Communicate to Play: Pragmatic Reasoning for Efficient Cross-Cultural Communication

Anonymous ACL submission

Abstract

In this paper, we study how culture leads to differences in common ground and how this influences communication. During communication, cultural differences in common ground during communication may result in pragmatic failure and misunderstandings. We develop our method Rational Speech Acts for Cross-Cultural Communication (RSA+C3) to resolve cross-cultural differences in common ground. To measure the success of our method, we study RSA+C3 in the collaborative referential game of Codenames Duet and show that our method successfully improves collaboration between simulated players of different cultures. Our contributions are threefold: (1) creating Codenames players using contrastive learning of an embedding space and LLM prompting that are aligned with human patterns of play, (2) studying culturally induced differences in common ground reflected in our trained models, and (3) demonstrating that our method RSA+C3 can ease cross-cultural communication in gameplay by inferring sociocultural context from interaction.

1 Introduction

An English speaker from the U.K. might refer to the storage space at the back of a car as the "boot", but an English speaker from the U.S. will likely take "boot" to mean a type of shoe. The confusion that would arise in communication between these speakers is an instance of pragmatic failure (Thomas, 1983). When humans communicate, however, they can often resolve such confusion by reasoning about the cultural background of their conversation partner, and correctly interpreting "boot" to refer to the appropriate concept. Our goal is to develop an AI system capable of pragmatic reasoning and able to adapt to new players during live interaction.

Existing research in cross-cultural communication focuses on single-turn interactions (Adilazuarda et al., 2024; Huang and Yang, 2023; He et al., 2024) or centers primarily on knowledge of cultural values or norms (Chiu et al., 2024; Huang and Yang, 2023). However, these works miss the central aspect of inferring and adapting to socio-cultural context through interaction (e.g. an American might infer that their conversation partner is British and use this to understand what the British person means when they say "boot"). To fill this gap, we introduce our method of Rational Speech Acts for Cross-Cultural Communication (RSA+C3). We study the effectiveness of our method by creating a test bed for culturally induced differences in common ground using the collaborative reference game Codenames Duet as described in Section 4.1.

First, we simulate players of Codenames Duet, using the dataset presented by (Shaikh et al., 2023) as training data for different cultures in Section 5. Then, we show that these simulated players can reflect the cultural differences present in the dataset in Section 6. Finally, we test how well our simulated players of different cultures can play Codenames with each other Section 7. Through these interaction experiments, we show that our method RSA+C3 can significantly improve the win rates of games of Codenames Duet over our baseline, showing that it is inferring socio-cultural context from the interaction.

2 Related work

We first discuss previous work that has expanded on the Rational Speech Acts framework (Degen, 2023; Goodman and Frank, 2016) and language games as a method of analyzing human dialogues, specifically in the context of conveying information concisely based on shared context.

Culture in NLP. Much work has been done to model cross-cultural differences using LLMs. State-of-the-art LLMs have been shown to struggle with multi-cultural reasoning (Chiu et al., 2024).

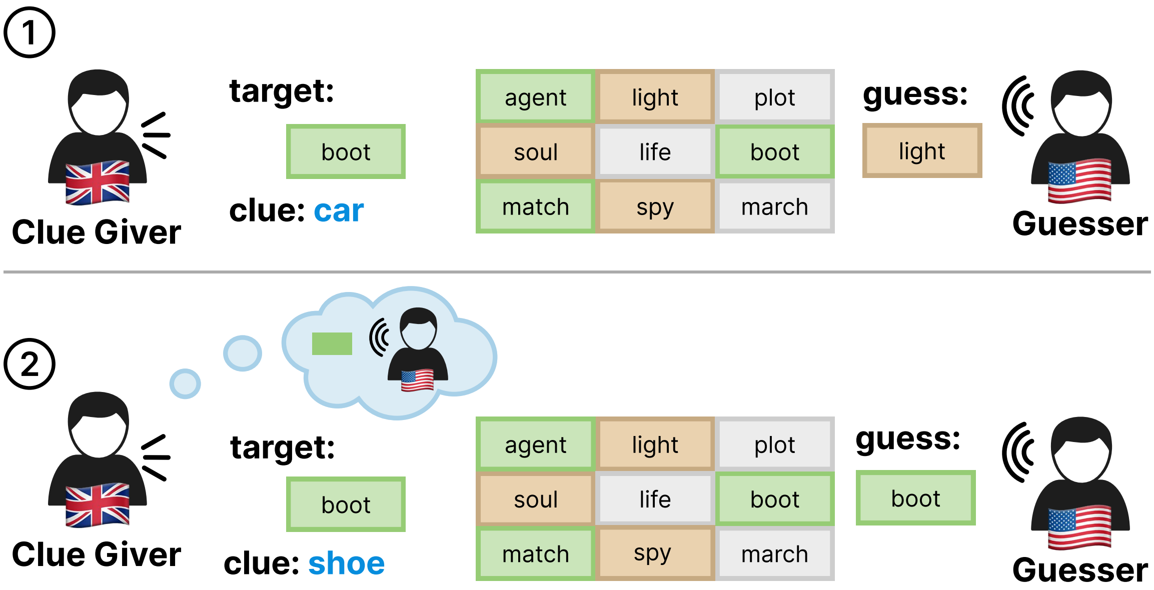


Figure 1: **RSA+C3: Rational Speech Acts framework with Cross-Cultural Communication**. Here we model interactions in Codenames Duet between the British clue giver and the American guesser. (1) In regular gameplay, the clue giver selects a **target** and generates a **clue** without considering the guesser's background. (2) Using RSA+C3, the giver considers what word the guesser may select based on their demographic background and generates a different **clue** accordingly. The **avoid** words will cause the game to end in an immediate loss and the **neutral** words have no effect on the success or failure of the game.

Though prompted LLMs might reflect some understanding of cultural norms, they fail to apply reasoning to downstream inferences (e.g. inferring differences in tip culture) (Huang and Yang, 2023) often producing toxic or heavily stereotyped text. Prompting such as in Niszczota and Janczak (2023) is not the only method to personalize LLMs, LLMs can be personalized using influence functions (He et al., 2024), fine-tuning (Li et al., 2024a). Culturally personalized LLMs provide a useful tool for content moderation (He et al., 2024; Li et al., 2024a,b) or sharing multi-cultural knowledge (Li et al., 2024b). Moreover, recent dataset and benchmark efforts (Fung et al., 2024) record a wide diversity of cultural norms. However, these papers focus mostly on norms and values (such as cultural traditions) rather than on the common ground shared between members of a culture. Norms and values refer to culturally correlated beliefs, whereas common ground refers to the assumed shared knowledge base. In contrast to the prior work, we seek to evaluate our models in their ability to infer socio-cultural differences in common ground through multi-turn interactions.

