Do dialogue representations align with perception? An empirical study

Sarenne Wallbridge and **Peter Bell** and **Catherine Lai**
Center for Speech Technology Research
University of Edinburgh, United Kingdom
{s1301730, peter.bell, c.lai}@ed.ac.uk

Abstract

There has been a surge of interest regarding the alignment of large-scale language models with human language comprehension behaviour. The majority of this research investigates comprehension behaviours from reading isolated, written sentences. We propose studying the perception of dialogue, focusing on an intrinsic form of language use: spoken conversations. Using the task of predicting upcoming dialogue turns, we ask whether turn plausibility scores produced by state-of-the-art language models correlate with human judgements. We find a strong correlation for some but not all models: masked language models produce stronger correlations than autoregressive models. In doing so, we quantify human performance on the response selection task for open-domain spoken conversation. To the best of our knowledge, this is the first such quantification. We find that response selection performance can be used as a coarse proxy for the strength of correlation with human judgements, however humans and models make different response selection mistakes. The model which produces the strongest correlation also outperforms human response selection performance. Through ablation studies, we show that pre-trained language models provide a useful basis for turn representations; however, fine-grained contextualisation, inclusion of dialogue structure information, and fine-tuning towards response selection all boost response selection accuracy by over 30 absolute points.

1 Introduction

Human language processing has intrigued researchers from numerous fields for centuries (Herder, 1772). The relatively recent convergence of philosophy, psycholinguistics, and information theory has produced valuable theories of language production and comprehension by framing people as predictive processors (Christiansen and Chater, 2016; Levy, 2008; Hale, 2001). In particular, Surprisal Theory posits that comprehension effort is directly related to the predictability of a linguistic unit in its context (i.e., its surprisal) (Hale, 2001), and the Smooth Signal Redundancy hypothesis (Aylett and Turk, 2004) (and related theories including Uniform Information Density (Fenk and Fenk, 1980; Levy and Jaeger, 2006) and Entropy Rate Constancy (Genzel and Charniak, 2002)) demonstrates that we produce linguistic signals that tend towards uniform distributions of information, or constant predictability.

Such theories rely on the conditional probability of observing a linguistic unit in a particular context. Traditionally, estimates of these probabilities were obtained through statistical models such as n-grams or PCFGs (Smith and Levy, 2013; Hale, 2001). By comparison, large scale language models (LLMs) like Transformers allow much greater degrees of context to be integrated into probability estimates (Vaswani et al., 2017). LLMs have enabled massive progress across the continuum of NLP tasks (Wang et al., 2019; Hu et al., 2020). As such, a recent field of research has investigated the application of LLMs for producing estimates of human surprisal. Language model quality (measured by perplexity) generally correlates with ability to predict aspects of human perception behaviour– psychometric predictive power (Wilcox et al., 2020; Frank et al., 2015; Levy, 2011; Goodkind and Bicknell, 2018).

The vast majority of psychometric predictive power studies are based on monologue-like data and involve perception of isolated sentences. However, the most natural, innate form of language-use is communicative interaction; we learn to hold conversations with very little direct instruction where as tasks related to the production or comprehension of monological data such as reading, writing, or presenting require years of conscious effort to master. Given the additional modelling capacity of current LMs, there has been growing interest in using them to apply classical language processing theories to more natural forms of language

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2696–2713
May 2-6, 2023

like dialogue (Vega and Ward, 2009; Doyle and Frank, 2015b,a; Giulianelli and Fernández, 2021; Giulianelli et al., 2021). Results from these works do not align neatly with findings about perception of isolated sentences. This raises interesting questions about differences between information transmission strategies employed in monological and conversational settings, and suggests that the the alignment of language models and human perceptual behaviour for monological and isolated linguistic data may not extend to dialogue.

At the same time, the field of dialogue modelling in NLP has grown substantially. Much of the recent work in this field is geared towards leveraging the structural differences between monologue and dialogue data (Lowe et al., 2015; Wu and Xiong, 2020). Although this has increased performance on dialogue modelling benchmark tasks such as response selection and dialogue generation, it is unclear whether increased performance on these tasks results in representations that align with human perception (Wolf et al., 2019; Liu et al., 2020).

To understand whether representations of dialogue align with human perception of dialogue, we ask how well language model outputs correlate with human perception of dialogue turn acceptability. To do so, we build on the novel perceptual task of rating dialogue turn plausibility proposed in our previous work (Wallbridge et al., 2022). First, we study the nature of human expectations in dialogue. By recasting the rating task as a discriminative one, we find evidence that, similar to linguistic acceptability judgements, human judgements of dialogue turn plausibility are probabilistic rather than deterministic (Lau et al., 2017). Next, we establish the psychometric predictive power of different context-dependent text representations with respect to this task by asking whether such representations are predictive of dialogue plausibility judgements. In previous work, we found a statistically significant but weak correlation between surprisal estimates from a generative dialogue-based language model and human plausibility scores for (context, response) pairs (Wallbridge et al., 2022). We use this paradigm here to investigate the predictive power of other language modelling styles. In doing so, we find a strong correlation between (context, response) scores from a masked LLM fine-tuned towards response-selection and human judgements. Interestingly, this model also achieves "superhuman" response selection performance in the sense that it obtains a higher accuracy than participants. This finding, however, motivated our analysis into response-selection dialogue models and under what conditions they align with human perception.

2 Language models and human language perception

Multiple large-scale comparisons between the predictive power of different families of a language models provide strong evidence for a relationship between a language model's quality and its predictive power for human comprehension behaviour. Behaviours include self-paced reading times and gaze duration (Wilcox et al., 2020; Goodkind and Bicknell, 2018; Meister et al., 2021), grammatical acceptability judgements (Richter and Chaves, 2020; Lau et al., 2017; Warstadt et al., 2019; Meister et al., 2021), and brain response data (Frank et al., 2015; Schrimpf et al., 2021). Other works provide similar evidence for a close relationship between surprisal and human language perception through improved language generation under cognitive-inspired constraints (Wei et al., 2021).

Although the majority of these studies are based on a constrained definition of perception using isolated sentences, they already reflect evidence that different comprehension tasks rely on different types of linguistic expectations. Wilcox et al. (2020) finds a dissociation between the syntactic generalisation capability of LMs which is heavily dependent on model architecture, and LM ability to predict human reading times. Similar results were found by Meister et al. (2021) – BERT is highly predictive of acceptability judgements but was described as "remarkably poor" for estimating reading times.

These findings suggest that the alignment of language model output and comprehension behaviours is unlikely to generalise to the perception of dialogue.

