SOCIAL SCAFFOLDS: A Generalization Framework for Social Understanding Tasks

Anonymous ACL submission

Abstract

Effective human communication in social settings is contingent on recognizing subtle cues, such as intents or implications. Without such cues, NLP models risk missing social signals, instead relying on surface patterns. We introduce SOCIAL SCAFFOLDS, an automated framework for facilitating generalization across social reasoning tasks by generating rationales that make these social cues explicit. Grounded in narrative modeling principles, we generate task-agnostic rationales that capture different perspectives, i.e. that of the speaker, the listener, and the general world-view. Our experimental suite showcases that providing rationales as augmentations aids task performance for both supervised fine-tuning and in-context learning paradigms. Notably, providing all three rationale types significantly improves cross-task performance in 44% of cases, and inferred speaker intent in 31.3% of cases. We conduct statistical and ablation analyses that show how rationales complement the input text and are used effectively by models.

1 Introduction

Computational modeling of human communication in social interactions remains a fundamental challenge. Most human communication employs indirect language whose meaning goes beyond the literal form of the text (Yerukola et al., 2024; Yusupujiang and Ginzburg, 2023; Markowska et al., 2023; Dutt et al., 2024). As Figure 1 shows, uncovering the sarcastic intentions of the speaker is necessary to infer implicit hate toward immigrants. Recognizing such subtle cues is crucial for many tasks, e.g., automated content moderation (Calabrese et al., 2024; Horta Ribeiro et al., 2023), intent resolution (Yerukola et al., 2024; Joshi et al., 2021), and others (Kim et al., 2024; Qian et al., 2024).

Our study investigates whether social *rationales*, i.e., textual explanations that make the implicit social meaning of the message apparent, can serve as scaffolds to transfer across different social reasoning tasks. Prior work has demonstrated that rationales can not only enhance task performance but also aid transfer across domains (Bhan et al., 2024; Dutt et al., 2024). We hypothesize that task-agnostic rationales **can also facilitate generalization across dialogue understanding tasks**. Since dialogues are often underspecified (Sap et al., 2022a), models trained solely on the utterance may rely on shallow surface cues correlated with task labels. Incorporating rationales can aid in transfer to unseen tasks by learning generalizable signals and reducing reliance on surface text.

To that end, we introduce SOCIAL SCAFFOLDS, an automated framework to facilitate generalization by generating social rationales. These rationales differ in spirit from the "task-specific explanations" used in NLI and commonsense reasoning (Zelikman et al., 2023; Wiegreffe et al., 2021) which do not necessarily capture pragmatic aspects. We explore rationales that are (i) general enough to be elicited for any dialogue, (ii) open-ended and not constrained to a reduced vocabulary set (like dialogue acts), (iii) task-agnostic, and (iv) capable of capturing different perspectives.

We test the utility of SOCIAL SCAFFOLDS for six distinct social interaction tasks such as negotiation and argumentation. We apply our framework to generate $\approx$ 200K rationales using both **proprietary and open-source LLMs**. Motivated by narrative modeling principles (Eisenberg and Finlayson, 2016; Hamilton, 2024), we explore rationales that reflect (i) the intentions and beliefs of the speaker (Dutt et al., 2024; Zhou et al., 2023), (ii) the effect of the utterance on the listener (Yusupujiang and Ginzburg, 2023), and (iii) the common world view that participants presuppose to be true (Mulcahy and Gouldthorp, 2016). We refer to these rationales as intentions (INT), hearer reactions (HR), and presuppositions (PreSup) respectively.

We test the impact of adding rationales on task

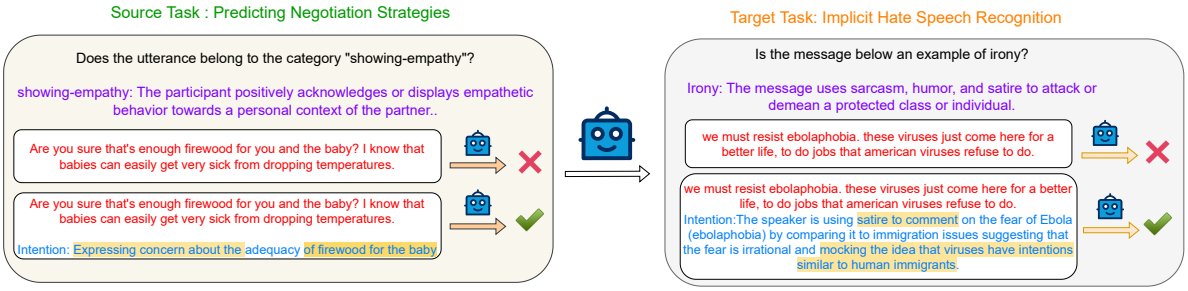


Figure 1: We illustrate the phenomena of indirect or subtle language usage in two scenarios; the scenario on the left corresponding to predicting negotiation strategies, whereas the scenario on the right corresponds to identifying different categories of hate. For both cases, we see how the model fails to associate the input message (in red) with the description of the label (in purple) since it is unable to capture the hidden cues in the message. Incorporating rationales, as additional inputs, can guide model prediction for both in-domain and cross-task settings.

performance in both supervised fine-tuning (SFT) and in-context learning (ICL) paradigms. Despite modest in-domain performance gains, **incorporating rationales significantly improves performance for both ICL and cross-task transfer.** In particular, rationales comprising all three perspectives (hereafter "ALL") yield significant cross-task transfer gains for 44% of the cases. Following closely are the speaker's intentions which improve cross-task transfer and ICL performance 30.5% and 31.3% of cases, respectively. Complex tasks characterized by a higher skew in label distributions and infrequent label categories benefit the most from rationales. We illustrate the benefits of adding the speaker's intentions on two tasks in Figure 1.

Comprehensive analyses show that our rationales are **task-agnostic**; how similar a rationale is to a task-specific label is not indicative of its task performance. Moreover, different categories of rationales (e.g. INT or PreSup) **capture different perspectives** as evidenced by their high soundness scores and low similarity. Our ablation studies and perturbation experiments highlight that the rationales **complement the input text** such that including both yields the best results. We also carry out qualitative analysis to probe the utility of certain tokens in the rationales in guiding model prediction.

Our contributions are the following:

- We propose SOCIAL SCAFFOLDS, a framework to facilitate generalization for different dialogue understanding tasks.
- We curate a dataset of 200K task-agnostic social rationales for six dialogue understanding tasks.
- We conduct extensive experiments to empirically demonstrate the utility of our framework.

Overall, our SOCIAL SCAFFOLDS framework show the promise of pragmatics-oriented data augmentation for social understanding and generalization. We will make our dataset and code public.

2 Related Work

We contextualize our work in the broader literature on generalization in dialogue tasks, as well as on the role of rationales in NLP.

2.1 Generalization in Dialogue

Generalization in dialogue is challenging because interactions are typically structured to accomplish a task rather than simply conveying information. Such a task-centered organization enables participants to rely heavily on implicit cues by omitting information they know to be shared among all participants (Dutt et al., 2024).

Mehri (2022) outlines different types of generalization imperative for dialogue. These include (i) new inputs arising from covariate shift or stylistic variation (Khosla and Gangadharaiah, 2022), (ii) new problems in dialogue modeling such as evaluation and response generation (Peng et al., 2020), (iii) new outputs and schemas corresponding to out-of-domain shift (Larson et al., 2019) and (iv) new tasks such as controlled generation or fact verification (Gupta et al., 2022).

In this work, we focus on generalization across **different dialogue understanding tasks** and investigate how rationales can act as scaffolds to bridge across tasks. Previous work on few-shot generalization in dialogue has benefited from large-scale multitask pre-training (Wu et al., 2020; Peng et al., 2021; Hosseini-Asl et al., 2020) or instruction tuning (Gupta et al., 2022; Wang et al., 2025; Sanh et al.; Wang et al., 2022). We propose an efficient solution that uses the underlying social cues in a

dialogue as augmentations to unify multiple tasks without the need for large-scale pretraining.

2.2 Rationales in NLP

In NLP, "rationales" [1] has long been used to refer to *textual explanations*, either generated by machines or humans (Camburu et al., 2018). Rationales serve several purposes, such as facilitating common sense and social reasoning (Zelikman et al., 2022; Majumder et al., 2022), explaining the predictions of neural models (Wiegreffe et al., 2021; Jayaram and Allaway, 2021; Zaidan et al., 2007), and assisting humans in their tasks (Das and Chernova, 2020; Joshi et al., 2023; Zhang et al., 2023).

Recent work has demonstrated the effectiveness of LLM in generating step-by-step explanations or rationales (Gurrapu et al., 2023) that subsequently benefit downstream tasks. (Rao et al., 2023; Wei et al., 2022; Zelikman et al., 2022). Rationales have also contributed to the OOD generalization (Majumder et al., 2022; Xiong et al., 2023; Joshi et al., 2022). Building upon this foundation, we frame rationales as the elicited verbalization of the underlying social signals that helps overcome some limitations of static text such as the omission of communicative intent (Sap et al., 2022b).

