CHARM: Cross-Cultural Harmonious Communication Multi-Modal Analysis via Norms and Emotion

Yi Fung[1]*, **Ziwei Gong**[2]*, **Zehui Wu**[2]*, **Yunfan Zhang**[2]*, **Charles Yu**[1]*
Zhaoyuan Deng[2], **Adil Soubki**[3], **Yanda Chen**[2], **Carl Vondrick**[2]
Owen Rambow[3], **Julia Hirschberg**[2], **Kathleen Mckeown**[2], **Heng Ji**[1]
[1]UIUC [2]Columbia [3]Stony Brook
sara.ziweigong@columbia.edu hengji@illinois.edu

Abstract

Communication between people from different cultures can be challenging due to a lack of understanding about the norms of another culture or the degree of emotion typically conveyed. If unaddressed, communication failure is a significant risk for such interactions. In this work, we propose a new paradigm for understanding and guiding human communication through detecting norm occurrences, emotional cues, and change-points in conversations leveraging multimedia signals. In particular, we present a framework for **cr**oss-cultural **h**armonious communication multi-modal **a**nalysis using no**r**ms and e**m**otions (CHARM), and show how each component achieves promising results on new tasks. We conclude by envisioning how CHARM can be used as a cultural interpreter for real-world human interactions.[1]

1 Introduction

We live in a highly interconnected, globalized world in which many organizations and individuals operate internationally and are in constant contact with diverse cultures. Communicative understanding, not simply of local languages but also of social customs and cultural backgrounds, lies at the heart of industrial, governmental, and cultural exchanges. In particular, cultural interpreters play a crucial role in various settings, including *healthcare*, where they help patients and providers communicate; *legal domain*, facilitating interaction between lawyers, judges, and clients; *social services*, to assist individuals of need access support; and *professional meetings*, promoting effective dialogue in international business and government diplomacy. Without such mediation, cross-cultural miscommunication can lead to misunderstandings, offense, disrupted negotiations, and even conflict (WHO, 2009; Legros and Cislaghi, 2020).

[1]CHARM is publicly released and actively maintained at https://github.com/yrf1/CHARM, along with a demo video for quick-start, calling for broader development.

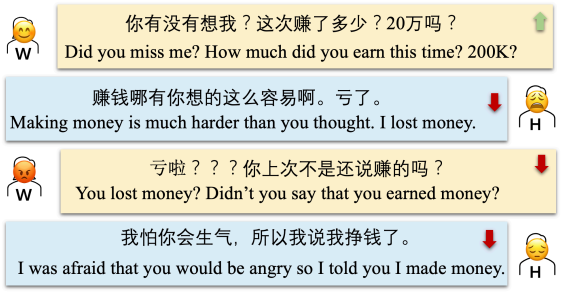

Figure 1: An example of a conversation that is likely headed toward a communication failure.

Despite advances in machine learning, current systems lack robust reasoning capabilities and knowledge of the underlying fine-grained norms guiding social interaction beyond surface-level linguistic semantics. Importantly, failure to recognize such cultural knowledge can lead to miscommunication and conflict. Figure 1 exemplifies this issue through a conversation headed towards failure, in which the wife expresses *anger* when the husband violates the norm: *Earn trust with your partner by being truthful*. In better support of cross-cultural dialogue, advancing AI from a tool to a true communication partner demands progress in recognizing sociocultural norms, interpreting emotions, as well as detecting and remediating communication breakdowns before they occur.

In this paper, we present a system that utilizes multimodal technologies (speech, vision, and natural language processing) for discovering social norms specific to individual cultures, identifying when such social norms are violated in conversation, and recognizing emotional reactions and changes across conversation. We show how these can be used to predict impending conversational failure for conversations in six different languages. Specifically, our contributions include components for:

- *Norm detection* – We evaluate state-of-the-art approaches for discovering social norms and

