Analysing the role of lexical and temporal information in turn-taking through predictability

Sean Leishman **Sarenne Wallbridge** **Peter Bell**
Institute for Language, Cognition and Computation
University of Edinburgh, UK
peter.bell@ed.ac.uk

Abstract

Turn-taking is a fundamental component of human communication and is signalled through complex cues distributed across lexical, temporal, and prosodic information. Full-duplex models of spoken dialogue integrate these information sources to produce impressive turn-taking behaviour. Yet, existing evaluations of their turn-taking capabilities do not address which information sources drive predictions. We present a systematic analysis of the role of lexical-temporal features on the predictability of turn structure by examining Pairwise-TurnGPT, a full-duplex model of spoken dialogue transcripts. Through PCA, mixed-effects modelling, and temporal surprisal analysis, we reveal context-dependent patterns: linguistic fluency paradoxically creates overconfidence at intermediate completion points, while turn-shift overlap dominates boundary detection. Our findings uncover where lexical-temporal information suffices and where additional cues become necessary, establishing a deeper understanding of how turn-taking cues are distributed and how to evaluate dialogue systems.

1 Introduction

Turn-taking—the systematic organization of who speaks when in conversation—is a fundamental component of human spoken communication. However, it is characterized by complexity; in spontaneous conversation, speakers overlap, interrupt, and backchannel while navigating disfluencies. Still, conversational partners coordinate with remarkable precision, exchanging turns with gaps of -100 to 500ms (Sacks et al., 1974). This intricate coordination relies on cues distributed across the lexical, temporal, and prosodic (the intonation, rhythm, and intensity of speech) channels, where their relative importance is context-dependent (Ford and Thompson, 1996; De Ruiter et al., 2006; Bögels and Torreira, 2015).

Recent full-duplex spoken dialogue models such as dGSLM (Nguyen et al., 2023), Moshi (Défossez et al., 2024), and SALM-Duplex (Hu et al., 2025) integrate information across these channels to generate artificial dialogue. Subjective human evaluations of interactive sessions suggest these systems can produce fluent conversational output (Veluri et al., 2024). However, fluent output does not guarantee realistic turn-taking behaviour.

Evidently, the degree to which these models can reflect realistic human turn-taking behaviour remains poorly understood. Many evaluations focus on human-AI interactions rather than human-human conversation and are primarily based on comparing corpus-level statistics of generated and real dialogue, which ignores variation in how models handle particular conversational phenomena; for example, do models distinguish between supportive backchannels and competitive turn-shifts? Importantly, such evaluations provide little insight into *which* information sources drive predictions about turn structure.

Beyond system evaluation, the predictability of turn-taking behaviour poses a broader psycholinguistic question: how much of turn structure is recoverable from different channels of information? If lexical and temporal cues already encode much of the structure of turn transitions, then prosody may primarily be important in ambiguous contexts. Conversely, systematic prediction failures can help localise where additional cues are necessary. Establishing the contribution of lexical-temporal features to the predictability of turn structure is therefore informative both for dialogue system development and for understanding how information is distributed in spoken conversation. In this work, we address this gap by asking : *"Under what conversational conditions are lexical-temporal features sufficient to predict turn structure, and where do they fail?"*.

We present a systematic analysis of turn-taking

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 1: Long Papers, pages 5998–6009

predictability using only lexical and temporal information by examining predictions from our previously proposed turn prediction model, PairwiseTurnGPT (Leishman et al., 2024). This system models the lexical content of spoken conversation, aligned across speakers using word-level timings. PairwiseTurnGPT is, to our knowledge, the only model that encodes temporally-aligned lexical content of each speaker in a dialogue, without collapsing them into a single serialised stream. As such, it is uniquely suited to isolating the contribution of lexical-temporal cues to turn-taking predictability.

Using PairwiseTurnGPT, we isolate the contribution of lexical-temporal features through three complementary analyses: we characterize the interdependencies of the lexical and structural features of conversational turns through PCA; we then apply mixed-effects modelling to quantify the effects of such features on boundary detection and within-turn uncertainty; finally we conduct novel temporal surprisal analysis to examine how predictions evolve throughout turns.

Our findings reveal patterns invisible in corpus aggregates. Linguistic fluency is paradoxically associated with more premature predictions for turn-endings, though additional context in longer turns mitigates this uncertainty. Turn-shift overlap is a dominant cue for boundary detection and can be differentiated from turn-internal phenomena such as backchannels and complete overlaps. However, lexical and temporal features cannot differentiate supportive from competitive overlapping speech within turns. Our results offer a computational characterisation of how much turn-taking structure is encoded in the lexical-temporal content of dialogue and demonstrate the value of context-sensitive analysis to reveal not just whether models work, but also when and how.

2 Background

2.1 Modelling spoken dialogue.

Traditional computational models of dialogue treated conversations as series of discrete, non-overlapping turns which differs fundamentally from the complex turn structure of spoken dialogue discussed in Section 2.3 (Ekstedt and Skantze, 2020). Full-duplex spoken dialogue models remove the need to serialize turns by simultaneously processing input and output (i.e., "listening" and "speaking"). Models such as dGSLM (Nguyen et al., 2023), Moshi (Défossez et al., 2024) and SALM-Duplex (Hu et al., 2025) achieve this by applying dual-channel architectures to underlying spoken language models (SLMs). Importantly, they differ in their approaches to integrating information sources—dGSLM operates on discrete speech units, Moshi aligns text with audio, and SALM-Duplex interleaves text and speech tokens.

2.2 Evaluating turn-taking.

Current evaluations of full-duplex systems fall into two main categories, neither of which are sufficient for diagnosing whether a model integrates the complex turn-taking cues in human conversation.