Our work shares builds upon the prior work of Dutt et al. (2024) which investigates the domain generalization capabilities of rationales for dialogue understanding tasks. Firstly, we investigate the generalization capabilities of rationales across different social understanding tasks and not simply across different domains for the same task. Secondly, we explore rationales that capture multiple perspectives whereas prior work has emphasized mostly on the speaker's intentions.

3 Modeling Framework

We present SOCIAL SCAFFOLDS, an automated framework that facilitates task generalization by generating different kinds of social rationales to capture the implicit information behind a message.

3.1 Rationale Types

We explore three distinct but complementary perspectives to generate the rationales. Motivated by prior work on narrative modeling, we present a one-to-one correspondence of the rationale category with the narrative perspective or point-of-view.

[1] While rationales can also refer to a subset of input tokens or words that contribute to a classification decision (Bao et al., 2018), we use it in the broader sense of textual explanations.

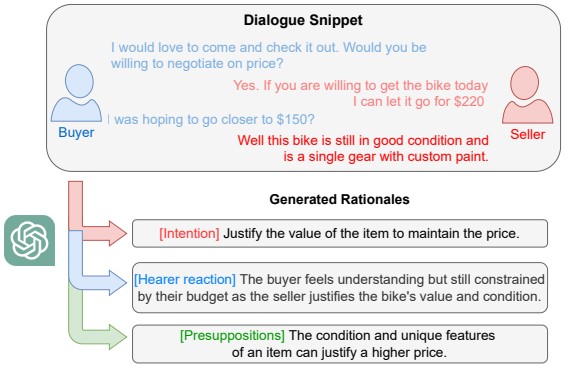


Figure 2: An overview of SOCIAL SCAFFOLDS for a negotiation snippet between a buyer and a seller. We prompt an LLM with the dialogue to generate rationales corresponding to the speaker's intentions, the hearer's reaction, and the presuppositions. For brevity, we show the generations for only the seller's last utterance.

Intentions: Intentions (or INT) refer to the speaker's hidden beliefs and desires, and correspond to the *first-person perspective*. They capture the implied meaning behind the speaker's utterance or signal the outcome the speaker wants (Dutt et al., 2024; Yusupujiang and Ginzburg, 2023).

Hearer Reaction: Hearer reactions (or HR) (Zhou et al., 2023; Sap et al., 2020) captures the effect the utterance might have on the listener(s). It provides insight into the listener's emotions or belief states, akin to second-order thinking, and corresponds to the *second-person perspective.*

Presuppositions: We use presuppositions (hereafter PreSup) to refer to general facts or truths that participants believe for the utterance to be credible. PreSup not only encapsulates common sense reasoning or social and communal norms often observed in practice (Perez Gomez, 2021; Kim et al., 2022), but also provides a de-contextualized and impersonal insight and thus serves as a *third-person perspective* (Mulcahy and Gouldthorp, 2016).

3.2 Rationale Generation Framework

We describe our prompting setup to automatically generate the different categories of rationales. Figure 2 presents a sample negotiation snippet with the corresponding intention, hearer reaction, and presupposition for the seller's last utterance.

SOCIAL SCAFFOLDS takes as input a multiparty dialog and generates rationales using a Large Language Model (such as GPT-4o) on an utterance-by-utterance basis. We employ a structured prompting framework to ensure that the generated rationale aligns with its corresponding utterance. We ad-

dress erroneous cases by prompting the framework to regenerate the rationales in an iterative manner. Additional details appear in Appendix Section B .

We generate each rationale category (e.g intentions or presuppositions) using our framework separately to prevent any ordering effects. We do not provide few-shot instances to avoid biasing the generations with previous seen examples unlike Dutt et al. (2024). Such a setting enables us to compare and contrast (i) different categories of rationales and (ii) rationales of the same category but generated by different LLMs. We explore both proprietary models such as GPT-4o and GPT-3.5-turbo and open-source LLMs such as Gemma-2-27B-it as the backbone of SOCIAL SCAFFOLDS.

3.3 Assessment of Rationale Quality

Since our framework automatically generates rationales without any human supervision, we develop a rigorous annotation manual to assess the validity of those generations based on three criteria: soundness, informativeness, and relevance. Additional details of these criteria appear in Appendix C

We score each rationale for each criteria using a Likert scale of 1 to 3, with 1 being the lowest and 3 the highest. Our two annotators or evaluators had a graduate level proficiency in English and at least five years of experience in computational linguistics and NLP. Due to the highly subjective nature of the task, we relied on these professional annotators as an alternative to crowd-sourcing or employing an automated annotation framework.

We compute the inter-rater reliability scores using the multi-item agreement measure of Lindell et al. (1999) and observe strong to moderate agreement on all three criteria: soundness (0.98), informativeness (0.76), and relevance (0.70). The mean scores of soundness, informativenss, and relevance are 2.95, 2.76, and 2.61 respectively, highlighting that the rationales are of sufficiently high-quality.

Our preliminary experiments (see Appendix F) highlight that the rationales of different categories differ substantially between themselves showcasing that each category captures distinct concepts. We observe a low similarity between the rationale and the corresponding utterance, signifying that the rationale generated captures information distinct from what is in the text. We also note that the rationales generated by different LLMs (i.e. specifically GPT-4o and Gemma-2-27B-it) bear a strong similarity to each other.

4 Experimental Setup

We outline the details of our methods or experimental setup for investigating the role of rationales in aiding generalization for understanding tasks. We describe the tasks, models, settings, and metrics.

4.1 Tasks and Datasets

We explore six dialogue understanding tasks, each instantiated with a distinct dataset, such that each task operates over a distinct domain. Moreover, these datasets have unique labels or categories to prevent any overlap between them. Such a setting would enable us to inspect the capabilities of rationales in a cross-task setting, where a model is trained for one task and then evaluated on another.

Our datasets include (i) P4G (Wang et al., 2019b) to identify persuasive strategies in charitable donations, (ii) CaSiNo (Chawla et al., 2021) to detect negotiation tactics during camping, (iii) Res_CB (Dutt et al., 2021) to categorize strategies employed to resist persuasion in online bargaining, (iv) EMH (Sharma et al., 2020) to understand different dimensions of empathy, (v) PROP (Jo et al., 2020) to categorize different kinds of argumentation, and (vi) IMP_HATE (ElSherief et al., 2021) to classify different kinds of implicit hate speech.

We present a brief overview of the dataset statistics in Table 1 and their corresponding distribution of labels in Figure 7 of the Appendix A. We observe that the datasets exhibit distinct characteristics, such as long conversations for P4G and PROP, and highly skewed labels for CaSiNo and Res_CB.

4.2 Configurations: SFT and ICL

We test the impact of rationales on downstream task performance in two distinct configurations. The first is a supervised fine-tuning (SFT) setup (Figure 3); we instruct-tune a pre-trained language model on a given source task (say persuasion) and then subsequently evaluate it on a new target task (say argumentation) in a 0-shot or few-shot setting. [2] The second is an in-context learning (ICL) setup, where we prompt an LLM with 0-shot or few-shot examples with the rationale as a control condition.

Since we investigate task transferability, it is imperative for us to map tasks with distinct label categories into a common shared space. **We format each task as binary classification**, such that the model outputs "Yes" or "No", depending on

[2] We also explore parameter efficient fine-tuning of LLama-3-8B-it as another SFT model. Details are in Appendix D

Dataset	Avg Words per Turn	Avg Turns per Dialog	# Turns	# Labels
P4G (Wang et al., 2019a)	10.75 / 13.76 / 11.53	18.74 / 15.45 / 17.9	4004 / 110 / 154	11 / 11 / 11
CaSiNo (Chawla et al., 2021)	21.53 / 20.29 / 26.50	5.42 / 4.88 / 5.02	4862 / 49 / 247	10 / 9 / 10
Res_CB (Dutt et al., 2021)	12.22 / 13.63 / 13.71	5.86 / 5.18 / 6.09	6348 / 160 / 160	8 / 8 / 8
PROP (Jo et al., 2020)	12.55 / 14.86 / 15.71	11.66 / 9.47 / 12.21	741 / 43 / 75	4 / 4 / 4
EMH (Sharma et al., 2020)	54.03 / 47.75 / 53.83	1 / 1 / 1	1823 / 104 / 112	3 / 3 / 3
IMP_HATE (ElSherief et al., 2021)	15.79 / 17.18 / 15.39	0 / 0 / 0	3182 / 156 / 153	6 / 6 / 6

Table 1: Dataset statistics across the train, validation, and test splits. Additional details in Appendix.

