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

Anonymous ACL submission

Abstract

Turn-taking is a fundamental aspect of human communication, essential for smooth and comprehensible verbal interactions. While recent advances in Large Language Models (LLMs) have shown promise in enhancing Spoken Dialogue Systems (SDS), existing models often falter in natural, unscripted conversations due to their being trained on mostly written language, and focus only on turn-final Transition Relevance Places (TRPs). This paper addresses these limitations by evaluating the ability of state-of-the-art LLMs to predict within-turn TRPs, which are crucial for natural dialogue but challenging to predict. We introduce a new and unique dataset of participant-labeled within-turn TRPs and evaluate the accuracy of TRP prediction by state-of-the art LLMs. Our experiments demonstrate the limitations of LLMs in modeling spoken language dynamics and pave the way for developing more responsive and naturalistic spoken dialogue systems.

1 Introduction

When humans interact verbally, they avoid speaking simultaneously and take turns to speak and listen, a process essential for mutual understanding and smooth communication (**??**). Unlike in formal settings with pre-assigned roles, in everyday conversation participants decide when to speak or listen on a per-turn basis (**?**). This *local management system* hinges on conversationalists' ability to recognize and anticipate so-called *Transition Relevance Places* (TRPs), which are points in the speaker's utterance where a listener could take over the role of speaker. To anticipate and recognize TRPs people use various lexico-syntactic, contextual, and intonational cues (**??**).

The ability to predict TRPs is therefore crucial for artificial conversational agents, as it enables them to take turns and provide verbal feedback signals with socially appropriate timing. Recent advances in Large Language Models (LLMs) have sparked interest in leveraging these models to improve turn-taking in Spoken Dialogue Systems (SDS) (**?**). Approaches like TurnGPT and RC-TurnGPT introduce probabilistic models to predict TRPs using contextual and speaker-identity information (**??**). However, these models struggle with unscripted interactions, often resulting in long silences or poorly timed feedback (**?**).

There are two critical issues with the current approaches. One is the optimistic assumption that LLMs that are trained predominantly on written language can generalize to spoken dialogue (**?**). A second and arguably more fundamental problem is that in dialogue corpora we can only unambiguously identify TRPs that occur with speaker switches, but not when TRPs occur *without* a speaker switch (i.e., within a speaker's turn). This means that we have no "ground truth" data about these "silent" TRPs. Having reliable data about this ground truth would allow us to evaluate and improve our spoken dialogue systems.

In this study, we address both of these issues. First, we collected a new and unique empirical dataset[1] based on human responses that allows us to identify and localize within-turn TRPs in recordings of natural conversation. Second, we used this dataset to evaluate how well current state-of-the art LLMs can predict these TRPs. This ability is important, as it would enable dialogue systems to initiate their turns and provide feedback with correct, human-like timing.

2 Theoretical Background

When humans verbally interact with each other, they avoid speaking at the same time, and take turns speaking and listening. This allows them to respond sequentially to each other's utterances

[1] Our dataset and processing code along with the elicitation protocol will be made publicly available upon publication.

(**??**), and facilitates mutual comprehensibility (**?**). However, in natural conversation, the alternation between speaker and listener roles is not managed by having pre-assigned time slots or a chairperson, like in more formal interactions (e.g., court proceedings or business meetings). Rather, it is a *locally managed* system (**?**). This means that speaker selection is managed by the participants themselves, on a per-turn basis. But how can participants in conversations manage to avoid speaking at the same time, or having long silences in which they are waiting for one another?

This is why conversationalists follow the rules of turn-taking as outlined in **?** (see also **??**). This local management system crucially depends on **?**'s notion of the *Transition Relevance Place* (TRP), which is a position in the current speaker's utterance at which a next speaker can, but is not obliged to, take over the role of speaker. Importantly, even the very short feedback-like turns like "hm-mm", called *backchannels* (**?**) or *continuers* (**?**), are only produced by listeners at TRPs. In the turn-taking literature, an important distinction is made between a *turn*, which is the entire contribution by one speaker, and a *Turn Constructional Unit* (TCU), which is an utterance by the speaker upto a TRP. A turn can consist of multiple TCUs, because at a TRP, the listener is not obliged to take over the floor. This also means that in a corpus of spoken dialogue, it is easy to identify the turn-final TRP, because that is where another speaker takes over the turn, but difficult to identify turn-medial TRPs (where the same speaker continues) because there is no overt recorded behavior that suggests the presence of a TRP.