2.1 Perception of dialogue acceptability

The majority of psycholinguistic language production theories have been developed based on monological data analysed from a generative linguistics perspective. However, dialogue perception differs from the comprehension of monological data in a number of fundamental aspects (Pickering and Garrod, 2004). Dialogue has been described as a game where participants only "win" if both un-

derstand the dialogue (Lewis, 1969), or as a joint process where interlocutors collaborate to build common ground (Clark and Wilkes-Gibbs, 1986). Context plays a much larger role in dialogue comprehension (Nieuwland and Berkum, 2006). As a concrete example, Fernandez and Ginzburg (2002) find that more than 11% of dialogue turns in the British National Corpus are non-sentential in isolation. Pragmatic context is particularly important. Conversational acts such as backchanneling or the use of adjacency and coordinate pairs tend to have low lexical information density, but are crucial for turn taking and grounding (Clark and Wilkes-Gibbs, 1986; Kawahara et al., 2016).

The challenges of integrating such context into language models has been highlighted by a handful of works investigating information transmission strategies in communicative contexts. Giulianelli et al. (2021) find that relevant contextual units in dialogue should be topically and referentially coherent, and that defining these units depends on the domain of discourse. Doyle and Frank (2015b) find that common ground is a crucial aspect to account for when modelling information distribution in Twitter dialogues. Vega and Ward (2009) present evidence of uniform information density in spoken dialogues, but note the importance of accounting for non-lexical information as additional context.

To model information strategies in communication, language models should align with human perception of dialogue acceptability. In this work, we propose using the task of dialogue utterance plausibility to quantify human perception of dialogue acceptability, and examine whether language models align with this aspect of perception.

2.2 Language models and dialogue

Similarly to psycholinguistics, the vast majority of NLP models have been developed based on monological data. However, interest in modelling interaction and dialogue is booming, fuelled by the development of commercial dialogue systems like Apple Siri[1] and Amazon Alexa[2]. Work related to dialogue perception is also gaining interest, focused on tasks such as dialogue act classification and dialogue coherence estimation (Shriberg et al., 1998; Tran, 2020; Cervone et al., 2018). However, it is unclear whether current dialogue-based LLMs produce perceptually meaningful representations.

[1]https://www.apple.com/siri

[2]https://developer.amazon.com/alexa

Response selection Regardless of the downstream application, response selection has become a pervasive task in dialogue modelling. A major reason for this is that it doesn't require annotated labels. This self-supervised task is used throughout dialogue modelling as both an evaluation metric and a training signal. Given some degree of dialogue history as an anchor, response selection involves selecting the upcoming response from a set of potential turns. This is a direct extension of quintessential next-sentence-prediction task which makes a strong assumption that there is a single correct upcoming turn (Devlin et al., 2019). In this work, we explore the extent to which this assumption holds when making predictions in dialogue.

Response selection models, also known as retrieval-based dialogue models, can be broadly separated into two classes: bi-encoder models that learn independent representations of responses and their respective contexts, and cross-encoders which encode the context and response as a single representation before scoring them (Henderson et al., 2020; Zhou et al., 2016; Wu et al., 2020, 2017a). The latter often involve tuning a pre-trained language model using dialogue-specific objectives (Wolf et al., 2019; Han et al., 2021; Xu et al., 2020). Such methods aim to make use of the representation capacity of pre-trained language models, while also leveraging important and unique structural features of dialogue data.

By encoding contexts and responses in isolation, bi-encoders achieve cheaper training and inference as representations can be cached. However, they enforce a strong independence assumption between contexts and responses. The separate encoders can provide a weak notion of position (i.e., the same lexical content may be encoded differently if it is a response or a piece of context), however cross-encoding captures much richer interactions.

Given the importance of response selection in dialogue modelling, we explore here how well this task aligns with the perception of spoken dialogue acceptability. We also perform ablation studies to better understand which aspects of model architecture are important for response selection.

3 Experiments

3.1 Psychometric predictive power in spoken dialogue transcripts

Data To investigate the relationship between representations of dialogue and human perception, we

make use of a dataset collected in our previous work (Wallbridge et al., 2022) which comprises of 100 (context, response) pairs extracted from the Switchboard Telephone Corpus, and their associated median plausibility scores. Switchboard is a corpus of over 2,400 spontaneous chit-chat style telephone conversations between 542 participants covering 70 topics. It includes manual transcriptions and turn segmentations (Godfrey et al., 1992). Each of the 100 stimuli consists of a set of speaker turns as context $c = [c_1, ..., c_k]$ and an upcoming response r where r is either the true upcoming turn in the dialogue, or a turn sampled from the corpus. Scores were collected by asking participants to rate how plausible r is in the context of c on a scale of 1-5 ("Very Unlikely" – "Very Likely"). Each dialogue context c was presented in 10 pairs: with the true upcoming turn, and 9 negative samples.

In this previous work, we found that people can make relatively accurate discriminative judgements regarding the true upcoming turn in a dialogue: using the mean score per stimuli as a proxy for turn selection, participants obtained an accuracy of 70% (Wallbridge et al., 2022). Plausibility scores were also compared to surprisal estimates from TurnGPT (Ekstedt and Skantze, 2020). Although the relationship between plausibility scores and TurnGPT surprisal estimates was found to be statistically significant, the correlation between them was weak.

Here, we investigate the effect of language modeling styles on psychometric predictive power, using both generative and retrieval-based language models. To do so, we train a range of language models towards the response selection task using the Switchboard dataset. We split the corpus by conversation into training, validation, and test sets (80%, 10%, 10%). Transcripts are all lower-cased and speech-based annotations such as pronunciation markers and speech events are removed.

Models:

Our models are all implemented in PyTorch (Paszke et al., 2019) and pre-trained language models are obtained from the Transformers library (Wolf et al., 2020).

To test generative model capabilities, we employ TurnGPT. This architecture extends the standard GPT-2 model of Radford et al. (2019) to dyadic interaction by fine-tuning a pretrained GPT-2 model with additional speaker embeddings and speaker tokens to encode conversational structure (Ekstedt and Skantze, 2020). For our experiments, we fine-tune the pretrained GPT-2 model from the Transformers library (Wolf et al., 2020) on our training portion of Switchboard using the augmented TurnGPT cross-entropy loss.[3]

Table 1: Correlations between human plausibility scores and LM response scores. Scores are based on turn-level surprisal estimates for TurnGPT, and the joint (context, response) response selection score for BERT-FP.

Model	**Metric**	ρ	p-value
TurnGPT	S_{total}	-0.302	0.002
	S_{mean}	-0.392	<0.001
	$S_{relative}$	-0.395	<0.001
	S_{max}	-0.463	<0.001
	S_{var}	-0.346	0.001
BERT-FP	RS Score	0.637	<0.001

As has been done in previous studies of psychometric predictive power, we obtain estimates of response surprisal from TurnGPT using various aggregates of token-level surprisal (Lau et al., 2017; Meister et al., 2021; Wallbridge et al., 2022). These include both global and local definitions of surprisal. Global metrics include $S_{total}, S_{mean}, S_{relative}$ while local metrics include S_{max}, S_{var}. See Appendix A for further details.

