Werewolf Among Us: Multimodal Resources for Modeling Persuasion Behaviors in Social Deduction Games

Bolin Lai[1*]   **Hongxin Zhang**[2*]   **Miao Liu**[3*]   **Aryan Pariani**[1*]   **Fiona Ryan**[1]
Wenqi Jia[1]   **Shirley Anugrah Hayati**[4]   **James M. Rehg**[1]   **Diyi Yang**[5]

[1]Georgia Institute of Technology [2]Shanghai Jiao Tong University
[3]Meta AI [4]University of Minnesota [5]Stanford University

{bolin.lai, apariani3, fkryan, wenqi.jia, rehg}@gatech.edu
icefox@sjtu.edu.cn, miaoliu@meta.com
hayat023@umn.edu, diyiy@cs.stanford.edu

Abstract

Persuasion modeling is a key building block for conversational agents. Existing works in this direction are limited to analyzing textual dialogue corpora. We argue that visual signals also play an important role in understanding human persuasive behaviors. In this paper, we introduce the first multimodal dataset for modeling persuasion behaviors. Our dataset includes 199 dialogue transcriptions and videos captured in a multi-player social deduction game setting, $26,647$ utterance level annotations of persuasion strategy, and game level annotations of deduction game outcomes. We provide extensive experiments to show how dialogue context and visual signals benefit persuasion strategy prediction. We also explore the generalization ability of language models for persuasion modeling and the role of persuasion strategies in predicting social deduction game outcomes. Our dataset can be found at https://persuasion-deductiongame.socialai-data.org. The codes and models are available at https://github.com/SALT-NLP/PersuationGames.

1 Introduction

As humans, from childhood, we develop the ability to attribute mental belief states to ourselves and others (Premack and Woodruff, 1978). Moreover, we constantly exhibit persuasive behaviors to influence and even reshape the belief states of others during our daily social interactions (Lonigro et al., 2017). An automatic system with the ability to understand human persuasion strategies and deduce human belief states may enable more proactive human-computer interaction, and facilitate collaborative decision-making processes.

Prior works targeted at understanding the persuasion strategies utilized on online forums like Reddit, crowd-funding platforms (Yang et al., 2019; Chen and Yang, 2021; Atkinson et al., 2019), and in 1-on-1 dialogues under simulated scenarios through the Amazon Mechanical Turk platform (Wang et al., 2019; Chawla et al., 2021).

However, the persuasive behaviors during naturalistic group discussions with face-to-face conversation remain unexplored. More importantly, daily human social interaction is multimodal by nature. Both verbal communication (*e.g.* language and audio) and non-verbal communication (*e.g.* gesture and gaze behavior) are essential for analyzing persuasive behavior.

Moreover, resources for understanding how persuasion strategies affect decision and deduction outcomes during social interactions are missing from the language technologies community.

To bridge these gaps, we introduce *the first multimodal benchmark dataset for modeling persuasive behaviors during multi-player social deduction games*. As shown in Fig. 1, our dataset is captured in a naturalistic setting where groups of participants play social deduction games[1]. Our dataset contains both video recordings and the corresponding dialogue transcriptions. The video data is sourced from both the Ego4D Social dataset (Grauman et al., 2022) and YouTube videos. Our dataset also has annotations for persuasion strategy at the utterance level and the voting outcome of each participant during the social deduction game.

We benchmark our dataset by providing comprehensive experimental results and analyzing the role

---

* denotes equal contribution.

[1]We consider two games in our dataset: One Night Ultimate Werewolf and The Resistance: Avalon. Appendix A describes the rules of these two games.
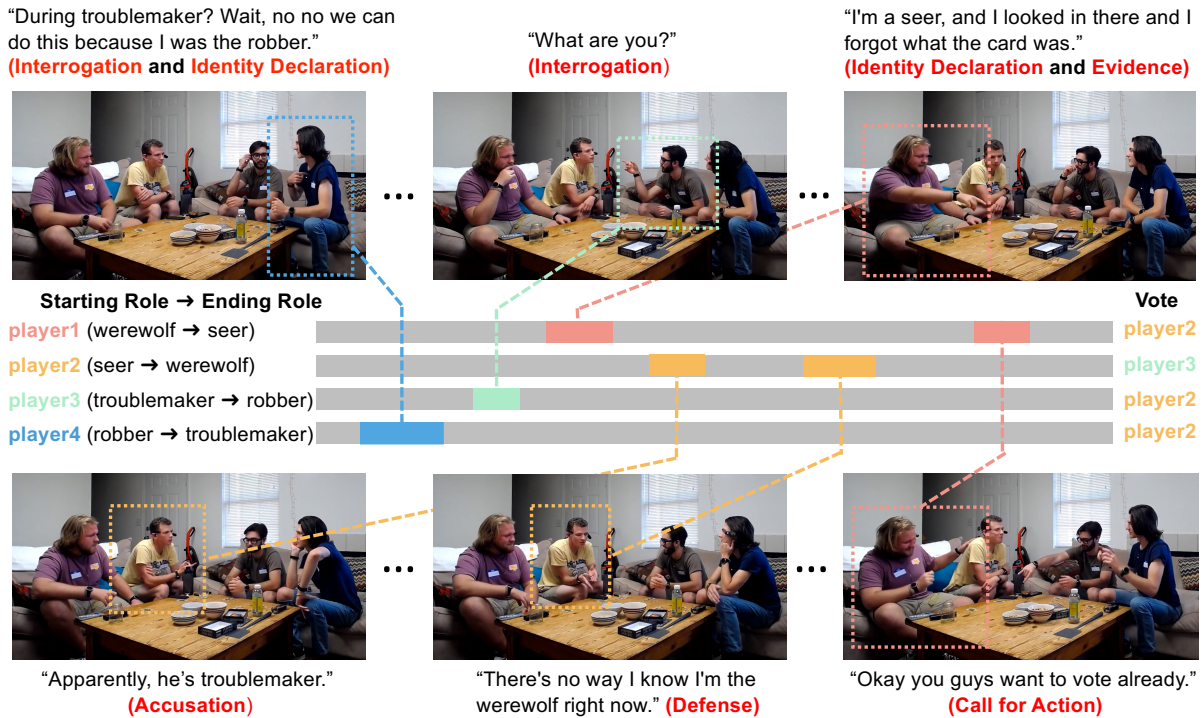
Figure 1: Demonstration of the six persuasion strategies included in our dataset and the corresponding video. The players are numbered as 1,2,3,4 from left to right. Players' roles might be changed during the game. In this example, player1's and player2's cards were swapped by the troublemaker, and player3's and player4's cards were swapped by the robber. Player 2 voted for player 3 at the end while the others voted for player 2.

of the video modality and contextual cues in designing computational models for persuasion behavior prediction. We also provide results to show how different computational models generalize across different data sources and different games. Our contributions are summarized as follows:

- We present the first multimodal dataset for persuasion modeling. Our dataset is collected in naturalistic social game scenarios with intensive face-to-face group conversations.

- We conduct comprehensive experiments to show the importance of context and visual signals for persuasion strategy prediction.

- We provide additional experimental results to investigate model generalization on the persuasion modeling task and discuss how persuasion strategy influences the game voting outcome.

2 Related Work

Persuasive Behaviors Understanding. A few previous works introduce datasets for the computational modeling of persuasion (Yang et al., 2019; Chen and Yang, 2021; Chawla et al., 2021; Wang et al., 2019; Luu et al., 2019; Atkinson et al., 2019). As summarized in Table 1, existing works collect persuasion language data from online platforms

Prior Works	Interaction	Modalities	Setting
(Yang et al., 2019)	Online	Text	Loan
(Chen and Yang, 2021)	Online	Text	Request
(Chawla et al., 2021)	1 on 1	Text	Negotiation
(Wang et al., 2019)	1 on 1	Text	Charity
(Luu et al., 2019)	1 on 1	Text	Debate
(Atkinson et al., 2019)	Online	Text	Reddit
Ours	Group Discussion	Text+Video +Audio	Deduction Game

Table 1: Previous datasets for computational persuasion.

where real-time communication is not available. Moreover, these datasets mainly contain 1-on-1 conversations and lack conversations among multiple speakers. In contrast to these prior efforts, our dataset targets capturing persuasive behaviors in a social group setting of 4 to 6 people.

Multimodal Social Interaction. A rich set of literature has addressed the problem of multimodal sentiment analysis (Xu et al., 2021b; Li et al., 2020; Lu et al., 2019; Xu et al., 2021a). We refer to a recent survey (Kaur and Kautish, 2022) for a more detailed discussion on this topic. The Ego4D Social benchmark (Grauman et al., 2022) includes the tasks of identifying who is looking at and talking to the camera wearer using video

6571

and audio. Bara et al. (2021) adopts a multimodal approach to understanding dialogue behavior in a simulated setting. The most relevant work is Bai et al. (2021), which adopts a multimodal method to predict the debate outcome from a TV show. In contrast, we address the challenging tasks of predicting the utterance-level persuasion strategy and the deduction outcome from a naturalistic conversation, which requires a richer understanding of high-level social behaviors.

