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

Muhammad Umair and **Vasanth Sarathy** and **JP de Ruiter**
Department of Computer Science
Tufts University
Medford, Massachusetts, USA
{muhammad.umair, vasanth.sarathy, jp.deruiter}@tufts.edu

Abstract

Turn-taking is a fundamental mechanism in human communication that ensures smooth and coherent verbal interactions. Recent advances in Large Language Models (LLMs) have motivated their use in improving the turn-taking capabilities of Spoken Dialogue Systems (SDS), such as their ability to respond at appropriate times. However, existing models often struggle to predict *opportunities* for speaking — called Transition Relevance Places (TRPs) — in natural, unscripted conversations, focusing only on turn-final TRPs and not within-turn TRPs. To address these limitations, we introduce a novel dataset of participant-labeled within-turn TRPs and use it to evaluate the performance of state-of-the-art LLMs in predicting opportunities for speaking. Our experiments reveal the current limitations of LLMs in modeling unscripted spoken interactions, highlighting areas for improvement and paving the way for more naturalistic 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 (Stivers et al., 2009; de Ruiter, 2019). Unlike in formal settings with pre-assigned roles, participants in everyday conversations decide when to speak or listen on a per-turn basis (Sacks et al., 1974). This *local management system* hinges on conversationalists' ability to recognize and anticipate so-called *Transition Relevance Places* (TRPs), which are points in a speaker's utterance that signal appropriate opportunities for the listener to speak. In other words, at a TRP, listeners have the opportunity, but are not obligated, to speak. Importantly, interlocutors anticipate and recognize TRPs using various lexico-syntactic, contextual, and intonational cues (de Ruiter et al., 2006; Bögels and Torreira, 2021).

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 generated interest in improving turn-taking capabilities in Spoken Dialogue Systems (SDS) using these models (Ni et al., 2021). Specifically, approaches like TurnGPT and RC-TurnGPT introduce probabilistic models to predict TRPs using contextual and speaker-identity information (Ekstedt and Skantze, 2020; Jiang et al., 2023a). However, most methods struggle to handle unscripted spoken interactions, often resulting in long silences or poorly timed feedback (Skantze, 2021).

There are two critical issues with the current approaches. First is the optimistic assumption that LLMs trained predominantly on written-first language can learn the complex dynamics of spoken-first language (Mahowald et al., 2024; Umair et al., 2024; Liesenfeld and Dingemanse, 2024), which are distinct in structure and use (e.g. Drieman 1962; Pilan et al. 2024). A second, more fundamental issue is that, while TRPs at speaker switches can be identified unambiguously, it is challenging to clearly identify TRPs *within* turns. This means that we have no 'ground truth' data about these 'silent' TRPs, where a listener could have responded but chose not to.

In this work, we address these issues by first developing a novel and unique empirical dataset[1] based on human responses that allows us to identify within-turn TRPs in natural conversation. Second, we use this dataset to establish the baseline performance of state-of-the-art LLMs to predict within-turn TRPs. This ability is vital for future dialogue systems to appropriately time their turns, use strategically timed silences to convey social

[1]The dataset collected as part of this work is publicly available at: https://osf.io/k5pc9/?view_only=5124d862448f4435b775d49a7b299d6d

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 15503–15514
November 12-16, 2024

cues, and maintain conversational flow.

2 Theoretical Background

When humans verbally interact with each other, they avoid speaking at the same time and instead take turns speaking and listening (Sacks et al., 1974). This allows them to respond sequentially to each other's utterances and facilitates mutual comprehensibility (Duncan, 1972; de Ruiter, 2019). However, the alternation between speaker and listener roles in natural conversation is not predetermined (e.g., allotted time slots in court proceedings). Rather, it is *locally* managed by speakers themselves on a per-turn basis (Stivers et al., 2009; Bögels and Torreira, 2015). 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?

Conversationalists follow rules imposed by a *universal* turn-taking model, proposed by Sacks et al. (1974) (see also Levinson, 1983; de Ruiter, 2019). This system crucially depends on the notion of the *Transition Relevance Place* (TRP), which is an opportunity in the current speaker's utterance at which a listener can, but is not obligated to, take over the role of speaker. Even short feedback-like turns, such as 'hmm', known as *backchannels* (Yngve, 1970) or *continuers* (Schegloff, 1982), are precisely timed to occur at TRPs. Importantly, TRPs are not a function of a speaker's intentions but a consequence of the turn-taking mechanism itself. This distinguishes speech at TRPs from interruptions or barge-ins, which can occur at any point in a speaker's utterance and are noticeable precisely because their timing does not meet normative expectations.

In the turn-taking literature, a crucial distinction is made between a *turn*, which is the entire contribution by one speaker, and a *Turn Construction Unit* (TCU), which ends at a TRP. Since a listener is not required to speak at every TRP, a turn can consist of multiple TCUs, with potentially multiple TRPs occurring within a turn. This implies that, by definition, turn-switches can only occur at a TRP. While it is relatively straightforward to identify turn-final TRPs – where the listener takes over – it is challenging to reliably locate turn-medial TRPs (where the speaker continues) due to the absence of observable cues suggesting the presence of a TRP.

