Attribution and Alignment: Effects of Local Context Repetition on Utterance Production and Comprehension in Dialogue

Aron Molnar[◇] **Jaap Jumelet**[◁] **Mario Giulianelli**[◁] **Arabella Sinclair**[◇]
[◇]Department of Computing Science, University of Aberdeen
[◁]Institute for Logic, Language and Computation, University of Amsterdam
a.molnar.19@abdn.ac.uk j.w.d.jumelet@uva.nl
m.giulianelli@uva.nl arabella.sinclair@abdn.ac.uk

Abstract

Language models are often used as the backbone of modern dialogue systems. These models are pre-trained on large amounts of written *fluent* language. Repetition is typically penalised when evaluating language model generations. However, it is a key component of dialogue. Humans use *local* and *partner specific* repetitions; these are preferred by human users and lead to more successful communication in dialogue. In this study, we evaluate (a) whether language models produce human-like levels of repetition in dialogue, and (b) what are the processing mechanisms related to lexical re-use they use during comprehension. We believe that such joint analysis of model production and comprehension behaviour can inform the development of cognitively inspired dialogue generation systems.

1 Introduction

Human production in dialogue is influenced by many factors within the recent conversational history, leading speakers to repeat recently used lexical and structural elements of their own and their partners' language. These factors can involve conceptual pacts speakers make in order to establish common ground (Brennan and Clark, 1996), priming of lexical or syntactic cues which influences their subsequent re-use (Bock, 1986), and other social, interpersonal, cognitive, or neural influences (Pickering and Garrod, 2005; Danescu-Niculescu-Mizil et al., 2012; Hasson et al., 2012; Fusaroli et al., 2014).

Language models, which are often used as the backbone of modern dialogue systems, should learn to attend to such factors in order to successfully mimic human linguistic behaviour in interaction. The pre-training data of these models typically contains *fluent* monologic language and little diverse dialogue data—and indeed one goal of building language generators is having them produce fluent language. A key aspect of achieving fluency is the avoidance of repetition: repetitions are typically thought of as evidence of degenerate production (Li et al., 2016a,b; Welleck et al., 2019; Holtzman et al., 2019).

Recent advances in conversational language models, such as *ChatGPT*, demonstrate neural models' impressive performance in producing human-like, proficient language. However, despite these advances, they are yet to display human-like communicative behaviour (i.e., adhering to Gricean maxims—the verbosity of such models can be high), and more nuanced, local, and partner-specific interactions. Humans in dialogue use specific communication strategies which rely on repetition, and, in particular, these are *local* and *partner-specific* (Schlangen, 2004; Pickering and Garrod, 2005; Sinclair and Fernández, 2023). We start from the desideratum that dialogue response generation models should also produce *human-like* levels of repetition. While excessive levels of repetition, designed to mimic alignment, can hinder naturalness (Isard et al., 2006; Foster et al., 2009), humans generally prefer generated dialogue that contains higher levels of alignment (Lopes et al., 2015; Hu et al., 2016), which also lead to more successful communication in human-human dialogue (Xi et al., 2021; Isard et al., 2006). Moreover, elements of alignment have been successfully incorporated in chat bots (Hoegen et al., 2019; Gao et al., 2019).

Investigating and understanding the mechanisms which drive more human-like patterns of repetition is critical to creating more human-like natural language generation and dialogue systems. We therefore study whether models reproduce the repetition behaviour humans display in spoken dialogue, and the extent to which this repetition is affected by contextual cues. In particular, we focus on locality effects, comparing repetition patterns of speakers with respect to their own, and their partner's language. We investigate language models' *production* behaviour, via measuring

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 254–273
December 6–7, 2023.

the extent to which they generate similar local repetitions to humans, and their *comprehension* behaviour, through measuring the salience they assign to a given portion of the local dialogue context when comprehending an utterance.

2 Background

2.1 Human Repetition and Alignment

Local repetition of shared language between speakers is one of many lower-level linguistic signals indicating the presence of interactive alignment between speakers (Pickering and Garrod, 2004a). It is thought to contribute to more successful communication (Pickering and Garrod, 2005) as it allows speakers to establish and maintain shared common ground (Brennan and Clark, 1996; Pickering and Garrod, 2004b). Developing local routines—shared sequences of repeated language (Pickering and Garrod, 2005; Garrod and Pickering, 2007)—can also indicate mutual understanding between speakers (Wilkes-Gibbs and Clark, 1992; Gallotti et al., 2017). Producing repeated language in dialogue, either at a word level, or, in the case of routines, a construction level, is influenced by many factors in the local context. Speakers can be *primed* by language they have been recently exposed to, which may, in addition to the coordination and alignment factors mentioned above, play a role in the choice to repeat language *locally* (Tooley and Traxler, 2010). Priming effects can take place at multiple levels (from phonetic, lexical and syntactic to gesture, gaze and body posture), and are well attested in human dialogue (Brennan and Clark, 1996; Pardo, 2006; Reitter et al., 2006a; Holler and Wilkin, 2011; Rasenberg et al., 2020).

Alignment and coordination between speakers in dialogue are often measured in terms of *local* linguistic 'alignment effects', i.e., whether adjacent utterances contain high linguistic overlap, and whether the incidence of repetitions decays with the distance between utterances (Reitter et al., 2006b; Xu and Reitter, 2015; Sinclair et al., 2018; Sinclair and Fernández, 2021; Giulianelli et al., 2022). Local shared construction use has been linked to more successful grounded communication (Fusaroli et al., 2014; Reitter and Moore, 2007, 2014; Ward and Litman, 2007; Friedberg et al., 2012; Sinclair and Schneider, 2021; Norman et al., 2022). Local alignment is also affected by whether a speaker repeats their own or their partner's language, both in humans and in human-agent dialogue settings (Reitter et al., 2006b; Sinclair et al., 2018; Duplessis et al., 2017; Sinclair et al., 2019). We focus our attention on these short term, local repetition effects and structure our analyses accordingly.

2.2 Understanding the Behaviour of Language Models

Analysing model *behaviour* is a key approach when investigating patterns of model repetition, for example, paradigms from psycholinguistics can be repurposed to this end (e.g., Futrell et al., 2019). During language comprehension, language models have been shown to be prone to structural priming effects, in a manner with parallels to findings in humans. In particular, recency of prime to target within the input context heavily influences the likelihood of the congruent structure (Sinclair et al., 2022). It is less clear, however, to what extent models are affected by priming and repetition during language *production*, or generation, and what the mechanisms are that drive their *comprehension* behaviour. One method for *explaining* model behaviour is to employ interpretability techniques such as attribution methods. Attribution methods (Covert et al., 2021) allow for a high-level explanation of model behaviour that aligns strongly with how humans explain their decision-making, i.e., based on counterfactual examples (Yin and Neubig, 2022): *how would the prediction have changed if a particular input feature was not present?* Attribution methods have been used to examine *linguistic* patterns in model behaviour, and it has been argued they provide more comprehensive insights than attention heatmaps (Bastings and Filippova, 2020), because attention only determines feature importance within a particular attention head, and not for model predictions as a whole (Jain and Wallace, 2019). Linguistic phenomena investigated using attribution methods include co-reference, negation, and syntactic structure (Jumelet et al., 2019; Wu et al., 2021; Nayak and Timmapathini, 2021; Jumelet and Zuidema, 2023). Within conversational NLP, feature attribution methods have been used to identify salient features in task-oriented dialogue modelling (Huang et al., 2020), dialogue response generation (Tuan et al., 2021), and turn-taking prediction (Ekstedt and Skantze, 2020). However, relatively little work involves these techniques used to analyse human alignment behaviour in dialogue, in terms of patterns of local repetition, which we make our focus.

3 Experimental Setup

In this study, we investigate (a) to what extent repetition patterns in dialogue can be explained in terms of the re-use of lexical material in the local context; (b) whether LMs learn to generate repetitions with properties similar to those observed in human interaction and (c) how this relates to generation quality, as well as (d) whether LMs are influenced by the presence of repetitions in the local context when comprehending dialogue utterances. This section introduces the dialogue data and the language models used to study these four questions.[1]

3.1 Corpora

We choose two high-quality, naturalistic dialogue corpora, transcribed from spoken human interactions, with different conversational dynamics and well attested local repetition patterns at a lexical and structural level (Reitter et al., 2006a; Sinclair and Fernández, 2021). Although larger scale conversational corpora exist, often these consist of more artificial interactions (e.g., very short or highly closed-domain).

Map Task. The Map Task corpus (Anderson et al., 1991) comprises 128 dialogues between speakers participating in a navigational task. Speakers have either an instruction giver or instruction-follower role: they either describe a route, or attempt to follow and mark the described route, on their map.

Switchboard. The Switchboard corpus (Godfrey et al., 1992) contains 1,155 dialogues between participants making conversation over the telephone about one of a pre-specified range of common conversational topics. Speakers in this setting have equal status, with no pre-defined roles.

Extracting sample contexts. We are interested in evaluating the extent to which repetition occurs at a *local* level, therefore we extract sample contexts of 10 utterances, using a sliding window approach. Of these, utterances 1-9 are the *context*, and utterance 10 is the *target* utterance which we investigate. Since we are interested in between- vs. within-speaker effects, we define utterances based on speech turns—i.e. each time a speaker changes, we consider this a new utterance. Details of the corpora and extracted samples are in Table 1.

	Switchboard	Map Task
Full dialogues	1,155	128
Number of utterances	86.64±39.1	207.62±103.2
Unique vocabulary	19,927	1,882
Samples *(of 10 utterances)*	8,705	2,395
Words per utterance	14.6 ± 18.95	8.39 ± 9.21

Table 1: Corpus statistics.

3.2 Language Models

We select three autoregressive neural language models for our analysis: DialoGPT (DGPT; Zhang et al., 2020), GPT2 (Radford et al., 2019), and OPT (Zhang et al., 2022). We select DGPT as a model specifically designed for dialogue (yet still trained on written language, which differs significantly from our transcribed spoken language); GPT2 as its estimates are shown to be predictive of comprehension behaviour, even more so than larger LM variants (Shain et al., 2022; Oh and Schuler, 2023); and OPT, which has demonstrated competitive performance across a range of benchmarks (Paperno et al., 2016; Park, 2023). We fine-tune for 20 epochs, using an early stopping technique to save the best performing model based on perplexity.[2]

4 Producing Repetitions

We expect human repetition patterns to be highly local, given prior results showing priming effects in the same corpora (e.g., Reitter and Moore, 2007; Sinclair et al., 2018; Sinclair and Fernández, 2021). We also expect repetition patterns to be modulated by which dialogue partner is being repeated. In particular, we expect between-speaker repetition patterns to be the strongest given that developing shared routines can signal alignment and coordination of speakers' mental models or interpersonal synergy (Pickering and Garrod, 2005, 2004a; Fusaroli et al., 2014). We firstly analyse locality and between- vs. within-speaker repetition in human-produced utterances, then investigate whether the same patterns occur in model generations.

4.1 Methods

4.1.1 Measures of Repetition

To differentiate between routines vs. shared language, we compute two main measures of lexical repetition, at the word level, and in terms of shared word sequences (*constructions*; see

[1] https://github.com/the-context-lab/attribalign

[2] More details of model sizes can be found in Appendix C.

