Aligning Backchannel and Dialogue Context Representations via Contrastive LLM Fine-Tuning

Anonymous ACL submission

Abstract

Backchannels (e.g., 'yeah', 'mhm', and 'right') are short, non-interruptive feedback signals whose lexical form and prosody jointly convey pragmatic meaning. While prior computational research has largely focused on predicting backchannel timing, the relationship between lexico-prosodic form and meaning remains underexplored. We propose a two-stage framework: first, fine-tuning large language models on dialogue transcripts to derive rich contextual representations; and second, learning a joint embedding space for dialogue contexts and backchannel realizations. We evaluate alignment with human perception via triadic similarity judgments (prosodic and cross-lexical) and a context–backchannel suitability task. Our results demonstrate that the learned projections substantially improve context-backchannel retrieval compared to previous methods. In addition, they reveal that backchannel form is highly sensitive to extended conversational context and that the learned embeddings align more closely with human judgments than raw WavLM features.

1 Introduction

Conversational feedback refers to short, non-interrupting responses signaling, e.g., attention, understanding, and surprise (Allwood et al., 1992). These responses streamline communication by establishing *common ground* (Clark and Schaefer, 1989; Fusaroli et al., 2017) and replacing explicit metalinguistic exchanges — for instance, substituting full answers to the question "Did you understand?", such as "Yes, continue" or "No, please repeat", with simple tokens like "yeah!" or "sorry?", respectively. Feedback is typically multimodal, involving vocalizations, gaze, and gestures (Bertrand et al., 2007; Truong et al., 2011; Ferré and Renaudier, 2017). Modeling these signals is crucial for building rapport in conversational systems (Axelsson et al., 2022). Vocal instances, such as 'yeah', 'uh-huh' or 'wow!', are commonly termed *backchannels* (Yngve, 1970).

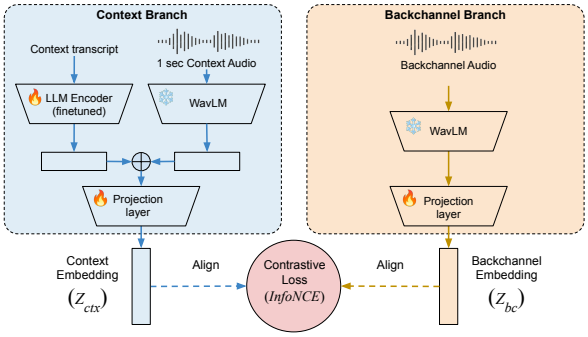

Figure 1: Architecture of the joint context-backchannel model. In this version of the model, both the context transcript and the context audio are passed through their respective encoders (fine-tuned LLM and pre-trained WavLM) before being concatenated and projected to a lower-dimensional space to form a context embedding.

Computational work has prioritized backchannel placement and timing (Heldner et al., 2013; Ruede et al., 2019; Ortega et al., 2020; Ishii et al., 2021) but the link between *form* and *meaning* is often overlooked. Prior work (Beňuš et al., 2007) shows that lexical choice and prosody both shape the pragmatic interpretation of backchannels — e.g., when distinguishing 'yeah!' from 'yeah?'. Consequently, inadequate use of backchannel forms risks resulting in pragmatically inappropriate responses.

To use and interpret backchannels effectively, robust representations are required. One promising approach embeds backchannel signals in a continuous space where distance reflects similarity. Earlier work (Qian and Skantze, 2024) showed that this is feasible using a contrastive learning framework that projects past context and feedback into a shared space. However, that work relied on simplistic mean-pooled text embeddings (e.g., BERT by Devlin et al., 2019) to encode dialogue context, limited to the last 4 seconds of the previous turn.

We extend this method by fine-tuning an autore-

gressive large language model (LLM) on dialogue transcripts (Section 4.1); the hidden state preceding a backchannel serves as a dense semantic representation of the context. This is fused with a WavLM (Chen et al., 2022) encoding of the last second of the corresponding speech (Section 4.2). The resulting architecture is shown in Figure 1.

Our contributions are fourfold. First, we show that fine-tuning an autoregressive LLM on spoken dialogue data substantially improves context encoding, which is needed for effective backchannel representations. Second, we highlight the importance of context length for these representations, indicating that the choice of backchannel form is a pragmatically complex phenomenon. Third, we bridge model and human semantics via downstream evaluation tasks grounded in perception data, showing that the learned representations align with human perception. Finally, we demonstrate that interpretable affective dimensions — Energy, Surprisal, and Polarity — can be recovered from the learned representations through linear projections.

2 Related work

Research on conversational feedback and backchannels has historically focused on placement and timing, utilizing concepts such as "feedback relevance spaces" (Heldner et al., 2013; Howes and Eshghi, 2021) and identifying backchannel-inviting cues (Gravano and Hirschberg, 2011). While early approaches relied on rule-based acoustic feature extraction (Koiso et al., 1998; Bertrand et al., 2007; Heldner et al., 2010; Poppe et al., 2010), more recent computational work has shifted toward neural methods, diversifying the potential tasks that can be solved, i.e., those based on backchannel form and meaning. Neural methods include prediction and classification tasks (Ruede et al., 2017; Wang et al., 2024; Park et al., 2024; Inoue et al., 2025; Fukunaga et al., 2025), contrastive learning frameworks (Qian and Skantze, 2024), textless generation (Mai and Carson-Berndsen, 2025), and multi-task learning approaches (Jang et al., 2021; Liermann et al., 2023).

Recent efforts have also targeted the collection, annotation, and analysis of backchannel data. Boudin et al. (2021) labeled feedback attributes (expectedness, valence, and specificity); Figueroa et al. (2022) provided annotations of feedback functions for Switchboard (Godfrey et al., 1992); Müller et al. (2022) focused on backchannels in multimodal groups; and Lin et al. (2025b) created a tri-modal dataset with backchannel and turn shift labels. Studies have also examined relative feedback perception (Qian et al., 2025) and virtual agent backchanneling behavior (Poppe et al., 2013).

Transformer-based large language models (LLMs) have been shown to capture long-range context and pragmatic cues essential for turn-taking (Ekstedt and Skantze, 2020) and response generation (Zhang et al., 2020). While modern LLMs power sophisticated chatbots (OpenAI's ChatGPT, Google's Gemini, etc.), these systems typically operate sequentially with clear turn demarcations, often lacking the naturalness required for real-time conversational feedback. Furthermore, these models are text-based and therefore cannot capture the prosodic aspects of contextual cues and backchannel forms.

In parallel, neural speech models have advanced the representation of prosody, which is critical for pragmatics (Ward et al., 2025). Models like wav2vec 2.0 (Baevski et al., 2020), HuBERT (Hsu et al., 2021), and WavLM (Chen et al., 2022) are trained in a self-supervised manner to learn continuous representations of speech. Recently, generative spoken language models (SLMs) have emerged as systems capable of processing and generating audio directly (Nguyen et al., 2023; Défossez et al., 2024), sometimes in a "full-duplex" fashion. Some fused SLMs combine text and speech embeddings (Arora et al., 2025). While promising, current SLMs often struggle to generate frequent, semantically controlled backchannels, and it is unclear whether they understand the nuances of backchannels coming from the user (Lin et al., 2025a).