Applications of RSA and Pragmatic Reasoning Previous work has incorporated context in the use of priors for modeling utterances via RSA, such as in using the perspective of a speaker to interpret motion verbs (e.g. "come" and "go") (Anderson and Dillon, 2019) and modeling connectives in utterances (e.g. "but" and "therefore") (Yung et al., 2016). RSA has also been studied as a model of human behavior through reference games, such as in differentiating ambiguous images via minimally distinguishing information (Frank, 2016). Beyond reference games and connective utterances, RSA has been used to study discourse, particularly in the use of indirect or polite phrases (Lumer and Buschmeier, 2022). Pragmatic reasoning plays a role in the arguments made during meetings of the UN (Kone, 2020), where the ambassadors reason about the context of the others. The framework of RSA assumes that common ground is shared between parties. Degen et al. (2015) adds an additional component where the probability of common ground not being shared is estimated and use to change predictions. However, they primarily use a high entropy backoff distribution to perturb predictions. For our method RSA+C3 in Section 3.2, we develop a way to utilize prior socio-cultural information (e.g. a person is British) to improve predictions.

Language Games for AI Language games have been frequently used as a test-bed for artificial intelligence and human-AI interaction (Hausknecht et al., 2020; Ammanabrolu et al., 2022; Wang et al., 2022). Previous work explored how language models interact in realistic social environments based on choose-your-own-adventure games, finding that agents could be steered towards valuing moral requirements rather than trading them off for greater rewards (Pan et al., 2023). Codenames has been studied in the simplified format of "Codenums", which replaced words with vectors to study non-linguistic attributes of the game via a deductive agent hierarchy that tracks the internal models of other players (Bills and Archibald, 2023). Clues for the game have been generated by ranking based on document frequency and existing word embedding models (Koyyalagunta et al., 2021). Sociolinguistic priors have been generated to account for the cultural context of the speaker in the simplified game "Codenames Duet" (Shaikh et al., 2023). We explore incorporating the speaker's sociocultural attributes across a varying set of games to explore how transferable these priors are and when this additional context could be clarifying versus superfluous.

3 Pragmatic Reasoning with the RSA Framework and RSA+C3

We formalize and describe the RSA framework as articulated in Degen (2023) and an extension to RSA used to represent differences in common ground. RSA formulates communication as a conversation between a listener and a speaker. For Codenames Duet, we treat the literal listener as the guesser and the pragmatic giver as the clue giver.

3.1 RSA: Rational Speech Acts Framework

In RSA formulations, the (abstract) *literal listener* L_0 interprets meaning based on literal semantics. In the context of Codenames Duet, this is equivalent to a guesser guessing to optimize semantic similarity. The *pragmatic speaker* or clue giver S_1 reasons about the literal listener by

$$P_{S_1}(c|g) \propto \exp(\alpha \cdot T(c|g))$$

$T(c|g)$ represents the utility of c for communicating target concepts g. T is a trade-off between the cost of an utterance and the informativeness of c.

$$U(c, g) = \ln\left(P_{L_0}(g|c) - \text{cost}(c)\right)$$

We will take the cost of the clue to be equivalent to the possibility of the guesser, or literal listener, choosing an avoid word (a word that will end the game) or a neutral word (a word that doesn't belong to any player's team).

3.2 RSA+C3: Rational Speech Acts for Cross-Cultural Communication

The RSA framework in Section 3.1 formalizes efficient communication, but does not account for instances where common ground is not shared. We introduce RSA+C3, a method that assumes that common ground is not shared and learns to interact with an interlocutor of a different culture through live interaction. To accomplish this, we provide the RSA+C3 pragmatic speaker S_1 with n different models representing literal listeners L_i of n different cultures. For each culture, we store a random variable w_i where $P(w_i)$ reflects the probability that the interlocutor shares the same culture, taking inspiration from (Degen et al., 2015). We estimate the probability $P(w_i)$ by calculating the probability that utterance g would have been chosen if the interlocutor shares the same culture and clue c was given. Let g be the utterance observed then we estimate:

$$P(w_i) = P_{L_i}(g|c, w_i)$$

Then, we select a literal listener L_i or guesser from the possible n cultures by finding the culture that maximizes $P(w_i)$ and estimate

$$P_{S_1}(c|g) \propto \exp(\alpha \cdot \ln(P_{L_i}(g|c) - \text{ cost } (c)))$$

Thereby selecting a clue c to maximize informativeness to a listener belonging to a culture i.

4 Task Data and Metrics

We introduce the dataset, game, and metrics we utilize in this paper to model cross-cultural communication.

4.1 Codenames Duet

Codenames Duet is a complex referential collaborative game featuring a *clue giver* and a *guesser* where the clues and guesses given are based on an assumption of common ground. The board consists of 25 words, nine *goal* words, three *avoid* words, and 13 *neutral* words. To win the game, the guesser must guess all *goal* words without guessing any *avoid* words. In a single turn, the *clue giver* chooses a subset of the *goal* words as their *targets* and provide a one-word clue that the guesser uses to guess the *target* words.

4.2 Dataset

To run our experiments, we utilize Codenames Duet and the Cultural Codes [1] dataset, which contains 794 Codenames Duet games across 153 players, along with survey results containing demographic information about each player (Shaikh et al., 2023).

4.3 Metrics

As we use LLMs and the word embedding space to simulate interactions in Codenames, we explore our modeled givers and guessers' alignments with human data from the dataset described in Section 4.2.

Giver metrics. In a single round, the clue giver must (1) select a set of target words from the goal words and (2) generate a clue to distinguish the intended targets from other words on the board. We define metrics for these two tasks:

- **Giver target accuracy** is the proportion of the human giver's target words that are also generated by the simulated giver.

$$\frac{\text{\# giver-aligned simulated targets}}{\text{\# human giver targets}}$$

- **Clue accuracy** is the proportion of the human giver's clues that are also generated by the simulated giver.

$$\frac{\text{\# giver-aligned simulated clues}}{\text{\# human giver clues}}$$

We sum the number of targets and clues across multiple rounds.

Guesser metrics. In a single round, the guesser selects words from the board that they believe correspond best to a given clue. We define metrics to study how well our simulated guesser aligns with both the behavior of the human guesser and the intentions of the human giver:

- **Guess accuracy** is the proportion of human guesses that are also generated by the simulated guesser.

$$\frac{\text{\# guesser-aligned simulated guesses}}{\text{\# human guesser guesses}}$$

- **Guesser target accuracy** is the proportion of targets intended by the human giver that are guessed by the simulated guesser.

$$\frac{\text{\# giver-aligned simulated guesses}}{\text{\# human giver targets}}$$

[1] https://github.com/SALT-NLP/codenames

As with the giver metrics, we sum the number of guesses and targets across rounds.

4.4 Interactive Evaluation

In our paper, our goal is to evaluate how simulated players of different cultures interact and collaborate to play Codenames Duet. Since Codenames Duet is a collaborative game, the main metric for whether two players are effectively communicating is the **win rate**. To ensure that a method does not increase the win rate simply by being evaluated on easier boards, we generated a fixed set of 100 boards and play a game on each board. We explain this further in Appendix E.