Section 4.1.2), with which we hope to capture between-speaker routines. We measure repetition between utterance pairs, at varying distances from one another within a given context sample. We define additional measures to capture established human dialogue behaviours.

Vocabulary Overlap. To compute vocabulary overlap, VO, we exclude punctuation, and calculate VO as the proportion of words w in the current turn t_c that also appear in a previous turn t_p:

$$VO = \frac{|w_{t_c} \cap w_{t_p}|}{|w_{t_c}|} \quad (1)$$

Construction Repetition. After extracting a shared inventory of constructions (Section 4.1.2) for a dialogue, we measure the proportion of repetition of shared constructions C as construction overlap CO as:

$$CO = \frac{|C_{t_c} \cap C_{t_p}|}{|w_{t_c}|} \quad (2)$$

Between vs. Within-Speaker Repetition. This binary measure describes whether the producer of utterance t_c and t_p is the same (*within*) or different (*between*).

Locality. We measure locality as the distance in utterance index between t_c and t_p. We take repetition decay, a negative effect of distance d on the shared constructions between t_c and t_p, as evidence of a local repetition effect.

Specificity. We calculate how sample-specific the extracted constructions are, and for each t_c, report average specificity of the repeated constructions. We measure specificity using pointwise mutual information (PMI), computed as follows:

$$PMI(c, s) = \log_2 \frac{P(c|s)}{P(c)} \quad (3)$$

Higher PMI indicates a construction c is more strongly associated with, or specific to, the sample s it occurs within due to the frequency of occurrence in this context being higher relative to its general usage.

4.1.2 Construction Extraction Procedure

To extract repeated constructions we make use of *dialign*, a framework for sequential pattern mining (Dubuisson Duplessis et al., 2017).[3] We then discard repeated expressions with fewer than two alphanumeric tokens (following Sinclair and Fernández, 2021). Repeated expressions consisting solely of punctuation or of more than half filled pauses are also excluded. We further discard constructions which contain *periods, commas and question marks*, to avoid constructions which include sentence boundaries: these do not contain the lexical elements we are interested in. We define the resulting shared lexicon as *constructions*. Table 2 provides details of their properties. [4]

	Switchboard			MapTask		
	M±Std	Med.	Max	M±Std	Med.	Max
Construction						
Length	2.1 ± 0.4	2.0	5	2.4 ± 0.8	2.0	11
Frequency	3.0 ± 1.2	3.0	6	3.3 ± 1.1	3.0	6
Rep. Dist.	3.6 ± 2.7	3.0	8	3.3 ± 2.7	3.0	8
Incidence	1.6 ± 1.1	1.0	10	2.0 ± 1.1	2.0	8
PMI	6.8 ± 3.4	6.6	11.5	7.2 ± 2.2	7.6	9.6
Utterance						
CO	0.004 ± 0.035	0.0	1.0	0.024 ± 0.13	0.0	2.8
VO	0.13 ± 0.23	0.008	1.0	0.13 ± 0.24	0.0	1.0

Table 2: Construction properties. Repetition distance (*Rep. Dist.*) measured in utterances.

4.1.3 Generating Dialogue Utterances

For each sample in our dataset of extracted dialogue excerpts, we precede each of the 9 utterances in the context with its speaker label, and append a final speaker label, corresponding to the upcoming target speaker, to the end. We then generate the target utterance using ancestral sampling (Bishop, 2006; Koller and Friedman, 2009) to study an unbiased representation of the model's predictive distribution. We set the maximum generation length to 64 tokens, and take the presence of a newline to indicate the end of an utterance, discarding any further generated text beyond this.[5] The resulting text we refer to as the target. To ensure that we take into account that a given context could support multiple targets—production variability is known to be high in dialogue (see, e.g., Giulianelli et al., 2023)—and to ensure our results are robust, we generate 5 utterances per context sample.

Evaluating generation quality. We measure the quality of a generated target utterance compared to the human reference in terms of their n-gram overlap (BLEU; Papineni et al., 2002) and semantic similarity (BERTScore; Zhang et al., 2019). We also

[3] https://github.com/GuillaumeDD/dialign

[4] Appendix E.1 contains examples of constructions and how they are repeated, Appendix D filled pauses.

[5] While the average token length for both datasets is relatively low, some utterances can be much longer. We analysed the distribution and select 64 as the maximum length since 95% and 99% of utterances fall below this length in Switchboard and in Map Task, respectively.

evaluate generations using perplexity, as computed using independent models, both independently of (PPL_{ii}), and conditioned on the context (PPL_{id}); we choose GPT-2 for the same reasons highlighted in Section 3.2, and Pythia (pythia-1.4b) (Biderman et al., 2023) for its open-source, highly performant properties. We additionally make use of MAUVE (Pillutla et al., 2021) to capture higher-level distributional differences between human- vs. model-produced text.

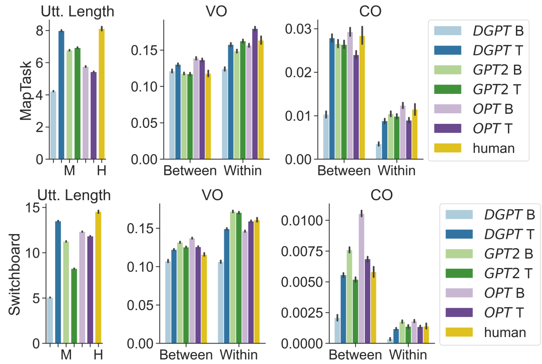


Figure 1: Human and model repetition properties. *B* indicates base models, *T* tuned models.

4.2 Analysis

4.2.1 Human vs. Model Repetitions

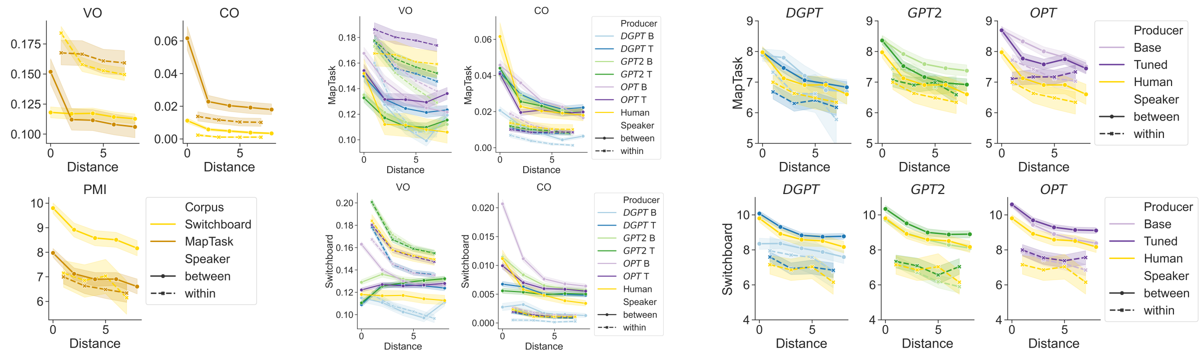
To analyse local production behaviour, we evaluate the extent to which human and model-produced utterances' CO is sensitive to between-speaker repetition, locality, and context-specificity.

The speaker being repeated affects CO and VO in humans and models. Dialogue partners differ in terms of what they repeat of their own vs. their partner's language (Reitter et al., 2006a; Sinclair et al., 2018), thus we expect to find differences in our human data. We also expect that if speakers make use of local routines (Pickering and Garrod, 2005), then between-speaker CO will be relatively higher. We observe that humans do indeed repeat constructions shared with their dialogue partner more so than they do those not shared (CO: Map Task: $t = 12.78$, $p < 0.05$. Switchboard: $t = 17.74$, $p < 0.05$). We observe the inverse effect for VO, showing speakers repeat their own language relatively more so than they do their dialogue partner (VO. Map Task: $t = -13.64$, $p < 0.05$. Switchboard: $t = -26.66$, $p < 0.05$). While models exhibit global human-like CO and VO patterns to some degree, for example GPT2 tuned is no different to human CO for within-speaker in Switchboard ($t = -0.18$, $p = 0.86$), and between-speaker in Map Task ($t = -1.86$, $p = 0.06$), these effects are not consistent across models or corpora. Figure 1 illustrates these results, details of statistical differences in Appendix E.

Humans produce repetitions locally. To evaluate the *local* effects of repetition, we employ linear mixed-effect models, including *dialogue, sample* and *speaker* identifiers as random effects.[6] We confirm that CO decays with the distance between a given utterance and those preceding it ($\beta = -0.001, p < 0.05, 95\%\ CI = [-0.001 : -0.001]$); this is not the case for VO (Figure 2a). Decay effects for CO are stronger for between-speaker repetition in both corpora. That is, speakers are more likely to repeat their partner's language locally. Interestingly, in Switchboard, decay effect are not observable when looking at the dialogue as a whole (Sinclair and Fernández, 2021). We hypothesise that other, less locally repeated constructions may drive down this effect when analysing the dialogues as a whole, or that some constructions may have multiple short bursts of local repetition over the course of a dialogue (Pierrehumbert, 2012).

Models learn some patterns of local repetition. We find that fine-tuned models learn turn-sensitive patterns of local repetition to some extent. Figure 2b demonstrates that models can learn similar patterns of local repetition to those observed in human dialogue. The most dramatic improvement in similarity to human behaviour is for DGPT. We find that in Switchboard, both models and humans show significant *local* repetition effects of CO independent of VO effects. Investigating CO in more detail, while human repetitions are sensitive to the length of the construction (longer constructions predict CO: $\beta = 0.035$, $p < 0.05$, $95\%\ CI = [0.025 : 0.045]$), this is not the case for models, for which the frequency of the repetition in the sample plays an important role in predicting CO (e.g. GPT2 repetition frequency: ($\beta = 0.01$, $p < 0.05$, $95\%\ CI = [0.007 : 0.013]$)). For Map Task, we find that humans repeat highly specific repetitions locally (CO $\beta = 0.006$, $p < 0.05$, $95\%\ CI = [0.003 : 0.009]$), however this is only true for GPT2 ($\beta = 0.001$, $p < 0.05$, $95\%\ CI = [0.0 : 0.002]$). Full model results in Appendix H.1.

[6] Full model output can be found in Appendix H. We include dialogue, sample and speaker as random effects, to allow for group-level variability in the linear model.



(a) Human CO, VO & PMI (b) Human vs. Model CO & VO. (c) Specificity (PMI) of repeated constructions.

Figure 2: Repetition effects for construction overlap CO and vocabulary overlap VO. Patterns of human vs. model repetition across contexts.

