Modeling Cross-Cultural Pragmatic Inference with Codenames Duet

Omar Shaikh[⋆†] **Caleb Ziems**[⋆†] **William Held** [‡] **Aryan J. Pariani** [‡]
Fred Morstatter [◇] **Diyi Yang** [†]
[†]Stanford University, [‡]Georgia Institute of Technology, [◇]USC Information Sciences Institute
{oshaikh, cziems, diyiy}@stanford.edu {wheld3, apariani3}@gatech.edu fred@isi.edu

Abstract

Pragmatic reference enables efficient interpersonal communication. Prior work uses simple reference games to test models of pragmatic reasoning, often with unidentified speakers and listeners. In practice, however, speakers' sociocultural background shapes their pragmatic assumptions. For example, readers of this paper assume NLP refers to "Natural Language Processing," and *not* "Neuro-linguistic Programming." This work introduces the CULTURAL CODES dataset, which operationalizes sociocultural pragmatic inference in a simple word reference game.

CULTURAL CODES is based on the multi-turn collaborative two-player game, *Codenames Duet*. Our dataset consists of 794 games with 7,703 turns, distributed across 153 unique players. Alongside gameplay, we collect information about players' personalities, values, and demographics. Utilizing theories of communication and pragmatics, we predict each player's actions via joint modeling of their sociocultural priors and the game context. Our experiments show that accounting for background characteristics significantly improves model performance for tasks related to both clue giving and guessing, indicating that sociocultural priors play a vital role in gameplay decisions.

1 Introduction

> "*Most of our misunderstandings of other people are not due to any inability to... understand their words... [but that] we so often fail to understand a speaker's* ***intention***."
>
> — **George Armitage Miller** (1974)

Certain pragmatic inferences can only be interpreted by individuals with shared backgrounds.

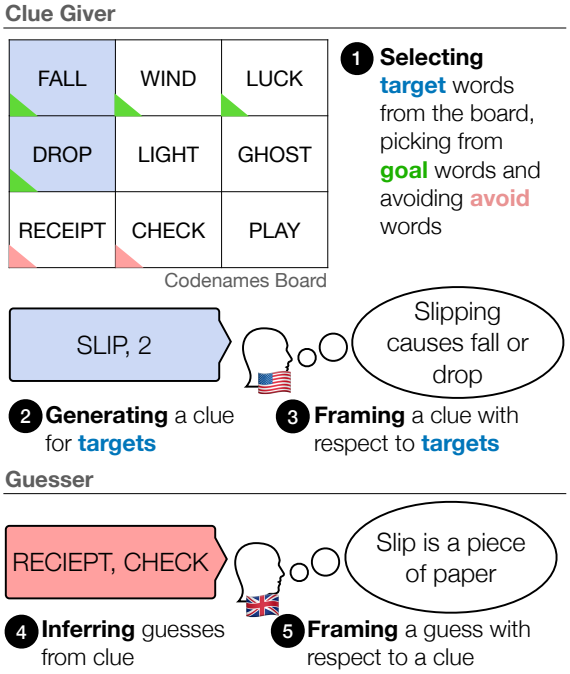

Figure 1: **An example interaction where difference in sociocultural background results in misinterpretation.** Steps 1-5 outline high-level gameplay tasks. THE CLUE GIVER targets the words *fall* and *drop*, giving the hint *slip*. THE GUESSER misinterprets *slip* as a piece of paper, guessing *reciept* and *check*.

For example, what researchers call *fun* may not be *fun* for kindergartners. Theories from sociolinguistics, pragmatics, and communication aim to explain how sociocultual background affects interpersonal interaction (Schramm, 1954)—especially since variation occurs across several dimensions: class (Bernstein, 2003; Thomas, 1983), age (Labov, 2011), gender (Eckert and McConnell-Ginet, 2013), race (Green, 2002), and more.

Rigorously modeling how culture affects pragmatic inference on *all* axes is understandably challenging. The board game *Codenames Duet* offers a more restricted setting of turn-based word reference between two players. In each round, THE CLUE GIVER provides a single-word clue;

⋆Equal contribution.

Findings of the Association for Computational Linguistics: ACL 2023, pages 6550–6569
July 9-14, 2023

then THE GUESSER must interpret this clue to select the intended word references on the game board. Ideal inferences come from the players' common ground—the set of shared beliefs between them (Clark, 1996). In practice, however, a player's behavior can be idiosyncratic. Each player has knowledge and experience that shape how they interpret clues and make guesses. When players' backgrounds differ, they may be more likely to misinterpret their partner, as seen in Figure 1.

Inspired by the above, we model the role of sociocultural factors in pragmatic inference with a new task and a series of ablation experiments. First, we describe the CULTURAL CODES dataset of cross-cultural *Codenames Duet* gameplay, with relevant background information from the players' demographics, personalities, and political and moral values (§3). Then, we deconstruct each action in a game into a distinct modeling task, taking inspiration from work on cross-cultural pragmatics (§4). Finally, we model each task with/without sociocultural priors, and highlight how player background improves model performance (§6). Our dataset and code is released publicly at `https://github.com/SALT-NLP/codenames`

2 Related Work

Cross-Cultural Pragmatics and NLP Pragmatics describes the nonliteral meaning that comes from context and social inference (Purpura, 2004; Thomas, 1983; Hatch et al., 1992). Although some pragmatic categories are universal (e.g., politeness), they can be expressed differently in sociocultural contexts (Taguchi, 2012; Shoshana et al., 1989; Gudykunst and Kim, 1984). When an intended meaning is misinterpreted, this is known as 'pragmatic failure' (Thomas, 1983)—often the result of misaligned reference frames or differences in common ground (Stadler, 2012; Crawford et al., 2017). Especially relevant to Codenames are communal lexicons, where common ground manifests in shared community vocabulary (Clark, 1998). Another axis of difference is between low/high-context cultures (Hofstede, 2001); high-context cultures rely more on shared background. Pragmatics also differs by age (Saryazdi et al., 2022), region, ethnicity, politics, and class (Thomas, 1983), as does theory of mind (Fiske and Cox, 1979; Miller, 1984; Shweder, 1984; Lillard, 1998, 1999).

Outside of work on politeness (Sperlich et al., 2016; Fu et al., 2020), sarcasm (Joshi et al., 2016), and irony (Karoui et al., 2017), the NLP subfield has not closely considered cross-cultural pragmatics. While there is work on understanding the role of individual culture—for example, learning demographic word vectors (Garimella et al., 2017), identifying deception/depression (Soldner et al., 2019; Loveys et al., 2018), or improving translation (Specia et al., 2016)—modeling **cross**-cultural pragmatic inference in communication remains a challenge (Hershcovich et al., 2022).

Still, a culture-free pragmatics has played a central role in various NLP tasks, from instruction-following (Fried et al., 2018), image captioning (Andreas and Klein, 2016), persona-consistent dialogue (Kim et al., 2020), and summarization (Shen et al., 2019). Much of this work is grounded in Bayesian models of cognition (Griffiths et al., 2008), with models like *Bayesian Teaching* (Eaves Jr et al., 2016), *Naive Utility Calculus* (Jara-Ettinger et al., 2016; Jern et al., 2017), and the *Rational Speech Acts* (RSA) model (Goodman and Frank, 2016; Franke and Jäger, 2016) that integrate language, world knowledge, and context to explain ideal pragmatic reasoning (Noveck, 2018) and grounded reference (Monroe et al., 2017). Instead of modeling socioculture in isolation, we model pragmatic inference, highlighting the role of culture in general interpersonal interaction.

Games as Testbeds for AI A significant body of work focuses on modeling optimal *strategy* across a wide set of games, including Go (Silver et al., 2016), Chess (Schrittwieser et al., 2020), Poker (Brown and Sandholm, 2017), Diplomacy (, FAIR), D&D (Callison-Burch et al., 2022; Zhou et al., 2022), and Mafia (Ibraheem et al., 2022). Reference games are growing in popularity as testbeds for AI. Tests for artificial pragmatic reasoning often rely on sequential language games, where two players leverage private knowledge either to compete Yao et al. (2021) or coordinate towards a common goal (Potts, 2012; Khani et al., 2018; Hawkins et al., 2015). In this vein, recent works have considered *Codenames* (Koyyalagunta et al., 2021; Kim et al., 2019; Jaramillo et al., 2020), *Connector* (Ashok Kumar et al., 2021; Kumar et al., 2021; Kovacs et al., 2022) *InfoJigsaw* (Khani et al., 2018), and image-based games (Bao et al., 2022). Word association games have been used in psychology to study semantic associations in cultural (Korshuk, 2007) and religious (Tikhonova, 2014) contexts. We utilize games to model the effect of

cross-cultural interactions on pragmatic inference.