Corpus-level metrics compare aggregate frequencies and cumulative durations of turn-taking events including interpausal units (IPUs), pauses, gaps, and overlaps between generated and human dialogue Nguyen et al. (2023); Défossez et al. (2024). However, these aggregate indicators mask important contextual distinctions; for example, high overlap frequency doesn't reveal whether overlaps function as backchannels or interruptions, and pause statistics cannot distinguish appropriate gaps from awkward silences.

LLM judge-based or event-classification approaches provide more nuanced evaluation. Frameworks like Talking Turns, and Full-Duplex-Bench propose using LLMs as judges to classify and evaluate the appropriateness of turn events Lin et al. (2025). FD-bench simulates conversations to assess the robustness of full-duplex systems to user-interruptions and noisy settings, using a judge LLM along with a host of metrics to score the timing and content of responses (Peng et al., 2025). While these approaches are sensitive to context, most focus on human-AI interaction where turn-taking differs from spontaneous human-human conversation (Gonzales, 1994; Levinson, 2016). Moreover, automatic judges are prone to specific failures and cannot identify the source of their judgements, leaving open the question of how full-duplex systems integrate information sources. Recent work already suggests that models exhibit differential capabilities across information modalities: SLMs have been shown to have limited awareness of temporal features of their generation (Chang et al., 2025), while Umair et al. (2024) demonstrate that text-only models struggle to distinguish within-turn transition relevance places from actual turn boundaries.

2.3 Linguistic theories of turn taking.

Linguistic theories characterize turn-taking as a predictive process where interlocutors anticipate turn boundaries from multimodal cues (Sacks et al., 1974; Duncan, 1972). Cues in spoken dialogue are distributed along lexical, temporal, and prosodic features where their relative importance is highly context-dependent. For example, Magyari and De Ruiter (2012) show that listeners use the content of projected utterance completions to accurately predict turn-completion points. Prosodic cues also play a crucial role (Gravano and Hirschberg, 2011; Bögels and Torreira, 2015). While Roddy et al. (2018); Ekstedt and Skantze (2022) show that turn-end prediction can be learned from prosodic features, both in isolation and in conjunction with lexical content, De Ruiter et al. (2006) find that end-of-turn prediction was only impaired by the removal of lexicosyntactic information but not intonational contours. Bögels and Torreira (2015) argue that while lexical-prosodic cues often converge, neither modality is universally sufficient.

Predictive complexity is particularly acute in spontaneous spoken dialogue which exhibits diverse turn-taking phenomena with distinct communicative functions. Backchannels—short utterances like "mm-hmm"—provide feedback without claiming the floor (Ward, 2004), requiring listeners to distinguish supportive overlap from competitive floor attempts. Complete overlaps, where one speaker's utterance begins and ends within another's turn, must be differentiated from turn-shift overlaps that signal genuine turn completion. Cooperative overlaps may complete the speaker's utterance, while competitive overlaps challenge for the floor. These phenomena are achieved through complex, context-dependent combinations of lexical, temporal, and prosodic cues (Schegloff, 2000), motivating our approach of isolating lexical and temporal features from prosodic realisation.

3 PairwiseTurnGPT

To isolate the contribution of lexical information, we examine predictability under PairwiseTurnGPT. Tokens are paired across streams based on word-level timing information to effectively model the complex interaction between speakers by incorporating overlapping speech.

Model architecture. Mirroring the full-duplex architecture in Nguyen et al. (2023), PairwiseTurnGPT is a dual-tower transformer network (GPT-2); weights are cloned between the towers, and each transformer block contains a multi-head cross-attention layer. As such, the prediction of each stream is conditioned on the conversational history of both speakers. The model is trained on the sum of cross-entropy losses for each speaker stream, where each tower consists of a GPT-2 model for each speaker in the dialogue. We note that while the GPT-2 model is not state-of-the-art, recent work has shown that more powerful GPT-4 class models make TRP prediction errors (Umair et al., 2024); this suggests that limitations are related to the informativeness of the lexical channel or model architecture rather than model capacity.

Turn-Level Annotation. Conversation analysis defines turn-constructional units (TCUs) as minimal complete contributions (Sacks et al., 1974) and Ford and Thompson (1996) shows that transition-relevance places (TRP) (i.e., where a turn may occur but need not) emerge from converging syntactic, intonational and pragmatic completion. However, operational definitions of turns do not always align with these theoretical concepts. A key advantage of PairwiseTurnGPT's dual-stream architecture is that it does not require commitment to a single such turn definition. Modelling dialogue as two time-aligned streams provides a grounding in acoustic events, which we use to label "turn-boundary" tokens.

We follow definitions by Ekstedt and Skantze (2020) to categorise specific turn-taking phenomena. We aggregate consecutive IPUs from the same speaker into a single turn unless interrupted by speech from the other speaker. Therefore, turn boundaries are defined where a speaker has yielded the floor, either through sufficient silence such that the other participant begins speaking, or through an actual speaker change. **Backchannels** involve one speaker interjecting a short utterance such as "hmm", "uh-huh" or "yeah" to provide feedback to the speaker (Ward, 2004). We define these based on their lexical content[1] and a pause of at least $1s$ between surrounding turns from the same speaker. **Complete overlap** occurs where one speaker begins and ends an utterance before the other speaker finishes theirs which encompasses a variety of heterogeneous phenomena — both competitive and cooperative overlaps. The remaining utterances are classified as general turns. The alignment process is demonstrated in Figure 1.