Models don't consistently produce speaker-specific repetitions. We find that while all models display significant CO speaker effects similar to humans, when taking into account other contextual factors, their behaviour with respect to *specificity* varies. While Figure 2c demonstrates that the PMI of constructions decays with distance, human speakers show no significant independent effect of PMI when predicting CO in either corpus. GPT2 exhibits the most similar behaviour to the human data in terms of the effect of distance and speaker on PMI in Map Task, however learns a significant negative relationship with PMI for Switchboard, not present in the human data. Full model results in Appendix H.1

		$PPL_m \downarrow$	$PPLg_{ii} \downarrow$	$PPLg_{id} \downarrow$	$PPLp_{ii} \downarrow$	$PPLp_{id} \downarrow$	BLEU	BertF1	Mve
SW									
GPT2	B	15.110	**3.770**	**2.870**	60.879	12.985	0.009	0.710	0.035
	T	**12.020**	3.830	2.880	**50.608**	**12.790**	**0.010**	**0.730**	**0.049**
OPT	B	37.540	**3.750**	2.870	54.706	12.799	0.010	0.700	0.052
	T	**15.130**	3.830	2.870	**45.488**	**12.635**	**0.014**	**0.733**	**0.069**
DGPT	B	6935.000	7.050	2.970	1323.338	14.064	0.000	0.656	0.006
	T	**10.910**	**3.570**	**2.870**	**41.700**	**12.735**	**0.016**	**0.730**	**0.049**
MT									
GPT2	B	16.170	**4.920**	**3.190**	**136.421**	18.353	0.006	0.679	0.101
	T	**7.930**	5.250	3.220	208.630	**18.193**	**0.014**	**0.702**	**0.245**
OPT	B	72.100	**5.270**	**3.210**	**199.344**	**18.189**	0.006	0.682	0.103
	T	**9.700**	5.730	3.240	294.677	18.384	**0.016**	**0.712**	**0.339**
DGPT	B	13014.000	6.670	3.280	998.832	19.852	0.002	0.662	0.041
	T	**8.050**	**5.320**	**3.220**	**235.385**	**18.007**	**0.016**	**0.699**	**0.176**

Table 3: Generation quality results. *SW*: SwitchBoard. *MT*: MapTask. PPL_m: Perplexity of the models under scrutiny on the analysis set. Perplexity of GPT2 ($PPLg_{ix}$) and PYTHIA ($PPLp_{ix}$) on model-produced utterances (ii independent of, and id dependent on context). *B*: base models, *T*: fine-tuned models. *Mve*: MAUVE score. **Bold** indicates the better value between base and fine-tuned variants.

4.2.2 Repetition vs. Quality

Finally, we investigate whether automatic NLG metrics capture human-likeness of repetition. This is an important aspect of naturalness in dialogue which the metrics are not explicitly designed for. Table 3 shows the relative generation quality of our base and fine-tuned models. Extended results can be found in Appendix B. All models demonstrate improvement with fine-tuning, although GPT2 base as an evaluator detects less difference than Pythia. This is expected, given their training data contains either little dialogue data, or a comparatively very different style of dialogue.

We find that the closer the levels of CO and VO are to human-produced language,[7] the higher *BertF1*, *BLEU*, and the lower the evaluation model perplexity both dependent and independent of the context. This correlation is strongest for GPT2 with $\rho = -0.395$, $p < 0.05$ for VO and $\rho = -0.258$, $p < 0.05$ for CO. This is perhaps to be expected for reference-based metrics, so we additionally inspect whether human-like CO levels correlate with MAUVE, a corpus-level metric, finding that more similar CO levels between human and model *inversely* correlate with MAUVE quality (above $\rho = 0.7$, $p < 0.05$ across models).[8] This tells us either that better corpus-level metrics need to be defined or, perhaps, that corpus-level evaluation is not really appropriate for dialogue where quality is determined by local and highly contextually dependent cues. This is in keeping with challenges in evaluating dialogue (Zhang et al., 2021; Liu et al., 2016), and suggests standard NLG evaluation approaches should be complemented by dialogue-specific metrics like the ones we use in our analysis.

[7] We measure this as the absolute value of the difference between human and model values.

[8] Table 9 in Appendix G provides a detailed breakdown of these results.

5 Interpreting Model Comprehension Behaviour

In the previous section, we investigated patterns of repetition in models' production behaviour. Now we turn our attention to their *comprehension* behaviour, making use of interpretability techniques to analyse what properties of the utterances in the context are more salient in determining expectations for a given target utterance. We expect models to learn patterns of turn-taking from the structure and contents of the context utterances (Wolf et al., 2019; Ekstedt and Skantze, 2020; Gu et al., 2020). We also expect that higher salience will be assigned to repetitions with local antecedents, in line with recency effects observed in model priming behaviour (Sinclair et al., 2022).

5.1 Methods

5.1.1 Feature Attribution

We obtain attributions over the dialogue context for a given target utterance, extracting scores for each token over the entire preceding context.[9] We are interested in examining behavioural patterns at the utterance level, in order to investigate the influence of their distance from the target, and design a measure to capture the *relative* boosting effects of the context for a given target utterance. This approach allows us to inspect attribution patterns across the context with respect to properties of the target utterance as a whole, allowing us to conduct similar, complementary analyses to the previous section.

A wide range of feature attribution methods exist (Lundberg and Lee, 2017; Murdoch et al., 2019). It remains an open question, however, which of these methods are most faithful with respect to the true model behaviour (Bastings et al., 2022). Some methods resolve this through defining theoretical properties that need to be satisfied by the method (Sundararajan et al., 2017). We focus on one such method, *DeepLift* (Shrikumar et al., 2017), which, besides its attractive theoretical properties, is also considerably more compute friendly than alternative attribution methods.

5.1.2 Attribution Aggregation Procedure

We design a measure that allows us to capture the relative effects that individual utterances in the local context have on models' utterance comprehension. Our measure aggregates over per-token attributions for a full utterance, returning relative prediction boosting effects of tokens within context utterances, speaker label tokens, and the target itself.

A given sample will consist of *speaker label tokens*, indicative of the change in speaker, e.g. *'A:'* and *'B:'*, the 9 context utterances, and the target utterance text. This can look like the following, with the speaker label tokens in orange, context utterances in dark blue, and the final target utterance of interest in light blue:

A: how are you? B: great, it's sunny A: about time B: agreed. A: I love sun B: me too A: makes me think of the beach B: the beach is great A: so great B:great, we should go to the beach!

Firstly, we create the feature attribution scores of each token in the input w_i with respect to the prediction of each token in the target utterance w_t:

$$\Phi \in \mathbb{R}^{|w_i| \times |w_t| \times n_{emb}} \tag{4}$$

Since feature attribution methods provide an importance score on the embedding level, we sum these scores along the embedding dimension n_{emb}.[10] Next, we sum the Φ matrix along the dimension of the tokens in the target utterance (w_t): creating a single score for each input token with respect to the target as a whole. Then, we create a single importance score for each individual input utterance or turn separator, denoted as a set T_i that contains the indices of the i^{th} utterance:

$$\Phi' \in \mathbb{R}^{|T|}, \quad \Phi'_i = \sum_{j \in T_i} \sum_k \sum_l \Phi_{j,k,l} \tag{5}$$

Note that the target utterance itself also yields importance scores of earlier tokens in the target with respect to later predictions.

The scores of Φ' are still unbounded, and can vary greatly between samples and models. We apply two further operations to allow sample and model comparison: we normalise the scores by the maximum absolute Φ' score, which maps the scores between -1 and 1, and we then centre the scores around the mean. This expresses the contribution of each element in the input as its *relative boosting effect* with respect to the other elements in the input

$$\Phi'' = \frac{\Phi'}{\max(|\Phi'|)} \tag{6}$$

$$\phi = \Phi'' - \text{mean}(\Phi'') \tag{7}$$

[9]For creating the attributions we make use of `Inseq` (Sarti et al., 2023) and `Captum` (Kokhlikyan et al., 2020).

[10]We could opt for the L2 norm as well, but this would hide negative contribution effects (Bastings et al., 2022).

5.2 Analysis

We now investigate model attribution patterns over the dialogue context. Our goal is to find out whether a model's comprehension behaviour exhibits robust patterns explainable through known psycholinguistic effects thought to influence human language producers, in particular *local, between-speaker* repetition patterns. While we are currently unable to understand precisely where humans place salience when comprehending, a large body of psycholinguistic research points to patterns of priming and alignment behaviour detectable from brain signals (Hasson et al., 2012; Futrell et al., 2019), and uses our understanding of the brain to inform analysis of neural language models (Hasson et al., 2020). We will contrast this analysis of model comprehension behaviour to the previous study of their production behaviour. We expect tuned models, the more human-like producers, to comprehend human language in a manner better predicted by factors thought to influence human processes—such as locality and priming effects—than base models.

5.2.1 Attributions Over Human Utterances

Humans and models display priming effects, which can be explained via accounts of residual activation, and they are sensitive to turn-taking (Ten Bosch et al., 2005; Tooley and Traxler, 2010; Ekstedt and Skantze, 2020; Sinclair et al., 2022). We thus expect attribution patterns to be sensitive to utterance position and speaker shifts within the context. Figure 3 shows how results change with fine-tuning.

Utterance comprehension is influenced by context locality in open domain dialogue. When comprehending utterances from a given speaker, models fine-tuned on Switchboard learn to attribute more salience to utterances in the nearby context, more strongly so when these are produced by the other speaker. This effect is strongest for GPT2 ($\beta = -0.009$, $p < 0.05$, $95\%\ CI = [-0.011 : -0.007]$). For Map Task, we do not see such a clear trend, with different behaviours between models. Even though evidence for sensitivity to utterance position and speaker shifts in comprehension is only found in one of the two corpora, this is an interesting result when juxtaposed to our analysis of production behaviour. It seems to indicate that while models learn to *understand* differences in speakers and in distance within the local context of open-domain dialogue, this does not always translate to human-likeness of *production* behaviour.

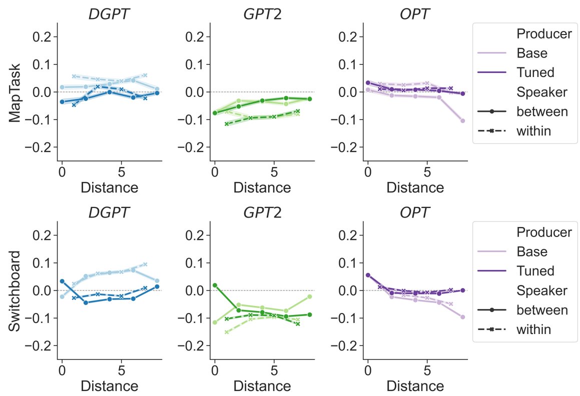


Figure 3: Relative attribution properties to human utterances over the dialogue context.

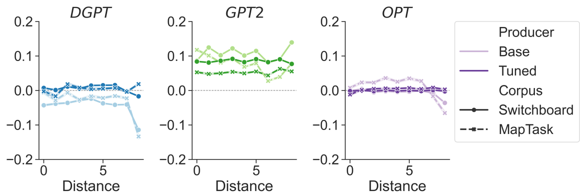


Figure 4: Relative attribution importance of speaker labels over the dialogue context.

Construction repetition in the local context predicts attribution patterns. High lexical repetition between context and target has been shown to boost priming effects in models (Sinclair et al., 2022), however, less is known about how this translates to attribution patterns. In line with priming results, we expect that attribution patterns over context utterances will be predicted by both construction and vocabulary overlap. We see mixed results across models, finding that only for Switchboard, GPT2 displays significant positive effect of CO ($\beta = 0.277$, $p < 0.05$, $95\%\ CI = [0.239 : 0.315]$) on attribution strength, independent of VO and distance effects. Surprisingly, however, the effect of VO on attribution strength is negative ($\beta = -0.308$, $p < 0.05$, $95\%\ CI = [-0.346 : -0.270]$). More remains to be done to precisely understand the relationship between the repetitions themselves and the local attribution patterns we observe, as well as to identify other factors driving this behaviour.