3 The CULTURAL CODES Dataset

This study has been approved by the Institutional Review Board (IRB) at the authors' institution. The purpose of the CULTURAL CODES dataset is to understand how measurable social factors influence dyadic communication *in English*. By collecting relevant participant background information, we aim to understand how these factors affect linguistic reasoning in a collaborative reference game.

3.1 *Codenames Duet* Game Overview

Codenames Duet is a collaborative variant of *Codenames* (Vlaada, 2015) designed for 2 players. The players share a 5×5 board of 25 common words. Each player has a distinct (but sometimes partially overlapping) map from words on the board to the following objectives: **goal**, **neutral**, and **avoid**. One player's map is hidden from the opposing player. The objective of the game is for both players to guess all of their partner's **goal** words without guessing any of their partner's **avoid** words, as doing so results in an immediate loss.

CULTURAL CODES uses an adapted version of *Codenames Duet*. With each turn, players alternate between the THE CLUE GIVER and THE GUESSER roles. To begin the turn, THE CLUE GIVER **(1)** selects one or more associated **goal** words as targets. Next, THE CLUE GIVER **(2)** provides a single word clue that relates to the associated target(s). This clue is displayed to THE GUESSER, along with the number of targets she should find. The THE CLUE GIVER also **(3)** provides a justifying *rationale* for the clue, describing the relationship between the clue and the target(s). This *rationale* is not displayed to the partner. Using the clue and the number of target words THE GUESSER **(4)** guesses targeted words. For each guess, THE GUESSER **(5)** provides a justifying *rationale* for the guess. After ending the turn, players alternate roles and continue until all **goal** words are selected for both sides, or players are eliminated for guessing an **avoid** word. An overview of roles is illustrated in Figure 1. In §4, we formalize actions **(1)-(4)** as distinct modeling tasks.

3.2 Selecting Board Game Words

All experiments are run on a strategically filtered subset of the 400 words from *Codenames Duet*. We select the 100 most abstract and semantically ambiguous board game words to elicit diverse responses from players. Since the *polysemy* (Ravin and Leacock, 2000) of a word—the number of related senses it includes—predicts the expected diversity of player responses, we retain only nouns with two or more senses in WordNet (Miller, 1992). Next, we rank polysemous words with Brysbaert et al. (2014)'s concreteness list, selecting the **100 most abstract words** according to the mean of their human concreteness scores (finalized list can be found in Appendix A.)

When a player starts a game, we initialize the board with a random subset of 25 words from the filtered 100. For each player, 9 words are randomly mapped to **goal**, 3 are **avoid**, and 13 are **neutral**.

3.3 Gameplay Data

To collect gameplay data, we modified an open-source implementation of *Codenames Duet*,[1] automatically pairing individuals who visited the game website. To source players, we relied on Amazon's Mechanical Turk. We provided MTurkers with an initial instruction video detailing rules and how to play. To be eligible for the task, Turkers had to get $\geq 80\%$ questions right on a qualifying quiz about Codenames rules and gameplay (Appendix D.1). Average game length was around 17.4 minutes, and MTurkers were paid $2.50 for every game.

Gameplay Attributes For each completed turn, we collected the following game state information from THE CLUE GIVER. Elements marked in gray were hidden from THE GUESSER.

> **Clue:** THE CLUE GIVER's clue c (e.g. c could be "*transport*" for the target "*car*").
>
> **Target Word(s):** (Hidden) The target words t_n (e.g. "*car*") that THE CLUE GIVER intended THE GUESSER to guess.
>
> **Target Word(s) Rationale(s):** (Hidden) A free-text phrase r_n, that describes the relationship between each target word t_n and the clue c (e.g. "*a car is a mode of transport*").

To summarize, each turn from THE CLUE GIVER results in a clue c and at least one target-rationale pair (t_n, r_n). On the other hand, we collect the following for **THE GUESSER**.

> **Guesses**: The guesses g_n that THE GUESSER selected for THE CLUE GIVER's clue c.

[1] https://github.com/jbowens/codenamesgreen

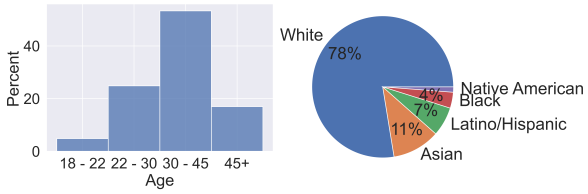

Figure 2: **Age (left) and Race (right) across our annotators.** Most of our annotators are between 30-45 and are White; however, we still see moderate representation across other racial groups and ages.

> **Rationale for Each Guess**: A free-text phrase r_n that relates the guess g_n to the clue c

Manual inspection revealed a wide range of rationales. To prevent models from exploiting variance, we instructed GPT-3 to normalize text, removing pronouns and determiners.[2] We provided few-shot examples of reformatted rationales and manually inspected normalized outputs. Additional preprocessing information can be found in Appendix B.

3.4 Sociocultural Priors and Worker Diversity

Because we aim to understand the role of sociocultural priors on gameplay, we asked Turkers to complete the standardized surveys below, which cover three broad dimensions: *demography, personality, and morality*.

Demographic Data (Figure 2) comes from both the annotation UI and in the task's qualifying questionnaires. In the UI, we asked Turkers for their numeric age, their country of origin, and whether English is their native language. These were required features, so we will denote them as $\textbf{Demo}_{\textbf{Req}}$. In the qualifier, we included an extended demographic survey with *age range, level of education, marital status*, and *native language* (Appendix D.2.1), which we will denote as $\textbf{Demo}_{\textbf{All}}$. We find that our annotator demographics are moderately diverse, mirroring Moss et al. (2020). Reported gender across annotators are evenly split: 53% identify as women, 47% identify as men, and 0% as other. Additional details are in Figure 2 and Appendix D.2.1.

Personality (Figure 3) surveys also offer insight into interpersonal interactions. We administer the Big 5 Personality Test (John et al., 1991), measuring a range of personality dimensions on a 5 point Likert Scale. Features include Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. Definitions are in Appendix D.2.2.

[2]We use the text-davinci-003 variant from OpenAI. Without GPT-3 normalization, we find that model performance is artificially inflated.

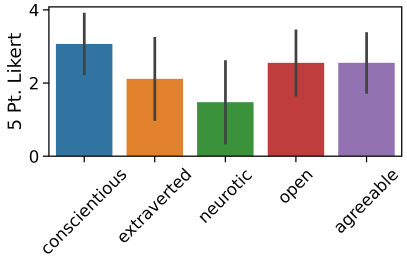

Figure 3: **Big 5 Personality (John et al., 1991) results across annotators.** Each personality dimension has a standard deviation ≈ 1, indicating a reasonable diversity across our annotator pool.

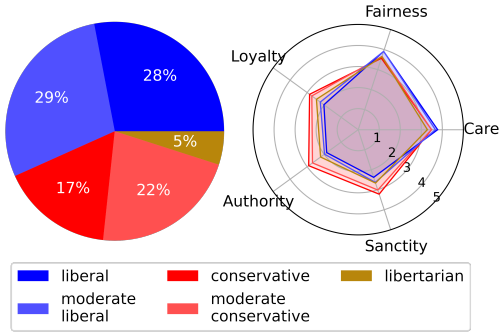

Figure 4: **Self-reported political leaning (left) and Haidt and Graham (2007)'s Moral Foundations Theory (right) across annotators.** A majority of our workers are liberal (57%), 39% are conservative, and the remaining 5% are libertarian. As observed in Haidt (2012), values like loyalty, authority, and sanctity are higher for conservative leaning annotators, while fairness is higher for liberal annotators ($p < 0.05$, t-test)

Moral and Political Leaning (Figure 4) also influences decision making processes. Therefore, we asked annotators to self-report their political leaning (liberal, conservative, libertarian, etc). While political leaning captures broad elements of annotator values, Haidt and Graham (2007)'s widely adopted Moral Foundations Theory (MFT) deconstructs values into individual foundations (Care/Harm, Fairness/Cheating, Loyalty/Betrayal, Authority/Subversion, and Sanctity/Degradation). Differences in each foundation can stem from cultural variation (Haidt, 2012). To record annotator leaning on MFT, we administer an abridged version of the Moral Foundations Questionnaire (Graham et al., 2008), which reports each dimension on a 5 point Likert scale (see Appendix D.2.3). Later, we refer to all recorded features as **Morality**.

