Queen Mary
University of London
Science and Engineering

Jožef Stefan Institute
Ljubljana, Slovenia

Computational Language Modelling for Cognitive and Social Science

Matt Purver

Computational Linguistics Lab, QMUL
Department of Knowledge Technologies, JSI

Queen Mary University of London / Jožef Stefan Institute

Language as a Probe

- Language helps us observe cognitive states
  - Attitudes
  - Biases
  - Mental health conditions

- Language helps us observe cognitive abilities
  - Interaction & communication quality
  - Relationships
  - Mental health conditions
  - Mental health treatment

- How can we use these to answer cognitive & social questions?

- What difference do today's Large Language Models (LLMs) make?

Language Models

- Generative
  - Trained to generate likely samples of a “language”

- Discriminative
  - Trained to discriminate between “languages”

Language Models

- Generative: e.g. BERT & GPT
  - Trained to guess a masked token

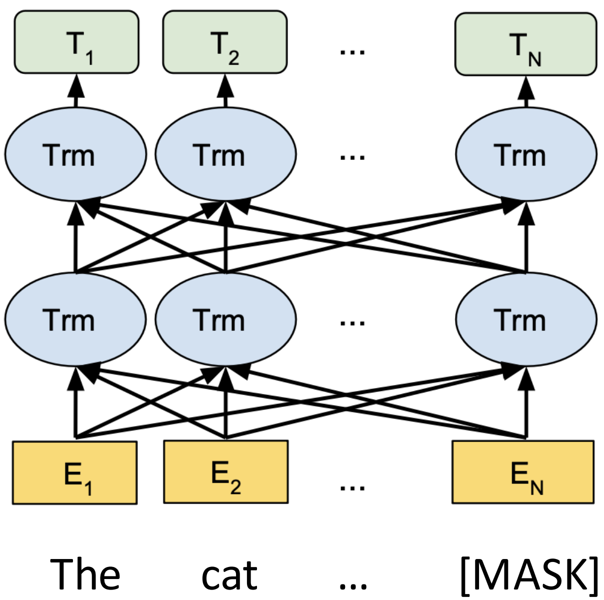
With one month to election day, the [redacted] between Donald Trump and Kamala Harris is the electoral equivalent of a bare-knuckle brawl.

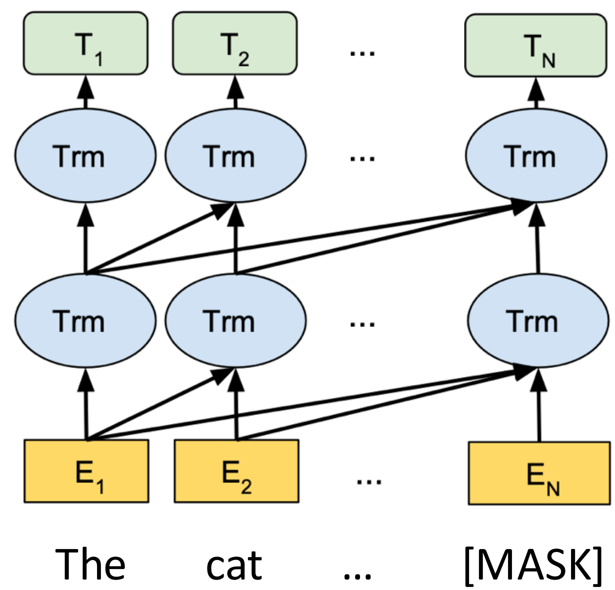
The race for the White House still appears deadlocked, both nationally and in battleground states, so [redacted] will be decided by the slimmest of margins - every new voter engaged, every undecided voter swayed, could help land a knock-out punch.

"In any super close [redacted], where the electorate is divided down the middle, a difference of a percentage point or two could be decisive," says David Greenberg, a presidential historian at Rutgers University.

Language Models

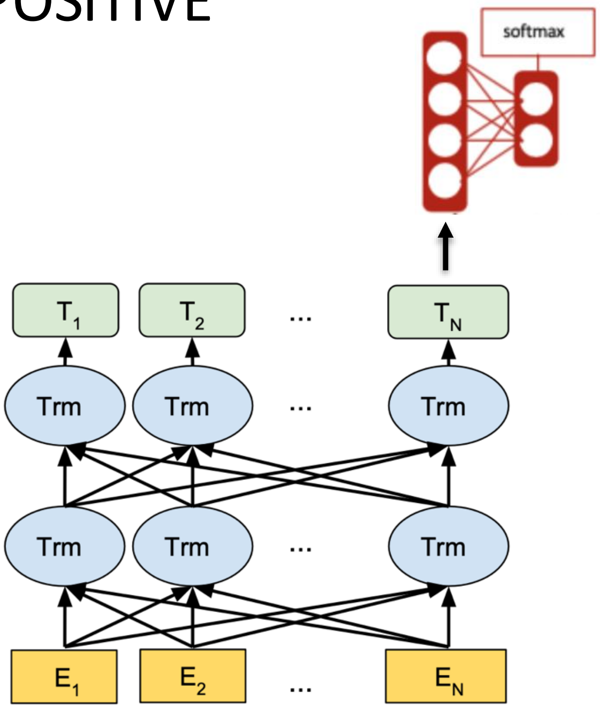
- Generative: e.g. BERT & GPT
  - Trained to guess a masked token
    - “the cat sits on the [MASK]”



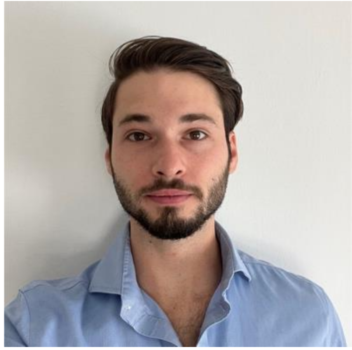


Language Models

- Discriminative: start with a generative model
  - Train another layer to predict a given label
    - “the weather will be sunny” → POSITIVE

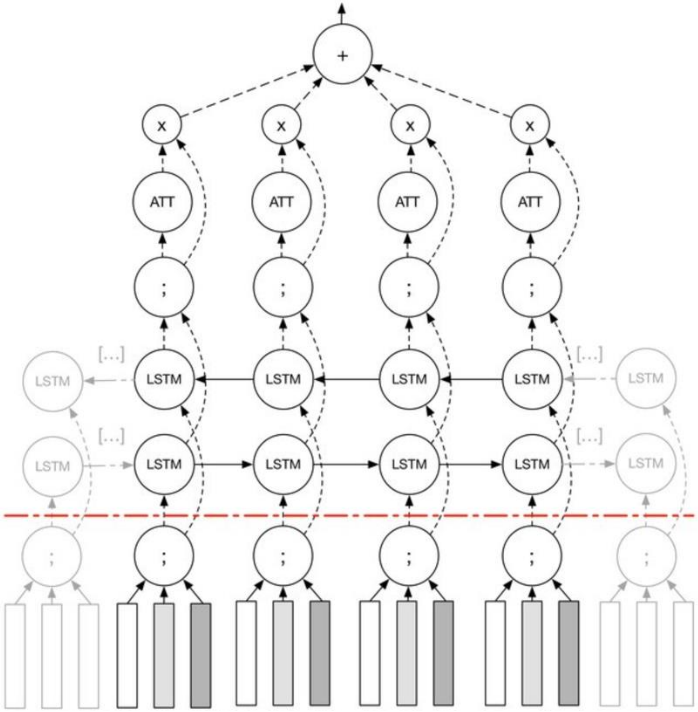


Tracking "depression"