Our work sits at the intersection of these fields. As the potential space of backchannel forms (considering both lexical choice and prosody) is very large, it is hard to discretize them in a meaningful way for pure next-token prediction. While previous research has shown how contrastive learning can be used to project context and backchannels into a shared space, it relied on simplistic context representations (Qian and Skantze, 2024). Our approach leverages the semantic depth of LLMs to capture the long-range pragmatic constraints of the preceding context while using audio embeddings to capture the prosodic nuances of both the end of the context and the backchannel response, directly addressing the gap between form and meaning in backchannels.

Learned context and backchannel embeddings

enable several practical applications. First, cosine similarity between contexts and backchannels can be used to rank candidate responses. Second, aligning the embedding space with interpretable dimensions (e.g., Energy, Surprisal, and Polarity) potentially enables semantic control for backchannel synthesis. Finally, projecting user backchannels into the same space allows for inferring their pragmatic meaning in a dialogue system setting.

3 Datasets

We use Fisher Part 1 [1] (Cieri et al., 2004) — a prevalent dataset consisting of 5,850, 3-to-10-minute-long telephone calls between native U.S. English speakers — for training, evaluation, and the perception study, utilizing its time-aligned transcripts for backchannel extraction. The other dataset forming the source of backchannels is FiCa (Figueroa et al., 2024; Carol Figueroa, 2024), which is a set of reenacted and spontaneously produced backchannels uttered by a single native English speaker, thereby eliminating speaker-dependent effects. FiCa was used for evaluation and comparison, in the form presented in Qian et al. (2025).

The set of backchannels included was chosen based on the frequency of their lexical form in the Fisher corpus: 'absolutely', 'ah', 'cool', 'definitely', 'exactly', 'good', 'mhm', 'mm', 'oh', 'okay', 'really', 'right', 'sure', 'uh-huh', 'wow', 'yeah', 'yep', 'yes'. Unlike Qian et al. (2025), we do not include responses in the form of 'no' and direct displays of non-understanding ('pardon', 'sorry', 'what') because these, under certain circumstances, are not regarded as backchannels but rather as full turns, answers, and clarification requests.

4 Models

Model training consisted of two stages. First, LLMs were fine-tuned on transcripts of Fisher conversations to learn textual context representations (Section 4.1). Second, a joint context–backchannel architecture was fine-tuned, leveraging text and speech for *context* modeling and speech alone for *backchannel* generation (Section 4.2).

The first stage corresponds to fine-tuning the **LLM encoder**, as illustrated in Figure 1, while the second stage involves fine-tuning the **projection layers**. Although the LLM fine-tuning stage does not capture prosodic variations of backchannels, we assume it learns how different conversational contexts shape expectations over the probability distribution of lexical backchannel forms (which carry distinct semantics). This enables the joint training stage to associate these lexical expectations with a richer representation of both prosodic and lexical realizations.

4.1 Fine-tuning LLMs for contextual semantic features

In this step, we compared the capability of open-source, state-of-the-art LLMs to model context semantics and to see how the length of prior context (expressed in *number of turns*) affects this. We compared Gemma 3 (Gemma Team et al., 2025; Google Deepmind, 2025), LLaMA 3.1 (Grattafiori et al., 2024; Meta AI, 2024), Qwen2.5 (Yang et al., 2025; Qwen, 2024) and Mistral (Jiang et al., 2023; Mistral AI, 2024). All models were fine-tuned for causal language modeling with a fixed set of hyperparameters (batch size = 2, max token length = 1024), using QLoRA (attention dimension = 32, alpha = 64, dropout = 0.05).

We split the 83,047 Fisher transcript snippets into training and test sets of equal size. First, we trained the models on the training set, with each transcript consisting of up to 50 turns including backchannels. To evaluate the models, we compared their average perplexity over all backchannels found in the test set by feeding in different numbers of preceding turns (1, 3, and 5). For comparison, this was also applied to the corresponding pre-trained models (with 5 preceding turns). The formatting of the transcripts and the calculations are described in Appendix A. The results of our experiments are shown in Table 1.

Average perplexity was computed based on the first token of each backchannel; the relative results were comparable when using a weighted average over all tokens, including those in multi-token words. Perplexity consistently decreased as context length increased, highlighting the importance of rich context for determining backchannel form. The three larger fine-tuned models exhibited similar performance, whereas the fine-tuned Gemma 3 4B yielded slightly higher perplexity, likely due to its smaller size. The pre-trained models showed considerably higher perplexity, demonstrating the necessity of fine-tuning; this was anticipated given the specific formatting of the data.

[1] https://catalog.ldc.upenn.edu/LDC2004T19, https://catalog.ldc.upenn.edu/LDC2004S13

Model	1	3	5	5 (pre)
Gemma 3 4B	11.45	9.88	9.30	54.35
LLaMA 3.1 8B	10.19	8.68	8.19	32.83
Qwen2.5 7B	10.37	8.87	8.40	55.33
Mistral 7B	10.87	9.30	8.76	30.91

Table 1: Average perplexity on the backchannels in the test set using 1, 3, and 5 preceding turns as context, across fine-tuned LLMs. For comparison, the perplexity scores by the corresponding pre-trained models are also included (for 5 turns). Perplexity is calculated on the first token of each backchannel word. For more details, see Appendix A.

4.2 Joint contrastive learning of context and backchannel embeddings

For contrastive learning, we created a joint architecture that projects context and backchannel vector representations into a shared space, forming *joint embeddings*, as illustrated in Figure 1. For **context**, we take the final layer's hidden representation that is used by the LLM head to predict the next token (the first token of the backchannel). This is concatenated with the last second of audio from the interlocutor's channel (ending at the onset of the backchannel), encoded with WavLM (Chen et al., 2022) and mean-pooled over its final layer. For ablation, we also compare this with using only the LLM or WavLM embeddings. These embeddings are then projected with an MLP to a context embedding (Z_{ctx}). For the **backchannel** embedding (Z_{bc}), we simply used the mean-pooled WavLM encoding of the audio with a linear projection.

Loss We use a symmetric InfoNCE-style contrastive loss, similar to the objective introduced in Oord et al. (2018) and later adopted in a symmetric form by Radford et al. (2021). Given a set of N pairs (a batch), the loss maximizes the cosine similarity between the representations of the matching (positive) pairs and minimizes the similarity for the non-matching (negative) pairs:

$$\mathcal{L} = \frac{1}{2}\Big(\mathcal{L}_{context} + \mathcal{L}_{backchannel}\Big)$$

$$\mathcal{L}_{context} = \frac{1}{N}\sum_{i=1}^{N} CE(S_{i,:}, i)$$

$$\mathcal{L}_{backchannel} = \frac{1}{N}\sum_{j=1}^{N} CE(S_{:,j}, j)$$

CE is the cross-entropy loss, and S is a temperature-scaled $N \times N$ cosine similarity matrix containing similarity scores between the joint embeddings of all contexts (rows) and all backchannels (columns). Considering that the goal is to maximize the diagonal, i.e., the similarity score of the ground truth pairs, this matrix can also be viewed as logits from the joint model; in this case, the task can be framed as a multiclass classification problem where the number of classes is equal to the number of true pairs (N). For each context, the model maximizes the score of the true backchannel among N candidates; the same applies to the backchannels trying to optimize the score of their own contexts. As can be seen, this method is largely affected by the batch size N.