To function, the turn-taking system requires listeners to not only recognize but also *anticipate* the occurrence of a TRP in the current speaker's contribution (Riest et al., 2015). Listeners process various turn-taking cues – primarily lexico-syntactic (de Ruiter et al., 2006), but also contextual and intonational cues (Bögels and Torreira, 2021) – incrementally to predict upcoming TRPs, ensuring that their responses are normatively timed. This anticipatory ability is essential not only for successful human interactions but also for designing artificial conversational agents capable of a) taking over the floor at the right moment and b) providing verbal feedback with correct timing.

Beyond the basic mechanisms, cultural variations also play a role in how turn-taking unfolds. While the fundamental mechanisms of turn-taking, such as the cues for taking or passing on turns (Stivers et al., 2009), are largely universal, cultural norms can shape the timing and style of these transitions (Schegloff, 1982). Therefore, understanding both the culture-invariant and culture-specific components of turn-taking is crucial for developing dialogue systems that are not only responsive but also adaptable across diverse cultural contexts.

3 Related work

Efforts to improve turn-taking in Spoken Dialogue Systems (SDS) have increasingly leveraged the linguistic capabilities of LLMs, driven by the need for these systems to manage natural unscripted interactions, particularly in multi-party settings (Ni et al., 2021). One notable approach is TurnGPT, which introduces a probabilistic model to predict Transition Relevance Places (TRPs) using turn-shift tokens based on both contextual and speaker-identity information (Ekstedt and Skantze, 2020). An extension of this approach is RC-TurnGPT, which incorporates the predicted responses of interlocutors, conditioning predictions on upcoming linguistic content (Jiang et al., 2023a). However, these methods so far do not generalize well to corpora of *unscripted* dialogue. As a consequence, current dialogue systems still tend to produce long silences and ill-timed feedback (Skantze, 2021).

Ablation studies on these LLMs suggest that previous linguistic content is generally sufficient for accurate prediction of turn-ends, and that the relative gain in accuracy diminishes with larger context windows—i.e., TRPs can often be predicted effectively using the local linguistic content of a turn. Additionally, although several approaches attempt to predict turn-ends using acoustic signals (e.g., Ek-

stedt and Skantze 2022a,b; Inoue et al. 2024), our work is grounded in human turn-taking literature, where linguistic cues are recognized as both necessary and sufficient for anticipating TRPs (de Ruiter et al., 2006).

4 Our approach

There are two common methods for identifying TRPs in recorded conversation corpora (e.g, Godfrey and Holliman 1993; Anderson et al. 1991; Kraaij et al. 2005). The first is to locate speaker changes, which are directly observable, and infer that these changes occur only at TRPs. The second is to have experts in conversation analysis manually annotate TRPs. While these methods are widely used, both have limitations. Speaker changes only account for a subset of all TRPs, as there are opportunities where a listener could respond but chooses not to i.e., within-turn TRPs (Threlkeld et al., 2022). Expert annotations, meanwhile, are subjective and do not align with the task faced by participants in real-time dialogue, who must predict TRPs instinctively 'on-the-fly'. In contrast, annotators analyze conversations retroactively, without engaging in the same anticipatory processing as active participants – leading to low ecological validity (see Albert and de Ruiter (2018)).

To address these limitations, we designed an experiment that engaged participants in natural conversations, asking them to produce auditory responses at any location they felt it was possible to respond—not necessarily where they would have responded in everyday interactions. By having multiple participants repeat this task, we gathered a wide range of responses for the same turns. While individual responses varied, the aggregated distribution (with a sufficiently large sample size) provides a reliable indicator of within-turn TRPs, reflecting how humans identify these opportunities in real conversations.

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

Corpus of natural conversations

To create a reliable dataset of participants' instinctive responses to TRPs, we first collected a corpus – named the *In Conversation Corpus*[2] (ICC) – containing high-quality recordings of informal dialogues in American English. The ICC consists of 93 conversations, each lasting approximately 25 minutes, and each featuring pairs of undergraduate students engaged in unscripted conversations. Participants sat in sound-proofed rooms separated by a glass window and communicated using microphones and headphones, ensuring that we could record high-quality audio with complete sound isolation per speaker i.e., no cross-talk. The recordings were first automatically transcribed using GailBot (Umair et al., 2022), following *Jeffersonian* transcription notation, and subsequently verified by human annotators.

From the 93 total conversations in the ICC, we initially selected 17 candidate conversations (approximately 425 minutes of talk) for further analysis. To focus specifically on participants' instinctive localization of within-turn TRPs, we further filtered these conversations to select 55 turns that we suspected contained at least two TCUs. Ultimately, this resulted in 28.33 minutes of talk being used in the data collection reported below.