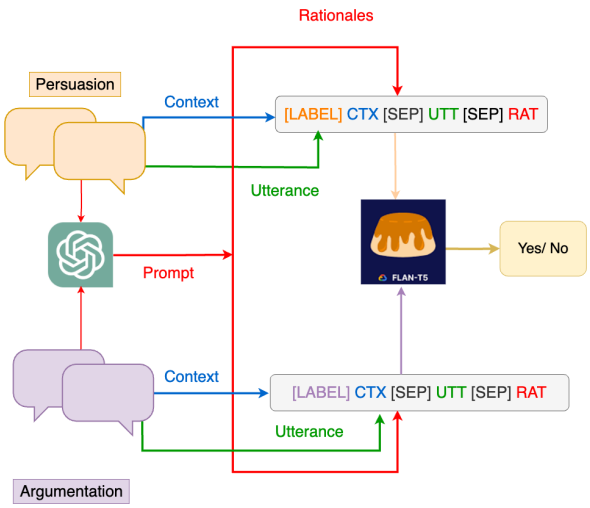


Figure 3: Overview of our SFT setting. For a source task, we instruction-tune FLAN-T5 with the label definition, dialogue context, utterance, and rationale as input and predict "yes" or "no" for the corresponding label. This model is then deployed for a new target task.

whether the utterance complies with the label definition. The input to the model is the label definition, the utterance, the dialog context, and the corresponding rationale. We adopt the binary classification framework for both SFT and ICL settings. Such a design would allow for a fair comparison of the two paradigms. Moreover, fine-tuned LMs with a single multiclass classification head is unlikely to generalize in a 0-shot setting. We show an example of how these tasks have been setup in Figure 1.

4.3 Models and Metrics

We use the base version of Flan-T5 (Chung et al., 2022) as our instruction-tuned model (i.e. SFT),[3] while Gemma-2-9B-it (Team, 2024) and LLama-3-8B-it (AI@Meta, 2024) serve as the models for ICL. These models have been trained to follow instructions and thus serve as strong baselines for the respective experimental paradigms. We measure the difference in performance from adding rationales as part of the input text (i.e., intentions, presuppositions, and hearer reaction) over only the utterance (which serves as the baseline).

Due to the skewed label distribution, we use the macro-F1 score as our evaluation metric for each of these six tasks. Following the recommendations in Dror et al. (2018), we employ the nonparametric bootstrap test of Berg-Kirkpatrick et al. (2012) to measure whether the rationale-augmented model's performance was statistically significant from the baseline. We reject the null hypothesis for cases with p-value ≤ 0.05. We perform each experiment for three seeds to account for variations over runs.

5 Results & Analysis

We present our experimental results using GPT-4o rationales. Appendix D highlights similar trends with the rationales generated by other LLMs.

5.1 Impact of Rationales in an SFT Setup

We inspect the impact of adding rationales on task performance in a supervised fine-tuning setup for both in-domain and cross-task transfer. The in-domain results serve to validate prior work that social rationales can enhance task performance whereas the transfer results showcase whether these rationales can facilitate task generalization.

In-domain Results: Table 2 shows that rationales improve in-domain performance on five of six tasks with significant gains for res_CB, PROP, and IMP_HATE, and a significant drop for EMH. The rationale with the greatest impact on performance varies across tasks (e.g. intentions are helpful for CaSiNo and res_CB, while the hearers' reaction aids P4G), implying that no individual category acts as a silver bullet. Nevertheless, **adding all three rationale categories (ALL) has the most in-domain benefit**, followed by intentions. Appendix D shows that our chosen FLAN-T5 model exhibits competitive in-domain task performance and surpasses prior baselines for all tasks.

[3]Due to limited computational budget, we focus on FLAN-T5 with more experiments on PEFT models in the Appendix.

Rationale	P4G	CaSiNo	res_CB	PROP	EMH	IMP_HATE
UTT	69.70 +/- 2.42	71.22 +/- 1.70	66.77 +/- 1.02	82.38 +/- 1.21	**90.91 +/- 0.13**	62.68 +/- 0.79
+ INT	69.36 +/- 1.45	**72.35 +/- 0.50**	**70.91 +/- 0.71**	84.66 +/- 1.07	89.35 +/- 1.35	67.91 +/- 1.49
+ HR	70.54 +/- 1.70	71.71 +/- 0.84	68.80 +/- 0.97	82.88 +/- 1.69	90.26 +/- 0.32	65.08 +/- 0.34
+ PreSup	68.12 +/- 2.30	71.81 +/- 1.39	69.69 +/- 1.51	80.11 +/- 2.86	89.37 +/- 0.16	62.88 +/- 2.55
+ ALL	**70.67 +/- 2.08**	70.68 +/- 1.12	67.72 +/- 2.59	**86.25 +/- 3.28**	90.46 +/- 1.12	**68.21 +/- 0.97**

Table 2: Performance of FLAN-T5 model in an in-domain setting with GPT-4o rationales across six tasks. The baseline includes only the utterance (UTT) which we compare by adding rationales, i.e. intentions (INT), hearer-reactions (HR), presuppositions (PreSup), and all three (ALL). We note the mean and s.d. across three runs.

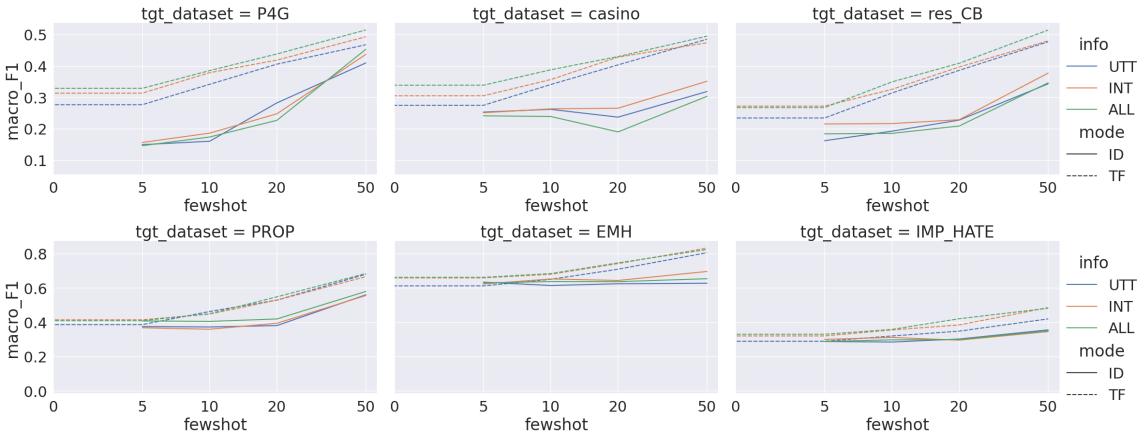


Figure 4: Impact of GPT-4o rationales on cross-task performance for different tasks and fewshot settings. TF and ID corresponds to the cross-task transfer and in-domain setting respectively. For better readability, we show results for only the intentions (INT) and all three categories (ALL).

Cross-Task Transfer Results: Our transfer experiments over the six tasks yield 30 unique source-target pairs. Figure 4 shows the aggregate impact of adding rationales for the six target datasets.[4] Against the utterance-only baseline, we see consistent and significant gains during transfer (in dotted lines) over the in-domain setting (in solid lines) for different zero-shot and few-shot cases. A similar trend is seen for PEFT models, albeit with not as pronounced gains (Figure 14 in Appendix D).

The impact of rationales is highest for target datasets that exhibit a high skew in their label distribution (such as P4G, res_CB, and IMP_HATE). Label-wise F1 scores in Figures 23 and 24 reveal that the rationales improve performance for impoverished label classes such as "foot-in-the-door" for P4G, "Self-Assertion" and "Self-Pity" for res_CB, and "threatening" for IMP_HATE. We thus posit that rationales help more complex dialogue tasks for both in-domain and cross-task settings.

We investigate whether a model's in-domain performance on a source task correlates with their transfer performance on a target task. Likewise, we explore whether rationales that yield in-domain gains are good predictors of transfer success. We observe negligible correlation in Table 15 on both fronts using Spearmann's ranked correlation. However, we observe from Figure 5 that adding intentions results in an overall positive impact for 28 of the 30 source target pairs.

Takeaway 1

Despite modest in-domain performance, rationales yield significant gains in transfer.

5.2 Impact of Rationales in an ICL Setup

Intentions improve performance on target datasets 91.7% of the time in an ICL paradigm (see Tables 3 and 19) across different few-shot settings and models. Presuppositions and hearer reactions fare better at 0-shot and 5-shot settings respectively. Surprisingly, adding ALL does not yield as significant gains as in the SFT setup. Nevertheless, SFT models in a cross-task transfer, with a mere 20 or 50 few-shot examples, can surpass ICL performance.

[4] Additional results for the HR and PreSup rationales are in Figures 10 and 11 of the Appendix.