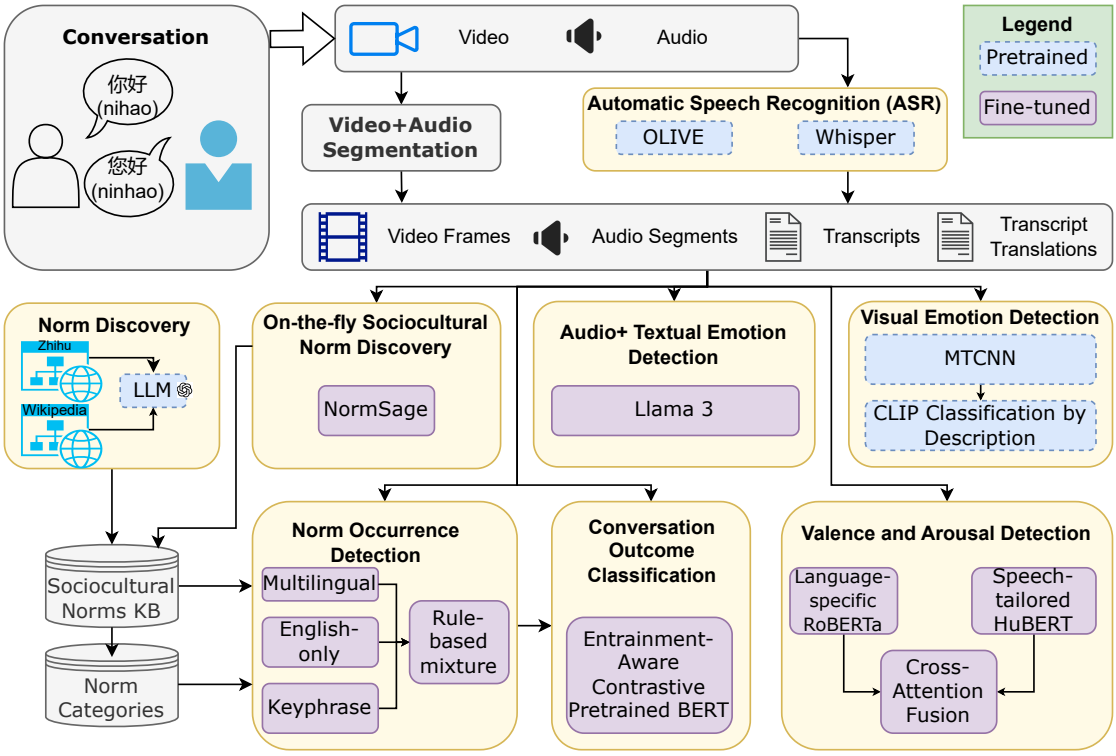

Figure 2: The system overview of emotion and norm-guided conversation success/failure profiling using CHARM.

grounding them to conversation instances, based on NormSAGE (Fung et al., 2023) and the pre-constructed multicultural norms library from CH-Wang et al. (2023); Fung et al. (2024). We further show their complementary nature in ensembling along with changepoint features, for the task.

- *Multi-modal emotion detection* – We achieve state-of-the-art results on the video emotion detection, and valence and arousal detection tasks.
- *Communication failure prediction* – We experiment with a variety of approaches to serve as a proxy for communication failure prediction, a novel task, and assess a contrastively pre-trained model for this task.

As broader impact, our integrated system and UI (CHARM) provide an AI-empowered approach facilitating practitioners to dynamically learn, interpret, and apply sociocultural norms, and enabling more seamless and effective cross-cultural communication across diverse real-world contexts. It helps overcome the prior challenge of nuanced cultural knowledge deeply embedded in societal practices and values being difficult to convey effectively in classroom settings, compared to language rules and vocabulary systematically taught.

2 System Overview of CHARM

The architecture of our multimedia conversation analysis system is illustrated in Figure 2.

Data Input As input, the system can take conversation videos, conversation audio only, or conversation text only. The current version of the system handles conversations in six languages (*e.g.,* Mandarin, Spanish, Russian, Japanese, Korean, Turkish).

Data Preprocessing Any audio data, including audio in videos, is first converted to text using the "OLIVE" speech recognition system (Lawson et al., 2016). OLIVE is optimized for real-time performance on standalone devices; it was chosen with the long term goal of real-time system deployment.

Communication Analysis Algorithmic Flow After data preprocessing, the system pipeline consists of a norm detection (ND) branch and an emotion detection (ED) branch (Sections 3, 4, respectively). Finally, we leverage the emotion and norm occurrence detection profiling to predict the conversation's outcome as a success or failure (Section 5).

3 Norm Detection

Determining whether social norms have been adhered to or violated during communication is an important first step for remediating human interaction conflicts. This task naturally decomposes into two sub-components: discovering applicable norm categories and content; then, grounding norm occurrences into specific conversations to detect when participants adhere to or violate norms. In this section, we start by defining the nature of norms. With a working knowledge of norms, we detail approaches to automatically discover social norm conventions by implicit context-aware knowledge elicitation of large language models while analyzing conversational data; and explicit norm extraction from information-rich web sources such as Wikipedia and Zhihu, a prominent QA platform amongst the Chinese community.