Dataset Using the backchannel forms listed in Section 3, 105,209 samples, i.e., context-backchannel pairs, were found in the Fisher dataset. Ground truth context-backchannel pairs form positive samples, while all other combinations serve as negative samples. Batches (sets of pairs) are shuffled and include multiple speakers for robustness and generalizability. We used an 80-10-10% train-validation-test split with mutually exclusive speakers and dialogues and ensured that the validation and test data were not used for LLM fine-tuning.

Hyperparameters With a predefined temperature of 0.07 (as used in Radford et al., 2021), we tuned the following hyperparameters within each category of **text embedding** (fine-tuned Gemma, LLaMA, Qwen, and Mistral, as well as the baseline GTE (Li et al., 2023)): **the number of layers** in the context encoder's MLP (1, 2, 3, 4), **embedding size** (64, 128, 256), and **batch size** (1024, 2048, 4096, 8192). GTE (General Text Embedding) operates with mean-pooled deep contextualized embeddings; it was used as a baseline due to its performance in Qian and Skantze (2024). We also found that projecting the feedback with a linear layer is sufficient, and larger MLPs decrease performance due to the simplicity of the feedback embeddings; therefore, no hyperparameter search was needed on the feedback side. We checked the best validation results within 20 epochs, based on *top-10%* accuracy (see *Metrics* below). The best hyperparameter configurations can be found in Appendix B.

Metric Optimizing cosine similarity scores can be viewed as a ranking task; for each context, the joint model provides a list of predictions in order of similarity. As mentioned before, this is equivalent

Text embeddings	Context modality		
	Audio + text	Text	Audio
Gemma 3 4B	45.6	43.3	29.4
LLaMA 3.1 8B	**49.8**	44.2	
Qwen2.5 7B	46.3	39.7	
Mistral 7B	44.7	**46.4**	
GTE baseline	33.4	21.2	-
random	10.0	10.0	10.0

Table 2: Top-10% test accuracy on the test set, in relation to context embedding modality and text embedding type (irrelevant for the last column). Here, the context consists of 5 turns.

Model	1	3	5	5 (pre)
Gemma 3 4B	**40.1**	40.3	45.6	40.9
LLaMA 3.1 8B	39.9	**46.5**	**49.8**	**41.4**
Qwen2.5 7B	39.7	42.4	46.3	40.9
Mistral 7B	37.3	43.6	44.7	41.3

Table 3: Top-10% test accuracy on the test set, for the full (audio + text) model, depending on how many preceding turns were given to the LLM. Pre-trained baselines with a context length of 5 turns are also provided.

to N-class classification, where the classes are the backchannels in the given batch, making accuracy a fitting metric. However, since backchannels are often very similar and batches are large, we consider *top-k* accuracy. Since this metric is relative to batch size, we use *top-k%* accuracy, similarly to Qian and Skantze (2024), which yields a random baseline of $k\%$. In this paper, we report $k = 10$.

Best hyperparameters Models favored mid-to-large batch sizes (2048, 4096) and generally smaller embedding sizes (64, 128), with no consistent pattern in number of layers. This suggests that dense, sometimes shallow embeddings suffice for modeling the context-backchannel relationships in our dataset. The preference for larger batch sizes was expected, as they improve discrimination between positive and negative pairs (Radford et al. (2021), for example, used a batch size of 32,768), but the question of why the largest chosen batch size (8192) is relatively underrepresented remains.

Results The top-10% test accuracy results are shown in Table 2. Autoregressive LLMs clearly outperform the SSL models in Qian and Skantze (2024), where HuBERT combined with GTE achieved 36.45%. Our GTE baseline performs slightly worse, likely due to a somewhat different selection of backchannels. Consistent with prior findings, combined modalities yield the best representations (with some fluctuations). There were no significant differences among the LLMs, suggesting that small LLMs are sufficient for this problem.

For ablation, we also examined the impact of context length (1, 3, and 5 turns) and LLM fine-tuning, as shown in Table 3. These results reflect the perplexities of the LLMs reported in Table 1, reinforcing the finding that LLM fine-tuning is an important first step, and that longer contexts are important to consider in order to predict backchannel forms. The best-performing architecture is henceforth referred to as *the joint model*.

5 Downstream tasks

While the results from the contrastive learning demonstrate internal alignment between context and backchannel embeddings, they do not guarantee that the learned embeddings correspond to human perception of semantic similarity, neither within or across the embedding spaces. To assess external validity, we collected human perception data for downstream evaluation.

5.1 Human perception data

First, we used the dataset presented by Qian et al. (2025), where participants had compared triplets of feedback responses with identical lexical forms and without context (from the multi-speaker *Fisher* and the single-speaker *FiCa* datasets), selecting the two most similar in terms of prosody and semantics. Using this dataset, we assessed representational quality by expecting the selected pair to be the closest in the embedding space, measured using cosine similarity. We call this the **prosodic backchannel similarity task**, as the lexical forms are identical.

We collected another perception dataset using *Fisher* samples that were not included in the training set for the LLM fine-tuning or the joint model. Participants evaluated a continuous, uninterrupted single-turn context (auditory and textual, 4-80 s) against three audio-only backchannels: *one ground truth* and *two distractors* from different speakers in different dialogues. Unlike the previously mentioned experiment, the candidate responses differed in lexical form. The participants were given three different tasks for each combination of context and

three responses, henceforth referred to as *stimulus set* (for details, see Appendix C and D):

Cross-lexical backchannel similarity task This is a triadic comparison task in which participants select the most similar pair of backchannels. It differs from the *prosodic backchannel similarity task* in that different lexical forms are presented. Here, too, perceived similarity is expected to align with proximity in the embedding space. Results of this task and the *prosodic backchannel similarity task* are reported in Section 5.2.

Context-backchannel matching task Participants rate each backchannel's suitability and naturalness in the given context (on a scale of 1-5). We leverage the ground truth to compare human accuracy against the model, evaluating the alignment of context and backchannel embeddings. Results are reported in Section 5.3.

Affective rating task Participants rate backchannels (1–5) on **Energy** (how energetic the response sounds), **Polarity** (how positive the response sounds), and **Surprisal** (how surprised the backchannel speaker sounds). These names were chosen to be fairly easy for annotators to understand. Polarity and Energy correspond to the notions of *valence* and *arousal*, commonly used in psychology, psycholinguistics, and affective computing (Kuppens et al., 2013; Warriner et al., 2013; Mollahosseini et al., 2019). Surprisal (the level of surprise) was chosen due to its importance as a backchannel function (Figueroa et al., 2022) and because it is not clearly captured by the other two dimensions — although they are slightly correlated, as we show later. We evaluate how well linear probes on projected backchannel embeddings predict these dimensions. Results are in Section 5.4.

In total, we collected 2100 data points (consisting of replies to all three tasks), with 6300 individual ratings of 1260 unique backchannels in 420 unique contexts.

5.2 Backchannel similarity tasks

For the similarity tasks, we selected samples (stimulus sets) with $\geq 80\%$ rater agreement on which backchannels are the most similar. We applied the best joint model's backchannel encoder to obtain embeddings (Z_{bc}), identified the most similar pair via cosine similarity, and calculated the agreement rate with human consensus. As a baseline, we use the mean-pooled WavLM embeddings fed into the backchannel projection layer, allowing us to isolate and quantify how much the projection improves the representational space.