Agent	Task Description	Input	Output	N
CLUE GIVER	***(1) Target Words*** Generate, from the **goal** p_i words, a subset of targets t_i. Targets are used to generate a single clue word.	{**goal**} $= \{\texttt{BOS}, p_1, p_2, ..., p_n, \texttt{EOS}\}$	{targets} $= \{\texttt{BOS}, t_1, t_2, ..., t_m, \texttt{EOS}\}$	7,961
	(2) Generating a Clue Generate a one word clue c_1 that relates selected target words while avoiding **avoid** a_i and neutral n_i words.	{**avoid**, neutral, targets} $= \{\texttt{BOS}, \texttt{AVO}, a_1, a_2, \ldots, a_o, \texttt{NEU}, n_1, ..., n_n \texttt{TGT}, t_1, t_2, \ldots, t_m, \texttt{EOS}\}$	{clue} $= \{\texttt{BOS}, c_i, \texttt{EOS}\}$	7,703
	(3) Framing a Clue Generate reasoning r that frames a candidate clue word c_i w.r.t. a target t_i word from the set of targets.	{targets, clue, target} $= \{\texttt{BOS}, \texttt{TGTS}, t_1, ..., t_n, \texttt{CLUE}, c_i, \texttt{TGT}, t_i, \texttt{EOS}\}$	{rationale} $= \{\texttt{BOS}, r, \texttt{EOS}\}$	9,519
GUESSER	***(4) Selecting Guess Words*** Generate a series of guesses $\{g_1, ..., g_m\}$ from the unselected words given a clue c_i.	{unselected, clue} $= \{\texttt{BOS}, \texttt{UN}, u_1, ..., u_n, \texttt{CLUE}, c_i, \texttt{EOS}\}$	{guesses} $= \{\texttt{BOS}, g_1, g_2, ..., g_m, \texttt{EOS}\}$	7,703
	(5) Framing Guesses Generate reasoning r that frames a guess g_i (from all guesses) w.r.t. clue c_i	{guesses, clue, guess} $= \{\texttt{BOS}, \texttt{GUESSES}, g_1, ..., g_n, \texttt{CLUE}, c_i, \texttt{GUESS}, g_i, \texttt{EOS}\}$	{rationale} $= \{\texttt{BOS}, r, \texttt{EOS}\}$	9,382
BOTH	***Predict Correct Guess*** Classify if CLUE GIVER message (using target, rationale, and clue) is correctly interpreted by the GUESSER	{unselected, target, rationale, clue} $= \{\texttt{BOS}, \texttt{UN}, g_1, ..., g_n, \texttt{TR}, t_i, r_i, \texttt{CLUE}, c_i, \texttt{EOS}\}$	$\{T, F\}$	9,519

Table 1: **Tasks associated with a turn in Codenames.** THE CLUE GIVER starts by selecting information to encode (in the form of a clue), and THE GUESSER decodes clues through guesses. In our experiments, we evaluate models with and without sociocultural priors. Task formulation (generation/classification) is underlined.

3.5 General Dataset Statistics

In total, we collect 794 games, with a total of 199 wins and 595 losses.[3] Games lasted an average of 9.7 turns, resulting in 7,703 total turns across all games. THE CLUE GIVER targeted an average of 1.24 words per turn. For all collected games, both players provided $\text{Demo}_{\textbf{Req}}$. For 54% of games, both players completed all background surveys; for the remaining 46% of games, at least one player completed all surveys. There were no games with *no* background information.

4 Tasks and Modeling

To investigate the role of sociocultural factors in pragmatic inference, we propose a set of tasks (Table 1) associated with THE CLUE GIVER (§4.1) and THE GUESSER (§4.2) roles. Concretely, we formalize each action into a conditional generation problem instead of classification, since outputs in CULTURAL CODES are unconstrained: actions and outputs depend on a changing board state.

4.1 Modeling THE CLUE GIVER

4.1.1 Selecting Target Words

To start, THE CLUE GIVER identifies target word(s) **(1)** on a board, which are later used to construct a target clue for the inference. Clues will target salient words, where salience is at least partially determined by the speaker's cultural background (Wolff and Holmes, 2011). Each set of targets is a subset of the remaining **goal** words for a given turn (targets $\subseteq$ **goal**)—we enforce this restriction in our annotation UI.

4.1.2 Giving a Clue

After selecting target words, THE CLUE GIVER must generate a common clue word across the targets **(2)**. Here, THE CLUE GIVER must select a prototypical word across the targets. Because cultural background plays a role in inference (Thomas, 1983), a clue should lie in players' common ground. Furthermore, the clue word should not lead the

[3] Some players went inactive before a game was completed. We only collect games that are reasonably long: greater than the 90^{th} percentile of incomplete games, or ≥ 7 turns.

Priors	Model	Target R-1	Guess R-1
No Priors	Random	0.60	0.65
	k-NN fastText	N/A	58.04
	T5	32.57	64.96
	BART	31.82	63.30
↓ ***With* Sociocultural Priors**			
$\text{Demo}_{\textbf{Req}}$	T5	32.71	67.25
	BART	29.45	65.18
$\text{Demo}_{\textbf{All}}$	T5	33.14	65.24
	BART	32.27	66.02
Personality	T5	33.61	65.56
	BART	28.55	63.14
Morality	T5	34.58	64.60
	BART	31.32	65.09
All	T5	33.38	66.31
	BART	30.17	64.78

Table 2: **Target (§4.1.1) & Guess (§4.2.1) Selection Generation Results**. We report only R-1 scores, since tasks must contain exact single-word matches to reference labels. Target Selection is maximized when using Morality priors, while Guess Selection is maximized by using only $\text{Demo}_{\textbf{Req}}$.

Priors	Model	Clue R-1	fastText *cos*
No Priors	Random	0.08	5.76
	k-1 fastText	0.00	10.33
	T5	23.86	40.38
	BART	23.00	40.97
↓ ***With* Sociocultural Priors**			
$\text{Demo}_{\textbf{Req}}$	T5	25.47	42.91
	BART	20.64	38.91
$\text{Demo}_{\textbf{All}}$	T5	25.74	42.07
	BART	21.45	39.45
Personality	T5	24.13	41.00
	BART	23.32	41.49
Morality	T5	26.54	43.31
	BART	23.59	41.39
All	T5	26.27	44.03
	BART	24.40	41.60

Table 3: **Clue Generation Results (§4.1.2)** We report R-1 scores and fastText *cos* similarities between the reference and generation, since outputs must be semantically close to or exactly match the reference labels. We find that **Morality** and **All** maximize performance over our metrics.

guesser to pick a **avoid** n_i or **neutral** e_i word, since these words can end the game or turn (see §3.1). Therefore, we also include **avoid** and remaining **neutral** words in our input.

4.1.3 Framing the Target Rationales

The relationship between the target and clue word plays a critical role in communication—how information is *framed* with respect to common ground can influence pragmatic success (Crawford et al., 2017). To this end, we model THE CLUE GIVER's framing of the rationale r for a specific target word t **(3)**, connecting the target t to the clue (c.f., §3.3). Because the framing is constructed in relation to every target word (if multiple are provided), we also encode all targets in the input.