3.1 Norm Definition

For our purposes, we consider a norm to be some rule of behavior with the following components.

1. A context C in which the norm applies.
2. A task or goal X one wishes to perform.
3. A level of force F with which the rule applies, as indicated by factors such as the presence or absence of negation, and the use of adverbs of degree, such as 'never,' 'rarely,' 'sometimes,' or 'always'.
4. A normative behavior Y one should do to achieve X.

As a result, we propose a novel representation of norms as: *In Context [C], if one wants to do [X, task or goal], one should/not [F, force], do [Y, behavior].* For example, at the start of a business meeting [C], if one wants to develop friendly relations [X], one should [F] greet the other participants [Y]. Norms can be grouped by the tasks or goals X that they pertain to, in order to form norm categories, such as apologizing or performing criticism. Given only a segment of text, audio, or video, the norm occurrence detection task can be considered the problem of identifying the categories of applicable norms, and whether they are being adhered to or violated. We explore techniques which are able to do this both with and without prior knowledge of the annotated norm categories.

3.2 Evaluation Data

We use a multimedia dataset consisting of naturally occurring conversations from speakers sourced across the web called LDCCCU (LDC, 2022). The data contains 15 second segments which are annotated for adherence or violation to five predefined norm categories, summarized in Table 1. While the majority of annotations are for videos, there are a number of sources files which contain only audio or text. The corpus totals 902 unique conversations with ~2,300 annotations provided by native speakers.

Norm Category	Adhere	Violate
Apology	69	6
Criticism	682	141
Greeting	136	10
Request	733	41
Persuasion	466	8
Total	2086	206

Table 1: Summary of norm occurrence annotations.

3.3 Norms Discovery

We explore norm discovery from two perspectives: implicitly from conversation context, and explicitly from the Zhihu QA forum. In both cases, we leverage the strong zero-shot and few-shot reasoning capabilities of prompting some large language model M such as from the GPT-series (Brown et al., 2020a).

Conversation-based norms discovery We follow the NORMSAGE framework (Fung et al., 2022) to prompt M with a natural language input consisting of a dialogue context and a directed question describing the norm discovery task. The specific prompt template used here is: *Given the conversation context: What are social norms applicable to the situation?* This prompting approach discovers norms, along with their categories (e.g., request, persuasion, apology, etc.), for various sociocultural scenarios.

Web-based norms discovery Zhihu is one of China's largest online knowledge communities. Similar to Quora, Zhihu is primarily a Q&A site where users can post questions, answer existing questions, and up-vote answers. Questions on Zhihu are often posed to inquire about the appropriate action to take in a given situation, e.g., "Is it Ok to greet your boss by their first name?", with other users indicating their views on appropriate behaviors in their answers. Thus we use Zhihu as a proxy of explicitly stated norms. We use the site search API to search for Questions (situations) related to our norm categories. Then we utilize

few-shot prompting of M to extract social norms from top answers to the relevant questions and label those norms with the norm category.

In contrast to Zhihu, characterized by dynamic social discussion, Wikipedia reflects established knowledge. We further leverage this source for massively multicultural fine-grained cultural norm library construction, following Fung et al. (2024).

3.4 Norm Grounding on Conversations

After we have obtained a set of norms as a reference for how people typically behave in society, we identify instances of norms in conversations, which we call grounding. For this task, we identify three sets of models which together cover the majority of norm grounding instances: **Multilingual**, **Translate-to-English**, and **Keyphrase**.

Multilingual natural language inference Some cultures adhere to specific mannerisms or social norms that may not be universally relevant in a broader context. To encode these mannerisms, we finetune a RoBERTa (Liu et al., 2019) model pretrained for the natural language inference (NLI) task. Then, we further fine-tune it based on 15-second sliding windows of conversation transcripts labeled for norm grounding. Since this training data is few and far between, we cannot only rely on pretraining or fine-tuning to extract the norm grounding signal.

Translate-to-English For some other cultures and sociocultural norms, mannerisms are largely shared based on politeness or general conversational maxims. To extract signal in these general cases, and especially when no training data exists, we train a larger NLI model based on Llama-3 (Grattafiori et al., 2024) which operates on translations of the conversation transcripts, also using 15-second sliding windows of the conversation.

