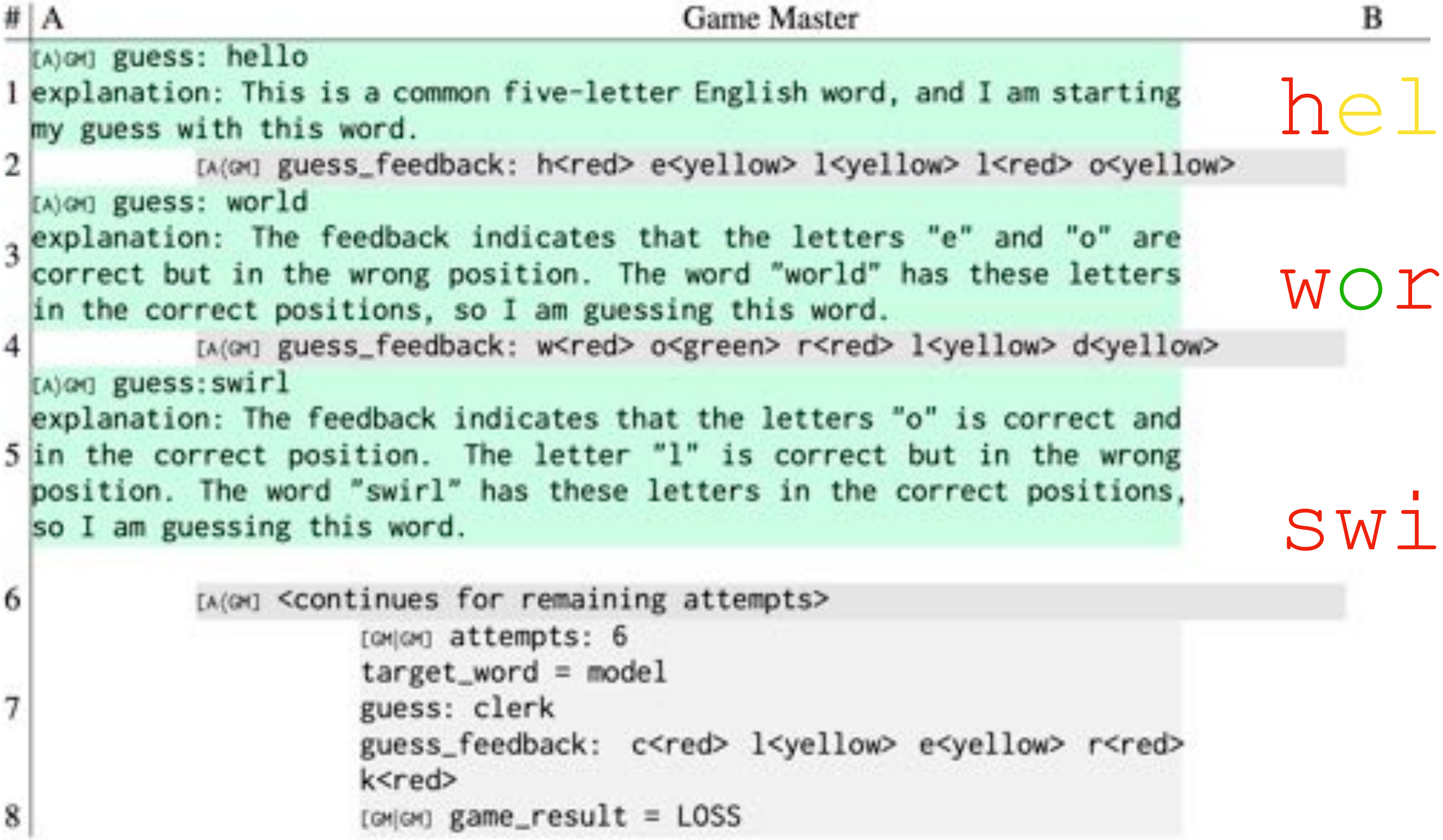


https://creativecommons.org/licenses/by-sa/4.0/deed.en https://commons.wikimedia.org/wiki/File:Dessin_scalpel.svg Petit B