	Gemma-2-9B-it						Llama-3-8B-it					
RAT	P4G	CaSiNo	res_CB	PROP	EMH	HATE	P4G	CaSiNo	res_CB	PROP	EMH	HATE
UTT	29.2	35.9	33.8	47.6	48.8	33.9	20.1	30.1	26.9	43.9	66.8	32.4
+ INT	31.3	38.5	35.4	**51.0**	55.0	**38.8**	20.9	31.3	**29.5**	**47.2**	67.2	32.7
+ HR	28.0	38.8	35.1	44.2	**56.6**	35.1	**21.2**	29.3	28.3	43.9	67.1	32.9
+ PreSup	32.3	40.5	**38.2**	45.2	53.3	38.3	20.1	**32.2**	28.2	45.6	**67.6**	**33.4**
+ ALL	**33.7**	**40.6**	33.9	43.8	55.5	37.1	21.1	28.9	27.9	44.8	**67.6**	31.9

Table 3: Zero-shot performance of models in an in-context learning setup with GPT-4o rationales.

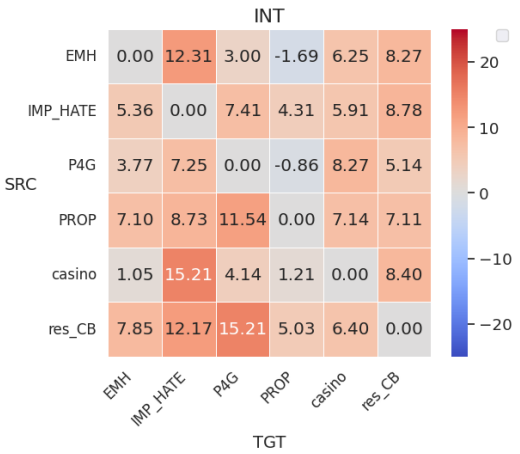


Figure 5: Net performance gains across different source and target tasks from adding speakers' intentions.

5.3 Factors affecting Task Performance

We inspect factors that impact performance at the instance-wise and global level for SFT and ICL.

Instance-wise Correlations: We investigate whether certain rationale characteristics correlate with task performance. These include (i) the length of the rationale, (ii) the length of the dialogue context, (iii) similarity between the rationale and the utterance, (iv) similarity between the rationale and the label description, (v) readability scores via Flesch's readability ease (Farr et al., 1951; Kincaid, 1975), (vi) valence, arousal, and dominance scores via the NRC lexicon (Mohammad, 2018), and (vii) emotional intensity, emotional polarity, and empathy scores (Wu et al., 2024).

We measure the point bi-serial correlation between each individual factor and instance-wise accuracy. A low (almost zero) correlation for all the factors in Table 22, signals that task performance is **not dependent on these data artifacts**. Our rationales are also **task-agnostic**; the similarity between a given rationale and the task-specific label is not predictive of task performance.

Global Generalization Characteristics: We perform a multivariate ANOVA analysis where our dependent variable is the relative change in performance from adding the rationales. Our covariates or independent variables include the rationale category, the LLM used to generate the rationales, the source dataset, and the target dataset [5], and the number of few-shot examples. We also include the pairwise interaction effects of these covariates. We note the F-statistic and their corresponding p-value for in-domain, cross-task and ICL setting respectively in Tables 23, 24, and 25 in the Appendix F. We consider covariates to have a significant effect when their corresponding p-values are ≤ 0.05.

In a nutshell, across all the different experimental setups, **the rationale category significantly influences task performance.** Unanimously across all settings, intentions yield the highest positive gains on average, followed by the hearer's reactions and then the presuppositions. We summarize the fraction of cases where adding rationales improves task performance for both SFT , which includes in-domain (ID) and cross-task transfer (TF) settings, and ICL setups in Figure 6.

Takeaway 2

Overall, the speaker's intentions have the greatest improvements on task performance.

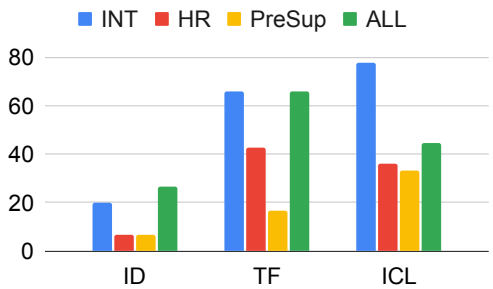


Figure 6: Fraction of cases where rationales improves performance for in-domain (ID), cross-task transfer (TF), and in-context learning settings (ICL).

[5] We have the source dataset only for cross task transfer

Dataset	Label	Utterance text	Rationale Text	CAT
casino	vouch-fair	hey buddy I hope we both end up with a good deal:)	Expressing hope for a **mutually beneficial outcome**	INT
IMP_HATE	white_grievance	but that wouldn't enable them to destroy white neighbourhoods .	There is a belief or concern that certain actions or policies could lead to **the destruction of white neighborhoods.**	PreSup
P4G	foot-in-the-door	Every little bit help.	EE feels reassured **that their small donation** is still valuable.	HR
P4G	foot-in-the-door	Every little bit help.	**Reassure the listener that any contribution is valuable.**	INT
res_CB	Self Pity	at this i can only pay about 1600 could you do that	Seller **realizes the buyer's budget constraints.**	HR

Table 4: We present instances across different datasets where adding the rationale information was crucial in predicting the correct label always. We compute Shapley values for each token in the rationale to observe its contribution to the model's decision; the highlighted portions correspond to high positive associations with the label.

5.4 Necessity and Sufficiency of Rationales

Having demonstrated the practical utility of adding rationales, we now examine if the information encoded in the rationale is sufficient or necessary.

Sufficiency Claims: We investigate the sufficiency claims of rationales, i.e. whether the rationales can meaningfully capture all the information in the utterance. We carry out two ablation experiments to examine the relative change in task performance compared to the baseline (i.e. when only the utterance is included). In the first experiment, we train the model using both the corresponding rationale and utterance, but provide only the rationale information during testing. In the second experiment, we omit out the utterance completely and train on only the rationales. For both cases, task performance degrades significantly highlighting that **rationales are insufficient by themselves and cannot match the baseline task performance**.

Necessary Claims: We investigate whether the rationale text is useful or necessary in guiding model prediction. We perform sensitivity analysis by perturbing the rationale in different ways such as synonym replacement or deletion. Additional details of our experiment appear in Appendix H. We note a deterioration in task performance as the proportion of text perturbed increases; specifically, deletions have the greatest impact while synonym replacement has the least (see Figure 22). Our findings thus highlight that models do indeed rely on the text in the rationales for classification.

Interpretability Analysis: We now carry out the interpretability analysis (Roth, 1988) using SHAPLEY for instances where the rationales consistently yielded the correct answer (i.e. across all seeds). We observe the SHAPLEY values specifically for these highlighted tokens in the rationales that guide model prediction. We present five examples of these instances in Table 4 for four datasets corresponding to different rationale categories. We observe that the highlighted tokens in the rationale text align with human intuition to explain the label category. For example, the phrase "destruction of white neighborhoods" acts as a signal for white aggression and "that their small donation" is a signal for foot-in-the-door strategy in Table 4. We present additional examples of these in the Appendix I.

Takeaway 3

Our generated rationales are task-agnostic, capture different perspectives, and complements the input utterance.

6 Conclusion

We introduce SOCIAL SCAFFOLDS, a framework for facilitating generalization across different dialogue understanding tasks via rationales. Motivated by narrative modeling principles, our rationales capture perspectives of the speaker, the listener, and the general world view. We apply our framework to generate $\approx$ 200K rationales spanning six distinct dialogue tasks. We design a comprehensive evaluation suite that spans 5,400 supervised fine-tuning and in-context learning experiments and demonstrate that rationales aid task performance in both experimental setups. In particular, incorporating only the speaker's intentions and all three rationale categories yields significant cross-task transfer gains (31.3% and 44.0% of the times). Our analysis also reveals that rationales are task-agnostic and complement the utterance.

Limitations

We highlight some of the potential limitations of our work.

(i) We have only focused on simple multi-label and multi-class classification tasks, and that too at an utterance level. We plan to investigate whether rationales can facilitate dialogue understanding at conversational-level and whether these social rationales can help generalize to new dialogue tasks such as response generation.

(ii) While we demonstrate the effectiveness of rationales at a dataset level for both supervised fine-tuning and in-context learning scenarios, we did not explore their effectiveness at a per-instance basis. Future work could entail identifying which cases benefit the most from adding rationales by employing LLM-as-a-judge.

(iii) We explore both proprietary and closed-source LLMs to generate these rationales. Although we released our entire 200K rationale database to promote future research in this space, we acknowledge that we cannot guarantee the reproducibility of generating the exact rationales due to the opaque nature of proprietary models.

(iv) Our proposed framework is simple in design and employs only a single LLM to generate rationales for a given conversation. One can envision developing a more rigorous agent-based framework that can automatically validate the quality of a rationale during the generation process, leading to higher grade rationales. We emphasize that while this is a promising research direction, it goes beyond the scope of the current work, where we focus more on the effectiveness of rationales on downstream task transfer and not how to generate the best possible rationales themselves.