[1]We use the list of backchannel responses defined in (Ekstedt and Skantze, 2020)

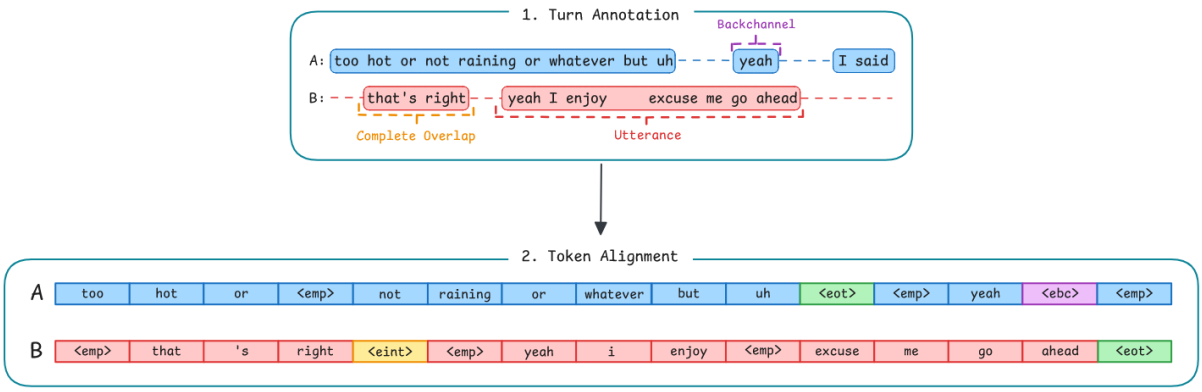

Figure 1: As in our prior work (Leishman et al., 2024), utterances are annotated as complete overlaps, backchannels and turns. <eot> tokens are added to the end of these utterances. Tokens are then time-aligned and <emp> tokens are inserted

Training Procedure and Setup. We train our own model on more data than the original implementation, combining the Telephone Speech Switchboard Corpus (Godfrey et al., 1992)[2] and Fisher English Training Speech (Cieri et al., 2004)—each of which contain 2430 and 11699 telephone conversations, respectively, between strangers. Non-verbal vocalisations are removed; partial words, mispronunciations and coinages are replaced with the intended word. We split the dataset randomly among train, validation and test sets of sizes (90/5/5) on the conversation level. Following the original implementation of PairwiseTurnGPT, we initialise a pre-trained *GPT2-base* model (12 layers, 12 attention heads and a hidden size of 768) [3] before fine-tuning on our spoken conversation data with a learning rate of $6.25e^{-5}$, a weight decay of 0.01 and a batch size of 4 for 2 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017) with sequences of maximum length 256.

4 Analysis of Turn-Taking Predictions

To investigate the source of complexity in turn-taking behaviour, we use a held-out evaluation set of Switchboard conversations and characterize each turn along two primary dimensions.

The **lexical components** encompass how turns are linguistically constructed:

- *Fluency*: KenLM-based negative log-likelihood score (Heafield, 2011). Predictable text is more likely to be grammatically well-formed and as a result, more fluent[4].
- *Coherence*: the cosine similarity of adjacent utterances' embeddings using the BERT's <cls> token. Utterances that share similar topics and concepts, with similar embeddings, are likely to be more logically connected [5].
- *Diversity*: inspired by Self-BLEU (Zhu et al., 2018) , we measure the lexical diversity of a turn relative to all turns in the same conversation. Turns that use similar words will have a higher scores while turns using a diverse vocabulary will have a lower score [6].

Turn-taking components capture temporal coordination and turn-taking phenomena:

- *Num. of Backchannels*: completed within the boundaries of the current turn
- *Num. of Complete Overlaps*: completed within the boundaries of the current turn
- *Turn-Shift Gap*: the time between the speaker ending their turn and the next speaker beginning their turn
- *Turn Length*: the number of tokens between the start of the turn and the end of the turn

4.1 Characterizing turn-taking complexity.

We begin by conducting Principal Component Analysis (PCA) on the seven turn-taking features to inspect conversational dynamics and set the stage for localising the contributions of lexical and temporal information to turn taking behaviour.

Table 1 reports the comparative feature loadings across a single component, indicating correlation

[2]We use the Mississippi State (Deshmukh et al., 1998) re-segmented transcripts for accurate word-level timings.

[3]Pre-trained weights were obtained from HuggingFace

[4]Trained using our Switchboard and Fisher training set.

[5]The model is fine-tuned using the Switchboard and Fisher training set with masked-language-modelling and next-sentence-prediction tasks.

[6]We calculate utterance BLEU scores by treating the target turn as the hypothesis and the rest of the utterances in the dialogue as the reference.

	PC1	PC2	PC3
Fluency Score	0.409	0.371	0.221
Diversity Score	-0.311	-0.111	0.248
Coherence Score	-0.007	0.847	0.242
Num. Complete Overlaps	0.314	-0.127	0.288
Num. Backchannels	0.533	-0.160	-0.081
Turn-Shift Gap	0.038	0.280	-0.860
Length	0.593	-0.111	-0.043
Turn-Taking Components	66.2%	35.2%	59.9%
Lexical Components	33.8%	64.8%	40.1%
Explained Variance	30.4%	15.5%	14.7%

Table 1: Feature contributions to each principal component. The first three principal components make up 60% of explained variance

between features (i.e., features with similar loading direction and magnitude on a component are positively correlated), and the distribution of variance across the components, providing information about redundancy between features (i.e., even spread of variance across many components indicates less redundancy).

The dominant source of conversational variation centres around structural turn properties. Accounting for 30.4% of the variance, PC1 shows strong positive loadings for turn length (0.593), backchannel frequency (0.533) and fluency (0.409). In other words, longer utterances tend to be more linguistically well-formed and are accompanied by more listener engagement through backchannels; which is to be expected as longer utterances have more backchannels opportunities.

The second component—explaining 15.5% of the variance—captures how well turns connect across speaker transitions. This dimension is dominated by coherence between turns (0.847) with moderate contributions from fluency (0.371) and turn-shift gap (0.280). The positive co-loadings suggest that longer inter-turn gaps are associated with more fluent utterances and greater coherence between turns. The differences between the two lexical features reflects the distinction between how semantically connected two adjacent utterances are (coherence), and whether an utterance conforms to typical language-use (fluency). Overall, we show that the temporal aspects of turn transitions are intrinsically linked to the underlying lexical content.