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



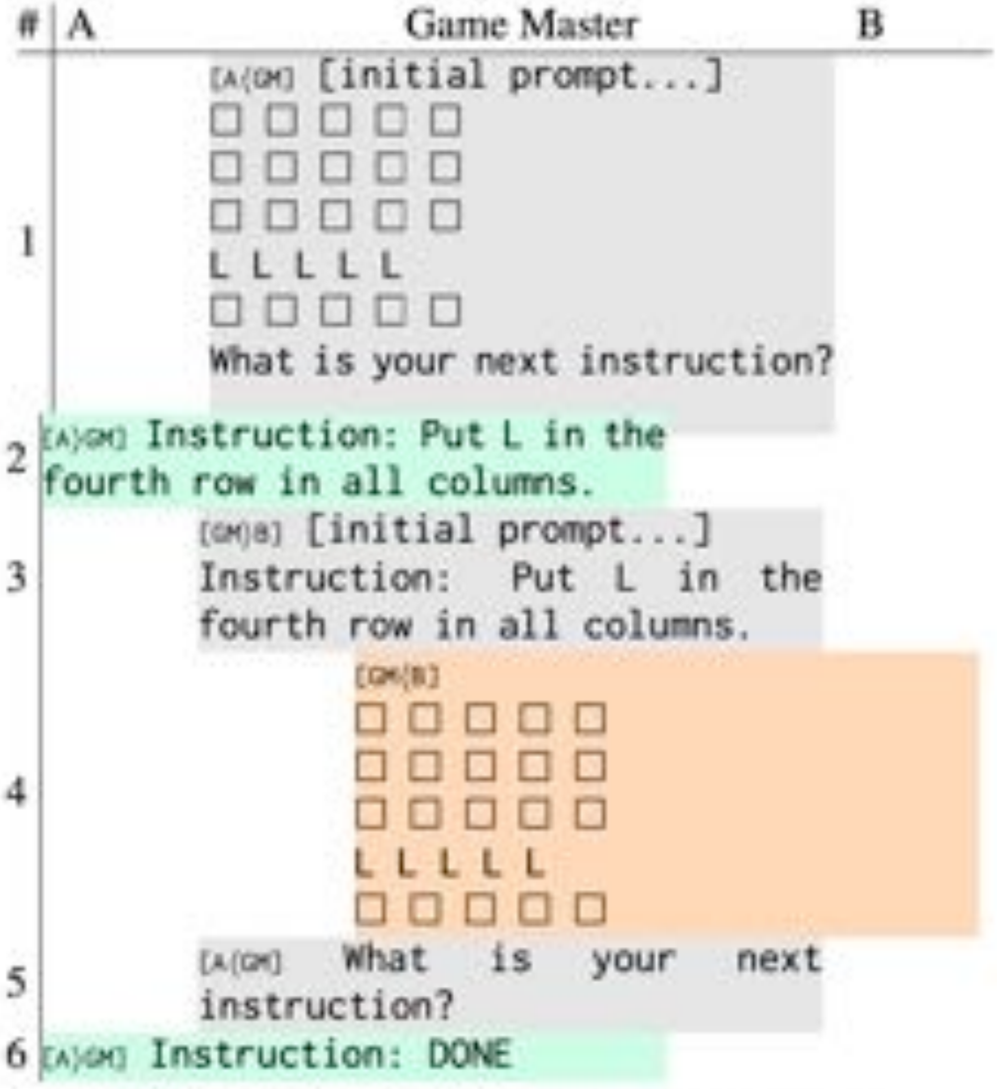
#	A	Game Master	B
1	[A⟩GM] guess: hello explanation: This is a common five-letter English word, and I am starting my guess with this word.		
2		[A⟨GM] guess_feedback: h<red> e<yellow> l<yellow> l<red> o<yellow>	
3	[A⟩GM] guess: world explanation: The feedback indicates that the letters "e" and "o" are 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