5 Modeling Codenames Players with Word Embeddings and LLMs

We explore two approaches to modeling our giver and guesser; trained word embeddings and prompting LLMs. We find that our giver and guesser based on word embeddings consistently outperform the few-shot prompted LLMs in accuracy on the human-selected guesses and targets, as illustrated in Figure 2.

5.1 Modelling the Guesser and Giver using Word Embeddings

The embeddings-based *literal guesser* selects the most likely words based on cosine similarity between the given clue c and the set of unseleted words U. For each unselected word u in U, the cosine similarity is given by

$$sim(c, u) = \frac{c \cdot u}{|c||u|}$$

Then for the literal guesser, we estimate

$$P_{L_0}(g|c) = \frac{\exp(sim(c, g))}{\sum_{u \in U} \exp(sim(c, u))}$$

and we select the g to be such that it maximizes $P_{L_0}(g|c)$. Similarly, we implement the embeddings-based *literal giver* by finding the clue c for target g such that the similarity between c and g is maximized.

$$c = \arg\max_{c} sim(c, g)$$

Finally, we select the target concept g by selecting

$$g = \arg\max_{g} \arg\max_{c} sim(c, g)$$

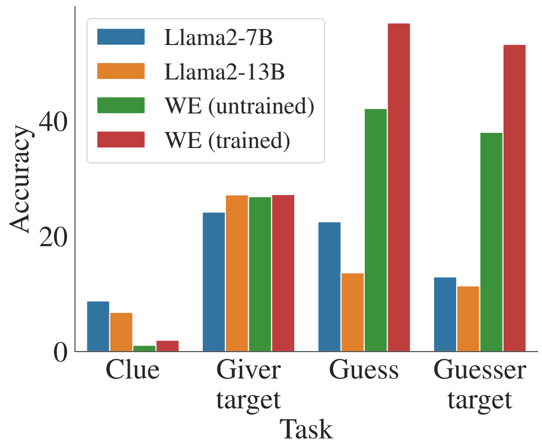


Figure 2: **Player modeling using LLM-prompting and trained word embeddings**. The efficacy of the Llama2 chat models at simulating human players, including both the giver and guesser, varied across model size and task. Trained word embeddings consistently outperformed untrained word embeddings and generally outperformed LLM-prompting with the exception of the giver clue selection task.

5.2 Training Word Embeddings

To train our word embeddings we use a linear layer f_θ on top of the GloVe model (Pennington et al., 2014) and compute the embedding of a word x as

$$\mathrm{E}(x) = f_\theta(\mathrm{GloVe}(x))$$

During training, we aim to model the lexicon of human players by increasing the similarity between the clue and the words selected by the humans while decreasing the similarity with other words on the board.

We formalize each turn as consisting of a clue c, a set of available words $\{w_1, \ldots, w_n\}$, and a set of selected words $S \subseteq \{1, \ldots, n\}$. The training objective is then defined as

$$\text{loss} = -\frac{1}{|S|} \sum_{i=1}^{n} \log \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)} \mathbb{1}\{i \in S\}$$

where u_i is the cosine similarity between w_i and c, scaled by temperature t:

$$u_i = \frac{\mathrm{E}(w_i) \cdot \mathrm{E}(c)}{|\mathrm{E}(w_i)||\mathrm{E}(c)|} \times \exp(t)$$

This objective is equivalent to a cross-entropy loss with equal probabilities across each selected word, and is modeled after the contrastive loss used in Radford et al. 2021.

5.3 Guesser and Giver Prompting

We chose to model the giver and guesser in Codenames using the Llama2 family of text and chat models (Touvron et al., 2023) due to these models being open-source.

We explore their models' accuracy across the metrics defined in Section 4.3 with few-shot prompts.

Giver. We first query the Llama2 chat models to generate a clue using a few-shot prompt as described in Appendix A.1. To allow for a diverse set of potential clues, we generated 5 clues per prompt, allowing for repeats. The clue giver then selects a target word for the guesser to select conditioned on the board state, as described in Appendix A.2.

Guesser. Using a provided clue, we model the codenames guesser by prompting a Llama2 chat model with:

```
You are playing Codenames and are the
    clue guesser. You need to select one
     word from {all words}. Given the
    clue {clue}, the most likely word is
```

We calculate the probability of a target word being generated from the list of possible target words as described in Appendix A.2.

6 Incorporating Cultural Context into Player Models

To model cross-cultural communication in Codenames Duet, we must first train models to reflect the cultural background of human players. In Section 6.1, we do this by training word embeddings using the technique described in Section 5.2 on data representing a specific demographic attribute (e.g. education). In addition, we demonstrate how few-shot prompting with cultural context can lead to higher performance - highlighting the influence of cultural priors on codenames play.

6.1 Training embedding spaces with cultural splits

To model players with different cultural backgrounds, we contrastively train embeddings using the technique in Section 5.2 on subsets of the Cultural Codes dataset. We split the dataset into subsets based on various demographic and cultural attributes. We split the dataset along the axes of education (high school & associate, bachelor, graduate), country (United States, foreign), native (true, false), political (liberal, conservative), age (under 30, over 30), and religion (Catholic, not Catholic). For some subsets of the dataset, we group the values of the cultural variables to obtain subsets with roughly equal amounts of data. We follow the pro-

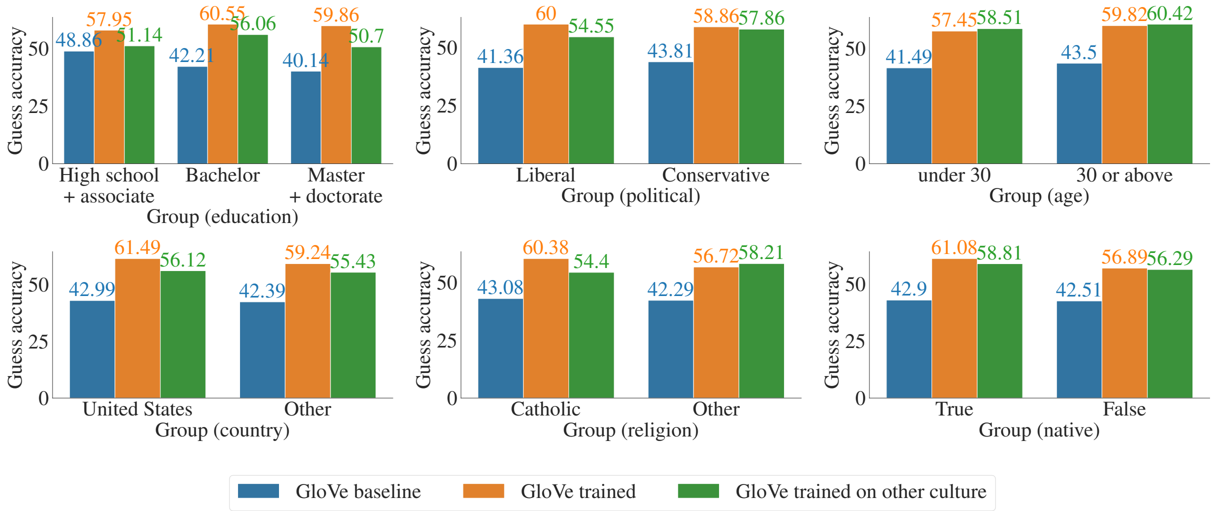


Figure 3: **Comparison of guess accuracy using embeddings trained on cultural splits against baseline GloVe embeddings and embeddings trained on different splits.** The large difference of 9% on the data of Master+Doctorate cultural split, between the GloVe trained on Master+Doctorate and GloVe trained on the remaining data (i.e. the difference between the orange and green bars) indicates that there are cultural patterns found in the Graduate+Bachelor data that do not occur in the remaining data. There are similar large differences in accuracy between GloVe trained on split and GloVe trained on the other split in the cultural splits on country and politics.