4.2 Modeling THE GUESSER

4.2.1 Selected Guesses

With the clue word, the THE GUESSER pragmatically infers THE CLUE GIVER's targets, selecting a sequence of corresponding guesses **(4)**. For this task, we model the sequence of all selected guesses, regardless of correctness. We input all *unselected*[4] words at the start of each turn for THE GUESSER, along with the provided clue. Like with Target Word Selection, guesses must be a subset of the unselected words ($\text{guesses} \subseteq \text{unselected}$); we enforce this during annotation.

4.2.2 Framing Guess Choice

Finally, THE GUESSER also provides framing rationale for their respective guesses, framing clues with respect to their guess **(5)**.

4.3 Predicting Pragmatic Success

So far, our tasks focus on *replicating* elements of a game turn: the Selected Guesses task (§4.2.1), for example, models both incorrect and correct guesses. However, we also wish to understand if an entire turn sequence results in a **successful** inference; differences in cross-cultural inferences can result in pragmatic failures (Thomas, 1983). We formulate this as binary classification.

Importantly, we only consider a guess correct if it is *intentional*. A guess is intentional *if and only if* the clue giver listed it as a target. If THE GUESSER selects a **goal** word that is *not* a target word, we count it as "incorrect." Like with guess generation, we encode unselected words in the input. Because we are not predicting the guess itself, we include

[4]Note that **goal**/**avoid**/**neutral** words differ across players. A **goal** word for one player can be **avoid** for another; game states are asymmetric. A clue from THE CLUE GIVER may also target a **goal** word for the THE GUESSER. As long as one does not guess a **avoid** word from the *opposing* player, the game continues. See §3.1.

Priors	Model	Target Framing					Guess Framing				
		R-1	R-2	R-L	BLEU	BScore	R-1	R-2	R-L	BLEU	BScore
No Priors	Random	14.08	3.80	13.88	3.46	86.88	8.31	1.01	8.07	0.80	85.88
	SBERT	53.14	23.10	49.13	20.04	92.24	40.49	10.82	33.57	10.53	89.31
	T5	69.22	36.82	64.13	34.11	94.52	54.67	19.65	47.22	17.40	91.25
	BART	66.20	31.85	59.84	30.09	93.72	52.36	17.27	44.49	14.72	90.85
↓ ***With* Sociocultural Priors**											
$\text{Demo}_{\textbf{Req}}$	T5	70.15	37.86	64.81	35.05	94.61	57.26	23.19	48.32	23.31	91.63
	BART	67.16	34.52	60.97	31.47	94.00	54.55	19.11	45.69	17.62	90.95
$\text{Demo}_{\textbf{All}}$	T5	70.40	38.14	64.98	35.07	94.60	57.22	23.14	48.36	21.05	91.59
	BART	66.14	32.21	59.72	31.36	93.88	52.43	16.51	43.52	13.23	90.78
Personality	T5	69.68	38.31	64.74	35.27	94.47	57.41	23.08	48.72	21.37	91.61
	BART	67.12	34.36	61.34	32.10	93.88	52.89	18.85	45.07	15.55	90.92
Morality	T5	69.82	37.96	64.35	34.53	94.63	58.06	23.67	48.85	22.62	91.76
	BART	67.78	34.47	61.49	32.25	94.25	53.46	18.49	45.73	14.95	90.93
All	T5	70.39	38.27	65.49	34.01	94.66	57.64	23.13	48.79	22.22	91.68
	BART	67.66	34.45	62.28	31.59	93.95	52.12	18.13	44.51	15.96	90.92

Table 4: **Framing Generation Results** for Target (§4.1.3) and Guess (§4.2.2) words. We find that the best models with sociocultural priors **universally** outperform their baseline counterparts. For Target Rationale Generation, jointly modeling all features yields highest improvements; Guess Rationale generation sees improvements when using Morality priors. Guess Rationale Performance sees higher relative/absolute improvement from baselines compared to Target Rationale Generation.

Priors	Random	BERT	RoBERTa	XLNet
None	0.50	0.57	0.57	0.57
↓ ***With* Sociocultural Priors**				
$\text{Demo}_{\textbf{Req}}$	–	0.52	0.55	0.52
$\text{Demo}_{\textbf{All}}$	–	**0.59**	0.63	0.62
Personality	–	0.57	**0.67**	**0.64**
Morality	–	0.57	0.64	0.61
All	–	0.57	0.65	0.63

Table 5: Macro F-1 scores for **Predicting Pragmatic Success (§4.3)**: models must predict if a guesser will guess correctly given the target word, target rationale, and clue. We use base variants of all models and experiment with ablations across different background characteristics.

target and rationale from THE CLUE GIVER.

4.4 Augmenting with Sociocultural Priors

We hypothesize that players' backgrounds influence Codenames gameplay. To this end, we encode background player information for each task. For each dimension described in §3.4, we encode an attribute/answer pair (e.g. `age: 22`) for each survey question. Then, we prepend all attributes to the encoded strings for each outlined task (§4), using a unique token to delimit attributes for THE CLUE GIVER and THE GUESSER.

$$\boldsymbol{in_{socio}} = \{\texttt{BOS}, \texttt{GIVER}, \text{Clue Giver}_{\text{Attr:A}}, \texttt{GUESSER}, \text{Guesser}_{\text{Attr:A}}\} + \boldsymbol{in}$$

If a player did not respond to a specific attribute, we replace the attribute/answer pair with `None`. From our sociocultural priors (§3.4), we have 5 ablations: $\text{Demo}_{\textbf{Req}}$, $\text{Demo}_{\textbf{All}}$, Personality, Morality, and All (concatenating and modeling all ablations). We additionally use *no* priors as a baseline, using $\boldsymbol{in}$ instead of $\boldsymbol{in_{socio}}$ to test our hypothesis.

5 Experiment Setup

Baselines and Dataset Splits For generation baselines, we use two Seq2Seq models: T5 (Raffel et al., 2020) and BART (Lewis et al., 2020). We optimize the associated language modeling objective across our tasks. Additionally, we experiment with two retrieval baselines for all generation tasks: (1) randomly selecting a generation from the train set and (2) selecting the nearest k-N inputs using pretrained SentenceBERT (Reimers and Gurevych, 2020) or fastText (Bojanowski et al., 2017). Retrieval baselines yield insight into how well off-the-shelf pretrained models capture sociocultural diversity. For classification, we experiment with BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), and XLNet (Yang et al., 2019). Models are base variants, and results are averaged over 5 runs.

For each task, we split *clue givers* into 80-10-10 train/val/test, since all tasks depend on initial clue giver choices. Importantly, **a single clue giver's**

data is not distributed across splits, since clue givers may reuse clues/strategies.

Evaluation Metrics We use a range of metrics to generation tasks. Rationale generation tasks (Target §4.1.3 & Guess §4.2.2) output entire sentences; therefore, we report F-1 scores from ROUGE-(1, 2, L) (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2020). For tasks that generate a single or set of words where order does not matter, (Guess Selection §4.2.1; Clue Generation §4.1.2), we report only ROUGE-1 and averaged word vector (fastText) cosine similarity.

6 Generation Results & Discussion

Including cultural priors improves modeling performance across **all** tasks. For generation problems, T5 generally outperforms BART, and our retrieval baselines lag behind more complex models. Finally, we conduct 🔎 *a qualitative analysis* of 20 random samples from each task.

Picking Targets and Guesses From our results (Table 2), we find that selecting guesses is an easier modeling task than picking target words, likely because the input for selecting a guess contains the clue word. Intuitively, selecting target words is more arbitrary than selecting a guess from a clue—especially since our generation task does not enforce guess correctness. Our models reflect this observation. Guess Selection has R-1 scores that are, on average, twice as good as Target Word Selection (Target 34 vs. Guess 66). Furthermore, Guess Selection only requires demographics ($\text{Demo}_{\textbf{Req}}$) to maximize performance, unlike **Morality** for Target Words. Regardless, both tasks see R-1 increase by ≈ 2 points over no prior baselines.

🔎 Looking at model outputs between the **None** and **Morality**, we observe that models generate words like *Well*/*Grace* instead of *Death*/*Poison* and vice versa, depending on player background.