(v) Likewise, while we observe the positive impact of our machine-generated rationales on task performance, and validate that the rationales are of sufficient high quality, further research is necessary to compare and contrast these machine-generated rationales from human-generated ones.

(vi) Our comprehensive experimental suite spans 810 in-domain , 4050 cross-task, and 540 in-context learning experiments. Subsequently, the majority of our experiments have been tested using our FLAN-T5 model for the SFT setup. While we do show similar trends using PEFT based models as well, we need to scale back on the number of the datasets due to our restricted computational budget. Even then our PEFT-based setting covers 180 in-domain and 360 cross-task runs.

Ethical Concerns

Our research relies on the responses generated by LLMs which are known to exhibit hidden biases in their representations. While during our experiments, we encountered no potential biases in terms of offensive language or stereotypes in the generated response for our controlled setting of social meaning detection, we implore practitioners and other researchers to conduct thorough analysis before adopting our particular prompting approach for the respective use-case. We also recognize the limitations of LLM in interpreting social meanings and clarify that our conclusions, based on probabilistic model outputs, do not construe absolute facts. Moreover, we stress that the application of LLM rationales, while beneficial within our controlled research environment for understanding human intent in utterances, should not be extended uncritically beyond these confines. The use of LLM rationales in broader contexts, especially as substitutes for human judgment and rationale, is not advocated.

References

AI@Meta. 2024. Llama 3 model card.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving machine attention from human rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An empirical investigation of statistical significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 995–1005, Jeju Island, Korea. Association for Computational Linguistics.

Milan Bhan, Jean-Noël Vittaut, Nicolas Chesneau, and Marie-Jeanne Lesot. 2024. Self-AMPLIFY: Improving small language models with self post hoc explanations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 10974–10991, Miami, Florida, USA. Association for Computational Linguistics.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pages 398–408, Bangkok, Thailand. Association for Computational Linguistics.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. *Advances in Neural Information Processing Systems*, 31.

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.

Hanjie Chen, Faeze Brahman, Xiang Ren, Yangfeng Ji, Yejin Choi, and Swabha Swayamdipta. 2023. REV: Information-theoretic evaluation of free-text rationales. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2007–2030, Toronto, Canada. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Devleena Das and Sonia Chernova. 2020. Leveraging rationales to improve human task performance. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 510–518.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Ritam Dutt, Sayan Sinha, Rishabh Joshi, Surya Shekhar Chakraborty, Meredith Riggs, Xinru Yan, Haogang Bao, and Carolyn Rose. 2021. ResPer: Computationally modelling resisting strategies in persuasive conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 78–90, Online. Association for Computational Linguistics.

Ritam Dutt, Zhen Wu, Jiaxin Shi, Divyanshu Sheth, Prakhar Gupta, and Carolyn Rose. 2024. Leveraging machine-generated rationales to facilitate social meaning detection in conversations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6901–6929, Bangkok, Thailand. Association for Computational Linguistics.

Joshua Eisenberg and Mark Finlayson. 2016. Automatic identification of narrative diegesis and point of view. In *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pages 36–46, Austin, Texas. Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

James N Farr, James J Jenkins, and Donald G Paterson. 1951. Simplification of flesch reading ease formula. *Journal of applied psychology*, 35(5):333.

Prakhar Gupta, Cathy Jiao, Yi-Ting Yeh, Shikib Mehri, Maxine Eskenazi, and Jeffrey Bigham. 2022. InstructDial: Improving zero and few-shot generalization in dialogue through instruction tuning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 505–525, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Sai Gurrapu, Ajay Kulkarni, Lifu Huang, Ismini Lourentzou, and Feras A. Batarseh. 2023. Rationalization for explainable nlp: a survey. *Frontiers in Artificial Intelligence*, 6.

Sil Hamilton. 2024. Detecting mode collapse in language models via narration. In *Proceedings of the First edition of the Workshop on the Scaling Behavior of Large Language Models (SCALE-LLM 2024)*, pages 65–72, St. Julian's, Malta. Association for Computational Linguistics.

Manoel Horta Ribeiro, Justin Cheng, and Robert West. 2023. Automated content moderation increases adherence to community guidelines. In *Proceedings of the ACM web conference 2023*, pages 2666–2676.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *Advances in Neural Information Processing Systems*, 33:20179–20191.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2021. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Sahil Jayaram and Emily Allaway. 2021. Human rationales as attribution priors for explainable stance detection. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5540–5554, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-aided annotation for fine-grained proposition types in argumentation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1008–1018, Marseille, France. European Language Resources Association.

Brihi Joshi, Aaron Chan, Ziyi Liu, Shaoliang Nie, Maziar Sanjabi, Hamed Firooz, and Xiang Ren. 2022. ER-test: Evaluating explanation regularization methods for language models. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 3315–3336, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Brihi Joshi, Ziyi Liu, Sahana Ramnath, Aaron Chan, Zhewei Tong, Shaoliang Nie, Qifan Wang, Yejin Choi, and Xiang Ren. 2023. Are machine rationales (not) useful to humans? measuring and improving human utility of free-text rationales. *arXiv preprint arXiv:2305.07095*.

Ratnesh Joshi, Arindam Chatterjee, and Asif Ekbal. 2021. Towards explainable dialogue system: Explaining intent classification using saliency techniques. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 120–127, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Sopan Khosla and Rashmi Gangadharaiah. 2022. Benchmarking the covariate shift robustness of open-world intent classification approaches. *AACL-IJCNLP 2022*, page 14.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022. ProsocialDialog: A prosocial backbone for conversational agents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 4005–4029, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jaekyeom Kim, Dong-Ki Kim, Lajanugen Logeswaran, Sungryull Sohn, and Honglak Lee. 2024. Auto-intent: Automated intent discovery and self-exploration for large language model web agents. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 16531–16541, Miami, Florida, USA. Association for Computational Linguistics.

JP Kincaid. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Chief of Naval Technical Training*.

Stefan Larson, Anish Mahendran, Joseph J Peper, Christopher Clarke, Andrew Lee, Parker Hill, Jonathan K Kummerfeld, Kevin Leach, Michael A Laurenzano, Lingjia Tang, et al. 2019. An evaluation dataset for intent classification and out-of-scope prediction. 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 1311–1316.

Michael K Lindell, Christina J Brandt, and David J Whitney. 1999. A revised index of interrater agreement for multi-item ratings of a single target. *Applied Psychological Measurement*, 23(2):127–135.

Bodhisattwa Prasad Majumder, Oana Camburu, Thomas Lukasiewicz, and Julian Mcauley. 2022. Knowledge-grounded self-rationalization via extractive and natural language explanations. In *International Conference on Machine Learning*, pages 14786–14801. PMLR.

Magdalena Markowska, Mohammad Taghizadeh, Adil Soubki, Seyed Mirroshandel, and Owen Rambow. 2023. Finding common ground: Annotating and predicting common ground in spoken conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8221–8233, Singapore. Association for Computational Linguistics.

Shikib Mehri. 2022. *Towards Generalization in Dialog through Inductive Biases*. Ph.D. thesis, Language Technologies Institute, Carnegie Mellon University.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

Melissa Mulcahy and Bethanie Gouldthorp. 2016. Positioning the reader: the effect of narrative point-of-view and familiarity of experience on situation model construction. *Language and Cognition*, 8(1):96–123.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2021. Soloist: Building task bots at scale with transfer learning and machine teaching. *Transactions of the Association for Computational Linguistics*, 9:807–824.

Baolin Peng, Chenguang Zhu, Chunyuan Li, Xiujun Li, Jinchao Li, Michael Zeng, and Jianfeng Gao. 2020. Few-shot natural language generation for task-oriented dialog. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 172–182.

Javiera Perez Gomez. 2021. Verbal microaggressions as hyper-implicatures. *Journal of Political Philosophy*, 29(3):375–403.

Cheng Qian, Bingxiang He, Zhong Zhuang, Jia Deng, Yujia Qin, Xin Cong, Zhong Zhang, Jie Zhou, Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2024. Tell me more! towards implicit user intention understanding of language model driven agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1088–1113, Bangkok, Thailand. Association for Computational Linguistics.

Kavel Rao, Liwei Jiang, Valentina Pyatkin, Yuling Gu, Niket Tandon, Nouha Dziri, Faeze Brahman, and Yejin Choi. 2023. What makes it ok to set a fire? iterative self-distillation of contexts and rationales for disambiguating defeasible social and moral situations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12140–12159, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Alvin E Roth. 1988. Introduction to the shapley value. *The Shapley value*, 1.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, et al. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022a. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022b. Neural theory-of-mind? on the limits of social intelligence in large LMs. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Minjia Wang, Pingping Lin, Siqi Cai, Shengnan An, Shengjie Ma, Zeqi Lin, Congrui Huang, and Bixiong Xu. 2025. Stand-guard: A small task-adaptive content moderation model. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 1–20.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019a. 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.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019b. 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.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. 2022. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. 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 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Sarah Wiegreffe, Ana Marasović, and Noah A. Smith. 2021. Measuring association between labels and free-text rationales. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language*

Processing, pages 10266–10284, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chien-Sheng Wu, Steven CH Hoi, Richard Socher, and Caiming Xiong. 2020. Tod-bert: Pre-trained natural language understanding for task-oriented dialogue. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 917–929.

