Characterizing Similarities and Divergences in Conversational Tones in Humans and LLMs by Sampling with People

Dun-Ming Huang[1, 2], **Pol van Rijn**[2], **Ilia Sucholutsky**[3], **Raja Marjieh**[4], and **Nori Jacoby**[2]

[1]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley
[2]Computational Auditory Perception Group, Max Planck Institute for Empirical Aesthetics
[3]Department of Computer Science, Princeton University
[4]Department of Psychology, Princeton University

Abstract

Conversational tones – the manners and attitudes in which speakers communicate – are essential to effective communication. Amidst the increasing popularization of Large Language Models (LLMs) over recent years, it becomes necessary to characterize the divergences in their conversational tones relative to humans. However, existing investigations of conversational modalities rely on pre-existing taxonomies or text corpora, which suffer from experimenter bias and may not be representative of real-world distributions for the studies' psycholinguistic domains. Inspired by methods from cognitive science, we propose an iterative method for simultaneously eliciting conversational tones and sentences, where participants alternate between two tasks: (1) one participant identifies the tone of a given sentence and (2) a different participant generates a sentence based on that tone. We run 100 iterations of this process with human participants and GPT-4, then obtain a dataset of sentences and frequent conversational tones. In an additional experiment, humans and GPT-4 annotated all sentences with all tones. With data from 1,339 human participants, 33,370 human judgments, and 29,900 GPT-4 queries, we show how our approach can be used to create an interpretable geometric representation of relations between conversational tones in humans and GPT-4. This work demonstrates how combining ideas from machine learning and cognitive science can address challenges in human-computer interactions.

1 Introduction

Conversational tones, the manner and attitude in which a speaker communicates, is essential to human communication (Yeomans et al., 2022; Saiewitz and Kida, 2018). Effective communication relies on people's understanding of conversational patterns and tones, and their ability to promptly react to them (Kreuz and Roberts, 1993; van Schendel and Cuijpers, 2015). Inability to do so results in the content of conversation being "lost-in-translation" between speakers of different languages and cultures (Yusifova, 2018). Notably, while traditionally the study of conversational tones involved only humans, the increasing prevalence of Large Language Models (LLMs) in everyday decision-making, especially their conversational ("Chat") variants, renders the study of conversational tones in LLMs necessary for human alignment (Ouyang et al., 2022; Rudolph et al., 2023; Sucholutsky et al., 2023b; Marjieh et al., 2023a). Developing tools for effectively characterizing conversational tones in humans and LLMs is hence essential for the development of human-centered AI, human-computer interaction research, and cognitive sciences (Figure 1A).

Background: Conversation Research. In conversation research, some literature engages explicitly with the composition and usage of conversational attitudes and linguistic markers (Fintel, 2006; Yeomans et al., 2022; Jakobson, 1960). A wider array of literature uses conversational analysis to investigate other dynamics that can affect the content of conversation, such as turn-taking (van Schendel and Cuijpers, 2015) and face-saving (Ting-Toomey et al., 1991; Oetzel et al., 2001), as well as other cross-cultural semantic differences that can lead to different behavior within the same conversational tone, such as refusal (Chang, 2009), shame (Kollareth et al., 2018), and politeness (Alemán Carreón et al., 2021; Ogiermann, 2009). On the other hand, the introduction of Large Language Models, especially its chatbot applications, brings attention to the alignment of conversational behavior in LLMs with that of human ideals (Ouyang et al., 2022; Rafailov et al., 2023), which can potentially contribute to the alignment of LLMs' perception and production of conversational tones with those of humans. Being able to effectively fine-tune LLMs also creates new opportunities to generate text style-

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 10486–10512

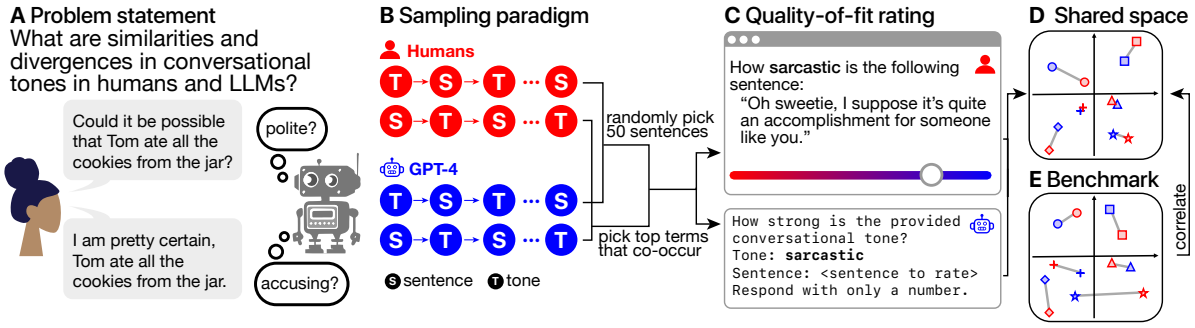


Figure 1: Summary of our approach. **A**: Problem statement. **B**: The Sampling with People paradigm that aims to collect a representative sample of conversational tones and sentences. **C**: A quality-of-fit rating procedure that allows us to obtain vector representations of conversational tones with respect to their usage context. **D**: A geometric representation of the shared embedding space across elicited domains (human, GPT). **E**: As an application of our obtained data, we benchmark a selection of popular unsupervised cross-domain alignment methods.

transfer corpora that specify sentences with specific conversational tones, such as politeness (Wang et al., 2022) and formality (de Rivero et al., 2023; Wang et al., 2019).

Challenge: biased apriori taxonomy. However, the domain of conversational tones is, like emotion (Schiller et al., 2023; Lindquist et al., 2022; Athanasiadou and Tabakowska, 1998) and color (Berlin and Kay, 1969), an instance of grounded semantics (Tannen, 1984; Semnani-Azad and Adair, 2013). While all participants observe the same stimulus (e.g., an emotional recording, a solid color, or in our case a sentence), people may use different words or labels to describe it (in our case conversational tones such as "polite", "excited", and "grateful") which makes it difficult to study grounded semantics at scale and especially across multiple languages or cultures. One challenge is that studying grounded semantics often involves adopting a predefined taxonomy, typically sourced from previous studies and curated by investigators (e.g.; colors (Adams and Osgood, 1973; Wang and Wang, 2016); facial emotion (Ekman, 1992); musical emotion perception (Juslin and Västfjäll, 2008; Palmer et al., 2013), concepts such as animal terms (Marti et al., 2023); sentiment of news items (Rozado et al., 2022) prosody (Sauter et al., 2010; Busso et al., 2008). Notably, many machine learning datasets also suffer from the same limitation of using a predefined list of using a predefined list of stimuli that can be outdated or unrepresentative of the correspond modality. Examples of such datasets span through realms of: object images (Deng et al., 2009; Krizhevsky, 2009), visual scenes (Zhou et al., 2017), sounds (Gemmeke et al., 2017), video and its categorizations (Kay et al., 2017), facial expressions (Goodfellow et al., 2013). This strategy is prone to researcher bias, potentially skewing the findings away from an accurate representation of labels as they occur in the real world and within a given culture (Kollareth et al., 2018; Henrich et al., 2010).

Challenge: biased stimulus set. Another challenge that almost all studies faced when studying grounded semantics is that they may use a constrained set of stimuli to be annotated (e.g. emotion (Cowen and Keltner, 2017; Cowen et al., 2019, 2020; Cowen and Keltner, 2020; Cowen et al., 2018); object recognition and similarities (Gifford et al., 2022; Hebart et al., 2019); word-associations (De Deyne et al., under review; De Deyne et al., 2019); musical perception (Juslin and Sloboda, 2013); facial expression (Khaireddin and Chen, 2021; Lin et al., 2021); prosody (El Ayadi et al., 2011; Batliner et al., 2008)). This introduces researcher bias, as curating the stimuli may influence the elicited labels, which we outline using the following example. Imagine an experiment where a particular semantic term can be associated with some class of objects (e.g., the term "red" can be used to describe red fruits). If the object class is not included in the predetermined list of objects (e.g., red fruits are not included in the list of objects), then the elicited terms will not include this association (we will conclude that "red" does not describe fruits), and it will be missing from the resulting semantic network. Bias in object selection can also occur in more subtle ways where a skew in the distribution of selected objects also skews the distribution of elicited terms, potentially even amplifying the initial bias. Furthermore, a large body of cross-cultural researchers suggests that studies

should not impose a terminology inherited from the experimenter or even from one group of studied agents (e.g., English speakers) on another agent or group of agents (e.g., Speakers of another language or demographics; Blasi et al., 2022; Barrett, 2020; Henrich et al., 2010).

Additional challenges. Finally, while some studies advocate for the exclusive use of large textual corpora and the extraction of semantic descriptors via data mining (Thompson et al., 2020), this indirect approach raises concerns about its ability to accurately represent the nuances of conversational tones as experienced in everyday human interactions. It is also difficult to rigorously compare humans and LLMs using such corpora because these same textual corpora are also the basis for LLM training. Notably, (Thompson et al., 2020) also studies the problem of aligning semantic networks of different individuals or groups in the context of cross-linguistic and cross-cultural comparisons. It turns out that this is a key part of the machine learning problem of automatic translation (Zinszer et al., 2016; Liu et al., 2021). Recent research has focused on aligning semantics in humans with Large Language Models (Sucholutsky et al., 2023b; Atari et al., 2023), with significant applications to designing human-computer interfaces (Hou et al., 2024) and AI safety.

Our approach. In light of these challenges, we propose a method that enables the characterization of conversational tones and their taxonomies in any target human population as well as LLMs, based on a human-in-the-loop Sampling with People (SP) technique (Sanborn et al., 2010; Griffiths and Kalish, 2005; Harrison et al., 2020) (Figure 1). Specifically, we propose an iterative procedure in which humans and LLMs are presented with sentences and are asked to label their conversational tones in an open-ended fashion (Figure 1B). The resulting conversational-tone terms are then presented to a new group of agents who are asked to produce sentences reflecting those conversational tones. This process is then repeated multiple times. With mathematical formalism, this process instantiates a Gibbs Sampler from the joint distribution of sentences and conversational tones in humans and LLMs (Harrison et al., 2020; Griffiths et al., 2024). Given the resulting sample, we derive representative sentences and tone taxonomies of our target population, then have an independent group of human evaluators and LLMs rate the extent to which each tone matched each sentence (Quality-of-fit Rating; Figure 1C). We use these to construct a geometric embedding that can be used to evaluate the alignment between human and LLM conversational tones (Figure 1D).

We show how our approach can be effectively used to reveal divergences in the representation of conversational tones between humans and LLMs. Moreover, we demonstrate how our new dataset and cross-evaluations can be used to benchmark unsupervised cross-domain semantic alignment methods used in existing work, and identify which of these work well for cases in which cross-evaluation is not possible (e.g., in multilingual scenarios; Figure 1E). Our method can be generalized to many more psycholinguistic modalities (e.g., sentiment, color), languages, and cultures beyond those involved in this paper. We believe it will help advance both human-machine alignment research as well as cross-cultural research.

2 Detailed Approach

2.1 Elicitation via Sampling with People

The core of our approach is the joint elicitation of a representative sample of conversational tones and sentences from both humans and LLMs. Specifically, we propose an iterative procedure that is composed of two steps per iteration. Step one, in which humans and LLMs are presented with sentences and are asked to classify their conversational tones in an open-ended fashion (Figure 1B). Step two, the resulting conversational tone descriptors (adjectives) are then presented to a new group of agents, from whom we ask to produce sentences that reflect those conversational tones. This process is then iterated multiple times. Formally, this process instantiates a Gibbs Sampler from the joint distribution of sentences and conversational tones, for any target population we choose to sample from, be it humans and LLMs (Griffiths et al., 2024). Therefore, by constraining the set of human participants to those from a specific cultural group, we expect to obtain a representative sample of psycholinguistic contents from the group of our participant population.

Here we draw inspiration from human-in-the-loop elicitation procedures (Griffiths et al., 2024). In these methods such as serial reproduction (Xu and Griffiths, 2010; Anglada-Tort et al., 2023; Jacoby and McDermott, 2017; Jacoby et al., 2024; Langlois et al., 2017, 2021), iterated learning (Xu et al., 2010; Griffiths and Kalish, 2005),

MCMCP (Sanborn et al., 2010) and GSP (Harrison et al., 2020; van Rijn et al., 2022b; Van Rijn et al., 2021; Van Geert and Jacoby, 2024; Marjieh et al., 2024; van Rijn et al., 2024), a Markov Chain is constructed by interspersing human decisions within a sampling chain to characterize latent representations in the human mind (e.g., perceptual prior, or subjective utility). In our novel approach, Sampling with People (SP), humans are recruited to perform two tasks (Figure 1B): (1) elicit a sentence based on a conversational tone ("S" task), and (2) annotate a conversational tone of a given sentence ("T" task). Under a probability theory framework, "S" and "T" tasks are essentially sampling operations, respectively from the conditional distribution of a sentence given a conversational tone $p(S|T)$, and the conditional distribution of a tone given a sentence $p(T|S)$. In practice, we run several parallel sampling chains, and in each trial, a participant is assigned to one chain and performs an "S" or "T" task as needed. This means that each sampling chain alternates between "S" and "T" tasks. Importantly, to satisfy the formulation of a Gibbs' Sampler, we design our paradigm to satisfy the Markovian property by constraining each participant to see only the output of the previous iteration (Harrison et al., 2020; Griffiths and Kalish, 2005).

