Large Language Models Know What To Say But Not When To Speak

Muhammad Umair, Vasanth Sarathy, J.P de Ruiter

Empirical Methods in Natural Language Processing (EMNLP) - 2024

Department of Computer Science
Tufts University

Turn-Taking Ensures Understanding

Smooth turn-taking in human interaction ensures a minimum level of understanding[1].

1 Turn-taking is rapid (200 ms on average)[2].

2 Speaker allocation occurs on a turn-by-turn basis[2].

3 Deviations from normative timing are used to convey social information[2].

Turn-Timing in Spoken Dialogue Systems

SDS **fail** to replicate human-like naturalistic turn **timing.**

This adversely affects user experience in several ways.[3, 4, 5]

1 The system interrupts speakers (e.g., untimely feedback).

2 The system cannot exploit norms to convey information (e.g., delayed response).

3 Humans attribute the system as being the trouble source and react in marked ways (frustration, amusement etc.).

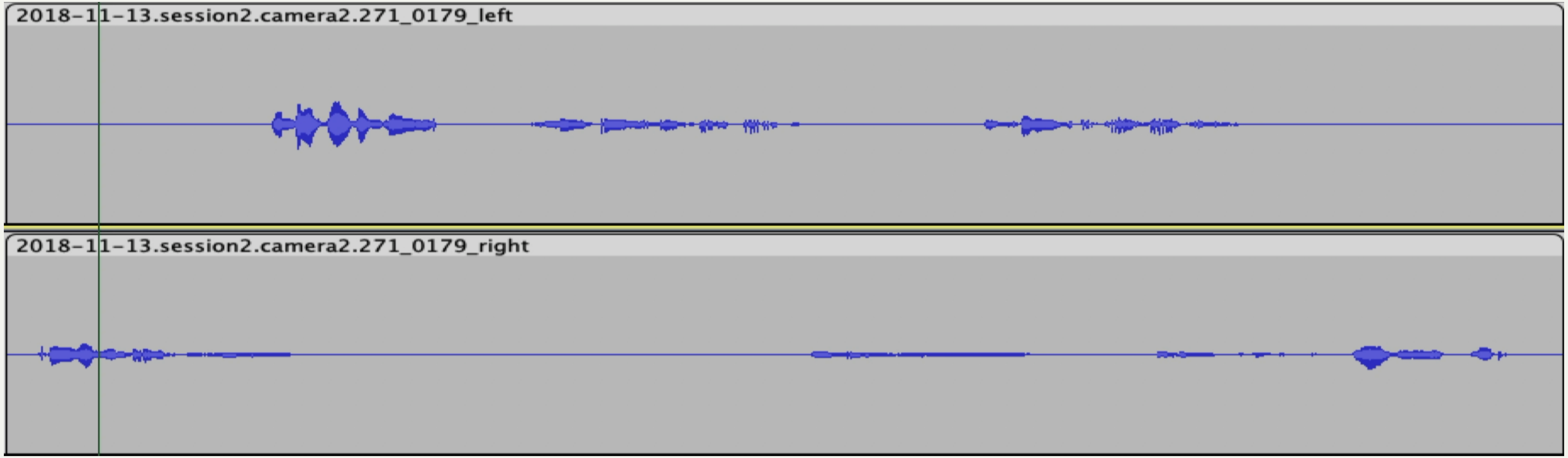
Can LLMs be Used to Improve *Naturalistic* Turn-Timing in SDS?

LLMs have shown promise in producing conversational content, and even identifying turn ends. However, LLM-based approaches face **two major challenges.**