Zhen Wu, Ritam Dutt, and Carolyn Penstein Rosé. 2024. Evaluating large language models on social signal sensitivity: An appraisal theory approach. In *The First Human-Centered Large Language Modeling Workshop*, page 67.

Weimin Xiong, Yifan Song, Peiyi Wang, and Sujian Li. 2023. Rationale-enhanced language models are better continual relation learners. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15489–15497, Singapore. Association for Computational Linguistics.

Akhila Yerukola, Saujas Vaduguru, Daniel Fried, and Maarten Sap. 2024. Is the pope catholic? yes, the pope is catholic. generative evaluation of non-literal intent resolution in LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 265–275, Bangkok, Thailand. Association for Computational Linguistics.

Zulipiye Yusupujiang and Jonathan Ginzburg. 2023. Unravelling indirect answers to wh-questions: Corpus construction, analysis, and generation. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 336–348, Prague, Czechia. Association for Computational Linguistics.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In *Human language technologies 2007: The conference of the North American chapter of the association for computational linguistics; proceedings of the main conference*, pages 260–267.

Eric Zelikman, Wanjing Ma, Jasmine Tran, Diyi Yang, Jason Yeatman, and Nick Haber. 2023. Generating and evaluating tests for k-12 students with language model simulations: A case study on sentence reading efficiency. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2190–2205, Singapore. Association for Computational Linguistics.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah Goodman. 2022. Star: Bootstrapping reasoning with reasoning. *Advances in Neural Information Processing Systems*, 35:15476–15488.

Yiming Zhang, Sravani Nanduri, Liwei Jiang, Tongshuang Wu, and Maarten Sap. 2023. BiasX: "thinking slow" in toxic content moderation with explanations of implied social biases. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4920–4932, Singapore. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

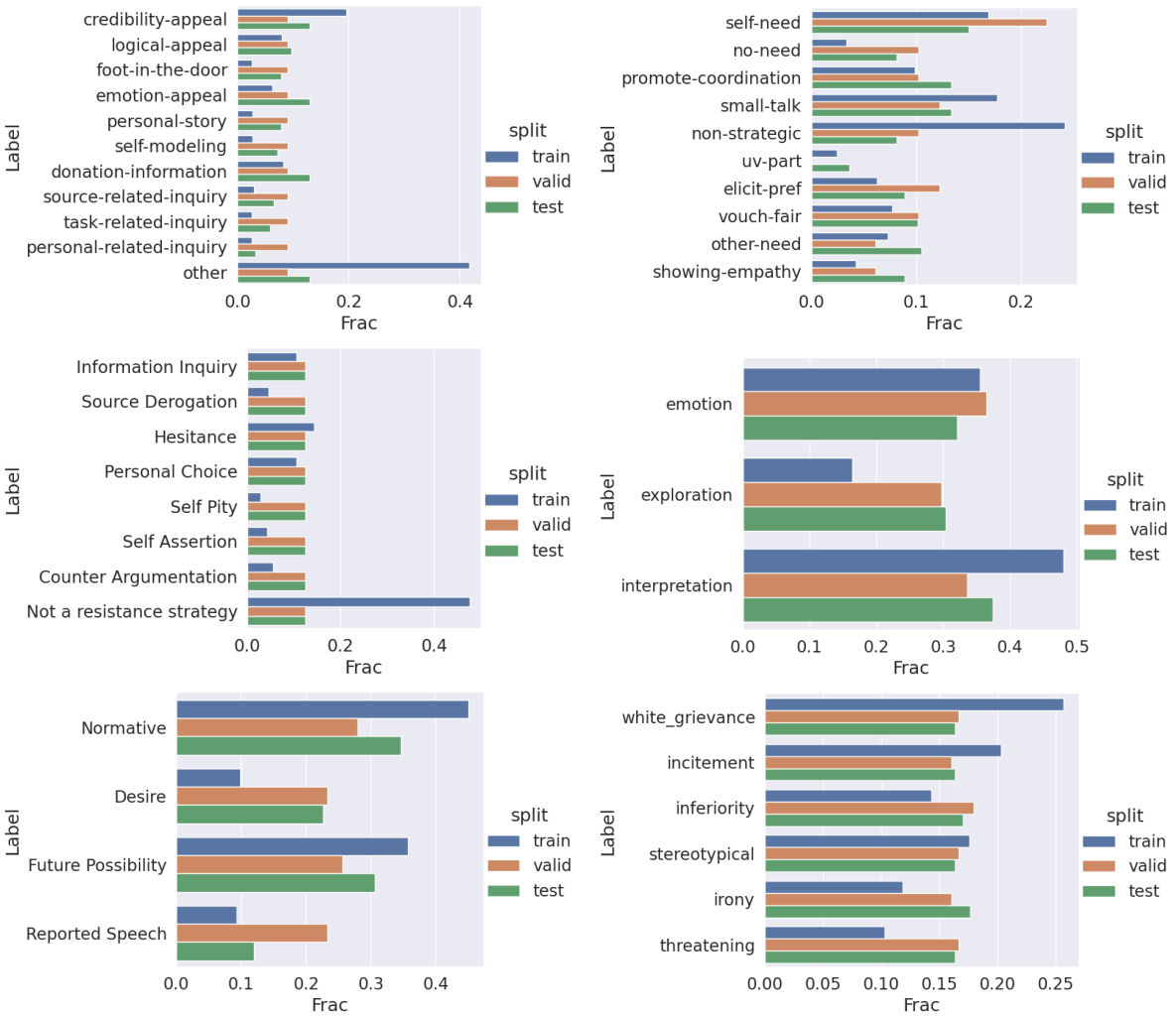


Figure 7: Distibution of labels across the different splits for the six datasets or tasks.

A Dataset Statistics

We describe in detail the six different datasets (or tasks) that we explore in this study. We showcase the distribution of the different labels across the different splits in Figure 7.

1. Persuasion - The task involves identifying persuasive strategies between two AMT workers where one adopts the role of the persuader and is expected to convince the other party (the persuadee) to donate to charity. We use the Persuasion for Good (P4G) dataset of Wang et al. (2019b).

2. Negotiation tactic - The negotiation task is grounded in the CaSiNo corpus of (Chawla et al., 2021), which consists of bargaining for campsite resources between crowd workers in a simulated camping setting. Dialogs contain various aspects of a realistic negotiation, such as building relationships, discussing preferences, exchanging offers, emotional expression, and persuasion with personal and logical arguments.

3. Resisting Strategies - Complementary to task of identifying persuasive attempts, the task proposed by Dutt et al. (2021) involves detecting resisting strategies, i.e. strategies employed to resist being persuaded by others. We focus on the Craigslist Bargain dataset (henceforth res_CB) which consists of simulated conversations between a buyer (BU) and a seller (SE) over an online exchange platform. Both are given their respective target prices and employ resisting strategies to negotiate the offer.

4. Empathy in mental health - We use the framework and dataset of Sharma et al. (2020) that characterizes the communication of empathy

