CanLanguageModels_Analysis

Definitions

- "parties are focused on a perceived injustice by the other party and the potential costs of ending an existing relationship" Page #definition

Applications

Technical Details

- "Real-world corpora involve rich and authentic conversations, but one rarely has access to each party's true underlying goals and emotions, making it difficult to quantify success or failure objectively" Page #dispute-theory
  - Benefits/drawbacks of observed disputes #dispute-theory
- " "humanmediator followed instructions (i.e., they role-played as a mediator)" " Page 6

Conceptual

- "anger in disputes provokes retaliation [21]" Page #dispute-theory , #evidence
- "here 71 participants selected the GPT-generated mediation compared to 35 for the human-constructed one" Page 8 #critique
  - Is it possible GPT did seem more profesisonal? #critique

Personal Insights

- "However, many of the disputes escalated and ended without agreement, thus forgoing their bonus. Even when agreements were reached, disputants often reported high frustration with their partner." Page 2 #question
  - Why did many disputes lead to conflict? #question
- "(1) whether or not the dispute ended in success or failure, and (2) whether or not the participants reported frustration with each other." Page 2 #question
  - How does frustration get mapped to anger? In the other paper, we have a 10-item tactics survey, but no label on GPT model for frustration #question
- "We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2 #question , #critique , #potential-gap

    - What kind of person typically needs to intervene in a dispute? Could the LLM have better suggestions because it does have knowledge of mediatiors?

If we changed the scenario, how would that affect preference of raters? #question, #critique, #potential-gap

- "importance to these issues via a payoff matrix, participants are free to assign their own importance to each issue." Page 3 #question
    - why move away from payoff matrix as in initial KODIS paper on the cross-cultural differences #question
- "ing LLMs and explore several prompt" Page 3 #question, #Further-exploration-needed
    - for what kinds of disputes? #question, #Further-exploration-needed
- "Fig. 3. GPT's selected reasons for intervention as a proportion of exchanges" Page 5 #Question
    - Does this also reflect the portion when ChatGPT interverena each time step? What is the 1-6 scale? #Question
- "Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 #question, #critique, #data-collection
    - But how do we define how a human interperets "exchange level"? #question, #critique, #data-collection
- "he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6 #question, #data-analysis
    - in what cases did they not? #question, #data-analysis
- "In this new task, we recruit N = 106 participants so that each mediation receives at least five annotations." Page 7 #question, #sampling
    - Any issues with sampling again and outcome qualioty/statistical effects? Is there more subjectivity now in the 20- to evaluate the justification? #question, #sampling
- "Concretely, we ask to what extent the participant disagrees or agrees (1-10) with three statements depicted in Table 2." Page 7 #question, #experiment
    - Why a 10 point scale for agreeability? Also, what were the comstrains for reason to intervene on giving the reaso? DO we need to include any ethical considerdations in the kinds of repsonses? #question, #experiment
- " "may infringe on one's freedom of speech if they shut down the conversation or intervene based on a prediction of future bad behavior. Further, prior work demonstrates LLMs struggle to generalize across cultures, specifically with emotion [16]" " Page 8

Literary Note To Lookup Later

- "come in the form of endowing the LLM's prompt with psychology-based negotiation strategies, as similar prior work exists with rule-based agents [19, 25]." Page 8 #potential-

gap , #question , #to-read

- is this for post dispute or during? #potential-gap , #question , #to-read

---

🧠 Key Ideas

-

⚙️ Project Relations

-

📄 Related Literature

-

🏷️ Tags

Unique tag Groups

#definition
#dispute-theory , #evidence
#dispute-theory
#question
#question , #critique , #potential-gap
#question , #Further-exploration-needed
#question , #data-collection
#critique , #data-collection
#emotion-recognition , #question , #experiment , #potential-gap , #LLM
#Further-exploration-needed , #experiment , #LLM
#data-analysis
#Question
#question , #experiment

#question , #experiment , #potential-gap
#question , #critique , #data-collection
#question , #data-analysis
#question , #sampling
#experiment , #question
#critique
#potential-gap , #question , #to-read

Tags Groups within Current Paper

#definition

Paper (1)	Related Annotations
CanLanguageModels_Analysis	"parties are focused on a perceived injustice by the other party and the potential costs of ending an existing relationship" Page

#dispute-theory , #evidence

Paper (1)	Related Annotations
CanLanguageModels_Analysis	"anger in disputes provokes retaliation [21]" Page

#dispute-theory

Paper (3)	Related Annotations
CanLanguageModels_Analysis	"Real-world corpora involve rich and authentic conversations, but one rarely has access to each party's true underlying goals and emotions, making it difficult to quantify success or failure objectively" Page
CanLanguageModels_Analysis	Benefits/drawbacks of observed disputes
CanLanguageModels_Analysis	"anger in disputes provokes retaliation [21]" Page , #evidence

#question

Paper (19)	Related Annotations
CanLanguageModels_Analysis	"However, many of the disputes escalated and ended without agreement, thus forgoing their bonus. Even when agreements were reached, disputants often reported high frustration with their partner." Page 2
CanLanguageModels_Analysis	Why did many disputes lead to conflict?
CanLanguageModels_Analysis	"(1) whether or not the dispute ended in success or failure, and (2) whether or not the participants reported frustration with each other." Page 2
CanLanguageModels_Analysis	How does frustration get mapped to anger? In the other paper, we have a 10-item tactics survey, but no label on GPT model for frustration
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2 , #critique , #potential-gap
CanLanguageModels_Analysis	"importance to these issues via a payoff matrix, participants are free to assign their own importance to each issue." Page 3
CanLanguageModels_Analysis	why move away from payoff matrix as in initial KODIS paper on the cross-cultural differences
CanLanguageModels_Analysis	"ing LLMs and explore several prompt" Page 3 , #Further-exploration-needed
CanLanguageModels_Analysis	for what kinds of disputes? , #Further-exploration-needed
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 , #critique , #data-collection
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? , #critique , #data-collection
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6 , #data-analysis
CanLanguageModels_Analysis	in what cases did they not? , #data-analysis
CanLanguageModels_Analysis	"In this new task, we recruit N = 106 participants so that each mediation receives at least five annotations." Page 7 , #sampling
CanLanguageModels_Analysis	Any issues with sampling again and outcome qualioty/statistical effects? Is there more subjectivity now in the 20- to evaluate the justification? , #sampling

Paper (19)	Related Annotations
CanLanguageModels_Analysis	"Concretely, we ask to what extent the participant disagrees or agrees (1-10) with three statements depicted in Table 2." Page 7 , #experiment
CanLanguageModels_Analysis	Why a 10 point scale for agreeability? Also, what were the comstrains for reason to intervene on giving the reaso? DO we need to include any ethical considerdations in the kinds of repsonses? , #experiment
CanLanguageModels_Analysis	"come in the form of endowing the LLM's prompt with psychology-based negotiation strategies, as similar prior work exists with rule-based agents [19, 25]." Page 8 #potential-gap , , #to-read
CanLanguageModels_Analysis	is this for post dispute or during? #potential-gap , , #to-read

#question , #critique , #potential-gap

Paper (1)	Related Annotations
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2

#question , #Further-exploration-needed

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"ing LLMs and explore several prompt" Page 3
CanLanguageModels_Analysis	for what kinds of disputes?

#question , #data-collection

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#critique , #data-collection

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 #question ,
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? #question ,

#emotion-recognition , #question , #experiment , #potential-gap , #LLM

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#Further-exploration-needed , #experiment , #LLM

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#data-analysis

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6 #question ,
CanLanguageModels_Analysis	in what cases did they not? #question ,

#Question

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Fig. 3. GPT's selected reasons for intervention as a proportion of exchanges" Page 5
CanLanguageModels_Analysis	Does this also reflect the portion when ChatGPT interverena each time step? What is the 1-6 scale?