The results, shown in Table 4, indicate that the projection indeed makes the space more discriminative for backchannel similarity. In particular, embeddings produced by the joint model's projection layer yield substantially higher agreement with human judgments than the mean-pooled WavLM baseline, demonstrating that the projection reshapes the acoustic representations in a way that better aligns with perceptual similarity. This suggests that the improvements are not merely due to the underlying pre-trained speech features, but rather to the task-specific structures learned during joint training. Overall, these results support the effectiveness of the backchannel projection in capturing fine-grained similarities that are salient to human listeners.

	Prosodic		**Cross-lexical**
	Fisher	**FiCa**	**Fisher**
Joint model	69.7	75.9	66.3
WavLM	61.7	70.8	56.6
Random	33.3	33.3	33.3

Table 4: Proportions of correct selections in the triadic backchannel similarity tasks (%).

5.3 Context-backchannel matching task

For the context–backchannel matching task, we computed the cosine similarity between the context and backchannel embeddings produced by our context and backchannel encoders (Z_{ctx} and Z_{bc}), selecting the backchannel with the highest similarity score for a given context. We measured the proportion of cases where the model correctly identified the ground truth backchannel, using 1 or 5 context turns. To compare with human performance, we averaged rater scores for each backchannel option, selecting the backchannel with the highest mean assigned score. We then calculated how often this choice matched the ground truth.

Results (Table 5) show that the model, surprisingly, substantially outperforms humans, even when it only has access to 1 context turn (just like humans had in the experiment). An alternative interpretation is that the relatively low human "performance" indicates that the same conversational context can allow for multiple valid backchannel

responses. Thus, treating the actual backchannel as ground truth should be done with caution.

Rater type	Score
Joint model (5 turns)	72.3
Joint model (1 turn)	62.0
Human (1 turn)	47.4
Random	33.3

Table 5: Proportions of correct selections in the context-backchannel matching task (%).

5.4 Affective rating task

Figure 2 displays the distributions of median ratings for each backchannel: Energy is centered but Surprisal skews lower, which is consistent with how many continuers (e.g., 'mhm', 'yeah', and 'right') do not express surprise. Polarity is predominantly neutral or positive, reflecting the fact that backchannels are less frequently used to express negative sentiments (Jurafsky et al., 1998). The correlation between these dimensions is also shown in Figure 2; this indicates that Surprisal and Polarity correlate the least, while Energy captures some aspects of both. For a more detailed analysis grouped by lexical category, see Appendix E.

To measure how well the backchannel embeddings (Z_{bc}) capture these dimensions, we split the samples (50/50 train/test) at random and fitted a linear ridge regression probe ($\alpha = 1.0$). We compared this against four baselines: raw mean-pooled WavLM embeddings, one-hot encoded lexical tokens, basic prosodic features (*pitch range* in semitones, *duration* in number of voiced frames), and the combination of lexical categories and prosodic features. Pitch was extracted using Reaper.

The results (Table 6) show that our learned embeddings capture these perceived dimensions better than WavLM, while simple prosodic features perform the worst. This indicates that the distributed representations integrate lexical and prosodic information more effectively than low-level acoustics or lexical identity alone. Consistent with our earlier findings in this paper, the backchannel projection layer further reshapes the embedding space in a way that improves alignment with human judgments.

6 Analysis of backchannel embeddings

The downstream tasks demonstrate the utility and validity of the learned context and backchannel

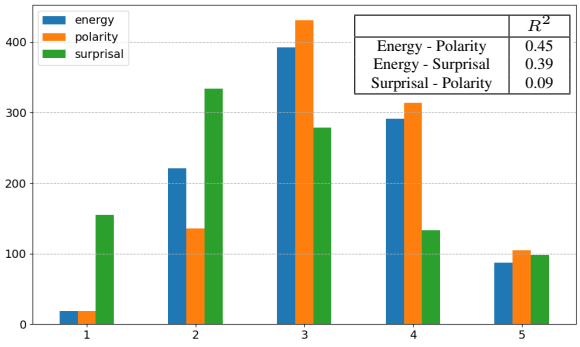

Figure 2: Distribution of median affective ratings over backchannels and correlation between mean ratings.

	Energy	Polarity	Surprisal
Joint model	0.465	0.341	0.552
WavLM	0.406	0.287	0.502
Lexical	0.193	0.204	0.432
Prosody	0.118	0.028	0.156
Lex. + Pros.	0.240	0.237	0.473

Table 6: Linear probe fit $\left(R^2\right)$ on the test set in the affective rating task.

embedding spaces. To investigate this further, we developed a tool to visualize backchannel embedding projections in different ways and listen to their prosodic realizations for both the Fisher and FiCa datasets (https://feedbackembeddings.github.io/demo1/).

Figure 3 shows a scatterplot of the representation space created by the tool. The backchannel embeddings are projected onto the *Surprisal* and *Polarity* affective dimensions (the least correlated dimensions) using the learned probes from Section 5.4. For clarity, we only show 'yeah', 'mm', 'exactly', and 'really'. The tool can also provide prosodic analysis over a region of backchannels, which is shown as rectangles in the plot. The reported metrics are average duration and average pitch range, as defined in Section 5.4.

The four lexical tokens exhibit distinct semantic tendencies: 'really' conveys high Polarity and Surprisal, whereas 'exactly' signals strong Polarity but low Surprisal. Both 'mm' and 'yeah' show weaker Polarity (with 'yeah' slightly stronger), while 'mm' generally expresses higher Surprisal than 'yeah'. Despite these lexical tendencies, there is substantial variation driven by prosodic realization, leading to considerable overlap between lexical clusters. Although higher Surprisal and Polarity are broadly

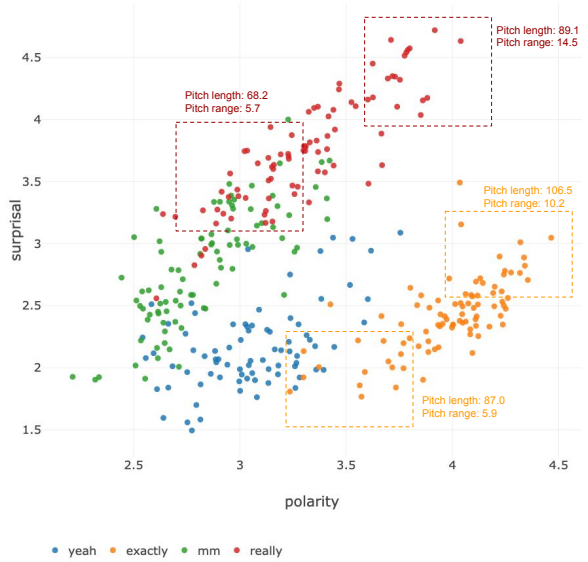

Figure 3: Backchannel samples from the Fisher corpus, embedded through the model, and projected to the Surprisal and Polarity affective dimensions. Rectangles show average pitch length (in voiced frames, 100 Hz) and pitch range (in semitones) for a certain region and for a specific lexical form.

associated with longer durations and wider pitch ranges, Table 6 shows that simple prosodic features alone are insufficient predictors, suggesting that the relevant prosodic cues are more nuanced. A more detailed analysis of the central tendencies and dispersion of lexical forms with respect to the affective dimensions is provided in Appendix E.

7 Discussion

The findings in this work reinforce the view of backchannels as pragmatically conditioned signals whose lexical and prosodic forms are important for their appropriateness and nuanced meaning. The strong effect of context length suggests that backchannel choice depends on discourse-level structure and expectations rather than solely on local acoustic or lexical cues, and that such dependencies can be effectively captured by fine-tuned autoregressive language models.