Computational Modeling of Deduction Games. Prior works have investigated computational models for social deduction games. One stream of work seeks to analyze strategies and develop AI agents that play deduction games using a game theory approach (Nakamura et al., 2016; Serrino et al., 2019; Chuchro, 2022; Braverman et al., 2008; Bi and Tanaka, 2016). These works focus on models of the state of the game alone and do not address understanding the dialogue and persuasive behaviors that often occur while playing. More relevantly, Chittaranjan and Hung (Chittaranjan and Hung, 2010) developed a model for predicting Werewolf game outcomes from player speaking and interrupting behaviors. Recently, (FAIR) introduced a game agent–CICERO, that achieves human-level performance in the Diplomacy game by leveraging a language model with planning and reinforcement learning algorithms. In contrast to these prior works, we present the first work for understanding persuasive behaviors in a group setting from a multimodal perspective.

3 Dataset

3.1 Data collection

To benchmark the generalization ability of the computational models, we collect our data from different sources, as detailed in this section. This work was approved by an Institutional Review Board.

Ego4D Dataset We first leverage a subset of Ego4D Social dataset (Grauman et al., 2022) for our study. This subset captures videos of groups of participants playing social deduction games. This subset contains 7.3 hours of videos with 40 games of One Night Ultimate Werewolf and 8 games of The Resistance: Avalon. Note that the Avalon data has a relatively small scale, and therefore is only used to evaluate the cross-domain game generalization ability of our models. To ensure all participants are visible in the frame, we use third-person videos

instead of the first-person videos from Ego4D for visual representation learning and transcription.

YouTube Video We retrieve the top search results for YouTube videos using the keywords of "one night ultimate werewolf" and "ultimate werewolf". We manually select from the searched videos to make sure the they adopt a similar game setup as the Ego4D data. Specifically, we filter all results with more than 5 players or fewer than 4 players, and those using game roles from the expansion package. We finally collect a final set of 14.8 hours of videos with 151 clips of completed games that adopt the same game setup as the Ego4D dataset and have fully visible game outcomes. We will release the YouTube URLs for the selected videos.

3.2 Data Annotation

Video Annotation Most Ego4D and YouTube videos contain multiple games. Therefore, we first annotate the starting time (when the game narration voice begins) and the ending time (right before the voting stage) of each game. We then ask the annotators to look through each game clip and annotate the starting role, ending role, and the voting outcome of each player.

Transcription We use an automatic transcription service *rev.com* to generate the transcript of each game clip. We further ask annotators to carefully examine the alignment of the videos and transcripts, and manually correct any errors in the transcripts. Please refer to Appendix B for more details.

Persuasion Strategy Annotation Inspired by prior psychology studies and other works on predicting persuasion strategies (Chawla et al., 2021; Carlile et al., 2018; Yang et al., 2019; Chen and Yang, 2021), we propose six persuasion tactics that are frequently adopted in social deduction games.

- **Identity Declaration**: *State one's own role or identity in the game.* This is a game-specific persuasion tactic.

- **Accusation**: *Claim someone has a specific identity or strategic behavior.* Accusation, similar to Undervalue-Partner (Chawla et al., 2021), is a generic proself behavior.

- **Interrogation**: *Questions about someone's identity or behavior.* Interrogation, is a proself strategy related to individual preferences.

Label	Example	Ego4D			YouTube		
		Count	**AUL**	α	**Count**	**AUL**	α
Identity Declaration	"I'll just come out and say I was a villager, so I have no idea what's going on."	293	9.87	0.90	1066	10.43	0.87
Accusation	"So James might be the werewolf."	669	11.28	0.74	2830	11.06	0.67
Interrogation	"Who did you rob?"	695	7.56	0.80	3407	7.66	0.90
Call for Action	"We shouldn't vote to not kill anyone. And then there could also be no werewolf."	236	9.99	0.78	1163	9.53	0.71
Defense	"I think that you accused me of being a Werewolf very quickly."	570	10.04	0.62	2696	9.75	0.80
Evidence	"If you swapped these two, he is not the werewolf."	489	11.45	0.75	1740	9.80	0.60

Table 2: Utterance-level persuasion strategy annotations. AUL refers to the average utterance length in terms of the number of words in an utterance and α refers to Krippendorff's alpha.

- **Call for Action**: *Encourage people to take an action during the game*. Call for Action relates to the coordination for persuasion (Chawla et al., 2021), which is a generic prosocial behavior.

- **Defense**: *Defend oneself or someone else against an accusation or defend a game-related argument*. An utterance demonstrates Defense when the persuader tries to use credentials to earn others' trust or justify their earlier decisions.

- **Evidence**: *Provide a body of game-related fact or information*. Evidentiality is a general persuasion tactic that has been widely studied in previous works (Carlile et al., 2018).

Following previous work (Chawla et al., 2021), we annotate the persuasion strategy at the utterance level. We provide a website annotation tool adopted from Hayati et al. (2020) for our annotation task (see Appendix C for details) to the annotators. To properly train the annotators, we first ask all three annotators to annotate the same subset of dialogues and compute inter-annotator agreement using the nominal form of Krippendorff's alpha (Krippendorff, 2018). We then discuss with the annotators on their disagreements and come up with a general rule to address the disagreements during the annotation process. We repeat the above process until the annotators reached a Krippendorff's alpha greater than 0.6 for each category. Despite the subjectivity of persuasion strategies, the previous work (Chawla et al., 2021) suggests that annotations from 3 people are reasonable enough to most humans if they have a Krippendorff's alpha greater than 0.6. In Table 2, we report the per-class Krippendorff's alpha value for the final round of

inter-annotator agreement calculation. After the annotator training phase is completed, we ask the three annotators to independently annotate the rest of the Ego4D and YouTube data.

Annotation Statistics Our dataset has $5,815$ utterances from the Ego4D data and $20,832$ utterances from the YouTube data. More than 49.2% of Ego4D utterances are labeled as no strategy because of the naturalistic social setting, while only 37.9% of YouTube utterances are labeled as no strategy since players from the YouTube videos are more proficient at the game and focused more on gameplay. Furthermore, as shown in Table 2, the adopted persuasion strategies have an imbalanced distribution, where "Accusation", "Interrogation", and "Defense" are the most frequent strategies for both the Ego4D and YouTube videos. The annotators are recruited from a Startup Data Platform dedicated to research projects. All annotators are paid hourly at a rate above the federal minimum.

4 Strategy Prediction

Given an utterance and its corresponding video segment, we seek to predict the persuasion strategies adopted in the utterance. We first leverage a pre-trained language model (Devlin et al., 2019; Liu et al., 2019) as the text encoder to obtain the utterance embedding, and a vision transformer (Fan et al., 2021) to obtain the visual embedding. We then concatenate the textual and visual features to predict the persuasion strategy. Additionally, we study the impact of textual context by including prior utterances as input.

"I started out as the seer and I looked at these two cards and I saw a werewolf here and then the drunk here."

Text Encoder ϕ $\mathcal{T}$ $\phi(\mathcal{T})$

$\oplus$

$\psi(\mathcal{V})$

Fixed $\mathcal{V}$ Video Encoder ψ

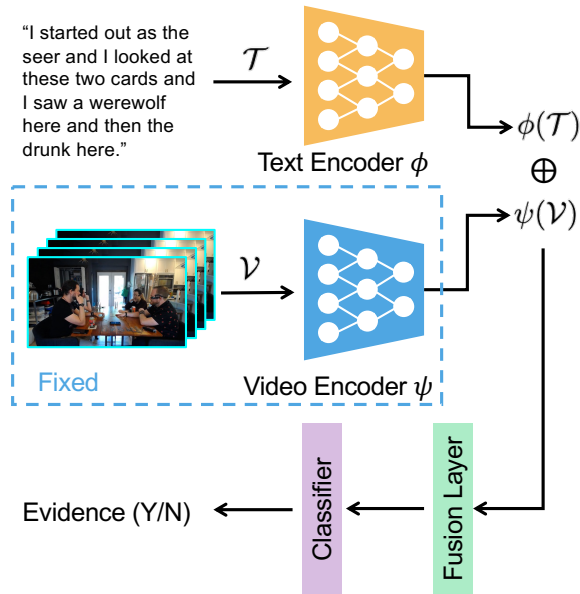Evidence (Y/N) ← Classifier ← Fusion Layer ←

Figure 2: Architecture of the independent model for each strategy. We fix the parameters in the video encoder and train the other modules end-to-end. $\oplus$ denotes the concatenation of two feature representations

4.1 Methodology

In our dataset, an utterance may be labeled with multiple persuasion strategies. For instance, *"I'm a villager and she is the werewolf."* is labeled as both *identity declaration* and *accusation*. Therefore, we formulate this task as a binary classification problem for each strategy and consider an utterance as non-strategic if it gets negative labels in all strategies. The most straightforward approach to solving this task is fine-tuning a pre-trained language model, which is referred to as *Base* model. In addition, we consider the following approaches:

Modeling with Context Embedding. Since some persuasion strategies cannot be easily recognized from one single utterance, we further consider a model with additional context (prior utterances) for each utterance. This is denoted as *Base + C*.