5.2.2 Attribution Over Special Tokens

While we are most interested in models' comprehension behaviour with respect to the utterance text in the context, we also investigate their behaviour over speaker labels. The effect of structural tokens on the performance and behaviour of LMs is an ongoing area of research (Wolf et al., 2019; Gu et al., 2020; Ekstedt and Skantze, 2020; Wallbridge et al., 2023). Speaker labels like *'A:'* and *'B:'* provide models with important information about the turn-taking dynamics of dialogues. Figure 4 shows that models learn, through fine-tuning, to attribute salience to speaker labels in a more *uniform* manner (note how the curves of tuned models are flatter). We find significant differences between base and tuned models in both corpora, with the highest boost in uniformity for DGPT (Switchboard: $\beta = 0.002$, $p < 0.05$, 95% $CI = [0.002 : 0.002]$, Map Task: $\beta = 0.005$, $p < 0.05$, 95% $CI = [0.005 : 0.005]$).[11] Speculatively, this could be taken as an indication that the models have learned to more consistently use these as structural markers of turn-taking. The discrepancy between the uniform attribution patterns over speaker labels and the decaying salience assigned to utterance text is an interesting finding that deserves more attention in future research.

6 Discussion & Conclusion

Repetition behaviour in dialogue, whether driven by local priming (Bock, 1986), alignment effects (Pickering and Garrod, 2004b), conceptual pacts (Brennan and Clark, 1996), or routinisation (Pickering and Garrod, 2005; Garrod and Pickering, 2007), is well attested in humans. In this study, we investigate the extent to which language models are sensitive to, and display the same *local*, *context-specific*, and *shared* patterns of construction repetition observed in human dialogue. We conduct an in-depth analysis using two corpora of English task-oriented and open-domain dialogue, and three autoregressive neural language models.

Analysing human interactions, we find that within highly local contexts (we consider dialogue samples consisting of 10 utterances), repetition effects decay with distance from antecedents, particularly when repetitions are between dialogue partners, rather than of a speaker's own language. This contrasts with and complements previous work finding no evidence of locality effects within Switchboard, the same open domain corpus, when considering dialogues as a whole rather than in short excerpts (Sinclair and Fernández, 2021), suggesting that some repeated constructions may occur in multiple short bursts (Pierrehumbert, 2012) over the course of a dialogue—a phenomenon that is not easily captured by more 'global' analyses.

We then evaluate model behaviour under two lenses: *production* behaviour, analysed in terms of the repetition of shared constructions (i.e., word sequences re-used by both dialogue participants) in model generations, and *comprehension* behaviour, measured by models' attribution of salience to contextual units when processing human-produced dialogue. We find that models learn, via fine-tuning, to generate more human-like patterns of construction re-use, although the degree to which repetitions are local, context-specific, and shared varies by model. We also find that while reference-based generation quality metrics correlate with the human-likeness of the repetitions produced, corpus-level metrics like MAUVE fail to capture this important aspect of dialogue quality. This highlights the need for more refined corpus-level approaches to statistical evaluation which take into account local and highly contextually dependent phenomena, or at least for their integration with instance-level analyses (Deng et al., 2022; Giulianelli et al., 2023). Making use of feature attribution techniques, which provide interpretations of models' comprehension behaviour, we then explore the extent to which models are sensitive to properties of the context thought to influence human propensity to produce *aligned* (i.e., locally repeated and context-specific) language. We observe that when comprehending utterances, tuned models assign salience to speaker labels in a more uniform manner, and that in open-domain dialogue, models learn to assign salience over the context in a more local manner.

We will follow up this study with experiments where our proposed attribution aggregation procedure is performed specifically over construction tokens in the target utterance. This may allow for more fine-grained interpretation of the relationship between repetitions and the observed local effects, as well as to investigate further psycholinguistic factors which may drive the tight coupling of local context and next utterance generation. We hope our experimental setup will inspire future work that attempts to create stronger connections between language model behaviour and findings from psycholinguistics. In particular, we look forward to seeing our attribution-based methodology being applied to other dialogue-specific phenomena, and the local, dyad-specific repetition measures we investigate applied to the development and evaluation of more adaptive and context-sensitive dialogue response generation systems.

[11] Full breakdown of results in Appendix H.2.

Limitations

Limitations of our work are that it is only conducted on English-spoken corpora, for two kinds types of dialogue context (conversational given a range of popular topics, and navigational task-oriented) and of that, native speakers of English only. Repetition patterns of dialogues in different conversational contexts, with language users of different cultures and in different languages may vary, and the patterns that models learn for these may also vary.

Acknowledgements

We would like to thank the anonymous reviewers for their thoughtful and useful reviews and comments. We also wish to thank Ehud Reiter for his useful comments on this work at an early stage. MG is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 819455).

References

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.

Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. 2022. "Will you find these shortcuts?" A protocol for evaluating the faithfulness of input salience methods for text classification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 976–991. Association for Computational Linguistics.

Jasmijn Bastings and Katja Filippova. 2020. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 149–155, Online. Association for Computational Linguistics.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.

Christopher M. Bishop. 2006. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg.

J Kathryn Bock. 1986. Syntactic persistence in language production. *Cognitive psychology*, 18(3):355–387.

Susan E Brennan and Herbert H Clark. 1996. Conceptual pacts and lexical choice in conversation. *Journal of experimental psychology: Learning, memory, and cognition*, 22(6):1482.

Ian Covert, Scott M. Lundberg, and Su-In Lee. 2021. Explaining by removing: A unified framework for model explanation. *J. Mach. Learn. Res.*, 22:209:1–209:90.

Cristian Danescu-Niculescu-Mizil, Lillian Lee, Bo Pang, and Jon Kleinberg. 2012. Echoes of power: Language effects and power differences in social interaction. In *Proceedings of WWW*, pages 699–708.

Yuntian Deng, Volodymyr Kuleshov, and Alexander M Rush. 2022. Model criticism for long-form text generation.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frederic Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 71–81.

Guillaume Dubuisson Duplessis, Chloé Clavel, and Frédéric Landragin. 2017. Automatic measures to characterise verbal alignment in human-agent interaction. In *18th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 71–81.

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.

Mary Ellen Foster, Manuel Giuliani, Amy Isard, Colin Matheson, Jon Oberlander, and Alois Knoll. 2009. Evaluating description and reference strategies in a cooperative human-robot dialogue system. In *Proceedings of the Twenty-first International Joint Conference on Artificial Intelligence (IJCAI-09)*.

Heather Friedberg, Diane Litman, and Susannah BF Paletz. 2012. Lexical entrainment and success in student engineering groups. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 404–409. IEEE.

Riccardo Fusaroli, Joanna Rączaszek-Leonardi, and Kristian Tylén. 2014. Dialog as interpersonal synergy. *New Ideas in Psychology*, 32:147–157.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of* *the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

M. Gallotti, M.T. Fairhurst, and C.D. Frith. 2017. Alignment in social interactions. *Consciousness and Cognition*, 48:253–261.

Xiang Gao, Yizhe Zhang, Sungjin Lee, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2019. Structuring latent spaces for stylized response generation. *arXiv preprint arXiv:1909.05361*.

Simon Garrod and Martin J Pickering. 2007. Alignment in dialogue. *The Oxford handbook of psycholinguistics*, pages 443–451.

Mario Giulianelli, Joris Baan, Wilker Aziz, Raquel Fernández, and Barbara Plank. 2023. What comes next? Evaluating uncertainty in neural text generators against human production variability. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2022. Construction repetition reduces information rate in dialogue. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 665–682, Online only. Association for Computational Linguistics.

J. Godfrey, E. Holliman, and J. McDaniel. 1992. Switchboard: telephone speech corpus for research and development . In *IEEE International Conference on Speech, and Signal Processing, ICASSP-92*, volume 1, pages 517–520.

Jia-Chen Gu, Tianda Li, Quan Liu, Zhen-Hua Ling, Zhiming Su, Si Wei, and Xiaodan Zhu. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2041–2044, New York, NY, USA. Association for Computing Machinery.

Uri Hasson, Asif A Ghazanfar, Bruno Galantucci, Simon Garrod, and Christian Keysers. 2012. Brain-to-brain coupling: a mechanism for creating and sharing a social world. *Trends in cognitive sciences*, 16(2):114–121.

Uri Hasson, Samuel A Nastase, and Ariel Goldstein. 2020. Direct fit to nature: an evolutionary perspective on biological and artificial neural networks. *Neuron*, 105(3):416–434.

Rens Hoegen, Deepali Aneja, Daniel McDuff, and Mary Czerwinski. 2019. An end-to-end conversational style matching agent. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*, pages 111–118.

Judith Holler and Katie Wilkin. 2011. Co-speech gesture mimicry in the process of collaborative referring during face-to-face dialogue. *Journal of Nonverbal Behavior*, 35:133–153.

Ari Holtzman, Jan Buys, Leo Du, Maxwell Forbes, and Yejin Choi. 2019. The Curious Case of Neural Text Degeneration. *CEUR Workshop Proceedings*, 2540.

Zhichao Hu, Gabrielle Halberg, Carolynn R Jimenez, and Marilyn A Walker. 2016. Entrainment in pedestrian direction giving: How many kinds of entrainment? *Situated dialog in speech-based human-computer interaction*, pages 151–164.

Xinting Huang, Jianzhong Qi, Yu Sun, and Rui Zhang. 2020. Generalizable and explainable dialogue generation via explicit action learning. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3981–3991, Online. Association for Computational Linguistics.

Amy Isard, Carsten Brockmann, and Jon Oberlander. 2006. Individuality and alignment in generated dialogues. In *Proceedings of the fourth international natural language generation conference*, pages 25–32.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Jaap Jumelet and Willem Zuidema. 2023. Feature interactions reveal linguistic structure in language models. pages 8697–8712.

Jaap Jumelet, Willem H. Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning, CoNLL 2019, Hong Kong, China, November 3-4, 2019*, pages 1–11. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Daphne Koller and Nir Friedman. 2009. *Probabilistic graphical models: Principles and techniques*. MIT press.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016a. A simple, fast diverse decoding algorithm for neural generation. *arXiv preprint arXiv:1611.08562*.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. 2016b. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

José Lopes, Maxine Eskenazi, and Isabel Trancoso. 2015. From rule-based to data-driven lexical entrainment models in spoken dialog systems. *Computer Speech & Language*, 31(1):87–112.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Anmol Nayak and Hari Prasad Timmapathini. 2021. Using integrated gradients and constituency parse trees to explain linguistic acceptability learnt by BERT. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 80–85, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Utku Norman, Tanvi Dinkar, Barbara Bruno, and Chloé Clavel. 2022. Studying alignment in a collaborative learning activity via automatic methods: The link between what we say and do. *Dialogue & Discourse*, 13(2):1–48.

Byung-Doh Oh and William Schuler. 2023. Why Does Surprisal From Larger Transformer-Based Language Models Provide a Poorer Fit to Human Reading Times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics - ACL '02*, pages 311–318.

Jennifer S Pardo. 2006. On phonetic convergence during conversational interaction. *The Journal of the Acoustical Society of America*, 119(4):2382–2393.

Daniel Park. 2023. Open-LLM-Leaderboard-Report.

Martin J. Pickering and Simon Garrod. 2004a. The interactive-alignment model: Developments and refinements. *Behavioral and Brain Sciences*, 27(2):212–225.

Martin J Pickering and Simon Garrod. 2004b. Toward a mechanistic psychology of dialogue. *Behavioral and Brain Sciences*, 27(2):169–190.