The third principal component reflects another aspect of how interlocutors interact. PC3 is negatively dominated by turn-shift gap (-0.860) with moderate positive contributions from the frequency of complete overlaps (0.288) and linguistic diversity (0.248). In total, it accounts for 14.7% of the variance. These loadings indicate that fast-paced turn-taking co-occurs with more constrained vocabulary choice and more simultaneous speech.

It is clear that the lexical and structural properties of turns vary in complex ways. As such, to understand the role of lexical and temporal information in turn-taking behaviour, we require a methodology that can account for such interactions and their effect on turn-taking predictability.

4.2 Mixed-Effects Modelling

We now directly quantify the contributions of turn-level features on the predictability of turn-taking using linear mixed-effects models to examine how features contribute individually and jointly.

Our models include fluency, coherence, diversity, number of backchannels, number of complete overlaps, turn-shift gap, and length as fixed effects with conversation ID as a random effect to control for potential effects of convergence (Cavalcanti and Skantze, 2025). As our dependent variable, we consider the model's behaviour when predicting the special <eot> token with two complementary metrics based on the probability assigned to the <eot> token by our autoregressive language model. Following previous work, we quantify this as surprisal (Smith and Levy, 2013; Goodkind and Bicknell, 2018; Wilcox et al., 2020); the surprisal of a particular unit u at time t can be defined as its Shannon entropy $s(u_t) = -\log(p(u_t|\mathbf{u_{<t}}))$, where $\mathbf{u_{<t}}$ denotes the preceding units. We consider:

1. TRUE s(<eot>), the surprisal of the <eot> token at actual turn-endings; and
2. IN-TURN s(<eot>), the minimum surprisal of the <eot> token at non-turn-end positions within a turn.

Lower TRUE s(<eot>) indicates that the correct prediction of an <eot> is assigned greater probability under our model, while lower IN-TURN s(<eot>) reflects the confidence of the model when *incorrectly* predicting an <eot> token. Taken together, these metrics provide a clearer understanding of turn-taking predictability: through turn-boundary recognition, TRUE s(<eot>) captures how well true turn-ends can be predicted by the dialogue model, while IN-TURN s(<eot>) reflects errors associated with overconfident turn predictions during turns. As longer utterances provide more opportunities for incorrect predictions, we note that IN-TURN s(<eot>) inherently reflects both turn-

Table 2: Statistical Results Comparison

Variable	TRUE s(<eot>)		IN-TURN s(<eot>)	
	Coef.	***p*-value**	**Coef.**	***p*-value**
Length	0.0815	0.0000	0.0468	0.0000
Number of Backchannels	0.0182	0.2831	-0.0173	0.0673
Number of Complete Overlaps	-0.0085	0.5196	-0.0428	0.0000
Turn Shift Gap	0.2614	0.0000	0.1045	0.0000
Coherence Score	0.0094	0.4695	-0.0110	0.1291
Fluency Score	0.0904	0.0000	-0.8408	0.0000
Diversity Score	-0.0299	0.0216	0.0806	0.0000

Table 3: Model Performance and Variance Explained by Each Parameter

Variable	TRUE s(<eot>)		IN-TURN s(<eot>)	
	Diff in R^2	**Variance %**	**Diff. in R^2**	**Variance %**
Length	0.0029	2.54	0.0010	0.14
Number of Backchannels	0.0002	0.18	0.0001	0.01
Number of Complete Overlaps	0.0001	0.09	0.0016	0.22
Turn Shift Gap	0.0643	56.40	0.0103	1.43
Fluency Score	0.0067	5.88	0.5729	79.48
Coherence Score	0.0001	0.09	0.0001	0.01
Diversity Score	0.0009	0.79	0.0061	0.85
Marginal R^2 (Fixed Effects)	**0.0970**	**85.09**	**0.7173**	**99.50**
Conditional R^2 (Full Model)	**0.1140**	**100.00**	**0.7209**	**100.00**

taking predictability and turn-length effects.

4.2.1 Main effects model

We first evaluate the independent effect of each feature on the predictability of turn-taking events. We also examine the explainable variance produced by each predictor by computing the difference in variance between the full model (trained with all predictors) and the variance of a reduced model with the predictor of interest removed (Nakagawa and Schielzeth, 2013).

Table 2 reveals fluency as the most influential factor in turn-taking predictability. The strong negative effect on IN-TURN s(<eot>) (-0.8408) suggests that fluency creates overconfidence, where pre-emptive turn-ending predictions are more likely in more fluent turns. Though surprising, this result is reinforced by variance decomposition in Table 3 where fluency accounts for 79.48% of the total explained variance in IN-TURN s(<eot>). Notably, this result does not extend to predictions of actual turn boundaries TRUE s(<eot>), where fluency accounts for 5.88% of the variance.

Turn-shift gap produces the strongest effect (0.2614) on TRUE s(<eot>) and explains a substantial 66% of variance. The positive coefficient indicates that overlapping speech onset serves as a dominant cue that the current turn is ending while gaps increase uncertainty. However, this reliance on temporal overlap is problematic for IN-TURN s(<eot>) where turn-shift gap is the second most influential individual factor (0.1045). When overlap occurs within turns, the premature turn-ending predictions become more likely.