The joint contrastive framework provides a useful abstraction for modeling backchannels in continuous space. Improvements over raw acoustic representations in perceptual similarity tasks indicate that the learned embedding space emphasizes pragmatically salient variation while down-weighting speaker- and token-specific idiosyncrasies. This is particularly evident in the cross-lexical similarity results, where different lexical forms are grouped according to perceived functional similarity rather than surface form alone.

As discussed earlier, while current generative SLMs should, in principle, be able to handle backchannels, they often struggle to generate and understand them correctly (Lin et al., 2025a). SLMs are trained to predict or synthesize the next segment of speech end-to-end, which entangles backchannel behavior with many other factors (lexical content, speaker identity, channel conditions) and makes fine-grained control of short feedback signals challenging. In contrast, we learn pragmatically meaningful representations in a joint embedding space that aligns dialogue contexts with backchannel realizations via a contrastive objective. This representation-centric formulation supports efficient retrieval/ranking, interpretable analysis (e.g., affective axes), and direct validation against human perceptual similarity — capabilities that are difficult to obtain from generation-focused objectives alone. Our approach could also be complementary to generative SLMs and could serve as a lightweight backchannel selection/control module within a more modular spoken dialogue system.

8 Conclusion

In this paper, we address the undermodeled link between the lexical–prosodic form of backchannels and their pragmatic appropriateness and interpretation in context. We introduced a two-stage approach that (i) fine-tunes an autoregressive LLM on spoken dialogue transcripts to obtain richer contextual representations and (ii) learns a shared embedding space that aligns dialogue contexts with acoustic backchannel realizations via contrastive learning. In retrieval-style evaluations, longer conversational context consistently improved performance, indicating that backchannel choice depends on pragmatics that extend beyond the immediately preceding turn.

Crucially, the learned projections produced backchannel embeddings that better match human perceptual structure than raw WavLM features: agreement increased in both prosodic and cross-lexical triadic similarity judgments. Linear probes further showed that the embedding space supports interpretable affective dimensions (Energy, Polarity, Surprisal), suggesting a path toward controllable feedback generation and pragmatic inference of user feedback in conversational systems.

Limitations

Modeling backchannels remains challenging due to their personality-dependent (Warriner et al., 2013) and idiosyncratic nature (Blomsma et al., 2024), as well as their dependence on specific dyadic dynamics (Cavalcanti and Skantze, 2025). Furthermore, usage varies by language (Heinz, 2003; Beňuš, 2016; Liesenfeld and Dingemanse, 2022), dialect (Wong and Peters, 2007; Kraaz and Bernaisch, 2022), and speaker proficiency (Cutrone, 2005; Shelley and Gonzalez, 2013; Cutrone, 2014; Lee, 2020; Heinz, 2003). In this work, we disregard these factors, experimenting with data collected from native speakers of United States English, irrespective of dialect. In the Fisher dataset, most participants did not know each other, making their dynamics more general and less clouded by previous acquaintance; however, we believe that future work needs to address individual, dyadic, and demographic differences to create more universally adaptive conversational agents.

While previous work on joint embeddings of backchannels (Qian and Skantze, 2024) has investigated different speech encoders, we have consistently used WavLM over alternatives such as wav2vec 2.0 and HuBERT, due to its superior performance on the SUPERB benchmark (Yang et al., 2021) and its robust capacity for modeling paralinguistic features (Chen et al., 2022). We also kept the audio context length short (1 second), given that the simple temporal mean-pooling we used cannot capture complex temporal dynamics or longer-term acoustic dependencies anyway. Furthermore, we kept the WavLM weights frozen throughout the training process to focus specifically on the alignment mechanism. While fine-tuning the audio backbone or employing different self-supervised speech models (or layers) might yield higher absolute performance metrics, we assume that the relative trends and the efficacy of the contrastive framework observed in this study would likely remain consistent.

Due to resource constraints, we fine-tuned smaller LLMs (4B and 7B), requiring 1-2 days on 8 NVIDIA RTX 3090 GPUs per model. We avoided LLM hyperparameter tuning, as the performance differences between model sizes were negligible. Focusing on context representations, we did not heavily optimize backchannel encoding; however, we found that a linear projection of audio-only encodings performed best.

Ethical considerations

As far as we are aware, our methods do not have any harmful effects, biases, or risks, apart from what can generally be expected from machine learning-based models (such as introducing demographic or linguistic biases). We did not develop new foundation models, nor have we created large-scale datasets that can be mass-deployed to cause intentional harm.

For data collection, participants were informed that we do not intend to store personal or sensitive data and that the information they provide is anonymized. No ethics approval was needed.

All coding was done in Python using Pytorch, and pre-trained models were downloaded from Hugging Face. The code and the dataset will be made public after the review process. AI assistants (ChatGPT and Gemini) were used to correct grammar and reformulate sentences in the paper. Additionally, they were used for coding in some instances.

References

Jens Allwood, Joakim Nivre, and Elisabeth Ahlsén. 1992. On the semantics and pragmatics of linguistic feedback. *Journal of Semantics*, 9(1):1–26.

Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2025. On the landscape of spoken language models: A comprehensive survey. *arXiv preprint arXiv:2504.08528*.

Agnes Axelsson, Hendrik Buschmeier, and Gabriel Skantze. 2022. Modeling feedback in interaction with conversational agents: A review. *Frontiers in Computer Science*, 4:744574.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 33:12449–12460.

Štefan Beňuš. 2016. The prosody of backchannels in Slovak. In *Proceedings of the 8th International Conference on Speech Prosody*, pages 75–79.

Štefan Beňuš, Agustín Gravano, and Julia Bell Hirschberg. 2007. The prosody of backchannels in American English. *Proceedings of the 16th International Congress of Phonetic Sciences (ICPhS 2007)*, pages 1065–1068.

Roxane Bertrand, Gaëlle Ferré, Philippe Blache, Robert Espesser, and Stéphane Rauzy. 2007. Backchannels

revisited from a multimodal perspective. In *Auditory-Visual Speech Processing*, pages 1–5.

Peter Blomsma, Julija Vaitonyté, Gabriel Skantze, and Marc Swerts. 2024. Backchannel behavior is idiosyncratic. *Language and Cognition*, 16(4):1158–1181.

Auriane Boudin, Roxane Bertrand, Stéphane Rauzy, Magalie Ochs, and Philippe Blache. 2021. A multimodal model for predicting conversational feedbacks. In *Text, Speech, and Dialogue*, pages 537–549. Springer International Publishing.

Carol Figueroa. 2024. FiCa speech dataset. Accessed: 2026-01-02.

Julio Cesar Cavalcanti and Gabriel Skantze. 2025. "Dyadosyncrasy", idiosyncrasy and demographic factors in turn-taking. *arXiv preprint arXiv:2505.24736*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Christopher Cieri, David Miller, and Kevin Walker. 2004. The Fisher corpus: A resource for the next generations of speech-to-text. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, volume 4, pages 69–71.

Herbert H. Clark and Edward F. Schaefer. 1989. Contributing to discourse. *Cognitive Science*, 13(2):259–294.

Pino Cutrone. 2005. A case study examining backchannels in conversations between Japanese-British dyads. *Multilingua, Journal of Cross-Cultural and Interlanguage Communication*, 24(3):237–274.