Using SP, we elicited a large database of tones and sentences from humans and LLMs (separately for each). We elicited 40 tones and 80 sentences with 955 participants, 90 chains, and 100 iterations each. The list of 40 conversational tones we investigate is the union of the top 24 conversational tones from each instance of SP experiments. We then took 40 random sentences from each of the humans and GPT, forming a balanced corpus of humans and GPT in terms of sentence sources. We preserve the representativeness of our sample from the internal distribution of sentences $P(S)$ by choosing the random sentences in a uniform sampling fashion. In the design of this study, we use a shared array of conversational tones to work with a consensus taxonomy, enabling direct comparability between the conversational behavior of humans and GPT-4 when prompted under the same language[1].

2.2 Annotation via Quality-of-fit Rating

Given the distributions from prior subsection, we have derived representative sentences and tone taxonomies. Then, we have an group of human evaluators independent from prior participants, as well as GPT rate the extent to which each tone matched each sentence (Quality-of-fit Rating; Figure 1B). We show how our approach can be effectively used to reveal divergences in the representation of conversational tones between humans and LLMs. Specifically, after SP sampling, we collect quality-of-fit ratings of all sentences with all tones (Figure 1C), and compute semantic similarity matrices of different tones. Then, for two tones t_i, t_j, let $\mathcal{R}_{i,j}$ denote the correlation of the two tones across the vector of average ratings of all sentences, such that t_i and t_j are similar if they have similar ratings across all sentences. We can compute such matrix $\mathcal{R}$ in either an intra-domain manner (only using embeddings from either humans or GPT but not both), or a cross-domain manner (where we exploit the shared list of sentences to compute the correlation between all conversational tone embeddings).

2.3 Geometric Representation of Conversational Tones

We use the resulting cross-domain similarity matrices to obtain a geometric representation of both human and GPT data within the same space (Figure 1D). We compute the full correlation of all 80 tone embeddings (40 conversational tones, each one with an embedding from human data and another from GPT data), and use Multidimensional Scaling (MDS; Carroll and Arabie, 1998; Anowar et al., 2021) to project them into a shared low-dimensional embedding space. The space thus represents not only the relation between tones within humans but also the way they relate to GPT, especially as the proximity of tones in the MDS Euclidean space corresponds to the proximity of tones in terms of their semantic similarity (Shepard, 1980). Therefore, tones that appear closer in the shared space are located nearer in Euclidean space.

2.4 Application: Benchmarking Semantic Alignment Methods

To demonstrate the usability of our alignment data, we show how it can be used to benchmark semantic alignment methods as ground truth when cross-annotation is unavailable (Figure 1E) and only intra-domain correlation matrices are used. Specifically, we benchmark the performances of (1) Gromov-Wasserstein Optimal Transport (Grave et al., 2018; Conneau et al., 2017; Kawakita et al., 2023), (2) Bilingual Lexicon Induction (Ruder

[1]From here on, all references to GPT-4 will be abbreviated as GPT.

et al., 2018; Artetxe et al., 2016, 2017, 2018a,b), and (3) Orthogonal Procrustes (Schönemann, 1966; Beauducel, 2018).

3 Method

3.1 Participants

Human Participants. We recruited N=1,339 participants for the four human experiments in this study (Appendix Table 1 provides the number of participants and responses for each experiment). Participants were recruited from the recruiting platform Prolific and provided informed consent under an approved protocol (see Ethics section for further information). Human experiments are implemented using Psynet (Harrison et al., 2020), a Python package for implementing complex online psychology experiments.

GPT experiments We used the June 13th, 2023 release of GPT-4. Overall, we ran 29,900 GPT queries across all experiments.

3.2 General Procedure

Each human experiment began with detailed instructions and practice trials. We emulated each of the human experiments with GPT-4 agents that used the very same procedure, using the experiment interface instructions as LLM prompts. The human and GPT-4 experiments were otherwise identical in their design. Appendix A contains the full instructions/prompts used in our experiments. All the data of the experiments, code for reproducible human and GPT experiments, and analysis script can be found here: `https://github.com/jacobyn/SamplingTonesACL`.

4 Results

4.1 Elicitation (Sampling with People)

Figure 2A shows the histogram of the 24 most popular conversational tones from 4,500 human and 4,500 GPT annotated sentences. We recruited 955 human participants for SP experiments. The collected histograms were reliable: the split-half reliability computed via bootstrapping[2] was high for both humans and GPT conversational tones (humans: r = 0.91, CI = [0.87, 0.93]; GPT: r = 0.87 CI = [0.73, 0.94][3]). The GPT histogram (Figure 2A) was much more concentrated (entropy of 3.10 bits CI = [3.06, 3.15] via bootstrapping) compared to that of humans (entropy of 5.48 bits CI = [5.43, 5.52]). Overall, there was some similarity between the histograms (r = 0.39 p = 0.006 CI = [0.3,0.454]), but also significant differences. Specifically, the prominent conversational tones had different weights: in humans the three most prominent conversational tones were "grateful" (mean 4.03% CI = [3.45%, 4.62%]), "excited" (2.79% CI = [2.35%, 3.28%]), then "happy" (2.64% CI = [2.18%, 3.1%]) whereas in GPT they were "excited" (mean 24.66% CI = [23.41%, 25.91%]), "grateful" (10.79% CI = [9.89%, 11.68%]), then "concerned" (6.79% CI = [6.05%, 7.52%]). The results highlight that the elicitation process results in a different distribution of terms used to describe conversational tones.

4.2 Annotation via Quality-of-fit Rating

To create a detailed semantic embedding based on the given sentences and tones, we conducted a further experiment involving both humans and GPT. In this study, participants evaluated all target sentences using a predefined list of 40 prominent conversational tones that were extracted from the terms elicited using the SP procedure above, ensuring a direct comparison by employing the same tones for both human and GPT assessments. As a result, every sentence is relabelled with all 40 tones regardless of the original tone it was assigned in the SRE step. The sentence set was evenly divided, with one half originating from humans and the other half generated by GPT.

We recruited an additional 275 human participants for these annotations. Participants rate the strength of conversational tones in a sentence using a Likert scale, resulting in 16,000 rating judgments. Likert scales were used because the degree to which a sentence represents a tone may vary and cannot be captured by categorical labels (Sucholutsky et al., 2023a). GPT data underwent the same rating process with GPT agents as raters. We then analyzed the correlation between tones by examining the rating vectors across the 80 sentences elicited from Section 4.1. Additionally, we find that the majority of sentences are labeled to possess at least one specific conversational tone: all human-originated sentences had at least one conversational tone with an average rating exceeding 2.89, and exceeding a rating of 2 for 95% of GPT-originated sentences. This would suggest that very few sentences were

[2]Throughout the paper, bootstrapping occurs with 5000 repetitions, unless specified otherwise.

[3]Throughout the paper all CI are reported as 95%.

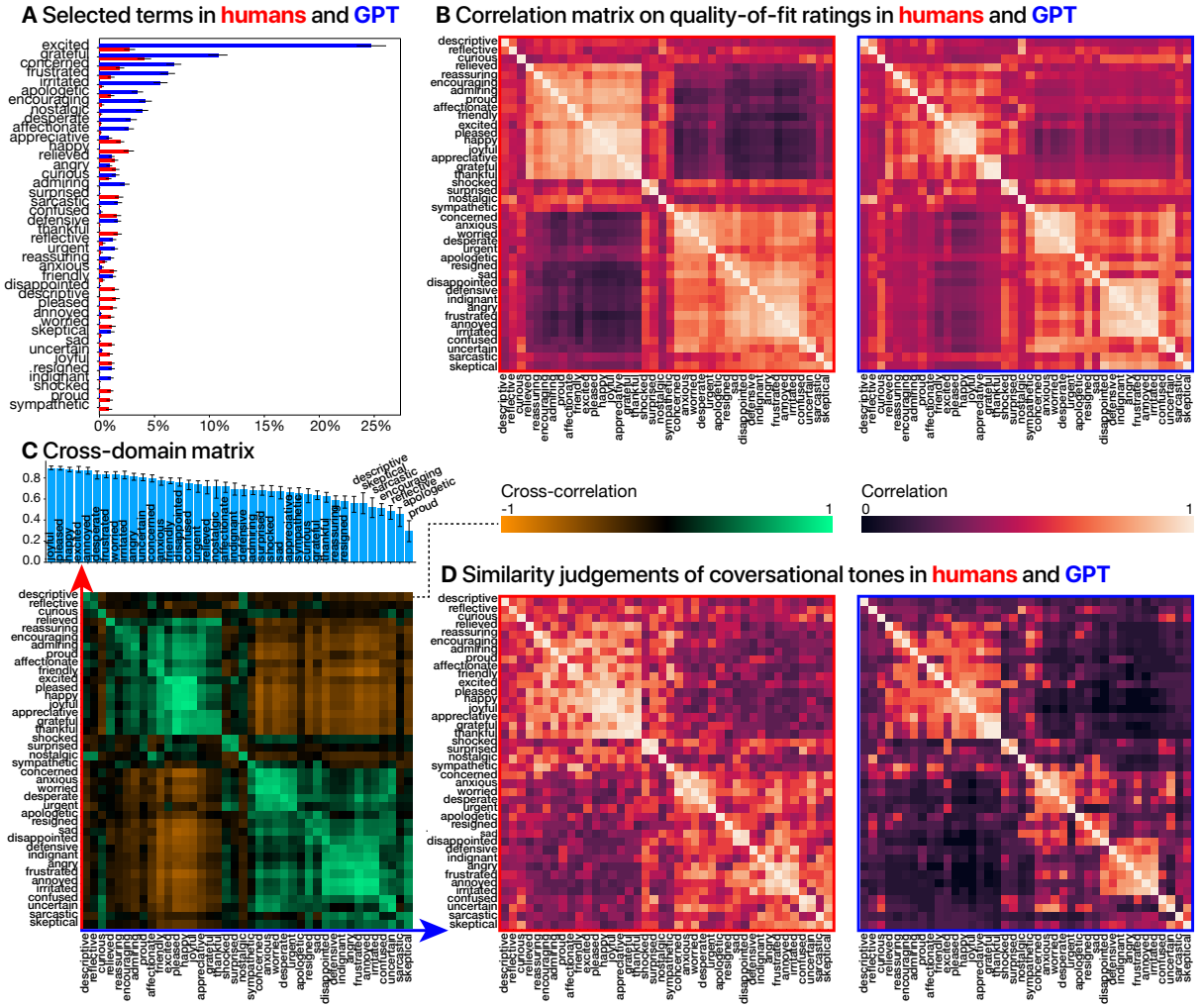


Figure 2: Results of Sampling with People and Quality-of-fit Rating paradigms and comparison to similarity judgments paradigm. **A**: Selection of most popular conversational tones from each of human and GPT instances, and their frequencies in respective samples (red for humans, blue for GPT). Error bars represent one standard deviation via bootstrapping. **B**: Correlation matrices of conversational tone quality-of-fit rating embeddings within humans (on the left) and within GPT (on the right). **C**: Cross-domain (Cross-correlation) matrix of human rating embeddings and GPT rating embeddings for conversational tones, and a bar plot showing the correlation between human ratings and GPT rating embeddings for each conversational tone word. Error bars represent one standard deviation via bootstrapping. **D**: Similarity judgment-derived similarity matrices of conversational tone from humans (on the left) and GPT (on the right). See enlarged version of this figure in the Appendix (Figure 17).

perceived as completely neutral with respect to all 40 tones.

Figure 2B presents the tone-correlation matrices for humans (left) and GPT (right). The matrices show reliable tone-similarity for both humans and GPT (humans: $r = 0.94$ CI [0.91, 0.96]; GPT ratings, $r = 0.86$ CI [0.78, 0.91]). It is visually apparent that there are two clusters of tones which roughly related to the valence of tones and that this structure is preserved in humans and GPT as we found a high correlation between the upper triangle of the two matrices ($r = 0.81$ CI [0.76, 0.85]).

To better understand how humans and GPT are similar and different in the structure of tone-similarity we now compute the cross-domain matrix, namely for each tone we correlated the vector of sentence ratings in humans and in GPT (Figure 2C). As expected, given that both humans and GPT show separately the pattern of valance clusters, the cross-domain matrix also showed this pattern. Interestingly, the main diagonal of this matrix shows that the alignment between tone ratings in humans and GPT varies significantly. Some tones such as "joyful", "pleased", and "happy" were highly correlated ($r = 0.89, 0.89, 0.87$ CI = [0.88, 0.91], [0.88, 0.91], [0.86, 0.90], respectively) suggesting that

these tones have a similar relation to other tones in both humans and GPT. However, other tones such as "proud", "apologetic", and "reflective" had relatively low alignment (r = 0.30, 0.46, 0.49 CI = [0.20, 0.39], [0.34, 0.52], [0.41, 0.54]).

Finally, our tone similarity calculations were somewhat indirect since we did not compare tone similarity directly but computed it based on the similarities of sentence ratings. To address this, we recruited a separate group of 71 participants. Each participant performed 50-60 trials and was asked to provide similarity ratings between two tones on a Likert scale. A similar procedure was applied for GPT. The resulting matrices (2D) indicate that direct similarity judgments also showed the valence block structure, but the similarity matrix appears to be noisier. Nevertheless, the data was still reliable: split-half correlations for the upper diagonal or the matrices were humans: r = 0.65 CI [0.63, 0.68], and that for GPT's similarity matrix to be mean 0.981 with CI [0.978, 0.987]. The high reliability of GPT partially originates from the monotone response that GPT provides in similarity judgments (unlike humans GPT agents provides the same response again and again for the same input). Importantly, the direct similarity matrices were aligned with the quality-of-fit rating approach (humans: r = 0.76 CI = [0.73, 0.78], GPT: r = 0.83 CI = [0.81, 0.84]). This is an independent validation that our approach does capture tone similarity structure. It also highlights the difficulty of working with direct similarity (Marjieh et al., 2023b).