#question , #experiment

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Concretely, we ask to what extent the participant disagrees or agrees (1-10) with three statements depicted in Table 2." Page 7
CanLanguageModels_Analysis	Why a 10 point scale for agreeability? Also, what were the comstrains for reason to intervene on giving the reaso? DO we need to include any ethical considerdations in the kinds of repsonses?

#question , #experiment , #potential-gap

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#question , #critique , #data-collection

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"?

#question , #data-analysis

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6
CanLanguageModels_Analysis	in what cases did they not?

#question , #sampling

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"In this new task, we recruit N = 106 participants so that each mediation receives at least five annotations." Page 7
CanLanguageModels_Analysis	Any issues with sampling again and outcome qualioty/statistical effects? Is there more subjectivity now in the 20- to evaluate the justification?

#experiment , #question

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#critique

Paper (5)	Related Annotations
CanLanguageModels_Analysis	"here 71 participants selected the GPT-generated mediation compared to 35 for the human-constructed one" Page 8
CanLanguageModels_Analysis	Is it possible GPT did seem more profesisonal?
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2 #question , , #potential-gap
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 #question , , #data-collection
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? #question , , #data-collection

#potential-gap , #question , #to-read

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"come in the form of endowing the LLM's prompt with psychology-based negotiation strategies, as similar prior work exists with rule-based agents [19, 25]." Page 8
CanLanguageModels_Analysis	is this for post dispute or during?

Tags Groups within All Papers

#definition

Paper (76)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"Disputes arise when one party in a relationship (an individual, group, or nation) claims that another party refuses to accept, thus threatening the future of the relationship [25]" Page , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	"egotiation (or Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), A. El Fallah Seghrouchni, Y. Vorobeychik, S. Das, A. Nowe (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (**www.ifaamas.org**). This work is licenced under the Creative Commons Attribution 4.0 International (CC-BY 4.0) licence. deal-making) involves coming together to create a new relationship (i.e., focus on opportunities for gains), and when parties can't reach a deal, they can seek other partners" Page , #negotiation
EmotionallyAwareAgentsDispute_Analysis	dispute resolution is NOT deal-making. I think deal-making is the same as negotiation (in the context of this paper). Disputes have invariant partners with the relationship quality as the outcome, whereas negotiation is focused on the mutual objective, with relationships possibly variant. , #negotiation

Paper (76)	Related Annotations
CanLanguageModels_Analysis	"parties are focused on a perceived injustice by the other party and the potential costs of ending an existing relationship" Page
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"Algorithm aversion–the tendency to ignore tool recommendations after seeing that they can be erroneous– originates from a lack of agency [29, 12] and lack of transparency of the algorithm [49]" Page 2 #algorithm-aversion ,
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	define lack of agency of the tool #algorithm-aversion ,
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"automation bias, on the other hand, will follow tool recommendations despite available (but unnoticed or unconsidered) information that would indicate that the recommendation is wrong" Page 2
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"Commission errors refer to instances where humans take action on the basis of an erroneous algorithmic recommendation, failing to incorporate contradictory external information into the decision process" Page 2
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	algorithm is wrong, human doesn't notice
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"adherence is taken to be synonymous with trust. Indeed, while there is no single commonly adopted definition of trust in the HCI literature, the term trust typically refers to a measure of, or the factor influencing, the degree to which the human is willing to delegate

Paper (76)	Related Annotations
	decision-making to the machine in absence of complete knowledge of the algorithmic pipeline [28, 51, 3, 37, 50]" Page 3 #adherence ,
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"Out-of-home placement refers to whether the child is placed out of the home following an investigation, and a future referral refers to a future call involving the child coming in to the hotline." Page 4 #ASFT ,
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	model predictions #ASFT ,
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"commit errors of omission—screening out shown low-risk cases that are assessed as high(er) risk" Page 6
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	failed to detect by human judgement where (model error) but needed higher vigilance
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	"hown high-risk cases that are assessed as lower risk" Page 7
de-arteagaCaseHumansintheLoopDecisions2020_Analysis	agree with model error, but actually not bad-> more consequences on child
gratchFieldAffectiveComputing_Analysis	"emotions" Page 2 #supportingevidence ,
gratchFieldAffectiveComputing_Analysis	"moods" Page 2
gratchFieldAffectiveComputing_Analysis	"interpersonal stances" Page 2 , #main-idea
gratchFieldAffectiveComputing_Analysis	"affective dispositions" Page 2
hardtEqualityOpportunitySupervised2016_Analysis	"oblivious: it depends only on the joint statistics of the predictor, the target and the protected attribute, but not on interpretation of individual features." Page
hardtEqualityOpportunitySupervised2016_Analysis	"protected attributes" Page

Paper (76)	Related Annotations
hardtEqualityOpportunitySupervised2016_Analysis	"Demographic parity requires that a decision—such as accepting or denying a loan application—be independent of the protected attribute." Page
hardtEqualityOpportunitySupervised2016_Analysis	"“oblivious”, in that it is based only on the joint distribution, or joint statistics, of the true target Y , the predictions Ŷ , and the protected attribute A." Page 2 , #theoretical-framework , #fairness
hardtEqualityOpportunitySupervised2016_Analysis	"Ŷ satisfies equalized odds with respect to protected attribute A and outcome Y , if Ŷ and A are independent conditional on Y ." Page 3
hardtEqualityOpportunitySupervised2016_Analysis	Equalized Odds
hardtEqualityOpportunitySupervised2016_Analysis	"property of a predictor Ŷ or score R is said to be oblivious if it only depends on the joint distribution of (Y , A, Ŷ) or (Y , A, R)" Page 4 , #fairness
hardtEqualityOpportunitySupervised2016_Analysis	"A predictor $\tilde{Y}$ is derived from a random variable R and the protected attribute A if it is a possibly randomized function of the random variables (R, A) alone. In particular, $\tilde{Y}$ is independent of X conditional on (R, A)." Page 5 , #pre
hardtEqualityOpportunitySupervised2016_Analysis	derived predictor , #pre
hardtEqualityOpportunitySupervised2016_Analysis	"predictor can be obtained as an derived threshold predictor" Page 10
hardtEqualityOpportunitySupervised2016_Analysis	a predictor is a classifier based on the thresholded R
hardtEqualityOpportunitySupervised2016_Analysis	"(Identical ROC Curves). We say that a score R has identical conditional ROC curves if Ca(t) = Ca′ (t) for all groups of a, a′ and all t ∈ R." Page 14

Paper (76)	Related Annotations
rudinInterpretableMachineLearning2022	"case-based reasoning" Page #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	define and find examples #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"disentanglement of neural networks" Page #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	define disentanglement #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"generative or causal constraints" Page #actionitem ,
rudinInterpretableMachineLearning2022	find examples and define #actionitem ,
rudinInterpretableMachineLearning2022	""Rashomon set"" Page #actionitem , #important ,
rudinInterpretableMachineLearning2022	Define #actionitem , #important ,
rudinInterpretableMachineLearning2022	"Interpretable predictive models, which are constrained so that their reasoning processes are more understandable to humans" Page 2 #Interpretable-machine-learning , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	high level defintion of interpretability the paper uses as basis for defintion. Could refine further #Interpretable-machine-learning , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	""Clever Hans" phenomenon" Page 2 #actionitem , #important , #blackbox ,
rudinInterpretableMachineLearning2022	Need to define in relation to impact of using black box models in practice #actionitem , #important , #blackbox ,