In contrast to the substantial effects of turn-shift overlaps, the frequency of specific turn-taking phenomena highlight more sophisticated effects. Complete overlaps and backchannels show minimal impact on TRUE s(<eot>) (-0.0085 and 0.0182, both non-significant). While the model effectively utilises turn-shift overlaps as an important cue for turn-boundary detection, it is less dependent on these turn-taking phenomena which fulfil more coordinative roles in turn management. This demonstrates that lexical-temporal features provide sufficient information to distinguish these phenomena as turn-internal that do not necessarily signal turn completion. However, this distinction is not perfect: both phenomena show small negative effects on IN-TURN s(<eot>) (-0.0428 and -0.0173). While substantially weaker than fluency and turn-shift timing effects, the presence of turn-internal overlapping speech still creates some residual uncertainty.

Lexical and temporal features have complex effects on predictability. Paradoxically, lexical fluency creates prediction difficulties by a strong negative effect on surprisal for premature turn-end predictions (IN-TURN s(<eot>)). On the other hand, fluency has a minimal effect on turn boundary detection (TRUE s(<eot>)) where turn-shift overlap is by far the dominant cue (66% of variance). However, our PCA analysis demonstrates that fluency co-loads with turn length, backchannel frequency, and complete overlap frequency on PC1, indicates that these features do not vary independently and that main-effects interpretations may not be sufficient for understanding the role of lexical information on turn taking predictability.

4.2.2 Fluency-based Interaction Model

Table 4: Statistical Results with Interaction Effects

Variable	TRUE s(<eot>)		IN-TURN s(<eot>)	
	Coef.	***p*-value**	**Coef.**	***p*-value**
Length	0.0733	0.3058	-0.3232	0.0000
Number of Backchannels	0.0954	0.1381	0.0460	0.1899
Number of Complete Overlaps	-0.0254	0.3243	-0.1546	0.0000
Turn Shift Gap	0.2599	0.0000	0.1079	0.0000
Coherence Score	0.0102	0.4333	-0.0123	0.0819
Diversity Self-BLEU	-0.0289	0.0290	0.0630	0.0000
Fluency Score	0.0445	0.5679	-0.3013	0.0000
Interaction Effects				
Fluency × Length	0.0192	0.8798	0.6414	0.0000
Fluency × Backchannels	-0.1514	0.2134	-0.1228	0.0643
Fluency × Complete Overlaps	0.0388	0.4277	0.2410	0.0000
Fluency × Turn Shift Gap	0.0185	0.1393	-0.0377	0.0000

Based on the interdependencies highlighted in Section 4.1, we introduce interaction terms between fluency and each turn-taking component.

Table 4 shows the results of the interaction model. The strong negative main effect of fluency on IN-TURN $s(\texttt{<eot>})$ decreases from -0.8408 in the base model to -0.3013 when interactions are included, indicating that the effect of fluency depends on wider dialogue context.

Fluency and length show the strongest interaction effect for IN-TURN $s(\texttt{<eot>})$ (0.6414). While both features reduce turn-taking predictability individually (main effects of -0.3013 and -0.3232 for IN-TURN $s(\texttt{<eot>})$, respectively), their combined effect is less detrimental. Additional lexical context of longer turns may help disambiguate intermediary and genuine completion points in fluent turns. Similarly, overlaps reduce the likelihood of premature turn-end predictions (IN-TURN $s(\texttt{<eot>})$) in fluent contexts (0.2410). When speech is linguistically well-formed, the model better differentiates between complete-overlaps and turn-shift overlaps that signal genuine turn completion. Though only marginally significant, backchannels show a negative effect in fluent contexts (-0.1228). Even if it may be supportive, such overlapping speech may be misinterpreted as a turn-boundary cue.

For TRUE $s(\texttt{<eot>})$, turn-shift gap remains the dominant effect and shows no significant interaction with fluency (0.0185). This suggests that the reliance on overlapping transitions as a cue for turn-boundaries is equally strong regardless of whether the lexical content of speech is fluent or not.

These interaction patterns reveal that fluency's effects are mediated by the wider dialogue context. Longer fluent turns are associated with improved turn-taking predictability while complete overlaps in fluent speech reduce overconfidence. However, backchannels in fluent contexts increase overconfidence. These results establish specific circumstances where lexical-temporal information fails to provide sufficient information for turn-taking prediction. Lexical-temporal features cannot reliably distinguish supportive and competitive overlapping speech, or fully disambiguate between plausible completion points in fluent turns. In these scenarios, turn-taking may rely on other cues such as prosodic realisation.

4.3 Turn Boundary Surprisal Analysis

Finally, we propose examining the evolution of surprisal throughout the course of a turn. This temporal surprisal analysis differs fundamentally from mixed effects modelling—which provides quantitative examination of turn-level effects—by providing an exploratory examination of token-level patterns within individual turns.

Specifically, we examine the EOT-Surprisal by plotting token-level surprisal against its position relative to <eot> in fig. 2. This approach is motivated by our PCA findings, which reveal a complex interplay between lexical and turn-taking features. Mixed-effects modelling revealed that fluency effects are mediated by conversational structure and most prominently by length. Therefore, it is important to question how the apparent overconfidence in turn-taking predictability manifests throughout turns and in different contexts.

We split turns into stratified groups based on the occurrence of turn-taking phenomena: backchannels, complete overlaps and turn-shift overlap. Furthermore, given the strong effects of fluency, we also segment turns by categorising turns as either low or high fluency based on the quantiles of fluency scores. For each stratification, we calculate EOT-surprisal at each token position, aligned based on their respective distance to <eot>. Surprisal is aggregated across all turns using smoothed averages to reveal typical trajectory patterns for <eot> prediction approaching the true turn-end. Figure 2 visualises a 30-token window based on the average turn length in Switchboard (28.1). Turns shorter than 30 tokens are zero-padded (padding is excluded from aggregate statistics); longer turns are right-aligned to the turn boundary. This ensures that all turns contribute to their analysis at positions where they have tokens.