We selected the ICC over publicly available dialogue corpora to ensure more natural and diverse turn-taking behaviors in our dataset. Although open-source datasets are well-annotated and widely studied, their data collection methods often limit the range and authenticity of conversational behaviors exhibited (Reece et al., 2023).

Empirical collection of estimated TRP locations.

To empirically determine the locations of TRPs, we created two mutually exclusive lists of *stimulus turns* from the filtered subset of the ICC (55 turns; 28.33 minutes of talk), ensuring that each turn was assigned to only one list. To mitigate potential ordering effects, we generated two additional lists in which the stimulus turns (see Figure 1) were presented in reverse order (i.e., the last turn appeared first, and so on). In total, we had four *stimulus lists*, each approximately 15 minutes long.

We recruited 118 native English speakers as participants[3], none of whom were experts in the turn-taking literature. Each participant was randomly assigned to one of the four stimulus lists and asked to verbalize brief backchannels (e.g., 'hmm', 'yes') whenever they felt it was appropriate. Each participant's responses were recorded on individual audio channels, synchronized with the stimulus audio to

[2]The ICC is not currently publicly available in its entirety due to restrictions by the Tufts University IRB.

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

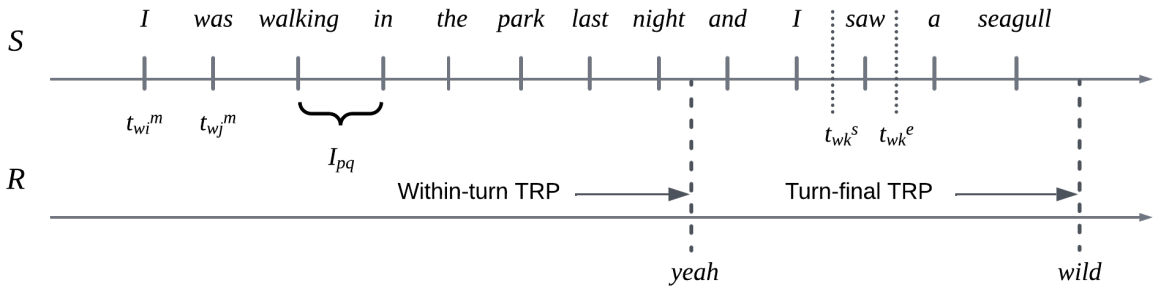

Figure 1: Participants listened to a stimulus (S) and produced auditory responses (R) to indicate their perception of TRPs. Each word in the stimulus ($w_1, t^s_{w_1}, t^e_{w_1}$) and the response ($\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}$) has a start and end time. Intervals are between adjacent words (I_{pq}).

maintain clarity and separation.

We used the phonetic analysis software Praat (Boersma and Van Heuven, 2001) and ELAN (Wittenburg et al., 2006) to *manually* locate the onset of each backchannel response across all participants. 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, outbreaths, laughter etc.) as responses. Since two of the lists were reversals of the originals, we merged the participant responses from these reversed lists with those from the original lists for analysis. On average, 59 participants responded to each stimulus turn, multiple times if they perceived multiple TRPs, resulting in an average of 159 responses per stimulus turn (see Table 1). This allows us to estimate both the likelihood of perceiving a TRP at a specific location and the distribution of those estimated response locations (see Figure 2). Refer to Appendix A for further details on the processing of stimulus lists.

	Stimulus Lists	
Metric	**List 1**	**List 2**
List duration (s)	846.3	853.5
# of words	2558	2693
# of participants	60	58
# of stimuli	28	27
Avg. stimulus duration (s)	30.5	31.7
# words per stimulus	91.3	99.7
Avg. # of responses per stimulus	156	162

Table 1: Participants listened to two stimulus lists and their reversals, each containing multiple turns. They indicated within-turn TRPs using brief auditory backchannels. Responses from the original and reversed lists were merged for analysis. The table summarizes statistics for each list. Note that # refers to number with the duration in seconds.

4.2 Within-Turn TRP Prediction Task

Preprocessing Multi-channel Audio Data

Since we are evaluating the ability of LLMs to recognize TRPs, we require a principled method to formalize our experimentally collected dataset by converting the audio data from two synchronized channels (stimulus and response) into a structured format suitable for analysis. From the recorded audio, we extract words along with their precise timing information, including start and end times. To ensure high accuracy, these timing details were manually annotated to the nearest tenth of a second for both the stimulus and participant responses.

Formally, we 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 each word belongs to a fixed vocabulary L, such that $\forall w_i \in S, w_i \in L$. The stimulus S also includes start ($t^s_{w_i}$) and end ($t^e_{w_i}$) times for each word. Participant responses are similarly defined as $R = \langle (\tilde{w}_1, t^s_{\tilde{w}_1}, t^e_{\tilde{w}_1}), \ldots, (\tilde{w}_M, t^s_{\tilde{w}_M}, t^e_{\tilde{w}_M}) \rangle$. Further, we calculate the temporal midpoint of each word as $t^m_{w_i} = (t^s_{w_i} + t^e_{w_i})/2$, and use these midpoints to create intervals, $I_{ij}, 1 \leq i, j \leq N, j = i+1$, between words. Using the temporal midpoint provides a more reliable estimate for determining whether a response is most reasonably associated with the preceding word.