cedure described in Appendix D, training for 25 epochs.

After training our embeddings, we evaluate the alignment of a literal guesser using these embeddings with the human guesses found in the hold-out validation set. The humans in the validation set are not the same humans in the training set, indicating that our predictions are extendable to other humans of a similar cultural background. Our results are displayed in Figure 3, with additional results in Appendix D.

6.2 Few-shot prompting with cultural context

We study how different axes of demographics included in the Cultural Codes dataset could inform alignment to the human guesser and the giver, with the LLM simulating the player. In both paradigms, we prompt the Llama2 chat models (Touvron et al., 2023) with a list of unselected words and a provided clue, asking the model to output the most likely target word. We provide information about the clue giver, as described in Appendix A.3, and study how often the model's giver alignment and guesser alignment. As illustrated in Figure 4, we find that including any demographic information improved alignment with the human guesser for the Llama-2-7B-Text model. Results vary for giver alignment and the 13B-Text model. Moreover, when studying the inclusion of cultural context in clue generation, we find that inclusion of all demographics increased performance in the 13B model while "leaning" (the political leaning and personality scores of the human players) increased performance for the 7B model, as shown in Figure 5. The increased performance under different cultural prompts underlines how cultural context influences the choices of the human guessers and givers in the dataset.

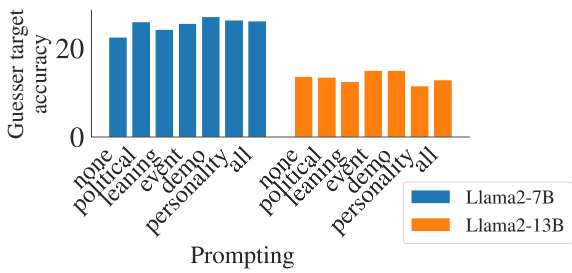


Figure 4: **Target guessing with cultural context**. Reranking potential target words based on the probabilities output by the Llama2 model simulating the clue giver and word guesser led to varying levels of guesser-aligned target word selections. Inclusion of cultural context (e.g. political leaning, personality) sometimes improved alignment with the guesser based on model size and selected demographic.

7 Cross-cultural Pragmatic Reasoning in Interaction

In Section 5 we demonstrated that a learned embedding space can accurately reflect human guesses from the Shaikh et al. (2023) dataset. In Section 6 we demonstrated how these models can reflect the preferences of different cultures. In this section, we aim to show how the RSA and RSA+C3 methods can improve performance for codenames players

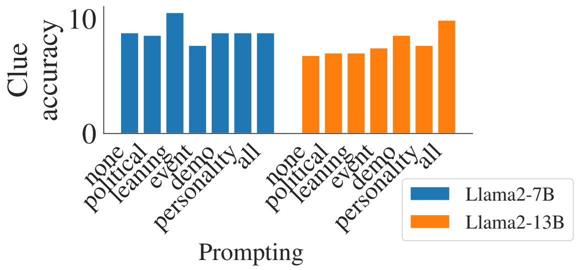


Figure 5: **Clue generation with cultural context**. Leaning notably led to an increase in accuracy for giver alignment for the 7B model while including all demographics for the 13B model led to more accurate giver-aligned generations.

of different cultures over our baseline literal player.

7.1 Clue Givers

To highlight the necessity of pragmatic reasoning, we introduce our three techniques for modeling the clue giver - the literal, RSA, and RSA+C3 clue givers.

Literal Clue Giver. We evaluate the literal clue giver as described in Section 5.1 that selects the clue c that is most similar in semantic similarity to the target g.

RSA Clue Giver. Recall from Section 3.1 how we defined P_{S_1} to be the probability distribution governing the actions of the pragmatic speaker. In Codenames Duet, the pragmatic speaker is the pragmatic clue giver. The clue giver must select the best clue c for the target concept g. The cost of the clue c is the probability that the guesser will instead guess avoid words $a \in A$ or neutral words $n \in N$. Therefore using P_{L_0} to refer to the probability distribution of the literal guesser we use

$$P_{S_1} \propto \exp(\alpha \cdot (\ln P_{L_0}(g|c) - \text{cost}(c))) \quad (1)$$

where

$$\text{cost}\ (c) = \max_{a \in A} P_{L_0}(a|c) - \delta \max_{n \in N} P_{L_0}(n|c) \quad (2)$$

where we introduce a neutral constant δ that governs how much to penalize the neutral words.

RSA+C3 Clue Giver. As we discuss in Section 3.2, the RSA method described does not account for differences in common ground, or in other words, culturally introduced differences in $P_{L_0}(g|c)$. As a result, we provide n word embedding models to model n distributions $P_{L_i}(g|c)$. We select culture L_i such that it maximizes $P(w_i)$ the posterior probability of the observed interactions if culture i is shared.

$$P(w_i) = P_{L_i}(g|c, w_i) \quad (3)$$

However, a critical component of modeling this for Codenames Duet, is that there must be memory of previous interactions. Therefore w_i is a smoothed average with smoothing constant β of the estimates $P(w_i)$ after each literal guesser L_i utterance. Therefore we update

$$P(w_{i_{\text{new}}}) = \beta \cdot P(w_{i_{\text{old}}}) + (1 - \beta) P_{L_i}(g|c, w_i)$$