4.3 Conversational Tone Representation (Multidimensional scaling)

In order to understand the organization of tone representations, we applied Multidimensional Scaling (MDS) to the combined within/across correlation matrix, such that each conversational tone possesses a 2-dimensional embedding in the resulting shared MDS space (Figure 3A).

We connected identical tones in GPT and humans with gray lines. Overall, it is visually apparent from Figure 3A that the structure of tones in humans and GPT is similar, which is consistent with the analysis of Figure 2. However, we also found differences in the proximity of tones in the shared space. The furthest tones away were "proud", "sad", and "thankful", while the closest were "uncertain", "desperate", and "conncerned" (Figure 3B).

To better interpret what the dimensions of difference are between humans and GPT and compare our results with respect to previous literature (Yeomans et al., 2022; Russell, 1980), we performed an additional experiment. 38 participants rated each tone from 1 to 5 on a Likert scale based on four theoretical dimensions: valence, emotional arousal, informational, and relational (overall 800 ratings). We also conducted a similar experiment with GPT (800 ratings). These features are defined in Appendix A.7. To observe how these ratings relate to the conversational tone embedding' MDS solutions, we projected the average rating over tones to the MDS (see Appendix B.3 for projection method using linear regression). Each theoretical dimension is represented by an arrow (direction) in MDS space, with the length of the arrow representing its relevance (measured by how much the dimension explained the variance among the points in the MDS).

The theoretical component of conversational tone that seems to explain most of the variance in their MDS solutions is "positive in valence" (humans: mean r = 0.71 CI = [0.69, 0.87]; GPT: mean r = 0.873 CI = [0.872, 0.878]), where the other terms explained significantly less variance (GPT: r = 0.45, 0.17, 0.13; humans: r = 0.37, 0.54, 0.3 for "relational", "aroused", "informational" respectively). This is consistent with the idea that the main axis of the quadratic shown by MDS solutions corresponds roughly to the positive/negative valence ("joyful", "happy", "excited", and "pleased" are in the bottom left, whereas "concerned", "worried", and "anxious" are in the top right).

From Figure 3A, however, we see that the human and GPT arrowmarks for the same conversational tone may have minor to medium directional differences, such as that for the feature "positive in valence" and "aroused" (indicated by a large cosine similarity; mean = 0.88 CI = [0.86, 0.94] and mean = 0.46 CI = [0.04, 0.53], respectively). Meanwhile, "relational" is consistently aligned (mean = 0.99 CI = [0.86, 0.99]), while "informational" is also strongly aligned (mean = 0.83 CI = [-0.26, 0.89]; see Appendix A.7). This suggests a deviation between the human and GPT understanding of these features. Furthermore, it validates Yeomans et al.'s chosen features for conversational tone composition, "informational" and "relational", as an aspect of well-aligned conversational understanding between humans and GPT (Yeomans et al., 2022). These results also allow us to further interpret the MDS of Figure 3A, for example, "proud" is located

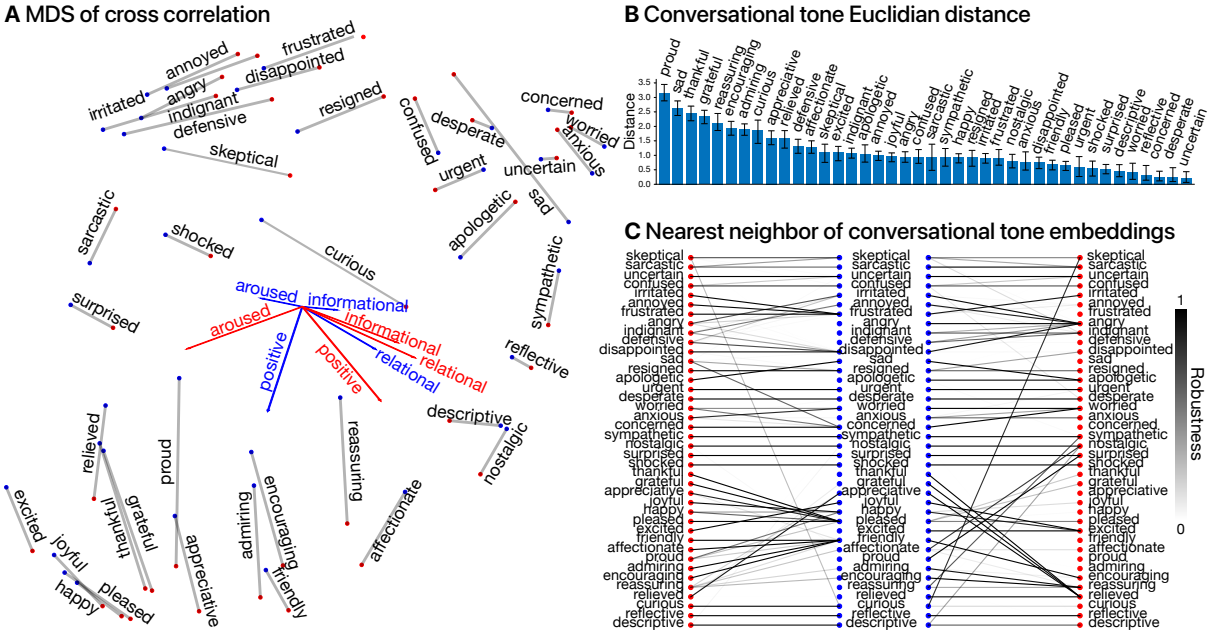


Figure 3: Cross-correlation alignment information. Blue points/arrowmarks in **A** and **C** represent GPT-originated data, while red represents human-originated instead. **A**: The MDS solution of applied to the combined within/across cross-domain (cross-correlation) matrix as a set of high-dimensional embedding to represent shared space of conversational tones embeddings across humans and GPT. Grey edges connect points representing the same conversational tone word. Arrow marks represent rating-derived dimensions of conversational tones. **B**: A barplot exhibiting the Euclidean distance between pairs of the same conversational tone embeddings in MDS space. Error bars represent one standard deviation via bootstrapping. **C**: A graph showing the nearest neighbor matches of conversational tone embeddings across humans and GPT. To measure robustness in matching, we bootstrapped the process 5000 times. Dark edges represent the frequency of its matching throughout bootstrap processes. See enlarged version of this figure in the Appendix (Figure 18).

farther away from neutral in the MDS space and stronger in the positive valence direction for humans. Thus, "proud" has a more neutral meaning for GPT. This shows how we can use Figure 3A to characterize how tones are conveyed in humans and GPT.

Finally, Figure 3C provides a mapping of similar tones across humans and LLMs. For each tone in one domain (humans or GPT), we present its nearest neighbor in another domain. This map is important because it allows us to "translate" conversational toes from humans to GPT and vice versa. As a result, Figure 3C can also help summarize distances in Figure 3A: the lines in Figure 3C represent concepts that are near one another in Figure 3A. Interestingly, we found cases where multiple human tones (e.g., "grateful", "joyful", "happy") were collapsed to a single LLM tone ("pleased"), suggesting that these terms were conveyed in a more limited way by GPT. We also found the reverse phenomenon that the GPT tones: "irritated", "annoyed", and "disappointed" collapsed to "angry" on the human side. This suggests that both humans and GPT have terms that are represented more broadly in the other group.

4.4 Application: Ground Truth for Benchmarking Semantic Alignment Methods

To demonstrate the utility of our data, we demonstrate how it can be used to benchmark semantic alignment methods. Note that in our approach we leverage the fact that we had the same taxonomy of tones for both humans and GPT. In many other use cases, this is not possible. For example, when translating words from one language to another it is impossible to ask speakers of one language to annotate words in a different language. Since our method has this kind of cross-linguistic information available, we use it as a source of ground truth to test methods that do not have access to this kind of information. Specifically, we survey the capability of frequently used unsupervised cross-domain alignment paradigms: Gromov-Wasserstein Optimal Transport (Grave et al., 2018), Orthogonal Procrustes Transformation (Schönemann, 1966),

and Bilingual Lexicon Induction (Ruder et al., 2018). We found that Bilingual Lexicon Induction via latent variable model (abbreviated hereon as BLI) is the best-performing approach. It recovers well the quality-of-fit rating similarity matrix entries (r = 0.81, CI = [0.81, 0.81]), followed by Orthogonal Procrustes Transformation (r = 0.56 CI = [0.54,0.58]) and Gromov-Wasserstein Optimal Transport (r = 0.54 CI = [0.53,0.58]). We found similar results for the recovery of k-Nearest-Neighbor structures (see Appendix Table 3). This indicates that not only is Bilingual Lexicon Induction superior at recovering the cross-domain proximity structure presented in our cross-domain ground truth, but also that it can preserve the intra-domain proximity structure of our embeddings after the alignment procedure. Then, in turn, this performance turns to show the superiority of our method in constructing a dataset that allows for almost-perfect reconstruction of similarity metrics for elements of the psycholinguistic modality.

5 Discussion

Using a cognitive science-inspired Sampling with People paradigm, we elicited tones and sentences for both humans and GPT creating a shared dataset of sentences and tones. Then, via quality-of-fit ratings, we further showed that the degree of alignment for different tones in humans and GPT varies. Tone alignment was high for tones such as "joyful" and "pleased" while it was significantly lower for tones such as "proud" and "apologetic". In a separate experiment, we found that similarity judgments were consistent with the resulting relationship between tones. Next, by projecting the rating vectors to a joint semantic space, we found variability in tone proximity, which is explained most by the well-known theoretical dimension of valence (Russell, 1980). Additionally, we provided a mapping of similar tones across humans and LLMs. We found cases where multiple human tones (e.g., "grateful", "joyful", "happy") were collapsed to a single LLM-proposed tone ("pleased"), as well as in the opposite direction (e.g., GPT's "irritated", "annoyed", and "disappointed" collapsed onto humans' "angry"), suggesting that both distributions ended up involving a hypernym for conversational tone categories. Finally, we demonstrate how our data can be used to benchmark methods for semantic alignment. We found that Bilingual Lexical Induction surpasses other (geometric) methods, suggesting that it would be appropriate for applications such as machine translation.

Our work opens up multiple avenues for future research. First, the same approach can be easily extended to other domains. For example, speakers of different languages and different cultures, as well as different language models. Second, our dataset can be used as a training signal for better aligning human and LLM conversational tones; for example, via an iterative process of refinements like reinforcement learning (Ji et al., 2024). Third, future work could look into whether we can use our alignment maps to predict performance in human-AI communication. More broadly, this work shows how combining approaches from machine learning and cognitive science provides routes for better understanding and resolving challenges in human-computer interactions.

Limitations

There are several technical limitations of our study that are important to highlight. First, we did not test a wide variety of LLMs and LLM parameters, including varying the prompt and temperature of the models. This limits the generalizability of our results, as the robustness of some of our findings may depend on these parameters. Second, we used a finite number of sampling chains with only 100 generations. Future work can explore what would be the effect of changing these hyperparameters. Finally, we only tested participants in the UK. It would be informative to test US participants as well as a wider range of speakers of other languages (Blasi et al., 2022).

More importantly, it is crucial to acknowledge that employing free elicitation methods could inadvertently generate sentences that reinforce societal biases, including racial and gender stereotypes. However, future research could investigate alternative filtering approaches, possibly involving human moderation, to actively reduce biases within the Sampling with People iterations (van Rijn et al., 2022a). Nonetheless, our approach can be used with participants in any language, which can help with creating AI systems for low-resource languages (Atari et al., 2023; Rathje et al., 2023). In particular, we are excited about the potential application of our approach to studying cross-cultural differences in tone of voice. We also believe that our research holds the potential to facilitate nuanced cross-cultural communication and support

the development of AI systems that communicate effectively with users from diverse backgrounds.

Ethics

We run our human experiment applying best practices in the responsible and ethical treatment of human subjects. We have deliberately reviewed the ACL Code of Ethics and confirm that our work conforms to all principles addressed in this code. All human experiment participants are paid $9 per hour and join the experiment after signing an informed consent of an approved protocol. Specifically, participants were recruited online via Prolific and provided consent in accordance with an approved protocol (MaxPlanck Ethics Council #2021 42).

In our Sampling with People human experiments, to avoid the appearance of profane words or triggering topics, we have added a profanity filter to our experiment such that responses containing profanity or vulgar words cannot be propagated along our sampling chains. Data was collected anonymously (beyond participants' Prolific IDs that were used for compensation). The published data was fully anonymized.

References

Francis M. Adams and Charles E. Osgood. 1973. A cross-cultural study of the affective meanings of color. *Journal of Cross-Cultural Psychology*, 4(2):135–156.

Elisa Claire Alemán Carreón, Hugo Alberto Mendoza España, Hirofumi Nonaka, and Toru Hiraoka. 2021. Differences in chinese and western tourists faced with japanese hospitality: a natural language processing approach. *Information Technology; Tourism*, 23(3):381–438.

Manuel Anglada-Tort, Peter MC Harrison, Harin Lee, and Nori Jacoby. 2023. Large-scale iterated singing experiments reveal oral transmission mechanisms underlying music evolution. *Current Biology*, 33(8):1472–1486.

Farzana Anowar, Samira Sadaoui, and Bassant Selim. 2021. Conceptual and empirical comparison of dimensionality reduction algorithms (pca, kpca, lda, mds, svd, lle, isomap, le, ica, t-sne). *Computer Science Review*, 40:100378.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. Generalizing and improving bilingual word embedding mappings with a multi-step framework of linear transformations. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pages 5012–5019.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018b. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798.