Modeling with Video Representation. We further leverage the non-verbal signals by combining video features with the text representation for persuasion modeling. We directly use a pre-trained Vision Transformer to extract video representations, and fuse the video and text representations before feeding them into the classification layer as shown in Fig. 2. We refer to this model as *Base + V*.

Late fusion of Video and Context. Finally, we adopt a late fusion model (*Base + C + V*) that incorporates both video features and context cues for persuasion strategy prediction.

4.2 Model Details

We perform our experiments using both **BERT** (Devlin et al., 2018) and **RoBERTa** (Liu et al., 2019) as backbones for the text encoder. We use the `bert-base-uncased` and `roberta-base` models from Huggingface (Wolf et al., 2020) in our implementation. We adopt MViT-B-24 (Fan et al., 2021) pretrained on Kinetics-400 (Kay et al., 2017) as the video encoder. Moreover, we also implement a multi-task model (Chawla et al., 2021) as an additional baseline, referred to as **MT-BERT**. Context and video features are incorporated into MT-BERT in the same way as BERT and RoBERTa.

Base Model. For base models, we obtain the textual input $\mathcal{T}$ from the current utterance only. Then we input $\mathcal{T}$ into a text encoder ϕ followed by a classifier to get the strategy prediction.

Base + C. We first concatenate the k previous utterances $C_1, C_2, \cdots, C_k$ with an *[EOS]* token to get context C, and then concatenate this with the current utterance U using a *[SEP]* token to get the final input $\mathcal{T}$. Formally, we have

$$C = C_1 \,[\text{EOS}]\, C_2 \,[\text{EOS}] \cdots C_k, \qquad (1)$$
$$\mathcal{T} = C \,[\text{SEP}]\, U. \qquad (2)$$

Base + V. We use video encoder ψ to extract visual representation $\psi(\mathcal{V})$ of the corresponding video clip $\mathcal{V}$. During training, video features are concatenated with the text representation $\phi(\mathcal{T})$ and fed into a fusion layer, which uses a linear mapping function W_F^T and an activation function $Tanh(\cdot)$. Finally, we apply a linear classifier W_P^T to obtain the prediction logits, which can be formulated as

$$logits = W_P^T \cdot Tanh\left(W_F^T \cdot (\phi(\mathcal{T}) \oplus \psi(\mathcal{V}))\right), \qquad (3)$$

where $\oplus$ denotes the concatenation of two vectors. Note that we fix the parameters of the video encoder during training. Please refer to Appendix D for more details on visual representation extraction.

Base + C + V. We further late fuse **+ C** with **+ V**. Formally, we denote the probability predictions of the two models after softmax as P_C and P_V. Then the output after linear combination is formulated as

$$P_{C,V} = (1 - \lambda)P_C + \lambda P_V, \qquad (4)$$

where λ is a scalar that balances P_C and P_V.

4.3 Training Details

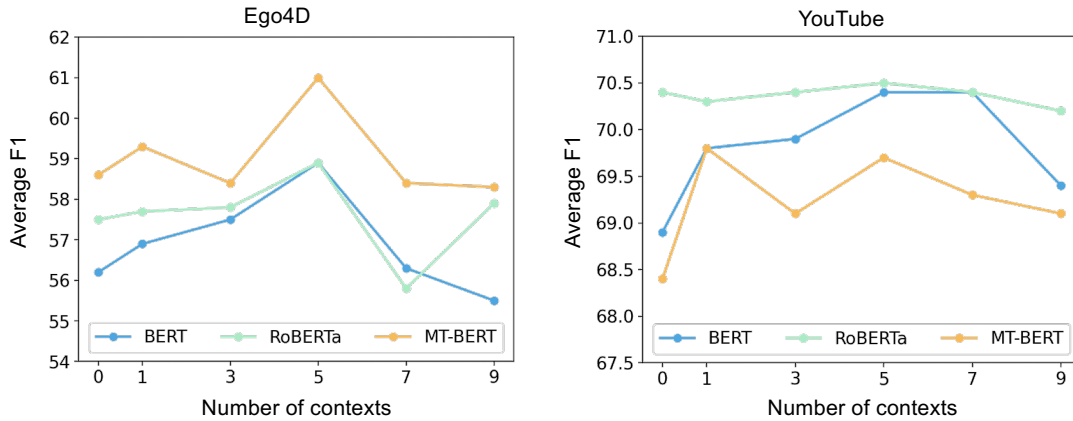All models are trained using cross-entropy loss. For training hyper-parameters, we do a grid search

Figure 3: Ablation study of adopting different context lengths for persuasion strategy prediction.

of learning rates in $\{1e-5, 3e-5, 5e-5\}$ and batch sizes in $\{16, 8\}$ for *Base* models. We then fix the optimal hyper-parameters for subsequent models incorporating context or videos. We train all models with the optimal learning rates and batch sizes for 10 epochs using AdamW (Loshchilov and Hutter, 2017) as the optimizer. We run all the experiments with three random seeds and report the average score and the standard deviation.

4.4 Experiment Results

Evaluation Metrics. Following (Chawla et al., 2021), we report the F1 score for each persuasion strategy category, the average F1 score of all categories, and Joint Accuracy. Note that the prediction is considered as correct when all the categories are predicted correctly in Joint Accuracy.

Ablations on Additional Context. We first present a systematic ablation study of how incorporating textual context may improve the performance of persuasion strategy prediction. Specifically, we feed a fixed length of previous utterances together with the current utterance into the backbone language encoders for classification. As shown in Fig. 3, the additional context can boost the performance of all baseline models. However, setting the context length too long may confuse the model, especially for categories that can be reliably predicted with the current utterance (*e.g.* Identity Declaration, and Interrogation). We present the per-class performance in Appendix E. Our empirical finding is that a context length of 5 can consistently improve the performance of all three baseline models. Therefore, we adopt a context length of 5 as a default setting for the rest of our experiments.

Modeling with Video Representation. We further study how incorporating video representation im-

proves the performance of persuasion modeling. The results are summarized in Table 3. Importantly, video features can improve the BERT model by 0.8% on both the Ego4D dataset and YouTube dataset. However, RoBERTa+V only beats the RoBERTa model by 0.2% on the YouTube dataset. This may be because the YouTube dataset has more training data which enables the RoBERTa model to learn a robust representation without video feature embedding. Interestingly, including video features has a larger performance boost on predicting "Accusation", "Interrogation", and "Call for Action", which is likely due to the more frequent non-verbal communication (*e.g.* pointing to someone, raising hands, turning the head) during these persuasive behaviors.

Off-the-shelf GPT-3 Inference. Prompting Large Language Models off-the-shelf to solve NLP tasks has received increasing attention (Brown et al., 2020). Here, we experiment with GPT-3-175B on our benchmark under three settings: zero-shot, one-shot and five-shot. Specifically, we use the `text-davinci-002` engine from OpenAI's API[2] with temperature 0 to produce a deterministic answer. The detailed templates for different settings are shown in Appendix G. The result is shown in Table 4. Using GPT-3 off-the-shelf achieves a non-trivial performance (Joint-A of 52.0 v.s. 38.8 for majority on YouTube data). Adding more examples further boosts performance, though it is still inferior to the fine-tuned models.