Paper (76)	Related Annotations
rudinInterpretableMachineLearning2022	"underlying distribution of data changes (called domain shift" Page 2 #Interpretable-machine-learning , #actionitem ,
rudinInterpretableMachineLearning2022	define domain shift, and whether thee are generalize frameworks for how this arises in 2-3 key domains were interperetability has a huge effect. #Interpretable-machine-learning , #actionitem ,
rudinInterpretableMachineLearning2022	"Clean" means that the data do not have too much noise or systematic bias" Page 4
rudinInterpretableMachineLearning2022	"Tabular" means that the features are categorical or real," Page 4
rudinInterpretableMachineLearning2022	"Raw" data is unprocessed and has a complex data type" Page 4
rudinInterpretableMachineLearning2022	"soft and hard interpretability constraints" Page 4 #Interpretable-machine-learning , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	define differences #Interpretable-machine-learning , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"This definition might require refinement, sometimes over multiple iterations with domain experts. There are many papers detailing these issues, the earliest dating from the mid-1990s [e.g., 157]." Page 5 #Interpretable-machine-learning , #important , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	What drives the definition of interpertabilitiy from a domain level and from the CS community? #Interpretable-machine-learning , #important , #Further-exploration-needed ,

Paper (76)	Related Annotations
rudinInterpretableMachineLearning2022	"Interpretability penalties or constraints can include sparsity of the model, monotonicity with respect to a variable, decomposibility into sub-models, an ability to perform case-based reasoning or other types of visual comparisons, disentanglement of certain types of information within the model's reasoning process, generative constraints (e.g., laws of physics), preferences among the choice of variables, or any other type of constraint that is relevant to the domain" Page 5 #Interpretable-machine-learning , #important , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	Need to understand why knowledge of these penalizes the outcome from (*) #Interpretable-machine-learning , #important , #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"coring systems are linear classification models that require users to add, subtract, and multiply only a few small numbers in order to make a prediction." Page 16 , #logical-model
rudinInterpretableMachineLearning2022	"Case-based reasoning is a paradigm that involves solving a new problem using known solutions to similar past problems [1]. It is a problem-solving strategy that we humans use naturally in our decision-making processes [219]" Page 25 , #case-based-reasoning
rudinInterpretableMachineLearning2022	"Nearest neighbor-based techniques. These techniques make a decision for a previously unseen test instance, by finding

Paper (76)	Related Annotations
	training instances that most closely resemble the particular test instance" Page 26 , #case-based-reasoning
rudinInterpretableMachineLearning2022	"Given a previously unseen test instance, they make a decision by finding prototypical cases (instead of training instances from the entire training set) that most closely resemble the particular test instance" Page 27 , #case-based-reasoning
rudinInterpretableMachineLearning2022	"Disentanglement" here refers to the way information travels through the network: we would perhaps prefer that all information about a specific concept (say "lamps") traverse through one part of the network while information about another concept (e.g., "airplane") traverse through a separate part." Page 31 , #disentanglement
rudinInterpretableMachineLearning2022	"n general, a PINN is a neural network that approximates the solution of a set of PDEs with initial and boundary conditions. The training of a PINN minimizes the residuals from the PDEs as well as the residuals from the initial and boundary conditions." Page 46 , #physics-machine-learning
rudinInterpretableMachineLearning2022	constraints based on laws of nature, PDE/ODES for material science, etc. , #physics-machine-learning
rudinInterpretableMachineLearning2022	"The Rashomon effect occurs when there are multiple descriptions of the same event [41] with possibly no ground truth" Page 48 , #Rashomon-Set

Paper (76)	Related Annotations
rudinInterpretableMachineLearning2022	"computing statistics of the Rashomon set in parameter space, such as the volume in parameter space, which is called the Rashomon volume" Page 51 , #Rashomon-Set
rudinInterpretableMachineLearning2022	"Unfortunately, these topics are much too often lumped together within the misleading term “explainable artificial intelligence” or “XAI” despite a chasm separating these two concepts [250]" Page 9 #question , #Further-exploration-needed , , #To-read
haleKODIS_NAACL_Annotatedpdf_Analysis	"Conflict dialogues are task036 oriented non-collaborative conversations" Page , #conflict
haleKODIS_NAACL_Annotatedpdf_Analysis	Definition of a conflict dialogue. there is a set of goals for both parties as well as competing interests. , #conflict
haleKODIS_NAACL_Annotatedpdf_Analysis	"disputes are 071 backward-looking, typically involving an existing 072 relationship that has gone badly." Page , #conflict
haleKODIS_NAACL_Annotatedpdf_Analysis	distinction between deal-making and disputes. Deal making is opportunistic benefits of the relationship while disputes are focused on the cost of ending the relationship. More emotionally driven, increases the effect of influence methods. , #conflict
haleKODIS_NAACL_Annotatedpdf_Analysis	"Dignity relates to the 350 Western ideal that each individual has intrinsic self351 worth and thus tends to be impervious to threats 352 from others." Page 5 , #multicultural
haleKODIS_NAACL_Annotatedpdf_Analysis	"Honor cultures tend to view self-worth 353 as something that

Paper (76)	Related Annotations
	must be claimed and defended 354 from external threats." Page 5 , #multicultural
haleKODIS_NAACL_Annotatedpdf_Analysis	"Face cultures also see self355 worth as conferred by others but see retaliation as 356 further eroding self-worth." Page 5 , #multicultural
changTroubleHorizonForecasting2019_Analysis	"conversational forecasting, which includes future-prediction tasks such as predicting the eventual length of a conversation" Page 2 , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"Antisocial behavior online comes in many forms, including harassment (Vitak et al., 2017), cyberbullying (Singh et al., 2017), and general aggression (Kayany, 1998)." Page 3
changTroubleHorizonForecasting2019_Analysis	antisocial behavior
changTroubleHorizonForecasting2019_Analysis	"This process, known as fine-tuning, reshapes the representation learned during pre-training to be more directly useful to prediction (Howard and Ruder, 2018)." Page 6 , #model-architecture , #to-learn

#dispute-theory , #evidence

Paper (1)	Related Annotations
CanLanguageModels_Analysis	"anger in disputes provokes retaliation [21]" Page

#dispute-theory

Paper (51)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"Disputes arise when one party in a relationship (an individual, group, or nation) claims that another party refuses to accept, thus threatening the future of the relationship [25]" Page #definition ,
EmotionallyAwareAgentsDispute_Analysis	"disputes evoke much stronger emotions than negotiations, particularly anger." Page
EmotionallyAwareAgentsDispute_Analysis	"anger provokes retaliation [48]" Page
EmotionallyAwareAgentsDispute_Analysis	"dispute literature shows how specific emotions convey specific intentions and examines how these conveyed intentions change within a text. This is challenging as interpreting each utterance depends on the context of prior utterances [47" Page #emotion-recognition , #contribution ,
EmotionallyAwareAgentsDispute_Analysis	Lead with emotional mapping to dispute intentions based on theory to analyze context in unseen disputes to forecast intention meaning. #emotion-recognition , #contribution ,
EmotionallyAwareAgentsDispute_Analysis	"To foreshadow our findings, we find that automatically recognized emotional expressions explain up to 45% of the variance in dispute outcomes (compared with 5% in prior negotiation research)" Page , #result
EmotionallyAwareAgentsDispute_Analysis	"Whereas social science findings have emphasized the role of anger in disputes, our findings highlight the importance of other emotional expressions, such as compassion." Page 2 #emotion-recognition , #contribution ,
EmotionallyAwareAgentsDispute_Analysis	"LMs perform substantially better in predicting dispute outcomes than previous text emotion recognition methods (highlighting the impact of several prompting strategies)" Page 2 #emotion-recognition , #contribution ,
EmotionallyAwareAgentsDispute_Analysis	"results replicate prior social science findings on the role of anger in escalation