Generating a Clue for Targets Moving to our clue generation models, we again find that including sociocultural priors improves model performance (Table 3). Highest R-1 scores (26.54) occur when using **Morality** as a prior, resulting in a ≈ 2 pt. R-1 and 4 pt. cos-similarity increase when compared to a no prior baseline. We also suspect that selecting target words and generating a hint are interrelated processes: annotators are likely thinking about clues/targets in parallel. Therefore, the same Morality prior results in maximized performance.

🔎 While there are themes related to Morality in clue differences for a target word (accident → death vs. lucifer; or fair → equal vs. good), we also find that generations are *more specific* given sociocultural priors. Consider these generated target → clue pairs ✓ with and ✗ without priors:

- match → ✗ game ✓ cricket
- bond → ✗ connection ✓ james
- undertaker → ✗ funeral ✓ wrestler

Each ✓ example generates a clue that relies on shared cultural background: specifically, knowing that cricket is a sport; that James Bond is a popular character; and that the Undertaker is a wrestler. More details can be found in Appendix C, Table 6.

Clue Generation Errors Across Sociocultural Subtypes Despite jointly modeling cross-cultural information, our performance is far from perfect. Generating successful clues is a core element of Codenames; however, our exact match accuracy on clue generation is only $\approx 26\%$. To understand errors, we sample 100 generated clues from the Clue Generation Task, and identify errors and differences between (socioculturally) generated clues and the ground truth label.

For 43 samples, we notice that sociocultural priors have *no effect* on clue generation; the output is identical to the *no prior* model for the given target word. In these instances, we suspect that our models fail to exploit common ground between a giver/guesser, yielding the same clue as without sociocultural priors. Upon further analysis, we observe that these errors occur frequently (37 samples) when *both* the clue giver and guesser are white or from North America. Because these demographics are already over-represented in our dataset, we suspect that the model simply ignores over-informative sociocultural priors.

Errors also occur because clues are over (20 instances, e.g. "guevera" instead of "overthrow") or underspecified (13 instances, e.g. "supernatural" instead of "monster") compared to the gold clue. In 21/33 of these instances, there is a demographic mismatch between the clue-giver and guesser: the clue-giver and guesser do not share race/country demographics. In contrast to having no effect, we suspect that models mispredict the common ground between guesser/giver. We also judge 18 generation errors to be of similar specificity to the target word—prefixes/suffixes of the gold label—or completely unrelated to the gold clue (6 instances).

Rationalizing Targets and Guesses Beyond generating target words and guesses, we ask models to explain how a target or guess is related to a clue word (e.g. James Bond is a movie character). Again, we find that providing contextual priors improves performance (Table 4). For Target Rationale Generation, models see maximized performance when **all** priors are included, while Guess Rationale generation sees improvements for **Morality.**

🔎 Like with Clue Generation, we find that improvements in Guess Rationale are from increased specificity (e.g. "actors are cast" → "actors are part of a cast"; "money is center" → "money is the center of everything"). While qualitative differences are clear for Guess Rationale, Target Rationale results are more subtle: improvements stem from minor variations in the type of framing ("a kind of" vs. "a type of") used by the annotator. Additional generations can be found in Appendix C, Table 7.

Classifying Pragmatic Failure We find that classification performance across each architecture is maximized when using sociocultural priors during training (Table 5). While BERT sees reduced improvement (an increase of only +0.02 F-1 over a no-prior baseline), XLNet and RoBERTa see maximum increases of +0.07 and +0.10 respectively. Both XLNet and RoBERTa see these improvements across the same **Personality** setting. Sociocultural priors improve performance across mirroring *and* evaluating pragmatic inference.

A Word on Word Vector Baselines Surprisingly, retrieving nearest words using a word vector approach (fastText) performs poorly for both Clue and Guess Generation (Tables 2 & 3). We suspect that pretrained vectors fail to capture sociocultural inference in word association tasks.

7 Conclusion

Language is grounded in rich sociocultural context. To underscore this context, we propose a setting that captures the diversity of pragmatic inference *across* sociocultural backgrounds. With our Codenames Duet dataset (7K turns across 156 players), we operationalize cross-cultural pragmatic inference. Across our experiments, we detail improvements in mirroring/evaluating inferences when using sociocultural priors. Our work highlights how integrating these priors can align models toward more socially relevant behavior.

8 Limitations

Cross-Cultural Inference Beyond Codenames Our work explores sociocultural pragmatic inference in a very limited setting, using a core vocabulary of just 100 words. Despite this limitation, we find significant diversity in our dataset; furthermore, our models successfully capture these diverse inferences. While a limitation of our work is its focus on a single setting, we expect domains outside of Codenames to see similar variance. Understanding and highlighting miscommunication in dialog—due to culture-dependent misinterpretation—is one such extension. These domains are likely much nosier than Codenames; we urge future work to further investigate them.

Spurious Correlations across Sociocultural Factors Across all tasks but one (Target Rationale Generation §4.1.3), jointly modeling all sociocultural priors does not result in the highest performing model. Because our sociocultural factors already correlate with each other (§3.4), we suspect that modeling all features may be redundant, adding spurious correlations and resulting in overfitting. Improved modeling methodology and careful regularization may address these issues; we leave these experiments for future work.

Bigger Models and Task Specific Modeling Currently, we evaluate small Seq2Seq models due to computational constraints; however, evaluation of 0-shot and few-shot performance on larger language models (e.g. GPT-3) is necessary. Given the changing state of the Codenames board—along with evidence that LLMs struggle with theory-of-mind-esque perspective taking (Sap et al., 2022)—our dataset can serve as a challenging benchmark for sociocultural understanding. However, successfully encoding game state into prompts for LLMs may require experimentation.

Finally, our current task formulation and modeling setup are straightforward: we simply encode all information *in-context* and do not assume recursive reasoning like in RSA (Goodman and Frank, 2016). Future work can explore these directions.

Human Evaluations Our evaluation is limited to automatic metrics and qualitative analysis. Evaluating cross cultural generation *depends* on the evaluator's own culture. Each generation depends on the player's sociocultural background; finding evaluators who match the player may be prohibitive.

9 Ethics

Broadly, our work models user background to determine the choices they make. While we focus on a fairly harmless setting (Codenames), our operationalization can be used in harmful ways (e.g. tracking and modeling user behavior without consent). Future work that uses sociocultural information should only be applied to settings where there is no foreseeable harm to end-users.

Furthermore, learning sociocultural associations can introduce positive and negative stereotypes; documenting and reducing harmful stereotypes is an important avenue for future work. Finally, we emphasize that our work is not evidence for *linguistic determinism*: sociocultural variation in language can influence but not **determine** thought.

Acknowledgements

We are thankful to the members of SALT Lab for their helpful feedback on the draft. We are also thankful for the helpful feedback from Jing Huang and Rishi Bommasani. Caleb Ziems is supported by the NSF Graduate Research Fellowship under Grant No. DGE-2039655. This research was supported, in part, by MURI-ONR-N00014-20-S-F003 on Persuasion, Identity, and Morality in Social-Cyber Environments, as well as a DARPA grant HR00112290103/HR0011260656.

References

Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1173–1182, Austin, Texas. Association for Computational Linguistics.

Abhilasha Ashok Kumar, Ketika Garg, and Robert Hawkins. 2021. Contextual flexibility guides communication in a cooperative language game. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 43.

Yuwei Bao, Sayan Ghosh, and Joyce Chai. 2022. Learning to mediate disparities towards pragmatic communication. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2829–2842.

Basil Bernstein. 2003. *Class, codes and control: Applied studies towards a sociology of language*, volume 2. Psychology Press.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Noam Brown and Tuomas Sandholm. 2017. Libratus: Beating top humans in no-limit poker. In *Neural Information Processing Systems*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior research methods*, 46(3):904–911.

Chris Callison-Burch, Gaurav Singh Tomar, Lara J Martin, Daphne Ippolito, Suma Bailis, and David Reitter. 2022. Dungeons and dragons as a dialog challenge for artificial intelligence. *ArXiv preprint*, abs/2210.07109.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H Clark. 1998. 4 communal lexicons.