As expected, all groups and conditions display a gradual decrease in <eot> surprisal (i.e. the likelihood of predicting the <eot> token) across token positions. The temporal evolution of these predictions reveals patterns that corroborate and extend beyond the mixed-effects findings.

At the actual turn-end (position 0), the "Overlaps & No Backchannel" group achieves the lowest surprisal across almost all turn-groups and fluency conditions, confirming the finding that overlapping speech—particularly the combination of turn-shift overlap with complete overlaps—is a strong turn-boundary signal affecting TRUE $s(\texttt{<eot>})$. Additionally, turn-shift overlap produces steeper surprisal decreases towards position 0 than turn-shift gaps across both fluency conditions, reinforcing the reliance of overlapping speech onset as a prediction

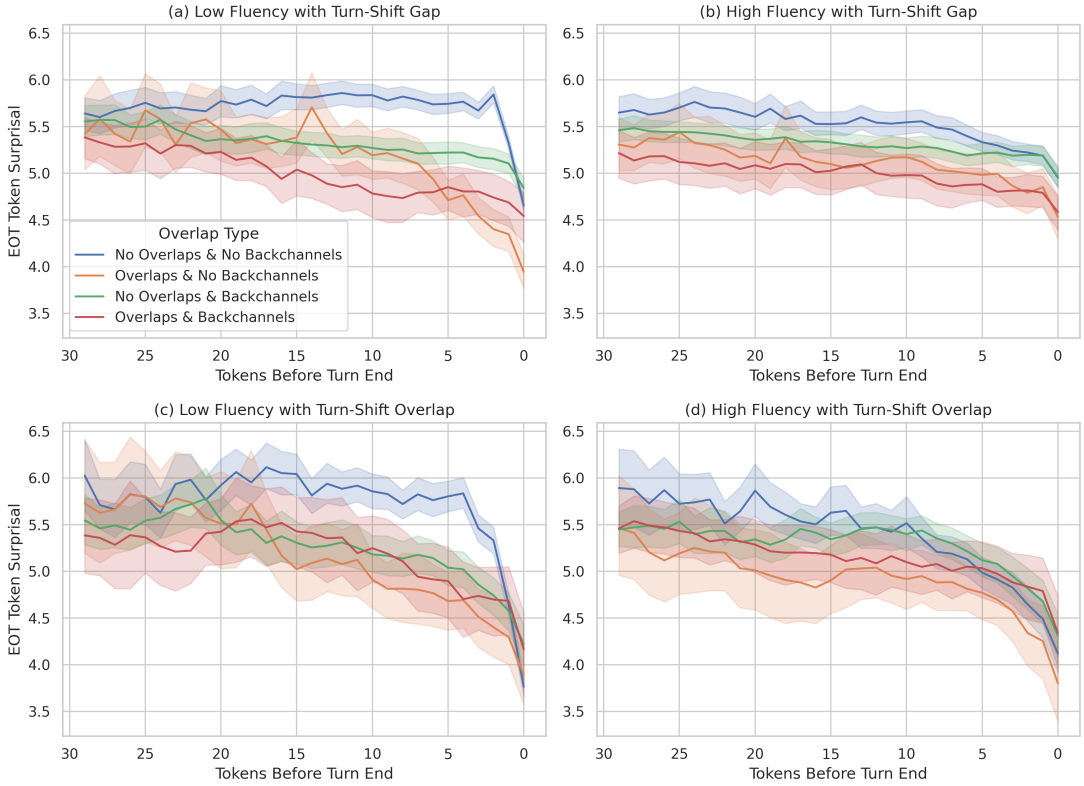

Figure 2: Plots of token-level surprisal approaching the ground-truth <eot> stratified by turn-taking phenomena and fluency: does/does not have overlaps/backchannels, turn-shift overlap/gap and high/low fluency. Plots are as follows (a) EOT-surprisal when turn-shifts are overlapping with low fluency; (b) EOT-surprisal when turn-shifts have a gap with high fluency; (c) the EOT-surprisal of the <eot> token when turn-shifts are overlapping with low fluency; and (d) the EOT-surprisal when turn-shifts have a pause with high fluency

cue regardless of lexical quality.

Examining the trajectory of predictions clarifies how fluency affects turn-taking prediction. Low-fluency conditions show greater variation in the surprisal trajectories caused by the presence of different types of turn-taking phenomena (Figure 2a,c). For example, when there is a turn-shift gap, the "No Overlap & No Backchannel" condition deviates from the gradual surprisal decrease pattern, maintaining consistent surprisal levels until 2-3 tokens from the true turn-end token position. This suggests that in disfluent contexts, the model maintains appropriate uncertainty until clear linguistic signals emerge near the true turn-end.

In contrast, high-fluency settings exhibit more consistent, gradual surprisal decrease across turn-types. High fluency with a turn-shift gap (Figure 2b) exhibits the most consistent <eot> surprisal where all turn-taking groups show a gradual decline and overall higher surprisal levels. These differences in temporal patterns point to different prediction processes that could explain the fluency effect identified in Section 4.2.2. For example, the lexical content of fluent speech may provide signals that dominate the cues from turn-taking phenomena, resulting in convergent trajectories. Disfluent speech, lacking clear intermediate completion points, may produce sharper, more decisive predictions concentrated near the actual boundary.

Across both fluency conditions, turn-shift gaps and overlaps have broadly similar effects. Turns with turn-shift overlaps (Figure 2c,d) show greater decreases in their surprisal as opposed to turn-shift gaps (Figure 2a,b). This analysis reinforces our finding that the onset of overlap at a turn-shift is a strong turn-completion cue, irrespective of lexical quality (i.e. fluency) but also demonstrates where the cue dominance occurs.