To function, this local management system crucially relies on the listener's ability to not only recognize, but also *anticipate* the occurrence of a TRP in the current speaker's contribution ahead of time (**?**). To accomplish this timely anticipation, listeners predominantly use lexico-syntactic (**?**), but also contextual and intonational cues (**?**). Listeners process these cues incrementally to predict (or project) TRPs. Importantly, this ability is also required in artificial conversational agents, to enable them to a) take over the floor at the appropriate moment, and b) to provide supportive verbal feedback to the listener with normatively correct timing.

3 Related work

Inspired by research on human turn-taking and advances in language processing through the successful application of predictive language models, there has been a recent focus on improving smooth turn-taking in Spoken Dialogue Systems (SDS) (**?**). Specifically, these recent approaches have sought to leverage linguistic knowledge learned by Large Language Models (LLMs). For example, TurnGPT introduces a probabilistic notion of TRPs that is operationalized by introducing turn-shift tokens that can be predicted using contextual as well as speaker-identity information (**?**). RC-TurnGPT is an extension of this model that conditions the probability of turn-shift tokens based on potential upcoming interlocutor responses, thereby taking their intention into account (**?**). However, these methods so far do not generalize well to corpora of unscripted dialogue. As a consequence, current dialogue systems systems still tend to produce long silences and produce ill-timed feedback signals (**?**).

4 Our approach

At present, we know of only two ways to identify TRPs in recorded conversations. One is to look at speaker changes in dialogue corpora (e.g., the Map Task, or Switchboard corpus (**??**)), and infer that where there is a speaker change, there must have been a TRP. The other is to have experts in the study of conversation annotate TRPs in transcripts of recorded conversations. The problem with the first approach is that, as mentioned above, we can only find TRPs associated with speaker changes. The second approach is not only subjective (although annotators often agree on clear cases of TRPs, they can disagree in more complex scenarios (**?**)), it also does not correspond with the task that dialogue participants (or for that matter, dialogue systems) face in their daily lives. In real life, people engaging in dialogue have to predict upcoming TRPs instinctively, "on the fly", with very little time to contemplate. This is not the same process as identifying them after the fact in transcripts of other people's dialogue.

4.1 Collection of data on human-detected TRPs in natural turns

4.1.1 Corpus of natural conversation

In order to get a reliable dataset of participants' instinctive, on the fly responses to TRPs in natural dialogue, we needed recorded turns from nat-

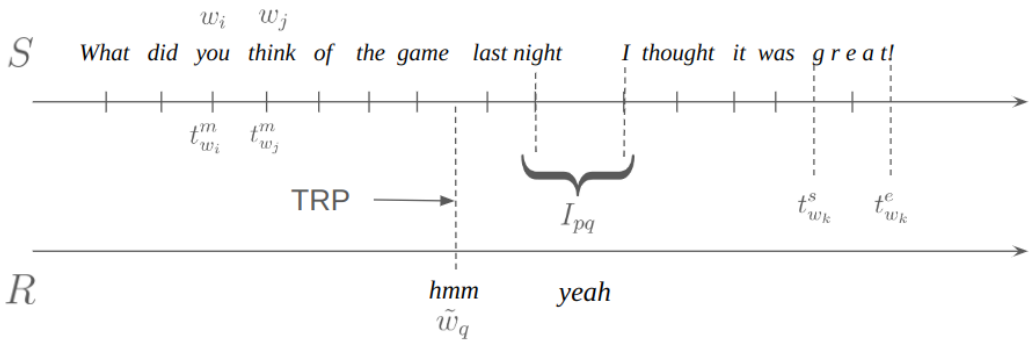

Figure 1: Participants listened to a stimulus (S) and produced auditory responses (R) to indicate their perception of TRPs. Each word in the stimulus ($w_1, t^s_{w_1}, t^e_{w_1}$) and the response ($\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}$) has a start and end time. We use the temporal midpoint($t^m_{w_i} = (t^s_{w_i} + t^e_{w_i})/2$) to discretize locations of both stimulus words and responses.