Tonia Crawford, Sally Candlin, and Peter Roger. 2017. New perspectives on understanding cultural diversity in nurse–patient communication. *Collegian*, 24(1):63–69.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Baxter S Eaves Jr, Naomi H Feldman, Thomas L Griffiths, and Patrick Shafto. 2016. Infant-directed speech is consistent with teaching. *Psychological review*, 123(6):758.

Penelope Eckert and Sally McConnell-Ginet. 2013. *Language and gender*. Cambridge University Press.

Meta Fundamental AI Research Diplomacy Team (FAIR)†, Anton Bakhtin, Noam Brown, Emily Dinan, Gabriele Farina, Colin Flaherty, Daniel Fried, Andrew Goff, Jonathan Gray, Hengyuan Hu, et al. 2022. Human-level play in the game of diplomacy by combining language models with strategic reasoning. *Science*, 378(6624):1067–1074.

Susan T Fiske and Martha G Cox. 1979. Person concepts: The effect of target familiarity and descriptive purpose on the process of describing others 1. *Journal of Personality*, 47(1):136–161.

Michael Franke and Gerhard Jäger. 2016. Probabilistic pragmatics, or why bayes' rule is probably important for pragmatics. *Zeitschrift für sprachwissenschaft*, 35(1):3–44.

Daniel Fried, Jacob Andreas, and Dan Klein. 2018. Unified pragmatic models for generating and following instructions. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1951–1963, New Orleans, Louisiana. Association for Computational Linguistics.

Liye Fu, Susan Fussell, and Cristian Danescu-Niculescu-Mizil. 2020. Facilitating the communication of politeness through fine-grained paraphrasing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5127–5140, Online. Association for Computational Linguistics.

Aparna Garimella, Carmen Banea, and Rada Mihalcea. 2017. Demographic-aware word associations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2285–2295, Copenhagen, Denmark. Association for Computational Linguistics.

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

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Koleva Spassena, and Peter H Ditto. 2008. Moral foundations questionnaire. *Journal of Personality and Social Psychology*.

Lisa J Green. 2002. *African American English: a linguistic introduction*. Cambridge University Press.

Thomas L Griffiths, Charles Kemp, and Joshua B Tenenbaum. 2008. Bayesian models of cognition.

William B Gudykunst and Young Yun Kim. 1984. *Communicating with strangers: An approach to intercultural communication*. Addison Wesley Publishing Company.

Jonathan Haidt. 2012. *The righteous mind: Why good people are divided by politics and religion*. Vintage.

Jonathan Haidt and Jesse Graham. 2007. When morality opposes justice: Conservatives have moral intuitions that liberals may not recognize. *Social Justice Research*, 20(1):98–116.

Evelyn Hatch et al. 1992. *Discourse and language education*. Cambridge University Press.

Robert XD Hawkins, Andreas Stuhlmüller, Judith Degen, and Noah D Goodman. 2015. Why do you ask? good questions provoke informative answers. In *CogSci*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural nlp. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013.

Geert H Hofstede. 2001. *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. sage.

Samee Ibraheem, Gaoyue Zhou, and John DeNero. 2022. Putting the con in context: Identifying deceptive actors in the game of mafia. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 158–168, Seattle, United States. Association for Computational Linguistics.

Julian Jara-Ettinger, Hyowon Gweon, Laura E Schulz, and Joshua B Tenenbaum. 2016. The naïve utility calculus: Computational principles underlying commonsense psychology. *Trends in cognitive sciences*, 20(8):589–604.

Catalina Jaramillo, Megan Charity, Rodrigo Canaan, and Julian Togelius. 2020. Word autobots: Using transformers for word association in the game codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 16, pages 231–237.

Alan Jern, Christopher G Lucas, and Charles Kemp. 2017. People learn other people's preferences through inverse decision-making. *Cognition*, 168:46–64.

Oliver P John, Eileen M Donahue, and Robert L Kentle. 1991. Big five inventory. *Journal of Personality and Social Psychology*.

Aditya Joshi, Pushpak Bhattacharyya, Mark Carman, Jaya Saraswati, and Rajita Shukla. 2016. How do cultural differences impact the quality of sarcasm annotation?: A case study of Indian annotators and American text. In *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 95–99, Berlin, Germany. Association for Computational Linguistics.

Jihen Karoui, Farah Benamara, Véronique Moriceau, Viviana Patti, Cristina Bosco, and Nathalie Aussenac-Gilles. 2017. Exploring the impact of pragmatic phenomena on irony detection in tweets: A multilingual corpus study. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 262–272, Valencia, Spain. Association for Computational Linguistics.

Fereshte Khani, Noah D. Goodman, and Percy Liang. 2018. Planning, inference and pragmatics in sequential language games. *Transactions of the Association for Computational Linguistics*, 6:543–555.

Andrew Kim, Maxim Ruzmaykin, Aaron Truong, and Adam Summerville. 2019. Cooperation and codenames: Understanding natural language processing via codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment*, volume 15, pages 160–166.

Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. 2020. Will I sound like me? improving persona consistency in dialogues through pragmatic self-consciousness. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 904–916, Online. Association for Computational Linguistics.

Alena Korshuk. 2007. Learning more about cultures through free word association data.

Collin J Kovacs, Jasper M Wilson, and Abhilasha A Kumar. 2022. Fast and frugal memory search for communication. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Divya Koyyalagunta, Anna Sun, Rachel Lea Draelos, and Cynthia Rudin. 2021. Playing codenames with language graphs and word embeddings. *Journal of Artificial Intelligence Research*, 71:319–346.

Abhilasha A Kumar, Mark Steyvers, and David A Balota. 2021. Semantic memory search and retrieval in a novel cooperative word game: A comparison of associative and distributional semantic models. *Cognitive Science*, 45(10):e13053.

William Labov. 2011. *Principles of linguistic change, volume 3: Cognitive and cultural factors*, volume 3. John Wiley & Sons.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Angeline Lillard. 1998. Ethnopsychologies: cultural variations in theories of mind. *Psychological bulletin*, 123(1):3.

Angeline Lillard. 1999. Developing a cultural theory of mind: The ciao approach. *Current Directions in Psychological Science*, 8(2):57–61.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv preprint*, abs/1907.11692.

Kate Loveys, Jonathan Torrez, Alex Fine, Glen Moriarty, and Glen Coppersmith. 2018. Cross-cultural differences in language markers of depression online. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 78–87, New Orleans, LA. Association for Computational Linguistics.

George A Miller. 1974. Psychology, language, and levels of communication. In *Human communication*. John Wiley.

George A. Miller. 1992. WordNet: A lexical database for English. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*.

Joan G Miller. 1984. Culture and the development of everyday social explanation. *Journal of personality and social psychology*, 46(5):961.

Will Monroe, Robert X.D. Hawkins, Noah D. Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.

Aaron J Moss, Cheskie Rosenzweig, Jonathan Robinson, and Leib Litman. 2020. Demographic stability on mechanical turk despite covid-19. *Trends in cognitive sciences*, 24(9):678–680.

Ira Noveck. 2018. *Experimental pragmatics: The making of a cognitive science*. Cambridge University Press.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Christopher Potts. 2012. Goal-driven answers in the cards dialogue corpus. In *Proceedings of the 30th West Coast Conference on Formal Linguistics*, pages 1–20. Cascadilla Proceedings Project.

James E Purpura. 2004. *Assessing grammar*, volume 8. Cambridge University Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Yael Ravin and Claudia Leacock. 2000. *Polysemy: Theoretical and computational approaches*. OUP Oxford.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 4512–4525, Online. Association for Computational Linguistics.

Maarten Sap, Ronan LeBras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. *ArXiv preprint*, abs/2210.13312.

Raheleh Saryazdi, Joanne Nuque, and Craig G Chambers. 2022. Pragmatic inferences in aging and human-robot communication. *Cognition*, 223:105017.

Wilbur Schramm. 1954. How communication works. *The process and effects of mass communication*, 3:26.

Julian Schrittwieser, Ioannis Antonoglou, Thomas Hubert, Karen Simonyan, Laurent Sifre, Simon Schmitt, Arthur Guez, Edward Lockhart, Demis Hassabis, Thore Graepel, et al. 2020. Mastering atari, go, chess and shogi by planning with a learned model. *Nature*, 588(7839):604–609.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. 2019. Pragmatically informative text generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4060–4067, Minneapolis, Minnesota. Association for Computational Linguistics.