To test retrieval-based methods, we use the BERT-FP architecture (Han et al., 2021). This cross-encoder obtains state-of-the-art response selection for written conversational benchmarks including the Ubuntu (Lowe et al., 2015), E-commerce (Zhang et al., 2018), and Douban (Wu et al., 2017b) datasets. The model encodes the joint (context, response) pair using pre-trained BERT (Devlin et al., 2019). We use `bert-base-uncased` from the Transformers library. The resulting BERT $[CLS]$ token is fed through a single-layer classifier to produce a response selection score – the relevance of the response given the context. We fine-tune BERT-FP towards response selection on our training portion of Switchboard using binary cross entropy loss.[4] See Appendix A for additional details of model training procedures. Rather than computing response surprisal from word-level model output, we obtain response plausibility scores directly from the BERT-FP model.

[3]We implement our model using the TurnGPT Github repository `https://github.com/ErikEkstedt/TurnGPT`

[4]We implement our model based on the BERT-FP Github repository `https://github.com/hanjanghoon/BERT_FP`

Table 2: Discriminative performance of people, TurnGPT, and BERT-FP.

Model	Metric	F1	R10@1
Human	Mean score	0.783	0.7
	Median score	0.800	0.8
TurnGPT	Max surprisal	0.435	0.3
	Mean surprisal	0.538	0.4
BERT-FP	Score	0.889	0.9

Input to both models consists of a set of speaker turns context $c = [c_1, ..., c_k]$ and response turn r where context turns are separated by a special $[eos]$ token to denote a change in speaker. As per the original BERT-FP implementation, we use a fixed number of context turns ($k = 3$) (Han et al., 2021).

3.1.1 Correlation with plausibility perception

Table 1 show the correlations between model response scores and human plausibility judgements. Similar to previous works (Lau et al., 2017; Meister et al., 2021) and our own (Wallbridge et al., 2022), we find a relatively weak correlation between TurnGPT surprisal metrics and plausibility scores, as well as variation in correlation strength between surprisal metrics.

Table 1 also shows the language model to have a larger effect than surprisal metric. We find a much stronger correlation between turn-plausibility scores from humans and scores predicted by our retrieval-based cross-encoder, BERT-FP, than for TurnGPT surprisal estimates. This indicates that the response selection task used to fine-tune BERT generates a latent space that matches some aspects of dialogue acceptability perception.

3.1.2 Discriminative performance

The plausibility scoring task can be reframed as response selection by taking the mean score for each (context, response) pair as the selection criterion, using each of the scoring metrics. Table 2 provides the response selection performance of our models, as well as a benchmark of human consensus performance for this task. We report both the standard response selection accuracy metric R_{10}@1 as well as F1 score which denotes the highest linear classification accuracy that can be achieved across all (context, response) pairs based on a given metric.

The response selection paradigm is used widely throughout dialogue modelling. But to our knowledge, this is one of the first investigations of how well humans can judge upcoming responses. People can execute the response selection task relatively well using a small amount of context, however their performance is not perfect in terms of R_{10}@1 accuracy.

Interestingly, we find that BERT-FP outperforms people at this task, achieving higher accuracy and F1 score. This "superhuman" performance is further evidence that the response-selection objective deviates from human perception of dialogue. We provide a some qualitative observations of deviations between model and human judgements below (see Error Analysis).

Although TurnGPT showed weak to moderate correlation with human judgements scores, it performs very poorly in the response selection task framing. Different surprisal definitions had relatively large impacts on the correlation strengths reported in Table 1, but these were not reflected in the F1 scores. TurnGPT is trained using cross-entropy loss over next-token prediction (autoregressive pre-training), while BERT-FP is based on the masked language modelling paradigm – these differences in underlying pre-training could help explain performance deviations and would be interesting to explore in future work. We discuss these findings further in section 4.

Error analysis: cross-encoders and humans Neither BERT-FP nor human raters achieve perfect performance on the response selection framing. However, their mistakes are different. The cross-encoder succeeds on all three questions where humans did not rate the true response highest, and vice versa for the question misranked by the model.

For all three stimuli that people misranked, the true response obtained the second highest score and had a mean plausibility score greater than 3. In two of these cases, the highest-scoring response was generic response ("mhm", "yeah") which is plausible in a wide range of contexts. In the final case, the highest scoring response was a question that shifted the topic of conversation. We noted this pattern in other stimuli: responses that initiated topic shifts – often questions – were often rated as somewhat likely by people but obtained low BERT-FP scores.

In comparison, for the stimuli misranked by BERT-FP, the true response obtained the 4th highest score. All four of the misranked stimuli included two or more context turns, suggesting that the amount of context is not necessarily a driving

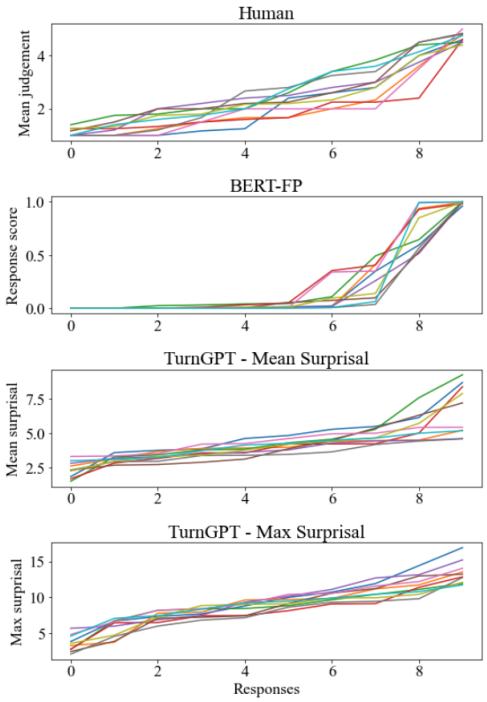

Figure 1: Sorted score sets per response selection stimuli for different models

factor in mistakes for either the model or people.

We also examine the rank correlation (Spearman r) between human ratings and model scores on a per-stimuli basis. Coefficients range from 0.33 to 0.93 ($\mu = 0.65 \pm 0.18$). The lowest correlations are obtained for stimuli where the model only scores a single response highly while the rest obtain scores close to 0. The distributions of BERT-FP scores across stimuli are more peaked than ratings produced by people (see Figure 1). Surprisal metrics from TurnGPT show score distributions more similar to human judgements, further highlighting differences between these models.