The temporal analysis corroborates the findings of section 4.2—for example, turn-shift overlap dominates boundary detection and operates regardless of fluency—while revealing when and how these effects unfold. Critically, it shows when uncertainty resolves: disfluent speech maintains uncertainty until sharp drops 2-3 tokens from turn-end, while fluent speech produces gradual decreases throughout. Differences in the temporal patterns between fluent and disfluent contexts sug-

gests differential reliance on information sources: disfluent contexts rely more on turn-taking cues while fluent contexts show lexical dominance.

5 Discussion & Conclusion

In this work, we systematically analyse the role of lexical and temporal information for predicting turn structures in spontaneous spoken conversations, with implications for both linguistic theory and the evaluation of computational models of dialogue. By analysing different aspects of turn structure prediction from PairwiseTurnGPT, a text model which preserves realistic temporal alignments of the lexical content in dialogue, our experiments establish where these features suffice and where additional cues, such as prosody, become necessary.

Most strikingly, we find that more fluent, well-formed speech is associated with greater predictive ambiguity about turn-ending (79.48% of IN-TURN s(<eot>) variance). This finding aligns with linguistic theories that frame conversation as series of turn-constructional units with multiple acceptable transition-relevance places (TRPs) (Sacks et al., 1974)—fluent speech may contain several linguistically complete points where turn-transfer may occur, while disfluent speech provides fewer plausible transition points. Our interaction analysis, motivated by the complexity highlighted in Section 4.2, shows that this effect is context-dependent: longer turns mitigate the problem through the accumulation of lexical context.

This context-dependence speaks to a fundamental question regarding the units of turn-taking. If lexical content is sufficient for turn-taking, we would expect fluent, well-formed speech to produce more certain predictions of turn boundaries. Instead, we see the opposite — fluency increases uncertainty. This pattern may be indicative of predicting multiple TRPs where turn-transfer could occur, while the actual turn-transition point can only be disambiguated by prosodic information.

Beyond fluency, our findings delineate both capabilities and limitations of lexical-temporal features. Notably, backchannels and complete overlaps have minimal impact on actual boundary detection (TRUE s(<eot>)), indicating partial success at disambiguating turn-internal phenomena from true turn-taking events. However, their presence is still associated with pre-emptive turn end-predictions, particularly backchannels (IN-TURN s(<eot>)). This functional ambiguity of supportive and competitive overlap likely depends on prosodic disambiguation; for example, backchannels use the same lexical content to achieve a range of communicative functions (Ward, 2004).

Our temporal analysis of surprisal in Section 4.3 reveals how predictability effects unfold dynamically. Disfluent speech maintains flat surprisal until 2-3 tokens from turn-end, while fluent turns show a more gradual decrease throughout. These contrasting temporal signatures suggest that different underlying prediction strategies: the model may rely more heavily on discrete closure markers in disfluent contexts, while fluent contexts maintain sustained ambiguity. This method provides a novel lens through which to explore the interaction of different information sources.

Our findings have direct implications for full-duplex models. Evaluation should move beyond corpus-level statistics; by stratifying performance, our analyses provide a framework for identifying settings where prosodic information may contribute to turn prediction. Similarly, these findings can be used to enhance training by exposing models to challenging scenarios like short, fluent utterances, which require cue integration.

This research establishes that lexical-temporal information can support sophisticated turn-taking behaviour, but with context-dependent limitations. Future research should examine whether prosodic features resolve the ambiguities we identify—particularly through controlled studies adding prosodic information to lexical-temporal baselines. Individual turn trajectory analysis or investigating finer-grained units such as turn-constructional units (TCUs), could test whether fluent turns indeed contain multiple plausible completion points, as our aggregate temporal patterns suggest. Conversely, investigating super-turn units may reveal how localised turn-taking behaviour is affected by overall conversational structure. This could be supported by an investigation to evaluate turn-taking across varying pause thresholds (Castillo-López et al., 2025) to test the robustness of our findings to different turn segmentation strategies.

Our approach of isolating information sources to identify their contributions to predictability provides a model for understanding multimodal language processing more widely, as well as how to model it computationally.

6 Limitations

This work is fundamentally, an exploratory study and there are many limitations, many of which we consider potential direction for future work.

Model We analyse predictions from a single model architecture, PairwiseTurnGPT based on GPT-2, which is not state-of-the-art. Modern language models may capture lexical-temporal patterns more effectively, making it difficult to determine whether our identified failures reflect genuine limitations of lexical-temporal features or simply are inadequate language modelling. More critically, we do not compare against models that integrate acoustic features, so we cannot distinguish truly unpredictable turn-endings and those that simply require prosodic information. Our findings establish that lexical-temporal features demonstrate specific failures, but cannot prove that prosodic information would be sufficient for rectifying them.

Interdependent features Our findings describe associations within naturally occurring dialogue. For example, longer turns tend to be more fluent and contain more backchannels which means that we cannot fully determine which feature drives observed effects. Experimental datasets with manipulated features could establish causality, but this study focuses specifically on turn-taking as it naturally occurs.

Lack of ground-truth of ambiguity and intermediate TRPs Within this analysis, we interpret IN-TURN s(<eot>), as reflecting ambiguity at potential, but incorrect, turn-end positions. However, this approach has two important shortcomings.

Firstly, we lack human judgments: we cannot fully determine whether model-defined ambiguity is (1) genuine ambiguity that humans experience, or (2) the models' inability to fully utilise available information. Secondly, our training data only contains actual turn-boundaries, not intermediate TRPs. Multiple intermediate TRPs are commonplace in longer utterances and are often lexically complete and therefore require prosodic disambiguation. However, proving this is not possible without using a TRP annotated dataset. Recently, Umair et al. (2024) collected a dataset of novel participant responses that reflect TRPs; though they only use single-stream models, they show that even state-of-the-art language models do not reflect human judgements. Applying our analysis of PairwiseTurnGPT predictions, which encode much longer contexts and turn structure, could offer a more generalisable description on the role of lexical and temporal information.