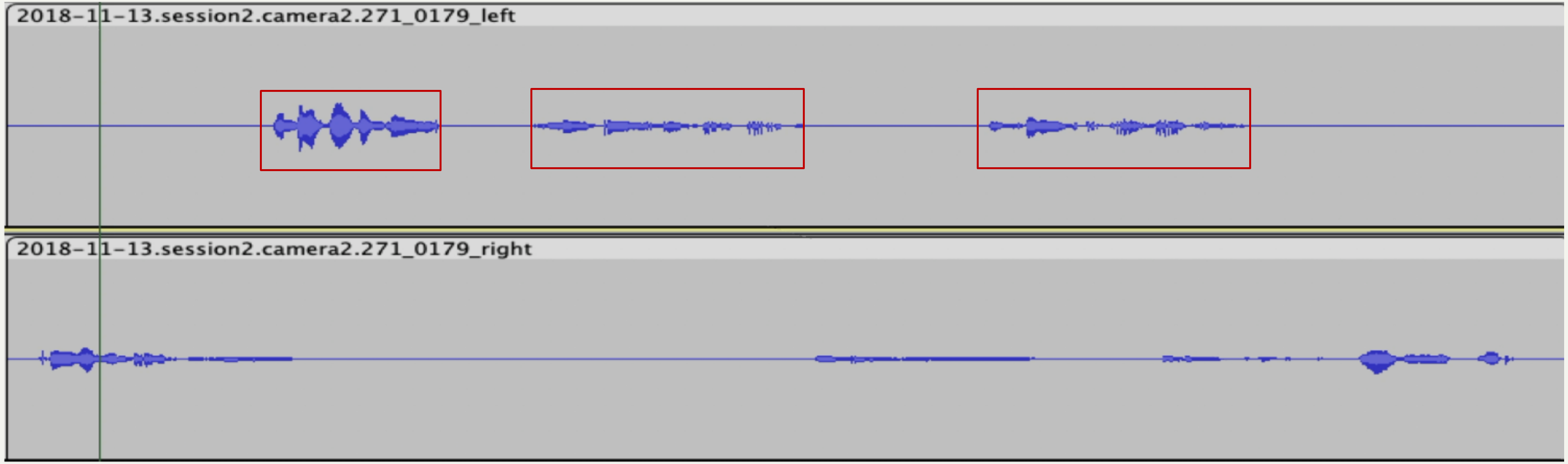
1 Opportunities to speak **within-turns** are difficult to identify.

2 Most LLMs are trained on written-first language, which differs significantly in structure and usage from spoken language.

Turn-Taking in Natural Conversation

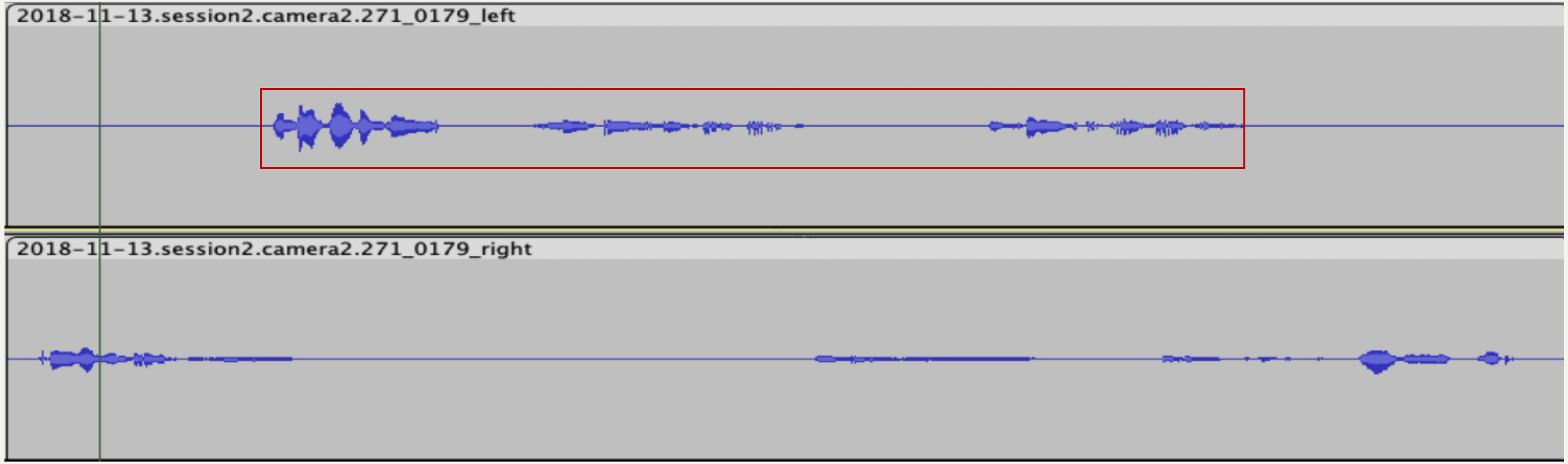

2018–11–13.session2.camera2.271_0179_left
2018–11–13.session2.camera2.271_0179_right

Atoms of Turn-Taking


2018-11-13.session2.camera2.271_0179_left
2018-11-13.session2.camera2.271_0179_right