Data Domain Generalization. We conduct additional experiments to show the generalization ability of language models on persuasion prediction. To begin with, we use the model trained on the

---

[2]https://beta.openai.com/

6575

	Method	Identity	Accusation	Interrogation	Call for Action	Defense	Evidence	Avg F1	Joint-A
Ego4D	BERT	82.6±1.1	48.8±4.8	82.8±0.2	39.4±9.6	29.3±5.5	54.2±2.5	56.2±2.5	65.1±1.6
	BERT + C	79.9±1.6	52.0±3.3	81.0±1.1	49.5±3.2	33.8±0.5	57.1±1.6	58.9±0.6	65.0±0.2
	BERT + V	81.5±3.5	52.1±1.9	83.3±1.6	42.4±3.8	28.4±5.1	52.8±1.0	56.7±1.2	64.5±1.2
	BERT + C + V	**84.5**±4.6	52.8±2.0	82.7±0.4	47.3±3.4	34.5±1.7	54.9±1.1	59.4±1.6	**66.5**±0.3
	RoBERTa	81.7±2.6	51.7±0.9	83.4±0.9	43.3±8.7	33.1±2.2	51.7±2.1	57.5±1.4	63.4±0.5
	RoBERTa + C	81.5±0.7	**59.4**±2.4	83.5±1.1	43.7±3.7	33.0±3.1	52.4±2.9	58.9±1.2	64.6±0.7
	RoBERTa + V	79.8±0.6	51.4±1.0	82.8±2.1	50.1±5.3	31.3±3.1	54.6±3.2	58.3±0.7	64.0±0.9
	RoBERTa + C + V	82.7±0.2	58.5±2.3	83.8±1.2	46.1±4.5	35.4±3.4	53.4±3.3	60.0±0.8	66.1±0.9
	MT-BERT	80.9±1.3	51.5±3.3	83.0±1.3	**56.6**±2.3	25.9±2.0	53.6±1.3	58.6±0.3	65.5±0.8
	MT-BERT + C	79.8±2.2	54.4±0.8	83.2±0.7	50.8±7.2	**36.5**±2.8	**61.5**±2.2	**61.0**±1.1	66.3±1.4
	MT-BERT + V	79.9±1.6	51.9±0.8	**84.8**±2.4	53.9±4.5	35.4±2.2	53.3±1.0	59.8±0.7	62.1±3.4
	MT-BERT + C + V	80.7±1.9	55.2±0.9	83.6±0.6	50.0±0.8	36.1±2.7	60.5±1.0	**61.0**±0.3	66.3±1.0
YouTube	BERT	80.2±1.6	64.7±1.1	89.6±0.4	77.2±2.5	43.5±1.0	58.3±0.7	68.9±0.0	64.6±0.8
	BERT + C	82.6±0.7	66.7±1.0	89.6±1.5	78.1±2.4	45.7±1.1	59.7±1.1	70.4±0.3	64.4±1.0
	BERT + V	82.4±0.5	65.4±1.4	89.7±0.1	78.0±0.8	45.3±2.8	58.4±1.3	69.9±0.4	66.2±0.5
	BERT + C + V	83.6±0.1	67.2±1.2	**90.2**±1.0	78.5±1.6	46.6±1.1	59.9±1.0	71.0±0.2	66.7±0.5
	RoBERTa	84.3±0.1	67.2±0.6	89.4±0.1	78.2±0.8	44.3±0.4	59.0±1.7	70.4±0.2	64.8±0.7
	RoBERTa + C	82.4±0.3	67.0±1.1	**90.2**±0.0	77.1±1.0	46.1±0.7	59.9±0.7	70.5±0.3	64.7±0.6
	RoBERTa + V	83.4±0.4	66.4±0.3	89.5±0.1	**78.7**±2.0	46.6±0.6	59.0±1.0	70.6±0.1	65.3±1.2
	RoBERTa + C + V	83.7±0.6	67.4±0.4	89.8±0.3	78.5±1.2	**48.2**±0.7	60.4±0.8	**71.3**±0.2	66.4±0.7
	MT-BERT	80.7±0.4	65.1±1.5	88.5±0.8	76.2±2.2	42.3±1.5	57.4±1.3	68.4±0.3	65.6±1.1
	MT-BERT + C	83.1±1.1	65.0±1.5	90.1±0.3	74.6±2.4	46.5±0.8	59.2±0.3	69.7±0.6	66.7±0.5
	MT-BERT + V	82.8±0.6	**68.5**±1.0	89.3±0.7	75.6±2.8	47.8±0.3	59.6±0.8	70.6±0.8	66.9±0.4
	MT-BERT + C + V	**84.4**±0.6	68.4±1.0	89.5±0.6	76.5±2.1	47.3±0.5	**60.6**±0.2	71.1±0.5	**68.1**±0.2

Table 3: Experimental Results on incorporating visual features for persuasion strategy prediction. We train an independent model for each category using BERT and RoBERTa backbones. Additionally, we also use the off-the-shelf Multi-Task BERT model (MT-BERT) (Chawla et al., 2021) to jointly predict all categories.

Setting	Ego4D		YouTube	
	Avg F1	Joint-A	Avg F1	Joint-A
Majority	0	52.5	0	38.8
Zero-Shot	35.4	58.5	40.3	52.0
One-Shot	40.7	56.3	47.2	53.2
Five-Shot	47.0	59.7	49.6	53.7

Table 4: GPT-3 results on Ego4D and YouTube data.

YouTube data to make predictions on the Ego4D Werewolf testing data without any fine-tuning. As shown in Fig. 4, the resulting model achieves better performance than models trained only on Ego4D in most cases, due to the larger amount of available training data from the YouTube dataset. This also suggests that, for the text modality, the domain gap between the Ego4D and the YouTube data is small. We further fine-tune the model trained on the YouTube data with the Ego4D training data, and the resulting model performs even better. These results suggest promise in leveraging the large body of videos available online as a pre-training source for persuasion modeling in naturalistic social interactions. Another finding from our experiments is

that the multi-task setting (MT-BERT) may compromise the model's generalization ability. We also find that including video representation cannot improve the model generalization ability (see more details in Appendix D), suggesting that the video modality domain gap between the two data sources is much larger than the text modality.

Game Domain Generalization. We also study the model generalization ability on another social deduction game – Avalon. Werewolf and Avalon are vastly different in the game rules and winning conditions, especially because Werewolf has only one voting round per game, while Avalon has multiple rounds per game. Therefore, the persuasion strategies adopted in Avalon have a different distribution from Werewolf (see Appendix H). We run inference on the Avalon data using models trained only on the Ego4D Werewolf data without fine-tuning. Results are shown in Figure 5. Despite the large domain gap between the two games, our models achieve decent performance on the Avalon data. However, we find incorporating additional context has marginal performance improvements, and may even compromise the performance of the RoBERTa
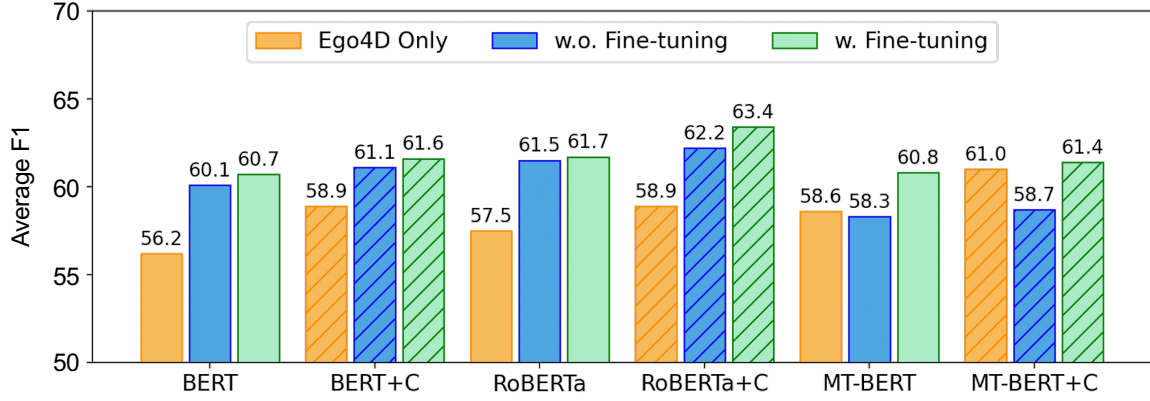
Figure 4: Data domain generalization experiments. We report the testing performance on the Ego4D dataset using models trained only on YouTube data (w.o. Fine-tuning), and trained on YouTube data and further fine-tuned with Ego4D data (w. Fine-tuning). We also report the performance (refer to Table 3) of the models trained only on Ego4D dataset (Ego4D Only) as comparison.
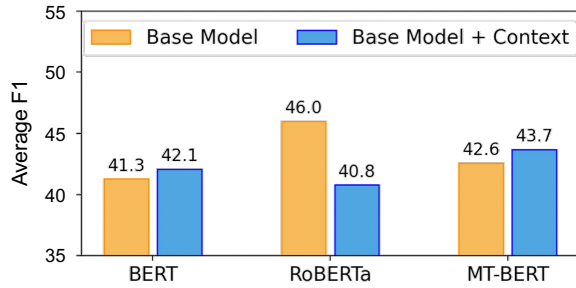

Figure 5: Game domain generalization experiments. We report the testing performance on Ego4D Avalon data using models trained only on Ego4D Werewolf data.

model. The detailed results of data and game domain generalization are shown in Appendix F.

5 Game Outcome Deduction

In addition to predicting persuasion strategies, we further model the human deduction process by predicting the voting outcomes of each pair of players, *i.e.*, whether player A (*voter*) votes for player B (*candidate*). Therefore, in a game of n players, there are C_n^2 (Combinations) of player pairs, corresponding to P_n^2 (Permutations) data points. We merge all data points from the Ego4D Werewolf data and YouTube data to enlarge the dataset size, and split the resulting data into 2741/427/827 samples for train/val/test sets. Since each player is only allowed to vote for one player, the resulting data has an imbalanced distribution, with 20.4% of the samples being positive (positive indicates the voter votes for the candidate).