We then estimate P_{S_1} the same way as in *eq.* (1) but using P_{L_i} so

$$P_{S_1}(c|g) \propto \exp(\alpha \cdot (\ln P_{L_i}(g|c) - \text{cost}(c)))$$

Then we select our clue to be

$$c = \arg\max_{c} P_{S_1}(c|g)$$

7.2 Interactive Evaluation Results

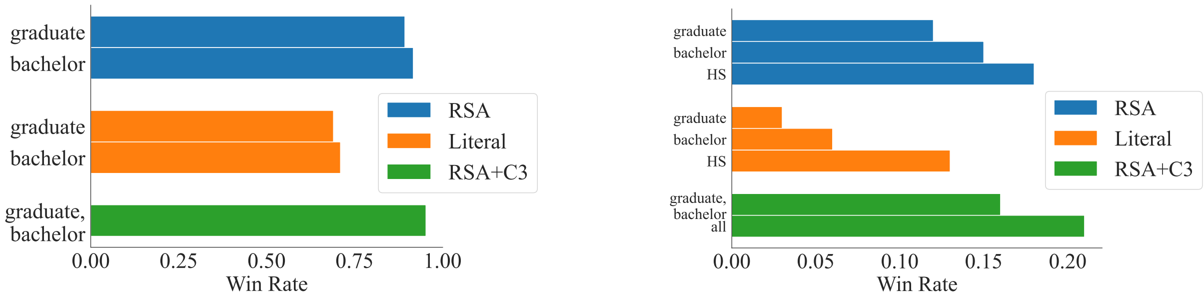
As described in Section 4.4, we evaluate the performance of two players of different cultures during interaction. To do this, we select the demographic in the dataset such that simulated players have the largest cultural difference as observed in Figure 3 - education.

We evaluate our literal, RSA, and RSA+C3 clue givers against two different guessers: a guesser trained to reflect a player with a high school or associates degree and llama-7b-chat prompted as described in Section 5.3. We evaluate with the llama-7b-chat-based guesser to simulate an unknown culture that the clue giver must adapt to. To ensure that players reflect different cultures we evaluate simulated players with a graduate or undergraduate degree when playing against the player with high school degree.

While the inclusion of the traditional RSA framework leads to significant improvements in contrast to the literal giver, our results demonstrate that including pragmatic reasoning and cross-cultural communication via RSA+C3 leads to a greater win rate regardless of whether the guesser is trained word embeddings or a prompted LLM.

8 Discussion

Using Codenames Duet as a testbed for studying cross-cultural communication, we demonstrated



(a) Word Embedding (High School) Guesser

(b) Llama2-Text-7B Guesser

Figure 6: **Interactive Evaluation across RSA, Literal, and RSA+C3 Guessers.** We evaluate RSA, Literal, and RSA+C3 givers across guessers simulated by word embeddings trainings and LLM prompting. In Figure 6a, we study interactions with a word embeddings guesser trained on data belonging to players whose highest level of education completed was high school. The "graduate, bachelor" RSA+C3 giver achieved the highest win rate, greater than RSA givers initialized on either "graduate" or "bachelor" alone. We used an LLM-prompted guesser in Figure 6b and found that the RSA+C3 giver initialized with all provided education options ("graduate, bachelor, HS") achieved the highest win rate, outperforming all RSA and Literal givers.

that our simulated players are capable of reflecting human gameplay and their sociocultural patterns. We utilize our player models reflecting different sociocultural backgrounds to emulate pragmatic failure in live gameplay. This enables us and future researchers to measure the collaborative ability between agents of different backgrounds - if the win rate of Codenames Duet is higher, then the difference in common ground is more easily overcome.

As the full complexity of cross-cultural communication cannot only be captured through Codenames Duet, directions for future work include applying these techniques to more complex utterances with more nuanced cultural differences and studying the resulting interactive gameplay.

Overall, we find that introducing cultural context as a way for givers and guessers to communicate in Codenames Duet gameplay increases alignment with human data based on the subset of culture involved. Our results across various methods of simulating players and different cross-sections of demographics demonstrate the significance of continuing to study the impact of cultural context in speaker and listener communication.

9 Limitations

In our paper, we train models to reflect various cultural attributes as shown in fig. 3 and evaluate our method RSA+C3 to resolve pragmatic failure due to cultural differences such as education level in fig. 6. However, the cultures are not equally represented in the cross-cultural codes dataset (Shaikh et al., 2023) we used with the participants being majority White (78%) and liberal (58%). Therefore some cultural differences are not as pronounced as they would be in a more balanced dataset.

10 Broader impacts statement

While cultural context can be a useful tool in informing clue generation and target selection in games like Codenames, we caution against leaning heavily on these demographics due to the potential for stereotype-based associations. Previous work has demonstrated the propensity for language models to incorporate biases into generations (Kotek et al., 2023). Although we are interested in seeing future work explore how culture can inform communication, allowing for both speakers and listeners to update their mental models of the other conversational participant, we acknowledge that leaning too heavily on these demographics can lead to potentially harmful assumptions.

References

Muhammad Farid Adilazuarda, Sagnik Mukherjee, Pradhyumna Lavania, Siddhant Singh, Ashutosh Dwivedi, Alham Fikri Aji, Jacki O'Neill, Ashutosh Modi, and Monojit Choudhury. 2024. Towards measuring and modeling" culture" in llms: A survey. *arXiv preprint arXiv:2403.15412*.

Prithviraj Ammanabrolu, Liwei Jiang, Maarten Sap, Hannaneh Hajizhirzi, and Yejin Choi. 2022. Aligning to social norms and values in interactive narratives. In *North American Chapter of the Association for Computational Linguistics (NAACL)*.

Carolyn Jane Anderson and Brian W. Dillon. 2019. Guess who's coming (and who's going): Bringing perspective to the rational speech acts framework. *Proceedings of the Society for Computation in Linguistics*, 2(20):185–194.

Joseph Bills and Christopher Archibald. 2023. A deductive agent hierarchy: Strategic reasoning in codenames. In *2023 IEEE Conference on Games (CoG)*, pages 1–8.

Yu Ying Chiu, Liwei Jiang, Maria Antoniak, Chan Young Park, Shuyue Stella Li, Mehar Bhatia, Sahithya Ravi, Yulia Tsvetkov, Vered Shwartz, and Yejin Choi. 2024. Culturalteaming: Ai-assisted interactive red-teaming for challenging llms'(lack of) multicultural knowledge. *arXiv preprint arXiv:2404.06664*.

Judith Degen. 2023. The rational speech act framework. *Annual Review of Linguistics*, 9:519–540.

Judith Degen, Michael Henry Tessler, and Noah D Goodman. 2015. Wonky worlds: Listeners revise world knowledge when utterances are odd. In *CogSci*.

Michael C Frank. 2016. Rational speech act models of pragmatic reasoning in reference games.

Yi Fung, Ruining Zhao, Jae Doo, Chenkai Sun, and Heng Ji. 2024. Massively multi-cultural knowledge acquisition & lm benchmarking. *arXiv preprint arXiv:2402.09369*.

Noah D. Goodman and Michael C. Frank. 2016. Pragmatic language interpretation as probabilistic inference. *Trends in Cognitive Sciences*, 20(11):818–829.

Matthew Hausknecht, Prithviraj Ammanabrolu, Marc-Alexandre Côté, and Xingdi Yuan. 2020. Interactive fiction games: A colossal adventure. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7903–7910.

Jerry Zhi-Yang He, Sashrika Pandey, Mariah L Schrum, and Anca Dragan. 2024. Cos: Enhancing personalization and mitigating bias with context steering. *arXiv preprint arXiv:2405.01768*.