3.2 Model ablation for response selection

Generative (TurnGPT) and retrieval-based (BERT-FP) models display significantly different propensities for psychometric predictive power and response selection performance. We use ablation studies to better understand which aspects of the retrieval-based model architecture are valuable for response selection. In particular, we consider

1. Independence of context and response encoding: bi- versus cross-encoders

2. Inclusion of dialogue structure: $[eos]$ tokens

3. Importance of fine-tuning.

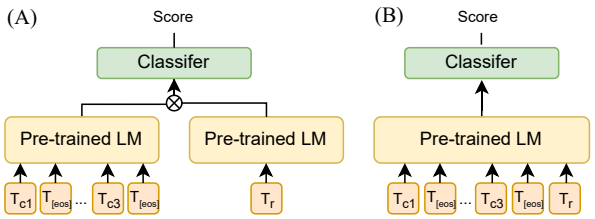

Figure 2: Architectures for (A) bi- and (B) cross-encoders. Input is shown in SBERT format; T denotes tokenized text. For BERT-based models, we prepend all input sequences with a $[CLS]$ token. Cross-encoder input also contains a $[SEP]$ token before and after T_r.

Models: We compare cross- and bi-encoder networks, depicted in Figure 2. The cross-encoder architecture is based on BERT-FP (Han et al., 2021) which encodes the joint context and response pair then feeds the resulting sequence representation through a classification network. We experiment with using both a single linear-layer classifier, or three fully-connected ReLU layers with dropout and a final sigmoid layer.

The bi-encoder style model follows the architecture presented in (Henderson et al., 2020). Each context and response is encoded in isolation using a shared, pre-trained language model. Context and response representations are then concatenated and fed through the same binary classifiers as the cross-encoder architectures.

We test scores based on both BERT $[CLS]$ tokens and SBERT sequence representations (Devlin et al., 2019; Reimers and Gurevych, 2019). SBERT representations are produced by mean pooling fine-tuned token representations from BERT. We use `bert-base-uncased` for BERT-based models and `all-MiniLM-L6-v2` for SBERT, both obtained from the Transformers library.

All models are trained towards the response selection task with a binary cross entropy loss on our training portion of Switchboard. We train models with the underlying LM in frozen and unfrozen scenarios.

3.2.1 Results

We present a simple baseline for response selection: minimizing the Euclidean distance between pre-trained SBERT representations of (context, response) pairs. This outperforms many of the ablated models discussed hereafter (see Table 3).

Independence of context and responses Although bi-encoders achieve cheaper training and inference by encoding contexts and responses in

Table 3: Response selection accuracy $R_{10}@1$ for models based on SBERT. The SBERT baseline (response selection based on euclidean distances between pre-trained SBERT representations of context and responses) achieves an $R_{10}@1$ of 0.480

Encoding Strategy	$R_{10}@1$ **accuracy**			
	Linear Classifier	**+ unfreeze**	**Non-linear Classifier**	**+ unfreeze**
Bi-encoder	0.102	0.105	0.455	0.507
Cross-encoder	0.236	0.465	0.307	0.472
Cross-encoder + [eos]	0.216	0.611	0.306	0.612

Table 4: Response selection accuracy $R_{10}@1$ for models based on BERT $[CLS]$

Encoding Strategy	$R_{10}@1$ **accuracy**			
	Linear Classifier	**+ unfreeze**	**Non-linear Classifier**	**+ unfreeze**
Bi-encoder	0.103	0.104	0.212	0.316
Cross-encoder + [eos]	0.324	0.675	0.346	0.673

isolation, they lack the rich contextualisation of cross-encoders. Previous work reports better response selection performance from cross-encoders for written dialogue (Urbanek et al., 2019). However, recent spoken dialogue research (Fuscone et al., 2020) as well as our own (Wallbridge et al., 2021) has shown that both people and models can make predictions about upcoming conversational turns based on small amounts of context. We therefore compare bi- and cross- encoding strategies to understand the value of such contextualisation in the domain of spoken dialogue.

Tables 3 and 4 show that bi-encoders require a classifier with some degree of non-linearity. Regardless of whether the underlying language model is frozen or not, bi-encoders with a linear classifier do not improve on chance performance. In comparison, the cross-encoder improves on chance even with a frozen language model and a linear classifier. This is the case for both SBERT- and BERT-based models, indicating that masked language model pre-training extracts information that is of some value for predicting upcoming dialogue turns.

When non-linearity is introduced in the classifier, bi-encoding outperforms cross-encoding by quite a margin on top of a frozen SBERT model. This may reflect similarities between response selection and the SBERT training objective of contrastive sentence similarity (Reimers and Gurevych, 2019). We verify that this result was not caused by the special $[eos]$ turn delimeter that wasn't seen during cross-encoder pre-training by training the same model without this token added to the input.

Regardless of the underlying LM, the performance gap between unfrozen bi- and cross-encoders indicates that the lack of contextualisation may cap bi-encoder performance. This difference confirms that fine-grained contextualisation is important for dialogue representations.

Dialogue structure: the importance of *[eos]* The effect of special tokens like $[eos]$ on the behaviour of large LMs is an active research area. Some works have found such tokens to have detrimental effects on the generalisability of generative models (Newman et al., 2020) but many recent models for dialogue representation rely on these tokens to capture structural characteristics of dialogue (Gu et al., 2020; Wolf et al., 2019).

Comparing the cross-encoder performance with and without $[eos]$ in Table 3 shows that explicit turn information is extremely valuable for response selection performance.

Fine-tuning The importance of pre-trained language models in NLP can't be understated, however spoken dialogue is fundamentally different to the monological text used for pre-training. Similar to previous work investigating the importance of fine-tuning with dialogue data, we find that pre-trained BERT representations do contain some useful information for response selection, but that fine-tuning towards the target task is still important (Noble and Maraev, 2021). In addition, we find that fine-tuning is much more beneficial when dialogue structure information is provided explicitly via the $[eos]$ token.

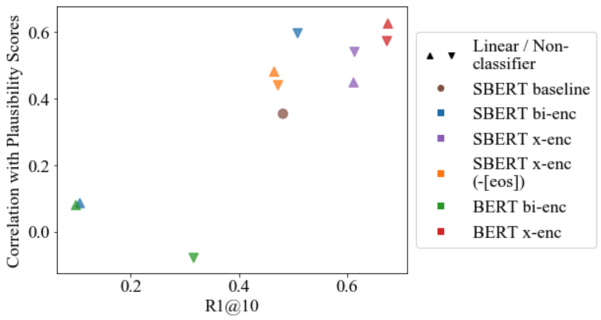

Figure 3: Response selection accuracy vs. psychometric predictive power

3.3 Is response selection performance a proxy for Psychometric Predictive Power?

Given the range of response selection results from our ablations, we consider the relationship between response selection performance and correlation with human plausibility judgements (PPP). We find a strong correlation between the two: $\rho = 0.84$. This relationship is marginally stronger if a larger range of response selection accuracies are considered; comparing response selection performance measured as $\sum_{k\in\{1,2,5\}}[R10@k]$ and PPP produces a correlation of $\rho = 0.86$. To ensure this correlation was not dominated by human response selection ability, we also computed this correlation across only negative responses and found a similarly strong correlation of model response selection performance and human scores ($\rho = 0.82$).