5.1 Method

For deduction modeling, we encode the input feature with three embeddings: a 7×1 vector rep-

resenting the persuasion strategy distribution (including non-strategy) adopted by the voter; a 7×1 vector representing the persuasion strategy distribution adopted by the candidate; and a 12×1 one-hot vector representing starting role of the voter (One Night Werewolf has 12 roles in total). Therefore, the input for deduction is a 26×1 vector. We use a simple logistic regression model for deduction modeling. To address the class imbalance of positive and negative samples, we train the model with weighted binary classification loss.

5.2 Experiment Results

Our model achieves an F1 of 32.7% and an AUC of 54.7%, outperforming random prediction, which obtains F1 and AUC of 28.6% and 50.0%, respectively. These results show the effectiveness of persuasion strategy usage and role as predictors for game-level outcomes. To analyze the contribution of persuasion strategy embedding and role embedding, we consider another model that only takes the persuasion strategy embeddings as inputs. This model achieves an F1 of 32.2% and an AUC of 54.6%. Overall we find that the persuasion strategy embedding is more informative for the predicting game outcomes than the role embedding.

We visualize the weights of logistic regression in Fig. 6. Interestingly, for positive prediction (the voter votes for the candidate), the weights of the candidate are higher than the voter. It indicates that a player's voting choice depends more on the candidate's behaviors. This confirms our intuition that players make their decisions based on candidates' arguments.

As for the negative prediction, we see that evidence is the most important strategy for the candi-
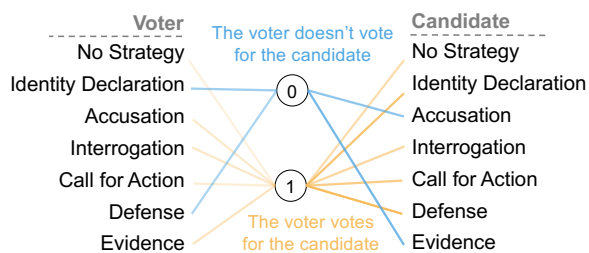
Figure 6: Weights visualization of persuasion strategies in logistic regression. The connection between a strategy and 0 means this strategy contributes to the prediction of 0 (*i.e.* the voter doesn't vote for the candidate). Likewise, the connection between a strategy and 1 denotes this strategy contributes to the prediction of 1 (*i.e.* the voter votes for the candidate). The transparency of lines corresponds to the weights of logistic regression. A less transparent line suggests a greater weight and more impact on the output.

date to negate suspicion. It confirms that players are inclined to trust those who provide more information and evidence to find the werewolf.

6 Conclusion and Future Work

In this work, we introduce the first persuasion modeling dataset with multiple modalities and rich utterance-level persuasion strategy annotations. We design a computational model that leverages both textual and visual representations for understanding persuasion behaviors in social deduction games. Our experiments show that visual cues benefit model performance on persuasion strategy prediction. We encourage future work to explore the role of the audio modality in persuasion modeling and to investigate joint learning of multimodal representations for the social persuasion setting.

Limitations

We only use pre-trained video transformers off-the-shelf to encode the videos, while more nuanced and specific utilization of other models can be explored to further improve the performance. There are also valuable egocentric videos and demographic statistics along with the Ego4D dataset that we have not yet incorporated in our approach. Due to the difficulty and cost of collecting videos with transcriptions and voting outcome annotations, the total number of games is insufficient to train a deep neural network for voting outcome deduction, though data augmentation techniques can be explored to mitigate this limitation.

Ethics Statement

How humans use persuasion strategies in their communication has been long studied in psychology, communication, and NLP (Hovland et al., 1953; Crano and Prislin, 2006; Petty and Cacioppo, 1986; Yang et al., 2019; Wang et al., 2019; Chen and Yang, 2021). We recognize that persuasion skills could be used for good and bad purposes. In this study, our goal is to study persuasive behaviors by multiple speakers through social deduction games. While we recognize that in both games of One Night Ultimate Werewolf and Avalon games players could use persuasion strategies for behaviors that perhaps are considered morally wrong, such as deception, bias, and emotional manipulation, our study does not encourage such behaviors. Instead, we aim to understand people's behavior in a group setting when persuasion happens. Having these persuasion skills could benefit people to perform well in their workplace, such as pitching their ideas, or advocating for peace-making (Simons, 1976).

For our data collection and annotation process, this study has been reviewed and approved by our institution's internal review board. We obtain consent from the players who are recorded and de-identify personally identifiable information (PII), as part of the Ego4D efforts. Moreover, to mitigate potential risks of harmful usage of this dataset in the future, we ask any users to sign an online agreement before using our resources for their research as follows: "*I will not use this dataset for malicious purposes (but not limited to): deception, impersonation, mockery, discrimination, manipulation, targeted harassment, and hate speech.*"

Acknowledgements

References

David Atkinson, Kumar Bhargav Srinivasan, and Chenhao Tan. 2019. What gets echoed? understanding the "pointers" in explanations of persuasive arguments. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2911–2921, Hong Kong, China. Association for Computational Linguistics.

Chongyang Bai, Haipeng Chen, Srijan Kumar, Jure Leskovec, and VS Subrahmanian. 2021. M2p2: Multimodal persuasion prediction using adaptive fusion. *IEEE Transactions on Multimedia*.

Cristian-Paul Bara, Sky CH-Wang, and Joyce Chai. 2021. MindCraft: Theory of mind modeling for situated dialogue in collaborative tasks. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1112–1125, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Xiaoheng Bi and Tetsuro Tanaka. 2016. Human-side strategies in the werewolf game against the stealth werewolf strategy. In *International Conference on Computers and Games*, pages 93–102. Springer.

Mark Braverman, Omid Etesami, and Elchanan Mossel. 2008. Mafia: A theoretical study of players and coalitions in a partial information environment. *The Annals of Applied Probability*, 18(3):825–846.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 621–631, Melbourne, Australia. Association for Computational Linguistics.

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

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 12648–12656.

Gokul Chittaranjan and Hayley Hung. 2010. Are you awerewolf? detecting deceptive roles and outcomes in a conversational role-playing game. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5334–5337. IEEE.

Robert Chuchro. 2022. Training an assassin ai for the resistance: Avalon. *arXiv preprint arXiv:2209.09331*.

William D Crano and Radmila Prislin. 2006. Attitudes and persuasion. *Annual review of psychology*, 57:345.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

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.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. 2021. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6824–6835.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012.

Shirley Anugrah Hayati, Dongyeop Kang, Qingxiaoyang Zhu, Weiyan Shi, and Zhou Yu. 2020. INSPIRED: Toward sociable recommendation dialog systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8142–8152, Online. Association for Computational Linguistics.

Carl Iver Hovland, Irving Lester Janis, and Harold H Kelley. 1953. *Communication and persuasion.* Yale University Press.

Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines*, pages 1846–1870.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+language omni-representation pretraining. In *Conference on Empirical Methods in Natural Language Processing*.

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.

Antonia Lonigro, Roberto Baiocco, Emma Baumgartner, and Fiorenzo Laghi. 2017. Theory of mind, affective empathy, and persuasive strategies in school-aged children. *Infant and Child Development*, 26(6):e2022.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Chujie Lu, Long Chen, Chilie Tan, Xiaolin Li, and Jun Xiao. 2019. Debug: A dense bottom-up grounding approach for natural language video localization. In *Conference on Empirical Methods in Natural Language Processing*.

Kelvin Luu, Chenhao Tan, and Noah A. Smith. 2019. Measuring online debaters' persuasive skill from text over time. *Transactions of the Association for Computational Linguistics*, 7:537–550.

Noritsugu Nakamura, Michimasa Inaba, Kenichi Takahashi, Fujio Toriumi, Hirotaka Osawa, Daisuke Katagami, and Kousuke Shinoda. 2016. Constructing a human-like agent for the werewolf game using a psychological model based multiple perspectives. In *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–8. IEEE.

Richard E Petty and John T Cacioppo. 1986. The elaboration likelihood model of persuasion. In *Communication and persuasion*, pages 1–24. Springer.

David Premack and Guy Woodruff. 1978. Does the chimpanzee have a theory of mind? *Behavioral and brain sciences*, 1(4):515–526.

Jack Serrino, Max Kleiman-Weiner, David C Parkes, and Josh Tenenbaum. 2019. Finding friend and foe in multi-agent games. *Advances in Neural Information Processing Systems*, 32.

Herbert W Simons. 1976. Persuasion. *Reading: Addison-Wesley*, 21.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5635–5649, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. 2021a. Vlm: Task-agnostic video-language model pretraining for video understanding. *arXiv preprint arXiv:2105.09996*.

Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. 2021b. Videoclip: Contrastive pre-training for zero-shot video-text understanding. In *Conference on Empirical Methods in Natural Language Processing*.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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 3620–3630, Minneapolis, Minnesota. Association for Computational Linguistics.

A   Game Rules

One Night Werewolf. In this game, players are divided into two teams – the team of villagers and the team of werewolves. In each game, players close their eyes in the night phase and take some actions (*e.g.* swapping cards) depending on their roles. Players' roles might be changed during the night, but they don't know their new roles except for a few special roles. Then all players open their eyes. The villager team needs to find the werewolf through communication and negotiation. The werewolf team must mislead the others and try to hide their identities. At the end of the game, everyone has to point out the most suspicious player. If at

least one werewolf is voted out, the villager team wins the game. Otherwise, the werewolf team wins. We refer to https://en.wikipedia.org/wiki/Ultimate_Werewolf#One_Night_roles for detailed explanations of game rules and roles.

The Resistance: Avalon. In this game, players are divided into two teams – the team of Minions and the team of Loyal Servants of Arthur. After shuffling and distributing cards to players, they secretly check their role cards and place them face down on the table. Each player will take turns serving as the Leader. In each round, the Leader proposes a Team to do a Quest, and all players are involved in discussing if the Team assignment is passed or rejected. After the Team Building phase, the approved Team will decide if the Quest is successful or not. In the Quest phase, the Good Team can only use the Quest Success card, and the Evil Team can use either Success or Fail card. The Good Team wins when three successful Quests are made, while the Evil Team wins when three failed Quests are made or the Evil players identify Merlin in the Good Team. We refer to https://en.wikipedia.org/wiki/The_Resistance_(game)#Avalon_variant for detailed explanations of game rules and roles.

B Transcription Interface

We provide a screenshot of the transcription tool (*rev.com*) in Fig 7. We upload video clips to this online platform for transcription. We also provide player names and roles involved in each game to make the transcription more accurate. As illustrated in Fig 7, they return the transcription of each utterance, the name of the speaker and the corresponding timestamp. Then we ask annotators to watch videos again and examine the alignment of videos and transcripts. Annotators also correct errors in speakers' names and texts.

C Annotation Interface

We provide a screenshot in Fig. 8 of our interface used by annotators to annotate utterance-level persuasion strategies.

D Details of Video Representation

We now introduce video representation extraction. Given an utterance U_i, We first localize the corresponding video segment using the utterance timestamp t_i, and then approximate the duration of the utterance by $d_i = t_{i+1} - t_i$. We have an average duration of 2 seconds on both the Ego4D and Youtube datasets. To tolerate some misalignment of videos and transcripts, we set a 2-second time window for utterances shorter than 2 seconds, and hence the final duration of an utterance U_i is $d_i' = \max(d_i, 2)$. Then we sample N frames out of the corresponding video segment with equal spacing, *i.e.*, $\mathcal{V} = \{\mathcal{V}_1, \mathcal{V}_2, \ldots, \mathcal{V}_N\}$. All videos in our dataset have an aspect ratio of 16:9. Hence we make three square crops on the left, center, and right of each frame to cover the entire view. Correspondingly, The visual embedding from the vision encoder is composed of three parts, *i.e.*, $\mathcal{V}_i = \{\mathcal{V}_i^{left}, \mathcal{V}_i^{center}, \mathcal{V}_i^{right}\}$. We input the left crops, center crops, and right crops into the video encoder separately and obtain the corresponding representations, *i.e.*, $\psi(\mathcal{V}) = \{\psi(\mathcal{V}^{left}), \psi(\mathcal{V}^{center}), \psi(\mathcal{V}^{right})\}$. The three video representations are flattened as a single vector when we concatenate them with the text representation. In our experiments, we adopt the 24-layer multiscale vision transformer (MViT) (Fan et al., 2021) pretrained on Kinetics-400 as the video encoder. The number of sampled frames N is set as 32 in our experiments. Note that we don't finetune the video encoder on our datasets because the performance of some models drops after finetuning due to overfitting.

In the experiments on Ego4D Werewolf data and Youtube data, the video features improve the performance prominently. However, the video feature does not necessarily help with model generalization. When we apply the model *Base+V* trained on Youtube data to Ego4D data, the average F1 of BERT and RoBERTa drop by 0.8% and 3.0%, respectively, while the F1 of MT-BERT increases by 1.1% after involving video features. This suggests a domain gap exists in the videos of the two datasets which is caused by the differences in the camera locations, angles of views, brightness in the room and etc. Video models are sensitive to these visual differences, resulting in limited performance in generalization on different datasets. In contrast, players communicate in a similar way in different conditions, so the pure text model generalizes better to other data.
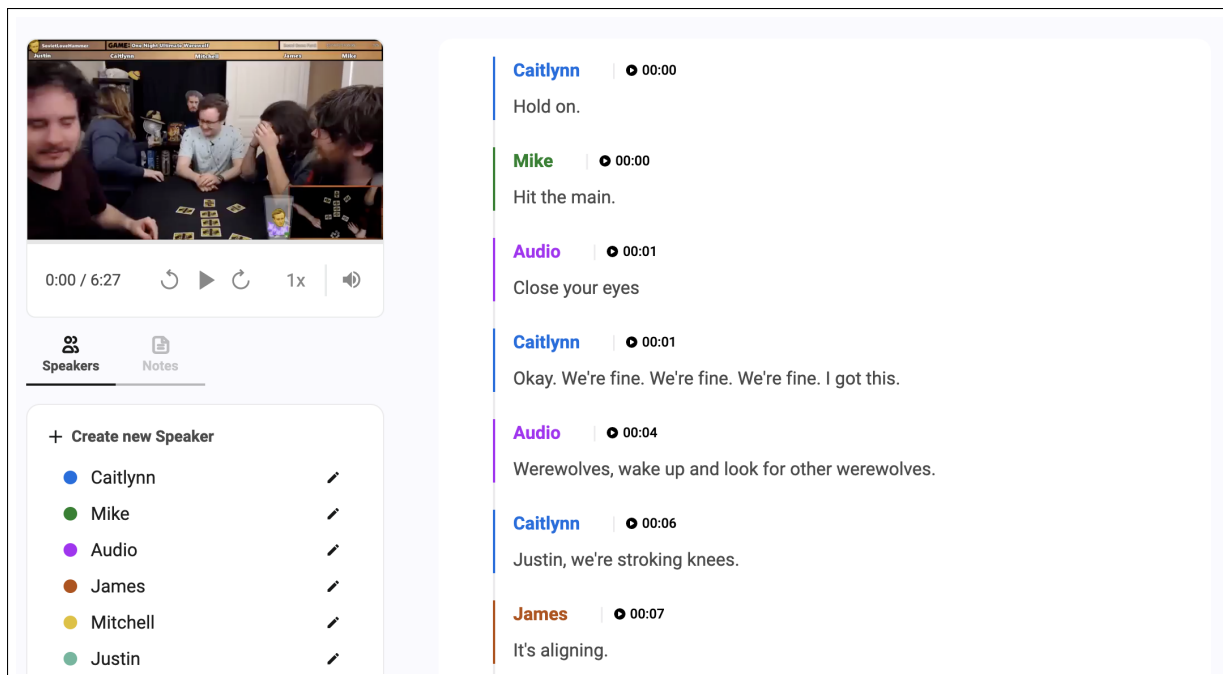
Figure 7: Screenshot of the transcription tool.

E Per-class Results for Experiments with Different Context Lengths

We showcase our experimental results including per-strategy scores on incorporating additional conversational context, in full detail in Table 5.

F Experiments of Domain Generalization

We demonstrate the detailed experiment results of data domain generalization (training models on Youtube data and testing on Ego4D Werewolf test set), as well as game domain generalization (training models on Ego4D Werewolf data and testing on Avalon data). Results are reported in Table 6 and Table 7, respectively.

G Detailed prompt templates used for GPT-3

For the prompt templates, we use the guideline and the persuasion strategy definitions provided to the annotators under the zero-shot setting, and append one/five more examples under one/five-shot setting. The detailed prompt template we used for GPT-3 inference is shown in Table 8.

H Persuasion Strategy Annotation on Avalon Games

Adjacent pie-charts comparing the distributions of annotated utterance-level persuasion strategies for One Night Ultimate Werewolf games and Avalon games in Ego4D are shown in Fig. 9. We can observe a different distribution of adopted persuasion strategies in the two games, suggesting a large game domain gap.