Pino Cutrone. 2014. A cross-cultural examination of the backchannel behavior of Japanese and Americans: Considerations for Japanese EFL learners. *Intercultural Pragmatics*, 11(1):83–120.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, pages 4171–4186.

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.

Gaëlle Ferré and Suzanne Renaudier. 2017. Unimodal and bimodal backchannels in conversational English. In *Proceedings of the 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017)*, pages 27–37.

Carol Figueroa, Adaeze Adigwe, Magalie Ochs, and Gabriel Skantze. 2022. Annotation of communicative functions of short feedback tokens in Switchboard. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference (LREC)*, pages 1849–1859.

Carol Figueroa, Marcel de Korte, Magalie Ochs, and Gabriel Skantze. 2024. Mhm... Yeah? Okay! evaluating the naturalness and communicative function of synthesized feedback responses in spoken dialogue. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, pages 544–553.

Yoshinori Fukunaga, Ryota Nishimura, Kengo Ohta, and Norihide Kitaoka. 2025. Backchannel prediction for natural spoken dialog systems using general speaker and listener information. In *Proceedings of Interspeech 2025*, pages 1078–1082.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H. Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. In *Proceedings of the Annual Meeting of the Cognitive Science Society (CogSci 2017)*, pages 2055–2060.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

J.J. Godfrey, E.C. Holliman, and J. McDaniel. 1992. SWITCHBOARD: Telephone speech corpus for research and development. In *Proceedings of ICASSP*, volume 1, pages 517–520.

Google Deepmind. 2025. Gemma 3 4B. Accessed: 2026-01-02.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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

Bettina Heinz. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics*, 35(7):1113–1142.

Mattias Heldner, Jens Edlund, and Julia Hirschberg. 2010. Pitch similarity in the vicinity of backchannels. In *Proceedings of Interspeech 2010*, pages 3054–3057.

Mattias Heldner, Anna Hjalmarsson, and Jens Edlund. 2013. Backchannel relevance spaces. In *Nordic Prosody XI*, pages 137–146.

Christine Howes and Arash Eshghi. 2021. Feedback relevance spaces: Interactional constraints on processing contexts in dynamic syntax. *Journal of Logic, Language and Information*, 30(2):331–362.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Koji Inoue, Divesh Lala, Gabriel Skantze, and Tatsuya Kawahara. 2025. Yeah, Un, Oh: Continuous and real-time backchannel prediction with fine-tuning of voice activity projection. In *Proceedings of NAACL (Volume 1: Long Papers)*, pages 7171–7181.

Ryo Ishii, Xutong Ren, Michal Muszynski, and Louis-Philippe Morency. 2021. Multimodal and multitask approach to listener's backchannel prediction: Can prediction of turn-changing and turn-management willingness improve backchannel modeling? In *IVA '21: Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, pages 131–138.

Jin Yea Jang, San Kim, Minyoung Jung, Saim Shin, and Gahgene Gweon. 2021. BPM_MT: Enhanced backchannel prediction model using multi-task learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3447–3452.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, and 1 others. 2023. Mistral 7B. *Preprint*, arXiv:2310.06825.

Dan Jurafsky, Elizabeth Shriberg, Barbara Fox, and Traci Curl. 1998. Lexical, prosodic, and syntactic cues for dialog acts. In *Discourse Relations and Discourse Markers*.

Hanae Koiso, Yasuo Horiuchi, Syun Tutiya, Akira Ichikawa, and Yasuharu Den. 1998. An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs. *Language and Speech*, 41(3-4):295–321.

Michelle Kraaz and Tobias Bernaisch. 2022. Backchannels and the pragmatics of South Asian Englishes. *World Englishes*, 41(2):224–243.

Peter Kuppens, Francis Tuerlinckx, James A. Russell, and Lisa Feldman Barrett. 2013. The relation between valence and arousal in subjective experience. *Psychological Bulletin*, 139(4):917.

Kanghee Lee. 2020. Backchannels as a cooperative strategy in ELF communications. *Korean Journal of English Language and Linguistics*, 20:257–281.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Wencke Liermann, Yo-Han Park, Yong-Seok Choi, and Kong Lee. 2023. Dialogue act-aided backchannel prediction using multi-task learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15073–15079.

Andreas Liesenfeld and Mark Dingemanse. 2022. Bottom-up discovery of structure and variation in response tokens ('backchannels') across diverse languages. In *Proceedings of Interspeech 2022*, pages 1126–1130.

Guan-Ting Lin, Jiachen Lian, Tingle Li, Qirui Wang, Gopala Anumanchipalli, Alexander H. Liu, and Hung-yi Lee. 2025a. Full-Duplex-Bench: A benchmark to evaluate full-duplex spoken dialogue models on turn-taking capabilities. *arXiv preprint arXiv:2503.04721*.

Yuxin Lin, Yinglin Zheng, Ming Zeng, and Wangzheng Shi. 2025b. Predicting turn-taking and backchannel in human-machine conversations using linguistic, acoustic, and visual signals. In *Proceedings of ACL (Volume 1: Long Papers)*, pages 15310–15322.

Long Mai and Julie Carson-Berndsen. 2025. Real-time textless dialogue generation. *arXiv preprint arXiv:2501.04877*.

Meta AI. 2024. LLaMA 3.1 8B. Accessed: 2026-01-02.

Mistral AI. 2024. Mistral 7B v0.3. Accessed: 2026-01-02.

Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. 2019. AffectNet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31.

Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Hali Lindsay, Patrick Gebhard, Elisabeth André, and Andreas Bulling. 2022. MultiMediate'22: Backchannel detection and agreement estimation in group interactions. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 7109–7114.

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

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Daniel Ortega, Chia-Yu Li, and Ngoc Thang Vu. 2020. Oh, Jeez! or uh-huh? a listener-aware backchannel predictor on ASR transcriptions. In *Proceedings of ICASSP*, pages 8064–8068.

Yo-Han Park, Wencke Liermann, Yong-Seok Choi, and Kong Joo Lee. 2024. Improving backchannel prediction leveraging sequential and attentive context awareness. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1689–1694.

Ronald Poppe, Khiet P. Truong, and Dirk Heylen. 2013. Perceptual evaluation of backchannel strategies for artificial listeners. *Autonomous Agents and Multi-Agent Systems*, 27(2):235–253.

Ronald Poppe, Khiet P. Truong, Dennis Reidsma, and Dirk Heylen. 2010. Backchannel strategies for artificial listeners. In *International Conference on Intelligent Virtual Agents*, pages 146–158. Springer.

Livia Qian, Carol Figueroa, and Gabriel Skantze. 2025. Representation of perceived prosodic similarity of conversational feedback. In *Proceedings of Interspeech 2025*, pages 374–378.

Livia Qian and Gabriel Skantze. 2024. Joint learning of context and feedback embeddings in spoken dialogue. In *Proceedings of Interspeech 2024*, pages 2955–2959.

Qwen. 2024. Qwen2.5 7B. Accessed: 2026-01-02.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139, pages 8748–8763.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2017. Enhancing backchannel prediction using word embeddings. In *Proceedings of Interspeech 2017*, pages 879–883.