Blum-Kulka Shoshana, Juliane House, and Gabriele Kasper. 1989. Cross-cultural pragmatics: Requests and apologies. *Grazer Linguistische Studien*.

Richard A Shweder. 1984. Anthropology's romantic rebellion against the enlightenment, or there's more to thinking than reason and evidence. *Culture theory: Essays on mind, self, and emotion*, pages 27–66.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. 2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Felix Soldner, Verónica Pérez-Rosas, and Rada Mihalcea. 2019. Box of lies: Multimodal deception detection in dialogues. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1768–1777, Minneapolis, Minnesota. Association for Computational Linguistics.

Lucia Specia, Stella Frank, Khalil Sima'an, and Desmond Elliott. 2016. A shared task on multimodal machine translation and crosslingual image description. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 543–553, Berlin, Germany. Association for Computational Linguistics.

Darcy Sperlich, Jaiho Leem, and Eui-Jeen Ahn. 2016. The interaction of politeness systems in Korean learners of French. In *Proceedings of the 30th Pacific Asia Conference on Language, Information and Computation: Oral Papers*, pages 163–171, Seoul, South Korea.

Stefanie Stadler. 2012. Cross-cultural pragmatics. *The Encyclopedia of applied linguistics*, pages 1–8.

Naoko Taguchi. 2012. Context, individual differences and pragmatic competence. In *Context, Individual Differences and Pragmatic Competence*. Multilingual Matters.

Jenny Thomas. 1983. Cross-cultural pragmatic failure. *Applied linguistics*, 4(2):91–112.

EV Tikhonova. 2014. Linguistic diagnosing of religious relationships through word association responses. In *Conference proceedings of international multidisciplinary scientific conference on social sciences and arts*, volume 3, pages 505–516.

Chvátil Vlaada. 2015. Codenames – rules - czech games edition | boardgame publisher.

Phillip Wolff and Kevin J Holmes. 2011. Linguistic relativity. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):253–265.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Yuan Yao, Haoxi Zhong, Zhengyan Zhang, Xu Han, Xiaozhi Wang, Chaojun Xiao, Guoyang Zeng, Zhiyuan Liu, and Maosong Sun. 2021. Adversarial language games for advanced natural language intelligence. In *Proceedings of AAAI*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Pei Zhou, Andrew Zhu, Jennifer Hu, Jay Pujara, Xiang Ren, Chris Callison-Burch, Yejin Choi, and Prithviraj Ammanabrolu. 2022. An ai dungeon master's guide: Learning to converse and guide with intents and theory-of-mind in dungeons and dragons. *ArXiv preprint*, abs/2212.10060.

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

A Finalized Codenames Word List

We sample from the following list of 100 words: *luck, grace, soul, fair, life, pass, revolution, change, charge, degree, force, code, genius, compound, time, wake, plot, draft, ghost, play, part, spell, well, point, link, mass, disease, sub, state, alien, space, mine, ray, millionaire, agent, bond, unicorn, figure, war, cycle, boom, sound, trip, centaur, death, club, crash, angel, cold, center, spring, round, date, press, cast, day, row, wind, fighter, embassy, beat, leprechaun, comic, pitch, mount, march, fall, undertaker, green, switch, strike, king, superhero, capital, slip, lead, check, lap, mammoth, air, match, spy, roulette, contract, witch, stock, light, drop, spot, novel, vacuum, cover, scientist, tag, conductor, field, racket, poison, ninja, opera.*

B Reformatting Rationales using GPT-3

Some annotators wrote verbose rationales (*I think fall happens after you slip*), while other annotators were more succinct (*fall after slip*). To prevent models from learning grammar variation across annotators, we normalize our text using GPT-3. We use the following prompt, using hand-written few-shot examples. Some of the examples are unchanged—we include them in the prompt to demonstrate positive examples to the model.

```
Normalize the text, removing
determiners like "the" and "a" at
the start of a sentence, along
with any pronouns.  Correct
spelling and grammar mistakes.
If possible, the final text
should be formatted with the clue
first and the target last or the
target first and the clue last.

clue:  "sub"
target:  "sandwich"
text:  "you can make a sub, which
is a type of sanwich"
output:  "sub is a type of
sandwich"

clue:  "die"
target:  "cliff"
text:  "you may die if you fall
off a cliff"
output:  "die if fall off a
cliff"

clue:  "explosion"
target:  "boom"
text:  "it makes sound"
output:  "explosion makes boom"

clue:  "superman"
target:  "superhero"
text:  "most famous superhero"
output:  "superman is most famous
superhero"

clue:  "night"
target:  "club"
text:  "i love night club"
output:  "night club is a kind of
club"

clue:  "horn"
target:  "air"
text:  "an air horn is a type of
horn"
output:  "air horn is a type of
horn"

clue:  "ivy"
target:  "poison"
text:  "poison ivy is a well
known plant"
output:  "poison ivy is a well
known plant"

clue:  "month"
target:  "march"
text:  "march is a month"
output:  "march is a month"

clue:  "{clue}"
target:  "{target}"
text:  "{text}"
output:  "
```

C Example Generations

Here, we include example generations for a subset of our tasks, illustrating the influence of sociocultural factors on generated Codenames gameplay.

C.1 Clue Generation

Below, we highlight more clues generated with-/without sociocultural priors. Note how some of the without generations are euro-centric: space →

nasa, {revolution, king} → war; adding priors creates more specific clues. However, this isn't always true: target words {pass, check} → leads to poker instead of overtake when conditioned on priors. We suspect that the average player in our pool is not aware of how {pass, check} are associated with poker, resulting in a more generic generation.

Target	Without	With	Gold
revolution, king	war	guevara	overthrow
check	mate	inspect	examine
space	nasa	galaxy	universe
compound	wall	house	together
pass, check	overtake	poker	go

Table 6: Clue generations with/without sociocultural priors, given target words on the board

C.2 Clue Framing

Additional generations can be found in Table 7. Again, we observe that adding sociocultural priors increases relation specificity.

D Annotation Task Details

D.1 Qualification Test

To qualify for the HIT, workers were required to complete a consent form detailing dataset collection and release; and were expected to watch an instructional video outlining game rules.

Then they had to pass the following qualifying test, answering at least 6 out of 7 questions correctly.

1. **True or False:** "angry dog" is an example of a clue you could give. [*Answer*: **False**]

2. **True or False:** you and your partner have different lists of black (assassin) words. [*Answer*: **True**]

3. **True or False:** it is possible to skip a turn without guessing. [*Answer*: **False**]

4. **True or False:** the tan "down" arrow indicates that you guessed the word wrong, while the tan "up" arrow indicates that your partner guessed it wrong. [*Answer*: **True**]

5. **Multiple Choice:** Which of the following kinds of phrases does not follow from our list of target rationales types? [*Answer*: **(b)**]
   (a) "a computer has a mouse"
   (b) "a doctor is smart"
   (c) "a dog is a kind of animal"
   (d) "a disease causes people to be sick"

6. **Multiple Choice:** How many guesses do you get (assuming there are still more words left to guess) [*Answer*: **(d)**]
   (a) you get three guesses each turn
   (b) the number of guesses you get is the same as the number of target words your partner's clue
   (c) as long as you keep picking green words, you can keep guessing, up to the number of target words in your partner's clue
   (d) as long as you keep picking green words, you can keep guessing without any limit, even if you guess more than the number of target words in your partner's clue

7. **Multiple Choice:** During the 8th timer token in the video, it looked like my grid froze and I couldn't make any more guesses. Why did this happen? [*Answer*: **(b)**]
   (a) I guessed an assassin word
   (b) I already guessed all my partner's words correctly
   (c) I clicked the "end game" button
   (d) My partner left the game

D.2 Demographic, Personality, and Moral Questionnaires

Before starting any HITs, workers also had to complete three standardized surveys about their moral foundations, personality, and demographic information. The survey questions and worker statistics are given as follows.

D.2.1 Worker Demographics

Questionnaire. Please answer these 8 questions about yourself.