	Method	Identity	Accusation	Interrogation	Call for Action	Defense	Evidence	Avg F1	Joint-A
Ego4D	BERT	**82.6**±1.1	48.8±4.8	82.8±0.2	39.4±9.6	29.3±5.5	54.2±2.5	56.2±2.5	65.1±1.6
	BERT + context1	80.4±0.8	49.0±3.6	82.6±0.4	46.4±8.3	29.1±2.8	53.8±3.6	56.9±0.7	64.0±0.9
	BERT + context3	81.7±1.3	51.8±4.7	81.1±2.6	45.7±4.6	32.5±2.4	52.2±0.9	57.5±1.6	64.8±1.0
	BERT + context5	79.9±1.6	52.0±3.3	81.0±1.1	49.5±3.2	33.8±0.5	57.1±1.6	58.9±0.6	65.0±0.2
	BERT + context7	80.7±3.1	47.4±5.4	80.7±1.7	38.6±12.0	34.7±2.2	55.3±0.7	56.3±2.4	63.5±1.0
	BERT + context9	77.9±0.1	47.5±2.7	78.5±1.9	43.0±5.3	31.8±0.9	54.1±3.5	55.5±0.7	63.7±0.9
	RoBERTa	81.7±2.6	51.7±0.9	83.4±0.9	43.3±8.7	33.1±2.2	51.7±2.1	57.5±1.4	63.4±0.5
	RoBERTa + context1	79.9±2.8	53.1±0.6	82.1±0.9	41.8±7.9	34.1±1.4	55.2±2.9	57.7±1.4	64.1±0.7
	RoBERTa + context3	81.7±1.0	53.9±3.9	82.3±1.6	39.6±9.0	35.4±2.9	54.0±3.8	57.8±2.1	64.1±1.5
	RoBERTa + context5	81.5±0.7	**59.4**±2.4	83.5±1.1	43.7±3.7	33.0±3.1	52.4±2.9	58.9±1.2	64.6±0.7
	RoBERTa + context7	78.6±1.7	55.5±0.5	80.6±0.4	38.2±4.0	30.1±4.9	51.9±3.2	55.8±1.2	62.4±2.3
	RoBERTa + context9	80.2±1.8	56.0±2.2	83.0±1.4	42.5±10.4	32.0±2.4	53.5±1.7	57.9±1.8	63.0±0.6
	MT-BERT	80.9±1.3	51.5±3.3	83.0±1.3	**56.6**±2.3	25.9±2.0	53.6±1.3	58.6±0.3	65.5±0.8
	MT-BERT + context1	79.2±2.2	53.3±2.3	**84.3**±0.6	52.9±2.9	31.1±5.8	55.0±3.2	59.3±0.4	66.0±2.0
	MT-BERT + context3	77.4±2.2	52.6±3.8	83.2±2.1	46.2±2.4	35.1±2.3	56.1±2.7	58.4±0.1	65.1±0.7
	MT-BERT + context5	79.8±2.2	54.4±0.8	83.2±0.7	50.8±7.2	**36.5**±2.8	**61.5**±2.2	**61.0**±1.1	**66.3**±1.4
	MT-BERT + context7	78.5±2.5	54.7±3.3	82.6±1.2	47.9±2.5	33.5±2.2	53.4±1.4	58.4±1.1	65.0±0.9
	MT-BERT + context9	78.2±2.2	54.7±1.6	82.1±0.4	47.8±3.8	30.5±5.8	56.3±1.0	58.3±0.5	64.8±0.8
YouTube	BERT	80.2±1.6	64.7±1.1	89.6±0.4	77.2±2.5	43.5±1.0	58.3±0.7	68.9±0.1	64.6±0.8
	BERT + context1	81.2±1.1	66.5±0.5	90.2±0.3	77.7±0.3	43.6±2.7	59.5±0.6	69.8±0.3	65.1±0.8
	BERT + context3	82.6±0.7	65.9±0.5	90.1±0.7	77.4±1.2	43.0±1.5	60.4±0.9	69.9±0.4	64.4±0.7
	BERT + context5	82.6±0.7	66.7±1.0	89.6±1.5	78.1±2.4	45.7±1.1	59.7±1.1	70.4±0.3	64.4±1.0
	BERT + context7	81.8±0.6	67.2±1.2	90.5±0.2	77.7±0.5	45.0±0.7	60.2±1.2	70.4±0.5	64.8±1.0
	BERT + context9	80.6±1.1	66.7±0.4	90.3±0.2	77.0±1.2	42.2±2.0	59.6±0.2	69.4±0.4	64.0±1.3
	RoBERTa	**84.3**±0.1	67.2±0.6	89.4±0.1	78.2±0.8	44.3±0.4	59.0±1.7	70.4±0.2	64.8±0.7
	RoBERTa + context1	83.3±0.2	67.0±0.3	89.9±0.2	**78.4**±0.9	43.4±2.7	59.7±0.5	70.3±0.5	65.7±1.0
	RoBERTa + context3	82.7±1.5	**67.8**±0.1	90.3±0.4	77.4±0.4	43.1±1.5	**61.0**±1.9	70.4±0.2	65.5±0.4
	RoBERTa + context5	82.4±0.3	67.0±1.1	90.2±0.0	77.1±1.0	46.1±0.7	59.9±0.7	**70.5**±0.3	64.7±0.6
	RoBERTa + context7	83.5±0.8	66.0±0.6	90.2±0.4	77.8±0.3	46.6±1.5	58.4±1.1	70.4±0.3	65.1±1.3
	RoBERTa + context9	82.9±2.0	66.6±0.7	**90.6**±0.2	75.5±0.5	**46.8**±0.9	58.9±1.1	70.2±0.3	64.9±0.9
	MT-BERT	80.7±0.4	65.1±1.5	88.5±0.8	76.2±2.2	42.3±1.5	57.4±1.3	68.4±0.3	65.6±1.1
	MT-BERT + context1	82.9±0.9	67.2±1.2	88.7±1.5	77.8±1.6	43.4±0.6	59.0±0.8	69.8±0.3	67.3±0.9
	MT-BERT + context3	80.5±2.5	65.9±1.5	89.9±0.2	75.2±1.4	44.9±1.9	58.3±0.4	69.1±0.6	65.8±0.9
	MT-BERT + context5	83.1±1.1	65.0±1.5	90.1±0.3	74.6±2.4	46.5±0.8	59.2±0.3	69.7±0.6	66.7±0.5
	MT-BERT + context7	82.1±0.8	67.3±0.1	89.4±1.1	76.0±0.9	43.2±1.3	57.8±0.5	69.3±0.3	**67.6**±0.7
	MT-BERT + context9	81.0±2.0	67.8±0.2	89.6±0.6	72.8±3.8	44.3±2.1	58.8±1.0	69.1±0.8	66.5±0.5

Table 5: Experimental Results on incorporating the conversational context of different lengths for persuasion strategy prediction

	Method	Identity	Accusation	Interrogation	Call for Action	Defense	Evidence	Avg F1	Joint-A
w.o. Fine Tuning	BERT	82.0±1.2	53.9±1.6	84.1±0.9	53.0±4.1	33.9±0.5	53.5±3.9	60.1±0.7	65.6±1.0
	BERT + C	83.6±0.8	55.7±1.1	85.6±1.0	46.5±2.7	34.5±2.5	60.9±3.2	61.1±0.9	65.6±1.3
	RoBERTa	86.9±0.9	57.0±1.4	85.0±2.0	53.5±3.6	31.5±1.3	55.0±1.3	61.5±0.8	66.0±0.7
	RoBERTa + C	82.5±2.4	56.3±2.7	86.2±0.6	50.7±4.1	37.6±1.8	59.6±1.8	62.2±1.3	67.6±0.6
	MT-BERT	80.6±2.9	50.4±3.6	83.5±1.7	45.3±10.6	34.7±0.7	55.2±3.2	58.3±2.7	64.8±2.1
	MT-BERT + C	81.8±2.1	53.8±2.7	83.4±1.9	44.1±8.2	35.9±1.7	53.5±2.1	58.7±2.3	66.3±2.0
w. Fine Tuning	BERT	82.0±1.4	54.9±0.9	82.8±0.6	53.0±1.9	29.9±1.0	61.7±0.7	60.7±0.2	68.1±0.2
	BERT + C	84.1±0.1	55.6±3.5	86.1±0.4	49.9±2.2	32.6±1.5	61.5±2.4	61.6±1.1	69.0±0.8
	RoBERTa	86.7±1.3	56.6±1.4	85.3±1.6	54.8±3.9	29.4±2.5	57.3±2.0	61.7±1.4	67.4±0.9
	RoBERTa + C	84.0±1.5	58.9±1.6	84.9±0.2	52.4±4.4	38.0±2.3	62.5±2.4	63.4±1.7	69.1±0.6
	MT-BERT	81.9±1.4	54.7±1.6	83.0±0.6	60.2±5.3	25.6±1.6	59.2±2.3	60.8±1.0	68.5±0.6
	MT-BERT + C	83.7±0.9	54.3±2.4	84.5±1.1	53.8±3.2	33.8±3.4	58.1±2.8	61.4±1.0	70.0±1.5

Table 6: Data domain generalization experiments. We train models on Youtube data and test on Ego4D testing set. Then we fine-tune the models on Ego4D training set and test again.