Paper (51)	Related Annotations
	and lay a foundation for autonomous agents that could detect escalation and intervene before a dispute ends in an impasse." Page 2 #emotion-recognition , #contribution ,
EmotionallyAwareAgentsDispute_Analysis	"parties can often find better solutions if they exchange information about each other's interests and find solutions that maximize joint gains [5, 56]." Page 2
EmotionallyAwareAgentsDispute_Analysis	"Neither line of work considers direct interactions between people nor do they consider the role of emotional expressions." Page 2 #critique ,
EmotionallyAwareAgentsDispute_Analysis	in relation to mediation and argument-based agent-agent negotiation #critique ,
EmotionallyAwareAgentsDispute_Analysis	"Most work uses simple dictionary-based approaches [40], though more recent work takes advantage of powerful transformer-based models [12]." Page 2 #emotion-recognition , #to-read ,
EmotionallyAwareAgentsDispute_Analysis	"hough we focused on dispute resolution, our findings could inform techniques that intentionally escalate conflicts." Page 8 , #potential-gap
EmotionallyAwareAgentsDispute_Analysis	Can we apply our results to other dispute types? , #potential-gap
EmotionallyAwareAgentsDispute_Analysis	"As parties are already linked," Page #question ,
EmotionallyAwareAgentsDispute_Analysis	What defines a link--furthermore, the type of the relationship? Historic agreement? Self-interests? #question ,
EmotionallyAwareAgentsDispute_Analysis	"Prior research on emotions in text focuses on a document as the unit of analysis and outputs coarse-grained "sentiment" rather than specific emotions like anger [59]." Page #question ,
EmotionallyAwareAgentsDispute_Analysis	Difference between intention and sentiment? #question ,
EmotionallyAwareAgentsDispute_Analysis	"figure climbs to 33% in the dialogues in the top quartile of self-reported

Paper (51)	Related Annotations
	frustration and drops to only 6% in the dialogues with the lowest." Page 4 , #result , #potential-gap
EmotionallyAwareAgentsDispute_Analysis	"This reinforces findings in the social sciences on how anger shapes conflict [30, 48]," Page 8 #question , #Further-exploration-needed ,
EmotionallyAwareAgentsDispute_Analysis	What is "new" result versus not something currently backed by the dispute literature? #question , #Further-exploration-needed ,
CanLanguageModels_Analysis	"Real-world corpora involve rich and authentic conversations, but one rarely has access to each party's true underlying goals and emotions, making it difficult to quantify success or failure objectively" Page
CanLanguageModels_Analysis	Benefits/drawbacks of observed disputes
CanLanguageModels_Analysis	"anger in disputes provokes retaliation [21]" Page , #evidence
haleKODIS_NAACL_Annotatedpdf_Analysis	"Subjective 423 perceptions of the outcome of a dispute are a bet- 424 ter predictor of future negotiation decisions than 425 the actual economic result (Brown and Curhan, 426 2012)." Page 5 , #subjective
haleKODIS_NAACL_Annotatedpdf_Analysis	"do ex534 pressions of anger provoke escalation and impasses 535 in disputes (as previously claimed), can negotiation 536 satisfaction be predicted by emotional expression 537 alone, and how does culture shape these findings?" Page 7 , #research-question
haleKODIS_NAACL_Annotatedpdf_Analysis	"(Curhan et al., 328 2006; Brown and Curhan, 2012)" Page 4 #to-read , #theoretical-framework ,
haleKODIS_NAACL_Annotatedpdf_Analysis	relates to agreeability and behavior theory to guide measurement variables in stidy #to-read , #theoretical-framework ,
changTroubleHorizonForecasting2019_Analysis	"The main difficulty in directly adapting these models to the supervised domain

Paper (51)	Related Annotations
	of conversational forecasting is the relative scarcity of labeled data:" Page 2 #challenge , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	What labels do we need for dispute context? #challenge , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"parency, precluding an analysis of how exactly CRAFT models conversational context." Page 8 #challenge , , #fairness , #potential-gap
changTroubleHorizonForecasting2019_Analysis	application of fairness? #challenge , , #fairness , #potential-gap
changTroubleHorizonForecasting2019_Analysis	"window of only two comments would miss the chain of repeated questioning in comments 3 through 6 of Figure 1)" Page 2 #question , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	What kinds of patterns can we consider for "spiraling"? How do we define structure of a window? #question , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"attack-containing conversation is paired with a clean conversation from the same talk page, where the talk page serves as a proxy for topic.3" Page 4 , #Further-exploration-needed , #unsupervised-learning
changTroubleHorizonForecasting2019_Analysis	look into techniques for correlation control- - what kinds of structures to consider for the text? , #Further-exploration-needed , #unsupervised-learning
changTroubleHorizonForecasting2019_Analysis	"order-sensitive representation of conversational context?" Page 8 #conversational-forecasting , , #question
changTroubleHorizonForecasting2019_Analysis	to what extent does order matter for disputes? #conversational-forecasting , , #question
changTroubleHorizonForecasting2019_Analysis	"ntuition that comments in a conversation are not independent events; rather, the order in which they appear matters (e.g., a blunt comment followed by a polite one feels intuitively different from a polite

Paper (51)	Related Annotations
	comment followed by a blunt one)." Page 8 , #idea , #question
changTroubleHorizonForecasting2019_Analysis	do we care about patterns for learning dispute structure? Relation to emotional recognition? , #idea , #question
changTroubleHorizonForecasting2019_Analysis	"A practical limitation of the current analysis is that it relies on balanced datasets, while derailment is a relatively rare event for which a more restrictive trigger threshold would be appropriate." Page 9 , #question
changTroubleHorizonForecasting2019_Analysis	does balanced matter for disputes? do we need to quantify how detailed impasses or spirals were? , #question
changTroubleHorizonForecasting2019_Analysis	"n reality, derailment need not spell the end of a conversation; it is possible that a conversation could get back on track, suffer a repeat occurrence of antisocial behavior, or any number of other trajectories." Page 9 #conversational-forecasting , , #potential-gap
changTroubleHorizonForecasting2019_Analysis	What can we say about spiral shapes and impasses? Can disputes esalate and descalate in such a way an impasse would not be reached? #conversational-forecasting , , #potential-gap
changTroubleHorizonForecasting2019_Analysis	"Antisocial behavior is a persistent problem plaguing online conversation platforms; it is both widespread (Duggan, 2014)" Page #question , #to-read , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	compare antisocial behavior to dispute characteristics-- are they distinct concepts, or is one a subclass of the other? #question , #to-read , , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"deception (Girlea et al., 2016; P´erez-Rosas et al., 2016; Levitan et al., 2018)" Page 3
changTroubleHorizonForecasting2019_Analysis	deception classification post-hoc

Paper (51)	Related Annotations
changTroubleHorizonForecasting2019_Analysis	"disagreement (Galley et al., 2004; Abbott et al., 2011; Allen et al., 2014; Wang and Cardie, 2014; Rosenthal and McKeown, 2015)" Page 3
changTroubleHorizonForecasting2019_Analysis	disagreement classification post-hoc

#question

Paper (103)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"As parties are already linked," Page , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	What defines a link--furthermore, the type of the relationship? Historic agreement? Self-interests? , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	"Prior research on emotions in text focuses on a document as the unit of analysis and outputs coarse-grained “sentiment” rather than specific emotions like anger [59]." Page , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	Difference between intention and sentiment? , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	"e model classified each utterance of the dispute in isolation, and we created an overall score for the dialog by summing across these labels afterward." Page 4
EmotionallyAwareAgentsDispute_Analysis	what does "in isolation" mean here?
EmotionallyAwareAgentsDispute_Analysis	"“No I do not because you clearly did not read the description” might be seen as neutral in isolation but more negative in the context of the preceding line “So you do not see a need to apologize to me for sending me the wrong jersey.”" Page 4 , #data-analysis