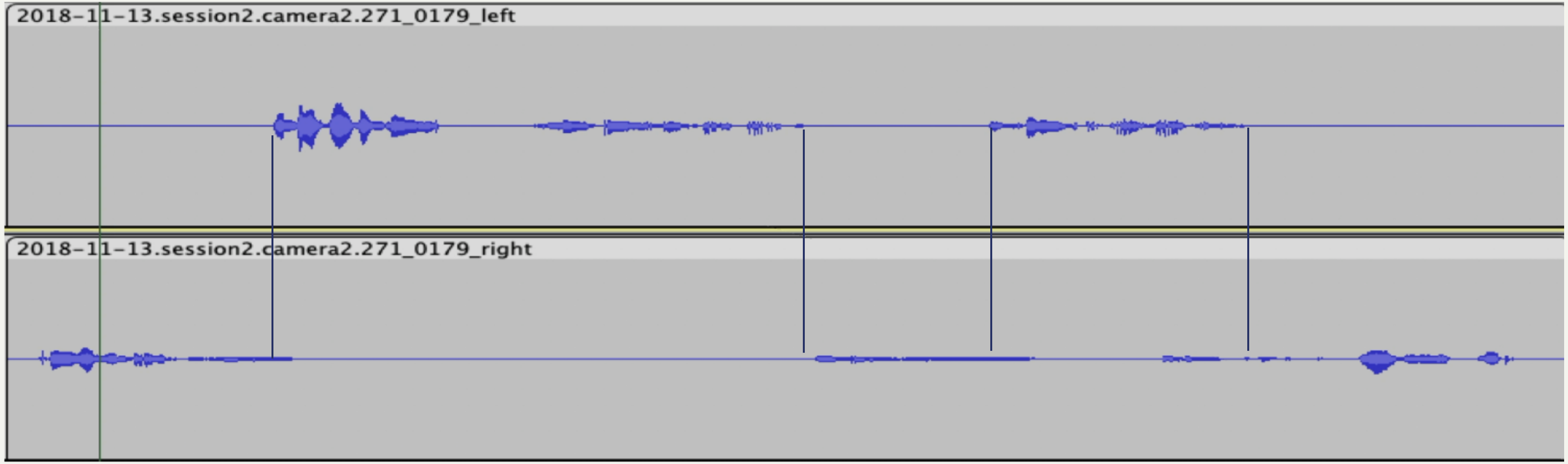
A *Turn Construction Unit* (TCU) is an **atom in turn-taking** that may be a word, phrase, or sentence that is standalone and makes full sense in the context[6,7].

Turns: Same-Speaker TCUs


2018-11-13.session2.camera2.271_0179_left
2018-11-13.session2.camera2.271_0179_right

A *turn* consists of one or more TCUs by the same speaker, typically within 1000 ms of each other.[6,7].

Opportunities for Transition


2018-11-13.session2.camera2.271_0179_left
2018-11-13.session2.camera2.271_0179_right

Transition Relevance Places (TRPs) present *opportunities* for turn-transition where a listener may, but is **not obligated**, to speak.

TRPs Between Turns (Switches)

TRPs where a speaker-switch occurred, which can be easily identified retroactively.

Speaker-1 I find that I am very tired after small physical activity,
I am wondering if perhaps I have a condition.

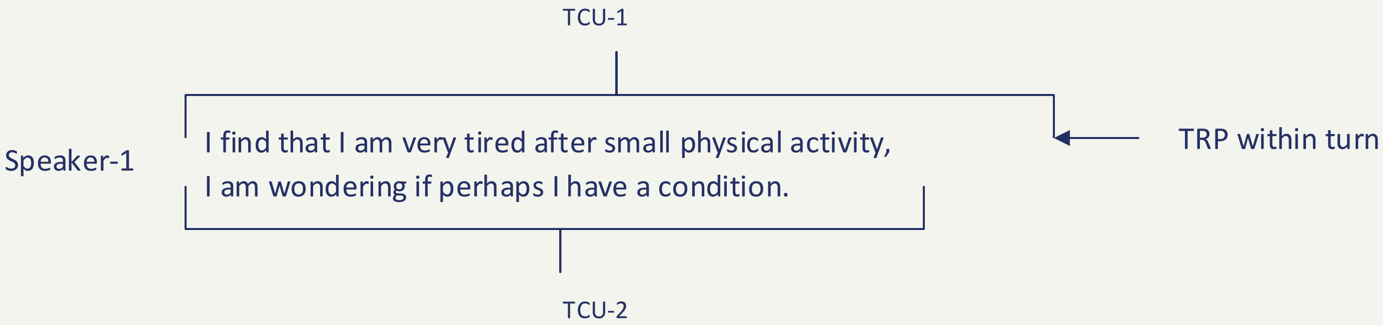
← TRP between turns

Speaker-2 I don’t think you do!