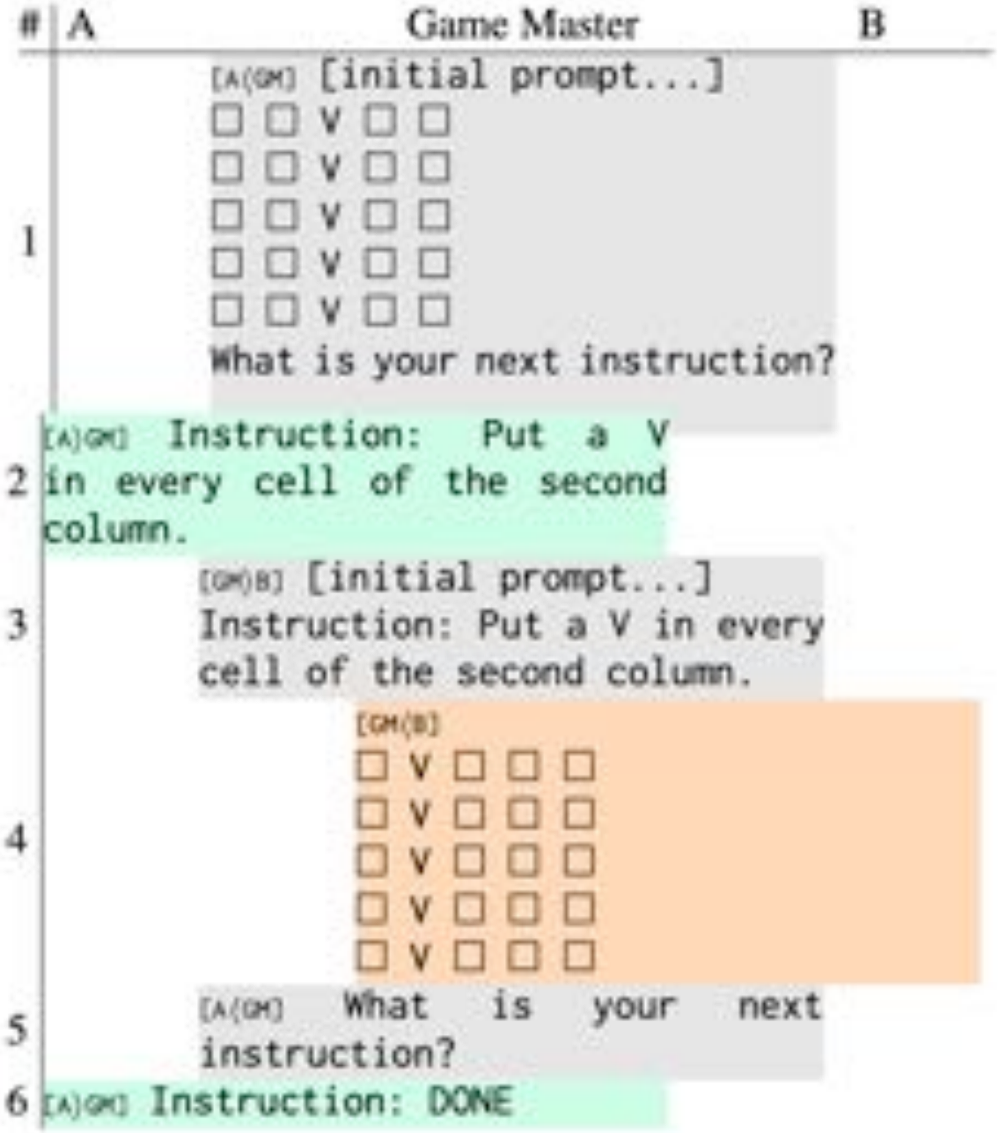
swirl

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		

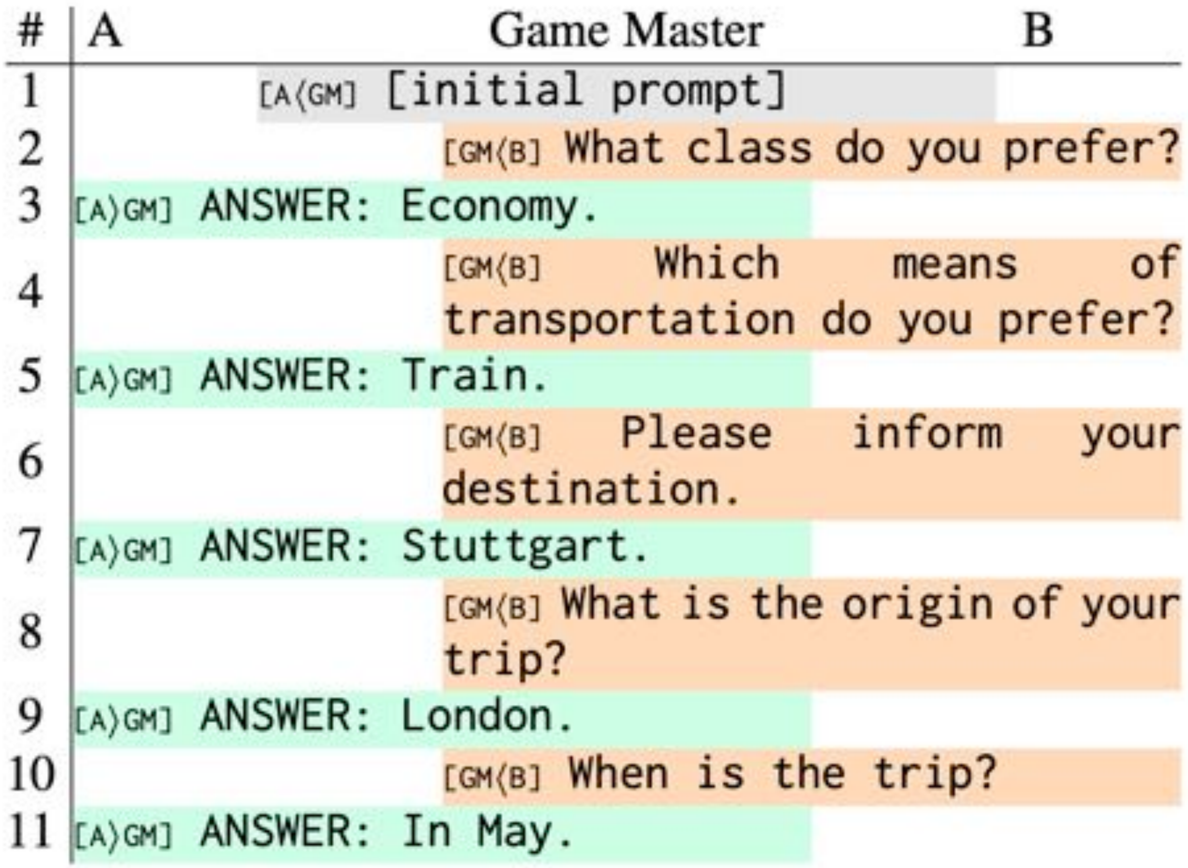
(a) Successful gameplay from both players