This indicates that response selection performance can be used as a coarse proxy for how well a model may correlate with human turn plausibility judgements. However, response selection performance patterns (e.g. SBERT cross-encoding outperforming bi-encoding) are not reflected in this relationship. Although the modelling capacity of the classifier (i.e., linear versus non-linear) did not have much effect on the response selection performance of cross-encoders (see pairs of points grouped around $R_{10}@1$ of $0.47, 0.61, 0.67$ in Figure 3), the impact on PPP is larger.

4 Discussion

Using the task of upcoming dialogue turn prediction, we presented a benchmark for human performance on the widely-used task of response selection: people are able to do this task, but not perfectly. To the best of our knowledge, this is the first investigation into the perceptual validity of response selection for spoken dialogue transcripts. Urbanek et al. (2019) measure human response selection using stimuli from written dialogues grounded in a virtual world, using sets of 20 responses. However, they note that the large number of negative samples makes obtaining reliable judgements from participants difficult. The rating paradigm proposed here avoids these issues, provides information about perceptual certainty, and is independent of the choice of negative samples.

Human performance showed that a fundamental assumption of response selection – that a single "correct" upcoming turn can be predicted from previous context – does not reflect the reality of perception with respect to dialogue acceptability. This result supports findings in cognitive science that linguistic perception of acceptability is intrinsically probabilistic (Lau et al., 2017; Chater et al., 2006).

We then used this quantification of dialogue acceptability perception to establish the psychometric predictive power of different language model families, finding that response selection performance can be used as a coarse proxy for alignment with this aspect of perception, but only to a certain extent. Strong correlation of BERT-FP scores and human plausibility ratings suggests that, as a training objective, the discriminative response selection task does align with some aspects of dialogue acceptability perception. However, this model achieved "superhuman" response selection performance. Qualitative differences in response scoring detailed in Section: Error Analysis also indicate that this disciminative training paradigm doesn't capture the open-ended nature of dialogue acceptability perception. On the other hand, TurnGPT produced a moderate correlation with plausibility judgements, but performed poorly at response selection. (Urbanek et al., 2019; Wu and Xiong, 2020) report similar findings regarding the response selection performance of generative models compared to retrieval-based models. We conducted preliminary analysis of quantitative and qualitative differences between autoregressive and masked LMs, however given that language model architecture is known to significantly affect psychometric predictive power measured by other perceptual tasks (Wilcox et al., 2020), we believe this to be a fruitful avenue for future investigation.

Finally, to investigate the strong correlation between the cross-encoder-style BERT-FP scores and perceptual judgements of upcoming dialogue turns, we ablated aspects of the model. In particular, we examined the importance of contextualisation

in dialogue representations, the value of including turn-taking structure, and whether pre-training is beneficial. Our bi-encoder models provided a small amount of dialogue structure information, however, even our best-performing bi-encoder only marginally outperforms our baseline of simply measuring the distance between SBERT embeddings for response selection. This may suggest that this method of encoding dialogue structure is too weak. Cross-encoding, on the other hand, provides much more granular contextualisation. We found cross-encoders to be much less sensitive to the capacity of the classifier than bi-encoders. Our cross-encoder outperformed the SBERT baseline and bi-encoders by a large margin when dialogue structure is explicitly encoded in the input (as $[eos]$ tokens). Inclusion of dialogue structure information and fine-tuning towards response selection boosted response selection accuracy by over 30 absolute percentage points, highlighting the value of turn structure information for response selection models.

5 Conclusions

In this study, we extend the concept of psychometric predictive power for language models to dialogue. We define the perceptual behaviour to be predicted in terms of expectations regarding upcoming dialogue turns, and ask whether different styles of language model are predictive of dialogue turn acceptability judgements from humans.

Using transcripts of spoken dialogue, we find a strong correlation between the two, but not for all models: masked language models produce higher correlations than their autoregressive counterparts. Ablation studies demonstrate that features of cross-encoders which enable fine-grained contextualisation are important for alignment with human acceptability judgements. Framing the turn acceptability rating task as discriminative response selection, we present a benchmark for human performance on this task, one of the first to our knowledge. Comparisons between human and model performance on this widely-used dialogue objective indicate that response selection can be used as a coarse proxy for alignment with perception of dialogue turn acceptability but only to a certain extent. Human perception of dialogue turn acceptability is inherently probabilistic and models make different mistakes compared to humans.

We hope that these findings encourage development of more perceptually-motivated evaluation and training paradigms in dialogue modelling.

6 Limitations

Psychometric predictive power has so far been contained to isolated sentence comprehension studies. This work takes steps to extend the notion of psychometric predictive power to more natural forms of language processing: comprehension of communication. Given that this is an exploratory study, there are many limitations, many of which we consider to be potential directions for future work.

The definition for perception used throughout this study is as generic as possible: predicting upcoming turns. However, there is ample evidence that linguistic processing is task-dependent (Huettig and Guerra, 2019; Huettig et al., 2020). Exploring other quantifications for the perception of dialogue acceptability would be an interesting avenue for future work. Similarly, features of communication such as turn-taking are known to vary across languages and cultures (Skantze, 2021). As such, is it unclear how well the findings presented in this work generalise to other styles and forms of conversation.

All of the turn representations in this work are based on a fixed number of turns as context, however we should also study whether the amount of context affects LM alignment with human judgments. In our previous work, we found the amount and type of context to affect human performance (Wallbridge et al., 2021), while Henderson et al. (2020) show that it affects response selection performance in models.

Previous works have studied surprisal estimates computed from different aggregates of word-level estimates from LMs (Lau et al., 2017; Meister et al., 2021; Wallbridge et al., 2022). These surprisal definitions are useful for testing psycholinguistic theories about information transmission strategies such as Uniform Information Density (Fenk and Fenk, 1980; Levy and Jaeger, 2006) and Entropy Rate Constancy (Genzel and Charniak, 2002), however we found stronger correlations with human judgements by using language models to compute plausibility scores directly. Understanding the gap between these scores and aggregate word-level surprisal estimates could inform more complex definitions of surprisal.

Although response selection is a pervasive training objective in dialogue representation research, there is a growing body of work refining this

task, and learning representations using additional dialogue-centric learning objectives such as detection of turn insertions or deletions (Qiu et al., 2021; Lee et al., 2021). Future work could consider how these training signals affect the psychometric predictive power of models with respect to both response selection and plausibility judgements.

This study has been restricted to the lexical component such dialogues. However, perception and production of linguistic signals are dependent on the modality available for transmission, and what information transmission channels it provides (Rowe, 1999; Alaçam et al., 2020). Perception of spoken dialogues and their transcripts has been shown to differ (Wallbridge et al., 2021, 2022), and findings for uni-modal data don't necessarily generalise to multi-modal settings (Bujok et al., 2022). We leave the extension of a similar analysis on the acoustic speech signal for future work.