Martin J Pickering and Simon Garrod. 2005. Establishing and using routines during dialogue: Implications for psychology and linguistics. *Twenty-first century psycholinguistics: Four cornerstones*, pages 85–101.

Janet B Pierrehumbert. 2012. Burstiness of verbs and derived nouns. *Shall We Play the Festschrift Game? Essays on the Occasion of Lauri Carlson's 60th Birthday*, pages 99–115.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. MAUVE: Measuring the Gap Between Neural Text and Human Text using Divergence Frontiers.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Marlou Rasenberg, Asli Özyürek, and Mark Dingemanse. 2020. Alignment in multimodal interaction: An integrative framework. *Cognitive science*, 44(11):e12911.

David Reitter, Frank Keller, and Johanna D. Moore. 2006a. Computational modelling of structural priming in dialogue. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 121–124, New York City, USA. Association for Computational Linguistics.

David Reitter, Frank Keller, and Johanna D. Moore. 2006b. Computational modelling of structural priming in dialogue. In *HLT-NAACL 2006 - Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics, Short Papers*.

David Reitter and Johanna D. Moore. 2007. Predicting success in dialogue. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 808–815, Prague, Czech Republic. Association for Computational Linguistics.

David Reitter and Johanna D. Moore. 2014. Alignment and task success in spoken dialogue. *Journal of Memory and Language*, 76:29–46.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, and Arianna Bisazza. 2023. Inseq: An interpretability toolkit for sequence generation models. *CoRR*, abs/2302.13942.

David Schlangen. 2004. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143.

Cory Shain, Clara Meister, Tiago Pimentel, Ryan Cotterell, and Roger Philip Levy. 2022. Large-scale evidence for logarithmic effects of word predictability on reading time.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning Important Features Through Propagating Activation Differences. *34th International Conference on Machine Learning, ICML 2017*, 7:4844–4866.

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Arabella Sinclair, Adam Lopez, and Dragan Gasevic. 2018. Does ability affect alignment in second language tutorial dialogue? In *Proceedings of the 19th annual SIGdial meeting on discourse and dialogue*, pages 41–50.

Arabella Sinclair, Kate McCurdy, Christopher G Lucas, Adam Lopez, and Dragan Gaševic. 2019. Tutorbot corpus: Evidence of human-agent verbal alignment in second language learner dialogues. *International Educational Data Mining Society*.

Arabella J Sinclair and Raquel Fernández. 2021. Construction coordination in first and second language acquisition. In *Proceedings of the 25th Workshop on the Semantics and Pragmatics of Dialogue*.

Arabella J Sinclair and Raquel Fernández. 2023. Alignment of code switching varies with proficiency in second language learning dialogue. *System*, 113:102952.

Arabella J Sinclair and Bertrand Schneider. 2021. Linguistic and gestural coordination: Do learners converge in collaborative dialogue?. *International Educational Data Mining Society*.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Louis Ten Bosch, Nelleke Oostdijk, and Lou Boves. 2005. On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47(1-2):80–86.

Kristen M Tooley and Matthew J Traxler. 2010. Syntactic priming effects in comprehension: A critical review. *Language and Linguistics Compass*, 4(10):925–937.

Yi-Lin Tuan, Connor Pryor, Wenhu Chen, Lise Getoor, and William Yang Wang. 2021. Local explanation of dialogue response generation. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 404–416.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. 2023. Do dialogue representations align with perception? An empirical study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2688–2705.

Arthur Ward and Diane Litman. 2007. Dialog convergence and learning. *Frontiers in Artificial Intelligence and Applications*, 158:262.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319*.

Deanna Wilkes-Gibbs and Herbert H Clark. 1992. Coordinating beliefs in conversation. *Journal of Memory and Language*, 31(2):183–194.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A transfer learning approach for neural network based conversational agents. *arXiv preprint arXiv:1901.08149*.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, and Daniel Weld. 2021. Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6707–6723, Online. Association for Computational Linguistics.

Yadong Xi, Jiashu Pu, and Xiaoxi Mao. 2021. Taming repetition in dialogue generation. *arXiv preprint arXiv:2112.08657*.

Yang Xu and David Reitter. 2015. An Evaluation and Comparison of Linguistic Alignment Measures. In *6th Workshop on Cognitive Modeling and Computational Linguistics, CMCL 2015 at the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2015 - Proceedings*.

Kayo Yin and Graham Neubig. 2022. Interpreting language models with contrastive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 184–198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Chen Zhang, Grandee Lee, Luis Fernando D'Haro, and Haizhou Li. 2021. D-score: Holistic dialogue evaluation without reference. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2502–2516.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. BERTScore: Evaluating Text Generation with BERT.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

A Contributions

Conceptualisation: AS. Methodology: AS, JJ. Software: AM. Experiments: AM, AS. Analysis: AM, AS, MG, JJ. Writing - Original Draft: AM, AS. Writing - Review & Editing: AS, JJ, MG. Supervision & Project Administration: AS. Order alphabetical.

B Language Model Fine-Tuning

We fine-tune GPT-2 (Radford et al., 2019), OPT (Zhang et al., 2022), and DialoGPT (Zhang et al., 2020) for 20 epochs, using an early stopping technique to save the best performing model (based on its perplexity). Table 4 shows the perplexity of all models, pre-trained and fine-tuned, on the evaluation set. Models significantly adapt to the domain in training, given the low fine-tuned perplexities.

C Language Model Sizes

The considered language models have the following number of parameters. GPT2: 124M, OPT: 125M, DGPT: 117M, PYTHIA: 1.4B.

D Filled Pauses

We define filled pauses using the part-of-speech tags in Map Task and Switchboard. **Map Task:** *uh-huh, er, um, mm-mm, eh, uh, mm, uh-uh, nah, mm-hmm, erm, ehm, huh, hmm, mmhmm*. **Switchboard:** *hm, huh, uh, um-hum, huh, huh-uh, uh, uh-huh, um*.

		PPL ↓	Prec	Rec	F1	BLEU	BP ↓	LR ↓	Mve	L±Std
SW										
GPT2	B	15.110	0.722	0.704	0.710	0.009	0.744	0.772	0.035	11.9 ± 14.7
	T	12.020	**0.745**	0.720	0.730	0.010	**0.496**	**0.588**	0.049	8.8 ± 10.5
OPT	B	37.540	0.703	0.702	0.700	0.010	0.859	0.868	0.052	13.0 ± 13.8
	T	15.130	0.737	**0.733**	**0.733**	0.014	0.824	0.838	**0.069**	12.6 ± 12.9
DGPT	B	6935.000	0.667	0.648	0.656	0.000	0.148	0.343	0.006	3.3 ± 3.5
	T	**10.910**	0.737	0.728	0.730	**0.016**	0.955	0.956	0.049	14.3 ± 15.8
MT										
GPT2	B	16.170	0.681	0.680	0.679	0.006	0.827	0.841	0.101	7.1 ± 6.2
	T	**7.930**	0.705	0.702	0.702	0.014	0.849	0.859	0.245	7.4 ± 6.1
OPT	B	72.100	0.686	0.681	0.682	0.006	0.701	0.738	0.103	6.1 ± 6.4
	T	9.700	**0.723**	**0.705**	**0.712**	0.016	**0.631**	**0.685**	**0.339**	5.7 ± 5.2
DGPT	B	13014.000	0.668	0.659	0.662	0.002	0.391	0.516	0.041	3.7 ± 2.8
	T	8.050	0.701	0.700	0.699	**0.016**	0.990	0.990	0.176	8.5 ± 7.9

Table 4: Post-training metrics of models. *SW*: Switchboard. *MT*: Map Task. Precision (*Prec*), recall (*Rec*) and *F1* are averages over multiple samples and part of BERTScore. *LR*: length ratio (BLEU). *BP*: brevity penalty (BLEU). *PPL*: Perplexity. *B*: base models. *T*: tuned models. *Mve*: MAUVE score. *L*: mean target utterance length (in words). **Bold** indicates best values across models per corpora per metric.

E Construction Repetitions

E.1 Construction Examples

Table 5 contains two dialogue excerpts with responses generated by a tuned OPT model. Phrases highlighted bold refer to constructions generated by the model.

Table 6 lists the most frequent constructions generated by fine-tuned models, grouped by locality. Local and global constructions are defined as having a repetition distance of ≤ 4 and > 4, respectively. The table contains the top three most frequent produced constructions per model, per dataset, per locality.

E.2 Repetition Properties

Tables 7 and 8 contain detailed repetition statistics with statistical significance test results. In both corpora, DGPT learns to best approximate human target lengths after fine-tuning (*TH* columns of all models: -15, -92.8, and -38.59 (t) for DGPT, GPT2, and OPT, respectively. $p < 0.05$ for all). It generates significantly longer responses ($t = -412.64$, $p < 0.05$). Models robustly generate more dialogue-specific shared constructions after fine-tuned on Switchboard (t: -109.41, 57.44, -19.15, $p < 0.05$). After fine-tuned on Map Task, models learn to generate less dialogue-specific constructions (t: 19.83, 27.43, 22.85, $p < 0.05$). Models learn to produce more *distant* shared constructions after trained on both open-ended and task-oriented dialogue data (*SW*: t: -4.34, -10.2, -20.6, *MT*: t: -10.76, -0.19 ($p \geq 0.05$, exception), -8.53, $p < 0.05$). DGPT exhibits higher lev-

Selected local dialogue sample excerpts
A: but . that is a very good point. i am from **west virginia** so i understand what you are saying there B: uh-huh. A: and it's it's a very good point. B: (*OPT*) uh-huh. yeah, uh, where **i was**, uh, i went to a church in **west virginia** and uh, they always wore suits (...)
B: and **there's a lot of** graft, like people trying to tell them, oh, giving them information that was free to them anyway if they just knew how to get it. A: uh, yeah, exactly. B: yeah. A: (*OPT*) yeah, **there's a lot of** corruption,
B: right, so that's that's right, so i'm coming back back down **the paper** again A: ah, yeah back down **the paper** B: uh-huh aye turn right A: (*DGPT*) and then you go through **the paper** and past the chapel, to the right of the page
A: **okay right**, you went down past burnt forest B: i went underneath burnt forest A: well, you weren't meant to B: well you said draw round the cottage A: **okay right**, you're meant to come down from the start B: (*OPT*) **okay right**
A: oh, yeah, yeah, yeah. B: **in the summer** or like in the easter time, like around now? A: (*HUMAN*) no, usually **in the summer** time.

Table 5: Example local repetitions produced by tuned models.

els of construction overlap (CO) after fine-tuned on both Switchboard and Map Task (both between and within speakers), closely approximating human patterns (*SW*: t: -23.09, -11.45, *MT*: t: -29.75, -14.75, $p < 0.05$). GPT2 and OPT generally learn to produce lower CO values, but they already exhibit highly human-like construction overlap scores in their pre-trained states (*SW*: t: 6.83, 2.68, 16.52, 3.18, $p < 0.05$, *MT*: t: -1.62, -1.4, 0.75, 1.05, $p \geq 0.05$).