Keyphrase heuristics Finally, we also find that in some cases, there are obvious indications of norm grounding based on keyphrases. For example, most greetings or expressions of gratitude can be predicted with higher accuracy using symbolic rules than via the prior NLI models.

To combine predictions from these separate models, we found that each model had strengths in different language/norm category pairs. A simple logistical classifier trained on top of them yielded worse results than explicitly exposing a single prediction. Thus, for known norms, we determine which prediction based on development set performance, whereas unseen norms and languages use only the Translate-to-English model.

4 Emotion, Valence and Arousal Detection

Accurately detecting emotional indicators provides critical insight into the dynamics of conversational participants, provides valuable information for assessing the success of the interaction and interpreting social norms. To identify emotions in multimodal conversations, we explored a range of strategies for predicting both categorical (emotion labels) and continuous dimensions (valence and arousal) across modalities. Valence captures the positivity or negativity of an utterance, while arousal reflects the intensity of the emotional state.

4.1 Emotion Detection

To capture emotions across modalities, we divide the task into visual and textual+audio emotion recognition. The visual model extracts emotion from facial expressions and environmental context, while the textual and audio model infers emotion from the transcript's semantic content and prosodic features. Together, they provide a richer interpretation of conversational emotion.

Video Emotion Detection To predict emotional categories from video, we adopt a "classification by description" approach (Menon and Vondrick, 2022), leveraging the implicit social knowledge encoded in large language models (Brown et al., 2020b; Chowdhery et al., 2022; Zhang et al., 2022). Specifically, we prompt an expressive GPT-series (Brown et al., 2020a) large language model M to generate natural language descriptions of each emotion class, and then use CLIP (Radford et al., 2021) to ground these descriptions in video frames. For each emotion $e \in \mathcal{E}$, where $\mathcal{E}$ is the set of emotion classes, we query M with the following prompt:

```
Q: Visually describe a {e} person in a
    picture.
```

The model M returns a natural language description of each emotion's visual expression. For example, for emotion "anger," M generates the following:

```
Q: Visually describe an angry person in
    a picture.
A: An angry person may have their brows
    furrowed, their mouth set in a tight
    line, and their eyes narrowed.
```

We generate 5 diverse responses (temperature=0.9) for each emotion, including descriptions for groups. Once we have visual descriptions for an emotion e, we compute their CLIP text embeddings and average them to get the text-embedding for the emotion T_e. For each video frame f, we use MTCNN (Zhang et al., 2016) to detect faces, predicting 'none' if none found. Otherwise, we compute the CLIP visual-embedding V_f and predict emotion $\hat{e} = \arg\max_{e \in \mathcal{E}} \phi(V_f, T_e)$, where ϕ is cosine similarity.

Text and Audio Emotion Detection Within our integrated cross-cultural communication system, SpeechCueLLM (Wu et al., 2024a) serves as the bridge between audio and textual analysis for emotion detection. It is a novel method that overcomes the limitations of traditional large language models (LLMs) in processing audio by converting key speech characteristics into natural language descriptions. Instead of relying on additional audio encoders or architectural changes, SpeechCueLLM extracts intuitive audio features—such as volume, pitch, and speaking rate—and translates these numerical cues into descriptive text. This transformation enables the LLM to perform multimodal emotion recognition by seamlessly integrating both verbal content and vocal nuances within a unified text-based prompt.

4.2 Valence and Arousal Detection

We employ the Multimodal Multi-loss Fusion Network (MMML) (Wu et al., 2024b) for valence and arousal detection. It integrates text and audio modalities to enhance sentiment analysis. The system employs two core components: a Feature Network and a Fusion Network. The Feature Network uses pre-trained models (RoBERTa (Cui et al., 2019) for text and HuBERT (Hsu et al., 2021) for audio) to extract high-quality features, avoiding hand-crafted methods. The Fusion Network combines modalities through cross-attention (projecting one modality into another's space) and self-attention (capturing intra-modality dynamics), followed by feed-forward layers to refine representations. Multi-loss training is applied, where each modality subnet and the fused output have separate losses, encouraging specialized learning while promoting collaboration. Additionally, context modeling processes historical utterances independently, improving performance by leveraging sequential dependencies.