References

Özge Alaçam, Xingshan Li, Wolfgang Menzel, and Tobias Staron. 2020. Crossmodal language comprehension—psycholinguistic insights and computational approaches. *Frontiers in neurorobotics*, 14:2.

Matthew P. Aylett and Alice Turk. 2004. The smooth signal redundancy hypothesis: A functional explanation for relationships between redundancy, prosodic prominence, and duration in spontaneous speech. *Language and Speech*, 47 1:31 – 56.

Ronny Bujok, Antje S. Meyer, and Hans Rutger Bosker. 2022. Visible lexical stress cues on the face do not influence audiovisual speech perception. In *Proceedings of Speech Prosody 2022*, pages 259–263.

Alessandra Cervone, Evgeny A. Stepanov, and Giuseppe Riccardi. 2018. Coherence models for dialogue. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association*, pages 1011–1015. ISCA.

Nick Chater, Joshua B. Tenenbaum, and Alan Loddon Yuille. 2006. Probabilistic models of cognition: Conceptual foundations. *Trends in Cognitive Sciences*, 10:287–291.

Morten H Christiansen and Nick Chater. 2016. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences*, 39 62:1–72.

Herbert H. Clark and Deanna Wilkes-Gibbs. 1986. Referring as a collaborative process. *Cognition*, 22 1:1–39.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Gabriel Doyle and Michael Frank. 2015a. Audience size and contextual effects on information density in Twitter conversations. In *Proceedings of the 6th Workshop on Cognitive Modeling and Computational Linguistics*, pages 19–28, Denver, Colorado. Association for Computational Linguistics.

Gabriel Doyle and Michael Frank. 2015b. Shared common ground influences information density in microblog texts. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1587–1596, Denver, Colorado. Association for Computational Linguistics.

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.

August Fenk and Günther Fenk. 1980. [constance and short-term memory - constance in speech information]. *Zeitschrift fur experimentelle und angewandte Psychologie*, 27 3:400–14.

Raquel Fernandez and Jonathan Ginzburg. 2002. Non-sentential utterances in dialogue: A: Corpus-based study. In *Proceedings of the Third SIGdial Workshop on Discourse and Dialogue*, pages 15–26, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. 2015. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Simone Fuscone, Benoît Favre, and Laurent Prévot. 2020. Neural representations of dialogical history for improving upcoming turn acoustic parameters prediction. In *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4203–4207. ISCA.

Dmitriy Genzel and Eugene Charniak. 2002. Entropy rate constancy in text. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 199–206, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Mario Giulianelli and Raquel Fernández. 2021. Analysing human strategies of information transmission as a function of discourse context. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 647–660, Online. Association for Computational Linguistics.

Mario Giulianelli, Arabella Sinclair, and Raquel Fernández. 2021. Is information density uniform in task-oriented dialogues? In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8271–8283, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

John Godfrey, Edward Holliman, and Jane McDaniel. 1992. Switchboard: telephone speech corpus for research and development. In *ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 1, pages 517–520 vol.1. IEEE Computer Society.

Adam Goodkind and Klinton Bicknell. 2018. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pages 10–18, Salt Lake City, Utah. Association for Computational Linguistics.

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 *CIKM '20: The 29th ACM International Conference on Information and Knowledge Management, Virtual Event, Ireland, October 19-23, 2020*, pages 2041–2044. ACM.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Janghoon Han, Taesuk Hong, Byoungjae Kim, Youngjoong Ko, and Jungyun Seo. 2021. Fine-grained post-training for improving retrieval-based dialogue systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1549–1558, Online. Association for Computational Linguistics.

Matthew Henderson, Iñigo Casanueva, Nikola Mrkšić, Pei-Hao Su, Tsung-Hsien Wen, and Ivan Vulić. 2020. ConveRT: Efficient and accurate conversational representations from transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2161–2174, Online. Association for Computational Linguistics.

Johann Gottfried Herder. 1772. *Abhandlung ÃŒber den Ursprung der Sprache ("Essay on the Origin of Language")*, pages 121–124. De Gruyter (2015).

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Falk Huettig and Ernesto Guerra. 2019. Effects of speech rate, preview time of visual context, and participant instructions reveal strong limits on prediction in language processing. *Brain Research*, 1706:196–208.

Falk Huettig, Ernesto Guerra, and Andrea Helo. 2020. Towards understanding the task dependency of embodied language processing: The influence of colour during language-vision interactions. *Journal of Cognition*, 3 1.

Tatsuya Kawahara, Takashi Yamaguchi, Koji Inoue, Katsuya Takanashi, and Nigel G. Ward. 2016. Prediction and generation of backchannel form for attentive listening systems. In *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association*, pages 2890–2894. ISCA.

Jey Han Lau, Alexander Clark, and Shalom Lappin. 2017. Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge. *Cognitive Science*, 41(5):1202–1241.

Hyunjae Lee, Jaewoong Yun, Hyunjin Choi, Seongho Joe, and Youngjune L. Gwon. 2021. Enhancing semantic understanding with self-supervised methods for abstractive dialogue summarization. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 796–800. ISCA.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106 3:1126–1177.

Roger Levy. 2011. Integrating surprisal and uncertain-input models in online sentence comprehension: formal techniques and empirical results. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1055–1065, Portland, Oregon, USA. Association for Computational Linguistics.

Roger Levy and T. Florian Jaeger. 2006. Speakers optimize information density through syntactic reduction. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 849–856. MIT Press.

David Kellogg Lewis. 1969. *Convention: A Philosophical Study*. Cambridge, MA, USA: Wiley-Blackwell.

Qian Liu, Yihong Chen, Bei Chen, Jian-Guang Lou, Zixuan Chen, Bin Zhou, and Dongmei Zhang. 2020. You impress me: Dialogue generation via mutual persona perception. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1417–1427, Online. Association for Computational Linguistics.

Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu dialogue corpus: A large

dataset for research in unstructured multi-turn dialogue systems. In *Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 285–294, Prague, Czech Republic. Association for Computational Linguistics.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. 2021. Revisiting the Uniform Information Density hypothesis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 963–980, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. 2020. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 33–44, Online. Association for Computational Linguistics.

Marius Mosbach, Maksym Andriushchenko, and Dietrich Klakow. 2021. On the stability of fine-tuning BERT: misconceptions, explanations, and strong baselines. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Benjamin Newman, John Hewitt, Percy Liang, and Christopher D. Manning. 2020. The EOS decision and length extrapolation. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 276–291, Online. Association for Computational Linguistics.

Mante S. Nieuwland and Jos J. A. Van Berkum. 2006. When peanuts fall in love: N400 evidence for the power of discourse. *Journal of Cognitive Neuroscience*, 18:1098–1111.