F Attributions To Target

We additionally analyse *Target vs. Context vs. Speaker Label* salience patterns. Regarding the *speaker labels* in the context (i.e., sequences containing non-utterance tokens: *A:*, <*eos*>), the effect of special or structural tokens on the performance and behaviour of LLMs is an ongoing area of research (Wolf et al., 2019; Gu et al., 2020; Wallbridge et al., 2023; Ekstedt and Skantze, 2020), we expect model attribution behaviour to be more similar between tuned models.

Figure 5: Attribution patterns for *Speaker labels* and *Utterances* in the dialouge Context (*Ctx*) during model comprehension of human Target (*Tgt*) utterances. The y-axis measures the *relative boosting effect*.

From Figure 5, we observe far higher variance in attribution over the target utterance than over the utterances in the context, with a similar relative difference between the speaker label in the target vs. those in the context. We observe very few consistent patterns across models in terms of relative boosting effects, except for *speaker label Ctx*, which becomes more relatively uniform (and closer to 0) with tuning. We observe that GPT2 learns to attribute relatively higher salience over the text in the context utterances than to that in the target. In other words, they learn to place relatively more importance on the target utterance itself (Switchboard: $t = -8.01$, $p < 0.05$; Map Task: $t = -14.42$, $p < 0.05$).

G Generation Quality

To perform a comparable correlation analysis of MAUVE scores and possibly influencing factors, we treat each model generation (we generate five responses to each sample) as a separate corpus. This allows us to compute multiple MAUVE scores for each model (instead of just one score that is based on all the model generations). For best practices, MAUVE requires at least a few thousand examples to run (the original paper uses 5000). Since we have $2,395$ samples in Map Task and $8,705$ samples in Switchboard, we select the number of samples used for MAUVE score computation to be $3,000$. We make use of all the Map Task samples for computation, and randomly sample model generations when we have more than $3,000$ examples available. We obtain five MAUVE scores for each model (base and fine-tuned), resulting in 30 scores for each corpus.

Table 9 shows a full breakdown of the most con-

distance		*Human*	*GPT2*	*OPT*	*DGPT*
MT	local	the diamond mine	the trout farm	the diamond mine	the abandoned cottage
		the concealed hideout	the diamond mine	the fallen pillars	have you got
		the rope bridge	to the left	to the left	the rift valley
	global	the pine forest	of the concealed hideout	edge of the map	outside of the
		don't have a	and a half	don't have a graveyard	a saloon bar
		the outlaws' hideout	two inches below where	of the walled city	up the map
SW	local	a lot of	a lot of	a lot of	a lot of
		i don't know	i don't know	i don't know	i don't know
		the peace corps	freedom of speech	one of the	the peace corps
	global	i used to	it was just	do you think	you're supposed to
		would be a	paying sales tax	i think it	i don't know if
		going to be	some of them	because i was	and a lot

Table 6: Example constructions from tuned models. *MT*: Map Task, *SW*: Switchboard. *Local*: repetition distance ≤ 4; *global*: repetition distance > 4.

	H	**DGPT** B	T	BH	TH	BT	**GPT2** B	T	BH	TH	BT	**OPT** B	T	BH	TH	BT
SW																
target len.	15.369	3.251	14.271	-174.840	-15.000	-412.640	11.925	8.802	-47.420	-92.800	108.160	13.026	12.599	-32.460	-38.590	14.090
constr. len.	2.176	2.117	2.185	-30.660	5.200	-55.900	2.196	2.186	11.070	5.750	9.400	2.239	2.215	33.810	21.410	19.790
PMI	8.520	8.053	8.821	-42.450	25.740	-109.410	8.424	8.907	-8.020	33.190	-57.440	9.147	9.303	53.330	67.020	-19.150
freq.	2.689	2.607	2.662	-21.530	-7.460	-22.690	2.778	2.672	24.660	-4.600	49.790	2.677	2.648	-3.230	-11.610	14.530
rep. dist.	3.525	3.363	3.891	-1.220	5.840	-4.340	3.586	3.990	0.980	7.040	-10.200	3.104	3.774	-6.870	3.950	-20.600
CO																
between	0.006	0.002	0.006	-16.910	-1.270	-23.090	0.008	0.005	6.830	-2.520	16.070	0.011	0.007	16.520	4.340	23.460
within	0.001	0.000	0.001	-9.860	-2.060	-11.450	0.002	0.001	2.680	-0.180	4.600	0.002	0.001	3.180	-0.400	6.340
VO																
between	0.116	0.107	0.122	-6.350	5.340	-15.770	0.132	0.125	12.700	7.920	8.530	0.137	0.126	18.620	8.920	17.100
within	0.161	0.106	0.149	-34.490	-7.960	-38.130	0.172	0.170	6.720	5.980	1.470	0.146	0.159	-10.800	-1.190	-16.190

Table 7: **Switchboard repetition statistics** with statistical significance tests. Red values indicate statistical *in*significance ($p \geq .05$). All values not highlighted red are statistically significant. The human (*H*), base model (*B*), and tuned model (*T*) columns contain averages. The base model–human (*BH*), tuned model–human (*TH*), and base model–tuned model (*BT*) comparison columns contain computed t-statistics. *Rep. dist.*: repetition distance. *Target len.*: target utterance length (in words). *Constr. len.*: construction length (in words). *Between/within*: between- and within-speaker. *Freq.*: frequency.

	H	**DGPT** B	T	BH	TH	BT	**GPT2** B	T	BH	TH	BT	**OPT** B	T	BH	TH	BT
MT																
target len.	8.607	3.701	8.488	-75.490	-1.710	-175.650	7.119	7.411	-22.220	-17.870	-10.990	6.062	5.670	-37.910	-44.360	15.530
constr. len.	2.373	2.272	2.240	-20.790	-28.610	11.740	2.321	2.287	-11.000	-18.390	13.830	2.427	2.403	11.210	6.270	8.260
PMI	7.063	7.339	7.113	18.580	3.220	19.830	7.652	7.341	39.130	18.180	27.430	7.956	7.722	60.480	44.730	22.850
freq.	3.249	2.980	2.999	-35.100	-32.780	-4.180	3.214	3.180	-4.590	-9.000	7.310	3.230	3.105	-2.470	-19.060	29.250
rep. dist.	3.281	2.736	3.554	-5.830	3.950	-10.760	3.439	3.447	2.270	2.390	-0.190	3.245	3.625	-0.530	4.840	-8.520
CO																
between	0.028	0.010	0.028	-20.600	-0.480	-29.750	0.027	0.026	-1.620	-1.860	0.320	0.029	0.024	0.750	-3.890	7.820
within	0.011	0.004	0.009	-14.300	-4.100	-14.750	0.010	0.010	-1.400	-2.380	1.370	0.012	0.009	1.050	-3.650	7.540
VO																
between	0.118	0.121	0.130	1.350	5.470	-6.160	0.118	0.117	0.020	-0.340	0.660	0.139	0.137	8.570	7.260	1.480
within	0.164	0.124	0.158	-13.920	-2.190	-19.590	0.149	0.162	-5.630	-0.380	-8.910	0.157	0.180	-2.370	5.050	-12.890

Table 8: **Map Task repetition statistics** with statistical significance tests. Red values indicate statistical *in*significance ($p \geq .05$). All values not highlighted red are statistically significant. The human (*H*), base model (*B*), and tuned model (*T*) columns contain averages. The base model–human (*BH*), tuned model–human (*TH*), and base model–tuned model (*BT*) comparison columns contain computed t-statistics. *Rep. dist.*: repetition distance. *Target len.*: target utterance length (in words). *Constr. len.*: construction length (in words). *Between/within*: between- and within-speaker. *Freq.*: frequency.

Metric	Type	Model	ρ	p
Construction Overlap	B	DGPT	0.914	0
Construction Overlap	B	GPT2	0.933	0
Construction Overlap	B	OPT	0.888	0.001
Construction Overlap	T	DGPT	0.698	0.025
Construction Overlap	T	GPT2	0.808	0.005
Construction Overlap	T	OPT	0.976	0
Prop. Repetition	B	DGPT	0.905	0
Prop. Repetition	B	GPT2	0.91	0
Prop. Repetition	B	OPT	0.944	0
Prop. Repetition	T	DGPT	0.637	0.047
Prop. Repetition	T	GPT2	0.747	0.013
Prop. Repetition	T	OPT	0.98	0

Table 9: MAUVE ρ correlation results. Metrics are the absolute value of the *difference* between model and human levels of CO and repetition, thus a positive correlation indicates an inverse correlation of the two metrics of human-likeness

sistent results across models. Since we are interested in general properties which apply to conversational corpora, we combine both Map Taskand Switchboardin this analysis. We find a strong ρ correlation across models, weakest for DGPT.

H Linear Mixed Effects Regression Results

To evaluate *local* effects, specifically the relationship between utterances in the context and the target utterance, we employ linear mixed-effect models, including *dialogue and sample* identifiers as random effects.

H.1 Production: Repetition Effects

To measure repetition effects we fit separate models for construction overlap CO, and vocabulary overlap VO, making these the dependent variables. We include dialogue and sample as random effects to allow for group-level variability in the linear model. We firstly investigate the effects of speaker, and distance. To measure repetition in the human data, we include speaker, and distance given speaker as fixed effects. To measure repetition in models, we follow the same process as for the human data, but adding model type (base or tuned) and their interaction with distance as additional fixed effects. Results for VO can be found in Table 10, and CO in Table 11.

We then conduct a second analysis, this time to investigate the impact of different properties of constructions on the CO effects. We include speaker, distance, construction length, specificity (PMI) and frequency as independent fixed effects. Results can be found in Table 12.

H.2 Comprehension: Attribution Effects

To measure Attribution strengths over the context utterances during model comprehension of human-produced target utterances, we made attribution the dependent variable.

H.3 Attribution Over Human Utterances

To investigate the effect of local context repetition on model attribution strengths to context utterance text during target utterance comprehension, we include speaker, distance, construction overlap, vocabulary overlap, average construction PMI, and construction frequency as fixed effects. Results can be found in Table 13.

H.4 Attribution Over Special Tokens

To investigate the effect of distance on model attribution to speaker labels within the context during target utterance comprehension, we include distance, model type (base or tuned) and their interaction as fixed effects. Results can be found in Table 14.