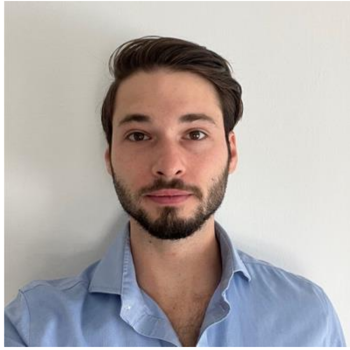


- (Tabak & Purver, EMNLP 2020)
- Learning to recognise the language of depression

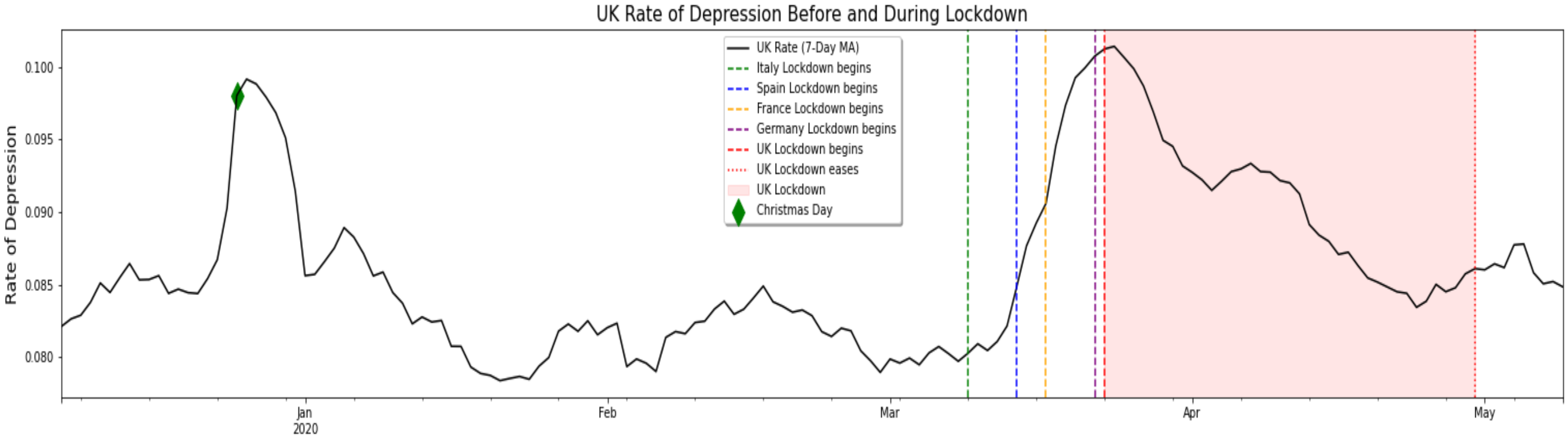
- Collect Twitter timelines with & without diagnosis statements
  - (this is a very noisy way to label data)
- Train a classifier to distinguish the two
- Bi-LSTM with self-attention
  - Per-timeline accuracy OK …
    (… but not great: F1 0.63)



Tracking "depression"



- (Tabak & Purver, EMNLP 2020)
- Learning to recognise the language of depression
- Tracking population depression over time by monitoring Twitter



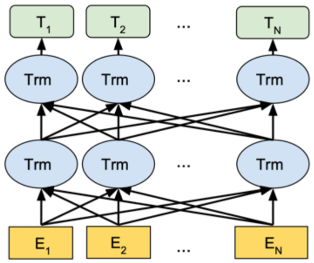
Healthcare applications

- Given the right data, this can be concretely useful …
- Therapy for depression & anxiety (Howes et al., 2014)
  - Diagnosis & severity prediction
  - Early dropout prediction
  - Therapist “quality” prediction

- Dementia diagnosis (Nasreen et al., 2019-21)
- Schizophrenia consultations (Howes et al., 2012)
  - Prediction of symptom severity
  - Prediction of treatment adherence

But: **what** do we learn?

Inside the box: Word Embeddings

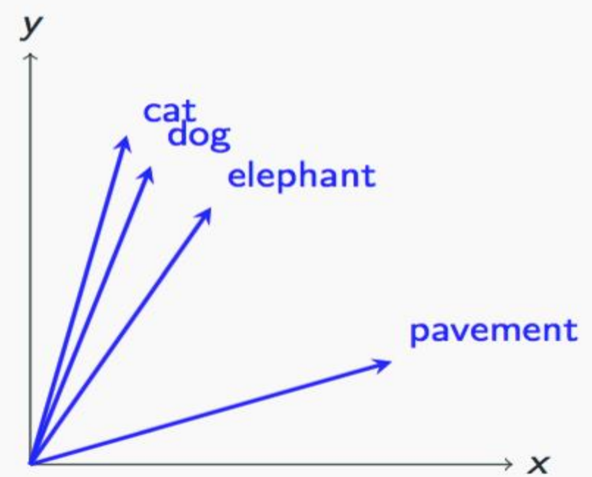


- Learned from word associations in very big datasets
  - (e.g. the web)
- 'Cat' & 'dog' are similar & appear in similar contexts

- Forced to capture lexical and sentential semantics:

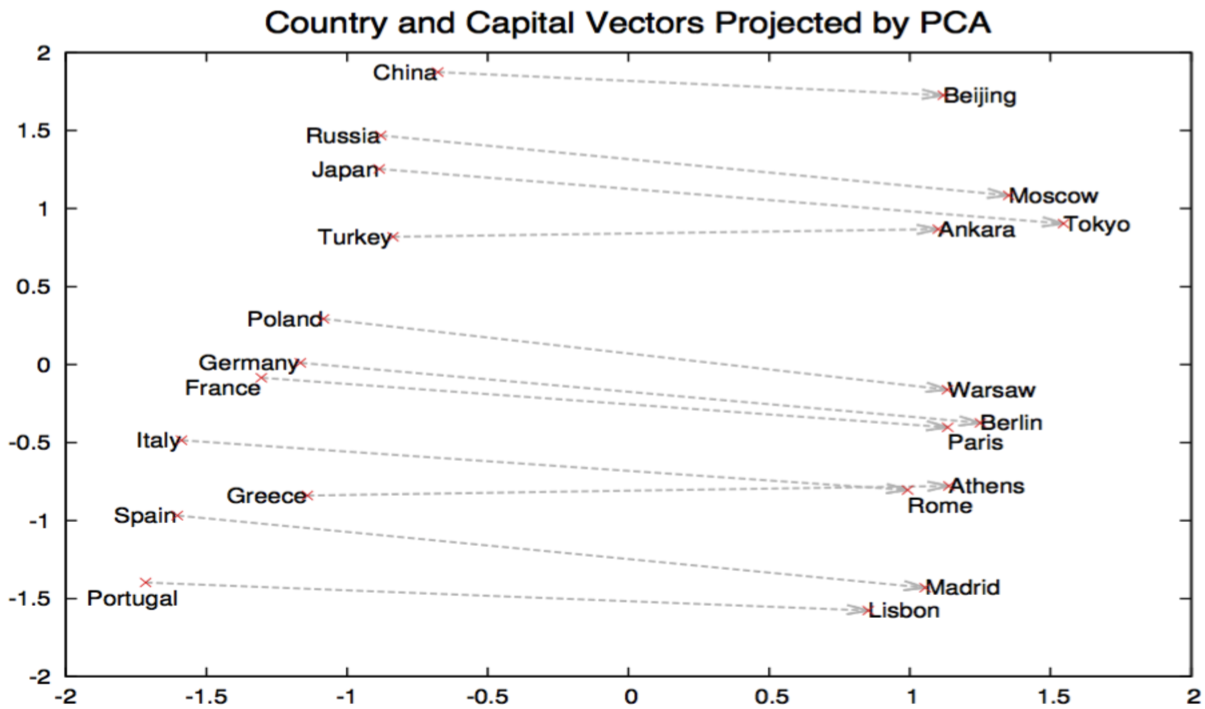
  the musician played the _BLANK_ very well
  the violinist played the _BLANK_ very well
  the actor played the _BLANK_ very well

- No need for any dataset labels!

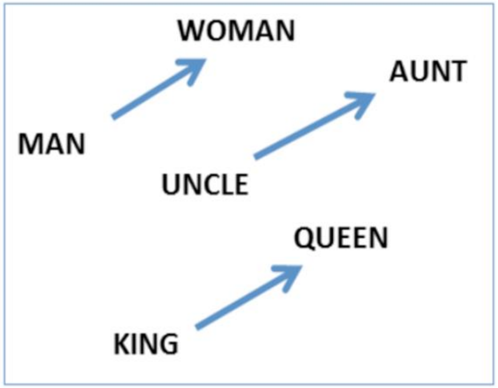



```
% wv.similarity('cat','dog') = 0.7609457089782209
% wv.similarity('cat','elephant') = 0.4638771410889477
% wv.similarity('cat','pavement') = 0.13728373264948163
```

Meaning and analogy



Meaning, analogy … and bias



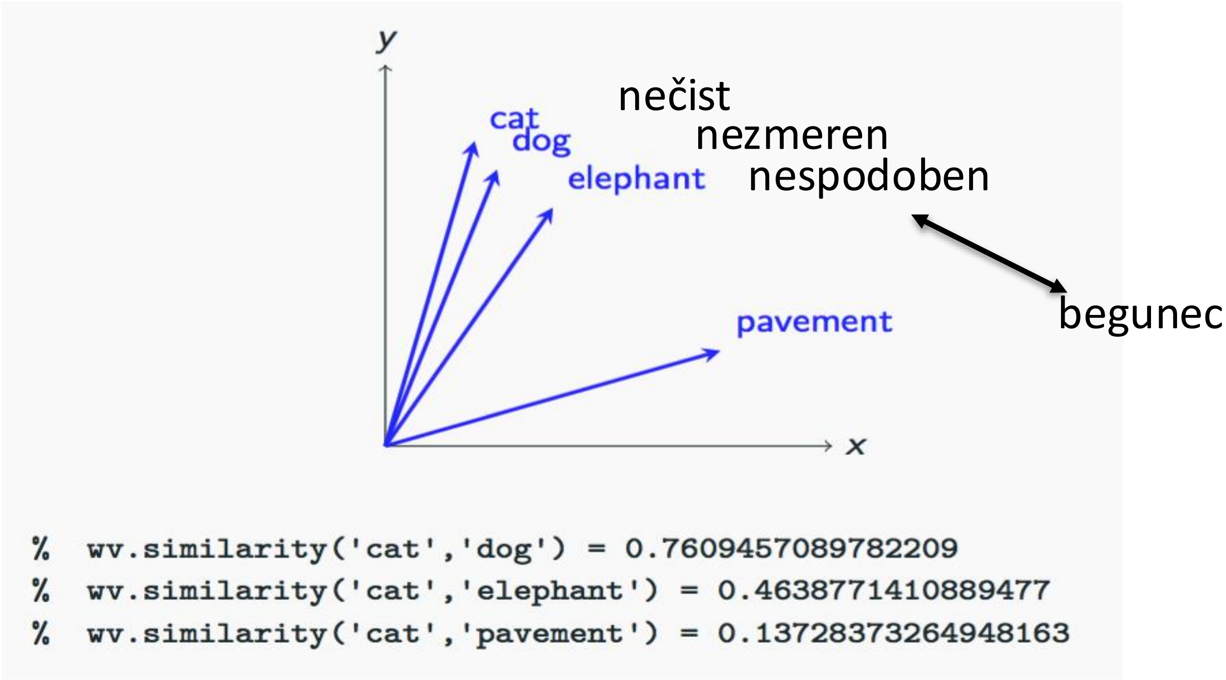
- `king – man + woman = queen`
- `uncle – man + woman = aunt`

- But: Bolukbasi et al (2016)
- `chuckle – man + woman = giggle`
- `pizza – man + woman = cupcakes`
- `surgeon – man + woman = nurse`
- `computer_programmer – man + woman = homemaker`
  - (Effects actually weaker than this suggests (see Nissim et al, 2020) – but they are real)

- Embeddings are **biased**: because language reflects society's biases

Measuring bias

- (Caporusso et al., JADT 2024)
- News media bias against social groups



- Compare news outlets with different political leanings

Measuring bias

Source	Number of documents	Number of words	Category
Train set			
Mladina	5772	2154366	left
Dnevnik	20443	5386894	left
All left	26215	7541260	/
24ur.com	26185	5715921	center
Slovenske novice	30	11059	center
All center	26215	5726980	/
Nova24TV	13095	7210277	right
Tednik Demokracija	13120	6028862	right
All right	26215	13239139	/
All	78645	26507379	/
Test set			
Delo	6553	2605103	left
Siol.net Novice	6553	2982801	center
Revija Reporter	6553	2484408	right
All	19659	8072312	/