Jing Huang and Diyi Yang. 2023. Culturally aware natural language inference. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7591–7609.

Nouhoum Kone. 2020. Speech acts in un treaties: A pragmatic perspective. *Open Journal of Modern Linguistics*, 10(6):813–827.

Hadas Kotek, Rikker Dockum, and David Sun. 2023. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, CI '23, page 12–24, New York, NY, USA. Association for Computing Machinery.

Divya Koyyalagunta, Anna Y. Sun, Rachel Lea Draelos, and Cynthia Rudin. 2021. Playing codenames with language graphs and word embeddings. *CoRR*, abs/2105.05885.

Cheng Li, Mengzhou Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024a. Culturellm: Incorporating cultural differences into large language models. *arXiv preprint arXiv:2402.10946*.

Cheng Li, Damien Teney, Linyi Yang, Qingsong Wen, Xing Xie, and Jindong Wang. 2024b. Culturepark: Boosting cross-cultural understanding in large language models. *arXiv preprint arXiv:2405.15145*.

Eleonore Lumer and Hendrik Buschmeier. 2022. Modeling social influences on indirectness in a rational speech act approach to politeness. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Paweł Niszczota and Mateusz Janczak. 2023. Large language models can replicate cross-cultural differences in personality. *arXiv preprint arXiv:2310.10679*.

Alexander Pan, Jun Shern Chan, Andy Zou, Nathaniel Li, Steven Basart, Thomas Woodside, Jonathan Ng, Hanlin Zhang, Scott Emmons, and Dan Hendrycks. 2023. Do the rewards justify the means? measuring trade-offs between rewards and ethical behavior in the machiavelli benchmark. *ICML*.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Omar Shaikh, Caleb Ziems, William Held, Aryan J. Pariani, Fred Morstatter, and Diyi Yang. 2023. Modeling cross-cultural pragmatic inference with codenames duet.

J. Thomas. 1983. Cross-Cultural Pragmatic Failure. *Applied Linguistics*, 4(2):91–112.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Ruoyao Wang, Peter Jansen, Marc-Alexandre Côté, and Prithviraj Ammanabrolu. 2022. Scienceworld: Is your agent smarter than a 5th grader? *arXiv preprint arXiv:2203.07540*.

Frances Yung, Kevin Duh, Taku Komura, and Yuji Matsumoto. 2016. Modelling the usage of discourse connectives as rational speech acts. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, pages 302–313, Berlin, Germany. Association for Computational Linguistics.

A Experiment details for simulating givers and guessers using LLMs

Here we elaborate on the framework for our experiments in clue and target selection using the Llama2 family of LLMs, as described in Section 5. For all of the following experiments, we used default hyperparameters as provided in the open-source Llama2 code [2] and model sizes of 7B and 13B. The following experiments were conducted over the validation set of the Cultural Codes dataset.

A.1 Clue generation

We prompted the 7B and 13B Llama2-Chat models to generate clues using the following few-shot prompt, allowing for a flexible free-form text generation informed by prior examples of a Codenames-style clue:

```
You are playing Codenames. You can only
    give clues which are one word. One
    clue will apply to multiple targets.
     Words to avoid are {avoid words}.
    Neutral words are {neutral words}.
    For the group of target words ['fall
    ', 'spring', and 'leaf'] the best
    clue is 'season'. For the group of
    target words ['round', 'cylinder']
    the best clue is 'circle'. For the
    target words {target words} the best
     clue is '
```

The target words were preselected from the Cultural Context dataset, allowing us to study the LLM's alignment with a human clue giver.

A.2 Target selection

Using the Llama2 Text models, we used the following prompt to extract potential target words.

```
You are playing Codenames and need to
    select a target word for your
    partner to guess. Words to avoid are
     {avoid words}. Neutral words are {
    neutral words}. Goal words are {goal
     words}. The best target word for
    your partner to guess is '
```

As the game is constrained to selecting target words from the set of goal words, we calculated the probability of the model generating each of the goal words as the completion to the prompt, then identified the most probable generations as the selected target words.

A.3 Target word selection under cultural context

We prompted the Llama2 Text models with the following prompt, optionally including the giver's demographics. Similar to our experiment with target selection in Appendix A.2, we selected the generation under the set of possible target words (i.e. restricted to the set of goal words) that had the highest probability.

```
You are playing Codenames. The possible
    words are {words}. Here is some
    information about the clue giver: {
    cultural context}. For the hint {
    clue}, the most likely target word
    is
```

As demographics were verbose, we provided them as a comma-separated list of values. For example, one possible prompt addition could be:

```
Here is some information about the clue
    giver: age: 29, gender: female,
    country: united states, native: true
    .
```

The demographics we used in Figure 4 consist of the demographic questions in the Cultural Codes dataset in Appendix D.2. We additionally extracted the political context from the broader political leaning category (abbreviated in the figure as "leaning").

Notably, we calculated accuracy for giver alignment versus guesser alignment with separate target words. Alignment with the giver meant selecting target words that were intended by the human giver for the guesser to select. Alignment with the guesser meant selecting target words that the human guesser selected given a similar set of information as provided in the prompt above, regardless of the giver's original intentions. As multiple target words could be selected per round, we computed the accuracy as the total number of correct target words divided by the total number of intended target words. Full results for both giver and guesser alignment can be found in Figure 7.

A.4 Clue generation under cultural context

We iterated on our clue generation experiments from Appendix A.1 by using a similar approach to Appendix A.3, drawing pre-specified demographics for the guesser to inform the giver's clues. We generated prompts of the following format:

```
You are playing Codenames. You can only
    give clues which are one word. One
    clue will apply to multiple targets.
     Words to avoid are {avoid words}.
    Neutral words are {neutral words}.
    Here is some information about the
    clue guesser: {cultural context}.
    For the group of target words ['fall
    ', 'spring', and 'leaf'] the best
    clue is 'season'. For the group of
```

[2] https://github.com/meta-llama/llama

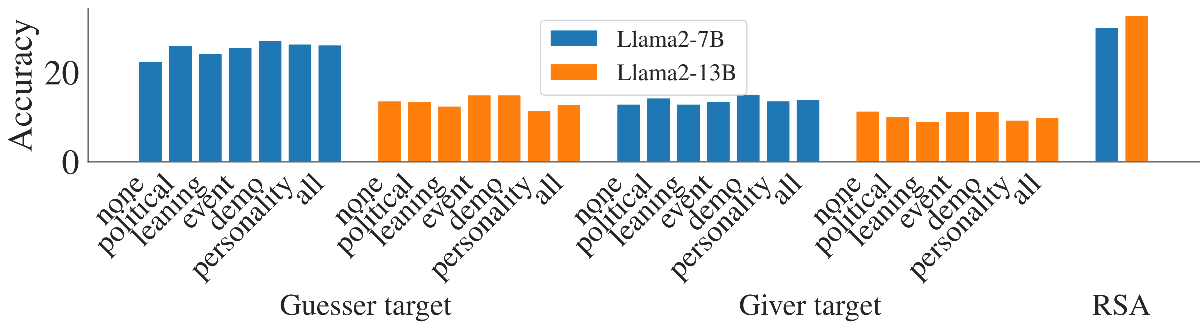


Figure 7: **Giver and guesser alignment for target selection**. RSA resulted in greater accuracy across both model sizes while model effectiveness varied across the cultural demographic that was included. Definitions of each cultural split can be found in Appendix D.2 of Shaikh et al. (2023).

```
    target words ['round', 'cylinder']
    the best clue is 'circle'. For the
    target words {target words} the best
     clue is '
```

A.5 Rational speech acts framework

In our extension of the RSA framework, we first queried the Llama2 chat models to generate a clue using the same clue generation prompt from Appendix A.1. To allow for a diverse set of potential clues, we generated 5 clues per prompt, allowing for repeat clues.