We also define a binary random variable $T_i \in \{0, 1\}$ for each interval I_{ij} indicating the occurrence (1) or absence (0) of a TRP after word w_i. The vector $\mathcal{T}_{R,S} = \langle T_1, \ldots, T_N \rangle$ subsequently acts as a collection of binary indicators representing whether a TRP occurred in each interval I_{ij} of a

stimulus S.

Finally, $\mathcal{T}_{R,S}^{Participants}$ represents a binary indicator of whether participants agreed that a TRP had occurred in each interval of a stimulus S. We determine participant agreement by calculating the proportion of participant responses $I_{ij}^{Proportion}$, based on their start times ($t_{\tilde{w}_i}^{s}$), that fall within each interval I_{ij}. We consider a TRP to have occurred if the proportion of responses for an interval exceeds a predefined threshold $\tau \in [0, 1]$, i.e., $I_{ij}^{Proportion} > \tau$ (we used $\tau = 0.3$). Note that the choice of τ is crucial, as it directly impacts $\mathcal{T}_{R,S}^{Participants}$: a larger τ requires a higher level of participant agreement for an interval to be marked as containing a TRP, while a smaller τ allows for a more relaxed consensus.

Task Definition

Broadly, the inference task can be defined as identifying between 0 and N TRPs in a stimulus S. However, it is important to consider that humans do not process entire turns as complete units; rather, we incrementally process speech and decide on the existence of TRPs at each point in time. To replicate this incremental processing in the inference task, we define a prefix $P_i = \langle w_1, \ldots, w_i \rangle$ as a sequence of words from the first to the i^{th} word, such that $\forall w_i \in P_i, w_i \in S$. We further define $\mathcal{P}_S$ as the set of all prefixes for a stimulus turn S, with $|\mathcal{P}_S| = N$

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

Definition 4.1 allows us to decompose each stimulus turn S into a set of binary string classification tasks. Notably, we assume that the value of T_i is independent of all prior TRP determinations, $T_1, \ldots, T_{i-1}$. While TRP determinations depend on multiple factors, in this paper, we focus solely on conditioning TRP determinations on the linguistic information provided by preceding words. See de Ruiter et al. (2006) and Riest et al. (2015) for experimental evidence that linguistic content is sufficient for TRP prediction.

5 Evaluation Metrics

Classification Metrics

We can evaluate the performance of a model for the within-turn TRP prediction task (see Definition 4.1) 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 less frequent 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 (Artstein and Poesio, 2008). There are a number of benefits 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 assess agreement between groups of models (Tang et al., 2024). Second, kappa is a measure of *reliability*, but not validity. It might be 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 (Wang et al., 2024). 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 (Fleiss, 1971). 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. Randolph (2005) proposed a kappa mea-

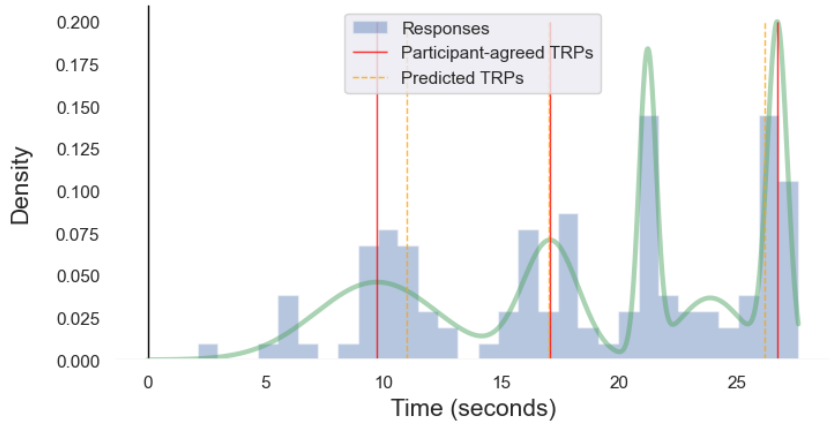

Figure 2: Distribution of participant responses, the times at which participants agreed a TRP occurred, and model predictions of TRPs for a single stimulus S. The dotted lines indicate that each participant-agreed TRP has some associated variance. The responses are binned between the temporal midpoint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

sure (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).

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

We calculate two variants of the Kappa statistic. The first, k_{free}^{all}, computes the kappa statistic across all intervals, as previously described. However, since our primary focus is on intervals where participants agreed that a TRP occurred—representing only a small portion of the total intervals—considering all intervals may result in an inflated kappa value, falsely indicating a high level of agreement. To address this, we also calculate k_{free}^{true}, which specifically evaluates the kappa statistic for intervals where a TRP was present. Moreover, k_{free}^{true} accounts for the density of participant responses by marking a model prediction as "correct" if it falls within a defined window around an interval where participants agreed there was a TRP (see Figure 3).