1. With what gender do you identify? {*Woman, Man, Transgender, Non-binary / non-conforming, Other*}

2. What is your age? {*0-17 years old, 18-22 years old, 22-30 years old, 30-45 years old, 45+*}

Target	Clue	Without	With	Gold
explode	boom	explode causes boom	bomb explodes **with a** boom	explosions make a boom sound
horse	unicorn	a unicorn is a horse	unicorn is **a type of** horse	unicorns are similar to horses
racket	tennis	tennis has racket	a racket **is used in** tennis	tennis uses a racket
day	month	day is month	month **has many** days	30 days in a month

Table 7: Example Rationales for Clues, with/without background priors. With priors, we observe that rationales become more specific, mentioning explicit relations between the target and clue.

3. Which best describes your race or ethnicity? {*African-American/Black, Asian, Latino or Hispanic, Native American, Native Hawaiian or Pacific Islander, White / Caucasian*}

4. In which continent are you located? {*North America, Central / South America, Europe, Africa, Asia, Australia*}

5. What is your highest level of education? {*Some High School / No Diploma, High School Diploma, Associate's Degree / Trade School, Master's Degree, Doctorate Degree*}

6. What is your marital status? {*Single and never married, Married or in a domestic partnership, Widowed, Divorced, Separated*}

7. Which of the following would you consider your native language {*English, Arabic, French, Mandarin, Spanish, Other*}

8. If applicable, please specify your religion {*Buddhism, Catholicism/Christianity, Hinduism, Islam, Judaism, Other*}

Results. Of the 153 unique players, 124 are from the U.S, 12 are from India, 8 are from Brazil, 3 from the U.K, 2 from Canada, and the rest are single players from the following 7 countries: Indonesia, Costa Rica, France, South Africa, Germany, and Portugal.

D.2.2 Worker Personality

Big 5 Personality Questionnaire. Please answer these 10 questions about yourself on the following scale: [-2] Strongly Disagree; [-1] Disagree; [0] Neutral; [1] Agree; [2] Strongly Agree.

1. I see myself as someone who does a thorough job.
2. I see myself as someone who is reserved.
3. I see myself as someone who is outgoing, sociable.
4. I see myself as someone who gets nervous easily.
5. I see myself as someone who has few artistic interests.
6. I see myself as someone who is relaxed, handles stress well.
7. I see myself as someone who tends to find fault with others.
8. I see myself as someone who is generally trusting.
9. I see myself as someone who tends to be lazy.
10. I see myself as someone who has an active imagination.

D.2.3 Moral Foundations And Political Leaning.

Moral Foundations Theory. Following Haidt and Graham (2007), we use the five-foundation theory of moral reasoning to understand our players' values and leanings. This theory does not give explicit definitions for the five foundations, but following recent work by Ziems et al. (2022), we can assume the following definition sketches:

1. **Care:** wanting someone or something to be safe, healthy, and happy.
   Harm: wanting someone or something to suffer physically, emotionally, socially, intellectually, or spiritually.

2. **Fairness:** wanting to see individuals or groups treated equally or equitably
   Cheating: wanting to see unfairness, injustice, bias, exclusion, or discrimination.

3. **Loyalty:** wanting unity and seeing people keep promises or obligations to an in-group.
   Betrayal: wanting to see people lie, abandon an in-group, or become isolated and divided.

4. **Authority:** wanting to respect social roles, duties, privacy, peace, and order.
   Subversion: wanting to see people disrespect, disobey or cause disorder, challenge the status-quo, and do what they do not have permission to do.

5. **Sanctity:** wanting people and things to be clean, pure, innocent, and holy.
   Degradation: wanting people to follow selfish or crude desires and do things that make them or others dirty, corrupt, sick, repulsive, or perverted.

Moral Foundations Questionnaire We use the associated Moral Foundations Questionnaire, which we shortened to 12 questions as follows.

Please answer 12 questions about "right" and "wrong." The prompts are the same in each case, but the considerations are different.

1. When you decide whether something is right or wrong, to what extent are the following considerations relevant to your thinking? Use the following scale: [0] Not at all relevant (It has nothing to do with my judgments of right and wrong); [1] Not very relevant; [2] Slightly relevant; [3] Somewhat relevant; [4] Very relevant; [5] Extremely relevant (It is one of the most important factors when I judge right and wrong)
   (a) Whether or not someone suffered emotionally.
   (b) Whether or not some people were treated differently than others.
   (c) Whether or not someone's action showed love for his or her country.
   (d) Whether or not someone showed a lack of respect for authority.
   (e) Whether or not someone violated standards of purity and decency.
   (f) Whether or not someone was good at math.
   (g) Whether or not someone cared for someone weak or vulnerable.
   (h) Whether or not someone acted unfairly.
   (i) Whether or not someone did something to betray his or her group.
   (j) Whether or not someone conformed to the traditions of society.

2. Which of the following best describes your political views?
   (a) Liberal
   (b) Moderate Liberal
   (c) Moderate Conservative
   (d) Conservative
   (e) Libertarian

D.3 Instructions for Writing Rationales

We explain that rationales should use at least 3 words to describe the connection between the clue and the target. Annotators were encouraged to be creative while trying to use one of the structures below. We imposed these structures for the sake of regularity.

1. `MERONYM` x has y
   (a) a dog has a tail
   (b) the pacific ocean has water

2. `HYPERNYM` x is a kind of y
   (a) bunkbed is a kind of bed
   (b) whisper is a kind of communication

3. `SYNONYM` x means the same thing as y
   (a) car means the same thing as automobile
   (b) sluggish means the same thing as slow

4. `ANTONYM` x means the opposite of y
   (a) civilian means the opposite of soldier
   (b) fast means the opposite of slow

5. `ADJECTIVE` x describes y
   (a) brave describes a firefighter
   (b) scary describes a clown

6. `AGENT` x does y
   (a) a star does twinkle police do make an arrest

7. `CAUSE` x causes y

(a) a bed causes people to sleep
(b) an oven causes food to bake
(c) a disease causes people to be sick

8. PATIENT x acts on y
   (a) a wrench acts on a bolt
   (b) a doctor acts on a patient
9. LOCATION x has an environment y
   (a) a star has an environment firmament

E Training and Hyperparameters

For our generation tasks, we perform use 5e-5 as our initial learning rate and perform a hyperparameter search over {1...20} epochs. For classification, we use the same splits and perform a hyperparameter sweep over learning rates ({1e-4, 5e-4, 1e-5, 5e-5, 1e-6, 5e-6}) and epochs ({1...15}). All models were trained on an NVIDIA A100 GPU. Across all experiments, GPU compute time was around 4-5 days.

F Artifact Details

We use several models in our paper for their intended retrieval or generation task. Each model has its own license and number of parameters, listed below:

1. T5 (Raffel et al., 2020), 220M parameters, is under the Apache 2.0 License.
2. BART (Lewis et al., 2020), 140M, is under the Apache 2.0 License.
3. fastText (Bojanowski et al., 2017) is under the MIT License.
4. SentenceBERT (Reimers and Gurevych, 2020), 33M variant, is under the Apache 2.0 License.
5. BERT (Devlin et al., 2019) base, 110M, is under the Apache 2.0 License.
6. XLNet (Yang et al., 2019) base, 110M, is under the Apache 2.0 License.
7. RoBERTAa (Liu et al., 2019) base, 123M, is under the Apache License 2.0.

We plan on releasing CULTURAL CODES and corresponding code under Creative Commons Attribution Share Alike 4.0 International. While our released dataset has extensive demographic information, we do not collect any identifiers that can uniquely isolate a person (e.g. name, MTurk ID, etc.)

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 8

☑ A2. Did you discuss any potential risks of your work?
Section 9

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract + Introduction

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 3 for our introduced dataset, and we cite all baseline models (Section 5)

☑ B1. Did you cite the creators of artifacts you used?
Section 5

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Appendix Section E

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Yes, Section 9

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Appendix E

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3 and Section 5

C ☑ Did you run computational experiments?

Yes, Section 5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix D and E

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 5 and Appendix D

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Our results are averaged across 5 runs; Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Yes, Section 3.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Yes, Section 3.3 and Appendix C

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 3.3

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Appendix C.1

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 3

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 3.4