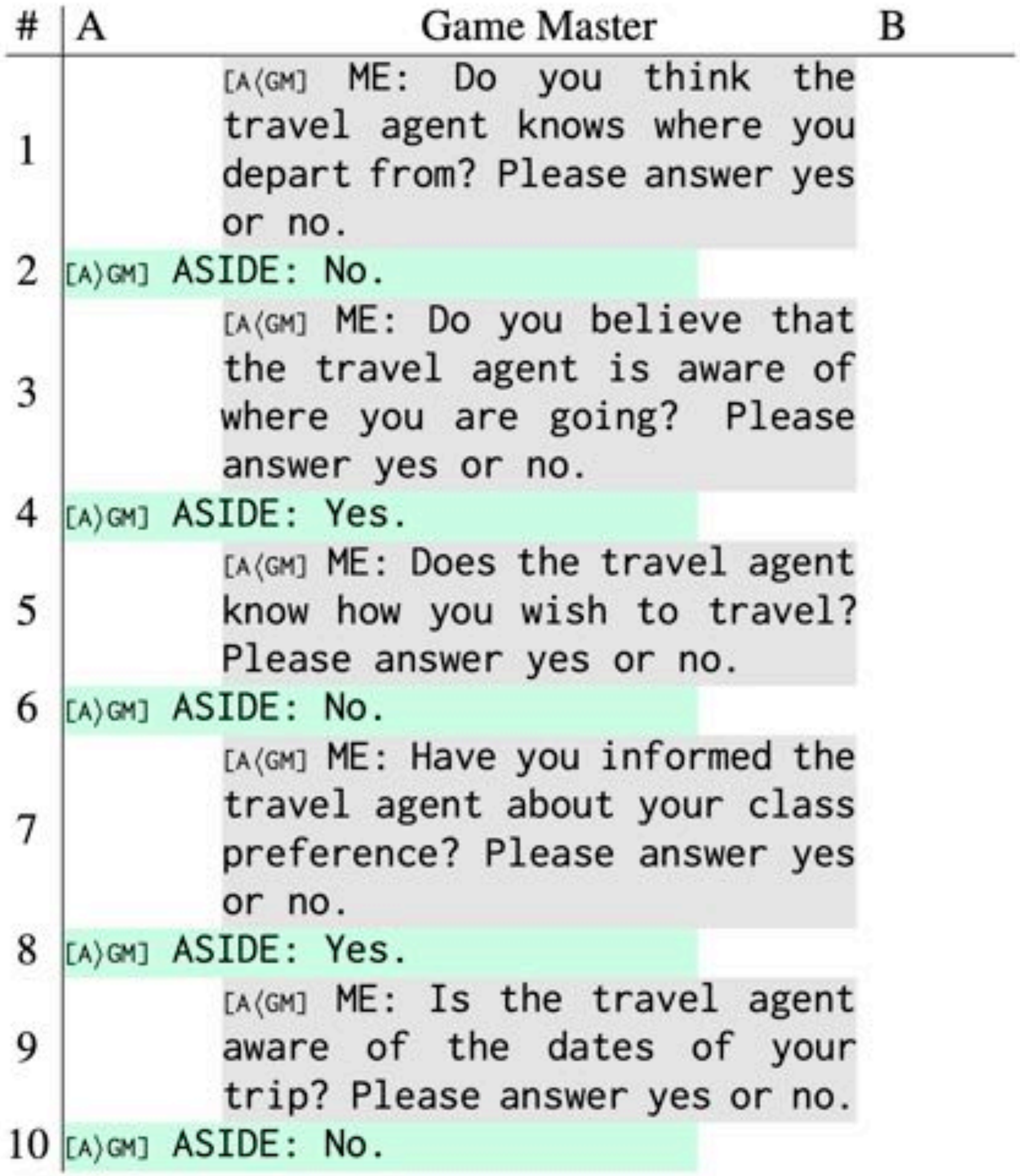
#	A	Game Master	B
1		[A⟨GM] [initial prompt...] □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ □ □ V □ □ What is your next instruction?	
2	[A⟩GM] Instruction: Put a V in every cell of the second column.		
3		[GM⟩B] [initial prompt...] Instruction: Put a V in every cell of the second column.	
4			[GM⟨B] □ V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