Mohammad Atari, Mona J Xue, Peter S Park, Damián E Blasi, and Joseph Henrich. 2023. Which humans?

A. Athanasiadou and E. Tabakowska. 1998. *Speaking of Emotions: Conceptualisation and Expression*. Cognitive linguistics research. Mouton de Gruyter.

H Clark Barrett. 2020. Towards a cognitive science of the human: cross-cultural approaches and their urgency. *Trends in cognitive sciences*, 24(8):620–638.

Anton Batliner, Stefan Steidl, and Elmar Noeth. 2008. *Releasing a thoroughly annotated and processed spontaneous emotional database: the FAU Aibo Emotion Corpus*.

André Beauducel. 2018. Recovering wood and mccarthy's erp-prototypes by means of erp-specific procrustes-rotation. *Journal of Neuroscience Methods*, 295:20–36.

Brent Berlin and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution*.

Damián E. Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. 2022. Over-reliance on english hinders cognitive science. *Trends in Cognitive Sciences*, 26(12):1153–1170.

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(4):335–359.

J Douglas Carroll and Phipps Arabie. 1998. Multidimensional scaling. *Measurement, judgment and decision making*, pages 179–250.

Yuh-Fang Chang. 2009. How to say no: an analysis of cross-cultural difference and pragmatic transfer. *Language Sciences*, 31(4):477–493.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alan Cowen, Hillary Elfenbein, Petri Laukka, and Dacher Keltner. 2018. Mapping 24 emotions conveyed by brief human vocalization. *American Psychologist*.

Alan Cowen, Xia Fang, Disa Sauter, and Dacher Keltner. 2020. What music makes us feel: At least 13 dimensions organize subjective experiences associated with music across different cultures. *Proceedings of the National Academy of Sciences*, 117:201910704.

Alan Cowen, Petri Laukka, Hillary Elfenbein, Runjing Liu, and Dacher Keltner. 2019. The primacy of categories in the recognition of 12 emotions in speech prosody across two cultures. *Nature Human Behaviour*, 3:1.

Alan S. Cowen and Dacher Keltner. 2017. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proceedings of the National Academy of Sciences*, 114:E7900 – E7909.

Alan S. Cowen and Dacher Keltner. 2020. What the face displays: Mapping 28 emotions conveyed by naturalistic expression. *The American psychologist*.

Simon De Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "small world of words" english word association norms for over 12,000 cue words. *Behavior Research Methods*, 51(3):987–1006.

Simon De Deyne, Sophie Warner, and Andrew Perfors. under review. Common words, uncommon meanings: Evidence for widespread gender differences in word meaning. In *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

Mariano de Rivero, Cristhiam Tirado, and Willy Ugarte. 2023. Formalstyler: Gpt-based model for formal style transfer with meaning preservation. *SN Computer Science*, 4(6):739.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6:169–200.

Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587.

Kai Von Fintel. 2006. Modality and language. In D. Borchert, editor, *Encyclopedia of Philosophy*, pages 20–27. Macmillan Reference.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA.

Alessandro T. Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M. Cichy. 2022. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754.

Ian J. Goodfellow, Dumitru Erhan, Pierre Luc Carrier, Aaron Courville, Mehdi Mirza, Ben Hamner, Will Cukierski, Yichuan Tang, David Thaler, Dong-Hyun Lee, Yingbo Zhou, Chetan Ramaiah, Fangxiang Feng, Ruifan Li, Xiaojie Wang, Dimitris Athanasakis, John Shawe-Taylor, Maxim Milakov, John Park, Radu Ionescu, Marius Popescu, Cristian Grozea, James Bergstra, Jingjing Xie, Lukasz Romaszko, Bing Xu, Zhang Chuang, and Yoshua Bengio. 2013. Challenges in representation learning: A report on three machine learning contests.

Edouard Grave, Armand Joulin, and Quentin Berthet. 2018. Unsupervised alignment of embeddings with wasserstein procrustes.

M.J. Greenacre. 2010. *Biplots in Practice*. Fundación BBVA.

Thomas L Griffiths and Michael L Kalish. 2005. A bayesian view of language evolution by iterated learning. In *Proceedings of the annual meeting of the cognitive science society*, volume 27.

TL Griffiths, Adam N Sanborn, R Marjieh, T Langlois, J Xu, and N Jacoby. 2024. Estimating subjective probability distributions.

Peter M. C. Harrison, Raja Marjieh, Federico Adolfi, Pol van Rijn, Manuel Anglada-Tort, Ofer Tchernichovski, Pauline Larrouy-Maestri, and Nori Jacoby. 2020. Gibbs sampling with people.

Martin N Hebart, Adam H Dickter, Alexis Kidder, Wan Y Kwok, Anna Corriveau, Caitlin Van Wicklin, and Chris I Baker. 2019. THINGS: A database of 1,854 object concepts and more than 26,000 naturalistic object images. *PLoS One*, 14(10):e0223792.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Yihan Hou, Manling Yang, Hao Cui, Lei Wang, Jie Xu, and Wei Zeng. 2024. C2ideas: Supporting creative interior color design ideation with large language model. *ArXiv*, abs/2401.12586.

Nori Jacoby and Josh H McDermott. 2017. Integer ratio priors on musical rhythm revealed cross-culturally by iterated reproduction. *Current Biology*, 27(3):359–370.

Nori Jacoby, Rainer Polak, Jessica A Grahn, Daniel J Cameron, Kyung Myun Lee, Ricardo Godoy, Eduardo A Undurraga, Tomás Huanca, Timon Thalwitzer, Noumouké Doumbia, et al. 2024. Commonality and variation in mental representations of music revealed by a cross-cultural comparison of rhythm priors in 15 countries. *Nature Human Behaviour*, pages 1–32.

Roman Jakobson. 1960. Closing statement: Linguistics and poetics. In *Style in Language*, pages 350–.

Jiaming Ji, Tianyi Qiu, Boyuan Chen, Borong Zhang, Hantao Lou, Kaile Wang, Yawen Duan, Zhonghao He, Jiayi Zhou, Zhaowei Zhang, Fanzhi Zeng, Kwan Yee Ng, Juntao Dai, Xuehai Pan, Aidan O'Gara, Yingshan Lei, Hua Xu, Brian Tse, Jie Fu, Stephen McAleer, Yaodong Yang, Yizhou Wang, Song-Chun Zhu, Yike Guo, and Wen Gao. 2024. Ai alignment: A comprehensive survey.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep Learning for Text Style Transfer: A Survey. *Computational Linguistics*, 48(1):155–205.

Patrik Juslin and John Sloboda. 2013. *Music and Emotion*, pages 583–645.

Patrik N Juslin and Daniel Västfjäll. 2008. Emotional responses to music: The need to consider underlying mechanisms. *Behavioral and brain sciences*, 31(5):559–575.

Genji Kawakita, Ariel Zeleznikow-Johnston, Naotsugu Tsuchiya, and Masafumi Oizumi. 2023. Comparing color similarity structures between humans and llms via unsupervised alignment.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. 2017. The kinetics human action video dataset.

Yousif Khaireddin and Zhuofa Chen. 2021. Facial emotion recognition: State of the art performance on fer2013.

Dolichan Kollareth, José-Miguel Fernandez-Dols, and James Russell. 2018. Shame as a culture-specific emotion concept. *Journal of Cognition and Culture*, 18:274–292.

Roger J. Kreuz and Richard M. Roberts. 1993. When collaboration fails: Consequences of pragmatic errors in conversation. *Journal of Pragmatics*, 19(3):239–252.

Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *arXiv preprint arXiv:1711.00043*.

Thomas Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. 2017. Uncovering visual priors in spatial memory using serial reproduction. In *CogSci*.

Thomas A Langlois, Nori Jacoby, Jordan W Suchow, and Thomas L Griffiths. 2021. Serial reproduction reveals the geometry of visuospatial representations. *Proceedings of the National Academy of Sciences*, 118(13):e2012938118.

Chujun Lin, Umit Keles, and Ralph Adolphs. 2021. Four dimensions characterize attributions from faces using a representative set of english trait words. *Nature Communications*, 12(1):5168.

Kristen A. Lindquist, Joshua Conrad Jackson, Joseph Leshin, Ajay B. Satpute, and Maria Gendron. 2022. The cultural evolution of emotion. *Nature Reviews Psychology*, 1(11):669–681.

Xiao Liu, Jing Zhao, Shiliang Sun, Huawen Liu, and Hao Yang. 2021. Variational multimodal machine translation with underlying semantic alignment. *Information Fusion*, 69:73–80.

Hans Peter Luhn. 1958. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165.

Raja Marjieh, Peter MC Harrison, Harin Lee, Fotini Deligiannaki, and Nori Jacoby. 2024. Timbral effects on consonance disentangle psychoacoustic mechanisms and suggest perceptual origins for musical scales. *Nature Communications*, 15(1):1482.

Raja Marjieh, Ilia Sucholutsky, Pol van Rijn, Nori Jacoby, and Thomas L Griffiths. 2023a. Large language models predict human sensory judgments across six modalities. *arXiv preprint arXiv:2302.01308*.

Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R. Sumers, Harin Lee, Thomas L. Griffiths, and Nori Jacoby. 2023b. Words are all you need? language as an approximation for human similarity judgments.

Louis Marti, Shengyi Wu, Steven T. Piantadosi, and Celeste Kidd. 2023. Latent Diversity in Human Concepts. *Open Mind*, 7:79–92.

Leland McInnes, John Healy, and James Melville. 2020. Umap: Uniform manifold approximation and projection for dimension reduction.

John Oetzel, Stella Ting-Toomey, Tomoko Masumoto, Yumiko Yokochi, Xiaohui Pan, Jiro Takai, and Richard Wilcox. 2001. Face and facework in conflict: a cross-cultural comparison of china, germany, japan, and the united states. *Communication Monographs*, 68(3):235–258.

Eva Ogiermann. 2009. Politeness and in-directness across cultures: A comparison of english, german, polish and russian requests. *Journal of Politeness Research*, 5(2):189–216.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Stephen E Palmer, Karen B Schloss, Zoe Xu, and Lilia R Prado-León. 2013. Music-color associations are mediated by emotion. *Proc. Natl. Acad. Sci. U. S. A.*, 110(22):8836–8841.

Paul Portner. 2009. *Modality*. Modality. OUP Oxford.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model.

Steve Rathje, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire Robertson, and Jay J Van Bavel. 2023. Gpt is an effective tool for multilingual psychological text analysis.

David Rozado, Ruth Hughes, and Jamin Halberstadt. 2022. Longitudinal analysis of sentiment and emotion in news media headlines using automated labelling with transformer language models. *PLOS ONE*, 17(10):1–14.

Sebastian Ruder, Ryan Cotterell, Yova Kementchedjhieva, and Anders Søgaard. 2018. A discriminative latent-variable model for bilingual lexicon induction.

Jürgen Rudolph, Shannon Tan, and Samson Tan. 2023. War of the chatbots: Bard, bing chat, chatgpt, ernie and beyond. the new ai gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, 6(1).

James A. Russell. 1980. A circumplex model of affect. *Journal of personality and social psychology*, 39(6):1161–1178.

Aaron Saiewitz and Thomas Kida. 2018. The effects of an auditor's communication mode and professional tone on client responses to audit inquiries. *Accounting, Organizations and Society*, 65:33–43.

Adam N. Sanborn, Thomas L. Griffiths, and Richard M. Shiffrin. 2010. Uncovering mental representations with markov chain monte carlo. *Cognitive Psychology*, 60(2):63–106.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Disa Sauter, Frank Eisner, Paul Ekman, and Sophie Scott. 2010. Cross-cultural recognition of basic emotions through nonverbal emotional vocalizations. *Proceedings of the National Academy of Sciences of the United States of America*, 107:2408–12.

Daniela Schiller, NC Alessandra, Nelly Alia-Klein, Susanne Becker, Howard C Cromwell, Florin Dolcos, Paul J Eslinger, Paul Frewen, Andrew H Kemp, Edward F Pace-Schott, et al. 2023. The human affectome. *Neuroscience & Biobehavioral Reviews*, page 105450.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31:1–10.

Zhaleh Semnani-Azad and Wendi L. Adair. 2013. Watch your tone ... relational paralinguistic messages in negotiation. *International Studies of Management & Organization*, 43(4):64–89.

Roger N. Shepard. 1980. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210:390 – 398.

Ilia Sucholutsky, Ruairidh M Battleday, Katherine M Collins, Raja Marjieh, Joshua Peterson, Pulkit Singh, Umang Bhatt, Nori Jacoby, Adrian Weller, and Thomas L Griffiths. 2023a. On the informativeness of supervision signals. In *Uncertainty in Artificial Intelligence*, pages 2036–2046. PMLR.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C. Love, Erin Grant, Iris Groen, Jascha Achterberg, Joshua B. Tenenbaum, Katherine M. Collins, Katherine L. Hermann, Kerem Oktar, Klaus Greff, Martin N. Hebart, Nori Jacoby, Qiuyi Zhang, Raja Marjieh, Robert Geirhos, Sherol Chen, Simon Kornblith, Sunayana Rane, Talia Konkle, Thomas P. O'Connell, Thomas Unterthiner, Andrew K. Lampinen, Klaus-Robert Müller, Mariya Toneva, and Thomas L. Griffiths. 2023b. Getting aligned on representational alignment.

Deborah Tannen. 1984. The pragmatics of cross-cultural communication. *Applied linguistics*, 5(3):189–195.