Transition Relevance Places (TRPs) are opportunities for turn transition that occur between TCUs.
Turn Construction Units (TCUs) are atoms of turn taking that encompass sentential, phrasal, and lexical constructions.

TRPs Within Turns (Continuations)

TRPs where a speaker switch might have but **did not** occur, which is difficult to identify retroactively.



TCU-1

Speaker-1

I find that I am very tired after small physical activity,
I am wondering if perhaps I have a condition.

TRP within turn

TCU-2

Transition Relevance Places (TRPs) are opportunities for turn transition that occur between TCUs.
Turn Construction Units (TCUs) are atoms of turn taking that encompass sentential, phrasal, and lexical constructions.

1

2

Locating Opportunities for Turn-Taking

Inquiry

How do we identify *opportunities* to speak (TRPs) *within turns* in natural conversations?

Approach

Develop an experimental paradigm to identify opportunities for transition in an ecologically valid manner.

Tufts UNIVERSITY

TRPs: The Data Problem

Existing corpora only **reliably** annotate TRPs at turn switches, which are a small subset of all TRPs[8].

1. There is individual variability in responses at TRPs.
2. TRPs within-turns are difficult to retroactively identify with high *ecological validity.*

We want to develop an experimental paradigm to annotate **TRPs within turns** with *high ecological validity* i.e., with a focus on generalizability.

Between turn TRPs are TRPs where a turn switch occurred.
Within-turn TRPs are TRPs where a listener might have, but did not, take a turn.

Stimulus Contains Multiple TRPs

A *stimulus* is a turn that was originally **one side of a dialogue.**
It contains multiple opportunities (TRPs) where the listener may respond.

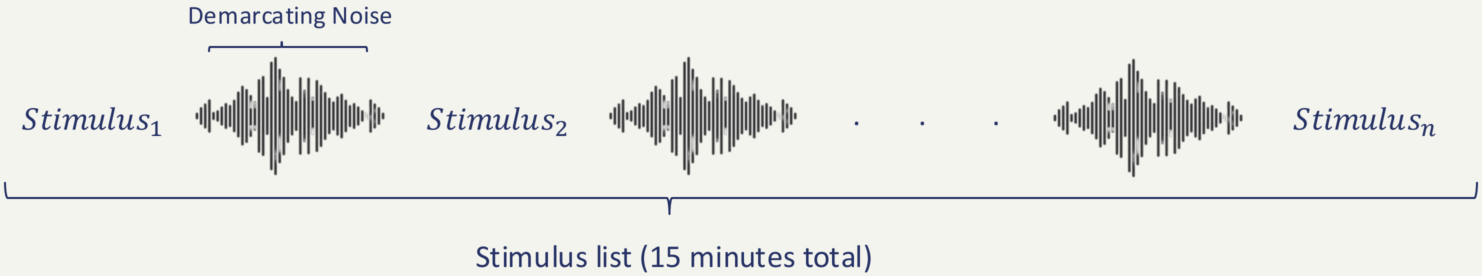
$Stimulus_i$

Data was collected from the **In Conversation Corpus (ICC)**, a high-quality speaker-separated corpus of natural interaction collected at the Tufts Human Interaction lab.

Stimulus Lists are a Set of Stimuli

A *stimulus list* is a collection of **independent** stimuli that are separated by a sound (to indicate a new stimulus).

We used four stimulus lists (two distinct lists and their reversals).