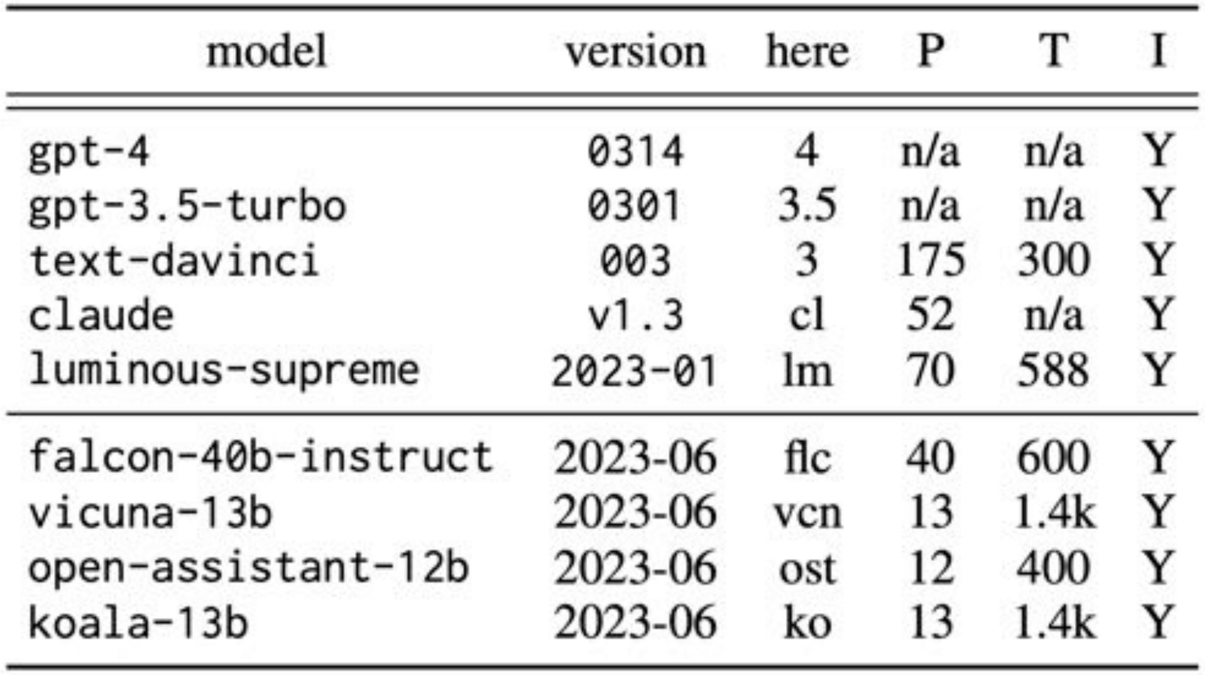


#	A	Game Master	B
1		[A⟨GM] [initial prompt]	
2		[GM⟨B] What class do you prefer?	
3	[A⟩GM] ANSWER: Economy.		
4		[GM⟨B] Which means of transportation do you prefer?	
5	[A⟩GM] ANSWER: Train.		
6		[GM⟨B] Please inform your destination.	
7	[A⟩GM] ANSWER: Stuttgart.		
8		[GM⟨B] What is the origin of your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