5 Conversation Outcome Classification

In this section, we present interpretable methods to classify conversation outcomes as *success* or *failure*, which is challenging due to limited labeled data.

Norms Social norm adherence may affect conversation outcome; conversations where social norms are adhered to are more likely to end with success, while conversations where social norms are violated are more likely to end with failure. We use social norm violation and adherence predictions to classify communication outcome. Specifically, we encode norms category and adherence/violations detected in Section 3 along with the transcript, and then train the model to predict change point using entrainment (Chen et al., 2023).

Emotion Surprisal We further hypothesize that derailment may more likely happen when the emotion in the most recent response is unexpected given the previous utterances. The key of our emotional surprisal method is a next-turn emotion predictor, which predicts the emotion of the next utterance based on the emotions of the dialog history. At test time, to predict whether a conversation is success or failure at turn T, we first use the emotion model built in Section 4.1 to identify the emotions $\{e_t\}_{t=1}^{T}$, and calculate emotion surprisal at turn T as $-\log p(e_T|\mathbf{e}_{t<T})$, where p denotes the next-turn emotion predictor. We fine-tune a linear model to predict communication outcome using the *1-dimensional* emotion surprisal score as the input.

6 Experiments and Results

Table 2 shows the performance of each component.

6.1 Norms Detection Evaluation

We compare language model prompting and finetuning approaches for norms detection on the LDC-CCU corpus. Given the scarcity of training annotations, NORMSAGE prompting with self-verification performs better than MACBERT (Cui et al., 2020) finetuned on LDCCCU textual input content and training labels only. The best performance is achieved by using an ensemble classifier that combines norm predictions from both zero-shot and finetuned approaches, with multi-modal emotion features. The results demonstrate the complementary nature of these sub-components in improving our reasoning of norm occurrences in conversation.

Component	Benchmark	Method	Score	Metrics
Norm Detection	LDCCCU	MACBERT	25.0	AUC
		MACBERT$^{+\text{Zhihu aug}}$	25.2	
		NORMSAGE	30.0	
		ensembling	**30.8**	
Video Emotion Detection	EMOTIC	Baseline	26.64	mAP
		Ours	**40.97**	
Audio+Text Emotion Detection	IEMOCAP	Text Baseline	70.11	mAP
		Ours	**72.6**	
Valence & Arousal Detection	CH-SIMS	Baseline	(80.1, 80.1, 39.6)	(Acc_2, F1 MAE)
		Ours	(**82.93**, **82.9**, **33.2**)	
	CMU-MOSEI	UniMSE (Hu et al., 2022)	(87.5, 87.46, 52.3)	
		Ours	(**88.02**, **88.15**, **49.22**)	
Conversation Outcome	CGA	End-to-End Baseline (100)	52.7	Acc
		Emotion Surprisal (2508)	60.1	
		Norms (2508)	**68.0**	

Table 2: Performance of each component on the listed benchmarks.

6.2 Emotion, Valence, and Arousal Evaluation

Emotion *For video,* we evaluate our method on the EMOTIC dataset (Kosti et al., 2020), focusing on the 7 Plutchik emotions represented in the data. Our approach achieves a mean average precision (mAP) of **40.97**, outperforming the baseline mAP by 14.33 points. *For text and audio,* evaluated on the popular benchmark dataset IEMOCAP (Busso et al., 2008), SpeechCueLLM demonstrates significant performance improvements. Under LoRA fine-tuning, it surpasses both the text-only and projection-based approaches by 2-3 F1 points. These results highlight the approach's robustness and effectiveness, especially when high-quality audio is available.

Valence and Arousal Our model outperformed the best previous models in the CH-SIMS (Yu et al., 2020b) and CMU-MOSEI (Bagher Zadeh et al., 2018) data sets for all metrics, including the two-class classification accuracy, the F1 and the mean absolute error (MAE) metrics.

6.3 Conversation Outcome Evaluation

We evaluate our entrainment contrastive pre-training model on the original Wikipedia Conversations Gone Awry (CGA) dataset (Zhang et al., 2018), where the task is to make an early prediction that a conversation will ultimately fail. CGA is a balanced dataset with 2508 training examples and 840 validation and test examples each. In evaluation, each conversation is fed to the model by adding one turn at a time, where a prediction can be made for every state of the conversation. If failure is predicted at any state of the conversation, the entire conversation is predicted to be failure.

End-to-End Baseline We compare our methods to an end-to-end baseline that directly predicts communication outcome based on conversation inputs.