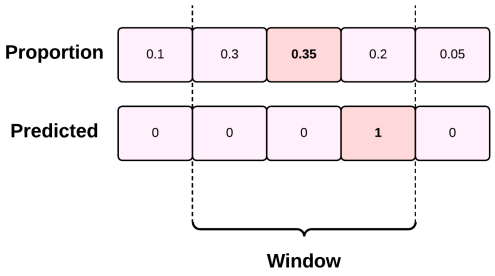

Figure 3: Example of participant response proportions and corresponding model predictions in each interval of a sample stimulus S. In this example, $\tau = 0.3$, which means that there is one one interval in which participants agree that a TRP has occurred. Due to variance in human indications of TRP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

Temporal Distance Metrics

Let $d_{i,j}^{S} \in 1, \ldots, N$ (see Equation 2) represent the minimum absolute distance, in terms of the number of intervals, between an interval where a response was predicted ($T_i^{Predicted} = 1$) and the closest interval in which participants agreed that a TRP occurred ($T_j^{Participants} = 1$). Furthermore, let $\mathcal{D}_S = \langle d_{i,j}^{S}, \ldots, d_{p,q}^{S} \rangle$ be a vector of these distances, with $|\mathcal{D}_S| = K$, where K is the total number of predicted TRPs.

As previously discussed (see Section 4.2), we consider an interval I_{ij} to contain a TRP ($T_j^{Participants} = 1$) if the proportion of participants that responded in that interval exceeds a predefined threshold ($I_{ij}^{Proportion} > \tau$). However, a model's prediction may not align perfectly with the exact interval I_{ij} where participants agreed that a TRP occurred. Instead, the prediction may fall in a nearby interval that still lies within an acceptable range, given the inherent variance in participant responses regarding the precise location of TRPs (e.g., Templeton et al. (2022)).

$$d_{i,j}^{S} = min(|i - j|) \forall j \in T_j^{Participants} = 1 \quad (2)$$

Therefore, we define two measures of temporal distance between the predicted TRP and the closest participant-agreed TRP location: the Normalized Mean Absolute Error (NMAE) and the Normalized Mean Square Error (NMSE). The NMAE provides a linear measure of distance, whereas the NMSE offers a quadratic measure.

$$NMAE = \sum_{i=1}^{|\mathcal{D}_S|} d_{i,j}^S$$

$$NMSE = \sum_{i=1}^{|\mathcal{D}_S|} (d_{i,j}^S)^2$$

However, these simple measures do not account for the *density* of responses around an interval where a TRP occurred. For instance, if the density of responses near this interval is high, it may be reasonable to expect that a TRP could also occur in the neighboring intervals. To incorporate this, we employ a windowed approach to calculate response density. For each interval where a participants agreed a TRP occurred ($I_{ij}^{Proportion} > \tau$), we center a window of size W on that interval. The density of responses is then defined as the proportion of participant responses within the entire window, which we use to compute the *density-adjusted* measure $NMAE_{DA}$.

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

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

6 Experiments and Results

We employed several state-of-the-art LLMs to perform the within-turn TRP prediction task described in Section 4.2. Our focus was on models pretrained on diverse datasets, some of which were explicitly designed with capabilities for spoken interaction.

To adapt LLMs for downstream tasks, two main strategies are commonly used: fine-tuning and in-context learning (ICL). Fine-tuning involves updating the weights of a pre-trained model to specialize it for a particular task, resulting in a single model tailored for that task. This approach is advantageous because it can accommodate training sets of any size, often leading to significant performance improvements. However, most state-of-the-art LLMs are not available for direct fine-tuning due to restricted open-source access and instead are accessible only through public APIs (Liesenfeld and Dingemanse, 2024).

Given these limitations, we employed in-context learning (ICL) as our adaptation strategy. Unlike fine-tuning, ICL does not require modifying model weights. Instead, it adapts the model to a specific task by using task demonstrations provided through prompts. However, it is crucial to note that ICL is highly sensitive to the formulation of these prompts, and optimizing prompts requires careful consideration and specific strategies (Chang et al., 2024).

We tested each model under two prompting conditions: expert and participant. In the expert condition, the model was provided with theoretical background on TRPs, similar to what an expert annotator might know. In the participant condition, the model was given a version of the instructions that the human participants received. These two prompting conditions explore how the level of provided information affects the model's performance on the TRP prediction task.

Table 2 shows the performance of multiple language models, including GPT-4 Omni, Phi3, Gemma2, Llama3.1, and Mistral, on the within-turn TRP prediction task averaged across all stimulus lists (see Section 4.1). We focus on GPT-4 Omni because it is the best overall performer, setting the benchmark for this challenging task, despite its significant shortcomings. While other models, like Mistral:7b and Phi3:14b, show strengths in specific metrics—such as lower NMAE (0.190) and favorable NMSE (5.091) in expert conditions—these are limited to isolated scenarios, and overall performance across metrics like precision, recall, and F1 score remains inferior to GPT-4 Omni.