Demarcating Noise

$Stimulus_1$ $Stimulus_2$ . . . $Stimulus_n$

Stimulus list (15 minutes total)

Participants Respond at TRPs

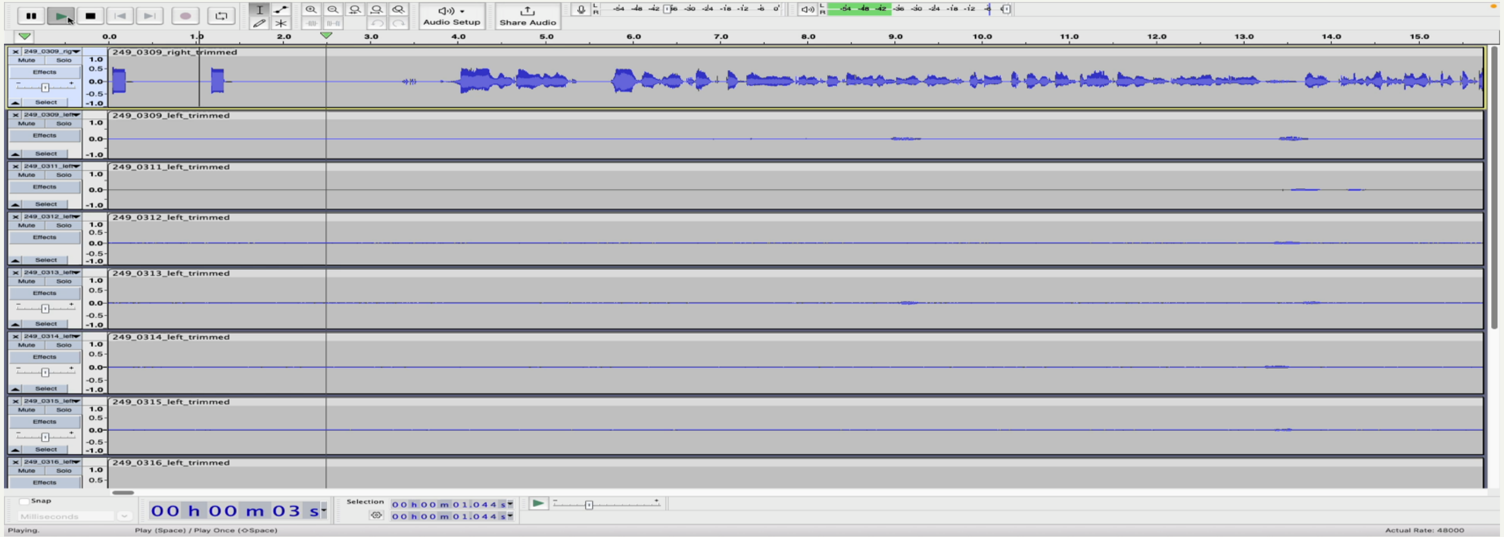
We recruited **120 participants such that 30 participants** responded to each list.

Participants were asked to listen to a stimulus list as if they were part of the dialogue, thus using the same anticipatory process in natural conversations.

Each Participant produced **one-word responses** at **as many points** as they judged to be appropriate.

Actual Participant Responses

We expect to see a **distribution** of participant responses centered around some ‘true’ within-turn TRP.



Locating Opportunities for Turn-Taking

1

Inquiry

We would now like to predict within-turn TRPs in **spoken language**. How effective are LLMs as a baseline for this prediction, given their training on vast amounts of written-first language?

2

Approach

We formulate a binary decision task for models to predict TRPs based on preceding linguistic information and measure performance through various metrics.

Formalism of Turn Components

A stimulus (S) has N words and K responses.

$$S = \langle (w_1, t^s_{w_1}, t^e_{w_1}), \ldots, (w_n, t^s_{w_n}, t^e_{w_n}) \rangle$$
$$R = \langle (\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}), \ldots, (\tilde{w}_M, t^s_{\tilde{w}_M}, t^e_{\tilde{w}_M}) \rangle$$

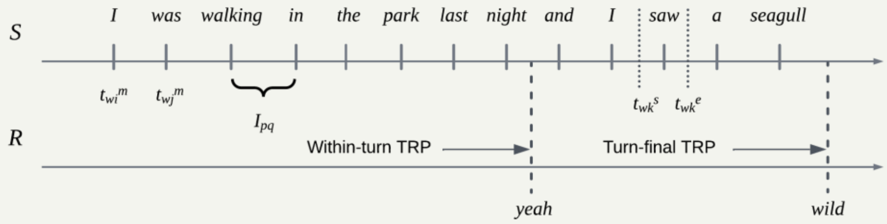
A Prefix is the set of words in S from the first to the i-th word.

$$P_i \quad = \langle w_1, \ldots, w_I \rangle ; \forall w_i \in P_i, w_i \in S$$