Norms The logistic regression model trained on norm violation and adherence counts achieves an accuracy of **68.0%**, which outperforms the end-to-end baseline of 65.8% while being significantly more interpretable. Results indicate that norm violations and adherence are indeed predictive of conversation outcomes.

7 Conclusions and Future Work

We present CHARM, a novel multi-modal communication assistance system for understanding cross-culture nuances and promoting harmonious human interaction. This system empowers users with novel profiling of social norm adherence/violations; emotional, valence, and arousal characterizations; as well as outcome classification, in conversations. We are investigating ways to further fuse different modalities in our components; for example, prosody, facial expression or gesture can be useful for norm and communication change prediction. We plan to study remediation actions, such as alternative dialogue utterance recommendation, in future work.

Ethical Considerations

We present a modular system which is intended to explore a novel NLP task: identifying norm violations, emotions, and communication failures in real time in cross-cultural conversations. Our system is a research prototype and produces results

that are sometimes incorrect, as we show in the results section. Given the potentially negative effects of incorrect communication (e.g.,angering or shaming discourse partners, early termination of dialog, miscommunication, discord, conflict, etc.), we do not intend for our system to be used as-is. Instead, we envision the initial use cases will be controlled laboratory experiments with human subjects who have been instructed about the system and its weaknesses. The system should only be used outside of the lab once its end-to-end effectiveness has been tested in realistic scenarios and should only be used with proper warnings to the users about its remaining possible shortcomings. While we intend to make the system available, we require downloaders to specifically acknowledge these limitations of the system and to commit to using it only for research and not as a deployed system. For future work, we intend to perform more extensive analysis on model biases that go beyond the implementation of prediction ensembling and correctness verification mechanisms, explored here and in Fung et al. (2022), respectively.

References

2022. CCU TA1 Mandarin/Chinese Development Annotation LDC2022E18. Web Download.

Muhammad Abdul-Mageed and Lyle Ungar. 2017. EmoNet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 718–728, Vancouver, Canada. Association for Computational Linguistics.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, Melbourne, Australia. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020a. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language Models are Few-Shot Learners. *arXiv:2005.14165 [cs]*. ArXiv: 2005.14165.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: interactive emotional dyadic motion capture database. *Language Resources and Evaluation*, 42:335–359.

Sky CH-Wang, Arkadiy Saakyan, Oliver Li, Zhou Yu, and Smaranda Muresan. 2023. Sociocultural norm similarities and differences via situational alignment and explainable textual entailment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3548–3564, Singapore. Association for Computational Linguistics.

Kushal Chawla, Jaysa Ramirez, Rene Clever, Gale Lucas, Jonathan May, and Jonathan Gratch. 2021. CaSiNo: A corpus of campsite negotiation dialogues for automatic negotiation systems. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3167–3185, Online. Association for Computational Linguistics.

Run Chen, Seokhwan Kim, Alexandros Papangelis, Julia Hirschberg, Yang Liu, and Dilek Hakkani-Tür. 2023. Identifying entrainment in task-oriented conversations. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *arXiv preprint arXiv:2004.13922*.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Ziqing Yang, Shijin Wang, and Guoping Hu. 2019. Pre-training with whole word masking for chinese bert. *arXiv preprint arXiv:1906.08101*.

Denis Emelin, Ronan Le Bras, Jena D Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718.

Maxwell Forbes, Jena D Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral

norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670.

Yi Fung, Tuhin Chakrabarty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2023. NORMSAGE: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15217–15230, Singapore. Association for Computational Linguistics.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition lm benchmarking.

Yi R Fung, Tuhin Chakraborty, Hao Guo, Owen Rambow, Smaranda Muresan, and Heng Ji. 2022. Normsage: Multi-lingual multi-cultural norm discovery from conversations on-the-fly. *arXiv preprint arXiv:2210.08604*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna

Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. 2024. The llama 3 herd of models.

Kun Han, Dong Yu, and Ivan Tashev. 2014. Speech emotion recognition using deep neural network and extreme learning machine. In *Interspeech 2014*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2020. Aligning ai with shared human values. In *International Conference on Learning Representations*.

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. *arXiv preprint arXiv:2106.07447*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. Unimse: Towards unified multimodal sentiment analysis and emotion recognition.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Ronak Kosti, Jose M Alvarez, Adria Recasens, and Agata Lapedriza. 2020. Context based emotion recognition using emotic dataset. *arXiv preprint arXiv:2003.13401*.