Overall, the performance of the best performing 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. While the kappa statistic across all intervals (k_{free}^{all} = 0.876) suggests good general agreement, the much lower kappa for participant-agreed TRP intervals (k_{free}^{true} = 0.263) highlights difficulties in accurately identifying participant-agreed TRPs. The NMAE (0.263) and NMSE (4.248) metrics further indicate substantial deviations between intervals

Model	Condition	Precision	Recall	F1 Score	k^{all}_{free}	k^{true}_{free}	*NMAE*	*NMSE*	$NMAE_{DA}$
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.

where the model predicted TRPs to the closest participant-agreed 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 participant-agreed 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 research on turn-taking has demonstrated that humans rely on various cues to achieve rapid and seamless turn-transitions in natural conversation by accurately predicting upcoming TRPs. This ability is crucial for minimizing response delays and avoiding overlapping speech, both of which are interactionally significant (Sacks et al., 1974; de Ruiter et al., 2006; Levinson and Torreira, 2015). Poorly timed turn-taking can negatively affect how utterances are interpreted; for example, longer response delays often signal reluctance or hesitation to deliver a dispreferred response (de Ruiter, 2019; Kendrick and Torreira, 2015). Current spoken dialogue systems (SDS), however, struggle to replicate human-like turn timing, resulting in reduced user satisfaction and diminished communicative effectiveness (Skantze, 2021).

State-of-the-art LLMs, pre-trained on large and diverse datasets, are well-suited for leveraging linguistic information—which has been shown to be sufficient for predicting opportunities for speech in humans—and increasingly, multimodal information, to perform a range of spoken language tasks (Ekstedt and Skantze, 2020; Jiang et al., 2023a,b). However, contrary to expectations, we find that the LLMs we tested underperform across multiple measures on a simple binary prediction task to identify within-turn TRPs when using In-Context Learning (ICL) as the adaptation strategy. This holds true even when providing essential background context through various prompts (expert versus participant). These findings point to a major issue: LLMs are currently unable to effectively utilize their extensive linguistic knowledge for unscripted

turn-taking in spoken interaction. This limits their application in dialogue systems by preventing these systems from accurately anticipating opportunities for speaker transitions.

Our work attempts to advance the performance of LLMs for turn-taking in spoken interaction. 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 provide evidence that high performance on written-language benchmarks does not necessarily translate to high performance on spoken language tasks, emphasizing the need for specialized evaluation in conversational settings. Third, we contribute a specialized dataset containing empirical, on-the-fly human judgments on where TRPs occur in natural conversation. This dataset is a valuable resource for the NLP research community, offering opportunities for targeted fine-tuning and evaluation of LLMs, and enabling the development of models that more closely replicate human conversational behavior.

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 selected LLMs on this task was 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, the models we used had access only to linguistic information, i.e., the words of a stimulus, whereas human participants had access to both prosodic and linguistic cues. Although humans can predict TRPs using only lexico-syntactic information (de Ruiter et al., 2006), computational models often perform better with multi-modal inputs (Roddy, 2021; Kurata et al., 2023). Despite this, our focus on text-only models is grounded in turn-taking literature, which establishes linguistic cues as necessary and sufficient for humans to anticipate TRPs (de Ruiter et al., 2006). Evaluating LLMs with only linguistic input was an essential step to determine whether they could replicate this human ability, particularly in spoken language contexts. Existing text-based models, such as TurnGPT and RC-TurnGPT, have struggled to generalize across diverse conversations, further emphasizing the need to isolate linguistic factors in our study. Future work should explore whether adding acoustic information can improve LLM performance on the TRP prediction task.

Second, we evaluated LLM predictions solely against participant responses from the ICC, a dataset specifically designed to capture naturalistic conversational behaviors. While we chose the ICC to avoid the limitations inherent in other corpora, it is essential to replicate our findings with commonly used dialogue datasets (e.g., Switchboard, ICSI, AMI, and SpokenWoz) to verify the broader applicability of our approach. This replication, however, is resource-intensive, as most of these corpora predominantly contain annotations for between-turn TRPs and lack detailed within-turn TRP annotations.

Third, we used In-Context Learning (ICL) as a task adaptation strategy for TRP prediction because fine-tuning was not feasible, given the restrictions on modifying the LLMs used in this work (Liesenfeld and Dingemanse, 2024). Although ICL has shown promise on certain tasks (Chang et al., 2024), its performance is highly sensitive to prompt design (Wei et al., 2023). It is possible that we may not have fully optimized our prompts, and it remains unclear how best to engineer prompts for the within-turn TRP prediction task. Future research should explore not only the potential benefits of

fine-tuning but also improved prompt engineering strategies to enhance model performance.

Finally, although LLMs can match human performance in qualitative coding tasks and provide justifications for their decisions (Dunivin, 2024), their reasoning often diverges from human reasoning (Bao et al., 2024). While our incremental binary labeling task allows us to track the LLMs' reasoning for TRP occurrences, we did not analyze the reported reasoning in this study. Future research should focus on analyzing these reasoning patterns, as they could offer valuable insights for designing more effective prompts to improve LLM performance.