Bill D. Thompson, Seán G. Roberts, and Gary Lupyan. 2020. Cultural influences on word meanings revealed through large-scale semantic alignment. *Nature Human Behaviour*, 4:1029 – 1038.

Stella Ting-Toomey, Ge Gao, Paula Trubisky, Zhizhong Yang, Hak-Soo Kim, Sung-Ling Lin, and Tsukasa Nishida. 1991. Culture, face maintenance, and styles of handling interpersonal conflict: A study in five cultures. *International Journal of Conflict Management*, 2:275–296.

Eline Van Geert and Nori Jacoby. 2024. Using gibbs sampling with people to characterize perceptual and aesthetic evaluations in multidimensional visual stimulus space.

Pol van Rijn, Harin Lee, and Nori Jacoby. 2022a. Bridging the prosody gap: Genetic algorithm with people to efficiently sample emotional prosody.

Pol van Rijn, Silvan Mertes, Kathrin Janowski, Katharina Weitz, Nori Jacoby, and Elisabeth André. 2024. Giving robots a voice: Human-in-the-loop voice creation and open-ended labeling. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pages 1–34.

Pol van Rijn, Silvan Mertes, Dominik Schiller, Piotr Dura, Hubert Siuzdak, Peter Harrison, Elisabeth André, and Nori Jacoby. 2022b. Voiceme: Personalized voice generation in tts. *arXiv preprint arXiv:2203.15379*.

Pol Van Rijn, Silvan Mertes, Dominik Schiller, Peter Harrison, Pauline Larrouy-Maestri, Elisabeth André, and Nori Jacoby. 2021. Exploring emotional prototypes in a high dimensional tts latent space. *arXiv preprint arXiv:2105.01891*.

Jef A. van Schendel and Raymond H. Cuijpers. 2015. Turn-yielding cues in robot-human conversation. AISB Convention 2015, Society for the Study of Artificial Intelligence and Simulation of Behaviour, 20-22 April 2015, Canterbury, 2015, AISB2015 ; Conference date: 20-04-2015 Through 22-04-2015.

Tieyan Wang and Tieshan Wang. 2016. Effects of color on expectations of drug effects: A cross-gender cross-cultural study. *Color Research & Application*, 42:n/a–n/a.

Xun Wang, Tao Ge, Allen Mao, Yuki Li, Furu Wei, and Si-Qing Chen. 2022. Pay attention to your tone: Introducing a new dataset for polite language rewrite.

Yunli Wang, Yu Wu, Lili Mou, Zhoujun Li, and Wenhan Chao. 2019. Harnessing pre-trained neural networks with rules for formality style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3573–3578, Hong Kong, China. Association for Computational Linguistics.

Jing Xu, Thomas Griffiths, and Mike Dowman. 2010. Replicating color term universals through human iterated learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 32.

Jing Xu and Thomas L Griffiths. 2010. A rational analysis of the effects of memory biases on serial reproduction. *Cognitive psychology*, 60(2):107–126.

Michael Yeomans, Maurice E. Schweitzer, and Alison Wood Brooks. 2022. The conversational circumplex: Identifying, prioritizing, and pursuing informational and relational motives in conversation. *Current Opinion in Psychology*, 44:293–302.

Pustakhanim Yusifova. 2018. Three layers of pragmatic failure across languages and cultures. *International Journal of English Linguistics*.

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Benjamin D. Zinszer, Andrew J. Anderson, Olivia Kang, Thalia Wheatley, and Rajeev D. S. Raizada. 2016. Semantic Structural Alignment of Neural Representational Spaces Enables Translation between English and Chinese Words. *Journal of Cognitive Neuroscience*, 28(11):1749–1759.

A Appendix: Additional Methods

A.1 Implementation of the Experiments

Human Experiments. Human experiments are implemented using Psynet (Harrison et al., 2020), a Python package for implementing complex online psychology experiments. PsyNet automates the online hosting of experiments and automatically pays participants through a variety of recruitment platforms, such as Prolific.

GPT Experiments. GPT experiments are implemented in Python and GPT responses are all elicited using its chat completion mode. We always used a temperature of 0.8 to elicit responses with higher variance. In all experiments, we used GPT-4 (the June 13th, 2023 release).

Licensed Code The implementations of unsupervised alignment methods that we reference for GWOT and latent variable model-based Bilingual Lexicon Induction follow respectively a CC BY-NC 4.0 license and a GPL-3.0 license.

Data Anonymity We confirm that all human-originated data provided with our submission are anonymized.

Computational Budget We used an AWS server to host our online human experiments. We used an AWS EC2 m5.2xlarge container for 90 hours for completing all necessary experiments once, and 1 AMD Ryzen 7 5800 CPU for 45 hours for all analyses and GPT requests once. No GPUs were used in the progress of this work.

A.2 Participants and Procedure

We recruited participants from the crowd-sourcing recruiting service Prolific (https://prolific.com/). Participants were compensated approximately 9 GBP per hour for their time. To reduce cultural differences between English-speaking participants, we target specifically English-speaking participants in one country. All participants satisfied the following three criteria: (1) Were over 18 years of age, (2) Lived in the United Kingdom, (3) Native English speakers. Participants were recruited online via Prolific and provided consent in accordance with an approved protocol (MaxPlanck Ethics Council #2021 42).

Sampling with People Experiment. We recruited 955 human participants for Sampling with People (SP) experiments. These experiments had the following procedure. Each participant completed 10 to 12 trials. For each trial, participants were randomly assigned to either annotate a sentence with a conversational tone (T trials; Figure 1A, 6b) or create a response sentence matching a displayed conversational tone (S trials; Figure 1A, 6a). There were 90 SP sampling chains, each with 100 iterations per chain (50 S and T trials). This results in exactly 4,500 generated sentences and 4,500 tone annotations.

Quality-of-fit Rating Experiment. We recruited 275 human participants for the Quality-of-fit Rating experiments. Participants performed 12 rating trials, where they rated each pair of sentence-tones on five Likert scales (1 being the weakest, 5 being the strongest) based on the strength of the conversational tone. For each sentence-tone pair, we collected approximately five ratings. Overall we collected 5 responses for all participants and sentence-tone pairs.

Similarity Judgement Experiment. We recruited 71 human participants for the experiment. Participants performed 50 to 60 trials where they were asked to provide five similarity judgments per pair of (not necessarily distinct) conversational tones, on a scale from 1 to 5, with 1 being the most dissimilar and 5 being the most similar. We then computed the resulting similarity for each conversation-tone pair (t_1, t_2), normalized to the scale of [0, 1].

Tone Feature Rating Experiment. We recruited 38 human participants for the experiment. In this experiment, participants performed 30 to 40 trials where they were asked to rate tone-feature pairs on a 1-to-5 Likert scale. We tested four tone features (positiveness in valence, emotional arousal, informational, relational; see below for definition). The definitions of these features which were provided to participants and GPT agents are listed in Appendix A.7. The resulting rating for each sentence-tone pair was the average of the obtained five ratings.

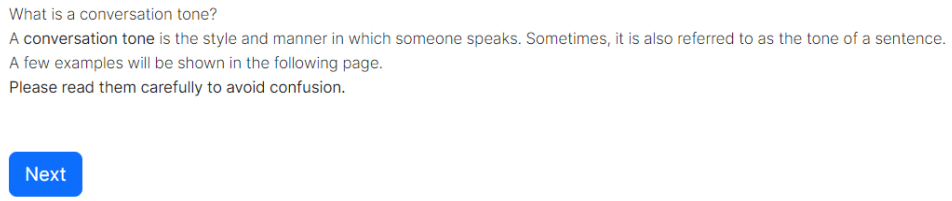
A.3 Explaining the Concept of Conversational Tone

In all human experiments, participants first receive instructions explaining what a conversational tone is and example sentences clarifying the concepts.

Upon entering the experiments, all participants are presented with this definition of conversational tone. After being presented with an operationalized definition of conversational tone, participants are provided examples that show usages of different conversational tones. Participants will be shown an example regarding detecting a conversational tone from a sentence, and an example of creating

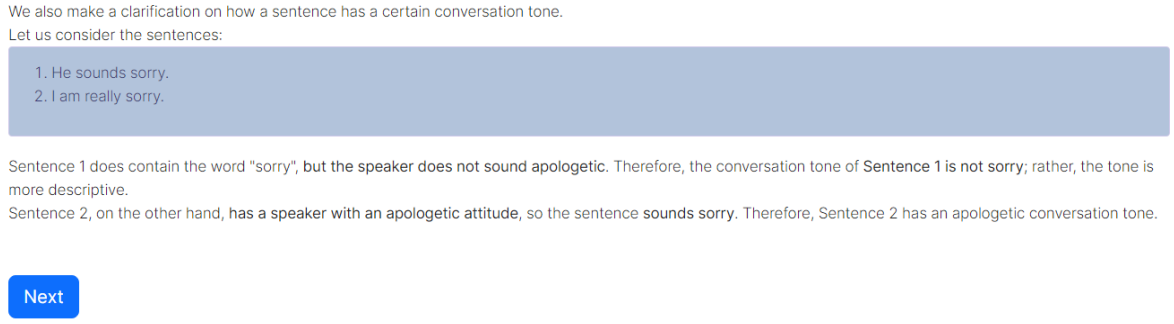
Experiment	# Human Participants	# Human Responses	# GPT Responses
SP Sampling	955	9000	9000
Quality-of-fit Rating	275	19470	16000
Similarity Judgment	71	4100	4100
Tone Feature Rating	38	800	800
Overall	1339	33370	29900

Table 1: Summary Table. Number of human participants, human judgments, and GPT judgments involved in our paper.



(a) Definition of conversational tones shown to participants.

(b) Example scenarios and appropriate example answers.



(c) Two examples explaining the concept of conversational tones.

Figure 4: Instructions shown to participants.

```
Create 50 data according to the following format:
{
    ''example_tone'': An adjective that describes a conversational tone,
    ''example_sentence_1'': A sentence with at least 7 words that has the a conversational tone
    of example_tone,
    ''example_sentence_1_explanation'': A 20~30 word explanation on 2~3 conversational tones of
    example sentence 1,
    ''example_sentence_2'': A different sentence with at least 7 words that has the a conversation
    tone of example_tone,
    ''example_sentence_2_explanation'': A 20~30 word explanation on 2~3 conversational tones of
    example sentence 1
}
```

(a) ChatGPT prompt for creating seeds of instructional examples on how to create sentences from conversational tones.

```
Create 50 data according to the following format:
{
    ''example_sentence'': A sentence with at least 7 words,
    ''example_tone_1'': An adjective representing a different conversational tone you observe
    from ''example_sentence'',
    ''example_tone_2'': An adjective representing a different conversational tone you observe
    from ''example_sentence'',
    ''example_tone_1_explanation'': Explain how you observed tone 1 from ''example_sentence''
    with 20 to 30 words,
    ''example_tone_2_explanation'': Explain how you observed tone 2 from ''example_sentence''
    with 20 to 30 words
}
```

(b) ChatGPT prompt for creating seeds of instructional examples on how to detect conversational tones from sentences.

Figure 5: ChatGPT prompts for creating seeds that generate the text seen in interface detailed at Figure 4b

a sentence that has a provided conversational tone. These interfaces are exemplified in Figure 4.

There is a pool of 50 items for each type of example, all created via ChatGPT. To mitigate bias resulting from examples, participants are shown a random item from each pool of examples. The prompt of creating such examples is exhibited as shown in Figure 5.

A.3.1 Consideration for human instructions and GPT prompts

In pilot experiments, several participants create sentences of a provided conversational tone T in the form of "<Subject> felt <T>". However, such a sentence may not necessarily convey conversational tone T. For example, for the conversational tone "apologetic", the sentence "He felt apologetic" communicates that a "He" is apologetic, but not that the speaker is apologetic, which contradicts the definition of conversational tone as "the speaker's attitude in a conversation". So, as an effort to prevent participants from providing such responses, an instruction as listed below is presented as in Figure 4c.

A.3.2 Slider Instruction

In some experiments, participants will need to rate some properties of conversational tones on a Likert scale of 1 to 5. For those experiments, after explaining the concept of conversational tones, they are presented with the following introduction to learn how to use the slider used for rating as depicted in Figure 7.

A.4 Experiment 1: SP Sampling

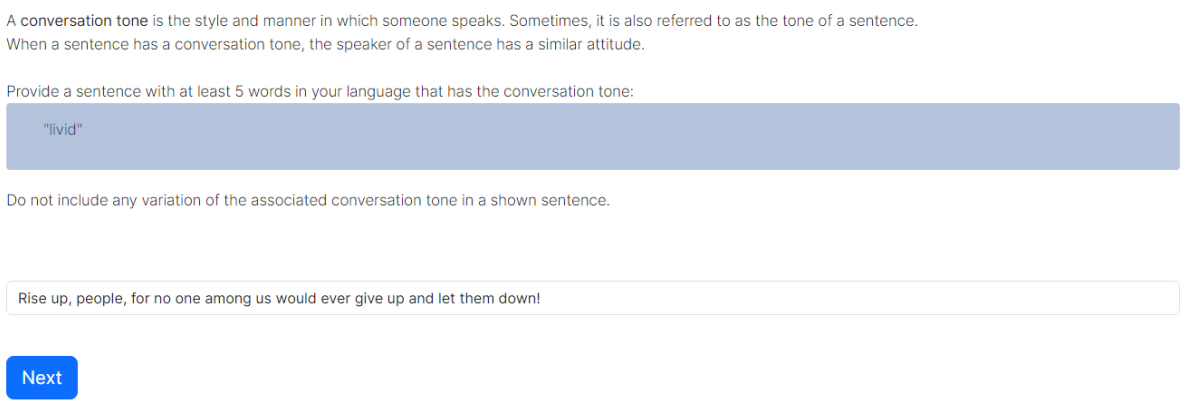
A.4.1 Human Experiment

This section describes the implementational details of SP Sampling (see Section 3) human and GPT experiments.