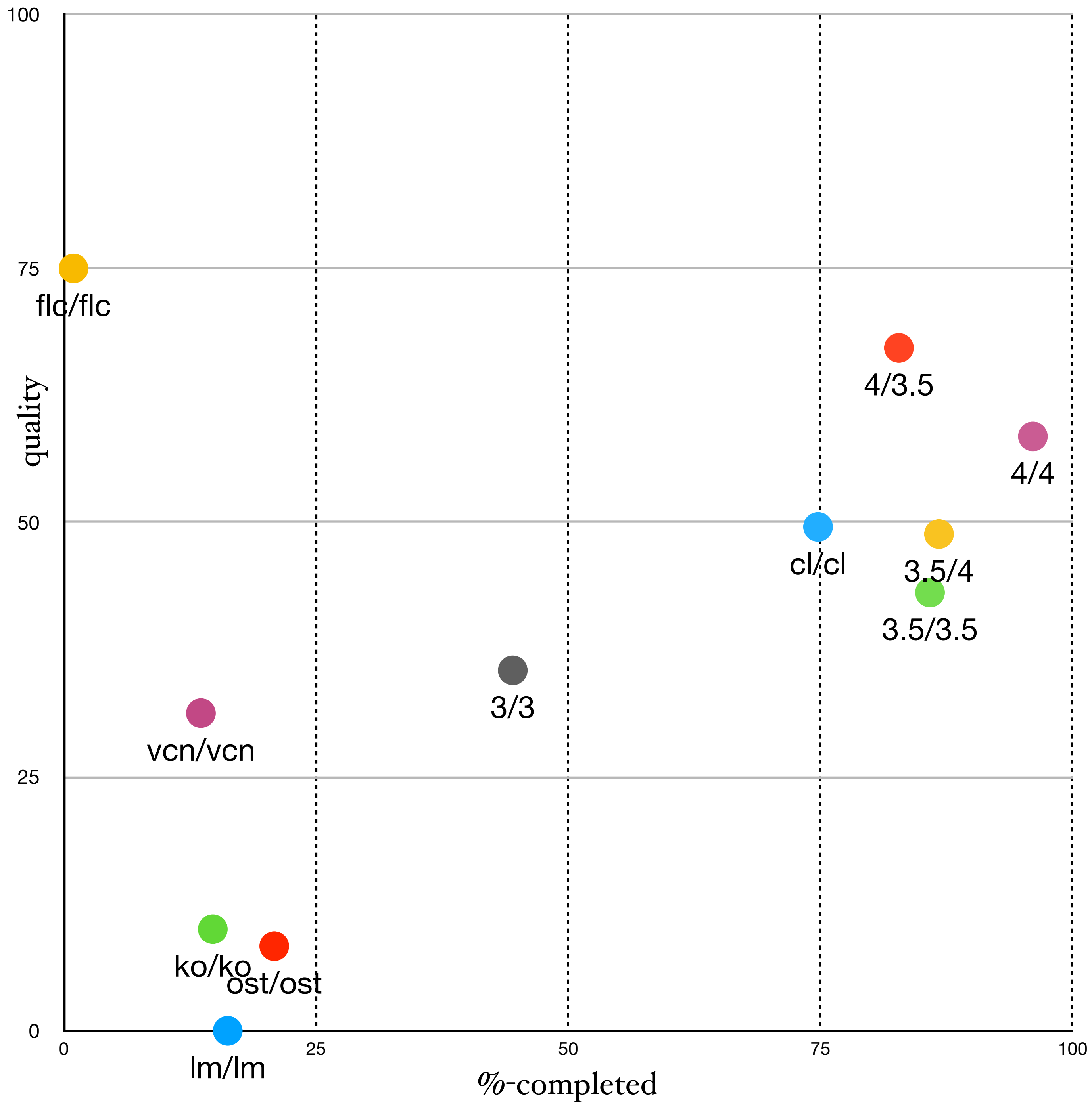
Running the benchmark

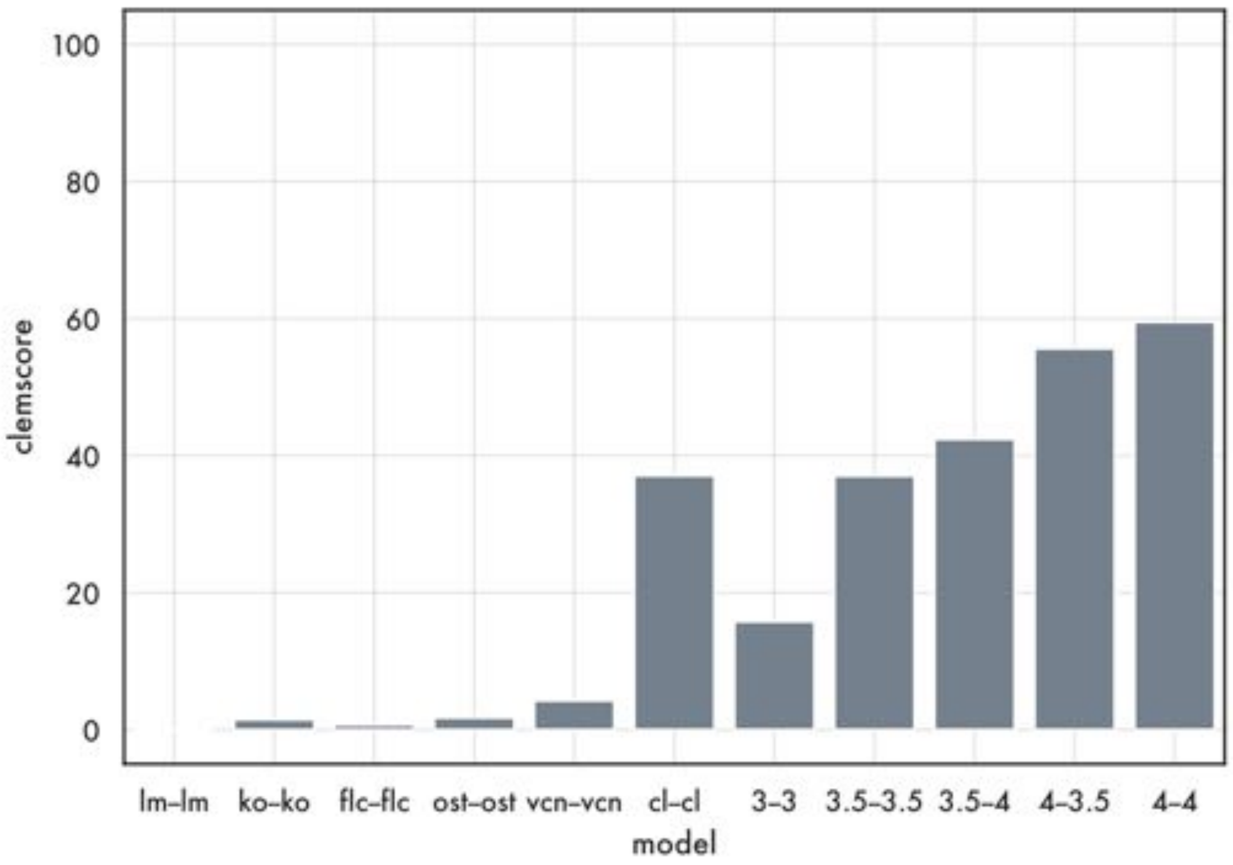


model	version	here	P	T	I
`gpt-4`	`0314`	4	n/a	n/a	Y
`gpt-3.5-turbo`	`0301`	3.5	n/a	n/a	Y
`text-davinci`	`003`	3	175	300	Y
`claude`	`v1.3`	cl	52	n/a	Y
`luminous-supreme`	`2023-01`	lm	70	588	Y
`falcon-40b-instruct`	2023-06	flc	40	600	Y