Paper (103)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	how many lines in advance is enough context? , #data-analysis
EmotionallyAwareAgentsDispute_Analysis	"conflict literature suggests that expressions of compassion are an important predictor of negotiated outcomes" Page 4 , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	Why does the scale not initially differentiate positive emotions as much? If a scale is adjusted for a specific use case, can we compare it well with prior research? , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	"o address concerns that the emotions of each dialogue turn can depend on prior turns [47]" Page 5 , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	Can we further refine how far back the context needs to be to get better accuracy? Further if we assume in isolation, how are we sure the model "forgets" previous input to make a fresh decision? , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	"address concerns with T5-Twitter's labels, we substituted T5-Twitter's love with compassion and added a neutral label to not force GPT to pick an emotion where none is apparent." Page 5 , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	is this too mach variation to compare performance with T5? I feel like this would have a lot of confounding variables in a better outcome. But the goal is to increase the accuracy to align better with the subjective value , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	"we assess how each emotion label correlates with self-reported frustration for each annotation method." Page 5 , #data-analysis

Paper (103)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"or anger while GPT assigns more diverse labels and utilizes neutral as a dampener. GPT suggests the dialogues contain far more anger than joy, especially for buyers." Page 5 , #data-artifacts
EmotionallyAwareAgentsDispute_Analysis	What is the actual distribution of emotions for the data set? , #data-artifacts
EmotionallyAwareAgentsDispute_Analysis	"Backwards Regression. A subsequent backward regression clarifies which expressions significantly contribute to predicting SVI (* p < .001, p < .01, and * p < .05)" Page 6 , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	for backwards comparison, how do we choose the p-level for this use case? Difference between the magnitude/ sign and the p-level? What is the scale based on? , #model-comparison
EmotionallyAwareAgentsDispute_Analysis	"This reinforces findings in the social sciences on how anger shapes conflict [30, 48]," Page 8 , #Further-exploration-needed , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	What is "new" result versus not something currently backed by the dispute literature? , #Further-exploration-needed , #dispute-theory
CanLanguageModels_Analysis	"However, many of the disputes escalated and ended without agreement, thus forgoing their bonus. Even when agreements were reached, disputants often reported high frustration with their partner." Page 2
CanLanguageModels_Analysis	Why did many disputes lead to conflict?
CanLanguageModels_Analysis	"(1) whether or not the dispute ended in success or failure, and (2) whether or not the participants reported frustration with each other." Page 2

Paper (103)	Related Annotations
CanLanguageModels_Analysis	How does frustration get mapped to anger? In the other paper, we have a 10-item tactics survey, but no label on GPT model for frustration
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2 , #critique , #potential-gap
CanLanguageModels_Analysis	"importance to these issues via a payoff matrix, participants are free to assign their own importance to each issue." Page 3
CanLanguageModels_Analysis	why move away from payoff matrix as in initial KODIS paper on the cross-cultural differences
CanLanguageModels_Analysis	"ing LLMs and explore several prompt" Page 3 , #Further-exploration-needed
CanLanguageModels_Analysis	for what kinds of disputes? , #Further-exploration-needed
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 , #critique , #data-collection
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? , #critique , #data-collection
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6 , #data-analysis
CanLanguageModels_Analysis	in what cases did they not? , #data-analysis
CanLanguageModels_Analysis	"In this new task, we recruit N = 106 participants so that each mediation

Paper (103)	Related Annotations
	receives at least five annotations." Page 7 , #sampling
CanLanguageModels_Analysis	Any issues with sampling again and outcome qualioty/statistical effects? Is there more subjectivity now in the 20- to evaluate the justification? , #sampling
CanLanguageModels_Analysis	"Concretely, we ask to what extent the participant disagrees or agrees (1-10) with three statements depicted in Table 2." Page 7 , #experiment
CanLanguageModels_Analysis	Why a 10 point scale for agreeability? Also, what were the comstrains for reason to intervene on giving the reaso? DO we need to include any ethical considerdations in the kinds of repsonses? , #experiment
CanLanguageModels_Analysis	"come in the form of endowing the LLM's prompt with psychology-based negotiation strategies, as similar prior work exists with rule-based agents [19, 25]." Page 8 #potential-gap , , #to-read
CanLanguageModels_Analysis	is this for post dispute or during? #potential-gap , , #to-read
gratchFieldAffectiveComputing_Analysis	"d) action tendencies (such as preparation for fight versus flight)" Page 2
gratchFieldAffectiveComputing_Analysis	How are these different from psychological changes, and do these vary across individuals or are there generalizations?
hardtEqualityOpportunitySupervised2016_Analysis	"‖p − q‖2 6 √ 2 · dK (R, R′)." Page 11 #Bayesian-methods , #formula ,
hardtEqualityOpportunitySupervised2016_Analysis	"The feasible set of false/true positive rates of possible equalized odds predictors is thus the intersection of the areas under the A-conditional ROC curves, and above the main diagonal (see Figure 2). Since for any loss function the optimal false/true-

Paper (103)	Related Annotations
	positive rate will always be on the upper-left boundary of this feasible set, this is effectively the ROC curve of the equalized odds predictors." Page 9 #explanation , , #ROC
hardtEqualityOpportunitySupervised2016_Analysis	"y = 1, the constraint requires that Ŷ has equal true positive rates across the two demographics A = 0 and A = 1. For y = 0, the constraint equalizes false positive rates." Page 3 , #Further-exploration-needed
hardtEqualityOpportunitySupervised2016_Analysis	"R ∈ [0, 1]." Page 8 #todo ,
hardtEqualityOpportunitySupervised2016_Analysis	see if this is same R as mentioned to be random variable? #todo ,
hardtEqualityOpportunitySupervised2016_Analysis	"In this case, any thresholding of R yields an equalized odds predictor (all protected groups are at the same point on the curve, and the same point in false/true-positive plane)." Page 8 , #analysis
hardtEqualityOpportunitySupervised2016_Analysis	ROC curve-- since we condition on the distribution of A, picking t selects a point on the ROC curve. we generate the same graph for all possible (A,Y) distributions , #analysis
rudinInterpretableMachineLearning2022	"uch issues with explanations have arisen with assessment of fairness and variable importance [258, 82] as well as uncertainty bands for variable importance [113, 97]" Page 10 , #Interpretable-machine-learning , #critique , #Explainability
rudinInterpretableMachineLearning2022	What are uncertainty bands? , #Interpretable-machine-learning , #critique , #Explainability
rudinInterpretableMachineLearning2022	"inflated by including many “obvious” cases" Page 7 • Why are "obvious' cases misleading?

Paper (103)	Related Annotations
	ROC- Distinguishability between classes
rudinInterpretableMachineLearning2022	"n interpretable robotic surgeon would be worse than its black box counterpart. The question ultimately becomes whether the Rashomon set should permit such an interpretable robotic surgeon–and all scientific evidence so far (including a large-and-growing number of experimental papers on interpretable deep learning) suggests it would." Page 9 , #blackbox , #Further-exploration-needed
rudinInterpretableMachineLearning2022	What do we mean by "worse"? Are we defining this in terms of accuracy? , #blackbox , #Further-exploration-needed
rudinInterpretableMachineLearning2022	"Unfortunately, these topics are much too often lumped together within the misleading term “explainable artificial intelligence” or “XAI” despite a chasm separating these two concepts [250]" Page 9 , #Further-exploration-needed , #definition , #To-read
rudinInterpretableMachineLearning2022	"But function approximators are not used in interpretable ML; instead of approximating a known function (a black box ML model), interpretable ML can choose from a potential myriad of approximatelyequally-good models, which, as we noted earlier, is called “the Rashomon set”" Page 11 , #Interpretable-machine-learning , #Explainability
rudinInterpretableMachineLearning2022	difference between model and function in this context , #Interpretable-machine-learning , #Explainability
rudinInterpretableMachineLearning2022	"linear models, which includes scoring systems (and risk scores)" Page 21 , #logical-models
rudinInterpretableMachineLearning2022	linear models are logical models? , #logical-models