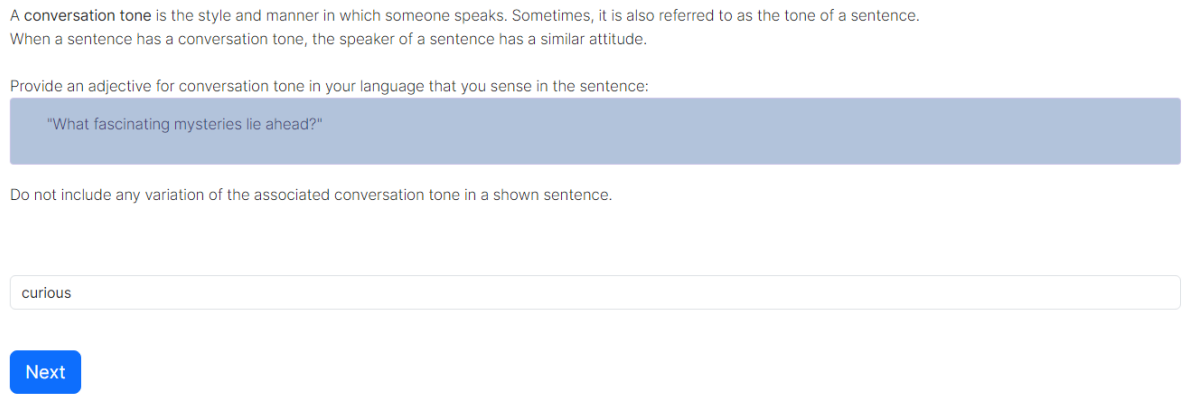
The participants first go through the general instructions (see Figure 4) and do two practice trials to familiarize themselves with the tasks they will be performing: (1) detecting a conversational tone from a sentence and (2) creating a sentence that conveys some provided conversational tone.

In the main experiment, each participant does ten trials. The human interface is shown in Figure 6

To avoid low-quality sentences, we automatically check the submitted sentence for the following criteria: (1) The sentence has to have more



(a) The interface of a S trial.



(b) The interface of a T trial.

Figure 6: The interfaces of a S trial and a T trial in our Sampling with People human experiment.

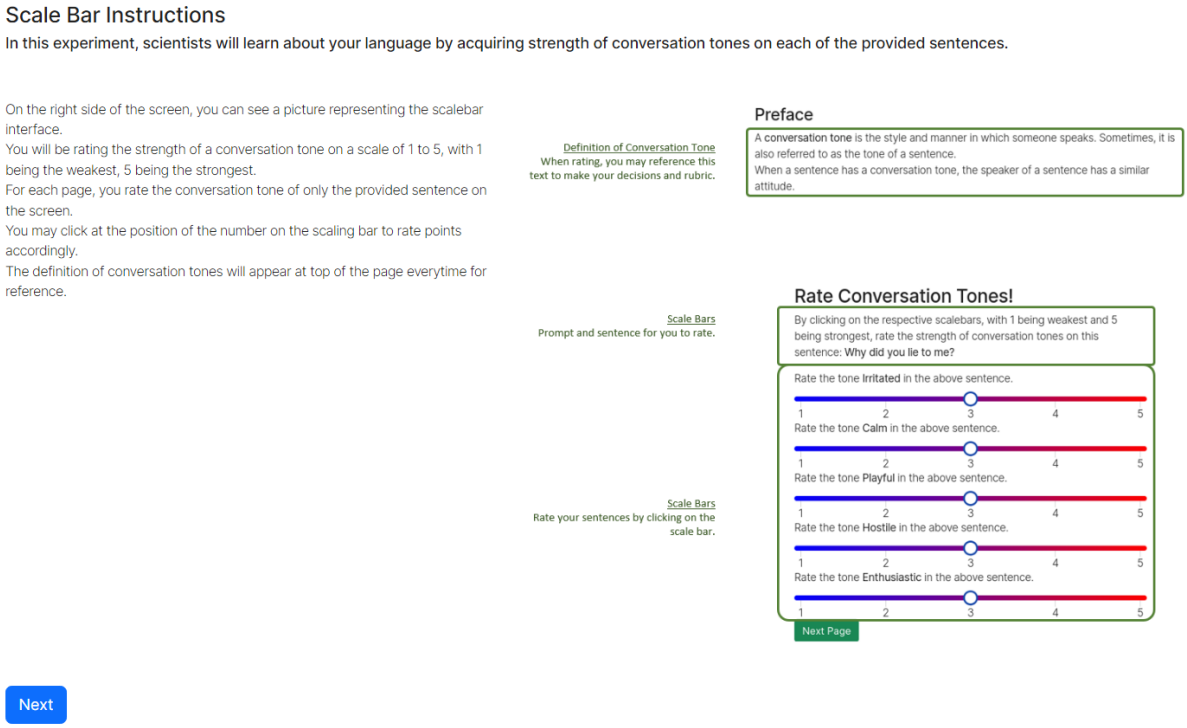


Figure 7: This page of the experiment details the instructions for using our designed scale interface for rating.

Response Type	Validation Criterion	Implementation
Sentence	Must have more than 5 words	`RegEx`
Sentence	Must be a grammatically correct sentence	`GingerIt`
Tone	The response can only contain alphabets and hyphens	`RegEx`
Tone	The response must be an adjective	`PyDictionary`
Tone	The response must be correctly spelled	`PyDictionary`
Both	Cannot contain any stemmed variation of prompt	`nltk`
Both	Cannot contain profanity	`profanity-check`

Table 2: Summary of Sampling with People response filters

than five words; (2) The sentence is grammatically correct (checked using `gingerit` [4]) (3) The sentence does not contain any stemmed variation of some word that is already in the provided conversational tone (for example "politely" for "polite") using `nltk`[5]; (4) Does not contain offensive or vulgar words checked with the Python package `profanity-check` [6].

The conversational tones were verified in a similar fashion: (1) The tone response did not contain any stemmed variation of some word that is already in the provided conversational tone or sentence, (2) The response must be an adjective using `PyDictionary`[7], and (3) It must not contain profanity. See Table 2 for a summary.

In the human experiment, we elicited 90 sampling chains, each with 100 iterations. Participants cannot revisit the same chain.

A.4.2 GPT-4 Experiments

The GPT prompts (see Figure 8) were nearly identical to the human experiments. In the GPT instance, we similarly elicited 90 sampling chains, each with 100 iterations.

A.5 Experiment 2: SP Quality-of-fit Rating

A.5.1 Human Experiment

After reading the general instruction (see Figure 4), participants first do two practice trials to familiarize themselves with the task they will be performing: rating the strength of a conversational tone when given a sentence. Then, participants proceed to the main experiment. Participants can only rate a conversational tone-sentence once, and each tone-sentence pair is rated by approximately 5 distinct participants. The strength of conversational tones in a sentence is rated on a Likert scale from 1 to 5, with 1 being the weakest and 5 being the strongest. The final rating of such a tone-sentence pair would be the average of all 5 ratings. An example of the rating interface is provided in Figure 9.

A.5.2 GPT-4 Experiments.

GPT receives the prompt for quality-of-fit rating as outlined in Figure 10.

A.6 Experiment 3: conversational tone Similarity Judgment

A.6.1 Human Experiment

After reading the general (see Figure 4) and experiment-specific instructions (see Figure 11a), participants proceed to the main experiment in which they judge the similarity of a pair of two conversational tones.

Each participant would rate the similarity of several distinct pairs of conversational tones on a Likert scale of 1 to 5, where 1 represents "Semantically dissimilar conversational tones" and 5 represents "Semantically similar conversational tones". Each pair of conversational tones would be rated five times, with the average score of those 5 ratings as the final similarity score for such pair. Only distinct pairs are rated. That means the similarity judgment of 40 conversational tones concerns only the 820 distinct pairs among all possible tuples of tones. And, similar to previous experiments, participants cannot rate any conversational tone pair more than once. An example of the rating interface follows in Figure 11b.

A.6.2 GPT Experiment

The prompt for similarity judgment that GPT receives is as outlined in 12.

[4] `gingerit`: https://github.com/Azd325/gingerit/blob/main/gingerit/gingerit.py

[5] https://www.nltk.org/

[6] `profanity-check`: https://pypi.org/project/profanity-check/

[7] https://pypi.org/project/PyDictionary/

```
A conversational tone is the style and manner in which someone speaks.
Provide an adjective for conversational tone in English that you sense in the following sentence:
<sentence of previous response>. Respond using only an adjective.
```

(a) GPT Prompt for sampling a conversational tone given a sentence.

```
A conversational tone is is the style and manner in which someone speaks.
Provide one sentence with at least five words in English that has the conversational tone:
<conversational tone of previous response>.
```

(b) GPT Prompt for sampling a sentence given a conversational tone.

Figure 8: Collection of GPT Prompts for GPT participant in sampling and rating aspects of SP paradigm.

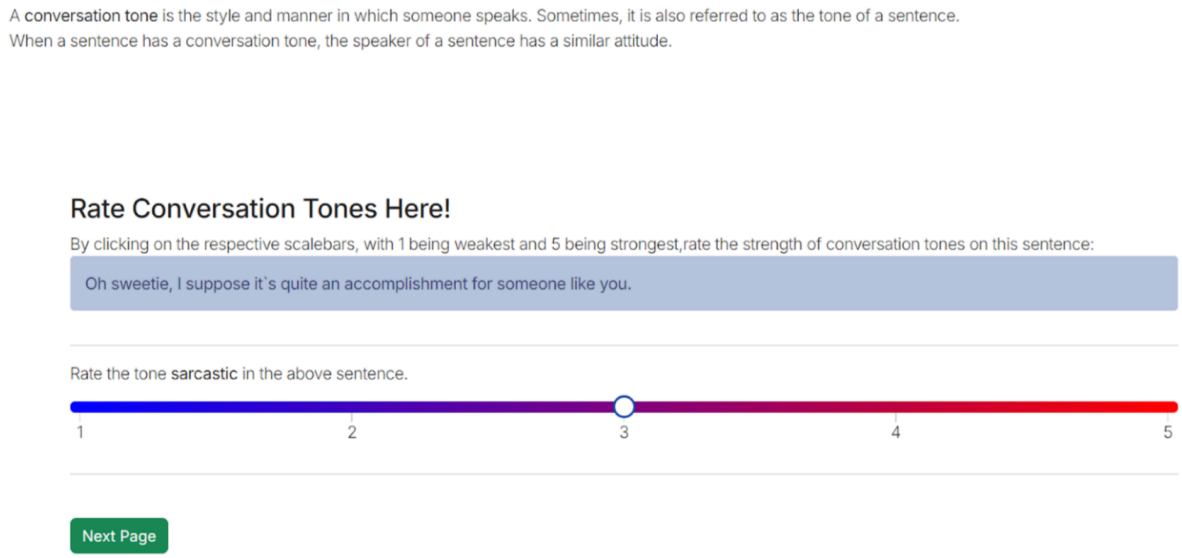
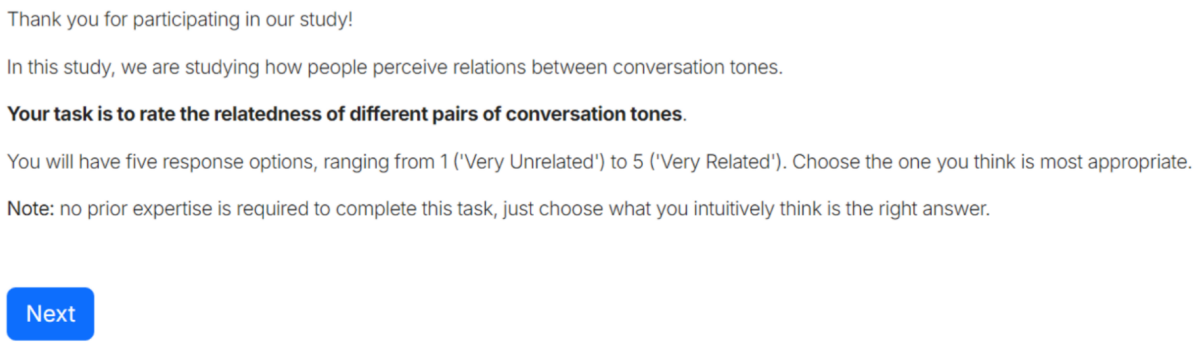


Figure 9: Example interface for quality-of-fit rating human experiment interface.

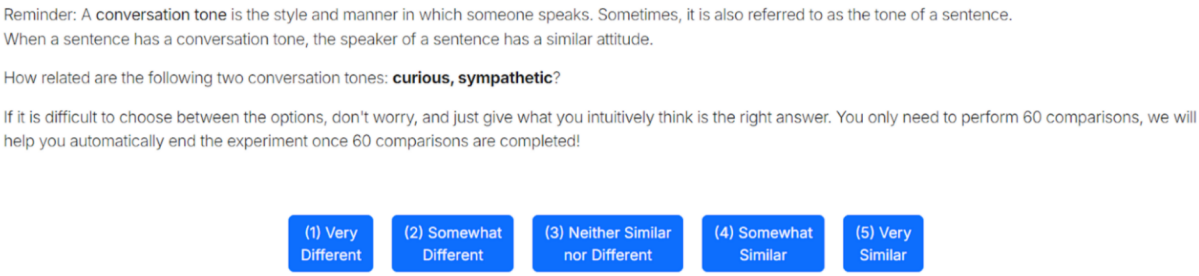
```
A conversational tone is the style and manner in which someone speaks.
On a scale of 1 to 5, with 5 being strongest, how strong is the provided conversational tone in
the following English sentence?
Tone: <conversational tone to rate>
Sentence: <sentence to rate>
Respond with only a number.
```

(a) GPT Prompt for rating the strength of a conversational tone on a sentence.