ural conversations with high audio quality and no cross-talk. Therefore, we collected a corpus, which we called the *In Conversation Corpus* (ICC), containing high quality recordings of natural informal dialogues in American English. In total, the ICC contains recordings of 93 conversations. Each conversation in the ICC is approximately 25 minutes long and features a pair of undergraduate students who engage in free and unscripted conversations. Participants were seated in two soundproofed rooms separated by a anti-reflective glass window, and communicated using a lapel microphone and headphones. The participants' speech was recorded on separate audio tracks, to avoid cross-talk (i.e., hearing the speech of one participant on the recording of the other). These conversations were then first transcribed using the transcription standards from Conversation Analysis using the automated transcription program GailBot (**?**) and subsequently checked and corrected by human annotators[2]. From the 93 total conversations in the ICC, we selected 17 conversations (approximately 425 minutes of talk) to be used in the empirical data collection reported below.

4.1.2 Empirical collection of estimated TRP locations.

To obtain participants' instinctive localization of the within-turn TRPs in spoken utterances, we selected 55 turns (28.33 minutes of talk) that each contained at least two TCUs from our collected corpus. Then we had 118 native speakers of English listen to these utterances, and asked them to verbalize a brief "backchannel" (like *hm-mm*, or *yes*) at every point in time that they thought this would be appropriate.[3] We recorded the presented stimulus turns on the right channel of a stereo recording, and the speech of the participant on the left channel. We used two different lists of stimulus turns, and also two lists that were the reverse of the original two lists, to counterbalance for possible order effects. Participants were randomly assigned to one of the four lists. Subsequently, we used phonetic analysis program Praat (**?**) as well as ELAN (**?**) to locate the onset of the different backchannels produced by the participants. Because we had an average of 59 participants respond to each stimulus, and each participant could respond multiple times (since there could be multiple TRPs in each stimulus), we had an average of 159 responses per stimulus turn. This allowed us to estimate both the probability that people would perceive a TRP at a specific location in the stimulus turn, as well as a distribution of the estimated location of that response.

4.2 Within-Turn TRP Prediction Task

Since we will be evaluating the ability of text-based LLMs to recognize TRPs, we will need a principled way to convert the data obtained from the two audio channels (stimulus and response) into text input for the LLM. Here we describe how the stimuli and responses are discretized and a formal definition of the text-based TRP prediction task.

Preprocessing Multi-channel Audio Data

Given audio data recorded from two synchronized channels, one for each of the stimulus and the participant response, we first extract the corresponding sequences of words and their timing information,

[2]The collected corpus is not publicly available due to restrictions in <anonymized> IRB regulations.

[3]This study was approved by the <anonymized> IRB (ID = STUDY00003236). Participants were undergraduates and were compensated as per IRB regulations.

	Stimulus Lists	
	List 1	List 2
List duration	846.33	853.47
# of words	2558	2693
# of participants	60	58
# of stimuli	28	27
Avg. stimulus duration	30.48	31.67
# words per stimulus	91.3	99.7
Avg. # of responses per stimulus	156	162

Table 1: We presented participants with two stimulus lists (and their reversals), with each list containing multiple stimulus turns, and asked them to indicate their instinctive localization of within-turn TRPs through brief auditory backchannels (e.g, *hm-mm,* yes *etc.*). This table shows various statistics for each list. Note that duration is in seconds and # refers to number. See Section 4.1.2 for further details.

including when the word was initiated, and when it was completed, along with the text of the word.[4]