Measuring bias

	CS		L-C	L-R	C-R
Migrant (female)	L: C: **R:**	0.116 0.074 **0.167**			
Migrant (male)	**L:** C: R:	**0.262** 0.219 0.227			
Migrant (general)	**L:** C: R:	**0.262** 0.219 0.228			
LGBTQIA+ (female)	L: C: **R:**	0.118 0.121 **0.134**			
LGBTQIA+ (male)	**L:** C: R:	**0.263** 0.217 0.198			
LGBTQIA+ (general)	**L:** C: R:	**0.245** 0.208 0.198			

Measuring dehumanization

- (Caporusso et al., LREC-COLING 2024)

Dehumanization

(A) Negative evaluation

1. Paragraph-level sentiment analysis
2. Migrant Vector Nearest neighbors valence

(B) Denial of agency

1. Migrant Vector Nearest neighbors dominance

(C) Moral disgust

1. Cosine similarity between the Migrant Vector and the Moral Disgust Vector

(D) Dehumanizing metaphor

1. Cosine similarity between the Migrant Vector and the Vermin/Parasite Vector

(after Mendelsohn et al., 2020)

Measuring dehumanization

- Compare Slovene news media across time

2014-2015
Migration crisis following the war in **Syria**

2022-2023
Migration crisis following the war in **Ukraine**

Specifically mentioning Ukraine

Not mentioning Ukraine

Measuring dehumanization

- Testing hypotheses – we expect:
  - H1: less dehumanization during Ukraine period than Syria period
  - H2: less dehumanization when mentioning Ukraine than when not

H1

"Moral disgust" vector:

Sim_UKRAINE > Sim_SYRIA

"Vermin" vector:

Sim_UKRAINE > Sim_SYRIA

H2

"Moral disgust" vector:

Sim_UKRAINE < Sim_OTHER

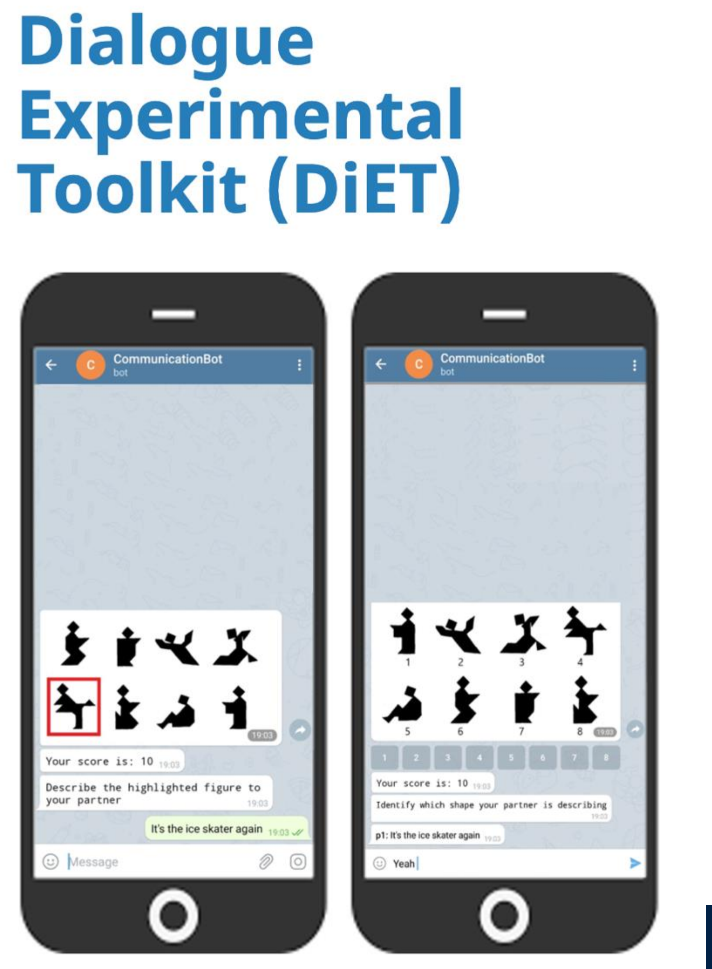
"Vermin" vector:

Sim_UKRAINE ≈ Sim_OTHER

- (i.e. less dehumanization when discussing Ukraine, but generally more over time)

Interaction as a Probe

- Things get more interesting with interactive language …



https://dialoguetoolkit.github.io/

https://clp-research.github.io/slurk

Modelling relationships

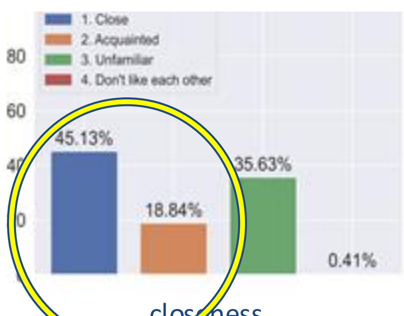
Setting 1: Private Conversations with Self-Reported Relationships

- Ask people to have a conversation in our platform (using Slurk)
- Ask them to fill in a form to identify their relationship in terms of closeness and respect

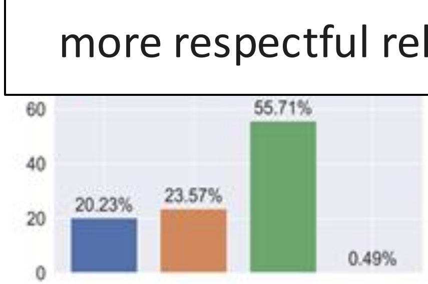
Setting 2: Public Conversations with Perceived Relationships

- Collect conversations from X (Twitter)
- Ask 3 annotators to labels the degree of closeness/respect they perceived from the conversation (the responder perspective)

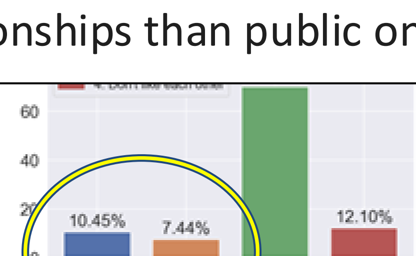
Private conversations lean towards closer and more respectful relationships than public ones



closeness



respect



closeness

respect

Modelling relationships

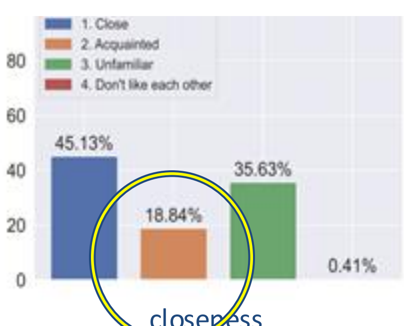
Setting 1: Private Conversations with Self-Reported Relationships