Figure 10: GPT Prompts used for quality-of-fit rating experiment.



(a) Interface that introduces human participants to similarity judgment experiment.



(b) Interface that represents the main body of similarity judgment human experiment.

```
People described conversational tones using words.
A conversational tone is the style and manner in which someone speaks,
and sometimes, it is also referred to as the tone of a sentence.
How similar are the conversational tones in each pair on a scale of 0-1
where 0 is completely dissimilar and 1 is completely similar?
        conversational tone 1: {tone_a}
        conversational tone 2: {tone_b}
Respond only with the numerical similarity rating.
```

Figure 12: GPT Prompt for rating the similarity of conversational tones.

```
A conversational tone is the style and manner in which someone speaks.
{explanation of rated tone feature using its definition in Table}
On a scale of 1 to 5, where 5 means strongest and 1 means weakest, how {feature} is the conversational tone
''{tone}''?
Respond with only a number.
```

Figure 13: GPT Prompt for rating the strength of features in conversational tones.

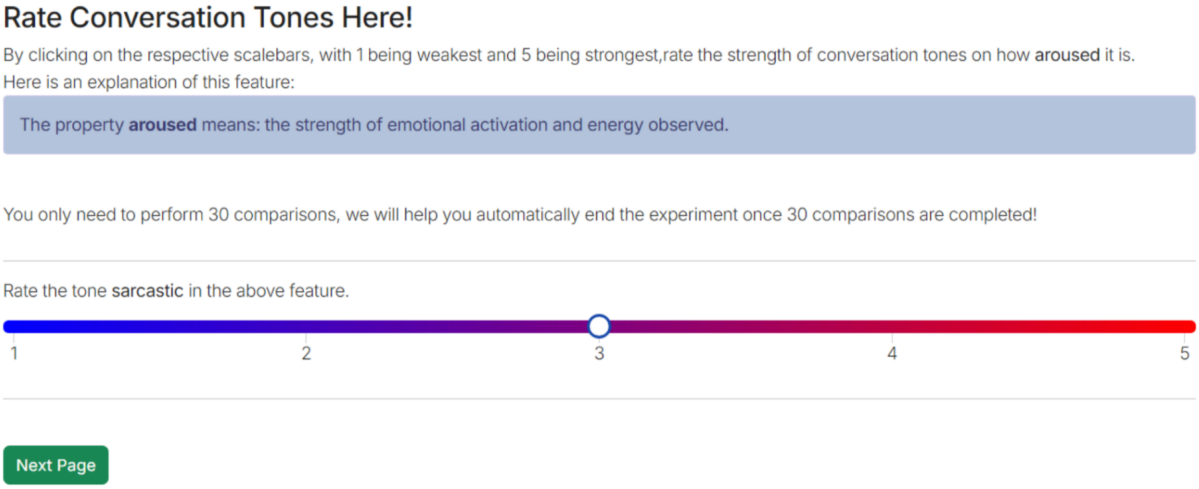


Figure 14: Interface for the main body of tone feature rating human experiment.

A.7 Experiment 4: conversational tone Feature Rating

A.7.1 Human Experiment

After reading the general instructions (see Figure 4), human participants proceed to the main experiment in which they rate a feature of conversational tones. Based on psycholinguistic literature (Yeomans et al., 2022; Fintel, 2006; Jin et al., 2022; Oetzel et al., 2001; Portner, 2009) we selected the following features along with their definitions:

- **positive in valence**: Positiveness in valence means the positiveness of emotional valence.
- **aroused**: Aroused means the amount of emotional arousal observed.
- **Informational**: Informational means the extent to which a speaker's motive focuses on giving and/or receiving accurate information.
- **Relational**: Relational means the extent to which a speaker's motive focuses on building the relationship.

The strength of features in a conversational tone is rated on a Likert scale from 1 to 5, with 1 being the weakest and 5 being the strongest. Participants are provided an interface for rating the features in conversational tones. See Figure 14 for a screenshot of the task.

A.7.2 GPT Experiment

The prompt for tone feature rating that GPT receives is as outlined in 13.

B Supplementary Statistical Analyses

B.1 SP Sampling

Sample Reliability. For testing the reliability of our elicited conversational tone distribution, for both human and GPT instances, we measure the split-half correlation of their conversational tone distributions from the acquired dataset of their SP instances. This split-half correlation is computed along the following procedure. First, we randomly partition a set of SP-gathered data into two halves. Then, we find the frequency of each conversational tone within the dataset's halves. At last, we compute the correlation between the frequency of conversational tones to be the split-half correlation of conversational tone distribution within an SP instance.

Over $N = 5000$ bootstrap processes for both the human and GPT instances, we measure the human conversational tone distribution split-half correlations to be $r = 0.91$ [0.87, 0.93], and the GPT conversational tone distribution split-half correlations to be $r = 0.87$ [0.73, 0.94].

Semantic Interpretation of Sentence Space Figure 16 shows the joint-embedding space via UMAP (McInnes et al., 2020) for sentences encoded using `distilbert-base-uncased` embeddings (Sanh et al., 2019) from both humans and GPT. Consistent with our findings regarding tones in the Result section, the distribution was much more concentrated (entropy of 5.05 bits [5.03, 5.07] via bootstrapping) compared with humans (entropy of 4.12 bits [4.10,4.15]). Figure 16 also shows different topics in different parts of the space, which also shows differences in the produced sentences for humans and GPT. From this figure, we observe high repetition of sentence literal content across many GPT-occupied locations of the shared sentence embedding space (e.g., excited to go to Disneyland), while in regions dominated by humans, we usually observe a higher variance of words used. The highlighted regions in Figure 16 show the sentence space shows dense semantic topics, (e.g., "gratefulness" in circle (ii)).

B.2 Quality-of-fit Rating Experiment

Sample Reliability of Correlation Matrices. We measure the sample reliability of human perception's and GPT perception's correlation matrix using the following procedure. First, for a set of gathered quality-of-fit ratings, we randomly partition such dataset by sentences. Then, within each partition, we produce a correlation matrix. Finally, we compute the correlation of these matrices (treated as vectors). We used 5000 bootstrapped dataset. For humans ratings, we find this correlation to be $r = 0.95$ [0.92, 0.96]; for GPT ratings, we find this correlation to be $r = 0.90$ [0.84, 0.93]. The cross-domain matrix itself has a halfsplit correlation of $r = 0.95$ [0.92, 0.96].

Sample Reliability of Similarity Judgment Experiment. First, for a set of gathered quality-of-fit ratings, we partitioned the ratings of each conversational tone into two halves and computed the quality-of-fit rating correlation matrix from each half of the quality-of-fit rating data. We then compute the correlation between these similarity matrices. We used 5000 bootstraps samples and found the split-half correlation of human's similarity ma-

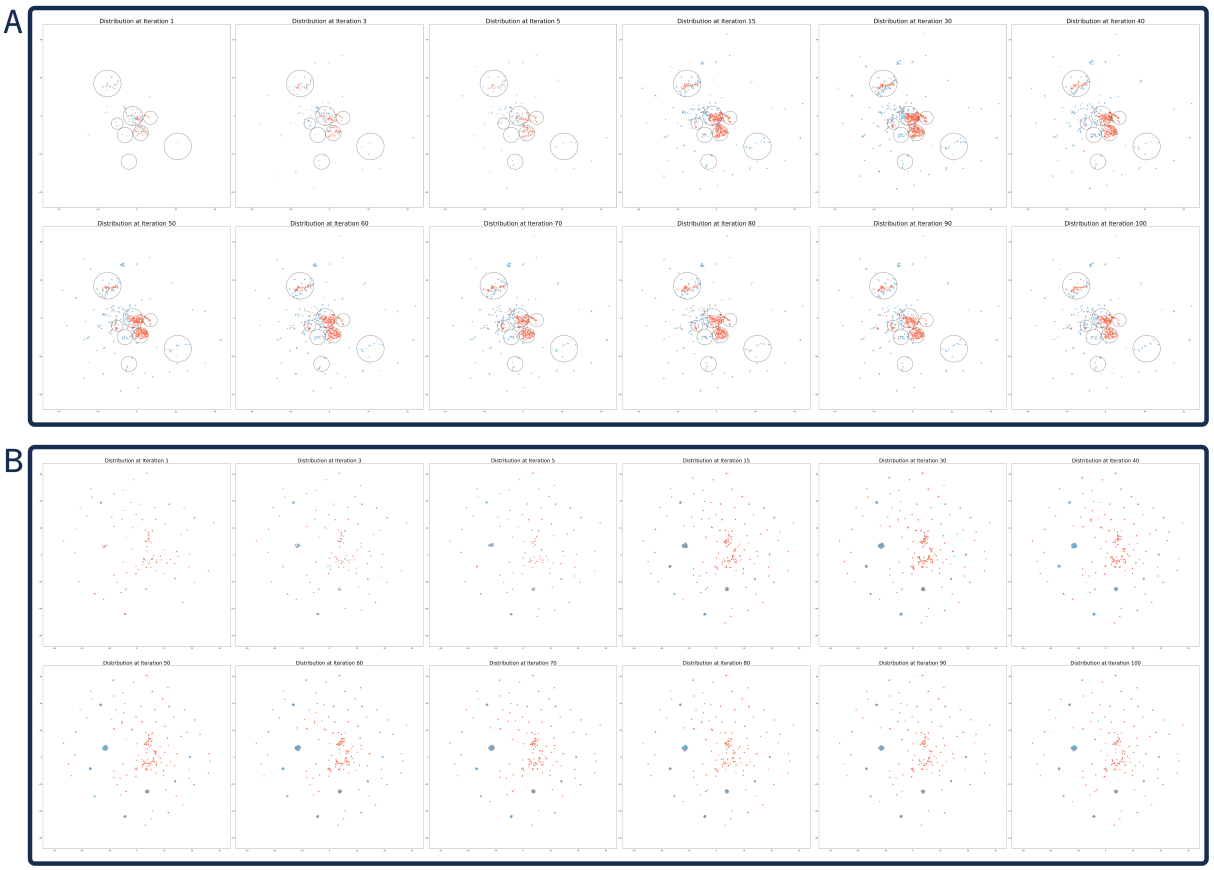


Figure 15: The dynamics of joint-embedding space for sentences and conversational tones throughout a selection of iterations. Embeddings are produced via first obtaining sentence-embeddings or tone word-embeddings using `distilbert-base-uncased` (Sanh et al., 2019) pretrained weights, and then projected onto a 2-dimensional space using `UMAP` manifold embedding (McInnes et al., 2020). **A**: The time evolution of joint-embedding sentence space. **B**: The time evolution of joint-embedding tone space.

trix to be $r = 0.72$ CI = [0.81, 0.84], and that for GPT's similarity matrix to be $r = 0.987$ CI = [0.978, 0.983].

B.3 Cross-Domain (Cross-Correlation) Analysis

Computation of Feature Arrowmarks. We used linear regression to regress the tone rating for each of the four theoretic dimensions using a projection technique and the responses of this experiment (MDS biplot (Greenacre, 2010) treating the tone feature ratings as biplot arrows as shown in Figure 3A. The concrete computation process is as follows. First, we construct a feature rating vector $\vec{f_i}$ for each vector. Then, we fit these features and the MDS embedding x coordinates ($\vec{x}$) using linear regression, arriving at some regression line: $\hat{\vec{x}} = \sum_i \alpha_i \vec{f_i}$ The coefficient α_i is then taken to be the x-axis direction of feature i's arrowmark. The same procedure was performed to compute the arrowmarks' y-axis direction. Note that when computing the arrowmark for humans' feature rating, we only fit the feature ratings to the humans' conversational tone MDS solution. GPT's arrowmark dimensions was only fitted to GPT's feature ratings too.

Cosine Similarity of Features. As performed in Section 4.3, we bootstrap over the cosine similarity of these feature vectors over different MDS solutions, and find that while the feature "informational" is consistently aligned with high cosine similarity in arrowmark direction across both groups (mean 0.98 CI = [0.97, 0.99]), the feature "relational" is not so strongly aligned (mean 0.6 CI = [0.57, 0.87]). Furthermore, the features "positive in valence" and "aroused" both observe negative cosine similarity in directions (respectively, mean -0.69 CI = [-0.71, -0.64]; mean -0.65 CI = [-0.69, -0.4]). This suggests a deviation between the human and GPT understanding of these features.