Label the Persuasion Strategy for the Dialogues during Social Deduction Game

Read the instruction given below carefully:
Your goal is to read the given dialog during the social deduction game. And then label the **strategy** someone used to persuade others by clicking the correponding check box. A sentence is labeled as **No Strategy** if it is not a strategy. We have 9 labels (including No Strategy).

Strategy Definition and Examples:

1. Identity Declaration

State one's own role or identity in the game
Example:
"I was the seer, and I know you were the robber."
"I'll just come out and say I was a villager, so I have no idea what's going on. "
"I was the robber, but I have no idead after the swap. "

2. Accusation

Claim someone has a specific identity or strategic behavior.
Example:
"I think you are lying."
"So James might be the werewolf."
"I think Sukeshi is the werewolf she is just too quiet."

3. Interrogation

Questions about someone's identity or behavior.
Example:
"Who did you rob?"
"Was Elliot actually a villager?"
"So, who does everyone hate the most?"

4. Call for Action

Encourage people to take an action during the game.
Example:
"Let's kill Sukeshi."
"We shouldn't vote to not kill anyone. And then there could also be no werewolf."
"But you never... You gotta listen to me."

5. Defense

Defending yourself or someone else against an accusation or defending a game-related argument.
Example:
"I'm not the werewolf."
"I didn't do shxt."
"I think that you accused me of being a Werewolf very quickly."

6. Evidence

Provide a body of game-related fact or information.
Example:
"I heard movement over here."
"If you swapped these two, he is not the werewolf."
"Well, because we did this and because people were moving cards, I started listening to that, so I heard a lack of movement from this part of the table."

7. No Strategy

Any sentences that do not fall into other categories are here. Clarification or discussion of game rules should also be considered "No-Strategy".
Example:
"All right."
"What does Tanner do?"
"Okay. Good point."

Please **carefully read the input text** first. Then, click the **appropriate strategy** of button for each sentence.
Once you have finished the annotation, click **Yes** to confirm you have labeled every statement. Then you can click the [Submit] button at the bottom to end the task. If the annotation file could not be downloaded, this means at least one of the entry has not been annotated. You can click on the last **Save Progress** button to find the entries that have not been annotated.

Step 1: Please enter your first name (required) and last name (optional) in the text boxes below:

First Name:
| Aryan |

Step 2: Upload the appropriately formatted text file (.txt) of your desired transcript here:

[Choose File] player2_Game9_1.txt

Step 3: (Optional) If you saved your annotation progress specific to the uploaded transcript file above in your last session and wish to continue from where you left off, please upload the annotations csv (.csv) file that you had downloaded and saved in your last session here:

[Choose File] No file chosen

Step 4: Choose appropriate persuasion strategy(ies) or "No Strategy" for the utterances in the dialogs below.

Turn	Utterance	Label	Notes	Save Progress?
1	Kevin (00:43): Sounds a little bit like Muse.	Identity Declaration Call for Action Evidence Accusation Defense No Strategy Interrogation	**Add Notes:**	Save Progress

Figure 8: The screenshot of persuasion strategy annotation interface

6584

Method	Identity	Accusation	Interrogation	Call for Action	Defense	Evidence	Avg F1	Joint-A
BERT	66.1±3.7	37.2±4.9	73.3±3.7	20.9±7.3	23.5±6.3	26.6±5.3	41.3±3.3	62.9±0.9
BERT + C	60.4±4.5	41.7±2.2	73.2±0.3	34.4±4.8	25.0±4.3	18.2±2.9	42.1±1.4	63.9±0.6
RoBERTa	63.0±3.2	45.9±3.0	73.1±2.8	33.0±12.1	33.6±5.7	27.5±1.2	46.0±1.8	64.3±0.5
RoBERTa + C	46.1±9.3	40.3±2.3	71.8±2.9	35.8±2.8	28.1±3.0	22.5±8.5	40.8±1.7	63.5±0.6
MT-BERT	56.3±3.6	38.2±4.6	73.1±1.9	39.3±5.7	23.5±8.0	25.3±1.0	42.6±1.7	63.4±0.2
MT-BERT + C	61.9±2.9	42.2±2.5	71.2±2.7	34.2±1.0	29.3±5.2	23.2±3.7	43.7±1.2	64.2±0.3

Table 7: Game domain generalization experiments. We train models on Ego4D Werewolf training set and test them on Avalon data.
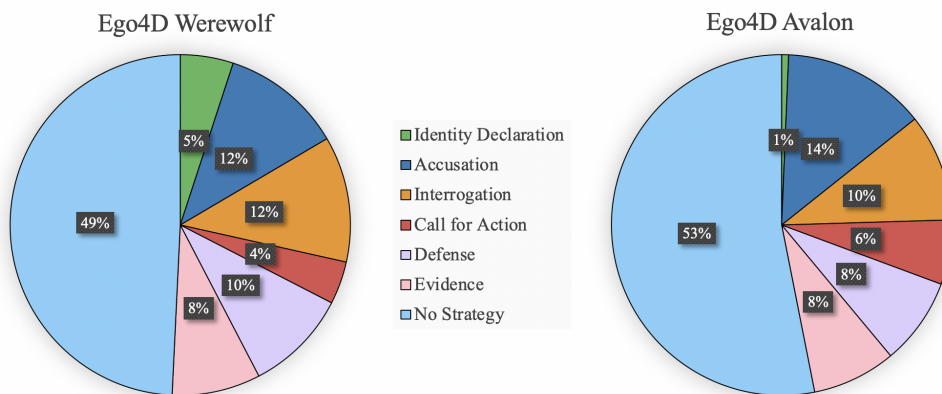

Figure 9: Persuasion strategy distributions for Ego4D Werewolf games and Ego4D Avalon games: Pie charts representing the percentage of annotated utterances for every persuasion strategy and "No Strategy" in One Night Ultimate Werewolf games and The Resistance: Avalon games in Ego4D.

zero-shot	Label the Persuasion Strategy for the Utterances in Dialogues during Social Deduction Game. Do not hesitate to select multiple strategies if one category can not summarize the given utterance.
	Strategy Definition:
	1. Identity Declaration: State one's own role or identity in the game 2. Accusation: Claim someone has a specific identity or strategic behavior 3. Interrogation: Questions about someone's identity or behavior 4. Call for Action: Encourage people to take an action during the game 5. Defense: Defending yourself or someone else against an accusation or defending a game-related argument 6. Evidence: Provide a body of game-related facts or information 7. No Strategy: Any sentences that do not fall into other categories are here. Clarification or discussion of game rules should also be considered "No-Strategy"
	Utterance: "$utterance$" Strategy:
one-shot	Label the Persuasion Strategy for the Utterances in Dialogues during Social Deduction Game. Do not hesitate to select multiple strategies if one category can not summarize the given utterance.
	Strategy Definition: *[same as above]*
	Utterance: "No, but in order to find it, I had to really tap around to find it." Strategy: Defense, Evidence
	Utterance: "$utterance$" Strategy:
five-shot	Label the Persuasion Strategy for the Utterances in Dialogues during Social Deduction Game. Do not hesitate to select multiple strategies if one category can not summarize the given utterance.
	Strategy Definition: *[same as above]*
	Utterance: "I'll just come out and say I was a villager, so I have no idea what's going on." Strategy: Identity Declaration
	Utterance: "So James might be the werewolf." Strategy: Accusation
	Utterance: "Did anybody do any swapping? Anybody willing to fess up to anything about swapping?" Strategy: Interrogation, Call for Action
	Utterance: "No, but in order to find it, I had to really tap around to find it." Strategy: Defense, Evidence
	Utterance: "Okay. Good point." Strategy: No Strategy
	Utterance: "$utterance$" Strategy:

Table 8: Prompt templates used for GPT-3, the variable within dollars is to be replaced with the corresponding value.

ACL 2023 Responsible NLP Checklist

A For every submission:

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

☑ A2. Did you discuss any potential risks of your work?
Ethics Statement

☑ A3. Do the abstract and introduction summarize the paper's main claims?
1

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

B ☑ Did you use or create scientific artifacts?

3; 4; Appendix C

☑ B1. Did you cite the creators of artifacts you used?
3.1; 4.2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
3.1; 4.2

☑ 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)?
3.1; 4.2

☑ 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?
3

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
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.
3; 5

C ☑ Did you run computational experiments?

4; 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?
4.2; 4.3

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?
4.4; 5.2; Appendix D E F

☑ 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?
4.4

☑ 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.)?
4.2; 4.3

D ☑ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**
3.2

☑ 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.?
Appendix B 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)?
3.2

☑ 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?
3

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
3.1

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
3.2