Robin Ruede, Markus Müller, Sebastian Stüker, and Alex Waibel. 2019. Yeah, right, uh-huh: A deep learning backchannel predictor. In *Advanced Social Interaction with Agents: 8th International Workshop on Spoken Dialog Systems*, pages 247–258. Springer.

Leah Shelley and Fernanado Gonzalez. 2013. Back channeling: Function of back channeling and L1 effects on back channeling in L2. *Linguistic Portfolios*, 2(1):9.

Khiet Phuong Truong, Ronald Walter Poppe, I.A. de Kok, and Dirk K.J. Heylen. 2011. A multimodal analysis of vocal and visual backchannels in spontaneous dialogs. In *Proceedings of Interspeech 2011*, pages 2973–2976.

Jinhan Wang, Long Chen, Aparna Khare, Anirudh Raju, Pranav Dheram, Di He, Minhua Wu, Andreas Stolcke, and Venkatesh Ravichandran. 2024. Turn-taking and backchannel prediction with acoustic and large language model fusion. In *Proceedings of ICASSP*, pages 12121–12125.

Nigel G. Ward, Divette Marco, and Olac Fuentes. 2025. Which prosodic features matter most for pragmatics? In *Proceedings of ICASSP*, pages 1–5.

Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. 2013. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods*, 45(4):1191–1207.

Deanna Wong and Pam Peters. 2007. A study of backchannels in regional varieties of English, using corpus mark-up as the means of identification. *International Journal of Corpus Linguistics*, 12(4):479–510.

An Yang, Bowen Yu, Chengyuan Li, Dayiheng Liu, and 1 others. 2025. Qwen2.5-1M technical report. *arXiv preprint arXiv:2501.15383*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, and 1 others. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. *Proceedings of Interspeech 2021*, pages 1194–1198.

Victor H. Yngve. 1970. On getting a word in edgewise. In *CLS-70*, pages 567–577. University of Chicago.

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.

A Transcripts for LLM fine-tuning

A.1 Spoken dialogue transcripts

For LLM fine-tuning, we used unpunctuated lowercase transcripts based on the ones provided with the Fisher dataset. A set of special characters were used to indicate speaker shifts and overlapping speech, however these were treated as ordinary characters from the perspective of the language model, so that no new tokens had to be introduced:

- **<A>** and **<B>**: the beginning of Speaker A or B's turn, respectively
- **Slash** (/): turn shift
- **Braces** ({}): the part of a given turn (Turn A) that overlaps with parts of the next turn (Turn B). There is always a corresponding part surrounded by square brackets in the next turn (Turn B), unless the current turn (Turn A) is the last one in the transcript.
- **Square brackets** ([]): the part of a given turn (Turn B) that overlaps with parts of the previous turn (Turn A). There is always a corresponding part surrounded by braces in the previous turn (Turn A), unless the current turn (Turn B) is the first one in the transcript.

Backchannels were located by searching for complete turns that consisted *only* of the backchannels we are investigating (see Section 3) and *only once*, i.e., we identified them only if they were surrounded by two turn shifts and a speaker token (with occasional braces and brackets).

Here are two example transcripts, with backchannels marked in bold:

Example transcript 1

<B> hi / <A> hello / <B> hi / <A> oh okay um my name is brandon um you / <B> right um my name's rajul mm / <A> okay okay so uh what do you think / <B> um i didn't quite catch the topic i mean i got the gist of it but um can you hear the topic of the day / <A> okay basically um how has corporate scandals affected you or do you think um what is what is the affect of corporate scandals on america um do you believe it's responsible for the m the mild recession that we we been having lately mm / <B> **right** / <A> **mhm** / <B> um well uh personally for me to the the solution and uh you know the aftermath of the uh uh the uh tech boom and the uh bubble thereafter because i think um probably um because uh i think because of my ignorance at that point in time about / <A> mhm mhm / <B> um about probably what i would call uh the truth as i see it wh what actually goes on in in the um in in the stock market as well as uh uh the biggest uh financial institutions uh in america um but initially e i i i really could not believe uh um things that were coming out uh in in the newspapers you know arthur anderson and uh citibank merrill lynch you know just n name a all those groups were like uh um being charged for um um for misguiding and misleading uh the investors / <A> mhm mhm / <B> and and thereby um in my words uh duping people uh of their money uh uh pretty blatantly { knowing } / <A> [**mhm**] /

Example transcript 2

<A> and you know we we uh all / <B> **yeah** / <A> of us hurt uh because of that and i believe it's contributed to the recession it's i mean the recess recessions happen because they happen i mean they they come in waves but / <B> **right** / <A> i don't i don't think that me i don't think that america takes white collar white collar crimes seriously and { it } / <B> [**right**] / <A> doesn't it doesn't try to it doesn't try to stop it at all i don't i think i don't know { h } / <B> [**right**] / <A> uh if you have a theory behind that because also i think the main people who are running the / <B> **mhm** / <A> the country the main people who are actually contributing to politicians who run the countr the country are all these companies { who are } / <B> [**right**] / <A> so rich and have all this money and so if they you know if they embezzle a few million dollars it doesn't hurt anyone cause we're still in power { th } / <B> [**right**] / <A> these are the um the people in power who are talking you know we're st we're still in power and like they're still paying us off so it doesn't really matter but everyone { but } / <B> [right right] / <A> everyone else hurts in the long run so i don't know /

In *Example transcript 1*, the penultimate turn overlaps with the last turn, which means that **{ knowing }** is uttered at the same time as **[mhm]** in the original audio. Other vocalized dialogue phenomena are included, e.g., repetitions and repairs, without special markers.

A.2 Defining the context

When restricting the number of past turns to, e.g., 2, the contexts corresponding to some of the above examples appear as follows (where everything before the backchannel in bold is treated as context):

<B> um i didn't quite catch the topic i mean i got the gist of it but um can you hear the topic of the day / <A> okay basically um how has corporate scandals affected you or do you think um what is what is the affect of corporate scandals on america um do you believe it's responsible for the m the mild recession that we we been having lately mm / <B> **right**

<A> okay basically um how has corporate scandals affected you or do you think um what is what is the affect of corporate scandals on america um do

Text embeddings	Context modality	
	Audio + text	Text
Gemma 3 4B	4, 128, 2048	2, 128, 4096
LLaMA 3.1 8B	3, 64, 4096	3, 128, 4096
Qwen2.5 7B	4, 64, 2048	1, 128, 2048
Mistral 7B	3, 64, 2048	1, 64, 8192

Table 7: Hyperparameters for the models (5 turns) in Table 2, listed in the order *number of layers*, *embedding size*, *batch size*. For audio-only, the optimal configuration was 2, 64, and 2048.

Model	1	3
Gemma 3 4B	2, 64, 2048	3, 64, 1024
LLaMA 3.1 8B	4, 256, 2048	2, 64, 4096
Qwen2.5 7B	3, 64, 2048	3, 64, 2048
Mistral 7B	3, 64, 2048	4, 64, 4096

Table 8: Hyperparameters for the models (text + audio, fine-tuned LLM context encoder) in Table 3 in the order *number of layers*, *embedding size*, and *batch size*. The number of past turns is 1 and 3.

`you believe it's responsible for the m the mild recession that we we been having lately mm / <B> right / <A>` **`mhm`**

`<A> mhm mhm / <B> and and thereby um in my words uh duping people uh of their money uh uh pretty blatantly { knowing } / <A> [` **`mhm`**

Thus, when computing the perplexity of the backchannel token, or the LLM context embedding, all text up to the backchannel (marked in bold) is used, including the current turn's speaker token and potential opening brackets.