More formally, we can define a single stimulus $S = \langle (w_1, t^s_{w_1}, t^e_{w_1}) \ldots, (w_N, t^s_{w_N}, t^e_{w_N}) \rangle$ of length N as a sequence of words w_i, where $\forall w_i \in S, w_i \in L$, from a fixed vocabulary L. S also includes timing information for the start ($t^s_{w_i}$) and end ($t^e_{w_i}$) for each word w_i. Similarly, we can define participant responses for each stimulus as $R = \langle (\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}) \ldots, (\tilde{w}_M, t^s_{\tilde{w}_M}, t^e_{\tilde{w}_M}) \rangle$. We used ELAN to *manually* annotate each word along with its timing information (to the nearest tenth of a second), for both the stimulus ($w_i, t^s_{w_i}, t^e_{w_i}$) and participant response ($\tilde{w}_i, t^s_{\tilde{w}_i}, t^e_{\tilde{w}_i}$) audios.[5] This allowed us to ensure that we used precise timing for words and did not accidentally consider other types of speech (e.g., in-breaths, out-breaths, laughter etc.) as responses. Further, we compute the temporal midpoint of words as $t^m_{w_i} = (t^s_{w_i} + t^e_{w_i})/2$. We use this mid-point to create intervals, $I_{ij}, 1 \leq i, j \leq N, j = i + 1$, between words. The midpoint, as opposed to the start or end of a word, is an estimate of the point before which a response may reasonably associated with the previous word.

Next, we determine the proportion of participant responses, based on their start time ($t^s_{\tilde{w}_i}$), that fall within each interval $I^{Proportion}_{ij}$. We consider a TRP to have occurred in an interval if the proportion of responses for that interval is greater than some threshold $\tau \in [0, 1]$.i.e., $I^{Proportion}_{ij} > \tau$. A TRP for an interval I_{ij} can be defined as $T_i \in \{0, 1\}$, a binary random variable that represents the occurrence (1) or lack thereof (0) of a TRP after word w_i, i.e., whether participant word $\tilde{w}_q$ was within an interval I_{ij}, where $t^m_{w_i} \leq t^s_{\tilde{w}_q} \leq t^m_{w_j}$. We can then define $\mathcal{T}_{R,S} = \langle T_1 \ldots T_N \rangle$ as a binary indicator of whether or not a TRP occurred in each interval I_{ij} of a stimulus S. Finally, we define $\mathcal{T}^{Participants}_{R,S}$ as the binary indicator of whether participants responded in all intervals of a stimulus S.

Note that the choice of τ is important since it directly affects $\mathcal{T}^{Participants}_{R,S}$. A large τ implies that we require a higher level of participant agreement for an interval to contain a TRP and vice versa. In this study, we used $\tau = 0.3$.

Task Definition

Broadly, the inference task can be defined as identifying between 0 and N TRPs in S. However, humans do not process all within-turn TRPs after hearing the entire stimulus. Instead, we make decisions about the existence of TRPs during the stimulus as we incrementally process it. Specifically, let a prefix $P_i = \langle w_1, \ldots, w_i \rangle$ be a sequence of words such that $\forall w_i \in P_i, w_i \in S$. We can define the set of all prefixes $\mathcal{P}_S$ for an stimulus S that contains all possible prefixes generated from S, where $|\mathcal{P}_S| = N$.

Definition 4.1 (Within-turn TRP Prediction) *Given a stimulus S, and set of all prefixes $\mathcal{P}_S$, determine $\mathcal{T}^{Predicted}_{R,S}$, where each $T_i \in \mathcal{T}^{Predicted}_{R,S}$ occurs after each of the prefixes $P_i \in \mathcal{P}_S$.*

This decomposes into a set of string classification tasks, with the turn incrementally presented with prefixes. Note that we assume each TRP is classified independently, which means we do not account for the possibility that a TRP might be conditioned on prior TRP determinations. There are several other factors that TRP determinations can be conditioned on, and for this paper, we will focus on conditioning TRP determinations on prior words in the turn (i.e., prefixes) only.[6]

[4] Due to space limitations, we o not provide all details regarding the audio segmentation process. We can share this information upon request.

[5] See Figure 1 for an example.

[6] See **?** and **?** for experimental evidence that linguistic content is sufficient for TRP prediction.