Dataset	Dialog History	Speaker: UTT	INT	PreSup	HR
CaSiNo (Chawla et al., 2021)	**mturk_agent_1:** I am running low on firewood. I need more to keep the fire going and cook food.	**mturk_agent_2:** Yeah I need firewood too. It's only 50 degrees at night on the mountain, even in the summer.	Agreeing on the need for firewood and providing additional context	It gets cold at night on the mountain even in the summer.	mturk_agent_1 feels validated in their need for firewood and understands the shared predicament.
P4G (Wang et al., 2019b)	**ER:** Save the Children is an amazing charity that helps kids who are in desperate need. **ER:** They can help with safety, education and more. **ER:** You can donate some of your earnings to this amazing charity. **EE:** I believe in this charity, but still wonder how much of the money I donate actually helps. **ER:** Every little bit makes a difference.	**ER:** When you have people who are so poor, it's amazing what a tiny amount can do.	Emphasizing the value of small contributions	Even minimal financial aid can greatly benefit those in extreme poverty.	EE is encouraged by the impact of small donations.
res_CB (Dutt et al., 2021)	**Buyer:** Hi there. I was looking for ads and this one caught my attention. Is it in a good and working condition? **Seller:** It is, it's been used a lot less than its age would suggest. I only rode it a few times a month. Are you interested?	**Buyer:** Yes very much so.	Express strong interest in the item	The buyer is expressing a strong interest in the item indicating a willingness to negotiate further.	The seller feels encouraged by the buyer's interest and is prompted to discuss the price.
IMP_HATE (ElSherief et al., 2021)		**Poster:** flynn's resignation set a dangerous precedent for the administration . #sessions is essential to justice 4 white america . he must not #resign	Expressing concern about the implications of Flynn's resignation and emphasizing the importance of Sessions to their view of justice for white America while urging that Sessions should not resign.	The resignation of a high-ranking official can have significant and potentially negative consequences for the administration.	The hearer may feel concerned about the implications of Flynn's resignation and the potential impact on the administration. They may also feel a sense of urgency or importance regarding Sessions' role and the need for him to remain in his position.
EMH (Sharma et al., 2020)	**Seeker:** Why do I always have good news followed by a shit night, followed by sitting up at 2am wanting to kill myself? Why is life so difficult? Why is it so impossible to be fucking happy for once in my shit fucking life? What's the point anymore?	**Responder:** well not for nothing but you made it extremely difficult to read your post by only using a period in the title. JUST saying not judging.	Pointing out the difficulty in reading the post due to formatting while attempting to clarify that they are not judging.	Clear communication is important for understanding and responding to others' concerns effectively.	The Seeker may feel invalidated or criticized as the Responder's comment focuses on the format of the post rather than addressing the Seeker's emotional distress.
PROP (Jo et al., 2020)	**S_1:** It is called the Constitution of the United States **S_2:** unfortunately, those few months gave us OBAMA **S_3:** We're going to win when we unite people with a hopeful, optimistic message **S_3:** we had high sustained economic growth	**S_3:** We created 1.3 million jobs	Emphasizing job creation	Creating jobs is a positive achievement.	Impression of job creation success

Table 5: Examples of rationales generated by GPT-4o for six utterances, each coming from a different dataset and task. For each utterance, we provide the dialog history and the corresponding intention, presupposition, and hearer reaction abbreviated as INT, PreSup, and HR respectively. The rationales score high on factuality, soundness, and relevance as evaluated by two annotators.

Table 6: Description of the resisting strategies used in our work for the res_CB (Dutt et al., 2021). Examples of each strategy are italicised.

Resisting Strategy	**Description**
Source Derogation	Attacks the other party or questions the item *Was it new denim, or were they someone's funky old worn out jeans?*
Counter Argumentation	Provides a non-personal argument/factual response to refute a previous claim or to justify a new claim. *It may be old, but it runs great. Has lower mileage and a clean title.*
Personal Choice	Provides a personal reason for disagreeing with the current situation or chooses to agree with the situation provided some specific condition is met. *I will take it for $300 if you throw in that printer too.*
Information Inquiry	Requests for clarification or asks additional information about the item or situation. *Can you still fit it in your pocket with the case on?*
Self Pity	Provides a reason (meant to elicit sympathy) for disagreeing with the current terms. *$130 please I only have $130 in my budget this month.*
Hesitance	Stalls for time and is hesitant to commit; specifically, they seek to further the conversation and provide a chance for the other party to make a better offer. *Ok, would you be willing to take $50 for it?*
Self-assertion	Asserts a new claim or refutes a previous claim with an air of finality/ confidence. *That is way too little.*

Table 7: Description of the negotiation strategies used in our work for Casino (Chawla et al., 2021). Examples of each strategy are italicised.

Negotiation Label	**Description**
self-need	Participant argues for creating a personal need for an item in the negotiation. *Yes. I'm actually taking a large group of people. Some friends and family are going and I kind of also wanted a bit of extra firewood.* :)
no-need	Participant points out that they do not need an item based on personal context. *I don't like food. my stomach is always full. I only drink water since im thirsty most of the time.*
promote-coordination	Participant promotes coordination between the two partners. *Alright so I think we can make a fair deal here where we both will be happy.* :)
small-talk	Participant engages in small talk while discussing topics apart from the negotiation in an attempt to build a rapport. *My mistake, hypothermia is messing with my brain.*
uv-part	Participant undermines the requirements of their opponent. *I understand that atleast you are going to be close to water, that will be our most important thing since we will be thirsty and you know kids and trying to tell them to ration the water...LOL*
elicit-pref	Participant provides an attempt to discover the preference order of the opponent *I get that and understand completely. I have a large number of mouths to feed making the food a necessity or all the firewood to cook whatever we hunt. How many you have?*
vouch-fair	Participant announces a callout to fairness for personal benefit, either when acknowledging a fair deal or when the opponent offers a deal that benefits them *hey buddy I hope we both end up with a good deal* :)
other-need	Participants discuss a need for someone else rather than themselves. *I would be willing to do that if I could have two of the waters? I didn't bring as much as I thought I would need because I forgot I would have my dog.*
showing-empathy	Participant positively acknowledges or displays empathetic behavior towards a personal context of the partner. *Are you sure that's enough firewood for you and the baby? I know that babies can easily get very sick from dropping temperatures.*
non-strategic	Utterance does not have any strategic element *oh well that's fantastic, so let me know what your deal is*

Table 8: Description of the different dimensions of empathy used in our work for EMH (**?**). Examples of each strategy are italicised.

Empathy Dimension	**Description**
emotion	Responder specifies the experienced emotions explicitly or communicates emotional reactions without the emotions being explicitly labeled. *I love you all. Genuinely the replies here made me feel much less alone. your kindness is beautiful*
interpretation	Responder specifies the inferred feeling or experience, or communicates understanding through descriptions of similar experiences, or mentions that they understand the situation. *I struggle with the same thing, but with my step dad. I've basically given up on it.*
exploration	Responder seeks to explore the specific experiences and feelings of the supporter, which can be specific to the situation or generic. *What kind of connections do you want / need?*

Table 9: Description of the argumentation labels used in our work for PROP (Jo et al., 2020). Examples of each strategy are italicised.

Argumentation Label	**Description**
Normative	A normative proposition is defined as a proposition where the speaker or someone else proposes that a certain situation should be achieved or that an action should be carried out. *Americans have a right to know if there are any conflicts of interest*
Desire	A desire proposition is defined as a proposition that explicitly claims that the speaker or someone else desires to own something, do something, or desires for a certain situation to be achieved. *Bernie Sanders' soft stance on guns is not what Americans want to hear right now*
Future Possibility	A future possibility proposition claims a possibility or prediction that something may be the case in the future. *the next president of the US could very well be a reality TV star*
Reported Speech	A reported speech proposition must convey an explicit or implicit predicate borrowed from a source external to the speaker. *Clinton's view that sometimes capitalism needs controls in order to deal with inequality were a lot more realistic*

Table 10: Description of the argumentation labels used in our work for IMP_HATE (ElSherief et al., 2021). Examples of each strategy are italicised.

Type of Hate	**Description**
white_grievance	The message includes frustration over a minority groups perceived privilege and casting majority groups as the real victims of racism. This language is linked to extremist behavior and support for violence *is zuma talking about seizing white land or not ? are you denying it ?*
incitement	The message includes flaunting in group unity and power or elevating known hate groups and ideologies. *india wud be a peaceful prosperous n progressive country only when each of its citizen is a hindu.that shud b d endall objective 4 d state*
inferiority	The message implies one group or individual is inferior to another, and it can include dehumanization, denial of a person's humanity, and toxic language that compares the target with disease, insects, animals . Related to assaults on human dignity, dominance, and declarations of superiority of the in group *have you ever written about islam being a crutch for a failed civilization? im new to your work, just wondering.*
stereotypical	The message associates a protected class with negative attributes such as crime, or terrorism or includes misinformation that feeds stereotypes and vice versa, like holocaust denial and other forms of historical negationism. *You can't be a person of colour; you're too pale!*
irony	The message uses sarcasm, humor, and satire to attack or demean a protected class or individual. *What's the one good thing about black people? They provide jobs for the prison guards*
threatening	The message conveys a speaker's commitment to a target's pain, injury, damage, loss or violation of rights, threats related to implicit violation of rights and freedoms, removal of opportunities, and more subtle forms of intimidation. *We have this huge military. Why don't we just go down there and create an ethno-state for whites. Most of the blacks weren't even there when South Africa was founded by whites!*

Table 11: Description of the persuasion labels used in our work for P4G(Wang et al., 2019b). Examples of each strategy are italicised.