`vicuna-13b`	2023-06	vcn	13	1.4k	Y
`open-assistant-12b`	2023-06	ost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

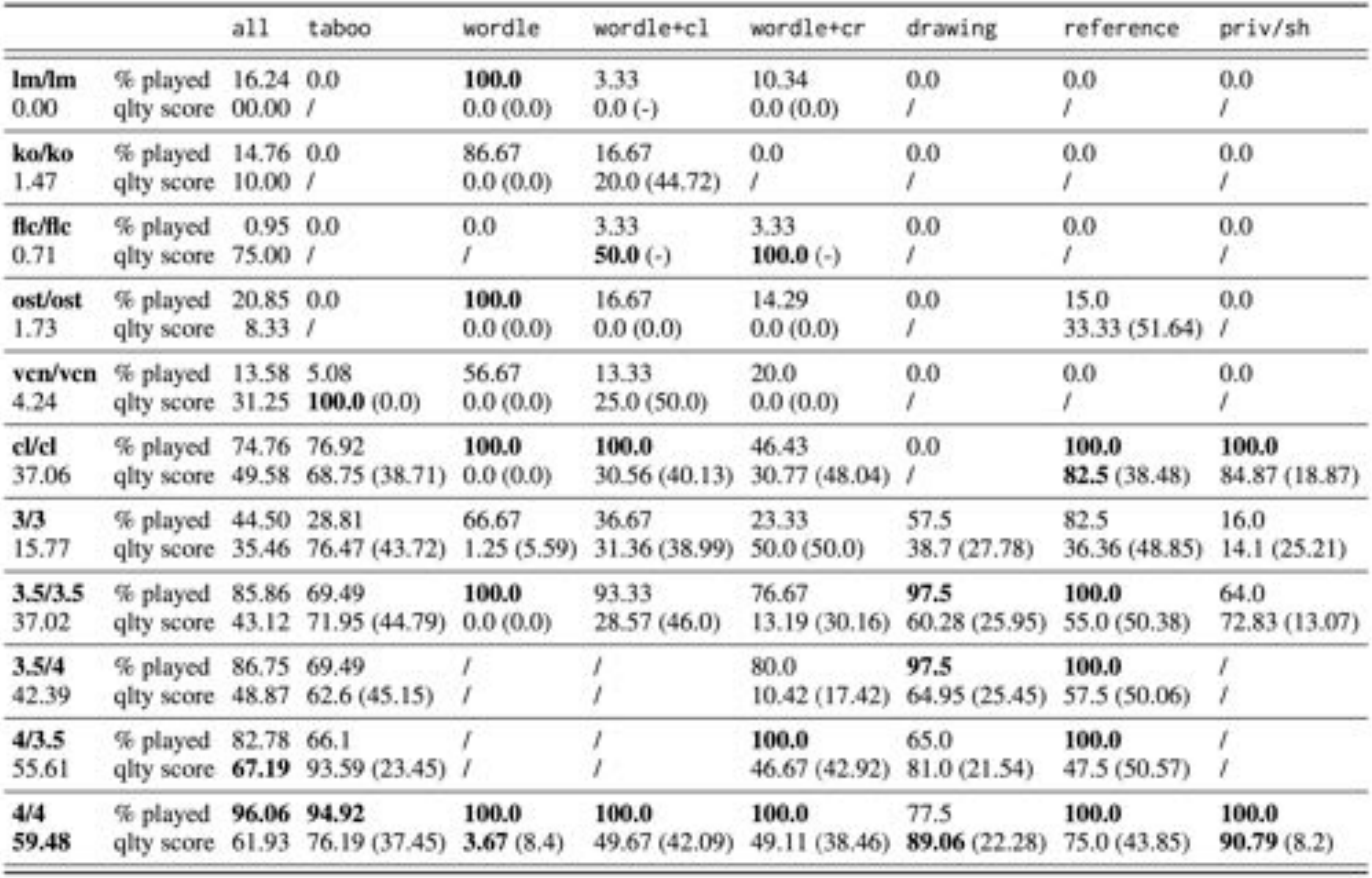
- Evaluated for:
  - *%* games played to completion
    [following formal rules]
  - quality of game play (only completed games)
    [reaching game-specific goal]