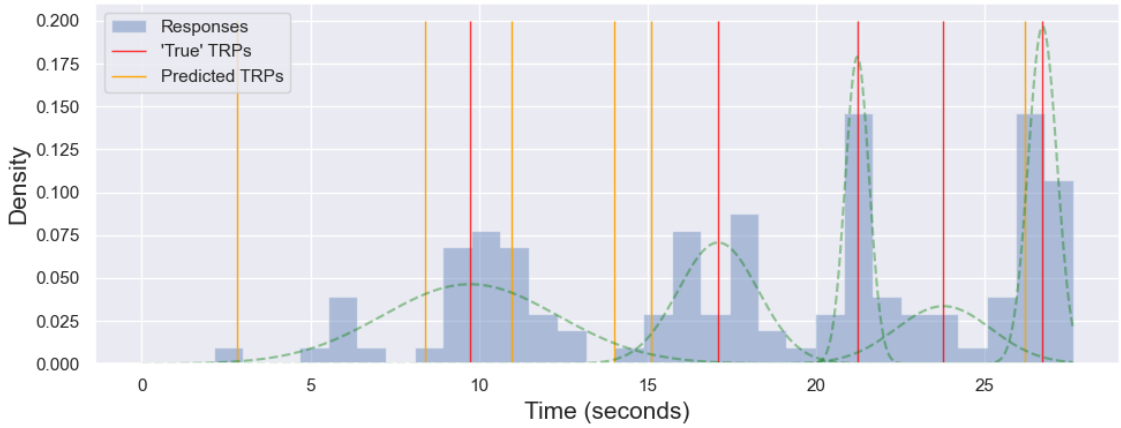

Figure 2: Distribution of participant responses, the times at which participants agreed a TRP occurred, and model predictions of TRPs for a single stimulus S. The dotted lines indicate that each 'true' TRP has some associated variance. The responses are binned between the temporal midpoint of words (see Section 4.2).

5 Evaluation Metrics

Classification Metrics

We can evaluate the performance of a model for the within-turn TRP prediction task (see Section 4.2) by comparing its predictions $\mathcal{T}_{R,S}^{Predicted}$ against the participants' indications of TRPs $\mathcal{T}_{R,S}^{Participants}$. It is important to note the imbalance inherent in the data i.e., intervals that contain TRPs are much lower than those that do not. In this case, we cannot use accuracy since a model that simply predicts the majority class for all intervals will have achieve a high value. Instead, the F1 score i.e., the harmonic mean of precision and recall, is well suited since it emphasizes models that perform well in identifying intervals that contain TRPs ($T_i = 1$), which are the vast minority of intervals.

Free-Marginal Multirater Kappa

Multirater Kappa statistics are often used in medical and behavioral sciences as a measure of agreement over chance between multiple raters (**?**). There are a number of benifits to using Kappa in the context of our work. First, most LLMs, especially smaller ones, lack consistency over multiple predictions generated with the same prompt. Additionally, since most state-of-the-art LLMs do not provide direct access to probability distributions, the kappa statistic can be used to directly compare multiple responses from the same model. In fact, it can also be used to asses agreement between groups of models (**?**). Second, kappa is a measure of *reliability*, but not validity. It might be case the case that groups of LLMs may agree with each other, but not with human participants. Therefore, the kappa statistic offers a way to compare predictions of LLMs to human evaluators (**?**). This is especially important when considering TRPs since the subjectivity of turn-taking decisions may lead to disagreement between raters (LLMs or humans), but might not necessarily indicate an incorrect prediction.

Fleiss' Kappa is typically used when there are multiple raters assessing a nominal variable (**?**). It assumes that the n raters know a priori the number of cases N that must be assigned to each category K. However, this assumption is not valid in our task, which consists of raters (the participants and the models) attempting to assign binary TRP categories across a number of cases (each interval is a case). Here, the rater does not know a priori the number of TRPs that occur in a specific stimulus. When this assumption does not hold, the value of Fleiss' kappa can change significantly based on the distribution of cases in each category, even when all other variables are held constant. **?** proposed a kappa measure (see Equation 1) that resolves this issue and does not make any assumptions about the number of cases in each category (number of TRPs in our case).

$$k_{free} = \frac{[\frac{1}{Nn(n-1)} \sum_{i=1}^{N} \sum_{j=1}^{K} n_{ij}^2 - Nn] - \frac{1}{k}}{1 - \frac{1}{k}} \tag{1}$$

We calculate two variants of the Kappa statistic. k_{free}^{all} simply calculates the kappa statistic across all intervals (or cases) as previously described. However, it is important to consider that we are pri-

marily interested in intervals in which participants' agreed that a TRP occurred, which are the vast minority of intervals. This may cause the kappa value to show an inflated level of agreement. To avoid this issue, we also calculate k^{true}_{free} i.e., the kappa statistic for intervals where there was a TRP. k^{true}_{free} also takes into account the density of participant responses by considering a model prediction to be 'correct' if it is within some window size of the 'true' label (see Figure 3).