	Switchboard						*Map Task*					
	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]
Human												
Intercept	0.119	0.002	58.807	0.000	0.115	0.122	0.137	0.004	33.787	0.000	0.129	0.145
S[T.same]	0.064	0.003	19.889	0.000	0.058	0.071	0.033	0.007	5.013	0.000	0.020	0.045
dist:S[diff]	-0.001	0.000	-1.868	0.062	-0.001	0.000	-0.005	0.001	-6.592	0.000	-0.006	-0.003
dist:S[same]	-0.005	0.001	-10.705	0.000	-0.006	-0.004	-0.002	0.001	-1.488	0.137	-0.004	0.000
GPT2												
Intercept	0.129	0.001	110.696	0.000	0.127	0.132	0.129	0.002	67.475	0.000	0.125	0.133
S[T.same]	0.076	0.002	48.199	0.000	0.073	0.080	0.050	0.003	19.480	0.000	0.045	0.056
type[T.tuned]	-0.011	0.001	-10.672	0.000	-0.013	-0.009	-0.002	0.002	-1.357	0.175	-0.006	0.001
dist:S[diff]:type[base]	0.000	0.000	2.142	0.032	0.000	0.001	-0.003	0.000	-9.877	0.000	-0.003	-0.002
dist:S[same]:type[base]	-0.008	0.000	-36.207	0.000	-0.009	-0.008	-0.008	0.000	-20.167	0.000	-0.008	-0.007
dist:S[diff]:type[tuned]	0.002	0.000	11.460	0.000	0.002	0.002	-0.002	0.000	-8.011	0.000	-0.003	-0.002
dist:S[same]:type[tuned]	-0.006	0.000	-28.161	0.000	-0.007	-0.006	-0.004	0.000	-10.058	0.000	-0.005	-0.003
OPT												
Intercept	0.147	0.001	147.422	0.000	0.145	0.149	0.158	0.002	69.367	0.000	0.153	0.162
S[T.same]	0.034	0.001	25.623	0.000	0.032	0.037	0.034	0.003	11.096	0.000	0.028	0.040
type[T.tuned]	-0.015	0.001	-16.526	0.000	-0.017	-0.013	-0.010	0.002	-5.213	0.000	-0.014	-0.007
dist:S[diff]:type[base]	-0.003	0.000	-19.647	0.000	-0.003	-0.003	-0.005	0.000	-14.935	0.000	-0.006	-0.004
dist:S[same]:type[base]	-0.008	0.000	-38.836	0.000	-0.008	-0.007	-0.009	0.000	-19.171	0.000	-0.009	-0.008
dist:S[diff]:type[tuned]	-0.001	0.000	-5.039	0.000	-0.001	-0.000	-0.002	0.000	-7.227	0.000	-0.003	-0.002
dist:S[same]:type[tuned]	-0.003	0.000	-12.382	0.000	-0.003	-0.002	-0.001	0.000	-2.042	0.041	-0.002	-0.000
DGPT												
Intercept	0.104	0.001	69.536	0.000	0.101	0.107	0.142	0.002	65.090	0.000	0.138	0.146
S[T.same]	0.047	0.002	27.535	0.000	0.043	0.050	0.027	0.003	9.267	0.000	0.021	0.032
type[T.tuned]	0.018	0.001	13.055	0.000	0.015	0.020	0.001	0.002	0.427	0.669	-0.003	0.005
dist:S[diff]:type[base]	0.001	0.000	3.648	0.000	0.000	0.001	-0.004	0.000	-11.628	0.000	-0.005	-0.003
dist:S[same]:type[base]	-0.007	0.000	-23.073	0.000	-0.008	-0.007	-0.010	0.000	-22.139	0.000	-0.011	-0.009
dist:S[diff]:type[tuned]	0.001	0.000	3.920	0.000	0.000	0.001	-0.004	0.000	-11.219	0.000	-0.004	-0.003
dist:S[same]:type[tuned]	-0.005	0.000	-22.278	0.000	-0.006	-0.005	-0.004	0.000	-9.171	0.000	-0.005	-0.003

Table 10: Repetition effects for Vocabulary Overlap VO. *S* indicates speaker, *type* indicates model type (base or fine-tuned), *diff* indicates whether the two utterances come from different speakers, or between-speaker repetition, and *same* indicates whether the two utterances come from the same speakers, or within-speaker repetition.

	Switchboard						*Map Task*					
	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]
Human												
Intercept	0.009	0.000	31.878	0.000	0.009	0.010	0.047	0.002	29.468	0.000	0.043	0.050
S[T.same]	-0.007	0.000	-14.930	0.000	-0.008	-0.006	-0.033	0.003	-12.807	0.000	-0.038	-0.028
dist:S[diff]	-0.001	0.000	-15.367	0.000	-0.001	-0.001	-0.005	0.000	-15.659	0.000	-0.005	-0.004
dist:S[same]	-0.000	0.000	-2.386	0.017	-0.000	-0.000	-0.001	0.000	-1.471	0.141	-0.001	0.000
GPT2												
Intercept	0.010	0.000	63.140	0.000	0.009	0.010	0.037	0.001	54.133	0.000	0.036	0.038
S[T.same]	-0.006	0.000	-27.845	0.000	-0.006	-0.005	-0.023	0.001	-25.390	0.000	-0.025	-0.021
type[T.tuned]	-0.003	0.000	-19.413	0.000	-0.004	-0.003	-0.000	0.001	-0.624	0.533	-0.002	0.001
dist:S[diff]:type[base]	-0.001	0.000	-19.494	0.000	-0.001	-0.000	-0.003	0.000	-21.228	0.000	-0.003	-0.002
dist:S[same]:type[base]	-0.000	0.000	-12.555	0.000	-0.001	-0.000	-0.001	0.000	-5.939	0.000	-0.001	-0.001
dist:S[diff]:type[tuned]	-0.000	0.000	-7.264	0.000	-0.000	-0.000	-0.003	0.000	-21.669	0.000	-0.003	-0.002
dist:S[same]:type[tuned]	0.000	0.000	2.012	0.044	0.000	0.000	-0.001	0.000	-5.276	0.000	-0.001	-0.001
OPT												
Intercept	0.016	0.000	103.178	0.000	0.015	0.016	0.043	0.001	58.941	0.000	0.042	0.045
S[T.same]	-0.011	0.000	-52.886	0.000	-0.011	-0.010	-0.024	0.001	-24.048	0.000	-0.025	-0.022
type[T.tuned]	-0.006	0.000	-32.546	0.000	-0.006	-0.005	-0.010	0.001	-13.559	0.000	-0.012	-0.009
dist:S[diff]:type[base]	-0.001	0.000	-49.486	0.000	-0.001	-0.001	-0.004	0.000	-26.986	0.000	-0.004	-0.003
dist:S[same]:type[base]	-0.001	0.000	-17.805	0.000	-0.001	-0.001	-0.002	0.000	-10.631	0.000	-0.002	-0.001
dist:S[diff]:type[tuned]	-0.001	0.000	-25.315	0.000	-0.001	-0.001	-0.002	0.000	-16.731	0.000	-0.002	-0.002
dist:S[same]:type[tuned]	0.000	0.000	8.118	0.000	0.000	0.000	-0.000	0.000	-0.706	0.480	-0.000	0.000
DGPT												
Intercept	0.004	0.000	21.791	0.000	0.003	0.004	0.022	0.001	33.796	0.000	0.020	0.023
S[T.same]	-0.004	0.000	-24.266	0.000	-0.004	-0.004	-0.019	0.001	-23.392	0.000	-0.021	-0.018
type[T.tuned]	0.003	0.000	16.913	0.000	0.003	0.003	0.013	0.001	19.424	0.000	0.012	0.014
dist:S[diff]:type[base]	-0.000	0.000	-10.319	0.000	-0.000	-0.000	-0.002	0.000	-19.909	0.000	-0.003	-0.002
dist:S[same]:type[base]	0.000	0.000	3.740	0.000	0.000	0.000	0.000	0.000	0.303	0.762	-0.000	0.000
dist:S[diff]:type[tuned]	-0.000	0.000	-10.197	0.000	-0.000	-0.000	-0.002	0.000	-17.875	0.000	-0.002	-0.002
dist:S[same]:type[tuned]	-0.000	0.000	-8.171	0.000	-0.000	-0.000	-0.001	0.000	-9.446	0.000	-0.002	-0.001

Table 11: Repetition effects for Construction Overlap CO. *S* indicates speaker, *type* indicates model type (base or fine-tuned), *diff* indicates whether the two utterances come from different speakers, or between-speaker repetition, and *same* indicates whether the two utterances come from the same speakers, or within-speaker repetition.

	Switchboard						*Map Task*					
	Coef.	Std.	z	$P > \|z\|$	$[0.025$	$0.975]$	Coef.	Std.	z	$P > \|z\|$	$[0.025$	$0.975]$
Human												
Intercept	0.074	0.021	3.505	0.000	0.033	0.116	0.099	0.028	3.554	0.000	0.045	0.154
S[T.same]	-0.006	0.011	-0.533	0.594	-0.029	0.016	-0.031	0.015	-2.061	0.039	-0.060	-0.002
dist	-0.003	0.001	-4.506	0.000	-0.005	-0.002	-0.004	0.001	-3.330	0.001	-0.006	-0.001
avg_constr_len	0.057	0.006	10.155	0.000	0.046	0.068	0.133	0.007	18.607	0.000	0.119	0.146
pmi_avg	0.001	0.001	0.865	0.387	-0.001	0.003	0.003	0.002	1.427	0.154	-0.001	0.008
freq_constr	-0.014	0.004	-3.392	0.001	-0.023	-0.006	-0.035	0.005	-7.074	0.000	-0.045	-0.025
BASE												
GPT2												
Intercept	0.048	0.010	4.629	0.000	0.028	0.068	0.109	0.014	7.533	0.000	0.081	0.137
S[T.same]	-0.026	0.006	-4.395	0.000	-0.037	-0.014	-0.017	0.008	-2.138	0.032	-0.033	-0.001
dist	-0.004	0.001	-8.614	0.000	-0.006	-0.003	-0.005	0.001	-5.689	0.000	-0.006	-0.003
avg_constr_len	0.058	0.003	19.832	0.000	0.052	0.064	0.127	0.004	29.966	0.000	0.119	0.135
pmi_avg	0.002	0.000	3.454	0.001	0.001	0.002	0.004	0.001	3.865	0.000	0.002	0.006
freq_constr	0.005	0.002	2.150	0.032	0.000	0.009	-0.016	0.003	-6.018	0.000	-0.022	-0.011
OPT												
Intercept	0.022	0.007	3.110	0.002	0.008	0.036	0.088	0.016	5.516	0.000	0.057	0.119
S[T.same]	-0.025	0.005	-5.151	0.000	-0.034	-0.015	-0.030	0.010	-3.134	0.002	-0.049	-0.011
dist	-0.004	0.000	-9.875	0.000	-0.004	-0.003	-0.007	0.001	-8.165	0.000	-0.008	-0.005
avg_constr_len	0.077	0.002	41.700	0.000	0.073	0.081	0.134	0.004	37.148	0.000	0.127	0.141
pmi_avg	0.001	0.000	3.862	0.000	0.001	0.002	0.004	0.001	3.105	0.002	0.001	0.006
freq_constr	-0.000	0.002	-0.232	0.816	-0.004	0.003	-0.003	0.003	-1.162	0.245	-0.009	0.002
DGPT												
Intercept	0.314	0.084	3.759	0.000	0.150	0.478	0.162	0.035	4.594	0.000	0.093	0.231
S[T.same]	-0.041	0.039	-1.059	0.290	-0.117	0.035	-0.011	0.017	-0.623	0.533	-0.044	0.023
dist	-0.010	0.004	-2.844	0.004	-0.017	-0.003	-0.006	0.002	-3.210	0.001	-0.010	-0.002
avg_constr_len	0.083	0.027	3.099	0.002	0.030	0.135	0.115	0.009	12.720	0.000	0.097	0.132
pmi_avg	0.000	0.007	0.059	0.953	-0.013	0.014	0.008	0.003	2.914	0.004	0.003	0.014
freq_constr	-0.019	0.009	-2.059	0.039	-0.037	-0.001	-0.002	0.007	-0.237	0.812	-0.015	0.012
TUNED												
GPT2												
Intercept	0.202	0.020	10.227	0.000	0.163	0.241	0.059	0.014	4.282	0.000	0.032	0.087
S[T.same]	-0.030	0.010	-2.920	0.004	-0.051	-0.010	-0.031	0.007	-4.447	0.000	-0.044	-0.017
dist	-0.005	0.001	-5.801	0.000	-0.007	-0.004	-0.006	0.001	-7.508	0.000	-0.007	-0.004
avg_constr_len	0.067	0.006	11.523	0.000	0.055	0.078	0.128	0.004	28.787	0.000	0.119	0.137
pmi_avg	-0.010	0.001	-11.189	0.000	-0.012	-0.008	0.004	0.001	4.017	0.000	0.002	0.005
freq_constr	0.004	0.004	1.032	0.302	-0.004	0.013	-0.011	0.003	-4.175	0.000	-0.016	-0.006
OPT												
Intercept	0.056	0.010	5.793	0.000	0.037	0.075	0.192	0.018	10.965	0.000	0.158	0.227
S[T.same]	-0.025	0.006	-4.117	0.000	-0.038	-0.013	-0.057	0.010	-5.581	0.000	-0.077	-0.037
dist	-0.003	0.000	-6.406	0.000	-0.004	-0.002	-0.006	0.001	-6.700	0.000	-0.008	-0.004
avg_constr_len	0.064	0.003	24.984	0.000	0.059	0.069	0.123	0.004	28.582	0.000	0.114	0.131
pmi_avg	0.001	0.000	3.123	0.002	0.001	0.002	-0.001	0.001	-1.085	0.278	-0.004	0.001
freq_constr	-0.004	0.002	-2.011	0.044	-0.009	-0.000	-0.022	0.003	-6.438	0.000	-0.029	-0.016
DGPT												
Intercept	0.023	0.009	2.429	0.015	0.004	0.041	0.124	0.015	8.252	0.000	0.094	0.153
S[T.same]	-0.015	0.005	-3.130	0.002	-0.024	-0.006	-0.026	0.007	-3.524	0.000	-0.040	-0.011
dist	-0.005	0.000	-10.320	0.000	-0.006	-0.004	-0.005	0.001	-5.817	0.000	-0.006	-0.003
avg_constr_len	0.054	0.003	18.517	0.000	0.048	0.059	0.110	0.005	22.849	0.000	0.100	0.119
pmi_avg	0.001	0.000	2.872	0.004	0.000	0.002	-0.002	0.001	-2.332	0.020	-0.004	-0.000
freq_constr	0.003	0.002	1.717	0.086	-0.000	0.007	-0.013	0.003	-4.412	0.000	-0.019	-0.007