Using these clues, we then queried the model to select a target word using the following prompt:

```
You are playing Codenames and are the
    clue guesser. You need to select one
     word from {all words}. Given the
    clue {clue}, the most likely word is
```

We calculated the probability of a target word being generated from the list of possible target words as described in Appendix A.2. Following both queries, we calculated the probability of the guesser's target word generation under a given clue as the sum of the individual probabilities of the target word being generated by the LlamaGuesser and the clue being generated by the LlamaGiver. Comparing these cumulative probabilities across all target word and clue pairs allowed us to *rerank* the probability of a given utterance.

As every prompt in the Cultural Codes dataset had the human giver's intended target words (sometimes multiple), we selected the top unique target words and calculated the accuracy of our LlamaGiver and LlamaGuesser together. Here, accuracy is based on alignment with the human giver. For clue selection, we selected the corresponding clue paired with the most probable target word.

B Additional embedding training results

B.1 Target accuracy

We evaluate the performance of trained embeddings in selecting correct targets, with results shown in Figure 8. Our method for training embeddings generally does not result in improved target accuracy. In fact, since the untrained GloVe embeddings perform better than human guessers in selecting the intended targets, training on human data decreases the target accuracy in many cases.

B.2 Improvement over baselines

We include our numerical results in Tables 1, 2, & 3, showing accuracy of trained embeddings compared to that of baselines.

C RSA Extensions

In a dialogue, there is both a *speaker* and a *listener*. The goal of the *speaker* is to communicate concepts that the *listener* aims to interpret. The standard RSA framework assumes that the speaker and listener share common ground (Degen, 2023). In cross-cultural communication, this assumption is false. We propose a method for modeling the repair process (Pickering and Garrod, 2004) of two speakers aiming to find common ground.

In RSA formulations, the (abstract) *literal listener* L_0 interprets meaning based on literal semantics. The *pragmatic speaker* S_1 reasons about the literal listener and chooses utterances to optimize informativeness while minimizing the cost (e.g. length). Formally, let w represent an abstract variable referred to as *world* in Degen (2023) and m stand for the meaning that the speaker wants to convey with their utterance u. Importantly, w can be instantiated by different situations or contexts in which the interlocutors find themselves.

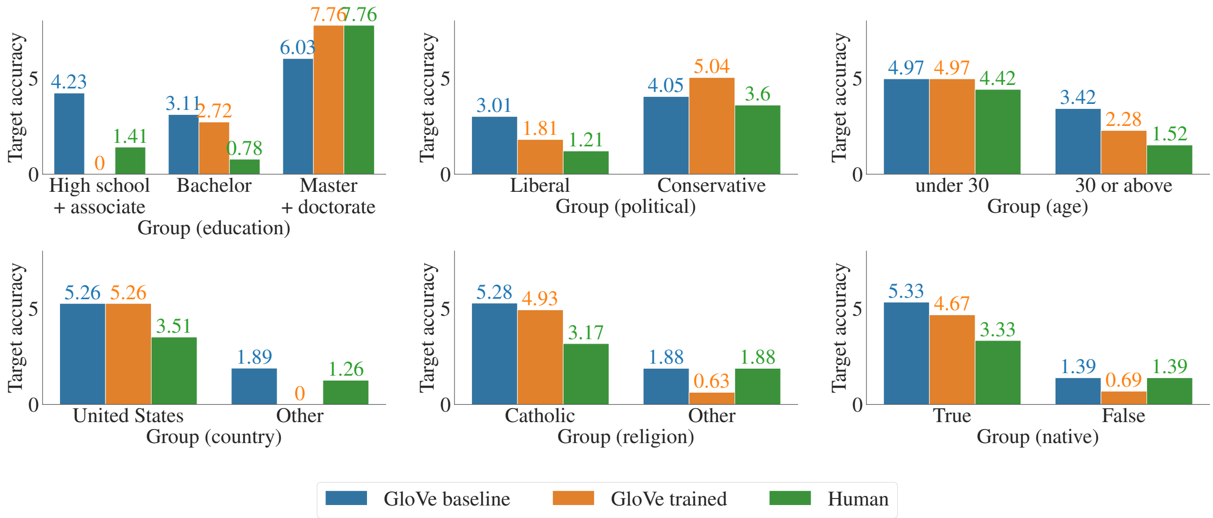


Figure 8: Comparison of target accuracy using embeddings trained on cultural splits against baseline GloVe embeddings. Target accuracy measures the performance of embeddings in correctly selecting the intended target words chosen by the clue giver. In green is the performance of the human guessers in the dataset.

Group	GloVe baseline guess acc.	GloVe trained guess acc.	% improvement
Education: high school, associate	48.86	57.95	49.13
Education: bachelor	42.21	60.55	18.6
Education: graduate	40.14	59.86	40.16
Gender: female	38.97	56.34	45.07
Gender: male	45.42	63.09	43.03
Country: united states	42.99	61.49	38.90
Country: foreign	42.39	59.24	43.45
Native: true	42.90	61.08	39.75
Native: false	42.51	56.89	42.38
Political: liberal	41.36	60.00	34.35
Political: conservative	43.81	58.86	33.83
Age: under 30	41.49	57.45	57.45
Age: over 30	43.50	59.82	59.82
Religion: catholic	43.08	60.38	60.38
Religion: not catholic	42.29	56.72	56.72
All	43.16	60.50	40.18

Table 1: Guess accuracy of trained embeddings across dataset splits.

Group	Same split guess acc.	Other split guess acc.	% improvement
Education: high school, associate	57.95	51.14	13.32
Education: bachelor	60.55	56.06	8.01
Education: graduate	59.86	50.70	18.07
Gender: female	56.34	56.81	—
Gender: male	63.09	58.50	7.85
Country: united states	61.49	56.12	9.57
Country: foreign	59.24	55.43	6.87
Native: true	61.08	58.81	3.86
Native: false	56.89	56.29	1.07
Political: liberal	60.00	54.55	9.99
Political: conservative	58.86	57.86	1.73
Age: under 30	57.45	58.51	—
Age: over 30	59.82	60.42	—
Religion: catholic	60.38	54.40	10.99
Religion: not catholic	56.72	58.21	—

Table 2: Comparison of guess accuracy when embeddings are trained on data from the same culture vs. data from different cultures.