Temporal Distance

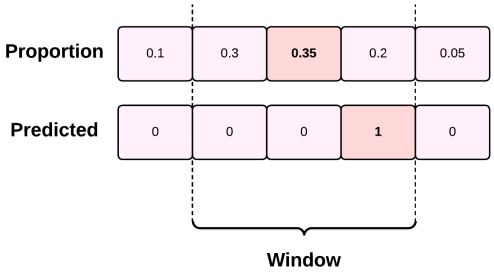

Figure 3: Example of participant response proportions and corresponding model predictions in each interval of a sample stimulus S. In this example, $\tau = 0.3$, which means that there is one one interval in which participants agree that a TRP has occurred. Due to variance in human indications of TRP locations, we may consider a correct prediction to have occured within some window of the 'true' TRP.

Let $d^S_{i,j} \in \{1, \ldots, N\}$ be the minimum absolute distance, in terms of number of intervals, between an interval in which a response was predicted $T^{Predicted}_i = 1$ to the *closest* 'true' TRP $T^{Participants}_j = 1$. Additionally, let $\mathcal{D}_S = \langle d^S_{i,j}, \ldots d^S_{p,q} \rangle$ be a vector of these distances such that $|\mathcal{D}_S| = K$, where K is the number of predicted TRPs.

$$d^S_{i,j} = min(|i-j|) \forall j \in T^{Participants}_j = 1 \quad (2)$$

When discretizing participant responses (see Section 4.1.2), we considered an interval as having a TRP if $I^{Proportion}_{ij} > \tau$. It may be the case that the model predicts a TRP not exactly at an interval where participants' agreed there was a TRP but in some surrounding interval. This is because there is inherent variance in human perceptions of the location of TRPs.

Therefore, we define two simple measures of temporal distance between the predicted and closest 'true' TRP location. the Normalized Mean Absolute Error (NMAE) and the Normalized Mean Square Error (NMSE). The NMAE is a linear measure of distance while the NMSE is a quadratic measure.

$$NMAE = \sum_{i=1}^{|\mathcal{D}_S|} d^S_{i,j} \quad (3)$$

$$NMSE = \sum_{i=1}^{|\mathcal{D}_S|} (d^S_{i,j})^2 \quad (4)$$

However, these simple measures do not take into account the *density* of responses surrounding an interval in which a TRP occurred. For example, if the density of responses around this interval is high, then we may reasonably expect a TRP to have occurred in the surrounding intervals. We use a windowed approach to calculate this density. For each interval in which a 'true' TRP occurred, $I^{Proportion}_{ij} > \tau$, we center a window of size W on that interval. The density of responses is then the proportion of participant responses in the complete window, which we use to compute the *density-adjusted* NMAE ($NMAE_{DA}$).

$$Density_S(I_{ij}, W) = \sum_{l=\frac{-W}{2}}^{\frac{W}{2}} I_{i+lj+l} \quad (5)$$

$$NMAE_{DA} = \sum_{i=1}^{|\mathcal{D}_S|} \frac{d^S_{i,j}}{Density_S(I_{ij}, W)} \quad (6)$$

6 Experiments and Results

We use a number of state-of-the-art LLMs to perform the binary labeling task outlined in Section 4.2. We are particularly interested in LLMs that are pre-trained on diverse datasets and have demonstrated spoken interaction capabilities. We aim to give LLMs a fair chance to falsify our hypothesis that LLMs that are not trained specifically for spoken language tasks will exhibit low performance on the task (**??**).