Aaron Lawson, Mitchell McLaren, Harry Bratt, Martin Graciarena, Horacio Franco, Christopher George, Allen R Stauffer, Chris Bartels, and Julien van Hout. 2016. Open language interface for voice exploitation (olive). In *INTERSPEECH*, pages 377–378.

Sophie Legros and Beniamino Cislaghi. 2020. Mapping the social-norms literature: An overview of reviews. *Perspectives on Psychological Science*, 15(1):62–80.

Diane Litman, Susannah Paletz, Zahra Rahimi, Stefani Allegretti, and Caitlin Rice. 2016. The teams corpus and entrainment in multi-party spoken dialogues. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1421–1431, Austin, Texas. Association for Computational Linguistics.

Ping Liu, Shizhong Han, Zibo Meng, and Yan Tong. 2014. Facial expression recognition via a boosted deep belief network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Sachit Menon and Carl Vondrick. 2022. Visual classification via description from large language models.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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. *arXiv:2103.00020 [cs]*. ArXiv: 2103.00020.

Guangyao Shen, Xin Wang, Xuguang Duan, Hongzhi Li, and Wenwu Zhu. 2020. Memor: A dataset for multimodal emotion reasoning in videos. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 493–502, New York, NY, USA. Association for Computing Machinery.

WHO. 2009. Changing cultural and social norms supportive of violent behaviour.

Zehui Wu, Ziwei Gong, Lin Ai, Pengyuan Shi, Kaan Donbekci, and Julia Hirschberg. 2024a. Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances.

Zehui Wu, Ziwei Gong, Jaywon Koo, and Julia Hirschberg. 2024b. Multimodal multi-loss fusion network for sentiment analysis.

Mingzhi Yu, Diane Litman, Shuang Ma, and Jian Wu. 2021. A neural network-based linguistic similarity measure for entrainment in conversations. *arXiv preprint arXiv:2109.01924*.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020a. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. 2020b. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotations of modality. *Association for Computational Linguistics*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.

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

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022a. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Caleb Ziems, Jane A Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022b. The moral integrity corpus: A benchmark for ethical dialogue systems. *arXiv preprint arXiv:2204.03021*.

A Related Work (Expanded)

Moral Assistance System There are multiple papers in building systems capable of providing moral suggestions and explanations, both for dialogues (Ziems et al., 2022b; Fung et al., 2022) and for general texts describing daily situations like Social-Chem-101 (Forbes et al., 2020) Moral Stories (Emelin et al., 2021), ETHICS (Hendrycks et al., 2020) and Delphi (Jiang et al., 2021). However, prior efforts mainly focus on textual data and overlook other conversational styles (*e.g.,* emotion, wording, tactic, etc.) which play an important role in communication success/failure. In our work, we also take speaker visual emotions and speech valence and arousal into consideration.

Norm Discovery Our goal of detecting norm occurrences is also related to the general problem of norms discovery. Social-Chemistry-101 (Forbes et al., 2020) and MIC (Ziems et al., 2022a) are dedicated to constructing static norm libraries from curated English data such as Reddit posts. Their data annotation process, however, tends to be labor-intensive and expensive. Our system is inspired by NormSage (Fung et al., 2022), which alternatively proposed an automatic norm discovery method by prompting large pre-trained language models.

Human Emotion Detection Our visual emotion detection component is also related to recent work in detecting human emotions and sentiments. Researchers have approached this task through natural language (Abdul-Mageed and Ungar, 2017), speech (Han et al., 2014), and visual data (Liu et al., 2014). Recently, there has been huge progress in emotion detection in multi-modal settings (Shen et al., 2020; Yu et al., 2020a; Poria et al., 2019).

Communication Outcome Classification Communication outcome classification is a relatively new area of research. Wikipedia Conversations Gone Awry (CGA) is a dataset labeled with binary conversation outcomes (success/failure) where the task is to make an early prediction that a conversation will ultimately fail (Zhang et al., 2018). Another communication outcome dataset is CaSiNo, which is a corpus of collected campsite negotiation dialogs annotated with negotiation outcomes (Chawla et al., 2021). Our work on entrainment-aware contrastive pre-training is inspired by Yu et al. (2021), who identified that pre-training with response matching produces model scores correlated with entrainment measures (Litman et al., 2016).