Paper (103)	Related Annotations
rudinInterpretableMachineLearning2022	"boosted models are not naturally sparse, and issues with bias arise under 1 regularization, as discussed in the scoring systems section." Page 23 , #Further-exploration-needed , #Generalized-Additive-Model
rudinInterpretableMachineLearning2022	What are proxies? Connects to how regularizaiton issues (stripping away too many small weights) if tree is not sparse can be an issue. , #Further-exploration-needed , #Generalized-Additive-Model
rudinInterpretableMachineLearning2022	"using a deep neural network that transforms the input space into a feature space where a kNN classifier will perform well (i.e., deep kNN). Papernot and McDaniel [229]" Page 26
rudinInterpretableMachineLearning2022	what is the relation to latent spaces here?
haleKODIS_NAACL_Annotatedpdf_Analysis	"This theoretical 103 framework distinguishes cultures by the degree to 104 which people's social identity is independent ver105 sus interdependent and thus shapes the importance 106 given to norms of reciprocity and honesty." Page 2 , #important , #theoretical-framework
haleKODIS_NAACL_Annotatedpdf_Analysis	what are norms? Why specifically recipricocity and honesty in the context of cultural analysis here. It seems the cultures are categorized according to values and how they relate to these norms? , #important , #theoretical-framework
haleKODIS_NAACL_Annotatedpdf_Analysis	"information 195 expressed in the dialog revealed participant's pri196 vate goals for the negotiation" Page 3 , #Further-exploration-needed
haleKODIS_NAACL_Annotatedpdf_Analysis	"Participants are next directed 217 to the dispute task implemented in Lioness Labs, 218 a software

Paper (103)	Related Annotations
	framework used for multi-participant 219 behavioral economic experiments" Page 3
haleKODIS_NAACL_Annotatedpdf_Analysis	in this general section consider the effects of randomization and some of those concepts from casual inference (?) from ISE 625. Do we care about individual characteristics in the survey that may be confounding variables in how "important" a dispute may be?
haleKODIS_NAACL_Annotatedpdf_Analysis	"otherwise, at one minute left, the 226 participant moves on and converses with an AI 227 counterpart." Page 3
haleKODIS_NAACL_Annotatedpdf_Analysis	is one minute enough time to resolve with an AI chatbot? 8 minutes total possible, but only 1 needed for AI?
haleKODIS_NAACL_Annotatedpdf_Analysis	"defines their goals in the negotiation" Page 4
haleKODIS_NAACL_Annotatedpdf_Analysis	do we set what the goals are for them in the negotiation?
haleKODIS_NAACL_Annotatedpdf_Analysis	"These points are 444 also used to determine the bonus." Page 6 , #confusion
haleKODIS_NAACL_Annotatedpdf_Analysis	I am unsure on what outcomes lead to "doubling". because if they have fixed weight and this formula holds the points + binary weights (I), then there is no doubling indicated with this linear additive utility. , #confusion
haleKODIS_NAACL_Annotatedpdf_Analysis	"We only exam- 565 ine human-human dyads; the final agreement (or 566 non-agreement) is excluded from the dialogues." Page 7
haleKODIS_NAACL_Annotatedpdf_Analysis	why is the outcome excluded?
haleKODIS_NAACL_Annotatedpdf_Analysis	"This is remarkable as participants forfeit 578 a cash bonus if they fail to achieve an agreement, 579 even though this was merely a simulated dispute." Page 8

Paper (103)	Related Annotations
haleKODIS_NAACL_Annotatedpdf_Analysis	Why was this the case? Was the reasoning for the high turnover of impasse accurate, or is there some bias/ influence in the study that may have lead to low resolution outcome
haleKODIS_NAACL_Annotatedpdf_Analysis	"Fig609 ure 3 illustrates the differences in occurrences of 610 emotions for the five countries we consider when 611 disputing against the same or a different country." Page 8
haleKODIS_NAACL_Annotatedpdf_Analysis	Is this based on the GPT turn-based rating?
changTroubleHorizonForecasting2019_Analysis	"Another compromising solution is to extract features from a fixed-length window, often at the start of the conversation" Page 2 , #fairness , #conversational-forecasting , #previous-work
changTroubleHorizonForecasting2019_Analysis	"window of only two comments would miss the chain of repeated questioning in comments 3 through 6 of Figure 1)" Page 2 , #dispute-theory , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	What kinds of patterns can we consider for "spiraling"? How do we define structure of a window? , #dispute-theory , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"longer windows risk missing the to-be-forecasted event altogether" Page 2 , #confusion , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	why do longer windows miss an event? , #confusion , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"while concomitantly being able to process the conversation as it develops (see Gao et al. (2018) for a survey)." Page 2 , #conversational-forecasting

Paper (103)	Related Annotations
changTroubleHorizonForecasting2019_Analysis	could something here be used as guide for the mediation response generation? , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"This is a useful property for the purposes of model analysis, and hence we focus on this as our primary dataset." Page 4
changTroubleHorizonForecasting2019_Analysis	considerations of hand annotated labels?
changTroubleHorizonForecasting2019_Analysis	"hidden state hcnon can then be viewed as an encoding of the full conversational context up to and including the n-th comment." Page 5 #model-architecture ,
changTroubleHorizonForecasting2019_Analysis	Can we vary token length to be more than 1 word #model-architecture ,
changTroubleHorizonForecasting2019_Analysis	"order-sensitive representation of conversational context?" Page 8 #conversational-forecasting , #dispute-theory ,
changTroubleHorizonForecasting2019_Analysis	to what extent does order matter for disputes? #conversational-forecasting , #dispute-theory ,
changTroubleHorizonForecasting2019_Analysis	"ntuition that comments in a conversation are not independent events; rather, the order in which they appear matters (e.g., a blunt comment followed by a polite one feels intuitively different from a polite comment followed by a blunt one)." Page 8 #dispute-theory , #idea ,
changTroubleHorizonForecasting2019_Analysis	do we care about patterns for learning dispute structure? Relation to emotional recognition? #dispute-theory , #idea ,
changTroubleHorizonForecasting2019_Analysis	"including those for which the outcome is extraneous to the conversation." Page 9 #conversational-forecasting ,

Paper (103)	Related Annotations
changTroubleHorizonForecasting2019_Analysis	what does this mean? the impact is outside of the actual conversation? #conversational-forecasting ,
changTroubleHorizonForecasting2019_Analysis	"A practical limitation of the current analysis is that it relies on balanced datasets, while derailment is a relatively rare event for which a more restrictive trigger threshold would be appropriate." Page 9 #dispute-theory ,
changTroubleHorizonForecasting2019_Analysis	does balanced matter for disputes? do we need to quantify how detailed impasses or spirals were? #dispute-theory ,
changTroubleHorizonForecasting2019_Analysis	"Antisocial behavior is a persistent problem plaguing online conversation platforms; it is both widespread (Duggan, 2014)" Page , #to-read , #dispute-theory , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	compare antisocial behavior to dispute characteristics-- are they distinct concepts, or is one a subclass of the other? , #to-read , #dispute-theory , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	"sequential neural models that make effective use of the intra-conversational dynamics (Sordoni et al., 2015b; Serban et al., 2016, 2017)," Page 2 , #to-read , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	reference for in-context generation , #to-read , #conversational-forecasting

#question , #critique , #potential-gap

Paper (1)	Related Annotations
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better

Paper (1)	Related Annotations
	rationale for intervening, and rated as providing a more effective mediation message to the disputants" Page 2