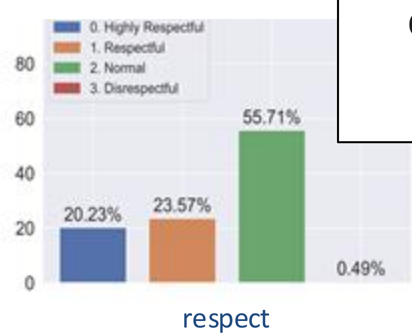
- Ask people to have a conversation in our platform (using Slurk)
- Ask them to fill in a form to identify their relationship in terms of closeness and respect

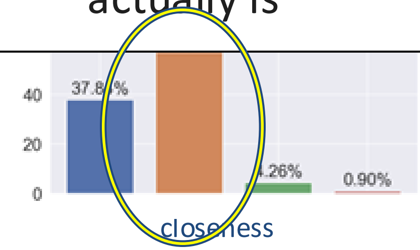
Setting 3: Private Conversations with Perceived Relationships

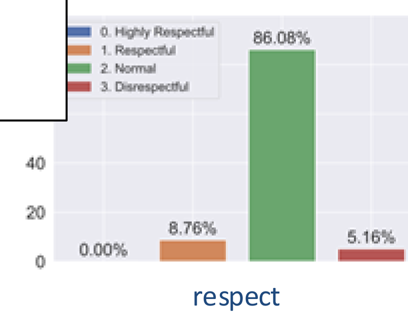
- Ask people from setting 2 to annotate conversations from setting 1

Third-party observers tend to perceive the degree of closeness as higher than it actually is









Computational models

- Fine-tuned PhayaThaiBERT -- Thai-specific 110-million parameter LM

Model	Task1: Closeness			Task2: Respect		
	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived	Setting 1 Private-Self	Setting 2 Public-Perceived	Setting 3 Private-Perceived
Baseline						
Majority-class Baseline	0.155	0.206	0.401	0.179	0.276	0.308
Naive Bayes Classifier	0.563	0.435	0.542	0.470	0.678	0.535
Logistic Regression	0.400	0.327	0.542	0.314	0.444	0.463
LMs						
XLM-R	0.604	0.420	0.498	0.200	0.675	0.432
WangChanBERTa	0.657	0.490	0.639	0.313	0.748	**0.761**
PhayaThaiBERT	**0.666**	**0.496**	**0.657**	**0.431**	**0.750**	0.712

Table 1: The f1 performance metrics of our social relationship models in the closeness and respect tasks across three conversational settings

- Hard to predict other-perceptions of closeness in public settings
  - Inter-annotator agreement is OK: κ = 0.61
- Hard to predict self-reports of respect in private settings
  - Agreement with self one month later is also poor: κ = 0.22

Inspecting model behaviour

1. Fine-tuned PhayaThaiBERT -- Thai-specific 110-million parameter LM
2. Calculate SHAP of selected relevant lexical features
   - **Pronouns**: a well-studied lexical feature known for their social functionality across many languages
   - **Sentence-final particles**: a lesser-known social-related feature observed in a narrower range of languages, primarily East and Southeast Asian languages
   - **Spelling variation**: a recent linguistic pattern that has gained recognition for its potential semantic functions in internet language

Inspecting model behaviour

- Hypothesis testing:
- Pronouns a pivotal contributor to predictions
  - 1st person pronouns contribute in all settings
  - 2nd person pronouns with private settings
  - 3rd person pronouns only with perceived closeness in private conversations
- Socially-related particles are important
- Some spelling variations really matter
  - Morphophonemic variation & non-standard pronouns

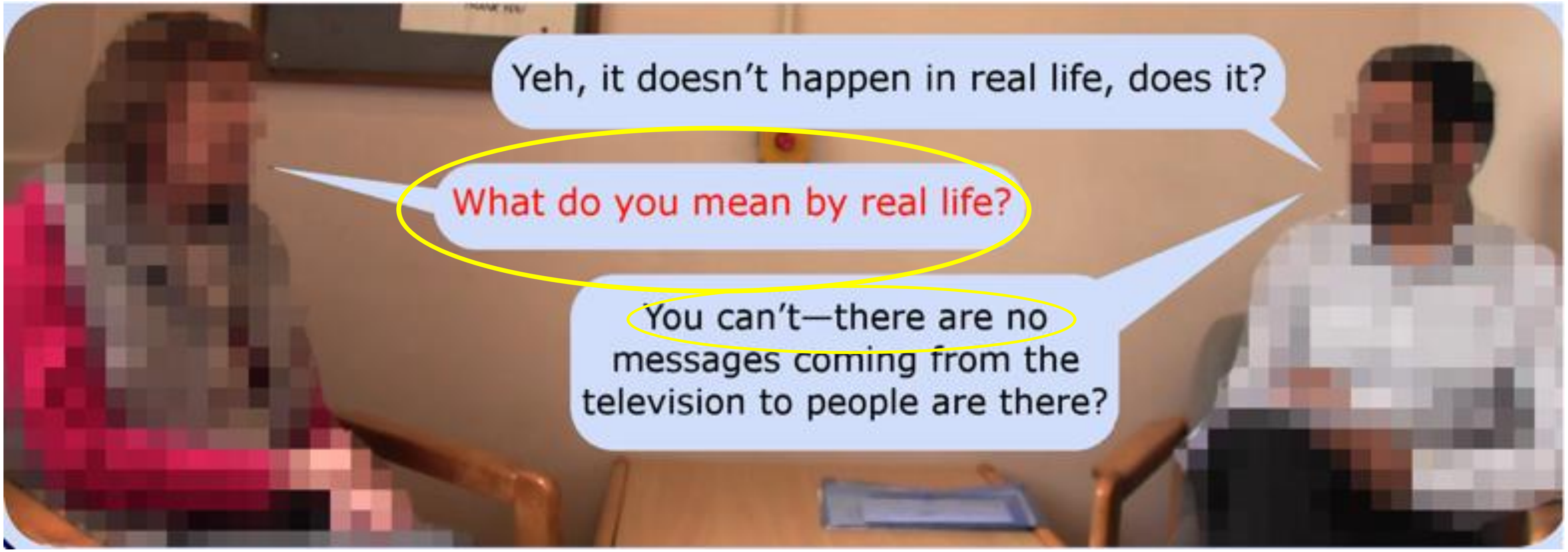
Lexical Features	Setting 1 Private-Self		Setting 2 Public-Perceived		Setting 3 Private-Perceived	
	Per token	Total	Per token	Total	Per token	Total
Reference						
Average per token	1.08	125.36	4.07	147.01	0.85	97.91
Pronoun						
All pronoun	1.13	4.05	4.52	9.47	1.60	5.65
» 1st person pronoun	1.25	2.85	5.15	7.73	1.14	2.56
» 2nd person pronoun	1.30	3.29	4.33	7.68	2.04	5.11
» 3rd person pronoun	0.71	1.31	3.47	5.61	1.71	3.14
» Singular pronoun	1.13	4.04	4.52	9.40	1.60	5.65
» Plural pronoun	1.07	1.07	4.30	5.73	0.49	0.49
» Pronoun in non-standard spelling	0.74	1.58	7.62	10.02	1.23	2.44
Sentence-final Particles						
All particles	1.75	8.81	4.16	7.54	0.93	4.68
» Socially-related particles	3.24	10.03	5.08	7.27	1.31	4.08
» Non-socially-related particles	0.85	2.97	3.47	5.45	0.69	2.43
» Particle in non-standard spelling	1.33	1.86	7.83	8.41	1.11	1.36
Spelling Variation						
All spelling variation	1.10	14.48	4.39	19.46	0.86	11.28
» Common misspelt words	0.83	1.29	3.80	5.24	0.80	1.24
» Morphophonemic variation	1.26	10.49	5.37	15.10	0.95	7.91
» Simplified variation	0.90	5.81	3.63	10.79	0.74	4.77
» Repeated characters	0.85	1.82	3.41	4.47	0.54	1.15