Explained Variance of Features. Additionally, we investigate the significance of each feature vec-

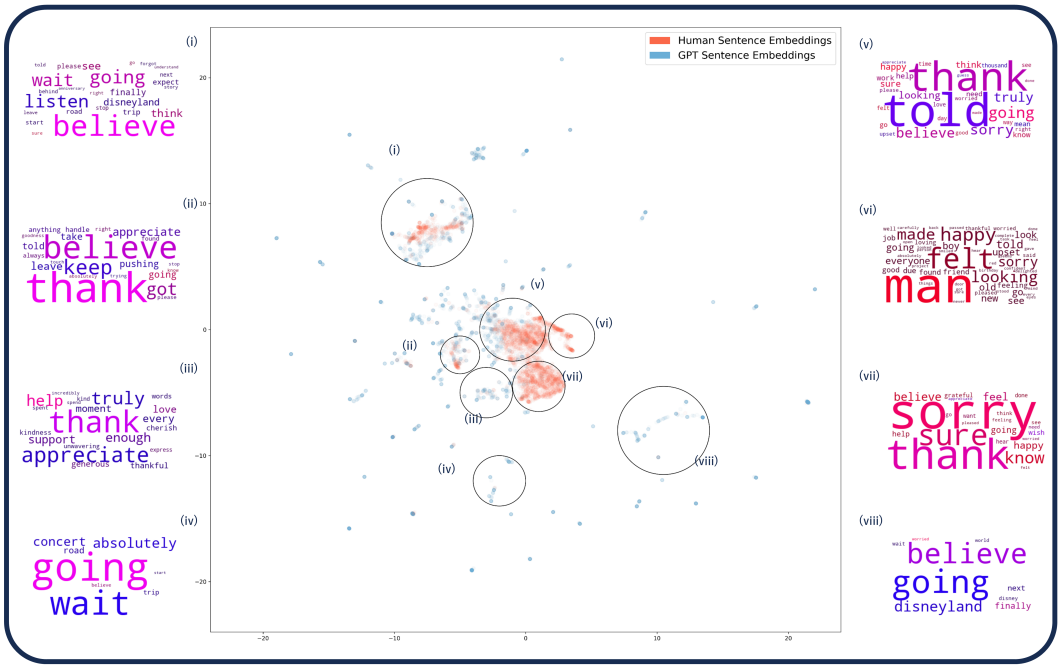


Figure 16: Sentence embedding. Word clouds show the frequency of words (right and left insets) in corresponding circles on the sentence UMAP embedding space (center). Red points resemble each sentence sampled from human instances, and blue points resemble GPT instance sentences. Brighter red and blue hues indicate respectively high TF-IDF (Luhn, 1958) scores in human, GPT sentences in each word cloud (i.e., bright purple words are highly frequent across humans and GPT).

tor by computing its "explained variance" within the shared embedding space. We compute the explained variance of a feature vector is computed as the variance of scalar projections of all MDS tone embeddings onto that feature vector. For GPT tone embeddings, the order of conversational tone features from highest to lowest explained variance is "positive in valence" (mean 0.873 CI = [0.872, 0.878]), "relational" (mean 0.45 CI = [0.42, 0.76]), "aroused" (mean 0.17 CI = [0.16, 0.24]), then "informational" (mean 0.126 CI = [0.12, 0.127]). For human tone embeddings, the order of features from highest to lowest explained variance is instead "positive in valence" (mean 0.71 CI = [0.69, 0.87]), "aroused" (mean 0.54 CI = [0.5, 0.79]), "relational" (mean 0.37 CI = [0.34, 0.76]), followed by "informational" (mean 0.3 CI = [0.27, 0.79]). In both spaces, we find "positive in valence" to be a dominant dimension of conversational tone embeddings, while humans and GPT do not fully agree upon the dominance of other directions.

C Hyperparameters in Alignment Paradigms

Gromov-Wasserstein Optimal Transport (GWOT). For GWOT, we adopted Grave et al.'s implementation (Grave et al., 2018; Conneau et al., 2017; Lample et al., 2017) using 500 iterations for its convex initiation, and a learning rate of 10, batch size of 10, regularization coefficient of 0.5, with 15 epochs for its stochastic iteration in GWOT procedure.

Bilingual Lexicon Induction via Latent Variable Model. For this method, we adopted Ruder et al.'s implementation (Ruder et al., 2018; Artetxe et al., 2016, 2017, 2018a,b). During lexicon induction, we used a backward direction, considering 5 nearest neighbors in translation retrieval. We did not use a seed dictionary. We also made small modifications to Ruder et al.'s implementation (model training batch size from 1000 to 5) to adapt the paradigm towards our smaller set of embeddings.

D Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, we sometimes used GPT for edits. After using this tool, the authors reviewed and significantly edited the content as needed and took full responsibility for the content of the publication. Additionally, we used *wordtune* https://www.wordtune.com/ and *Gram-*

Performance Category	k	Procrustes	GWOT	BLI
Domain Similarity Preservation (Human)	N/A	0.625 [0.607, 0.642]	0.625 [0.607, 0.642]	**0.8** [0.8, 0.8]
Domain Similarity Preservation (GPT)	N/A	0.499 [0.473, 0.513]	0.454 [0.405, 0.519]	**0.82** [0.82, 0.82]
kNN Matching Rate w.r.t. CC Alignment	1	0.392 [0.325, 0.463]	0.339 [0.293, 0.388]	**0.657** [0.638, 0.675]
	2	0.41 [0.363, 0.460]	0.354 [0.288, 0.397]	**0.635** [0.606, 0.65]
	3	0.456 [0.408, 0.496]	0.403 [0.350, 0.461]	**0.653** [0.635, 0.667]
	4	0.499 [0.451, 0.541]	0.461 [0.397, 0.520]	**0.705** [0.686, 0.719]
	5	0.531 [0.476, 0.568]	0.503 [0.422, 0.563]	**0.730** [0.709, 0.745]

Table 3: Table of benchmarking results on proposed metrics for unsupervised cross-domain alignment methods. Procrustes: Orthogonal Procrustes Transformation. GWOT: Gromov-Wasserstein Optimal Transport (Grave et al., 2018). BLI: Bilingual Lexicon Induction via Latent Variable Model (Ruder et al., 2018). Results are aggregated across 100 seeds for stochastic methods.

marly (https://www.grammarly.com/) to check syntax and proofread the document. Writing the experimental code we used code-suggestions by *Microsoft Copilot* ((https://copilot.microsoft.com/). We reviewed all suggestions to make sure they reflected our intentions.

E Enlarged Figures from Main Paper

In this section, we attach enlarged Figures 2, 3 as Figure 17, 18 from the main paper for readability.

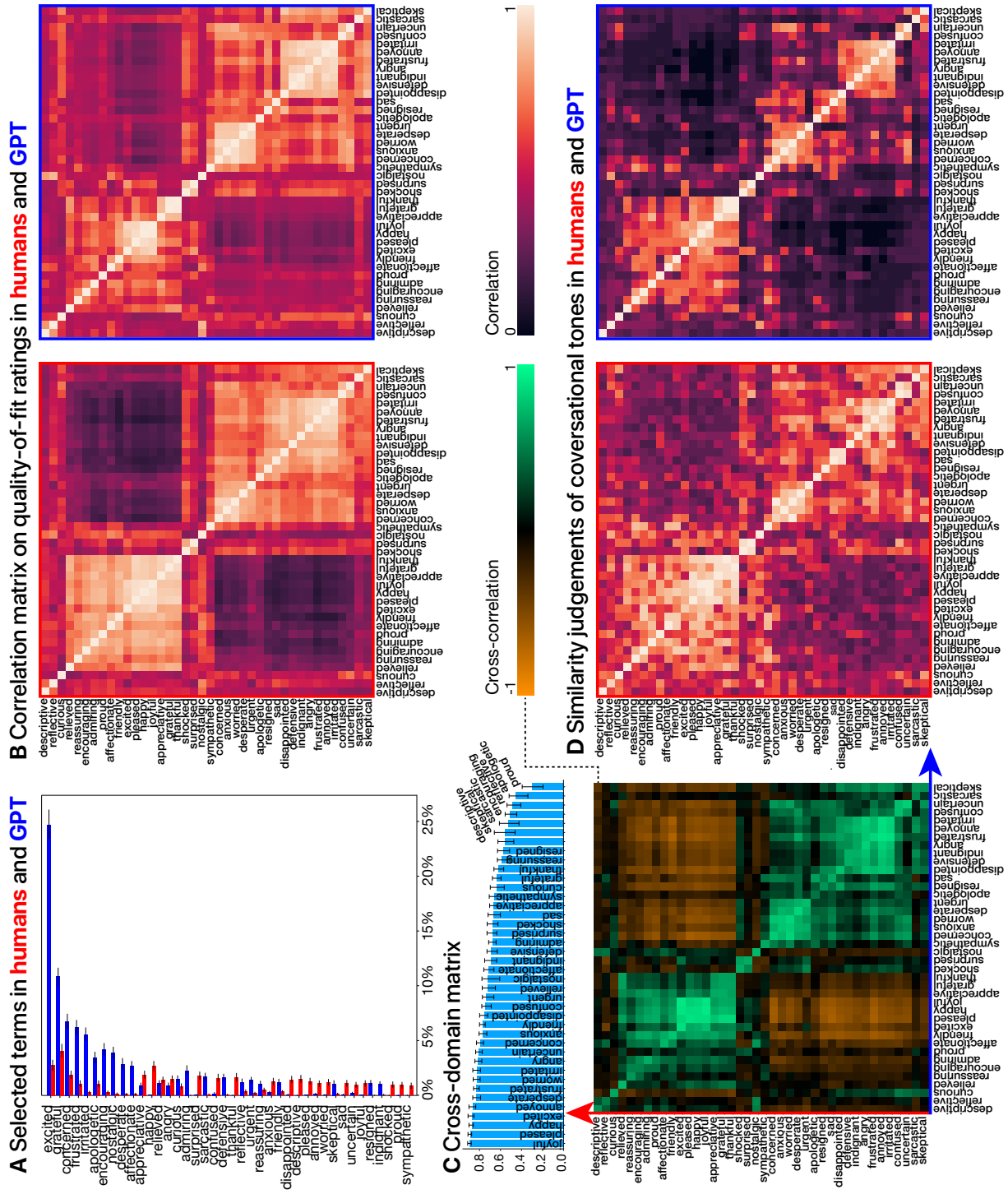


Figure 17: Enlarged version of Figure 2

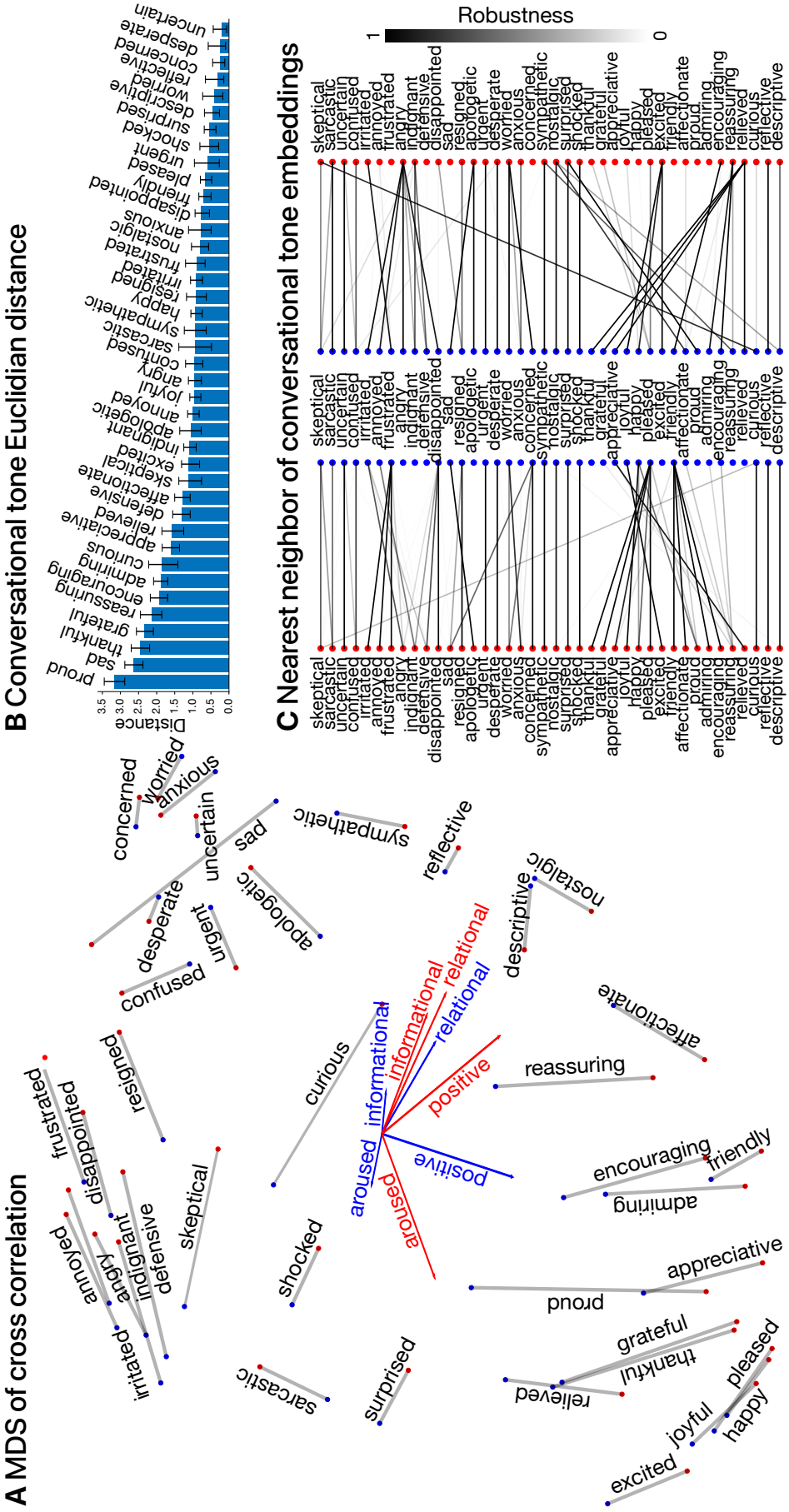


Figure 18: Enlarged version of Figure 3