#question , #Further-exploration-needed

Paper (9)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"This reinforces findings in the social sciences on how anger shapes conflict [30, 48]," Page 8 , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	What is "new" result versus not something currently backed by the dispute literature? , #dispute-theory
CanLanguageModels_Analysis	"ing LLMs and explore several prompt" Page 3
CanLanguageModels_Analysis	for what kinds of disputes?
hardtEqualityOpportunitySupervised2016_Analysis	"y = 1, the constraint requires that Ŷ has equal true positive rates across the two demographics A = 0 and A = 1. For y = 0, the constraint equalizes false positive rates." Page 3
rudinInterpretableMachineLearning2022	"Unfortunately, these topics are much too often lumped together within the misleading term “explainable artificial intelligence” or “XAI” despite a chasm separating these two concepts [250]" Page 9 , #definition , #To-read
rudinInterpretableMachineLearning2022	"boosted models are not naturally sparse, and issues with bias arise under 1 regularization, as discussed in the scoring systems section." Page 23 , #Generalized-Additive-Model
rudinInterpretableMachineLearning2022	What are proxies? Connects to how regularizaiton issues (stripping away too many small weights) if tree is not sparse can be an issue. , #Generalized-Additive-Model
haleKODIS_NAACL_Annotatedpdf_Analysis	"information 195 expressed in the dialog revealed participant’s pri196

Paper (9)	Related Annotations
	vate goals for the negotiation" Page 3

#question , #data-collection

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#critique , #data-collection

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 #question ,
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? #question ,

#emotion-recognition , #question , #experiment , #potential-gap , #LLM

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#Further-exploration-needed , #experiment , #LLM

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#data-analysis

Paper (16)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"They used a dictionary-based approach and several pre-trained models, finding the best results for a T5 model fine-tuned on Twitter corpus [49]" Page 4
EmotionallyAwareAgentsDispute_Analysis	"“No I do not because you clearly did not read the description” might be seen as neutral in isolation but more negative in the context of the preceding line “So you do not see a need to apologize to me for sending me the wrong jersey.”" Page 4 #question ,
EmotionallyAwareAgentsDispute_Analysis	how many lines in advance is enough context? #question ,
EmotionallyAwareAgentsDispute_Analysis	"we assess how each emotion label correlates with self-reported frustration for each annotation method." Page 5 #question ,
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6 #question ,
CanLanguageModels_Analysis	in what cases did they not? #question ,
haleKODIS_NAACL_Annotatedpdf_Analysis	"Finally, 117 we assess several theoretical mechanisms surround118 ing the dispute, including process variables (e.g., 119 what tactics did parties use, what emotions were 120 expressed, and did parties understand

Paper (16)	Related Annotations
	their part121 ner's interests?) and outcome variables, including 122 objective and subjective measures concerning the 123 outcome of the dispute." Page 2 #methodology , , #variables
haleKODIS_NAACL_Annotatedpdf_Analysis	"Post130 conflict measures include beliefs about whether 131 their partner was human or AI and attitudes towards 132 using AI technology for such applications." Page 2 #methodology , , #variables
haleKODIS_NAACL_Annotatedpdf_Analysis	"cultural variability, 284 we use "elicited preferences," meaning that partic285 ipants are provided a fixed number of points they 286 can allocate across the four issues" Page 4 #methodology ,
haleKODIS_NAACL_Annotatedpdf_Analysis	choice modeling #methodology ,
haleKODIS_NAACL_Annotatedpdf_Analysis	"The concept of "integrative 392 potential" measures the potential for joint gains 393 (which may or may not be realized)." Page 5 #methodology , , #variables
haleKODIS_NAACL_Annotatedpdf_Analysis	this is related to the monetary bonus of good performance and the participant's self-reported goals. Utility has multiple possible solutions-- what goal they expect may change in weight #methodology , , #variables
haleKODIS_NAACL_Annotatedpdf_Analysis	"They are asked to do the same for 411 their partner. The distance between these estimates 412 and their partner's actual preferences can serve as 413 a measure of perspective-taking accuracy." Page 5 #methodology , , #self-report , #perspective-taking
haleKODIS_NAACL_Annotatedpdf_Analysis	This is compared to the initial assignment, but is the scale different? #methodology , , #self-report , #perspective-taking
haleKODIS_NAACL_Annotatedpdf_Analysis	"GPT4o (run on 06/28/2024) (Achiam et al., 548 2023) is prompted to annotate each dialogue 549 turn." Page 7 #emotion , #methodology ,
haleKODIS_NAACL_Annotatedpdf_Analysis	annotate text in turn-based manner #emotion , #methodology ,

#Question

Paper (24)	Related Annotations
CanLanguageModels_Analysis	"Fig. 3. GPT's selected reasons for intervention as a proportion of exchanges" Page 5
CanLanguageModels_Analysis	Does this also reflect the portion when ChatGPT interverena each time step? What is the 1-6 scale?
rudinInterpretableMachineLearning2022	"aim to help readers avoid common but problematic ways of thinking about interpretability in machine learning." Page 3 #Further-exploration-needed , #Interpretability-Principles ,
rudinInterpretableMachineLearning2022	Based on the persona, what are common "problematic interpretations" in relation to the domain? #Further-exploration-needed , #Interpretability-Principles ,
rudinInterpretableMachineLearning2022	"and that each feature is a meaningful predictor of the output on its own." Page 4 #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	Can tabular data have continuous values? what is meant by real, and how do we know it is a meaningful predictor? #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"with interactions" Page 4 #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	what interactions in this context? Do the categories potentially have overlap themselves? #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"complex interactions (i.e., more than quadratic" Page 4 #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	What are complex interactions? #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"small enough that they can be memorized by humans." Page 4 #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	What kind of data is easy to memorize by people? #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"While solutions of (*) would not necessarily be sufficiently interpretable to use in practice, the constraints would generally help us find models

Paper (24)	Related Annotations
	that would be interpretable (if we design them well), and we might also be willing to consider slightly suboptimal solutions to find a more useful model." Page 4 #Further-exploration-needed , , #technical
rudinInterpretableMachineLearning2022	Why is the solution set not fully interperetable? #Further-exploration-needed , , #technical
rudinInterpretableMachineLearning2022	"Equation (*) can be generalized to unsupervised learning, where the loss term would simply be replaced by a loss term for the unsupervised problem, whether it is novelty detection, clustering, dimension reduction, or another task." Page 4 , #unsupervised-learning , #interesting
rudinInterpretableMachineLearning2022	What is meant by modifiying the loss term? , #unsupervised-learning , #interesting
rudinInterpretableMachineLearning2022	"Choices of model form (e.g., the choice to use a decision tree, or a specific neural architecture) are examples of interpretability constraints." Page 5 #Interpretable-machine-learning , #Further-exploration-needed , , #interpretability-metrics
rudinInterpretableMachineLearning2022	How does choice of the model change the penalty? It is the input to the formula #Interpretable-machine-learning , #Further-exploration-needed , , #interpretability-metrics
rudinInterpretableMachineLearning2022	"fully interpretable and partially interpretable models, often preferring the former" Page 5 #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	The choice of model partially drives the desirability of how interpretable it is. What factors of a model do we consider, and is there a tradeoff in the penalty term? would it be smaller where the desirability for fully interpretable models is high? #Further-exploration-needed ,
rudinInterpretableMachineLearning2022	"decisions where an explanation would be trivial and the model is 100% reliable" Page 5 , #confusing
rudinInterpretableMachineLearning2022	Is this based on the classification/output of the model? does the model *need* 100% reliability? , #confusing

Paper (24)	Related Annotations
rudinInterpretableMachineLearning2022	"COMPAS depends on race other than through age and criminal history [13, 258]" Page 7 #Interpretable-machine-learning , #To-read , , #bias
rudinInterpretableMachineLearning2022	what does is mean by racial dependence? #Interpretable-machine-learning , #To-read , , #bias

#question , #experiment

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Concretely, we ask to what extent the participant disagrees or agrees (1-10) with three statements depicted in Table 2." Page 7
CanLanguageModels_Analysis	Why a 10 point scale for agreeability? Also, what were the comstrains for reason to intervene on giving the reaso? DO we need to include any ethical considerdations in the kinds of repsonses?

#question , #experiment , #potential-gap

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#question , #critique , #data-collection

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6

Paper (2)	Related Annotations
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"?