Table 2: The average of absolute SHAP values of three lexical features in **closeness tasks** across 3 conversational settings from **fine-tuned PhayaThaiBERT**. The values highlighted in grey denote values exceeding the SHAP values of their respective random baseline

So …

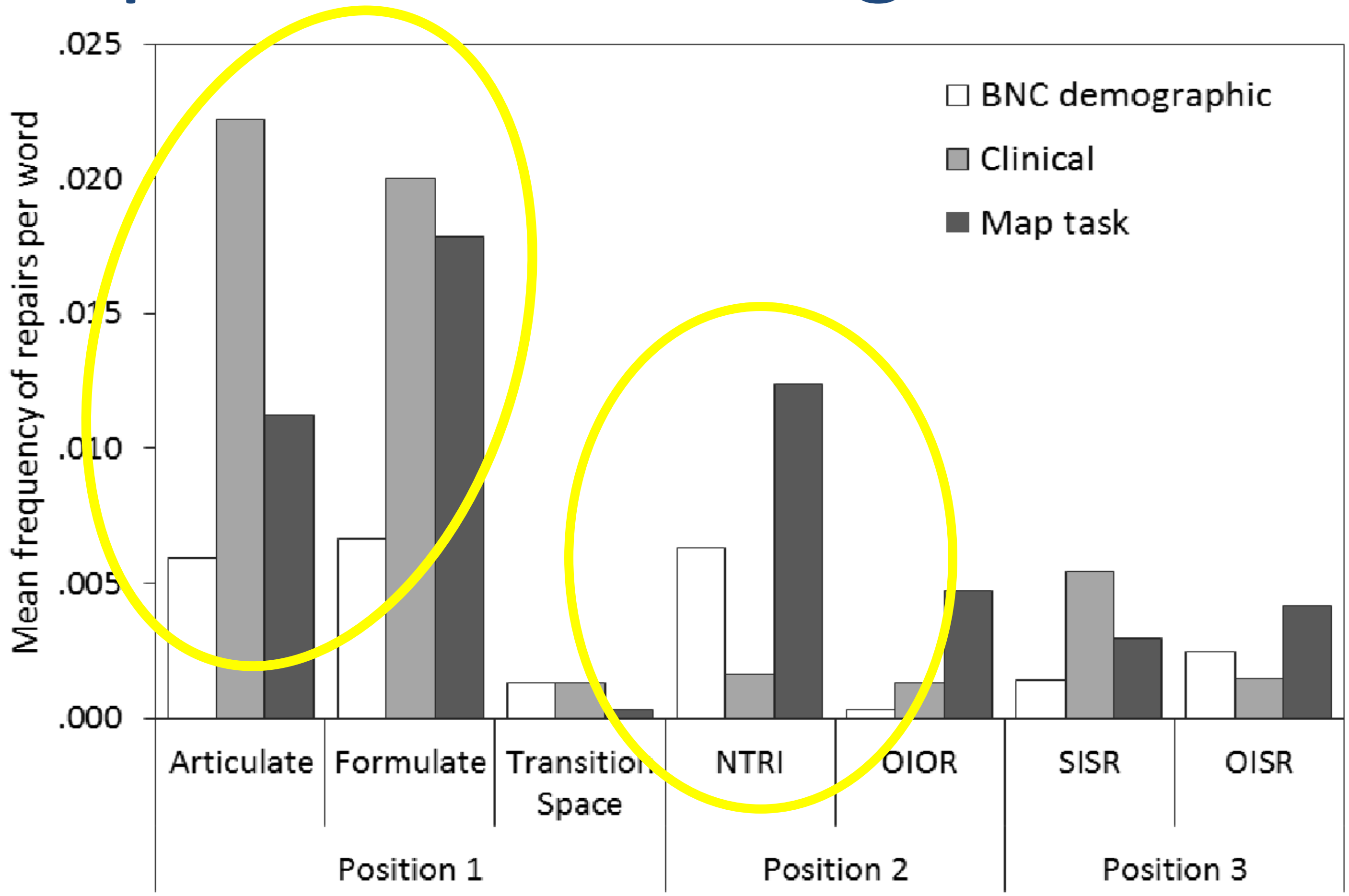
- When the linguistic phenomena are simple, LLMs can find them and use them …
- … but what if they’re not?