Table 12: Repetition details for *CO* taking into account length, sepcificity (PMI) and construction frequency (freq). *S* indicates speaker, *type* indicates model type (base or fine-tuned), *diff* indicates whether the two utterances come from different speakers, or between-speaker repetition, and *same* indicates whether the two utterances come from the same speakers, or within-speaker repetition.

	Switchboard						*Map Task*					
	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]
BASE												
GPT2												
Intercept	0.399	0.010	39.506	0.000	0.380	0.419	0.457	0.016	28.858	0.000	0.426	0.488
S[T.same]	0.003	0.006	0.493	0.622	-0.009	0.014	-0.015	0.008	-1.752	0.080	-0.031	0.002
dist_from_prev_turn	0.002	0.001	3.559	0.000	0.001	0.003	-0.000	0.001	-0.199	0.842	-0.002	0.002
constr_overlap	0.323	0.015	22.127	0.000	0.294	0.351	0.190	0.024	7.797	0.000	0.142	0.237
vocab_overlap	-0.383	0.013	-30.143	0.000	-0.408	-0.358	-0.198	0.023	-8.626	0.000	-0.243	-0.153
pmi_avg	0.003	0.001	5.488	0.000	0.002	0.004	-0.001	0.001	-1.038	0.299	-0.004	0.001
freq_constr	0.008	0.002	3.090	0.002	0.003	0.012	0.002	0.003	0.725	0.469	-0.004	0.008
OPT												
Intercept	0.534	0.012	46.370	0.000	0.511	0.556	0.516	0.018	29.281	0.000	0.481	0.551
S[T.same]	-0.002	0.007	-0.338	0.736	-0.016	0.011	0.039	0.008	4.822	0.000	0.023	0.055
dist_from_prev_turn	-0.014	0.001	-22.485	0.000	-0.016	-0.013	-0.008	0.001	-7.799	0.000	-0.010	-0.006
constr_overlap	0.039	0.017	2.258	0.024	0.005	0.072	0.035	0.021	1.716	0.086	-0.005	0.076
vocab_overlap	-0.041	0.014	-2.928	0.003	-0.068	-0.013	-0.034	0.020	-1.704	0.088	-0.073	0.005
pmi_avg	0.000	0.001	0.065	0.949	-0.001	0.001	-0.000	0.001	-0.217	0.828	-0.003	0.002
freq_constr	0.001	0.003	0.341	0.733	-0.005	0.006	-0.000	0.003	-0.119	0.905	-0.007	0.006
DGPT												
Intercept	0.524	0.071	7.365	0.000	0.384	0.663	0.482	0.041	11.645	0.000	0.401	0.563
S[T.same]	-0.024	0.036	-0.647	0.518	-0.095	0.048	0.061	0.020	3.071	0.002	0.022	0.100
dist_from_prev_turn	0.012	0.004	2.871	0.004	0.004	0.020	0.007	0.003	2.704	0.007	0.002	0.012
constr_overlap	0.018	0.083	0.215	0.829	-0.145	0.181	-0.086	0.052	-1.656	0.098	-0.187	0.016
vocab_overlap	-0.023	0.085	-0.275	0.784	-0.191	0.144	0.095	0.047	2.018	0.044	0.003	0.188
pmi_avg	0.001	0.007	0.174	0.861	-0.013	0.016	0.007	0.003	2.116	0.034	0.001	0.014
freq_constr	-0.011	0.009	-1.218	0.223	-0.028	0.007	-0.017	0.008	-2.032	0.042	-0.033	-0.001
TUNED												
GPT2												
Intercept	0.463	0.017	26.730	0.000	0.429	0.497	0.436	0.015	29.226	0.000	0.406	0.465
S[T.same]	-0.033	0.009	-3.510	0.000	-0.051	-0.014	-0.013	0.008	-1.590	0.112	-0.030	0.003
dist_from_prev_turn	-0.009	0.001	-9.436	0.000	-0.011	-0.007	0.001	0.001	1.416	0.157	-0.001	0.003
constr_overlap	0.277	0.020	14.149	0.000	0.239	0.315	0.183	0.024	7.511	0.000	0.135	0.230
vocab_overlap	-0.308	0.019	-15.922	0.000	-0.346	-0.270	-0.202	0.022	-9.113	0.000	-0.245	-0.159
pmi_avg	0.001	0.001	1.018	0.309	-0.001	0.003	-0.001	0.001	-0.753	0.451	-0.003	0.001
freq_constr	0.007	0.004	1.729	0.084	-0.001	0.015	0.006	0.003	1.963	0.050	0.000	0.013
OPT												
Intercept	0.528	0.013	39.783	0.000	0.502	0.554	0.494	0.017	29.608	0.000	0.461	0.526
S[T.same]	-0.004	0.008	-0.499	0.618	-0.020	0.012	0.002	0.009	0.234	0.815	-0.015	0.019
dist_from_prev_turn	-0.004	0.001	-5.376	0.000	-0.005	-0.002	0.001	0.001	1.536	0.124	-0.000	0.003
constr_overlap	0.021	0.019	1.129	0.259	-0.016	0.058	-0.022	0.021	-1.026	0.305	-0.063	0.020
vocab_overlap	-0.039	0.016	-2.508	0.012	-0.070	-0.009	0.012	0.021	0.575	0.566	-0.029	0.052
pmi_avg	-0.001	0.001	-1.377	0.168	-0.002	0.000	-0.001	0.001	-0.568	0.570	-0.003	0.002
freq_constr	0.001	0.003	0.195	0.845	-0.006	0.007	0.004	0.003	1.108	0.268	-0.003	0.011
DGPT												
Intercept	0.472	0.013	35.438	0.000	0.446	0.498	0.445	0.017	25.447	0.000	0.411	0.479
S[T.same]	0.003	0.008	0.401	0.689	-0.012	0.019	-0.006	0.010	-0.637	0.524	-0.026	0.013
dist_from_prev_turn	0.001	0.001	1.285	0.199	-0.001	0.003	0.005	0.001	4.126	0.000	0.002	0.007
constr_overlap	0.022	0.021	1.039	0.299	-0.019	0.063	0.064	0.028	2.305	0.021	0.010	0.118
vocab_overlap	-0.046	0.017	-2.748	0.006	-0.079	-0.013	-0.055	0.025	-2.225	0.026	-0.104	-0.007
pmi_avg	0.001	0.001	1.169	0.242	-0.001	0.002	-0.002	0.001	-1.264	0.206	-0.004	0.001
freq_constr	0.001	0.003	0.360	0.719	-0.005	0.008	0.011	0.004	2.716	0.007	0.003	0.019

Table 13: Attribution effects over human utterances. *S* indicates speaker, *type* indicates model type (base or fine-tuned), *diff* indicates whether the two utterances come from different speakers, or between-speaker repetition, and *same* indicates whether the two utterances come from the same speakers, or within-speaker repetition. constr_overlap indicates CO, vocab_overlap indicates VO, PMI indicates specificity, and freq, frequency of shared constructions.

	Switchboard						*Map Task*					
	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]	Coef.	Std.	z	$P > \|z\|$	[0.025	0.975]
GPT2												
Intercept	0.552	0.000	2122.312	0.000	0.551	0.552	0.554	0.001	878.909	0.000	0.552	0.555
m_type[T.tuned]	-0.009	0.000	-42.336	0.000	-0.009	-0.008	-0.029	0.001	-53.563	0.000	-0.030	-0.028
dist	0.000	0.000	16.487	0.000	0.000	0.001	-0.004	0.000	-48.544	0.000	-0.004	-0.004
dist:m_type[T.tuned]	-0.001	0.000	-13.645	0.000	-0.001	-0.000	0.004	0.000	37.490	0.000	0.004	0.004
OPT												
Intercept	0.502	0.000	1599.293	0.000	0.502	0.503	0.519	0.001	730.825	0.000	0.518	0.520
m_type[T.tuned]	-0.003	0.000	-11.565	0.000	-0.003	-0.002	-0.020	0.001	-26.957	0.000	-0.021	-0.018
dist	-0.001	0.000	-37.286	0.000	-0.001	-0.001	-0.003	0.000	-31.255	0.000	-0.004	-0.003
dist:m_type[T.tuned]	0.001	0.000	26.777	0.000	0.001	0.002	0.004	0.000	26.279	0.000	0.004	0.004
DGPT												
Intercept	0.488	0.000	1079.600	0.000	0.488	0.489	0.501	0.001	550.576	0.000	0.499	0.503
m_type[T.tuned]	0.017	0.000	42.653	0.000	0.017	0.018	-0.003	0.001	-2.734	0.006	-0.005	-0.001
dist	-0.003	0.000	-37.818	0.000	-0.003	-0.002	-0.004	0.000	-29.147	0.000	-0.004	-0.004
dist:m_type[T.tuned]	0.002	0.000	22.719	0.000	0.002	0.002	0.005	0.000	25.426	0.000	0.005	0.005

Table 14: Attribution effects over speaker labels. *m_type* indicates model: either base or tuned. *dist* indicates distance between context and target utterances.