B Best hyperparameters

The hyperparameters of the best models for each configuration in Section 4.2 are shown in Table 7 and 8.

C Perception study

In the perception study, participants were given three tasks, with the last one consisting of three separate questions regarding Energy, Polarity and Surprisal. In total, five questions were asked for each stimulus set. The participants were shown a range of examples at the beginning of the study.

The perception study was conducted on 100 native speakers of North American (U.S.) English who reported English as their primary — most frequently used — language and the United States as their primary place of residence during their first 18 years. The participants had no hearing difficulties. Each participant received 21 stimulus sets and two additional sets for attention checks. Each stimulus set was seen by at least three subjects.

Participants were recruited via Prolific [2]. The median completion time was approximately 56 minutes, and the participants were compensated with Prolific's default reward per hour.

General instruction: Listen to the context and the feedback responses. The feedback responses are repeated at the beginning of each question where you have to rate them individually.

Question 1: Please rate the feedback responses based on how well they match the context (min: 1, max: 5).

Question 2: Choose the two responses that are the most similar to each other.

Question 3: Rate the energy level: how energetic is the response (min: 1, max: 5)?

Question 4: Rate polarity: how positive is the response (min: 1, max: 5)?

Question 5: Rate surprisal: how surprised does the feedback speaker sound (min: 1, max: 5)?

D Informed consent and general information (verbatim)

For each question, you will hear:

- one **context** clip (Speaker 1)
- three short **feedback** clips (Speaker 2), labeled as 1, 2 and 3

Your task

1. **Rate compatibility**
   - For each feedback, rate how well it fits the context as a possible response
   - Scale: 1 = not at all, 5 = extremely
   - Consider factors like naturalness, appropriateness and expectedness
2. **Choose similar feedback responses**

[2] https://www.prolific.com

- Pick the two feedback responses most similar in role, attitude, intention and emotion (not just meaning!), given the current context
- Options: 1–2, 2–3 or 1–3
- Consider factors like how replaceable they are with each other! Take into account both sound and meaning.

3. **Rate energy**
   - How energetic does the feedback speaker sound?
   - Scale: 1 = not at all, 5 = extremely
4. **Rate polarity**
   - How positive is what the feedback speaker said?
   - Scale: 1 = not at all, 5 = extremely
5. **Rate surprisal**
   - How surprised does the feedback speaker sound?
   - Scale: 1 = not at all, 5 = extremely

Instructions

- You will see some examples on the first six pages. Please, read them carefully.
- **You have to wait a few seconds for the survey to load.**
- **Evaluate feedback together with context**, not in isolation. For better judgment, play each feedback **immediately after the context**.
- Consider tone, emotion, and purpose — but **ignore** differences in recording quality and loudness.

Listening requirements

- Use **headphones** in a calm environment.
- Replay samples as needed, but you **cannot return** to earlier ones.
- The volume of the context can be adjusted.
- If an audio file doesn't load or doesn't play, wait a few seconds. Once loaded, it should be possible to play it without delay. If it still does not work, please take a note of where the issue arose and contact us.

Attention checks

- Very short or long completion times may lead to rejection.
- Failing attention checks results in rejection. Listening activity *is recorded* for validation.
- The survey will not end if you fail the attention checks, and your answers, including the checks, will be evaluated after your submission.

Other notes

- Answers will help study how people perceive *backchannels* in conversation.
- Audio clips are from a dataset and do not represent our views.
- Responses are anonymized; no personal/sensitive data is collected.
- You may stop anytime or contact us via Prolific if needed.

By continuing, you confirm voluntary participation and acceptance of these conditions.

Thank you!

E Distribution of affective ratings

The mean and standard deviation over the median scores for each backchannel, grouped by lexical category, are shown in Table 9. These provide insight into the typical human perception of each category and how it varies within each group. The data suggests a hierarchy from high-arousal, expressive words (top of the table) to low-arousal, passive continuers (bottom of the table).

Highest Energy: The token "wow" (3.74 ± 0.70) has the highest average Energy, indicating that it is perceived as the most intense or aroused response in general. It is closely followed by strong, occasionally enthusiastic agreements like "absolutely" (3.67) and "exactly" (3.67).

Lowest Energy: The tokens "mhm" (2.70 ± 0.68), "yeah" (2.83), and "right" (2.87) are at the bottom. This aligns with their typical linguistic function as "passive" continuers, i.e., simple signals that the listener is still present, without adding significant emotion.

Most positive: "Absolutely" (4.03 ± 0.67) and "definitely" (3.98 ± 0.88) are the most positive tokens. This suggests that multi-syllabic, explicit agreement words carry more positive weight than short sounds.

Least positive (most neutral): The lowest Polarity scores belong to "mm" (2.71) and "mhm" (2.92). These scores are likely lower not because they are negative, but because they are highly neutral (our selection does not, for the most part, include backchannels that are usually perceived as expressing negative sentiment).

Most surprised: Words like "wow" (3.92), "really" (3.84), "ah" (3.82), and "oh" (3.69) all have very high Surprisal scores. These are typically used when the listener is reacting to new, shocking, or interesting information.

Least surprised: Words like "mhm" (2.04), "yep" (2.21), "right" (2.22), and "sure" (2.30) have low Surprisal scores. These are used to confirm known information or simply to agree.

High variance/variability: Certain tokens have consistently high standard deviations, suggesting that they can be pronounced in various ways and that their meaning depends heavily on how they are said (tone/prosody) rather than just on the word itself. Examples include "oh", "mm", and "cool".

Table 9: Mean ± standard deviation per backchannel lexical token. Calculated on the median rating of each backchannel.

token	energy	polarity	surprisal
wow	3.74 ± 0.70	3.64 ± 0.61	3.92 ± 0.71
absolutely	3.67 ± 0.73	4.03 ± 0.67	2.45 ± 0.56
exactly	3.67 ± 0.61	3.96 ± 0.57	2.41 ± 0.66
ah	3.57 ± 0.74	3.27 ± 0.67	3.82 ± 0.78
really	3.56 ± 0.74	3.22 ± 0.72	3.84 ± 0.74
definitely	3.51 ± 0.70	3.98 ± 0.88	2.34 ± 0.44
oh	3.45 ± 0.86	3.17 ± 0.87	3.69 ± 0.89
good	3.34 ± 0.71	3.55 ± 0.59	2.84 ± 0.63
yes	3.11 ± 0.73	3.39 ± 0.65	2.21 ± 0.62
cool	3.07 ± 0.83	3.49 ± 0.69	2.65 ± 0.81
okay	3.02 ± 0.64	3.21 ± 0.58	2.54 ± 0.69
uh-huh	3.00 ± 0.65	3.07 ± 0.69	2.38 ± 0.67
yep	2.95 ± 0.62	3.30 ± 0.57	2.21 ± 0.61
sure	2.89 ± 0.71	3.34 ± 0.64	2.30 ± 0.71
mm	2.88 ± 0.82	2.71 ± 0.70	2.82 ± 0.95
right	2.87 ± 0.66	3.30 ± 0.66	2.22 ± 0.52
yeah	2.83 ± 0.62	3.14 ± 0.55	2.20 ± 0.60
mhm	2.70 ± 0.68	2.92 ± 0.61	2.04 ± 0.58