Schizophrenia & **Repair**



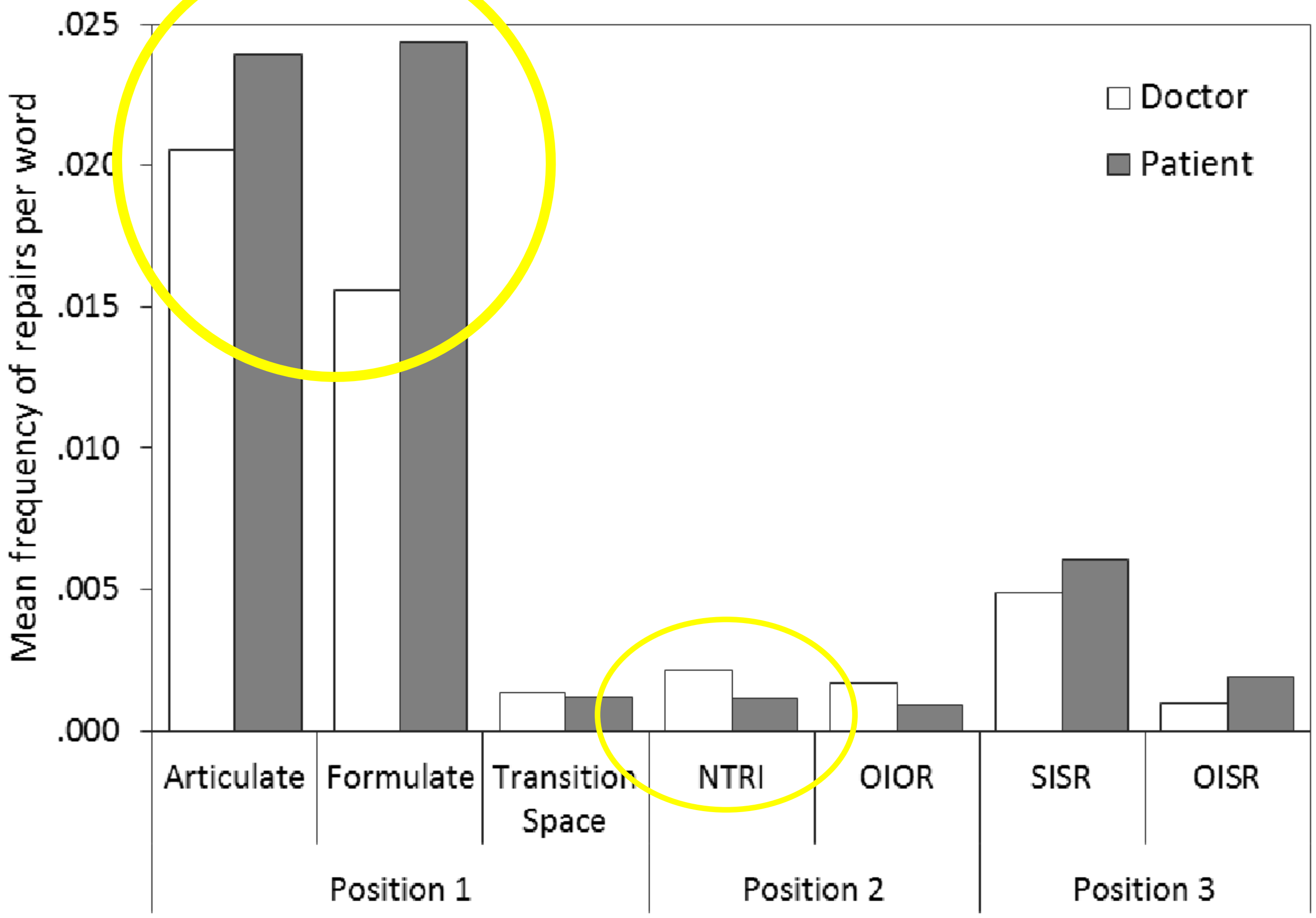
- Schizophrenia study: manual linguistic analysis
  - Significant role of *repair*
  - Patient-initiated other-repair & self-repair

Compare other dialogue contexts



- Therapy: more self-repair, less other-repair & initiation

Patient-doctor comparison



- Patients: more self-repair, less other-repair & initiation

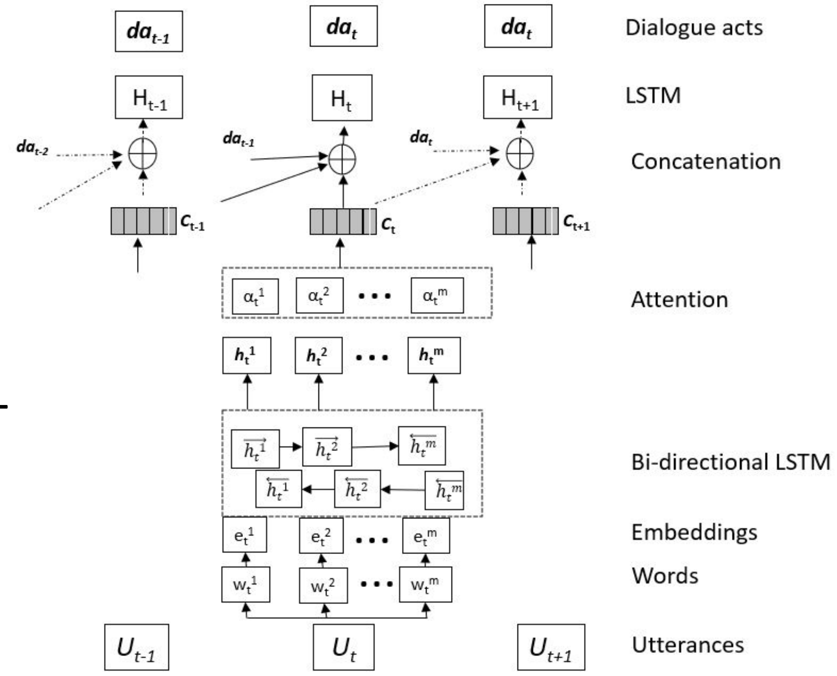
But …

- Experiments with automatic other-repair detection didn't help:
  - A very sparse problem (e.g. <1% of turns)
  - Only 35-44% F-scores on real data (above 20-36% baselines)
  - Needs a general measure of parallelism
  - Needs vocabulary-independence