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.

11 Acknowledgements

This research was supported in part by Other Transaction award HR00112490378 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program.

We would also like to acknowledge Grace Hustace for her contributions to data collection and processing, and Dr. Julia Mertens for her input during early-stage discussions, both of whom are affiliated with the Human Interaction Lab at Tufts University.

References

Saul Albert and Jan-Peter de Ruiter. 2018. Improving human interaction research through ecological grounding. *Collabra: Psychology*, 4(1):24.

Anne H Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, et al. 1991. The hcrc map task corpus. *Language and speech*, 34(4):351–366.

Ron Artstein and Massimo Poesio. 2008. Inter-coder agreement for computational linguistics. *Computational linguistics*, 34(4):555–596.

Guangsheng Bao, Hongbo Zhang, Linyi Yang, Cunxiang Wang, and Yue Zhang. 2024. Llms with chain-of-thought are non-causal reasoners. *arXiv preprint arXiv:2402.16048*.

Paul Boersma and Vincent Van Heuven. 2001. Speak and unspeak with praat. *Glot International*, 5(9/10):341–347.

Sara Bögels and Francisco Torreira. 2021. Turn-end estimation in conversational turn-taking: the roles of context and prosody. *Discourse processes*, 58(10):903–924.

Sara Bögels and Francisco Torreira. 2015. Listeners use intonational phrase boundaries to project turn ends in spoken interaction. *Journal of Phonetics*, 52:46–57.

Kaiyan Chang, Songcheng Xu, Chenglong Wang, Yingfeng Luo, Tong Xiao, and Jingbo Zhu. 2024. Efficient prompting methods for large language models: A survey. *arXiv preprint arXiv:2404.01077*.

Jan P. de Ruiter. 2019. Turn-taking. *The Oxford Handbook of Experimental Semantics and Pragmatics*, page 536–548.

Jan P. de Ruiter, H. Mitterer, and N. J. Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Gerard HJ Drieman. 1962. Differences between written and spoken language: An exploratory study. *Acta Psychologica*, 20:36–57.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2):283–292.

Zackary Okun Dunivin. 2024. Scalable qualitative coding with llms: Chain-of-thought reasoning matches human performance in some hermeneutic tasks. *arXiv preprint arXiv:2401.15170*.

Erik Ekstedt and Gabriel Skantze. 2020. TurnGPT: a transformer-based language model for predicting turn-taking in spoken dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2981–2990, Online. Association for Computational Linguistics.

Erik Ekstedt and Gabriel Skantze. 2022a. Voice Activity Projection: Self-supervised Learning of Turn-taking Events. In *Proc. Interspeech 2022*, pages 5190–5194.

Erik Ekstedt and Gabriel Skantze. 2022b. Voice activity projection: Self-supervised learning of turn-taking events. In *Interspeech 2022*, pages 5190–5194.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

John Godfrey and Edward Holliman. 1993. Switchboard-1 release 2 LDC97S62. *Linguistic Data Consortium*.

Koji Inoue, Bing'er Jiang, Erik Ekstedt, Tatsuya Kawahara, and Gabriel Skantze. 2024. Multilingual turn-taking prediction using voice activity projection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11873–11883, Torino, Italia. ELRA and ICCL.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023a. Response-conditioned turn-taking prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12241–12248, Toronto, Canada. Association for Computational Linguistics.

Bing'er Jiang, Erik Ekstedt, and Gabriel Skantze. 2023b. What makes a good pause? investigating the turn-holding effects of fillers. *arXiv preprint arXiv:2305.02101*.

Kobin H Kendrick and Francisco Torreira. 2015. The timing and construction of preference: A quantitative study. *Discourse Processes*, 52(4):255–289.

Wessel Kraaij, Thomas Hain, Mike Lincoln, and Wilfried Post. 2005. The ami meeting corpus. In *Proc. International Conference on Methods and Techniques in Behavioral Research*, pages 1–4.

Fuma Kurata, Mao Saeki, Shinya Fujie, and Yoichi Matsuyama. 2023. Multimodal turn-taking model using visual cues for end-of-utterance prediction in spoken dialogue systems. *Proc. Interspeech 2023*, pages 2658–2662.

Stephen C Levinson. 1983. Pragmatics. *Cambridge UP*.

Stephen C Levinson and Francisco Torreira. 2015. Timing in turn-taking and its implications for processing models of language. *Frontiers in psychology*, 6:136034.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative ai: open-washing and the eu ai act. In *Seventh Annual ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT 2024)*. ACM.

Kyle Mahowald, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*.

Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, Vinay Vishnumurthy Adiga, and E. Cambria. 2021. Recent advances in deep learning based dialogue systems: a systematic survey. *Artificial Intelligence Review*, 56:3055–3155.