T

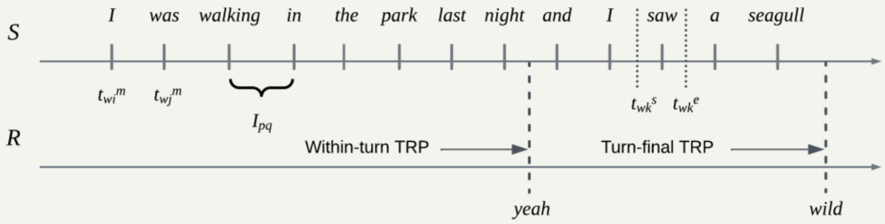
T_i is a binary R.V for intervals ($I_{i,j}$) between words, and $T_{R,S}$ as the set of predictions after each word.

$$I_{i,j}, 1 \leq i, j \leq N, j = i + 1$$
$$T_{R,S} = \langle T_1, \ldots, T_N \rangle$$



Formalism of Turn Components

Task Definition: *Given a stimulus S, and the set of all prefixes P_S, where each T_i in $T_{R,S}$ occurs after each of the prefixes in P_i in P_S*



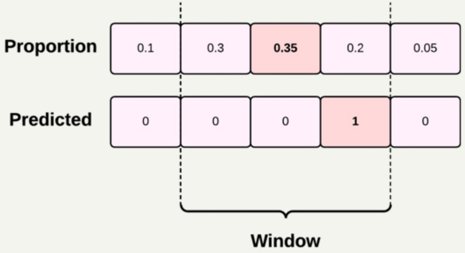
Evaluation Metrics

We evaluated the ability of models to predict TRPs using a range of metrics.

1 **Classification Metrics** measure binary labeling task performance, while considering class imbalance.

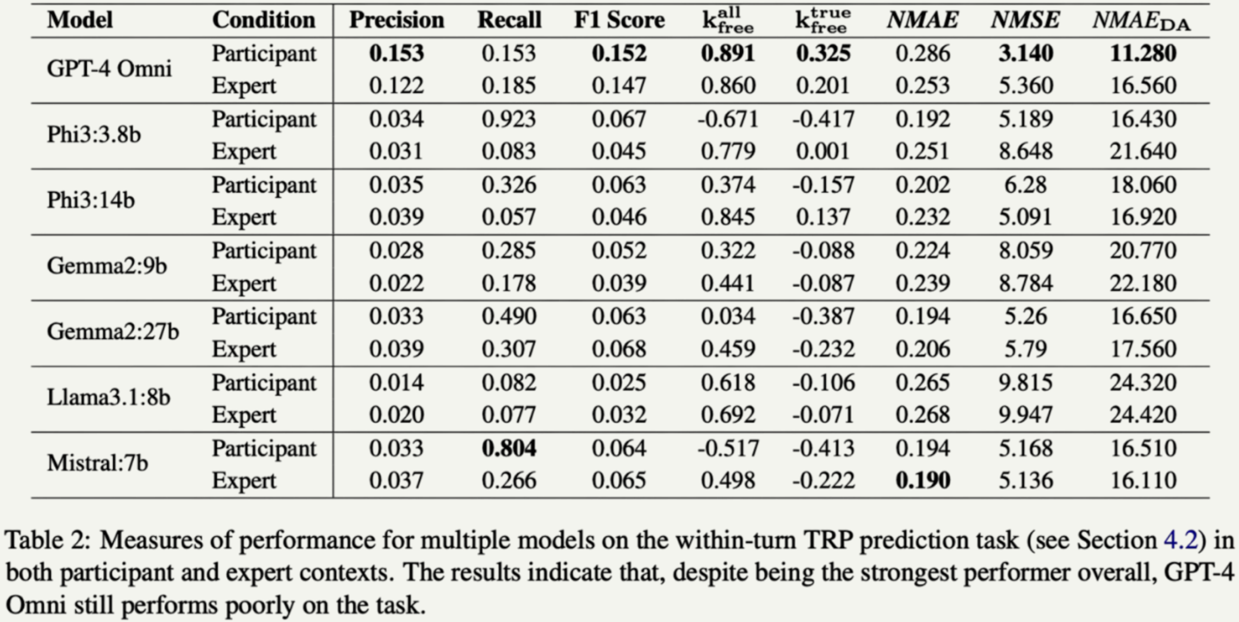
2 **Temporal Metrics** provide a measure of how far away model predictions participant-agreed TRPs.