Running the benchmark





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. (Compare 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

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)
  - 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

Structure

- Part I: "Understanding Understanding"
  - Our Research Programme
  - Some Systematic Insights
  - Some Recent Projects

- The *task* of LU has internal structure
- Realising this is useful (at the very least) for benchmarking

- Part II: Probing Chat-Optimized LLMs Through Gameplay
  - Games & Capabilities
  - `clemgame`
  - `clembench`

Structure

- Part I: “Understanding Understar
    - Our Research Programme
    - Some Systematic Insights
    - Some Recent Projects

- There’s more to LU than NLU
- The *task* of LU has internal structure
- Realising this is useful (at the very least) for benchmarking
- Also sketches path from NLP-NLU to SLU

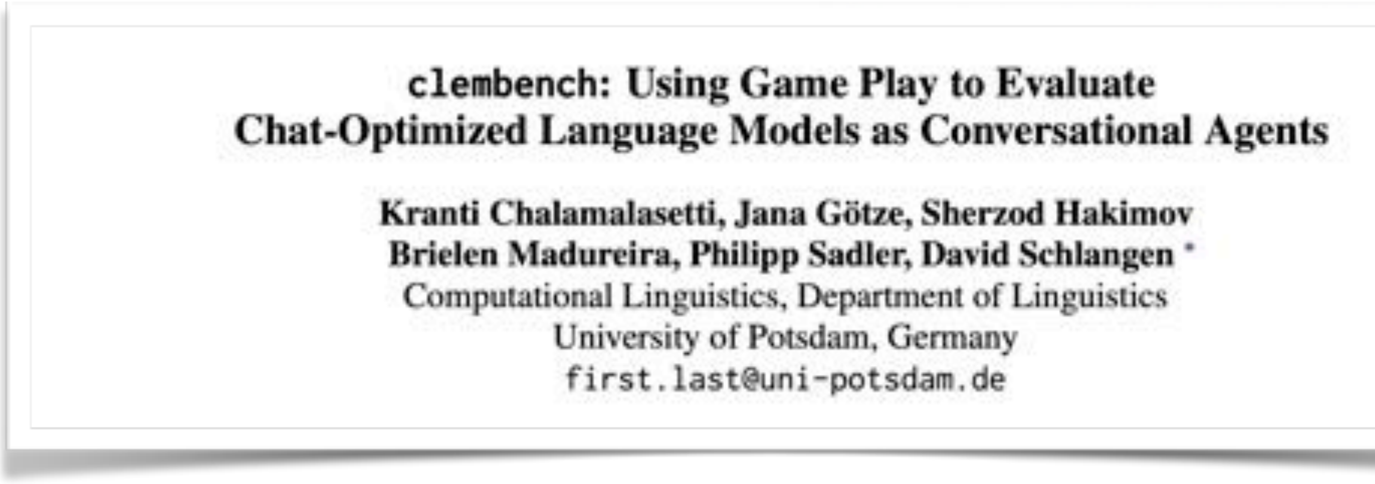
- Part II: Probing Chat-Optimized
    - Games & Capabilities
    - `clemgame`
    - `clembench`

- cLLMs do go some steps towards handling context-dependent language use
- using formal constraints to rein in LLMs might be worthwhile strategy for more (real-)task-oriented SIAs…

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.



`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`

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

List of References for the Talk “Understanding Understanding”

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

References

Andreas, Jacob (Dec. 2022). “Language Models as Agent Models”. In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 5769–5779.

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 (2023). “clembench: Using Game Play to Evaluate Chat-Optimized Language Models as Conversational Agents”. In: *CoRR*.

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.

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.

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.

Kontogiorgos, Dimosthenis, Elena Sibirtseva, Andre Pereira, Gabriel Skantze, and Joakim Gustafson (2018). “Multimodal Reference Resolution In Collaborative Assembly Tasks”. In: *Proceedings of the 4th International Workshop on Multimodal Analyses Enabling Artificial Agents in Human-Machine Interaction*.

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 (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.

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.

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`.

— (2023b). “What A Situated Language-Using Agent Must be Able to Do: A Top-Down Analysis”. In: *CoRR* abs/2302.08590. arXiv: `2302.08590`.

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`.