Ildiko Pilan, Laurent Prévot, Hendrik Buschmeier, and Pierre Lison. 2024. Conversational feedback in scripted versus spontaneous dialogues: A comparative analysis. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 440–457, Kyoto, Japan. Association for Computational Linguistics.

Justus J Randolph. 2005. Free-marginal multirater kappa (multirater k [free]): An alternative to fleiss' fixed-marginal multirater kappa. *Online submission*.

Andrew Reece, Gus Cooney, Peter Bull, Christine Chung, Bryn Dawson, Casey Fitzpatrick, Tamara Glazer, Dean Knox, Alex Liebscher, and Sebastian Marin. 2023. The candor corpus: Insights from a large multimodal dataset of naturalistic conversation. *Science Advances*, 9(13):eadf3197.

Carina Riest, Annett B Jorschick, and Jan P de Ruiter. 2015. Anticipation in turn-taking: mechanisms and information sources. *Frontiers in psychology*, 6:89.

Matthew Roddy. 2021. *Neural Turn-Taking Models for Spoken Dialogue Systems*. Ph.D. thesis, Ph. D. thesis, Trinity College Dublin, Dublin.

Harvey Sacks, Emanuel A. Schegloff, and Gail Jefferson. 1974. A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Emanuel A Schegloff. 1982. Discourse as an interactional achievement: Some uses of 'uh huh'and other things that come between sentences. *Analyzing discourse: Text and talk*, 71:71–93.

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

Tanya Stivers, Nicholas J Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federico Rossano, Jan Peter De Ruiter, Kyung-Eun Yoon, et al. 2009. Universals and cultural variation in turn-taking in conversation. *Proceedings of the National Academy of Sciences*, 106(26):10587–10592.

Liyan Tang, Igor Shalyminov, Amy Wong, Jon Burnsky, Jake Vincent, Yu'an Yang, Siffi Singh, Song Feng, Hwanjun Song, Hang Su, Lijia Sun, Yi Zhang, Saab Mansour, and Kathleen McKeown. 2024. TofuEval: Evaluating hallucinations of LLMs on topic-focused

dialogue summarization. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4455–4480, Mexico City, Mexico. Association for Computational Linguistics.

Emma M Templeton, Luke J Chang, Elizabeth A Reynolds, Marie D Cone LeBeaumont, and Thalia Wheatley. 2022. Fast response times signal social connection in conversation. *Proceedings of the National Academy of Sciences*, 119(4):e2116915119.

Charles Threlkeld, Muhammad Umair, and Jp de Ruiter. 2022. Using transition duration to improve turn-taking in conversational agents. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 193–203, Edinburgh, UK. Association for Computational Linguistics.

Muhammad Umair, Julia Beret Mertens, Saul Albert, and Jan Peter de Ruiter. 2022. Gailbot: An automatic transcription system for conversation analysis. *Dialogue & Discourse*, 13(1):63–95.

Muhammad Umair, Julia Beret Mertens, Lena Warnke, and Jan P. de Ruiter. 2024. Can language models trained on written monologue learn to predict spoken dialogue? In review.

Li Wang, Xi Chen, XiangWen Deng, Hao Wen, MingKe You, WeiZhi Liu, Qi Li, and Jian Li. 2024. Prompt engineering in consistency and reliability with the evidence-based guideline for llms. *npj Digital Medicine*, 7(1):41.

Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. 2023. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*.

Peter Wittenburg, Hennie Brugman, Albert Russel, Alex Klassmann, and Han Sloetjes. 2006. ELAN: a professional framework for multimodality research. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy. European Language Resources Association (ELRA).

Victor H Yngve. 1970. On getting a word in edgewise. In *Papers from the sixth regional meeting Chicago Linguistic Society, April 16-18, 1970, Chicago Linguistic Society, Chicago*, pages 567–578.

A Stimulus Preparation Procedure

To prepare the stimulus lists used in this study, we first randomly sampled approximately 60% of the conversations from the ICC, which comprises 93 conversations in total. We selected 17 conversations (425 minutes of talk) that met specific criteria for eliciting participant responses. These criteria included minimal background noise, no recording artifacts, and no cross-talk, ensuring that these factors would not influence participant responses.

From each selected conversation, we isolated a single audio channel representing the speech of one speaker and used ELAN to manually mark the start and end of selected turns. Turns were chosen if they were judged, by an expert annotator, to contain at least two TCUs, thus ensuring the presence of at least one within-turn TRP. The segmentation process involved two stages: an initial pass to mark provisional turn boundaries, followed by a second pass to verify and refine these boundaries. Ambiguous segments, such as those with extended pauses or unclear speaker intent, were excluded to ensure the quality of the stimuli. Ultimately, we selected 55 turns, totaling 28.33 minutes of speech.

Finally, the segmented turns were organized into stimulus lists, with each turn separated by an audible beep to indicate the start of a new turn. To minimize potential ordering effects, the turns within each list were arranged in random order. In total, four stimulus lists were generated: two with the original turn sequences and two with the sequences in reverse order. Each of the four lists was approximately 15 minutes long.