Group	Human target acc.	GloVe baseline guess acc.	GloVe trained guess acc.	% improvement
Education: high school, associate	1.41	4.23	0.00	—
Education: bachelor	7.78	3.11	2.72	—
Education: graduate	7.76	6.03	7.76	28.6
Gender: female	1.12	4.47	2.80	—
Gender: male	3.77	3.77	3.77	0.00
Country: united states	3.51	5.26	5.26	0.00
Country: foreign	1.26	1.89	0.00	—
Native: true	3.33	5.33	4.67	—
Native: false	1.39	1.39	0.69	—
Political: liberal	1.21	3.01	1.81	—
Political: conservative	3.60	4.05	5.04	24.22
Age: under 30	4.42	4.97	4.97	0.00
Age: over 30	1.52	3.42	2.28	—
Religion: catholic	3.17	5.28	4.93	—
Religion: not catholic	1.88	1.88	0.63	—
All	2.70	4.05	3.60	—

Table 3: Target accuracy of trained embeddings across dataset splits.

The joint probability distribution of these variables, conditioned on w, factorizes as

$$P(m,u|w) = P(m|w)P_{S_1}(u|w,m), \quad (4)$$

where P_{S_1} is governed by speaker S_1. The goal of pragmatic listener L_1 is to comprehend the meaning m and infer meaning m given w and S_1's utterance u. Using Bayes's rule, this probability is proportional to

$$P_{L_1}(m|w,u) \propto P(m|w)P_{L_1}(u|w,m). \quad (5)$$

The subtle assumption made by this equation is that the probability over meanings, given world, is independent of the interlocutor, and thus L_1 reasons about it the same way the speaker does. We believe that this is *not true*. The response, and therefore a meaning to communicate, to a situation depends tightly on the speaker, and can be shaped by factors such as cultural or demographic background. Hence, in the context of cross-cultural communication, Eq. (4) should be written as

$$P(m,u|w) = P_{S_1}(m|w)P_{S_1}(u|w,m),$$

and Eq. (5) would read

$$P_{L_1}(m|w,u) \propto P_{L_1}(m|w)P_{L_1}(u|w,m).$$

In this paper, we will model two different *literal listeners* and respective *pragmatic speakers* with overlapping but not identical prior beliefs. We will model the different literal listeners and pragmatic speakers using prompting and/or training. Therefore these pragmatic speakers will have different subjective prior beliefs, reflecting the scenario of cross-cultural communication. We then seek to learn a *pragmatic listener* with incorrect or without access to the prior beliefs of the *pragmatic speaker*.

$$P_{L_1}(m,w|u) = P_{S_1}(u|m,w) \cdot P(m|w) \cdot P(w)$$

Where the variable captures whether the world is normal or wonky such that:

$$P(m|w) \propto \begin{cases} P_{usual}(m) & \text{if not } w, \\ P_{backoff}(m) & \text{if } w \end{cases}$$

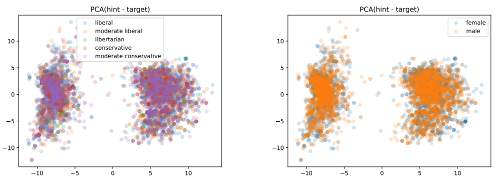
In this case, P_{usual} is the prior probability in the scenario where the world is "normal" and $P_{backoff}$ is the prior probability where the world is "wonky". This backoff probability is a uniform distribution. The value of w is inferred from the utterances u of the pragmatic speaker S_1 by the pragmatic listener L_1 based on how unlikely the utterances u are in the context of the pragmatic listener's prior beliefs. To calculate the posterior beliefs of the pragmatic listener about the meaning w

$$P_{L_1}(m|w) \propto \sum_w P_{L_1}(m,w|u)$$

The pragmatic listener's posterior probabilities are a mixture of the computation and a backoff prior based on how likely it is that w is true and the world is "wonky". In cross-cultural communication, the "wonky" world represents the case where the assumed common ground does not exist or is different in some way. In this paper, we hypothesize that RSA and the concept of wonky world can assist in understanding cross-cultural communication in the context of Codenames Duet and predict when common ground is not held between agents.

D Data analysis across clue giver attributes

We attempt to see if the obtained clusters align with existing classes of clue givers that are recorded in the data set. We consider the following labels: *nativeness* - (whether one is an English native speaker or not), *political leaning* (conservative, moderate conservatism, libertarian, moderate liberal, liberal), *race* (Asian, Black, Native American, Hispanic/Latino, White), *conscientious* (a score in range 1-4), and *gender* (male or female). Unfortunately, as we illustrate in Figure 9 for political leaning and gender, we haven't found classes that significantly align with any of the K-Mean clusters. While it is possible that we have not run these tests with classes that would display such an alignment, it is also possible that the clusters are formed by features that involve non-trivial interactions between the socio-cultural background information variables. It is also possible that this misalignment is driven by class imbalances within the dataset. For example, we found that approximately 70% of the contributors were White, leaving little room for the other races. In this case, the contribution to the total variance of the dataset coming from the minorities may be insignificant, and thus lost in PCA projections. This is further confirmed by our linear probing experiments (see Table 4); here, using the representations projected onto the first 5 PCA dimensions, we train logistic-regression (linear) classifiers and contrast them with the fraction



(a) Political leaning (b) Gender

Figure 9: **Scatter-plots of *target-hint* difference from GPT after PCA transformation with the first 2 principal components.** Here, we attempt to align with the political leaning and gender labels.

of the data occupied by the majority class. We find that the accuracies at convergence follow closely simply that of the fixed majority vote.

	GloVE_t-h	GPT_t-h	GPT_r	Majority
nativeness	**0.766**	0.759	0.762	0.765
political	0.38	**0.397**	0.387	0.386
race	0.676	**0.692**	0.667	0.685
consc.	0.353	0.336	**0.356**	0.356
gender	0.518	**0.556**	0.525	0.551

Table 4: Accuracy scores of a logistic regression (linear) classifier, averaged over 5 random seeds, together with the proportion of the data occupied by the majority of a considered class. The features were derived from GloVE *target-hint*, GPT *target-hint*, and GPT *rationale*.

E Interactive Evaluation Experiments

We run experiments with 1 target, because of higher win rates. We ran the experiments for Llama2-7B-Text for 100 games and the one for the High School guesser for 1000 games. We ran less games under Llama due to time restrictions.

To make sure that the games all occur on the same set of boards, we generate a fixed set of boards to be used for each experiment. We do this by generating a set of n board each with a unique seed and hold the seeds constant. This allows us to easily scale up a number of boards while ensuring that the boards are the same for each run and each experiment.