Bill Noble and Vladislav Maraev. 2021. Large-scale text pre-training helps with dialogue act recognition, but not without fine-tuning. In *Proceedings of the 14th International Conference on Computational Semantics (IWCS)*, pages 166–172, Groningen, The Netherlands (online). Association for Computational Linguistics.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 8024–8035.

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

Yao Qiu, Jinchao Zhang, Huiying Ren, and Jie Zhou. 2021. Challenging instances are worth learning: Generating valuable negative samples for response selection training. *ArXiv*, abs/2109.06538.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. Technical report, OpenAI.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Stephanie Richter and Rui Chaves. 2020. Investigating the role of verb frequency in factive and manner-of-speaking islands. In *Proceedings of the 42th Annual Meeting of the Cognitive Science Society*. cognitivesciencesociety.org.

Candy Rowe. 1999. Receiver psychology and the evolution of multicomponent signals. *Animal behaviour*, 58 5(5):921–931.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45).

Elizabeth Shriberg, Andreas Stolcke, Daniel Jurafsky, Noah Coccaro, Marie Meteer, Rebecca Bates, Paul Taylor, Klaus Ries, Rachel Martin, and Carol Van Ess-Dykema. 1998. Can prosody aid the automatic classification of dialog acts in conversational speech? *Language and speech*, 41(3-4):443–492.

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

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

Trang Tran. 2020. *Neural Models for Integrating Prosody in Spoken Language Understanding*. Ph.D. thesis, University of Washington.

Jack Urbanek, Angela Fan, Siddharth Karamcheti, Saachi Jain, Samuel Humeau, Emily Dinan, Tim Rocktäschel, Douwe Kiela, Arthur Szlam, and Jason Weston. 2019. Learning to speak and act in a fantasy text adventure game. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 673–683, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. 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 5998–6008.

Alejandro Vega and Nigel G. Ward. 2009. Looking for entropy rate constancy in spoken dialog. Technical report, University of Texas at El Paso.

Sarenne Wallbridge, Peter Bell, and Catherine Lai. 2021. It's not what you said, it's how you said it: Discriminative perception of speech as a multichannel communication system. In *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, pages 2386–2390. ISCA.

Sarenne Wallbridge, Catherine Lai, and Peter Bell. 2022. Investigating perception of spoken dialogue acceptability through surprisal. In *Interspeech 2022, 23nd Annual Conference of the International Speech Communication Association, Incheon, South Korea*, pages 4506–4510.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Jason Wei, Clara Meister, and Ryan Cotterell. 2021. A cognitive regularizer for language modeling. 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 5191–5202, Online. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger P. Levy. 2020. On the predictive power of neural language models for human real-time comprehension behavior. In *Proceedings of the 42nd Annual Meeting of the Cognitive Science Society*, page 1707–1713.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. Transfertransfo: A transfer learning approach for neural network based conversational agents. *ArXiv*, abs/1901.08149.

Chien-Sheng Wu, Steven C.H. Hoi, Richard Socher, and Caiming Xiong. 2020. TOD-BERT: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929, Online. Association for Computational Linguistics.

Chien-Sheng Wu and Caiming Xiong. 2020. Probing task-oriented dialogue representation from language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5036–5051, Online. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017a. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2017b. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 496–505, Vancouver, Canada. Association for Computational Linguistics.

Ruijian Xu, Chongyang Tao, Daxin Jiang, Xueliang Zhao, Dongyan Zhao, and Rui Yan. 2020. Learning an Effective Context-Response Matching Model with Self-Supervised Tasks for Retrieval-based Dialogues. *ArXiv preprint*, abs/2009.06265.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, Hai Zhao, and Gongshen Liu. 2018. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3740–3752, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. 2016. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, Austin, Texas. Association for Computational Linguistics.

A Language Models

A.1 Fine-tuning language models

We employ a range of language models to explore alignment with dialogue acceptability perception and response selection performance, including autoregressive, masked, and sequence-level models.

TurnGPT TurnGPT is a variant of GPT-2 (Radford et al., 2019), an autoregressive Transformer-based model (Vaswani et al., 2017) with 117M parameters. As reported in the main paper, our implementation is based on the pretrained GPT-2 model from the Transformers library (Wolf et al., 2020). We fine-tune the GPT-2 model following the procedure laid out by (Ekstedt and Skantze, 2020). Use our training portion of Switchboard, we train with the cross-entropy loss of the original GPT-2 model and default parameters. We fine-tune with early stopping, and achieve our best model in terms of validation loss after 2 epochs of training.

BERT-FP We employ BERT-FP to test retrieval-based dialogue models (Han et al., 2021). The architecture includes a simple linear MLP on top of a BERT model (Devlin et al., 2019). The original paper presents a two-stage training strategy for response selection. First, the model is post-trained using the task of utterance relevance classification, then it is fine-tuned towards response selection. To facilitate comparisons across architectures, we report experiments involving fine-tuning with no post-training in the main paper. However, we find that following the full procedure of post-training before fine-tuning increases response selection accuracy $R_{10}@1$ by 3 absolute points ($R_{10}@[1, 2, 5] = 0.705, 0.840, 0.975$).

As reported in the main paper, we implement this model using Pytorch and the pre-trained `bert-base-uncased` BERT model from the Transformers library. The full model has 109,483,777 trainable parameters. Fine-tuning is done with a cross-entropy loss. The response selection training set consists of an even split of positive and negative (context, response) pairs. Negative pairs are generated by sampling a turn from the elsewhere in the training set. To maintain consistency with BERT pre-training, joint (context, response) pairs are presented in the following format:

$$([CLS], c_1, [eos], ..., c_k, [eos], [SEP], r, [SEP]) \quad (1)$$

where the $[eos]$ token represents the end of a speaker's turn.

We follow the fine-tuning procedure from the BERT-FP paper. First, this involves splitting conversations into short (context,response) segments, resulting in 267,562 samples for training the cross- and bi-encoders. Second, fine-tuning is implemented using recent "BERTology" research to mitigate one of the pervasive issues with fine-tuning large, pre-trained LMs: training instability (Mosbach et al., 2021; Merchant et al., 2020). Compared to the BERT fine-tuning approach presented in Devlin et al. (2019), this involves using smaller learning rates with bias correction to avoid vanishing gradients early in training, and increasing the number of epochs. We use a batch size of 16 and make use of the Pytorch AdamW optimizer with an initial learning rate of $1e-5$. With early stopping, out best model in terms of validation loss is achieved at epoch 8.

Bi- and Cross-encoders For our ablation studies, we test a number architectural choices, including bi- versus cross-encoder set-ups, linear versus non-linear classifiers, and sequence representations from the BERT $[CLS]$ token or SBERT representations (`bert-base-uncased` for BERT-based models and `all-MiniLM-L6-v2` for SBERT; both models are both from the Transformers library). These are all build on the same codebase as our BERT-FP experiments. We fine-tune all of these architectures following the procedure described above for BERT-FP. As reported in the main paper, the classifiers consist of either a single linear-layer classifier, or three fully-connected ReLU layers with dropout and a final sigmoid layer.