These limitations are avenues of future work: to further test the sufficiency of lexical-temporal features; to additionally identify where prosodic information is sufficient along with lexical information and how model-identified ambiguity aligns with human judgement.

References

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

Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. 2025. A survey of recent advances on turn-taking modeling in spoken dialogue systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pages 254–271, Bilbao, Spain. Association for Computational Linguistics.

Julio Cesar Cavalcanti and Gabriel Skantze. 2025. "Dyadosyncrasy", Idiosyncrasy and Demographic Factors in Turn-Taking. In *Interspeech 2025*, pages 1093–1097.

Kai-Wei Chang, En-Pei Hu, Chun-Yi Kuan, Wenze Ren, Wei-Chih Chen, Guan-Ting Lin, Yu Tsao, Shao-Hua Sun, Hung-yi Lee, and James Glass. 2025. Gametime: Evaluating temporal dynamics in spoken language models. *arXiv preprint arXiv:2509.26388*.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The fisher corpus: a resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Jan-Peter De Ruiter, Holger Mitterer, and Nick J Enfield. 2006. Projecting the end of a speaker's turn: A cognitive cornerstone of conversation. *Language*, 82(3):515–535.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Neeraj Deshmukh, Aravind Ganapathiraju, Andi Gleeson, Jonathan Hamaker, and Joseph Picone. 1998. Resegmentation of switchboard. In *ICSLP*. Syndey.

Starkey Duncan. 1972. Some signals and rules for taking speaking turns in conversations. *Journal of personality and social psychology*, 23(2):283.

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. 2022. How Much Does Prosody Help Turn-taking? Investigations using Voice Activity Projection Models. In *Proceedings of the 23rd Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 541–551. Association for Computational Linguistics.

Cecilia E. Ford and Sandra A. Thompson. 1996. *Interactional units in conversation: syntactic, intonational, and pragmatic resources for the management of turns*, page 134–184. Studies in Interactional Sociolinguistics. Cambridge University Press.

John J. Godfrey, Edward Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1:517–520 vol.1.

Patrick Gonzales. 1994. Talk at work: Interaction in institutional settings (studies in interactional sociolinguistics 8) edited by paul drew and john heritage. new york: Cambridge university press, 1992. 580 pp. *Issues in Applied Linguistics*, 5(1).

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)*, pages 10–18.

Agustín Gravano and Julia Hirschberg. 2011. Turn-taking cues in task-oriented dialogue. *Computer Speech & Language*, 25(3):601–634.

Kenneth Heafield. 2011. KenLM: Faster and smaller language model queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Ke Hu, Ehsan Hosseini-Asl, Chen Chen, Edresson Casanova, Subhankar Ghosh, Piotr Żelasko, Zhehuai Chen, Jason Li, Jagadeesh Balam, and Boris Ginsburg. 2025. Efficient and direct duplex modeling for speech-to-speech language model. *arXiv preprint arXiv:2505.15670*.

Sean Leishman, Peter Bell, and Sarenne Wallbridge. 2024. Pairwiseturngpt: a multi-stream turn prediction model for spoken dialogue. In *Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue - Full Papers*, Trento, Italy. SEMDIAL.

Stephen C Levinson. 2016. Turn-taking in human communication–origins and implications for language processing. *Trends in cognitive sciences*, 20(1):6–14.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H Liu, and Hung-yi Lee. 2025. Full-duplex-bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv e-prints*, pages arXiv–2503.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *ArXiv*, abs/1711.05101.

Lilla Magyari and Jan P De Ruiter. 2012. Prediction of turn-ends based on anticipation of upcoming words. *Frontiers in psychology*, 3:376.

Shinichi Nakagawa and Holger Schielzeth. 2013. A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in Ecology and Evolution*, 4(2):133–142.

Tu Anh Nguyen, Eugene Kharitonov, Jade Copet, Yossi Adi, Wei-Ning Hsu, Ali Elkahky, Paden Tomasello, Robin Algayres, Benoit Sagot, Abdelrahman Mohamed, and 1 others. 2023. Generative spoken dialogue language modeling. *Transactions of the Association for Computational Linguistics*, 11:250–266.

Yizhou Peng, Yi-Wen Chao, Dianwen Ng, Yukun Ma, Chongjia Ni, Bin Ma, and Eng Siong Chng. 2025. FD-Bench: A Full-Duplex Benchmarking Pipeline Designed for Full Duplex Spoken Dialogue Systems. In *Interspeech 2025*, pages 176–180.

Matthew Roddy, Gabriel Skantze, and Naomi Harte. 2018. Investigating speech features for continuous turn-taking prediction using lstms. *arXiv preprint arXiv:1806.11461*.

Harvey Sacks, Emanuel Schegloff, and Gail Jefferson. 1974. A simple systematic for the organisation of turn taking in conversation. *Language*, 50:696–735.

Emanuel Schegloff. 2000. Overlapping talk and the organization of turn-taking for conversation. *Language in Society*, 29.

Nathaniel J Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Muhammad Umair, Vasanth Sarathy, and Jan Ruiter. 2024. Large language models know what to say but not when to speak. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 15503–15514, Miami, Florida, USA. Association for Computational Linguistics.

Bandhav Veluri, Benjamin N Peloquin, Bokai Yu, Hongyu Gong, and Shyamnath Gollakota. 2024. Beyond turn-based interfaces: Synchronous LLMs as full-duplex dialogue agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21390–21402, Miami, Florida, USA. Association for Computational Linguistics.

Nigel Ward. 2004. Pragmatic functions of prosodic features in non-lexical utterances. In *Speech prosody*, volume 4, pages 325–328.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. *arXiv preprint arXiv:2006.01912*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.