There are a number of strategies for adapting LLMs for downstream tasks. Fine-tuning is the process of updating the weights of a pre-trained LLM and results in a single specialized model for a specific task. A major benefit of this approach is that training sets of arbitrary size can be used, which have been shown to drastically improve performance. However, most state-of-the-art LLMs do

not allow open-source access to fine-tune models (**?**) and instead allow limited access through public facing APIs. Therefore, we employ *In-context Learning* (ICL) as a task adaptation strategy that does not update model weights. Instead, ICL adapts a model to a downstream task through task demonstrations, which can be performed via prompts. It is important to note that ICL is highly sensitive to prompt choice and that prompts must be optimized through a number of strategies (**?**).

We use GPT-4 Omni, a variant of GPT-4 that is able to process multi-modal information using a single end to end model (**?**), and has achieved higher performance on benchmarks compared to other models (e.g., Gemini, Llama etc.)[7]. We tested the model under two prompting conditions: *expert* and *participant*. In the expert condition, the model was provided theoretical background on TRPs, similar to what an expert annotator might know. In the participant condition, the model was just given an optimized version of the instructions that the human participants received.

Metric	**Average Value**	
	Participant	Expert
Precision	0.126	0.147
Recall	0.153	0.185
F1 Score	0.138	0.164
k^{all}_{free}	0.891	0.860
k^{true}_{free}	0.325	0.201
NMAE	0.286	0.253
NMSE	3.135	5.36
$NMAE_{DA}$	11.28	16.56

Table 2: Measures of performance of GPT-4 Omni on the discrete labeling task (see Section 4.2) in both the participant and expert contexts.

Table 2 shows the results of using GPT-4 Omni to perform the within-turn TRP prediction task averaged across all stimulus lists (see Section 4.1). Overall, the performance of the model reveals significant shortcomings. First, the model exhibits low precision (0.137) and recall (0.169), leading to a low F1 score (0.151), indicating frequent false positives and missed TRPs.[8] While the kappa statistic across all intervals (k^{all}_{free} = 0.876) suggests good general agreement, the much lower kappa for true TRP intervals (k^{true}_{free} = 0.263) highlights difficulties in accurately identifying true TRPs. The NMAE (0.263) and NMSE (4.248) metrics further indicate substantial deviations between intervals where the model predicted TRPs to the closest 'true' TRP. The high density-adjusted NMAE ($NMAE_{DA}$ = 13.92) highlights even greater errors when considering the density of participant responses near intervals in which TRPs occurred.

There are also differences between the participant and expert conditions. The expert condition yielded higher precision (0.147) compared to the participant condition (0.126), indicating more accurate identification of TRPs. The expert condition also achieved higher recall (0.185 vs. 0.153), suggesting a better ability to detect intervals in which TRPs occur. The F1 score, balancing precision and recall, was slightly higher in the expert condition (0.164) than in the participant condition (0.138). Kappa statistics also showed variability: k^{all}_{free} was higher for participants (0.891 vs. 0.860), reflecting stronger overall agreement, while k^{true}_{free} was higher for participants (0.325 vs. 0.201), indicating better performance in correctly identifying true TRPs. Error metrics further demonstrated that the expert condition had lower NMAE (0.253 vs. 0.286) but higher NMSE (5.36 vs. 3.135) and significantly greater density-adjusted NMAE ($NMAE_{DA}$ = 16.56 vs. 11.28). These results suggest that while the expert prompts provided more theoretical accuracy, the participant prompts offered more practical relevance and alignment with true TRPs.

7 Discussion

Half a century of turn-taking research has shown that humans use a number of cues to achieve rapid and seamless turn-transitions in natural conversation by predicting upcoming TRPs. This is vital in minimizing response delays and overlapping speech and is interactionally consequential. Ill-timed turn-taking behavior can have unintended consequences for the interpretation of utterances. For instance, longer delays in responding to an utterance often signal reluctance to produce a dispreferred response (**??**). Current SDS are unable to replicate human-like turn timing (**?**), which decreases user satisfaction and communicative accuracy.

[7]Our initial experiments revealed that the Gemini and Llama models achieved substantially lower performance than GPT-4 Omni in the discrete labeling task. For brevity, we only report the results for the best performing model.