To match the BERT input format, joint (context, response) pairs are presented in the following format for SBERT models:

$$(c_1, [eos], c_2, [eos], c_3, [eos], r) \quad (2)$$

A.2 TurnGPT surprisal definitions

We compute response scores from TurnGPT using various sequence-level surprisal definitions from previous works (Lau et al., 2017; Meister et al., 2021; Wallbridge et al., 2022). These include global surprisal metrics which account for surprisal at the turn level ($S_{total}, S_{mean}, S_{relative}$), and local netrics which provide more granular information at the token level (S_{max}, S_{var}). For a given (context c, response r) pair, each metric is defined as follows:

$$S_{total}(\mathbf{r}|\mathbf{c}) = \sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{mean}(\mathbf{r}|\mathbf{c}) = \frac{1}{N}\sum_{n=1}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{relative}(\mathbf{r}|\mathbf{c}) = S_{mean}(\mathbf{r}|\mathbf{c}) - S_{mean}(\mathbf{r})$$

$$S_{max}(\mathbf{r}|\mathbf{c}) = \max[S(r_n|\mathbf{r}_{<n}, \mathbf{c})]$$

$$S_{var}(\mathbf{r}|\mathbf{c}) = \frac{1}{N-1}\sum_{n=2}^{N}[S(r_n|\mathbf{r}_{<n}, \mathbf{c}) - S(r_{n-1}|\mathbf{r}_{<n-1}, \mathbf{c})]^2$$

B Error Analysis: Corpus Excerpts

Tables 5, 6, 7, 8 contain examples of the stimuli collected in our previous work (Wallbridge et al., 2022) along with associated scores from both human participants, and the BERT-FP language model. Each stimuli consists of a variable number of conversational turns as context (1-4), and 10 potential upcoming responses (one of which is the true response in the particular conversation). Participants were presented with a context and a single response, and were tasked with rating how plausible the response was given the context.

Tables 5, 6, 7 show human misrankings while Table 8 contains the stimuli misranked by BERT-FP.

Table 5: Example of response selections stimuli ($sw3512\_14$) with human and BERT-FP scores.
Table (A) contains the set of conversational turns provided as context. Table (B) contains the set of 10 responses– these include the True upcoming turn for the given context and the 9 negative samples. Each response is displayed with it's true response label, the median plausibility score it recieved from participants (collected in our previous work (Wallbridge et al., 2022) study), and its BERT-FP response score.

(A) Context		
	Speaker A	that's right
	Speaker B	and then it's just so noisy that you can't visit
	Speaker A	oh i know
	Speaker B	know and normally when i'm eating out i you know with people and i wanna sit and talk i mean

(B) Responses	Human	BERT-FP	Transcript
Negative	4.50	0.57	mhm
True	4.40	0.99	that's the half the fun is the conversation right
Negative	3.83	0.60	dear
Negative	3.40	0.15	true you know and you have to start thinking about is it is it worth spending the money to go see it or shall i just wait
Negative	2.60	0.12	hello
Negative	2.00	0.17	but a lot of the stuff they do really you know evidently is pretty easy but i've just never
Negative	2.00	0.00	was lonely and she needed company for her mother and so she opened a nursing home and initially started with eight ladies
Negative	1.80	0.02	it's uh it's about an eight or nine hour drive really i make it in two days because i i don't push it
Negative	1.75	0.03	yeah when he played Danny Boy it just almost brought tears to your eyes because he can make that flute sing
Negative	1.40	0.01	yeah it's raining out here and i just steam cleaned my carpet today and i really don't wanna let the dog in

Table 6: Example of response selections stimuli ($sw4040\_20$) with human and BERT-FP scores.

(A) Context	Speaker A	say you went back to school
	Speaker B	the wife went back to school
	Speaker A	i see uhuh
	Speaker B	he's been there for almost a year now

(B) Responses	Human	BERT-FP	Transcript
Negative	4.83	0.51	yeah
True	4.50	0.94	um gee i think where where does she go to school
Negative	3.00	0.06	what kind of things have you read
Negative	2.60	0.40	but that that's if
Negative	2.25	0.04	um she was referred to me by a couple of people and she turned out to be wonderful i couldn't have asked for anything better i don't think
Negative	2.20	0.01	it's so much easier to sit there and besides i can be doing other things and still listening to the news
Negative	2.00	0.01	you you say show music like Broadway musical type show music
Negative	2.00	0.02	and you they find everything i mean they find out everything about you they want to know your you know where you live what you do what you know and some of the questions
Negative	1.50	0.07	the starter was Bosch American so
Negative	1.17	0.00	that's there for direct yeah the direct sun beating on it yeah that's right

Table 7: Example of response selections stimuli ($sw3959\_32$) with human and BERT-FP scores.

(A) Context	Speaker A	uhuh oh my
	Speaker B	down in Houston for um several years seven years and then uh my son is a CPA another has a business degree

(B) Responses	Human	BERT-FP	Transcript
Negative	4.80	0.63	before that did he go somewhere else or
True	3.60	0.93	my
Negative	2.33	0.00	that uh lots lots of new uh newsmen were created during that so it will be interesting to see what happens
Negative	2.00	0.00	you know with um the US funding Israel and you had the um Soviet Union funding the Arab countries
Negative	1.67	0.58	um yeah
Negative	1.67	0.02	uh what kind of lawn and garden work do you wind up doing
Negative	1.50	0.08	something that we're looking forward to
Negative	1.25	0.00	painted on T-shirts or sweat shirts at all
Negative	1.00	0.01	i do know of a way to get around the computer generated calls
Negative	1.00	0.00	we try to stay away from those things which might have uh salmonella in them

Table 8: Example of response selections stimuli ($sw2303\_59$) with human and BERT-FP scores.

(A) Context	Speaker A	so they're going to be
	Speaker B	that's another and that's another interesting question should judges be elected or appointed

(B) Responses	Human	BERT-FP	Transcript
True	4.60	0.18	that's true that's true well
Negative	4.00	0.50	hm
Negative	2.80	0.90	idea
Negative	2.60	0.01	anyway what do you think we've gained from the space flights
Negative	2.40	0.49	makes you wish they had uh still had indentured servitude for this sort of thing
Negative	1.25	0.12	mean like those you know twenty thousand dollar toilets
Negative	1.17	0.14	well the the you know those little arms are supposed to twist almost any
Negative	1.00	0.00	yeah it's nothing but woods up here
Negative	1.00	0.00	it's uh it's about an eight or nine hour drive really i make it in two days because i i don't push it
Negative	1.00	0.00	i i i used to exercise at night and i found that you don't come home tired you come home with a new found energy so you