#question , #data-analysis

Paper (5)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"“No I do not because you clearly did not read the description” might be seen as neutral in isolation but more negative in the context of the preceding line “So you do not see a need to apologize to me for sending me the wrong jersey.”" Page 4
EmotionallyAwareAgentsDispute_Analysis	how many lines in advance is enough context?
EmotionallyAwareAgentsDispute_Analysis	"we assess how each emotion label correlates with self-reported frustration for each annotation method." Page 5
CanLanguageModels_Analysis	"he humanmediator followed instructions (i.e., they role-played as a mediator);" Page 6
CanLanguageModels_Analysis	in what cases did they not?

#question , #sampling

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"In this new task, we recruit N = 106 participants so that each mediation receives at least five annotations." Page 7
CanLanguageModels_Analysis	Any issues with sampling again and outcome qualioty/statistical effects? Is there more subjectivity now in the 20- to evaluate the justification?

#experiment , #question

Paper (0)	Related Annotations

Dataview: No results to show for table query.

#critique

Paper (35)	Related Annotations
EmotionallyAwareAgentsDispute_Analysis	"Neither line of work considers direct interactions between people nor do they consider the role of emotional expressions." Page 2 , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	in relation to mediation and argument-based agent-agent negotiation , #dispute-theory
EmotionallyAwareAgentsDispute_Analysis	"find that the subjective outcome of a dispute can be predicted from emotional expressions alone, ignoring the actual content of the dialog." Page 8 , #result , #potential-gap
EmotionallyAwareAgentsDispute_Analysis	"we also find evidence that “spirals of compassion” might reverse these effects (see Fig. 6(b)). Together, our findings suggest agents could intervene early to encourage participants to avoid costly escalation (see [44, 52])." Page 8 , #result , #potential-gap
EmotionallyAwareAgentsDispute_Analysis	What do we mean by "reverse the effect?" , #result , #potential-gap
CanLanguageModels_Analysis	"here 71 participants selected the GPT-generated mediation compared to 35 for the human-constructed one" Page 8
CanLanguageModels_Analysis	Is it possible GPT did seem more profesisonal?
CanLanguageModels_Analysis	"We find that the LLMs were rated significantly better in predicting when to intervene, rated as providing a better rationale for intervening, and rated as providing a more effective mediation

Paper (35)	Related Annotations
	message to the disputants" Page 2 #question , , #potential-gap
CanLanguageModels_Analysis	"Given the human-mediated dialogs alongside the GPT-mediated ones1, we move to subjective evaluations of each." Page 6 #question , , #data-collection
CanLanguageModels_Analysis	But how do we define how a human interperets "exchange level"? #question , , #data-collection
rudinInterpretableMachineLearning2022	"here is no scientific evidence for a general tradeoff between accuracy and interpretability when one considers the full data science process for turning data into knowledge." Page 6 #Interpretable-machine-learning ,
rudinInterpretableMachineLearning2022	"In these domains, neural networks generally find no advantage." Page 8 #blackbox ,
rudinInterpretableMachineLearning2022	relate to rashomon set. Tabular data has good performance without blackbox methods. What would necessitate the trade off to obfsucate interpretability to choose a blackbox model? #blackbox ,
rudinInterpretableMachineLearning2022	"uch issues with explanations have arisen with assessment of fairness and variable importance [258, 82] as well as uncertainty bands for variable importance [113, 97]" Page 10 #question , #Interpretable-machine-learning , , #Explainability
rudinInterpretableMachineLearning2022	What are uncertainty bands? #question , #Interpretable-machine-learning , , #Explainability
rudinInterpretableMachineLearning2022	"posthoc explanation, called saliency maps" Page 10 , #Explainability
rudinInterpretableMachineLearning2022	posthoc processing is not good , #Explainability

Paper (35)	Related Annotations
rudinInterpretableMachineLearning2022	"Typographical errors in input data are a prime example of this issue" Page 10 #blackbox ,
rudinInterpretableMachineLearning2022	"Each scoring system was created using a different method involving different heuristics. Some of them were built using domain expertise alone without data, and some were created using rounding heuristics for logistic regression coefficients and other manual feature selection approaches to obtain integer-valued point scores [see, e.g., 175]." Page 17 , #logical-model
rudinInterpretableMachineLearning2022	"When rounding, we lose all signal coming from all variables except the first two. The contribution from the eliminated variables may together be significant even if each individual coefficient is small, in which case, we lose predictive performance." Page 18 , #logical-model
rudinInterpretableMachineLearning2022	"1 regularization does more than make the solution sparse, it also imposes a strong 1 bias." Page 18 , #logical-model
rudinInterpretableMachineLearning2022	"Some of these constraints may be able to be placed into the mathematical program, but it is still not clear whether the solution of the optimization problem one solves would actually be close to the solution of the optimization problem we actually care about." Page 19 , #logical-model
rudinInterpretableMachineLearning2022	"those techniques often require a substantial amount of distance computations (e.g., to find out the nearest neighbors of a test input), which can be slow in practice. Also, it is possible that the nearest neighbors may not be particularly good representatives of a class, so that reasoning about nearest neighbors may not be interpretable" Page 27 , #case-based-reasoning

Paper (35)	Related Annotations
rudinInterpretableMachineLearning2022	"Part-based prototypes. One issue that arises with both nearest neighbor and prototype techniques is the comparison of a whole observation to another whole observation." Page 27 , #case-based-reasoning
rudinInterpretableMachineLearning2022	"n that sense, vectors in the latent space are “impure” in that they do not naturally represent single concepts [see 58, for a detailed discussion]." Page 32 , #disentanglement
rudinInterpretableMachineLearning2022	we align the axis of the neurons along latent spaces of classes- loading the concepts may take a lot of time if the network has a lot of neurons , #disentanglement
rudinInterpretableMachineLearning2022	"but this would not mean that the information flow concerning each concept goes only though that concept’s designated path through the network;" Page 33 , #disentanglement
rudinInterpretableMachineLearning2022	"abeled datasets for computer vision have a severe labeling bias: we tend only to label entities in images that are useful for a specific task (e.g., object detection), thus ignoring much of the information found in images." Page 36 #unsupervised-learning , , #disentanglement
rudinInterpretableMachineLearning2022	"distances from the original space when creating low-dimensional embeddings. But distances between points behave differently in high dimensions than in low dimensions, leading to problems preserving the distances." Page 42 , #dimension-reduction
rudinInterpretableMachineLearning2022	also dependent on assumptions about the accuracy of distance metric for data-- not god for neuron weights. not choosing which hyperparameters are important can lead to loss of global data. , #dimension-reduction

Paper (35)	Related Annotations
rudinInterpretableMachineLearning2022	"When their perplexity parameter or the number of nearest neighbors is not chosen carefully, algorithms can fail to preserve the global structure of the mammoth (specifically, the overall placement of the mammoth's parts), and they create spurious clusters (losing connectivity between parts of the mammoth) and lose details (such as the toes on the feet of the mammoth)" Page 44 , #dimension-reduction
rudinInterpretableMachineLearning2022	"e size of the Rashomon set differs from all of these quantities in fundamental ways, and it is important in its own right for showing the existence of simpler models." Page 50 #Further-exploration-needed , , #Rashomon-Set
rudinInterpretableMachineLearning2022	fundemn #Further-exploration-needed , , #Rashomon-Set
changTroubleHorizonForecasting2019_Analysis	"nly identify antisocial content after the fact limits their practicality as tools for assisting pre-emptive moderation in conversational domains." Page , #conversational-forecasting
changTroubleHorizonForecasting2019_Analysis	Current state of research has not explored real time moderation as in depth as post-hoc analysis case. , #conversational-forecasting

#potential-gap , #question , #to-read

Paper (2)	Related Annotations
CanLanguageModels_Analysis	"come in the form of endowing the LLM's prompt with psychology-based negotiation strategies, as similar prior work exists with rule-based agents [19, 25]." Page 8
CanLanguageModels_Analysis	is this for post dispute or during?