[8]Here, we averaged the score for the participant and expert modes.

LLMs, with their extensive pre-training on every large and diverse datasets, are well-suited to use linguistic (and increasingly multi-modal) information to perform various spoken language tasks (**???**).

However, we find that LLMs underperform across several measures on a simple binary labeling task to predict within-turn TRPs. This holds true even when providing important background context for the task through various prompts (expert versus participant), showcasing this limitation despite employing a widely accepted task adaptation strategy. This limitation points to a major issue: LLMs are as yet not able to effectively utilize their vast linguistic knowledge in the domain of turn-taking. This severely limits their utility in improving the turn-taking timing of SDS.

Our work offers several avenues to further explore and address this issue. First, by empirically demonstrating that current LLMs struggle with TRP prediction despite their extensive pre-training, we expose a critical bottleneck that needs to be addressed. Second, we show that high performance on written-language benchmarks does not necessarily translate to high performance on spoken language tasks. Our creation of a specialized dataset with empirical, on-the-fly human judgements on where TRPs are in natural conversational turns offers a valuable resource for the NLP research community. This dataset allows for targeted fine-tuning and evaluation of LLMs, enabling researchers to develop models that can better mimic human conversational patterns.

8 Conclusion

Even though Large Language Models show impressive performance on a range of challenging language-related tasks, it is as yet unclear whether they can be employed for determining when they can start producing their turn in spoken dialogue at a socially appropriate time. This would require them to have human-level ability to predict Transition Relevance Places, locations in speaker's contribution where they may take over the turn and start speaking. To test this ability in state-of-the-art LLMs, we collected data from humans that perform this task on-the-fly, and compared the performance of the LLMs with that of the human participants. It turned out that the performance of LLMs on this task were far below the level of that of the human participants. Apparently, the pre-training of LLMs on vast amounts of written data was not sufficient to generalize to this particular task. Possible causes for the disappointing performance could be that we haven't found the optimal prompts, and/or that the models would either need more spoken dialogue input during pre-training, or explicit fine-tuning on spoken dialogue data. Either way, the dataset that we have developed will allow researchers in the area of human-machine turn-taking to explore ways to improve the models' performance on this crucial task.

9 Limitations

We acknowledge several limitations in our work. First, our model was provided only with linguistic information, whereas our human participants had access to both prosodic and linguistic cues. While humans can predict TRPs using only lexico-syntactic information (**?**), computational models typically perform better with multi-modal inputs (**??**). Future work could explore if the performance of LLMs can be improved by providing non- and paraverbal cues in addition to the lexico-syntactic information.

Second, although LLMs can match human performance in qualitative coding tasks and provide justifications for their decisions (**?**), their reasoning can differ from human reasoning (**?**). While our incremental binary labeling task allows tracking of LLMs' reasoning for TRP occurrences, we did not analyze the LLM-reported reasoning in the current study. This reasoning may provide cues for providing more effective prompts to the LLM.

Third, we utilized In-Context Learning (ICL) for task adaptation, which has shown high performance in some tasks (**?**), but fine-tuning on conversational data that includes diarization information could potentially improve model performance. This was not feasible due to current restrictions on fine-tuning GPT-4 Omni and similar LLMs (**?**). We aim to explore this as fine-tuning capabilities become available.

Finally, ICL performance is highly sensitive to prompt design (**?**). We may not have optimized our prompts sufficiently, and it remains unclear how best to engineer prompts for within-turn TRP prediction. This, too, can be addressed in further research.

10 Ethical Impact Statement

Value alignment is a key concern shared by researchers and end-users of large language models.

Being able to understand and model the values and normative expectations of not only the contents of speech, but the underlying communicative process itself is important to reduce the risk of misunderstandings, false attributions, and unmet normative expectations. Our work attempts to mitigate these shortcomings and provide the basis for understanding these normative nuances in communicative behavior. Our goal in releasing our corpus and these findings is to facilitate and further research in this domain. We hope to continue exploring the challenges in modeling turn-taking and evaluating the performance of large language models so as to highlight the strengths and weaknesses of using LLMs for spoken dialogue systems to researchers and practitioners.