Dementia & **Repair**

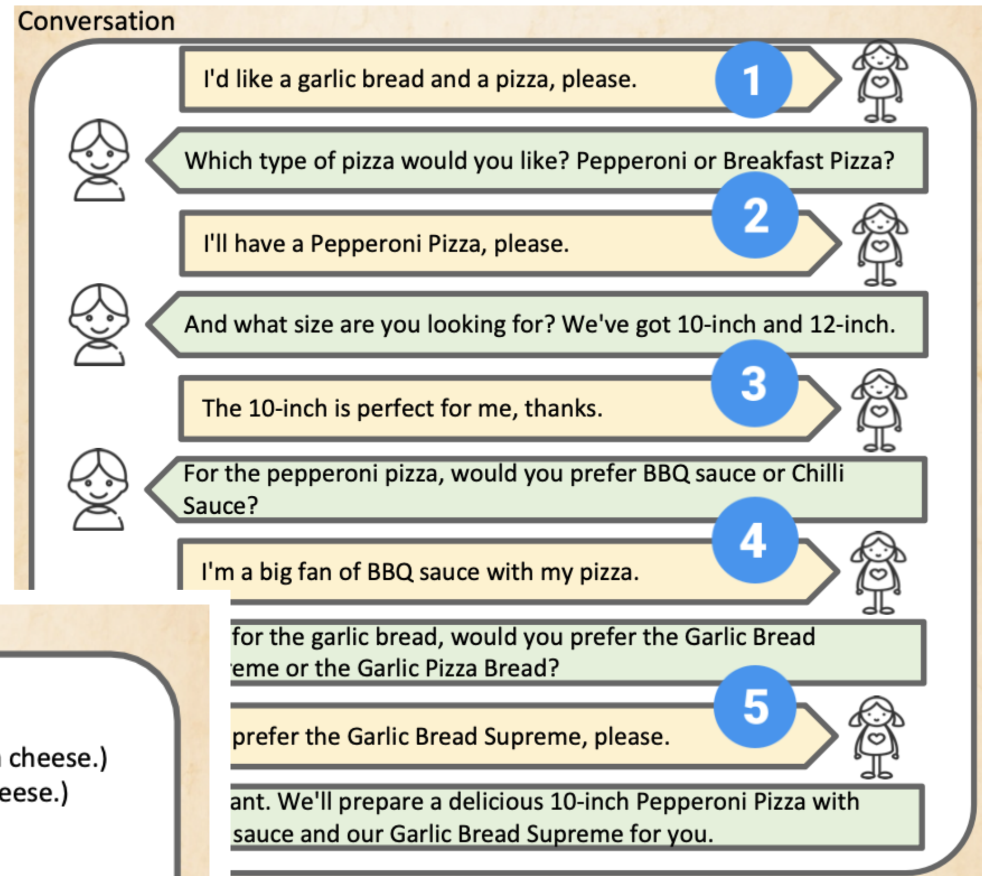
- Repair also significant with dementia & cognitive impairment
  - Self-repair: individual cognitive difficulties
  - Other-repair: lack of understanding, avoidance/delay strategies, prompting from others …

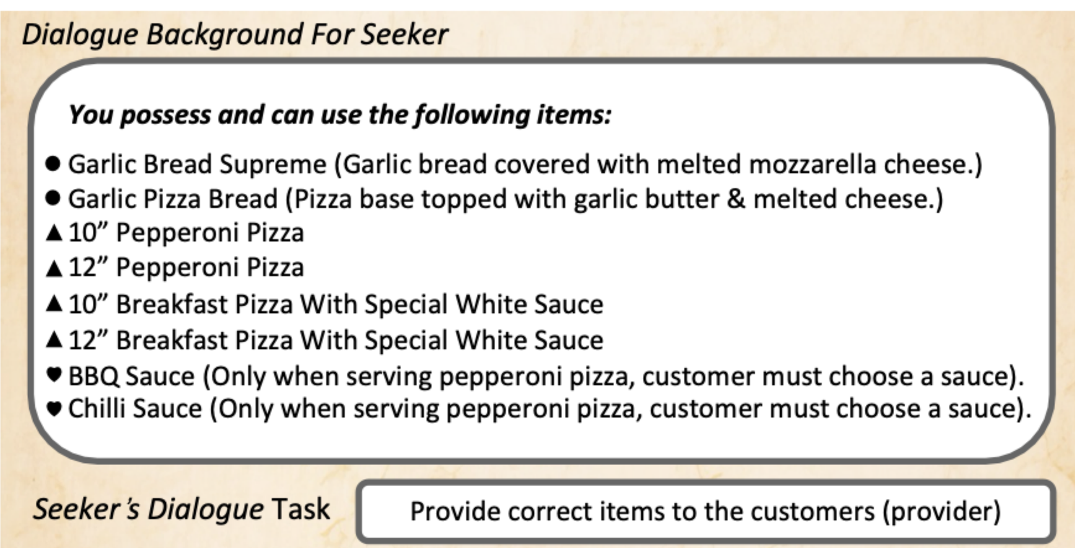
- Structured NN+CRF to detect relevant dialogue acts (Nasreen & Purver, 2019-2021)
  - SotA performance by some distance
  - But still not great: 0.5-0.6 macro F1
  - (This gives 0.7-0.8 F1 in diagnosis)



Can LLMs help?

- The new generation of LLMs can help, right?
  - Actually, not very much!
- New benchmark for clarification behaviour (Gan et al., 2024 & forthcoming)
- LLAMA3.1 405B gets only 60%

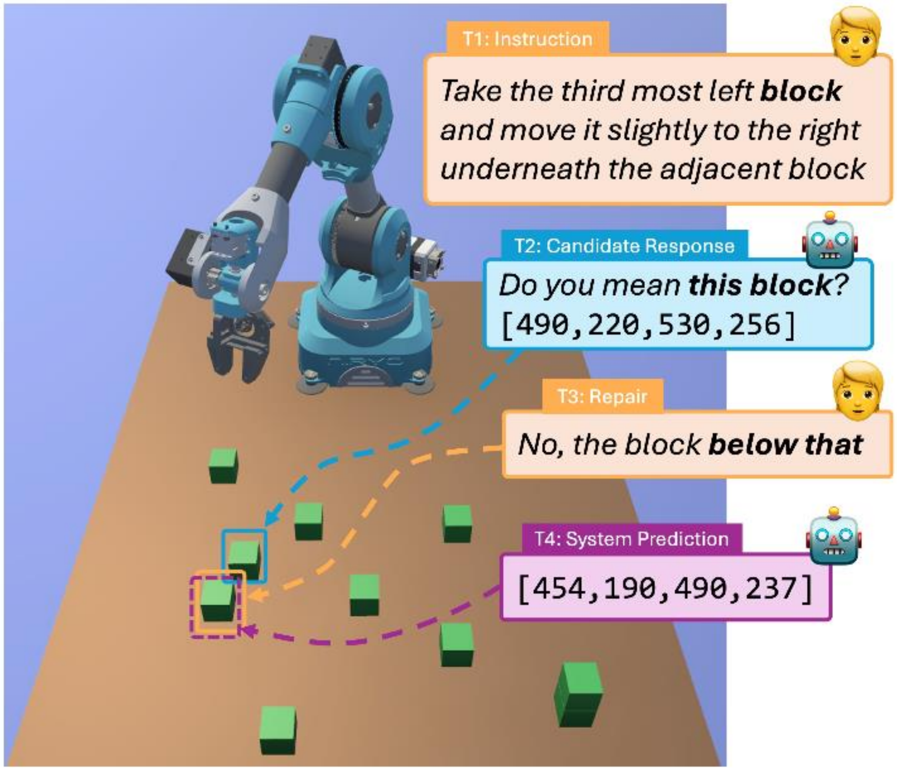




Can LLMs help?

- See also (Chiyah-Garcia et al., 2024)

- GPT-4o 26-50% accuracy
- (Humans 68-75% accuracy)



Where next for LLMs?

- For this kind of work, we need:
  - Language models that are inspectable
  - Language models for rare but important phenomena

- How do we get there?
  - Improved training regimes?
  - More suitable benchmarks (datasets, metrics)?
  - Improved explainability methods
  - Better understanding of how we want to use LLMs!