2 **Agreement Metrics** determine how well models agree with each other and participants over chance.



Proportion	0.1	0.3	**0.35**	0.2	0.05
Predicted	0	0	0	**1**	0

Results



Model	Condition	Precision	Recall	F1 Score	$k_{\text{free}}^{\text{all}}$	$k_{\text{free}}^{\text{true}}$	*NMAE*	*NMSE*	*NMAE*$_{\text{DA}}$
GPT-4 Omni	Participant	**0.153**	0.153	**0.152**	**0.891**	**0.325**	0.286	**3.140**	**11.280**
	Expert	0.122	0.185	0.147	0.860	0.201	0.253	5.360	16.560
Phi3:3.8b	Participant	0.034	0.923	0.067	-0.671	-0.417	0.192	5.189	16.430
	Expert	0.031	0.083	0.045	0.779	0.001	0.251	8.648	21.640
Phi3:14b	Participant	0.035	0.326	0.063	0.374	-0.157	0.202	6.28	18.060
	Expert	0.039	0.057	0.046	0.845	0.137	0.232	5.091	16.920
Gemma2:9b	Participant	0.028	0.285	0.052	0.322	-0.088	0.224	8.059	20.770
	Expert	0.022	0.178	0.039	0.441	-0.087	0.239	8.784	22.180
Gemma2:27b	Participant	0.033	0.490	0.063	0.034	-0.387	0.194	5.26	16.650
	Expert	0.039	0.307	0.068	0.459	-0.232	0.206	5.79	17.560
Llama3.1:8b	Participant	0.014	0.082	0.025	0.618	-0.106	0.265	9.815	24.320
	Expert	0.020	0.077	0.032	0.692	-0.071	0.268	9.947	24.420
Mistral:7b	Participant	0.033	**0.804**	0.064	-0.517	-0.413	0.194	5.168	16.510
	Expert	0.037	0.266	0.065	0.498	-0.222	**0.190**	5.136	16.110

Table 2: Measures of performance for multiple models on the within-turn TRP prediction task (see Section 4.2) in both participant and expert contexts. The results indicate that, despite being the strongest performer overall, GPT-4 Omni still performs poorly on the task.

Takeaways

1 LLMs **struggle** to predict within-turn TRPs despite various ICL strategies and their pre-training on vast amounts of language.

2 High performance on written language tasks **does not** translate to high performance on normative spoken language tasks.

3 Our research highlights this gap, which **limits** dialogue systems ability to use non-verbal cues to provide social information.

4 We contribute a specialized empirical **dataset of participant-labeled TRPs** and establish **baseline performance** on the TRP prediction task.

Tufts UNIVERSITY

Acknowledgements

Dr. Julia Mertens

Grace Hustace

hi Lab

Tufts UNIVERSITY

References

[1] Stivers, T., Enfield, N. J., Brown, P., Englert, C., Hayashi, M., Heinemann, T., Hoymann, G. Rossano, F., De Ruiter, J. P., Yoon, K.-E., et al. Universals and cultural variation in turn-taking in conversation. Proceedings of the National Academy of Sciences 106, 26 (2009), 10587–10592.

[2] Levinson, Stephen C., and Francisco Torreira. "Timing in turn-taking and its implications for processing models of language." *Frontiers in psychology* 6 (2015): 731.

[3] Skantze, Gabriel. "Turn-taking in conversational systems and human-robot interaction: a review." *Computer Speech & Language* 67 (2021): 101178.

[4] Ram, Ashwin, et al. "Conversational ai: The science behind the alexa prize." *arXiv preprint arXiv:1801.03604* (2018).

[5] Majlesi, Ali Reza, et al. "Managing turn-taking in human-robot interactions: The case of projections and overlaps, and the anticipation of turn design by human participants." *Social Interaction. Video-based Studies of Human Sociality* 6.1 (2023).

[6] Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. A simplest systematics for the organization of turn-taking for conversation. Language, 50(4):696–735, 1974. doi: 10.2307/412243.

[7] de Ruiter, J. P. Turn-taking. The Oxford Handbook of Experimental Semantics and Pragmatics (Mar 2019), 536–548

[8] Serban, I. V., Lowe, R., 0002, P. H., Charlin, L., and Pineau, J. A survey of available corpora for building data-driven dialogue systems: The journal version. Dialogue and Discourse 9, 1 (2018), 1–49