Persuasion Label	**Description**
credibility-appeal	Refers to the uses of credentials and citing organizational impacts to establish credibility and earn the persuadee's trust *It is the worlds first global charity for children, and have credentials to back them up.*
logical-appeal	Refers to the use of reasoning and evidence to convince others. *You are donating money you don't even have yet so it is not like you are missing something.*
foot-in-the-door	Refers to the strategy of starting with small donation requests to facilitate compliance followed by larger requests." *Are you sure, you can do as little as 5 cents???*
emotion-appeal	Refers to the elicitation of specific emotions to influence others in the form of story-telling, empathy, guilt, or anger" *It broke my heart to see that famous photograph of a child with a vulture sitting next to it.*
personal-story	Refers to the strategy of using narrative exemplars to illustrate someone's donation experiences or the beneficiaries' positive outcomes, which can motivate others to follow the actions." *I have three children myself, and the welfare of children around the world is a very important cause to me.*
self-modeling	Refers to the strategy where the persuader first indicates their own intention to donate and chooses to act as a role model for the persuadee to follow" *I think I am going to give a small portion of my hit payment to save the children.*
donation-information	Refers to providing specific information about the donation task, such as the donation procedure, donation range, etc." *The research team will collect all donations and send it to Save the Children.*
source-related-inquiry	Asks about the persuadee's opinion and expectation related to the task." *Iḿ alright, just reading up on this organization called "Save the Children".. have you heard about it?*
task-related-inquiry	Asks if the persuadee is aware of the organization (charity) *Do you need more info about this program?*
personal-related-inquiry	Asks about the persuadee's previous personal experiences relevant to charity donation" *I imagine hospitals are very strict about who gets to be with the little ones.*
other	Does not conform to any persuasion category *I am homeless and at Mcdonalds on the wifi.*

in text-based conversations. The task involves detecting different dimensions of empathy in text-based mental health support, i.e., empathy expressed or communicated by peer supporters in their textual interactions with seekers.

5. Argumentation - We formalize the task of argumentation into identifying different kinds of proposition in rhetorical debates. We use the data set of Jo et al. (2020) which consists of four categories of propositions: normative statements, desires statements, statements about future possibilities, and reported speech.

6. Implicit Hate Speech Detection - The task involves identifying different categories of covert or indirect language that disparages a particular individual or group based on certain protected attributes (ElSherief et al., 2021). Some instances include irony, inferiority language, and incitement to violence, among others.

We also provide descriptions of the label categories for each dataset along with an example of each for res_CB, Casino, EMH, PROP, IMP_HATE, and P4G in the Tables 6, 7,8, 9, 10, and 11 respectively.

B Prompting Framework Description

SOCIAL SCAFFOLDS takes as input a multiparty dialog and generates rationales on an utterance-by-utterance basis. This is achieved using a Large Language Model (such as GPT-4o) that goes over each utterance in the conversation and generates the corresponding rationale. We instruct the framework to generate the outputs in a structured format, i.e. the rationales are generated in the form of a CSV file and aligned with the corresponding speaker and utterance index. These checks and measures help ensure that each utterance has a corresponding rationale and enables us to revisit erroneous cases. We address those misaligned dialogs by simply prompting the framework to regenerate the rationales for those dialogs in an iterative fashion. We stop after 3 iterations.

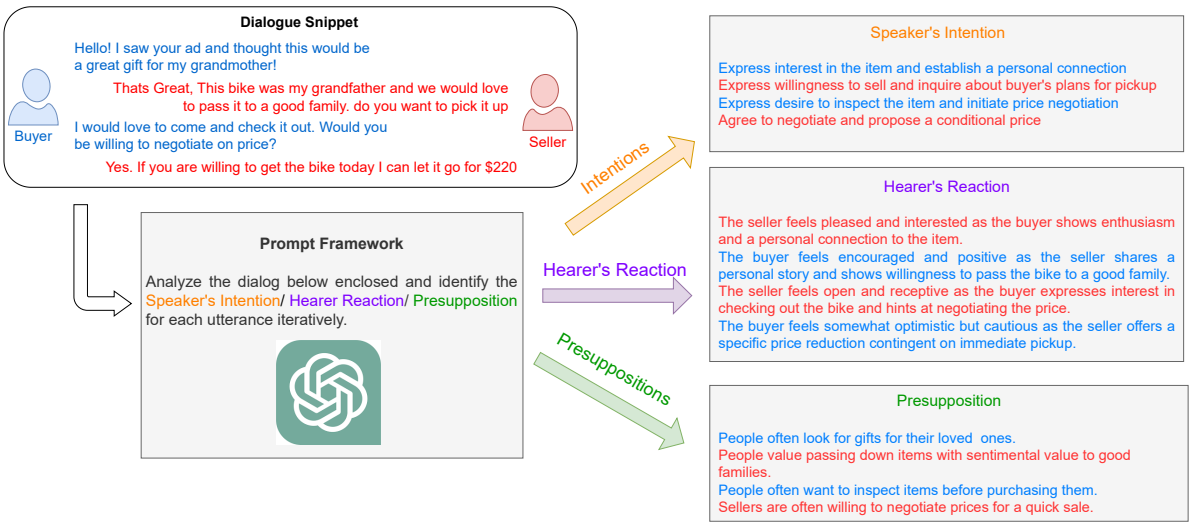


Figure 8: An overview our rationale generation framework SOCIAL SCAFFOLDS. We present a dialogue snippet between a buyer and a seller, shown in blue and red. We prompt an LLM with the dialogue snippet to generate the speaker's intentions, the hearer's reaction, and the presuppositions in orange, purple, and green, respectively.

Prompt for generating rationales

System message: "You are a helpful agent that produces consistent and structured output."
User message: "Analyse the dialogue below enclosed within the <dialog> and </dialog> tags and identify the speaker's intentions for each utterance. Go over the entire conversation on an utterance-by-utterance basis without grouping or skipping, and generate the corresponding speakers' intentions for each utterance iteratively. Return the result in the format of an csv file, with the headers corresponding to the following columns: i.e. utterance_idx, speaker, and intentions."

Figure 9: The prompt we pass to our framework to generate the rationales of a corresponding category.

We reuse the prompting framework to generate each category of rationale separately. The motivation for our design choice is two-fold. Firstly, we wish to observe whether the different rationale categories can capture distinct concepts; by forcing the framework to generate the rationales together would make it sensitive to ordering effects, for e.g. if the intentions are generated first, then those intentions would influence the generation of presuppositions. Secondly, our framework is easily generalizable to new categories of rationales.

Additionally, we do not provide any few-shot instances for in-context learning while generating these rationales to avoid biasing the generations with previously seen examples as in Dutt et al. (2024). Overall, our framework enables us to compare and contrast not only different categories of rationales with each other but also the same categories of rationales generated by different LLMs. We show an example of the prompt to generate the rationales in Figure 9.

C Annotation Guidelines

C.1 Metrics for Annotating Rationales

Since our framework automatically generates rationales without any human supervision, we develop a rigorous annotation framework to assess the validity of generations. To validate the quality of rationales, we define the following three criteria: soundness, informativeness, and relevance.

Soundness: Soundness reflects whether the rationale adheres to the definition provided during prompting, i.e. whether the generated rationale reflects the speaker's intentions, the hearer's reactions, and the presuppositions about the world. In some cases, the rationale generated might not contain any additional subtext beyond the literal rephrasing of the utterance. Such instances are scored high on soundness.

Informativeness: The information conveyed by the rationales should comply with the context of the current dialogue. The information should be cor-

Table 12: Instances of annotator disagreement for the different datasets

Dataset	Utterance	Rationale	Category	A	B
		Relevance Disagreements			
IMP_HATE	#dayafterchristmas in america during the current year . join the #altright white people . #malls	There is a significant cultural or social event related to the day after Christmas in America that is relevant to the alt-right and white people and it involves malls.	PreSup	2	3
PROP	CLINTON would ask them then draw their own conclusion	Clinton believes voters should make an informed decision based on her record.	PreSup	2	3
casino	I completely understand where you are coming from. How much food do you need?	mturk_agent_1 feels understood and is prompted to specify needs.	HR	2	3
P4G	Almost 87% of all dollars donated go directly the programs supporting the children, so they have a very low overhead.	A high percentage of donations go directly to the programs indicating low administrative costs.	PreSup	3	2
		Informativeness Disagreements			
res_CB	Look it up, these are very high end bikes	Encouraging the buyer to verify the information	INT	3	2
EMH	Be strong my friend, Everytime I come in here, I don't feel so alone. This may sound selfish, but it's good for me to know that there's people out there that feels the pain that I feel every single day. Anyway, have a great day, my friends.	Offering encouragement and sharing personal experience to provide comfort	INT	3	2

rect, i.e. rationale should not exhibit hallucination, (present additional information that has not been encountered so far in the dialogue), and complete, i.e. they should not omit important information that could change the meaning of the utterance.

Relevance: A rationale is relevant when it goes beyond the utterance text and presents information that is not only factual and sound but also provides additional subtext. We include this metric to assess whether the rationale is useful or not for the current scenario by providing important information or cues that are not directly observable.

We score each rationale based on soundness, informativeness, and relevance using a Likert scale of 1 to 3, with 1 being the lowest and 3 the highest. The evaluations were carried out by two annotators with a graduate level proficiency in English and at least five years of experience in computational linguistics and NLP. Due to the highly subjective nature of the task, we relied on these professional annotators as an alternative to crowd-sourcing or employing an automated annotation framework. We also follow the appropriate protocols to assure the annotation and data aligned with institutional approval guidelines.

We compute inter-rater reliability scores (IRR) using the multi-item agreement measure of Lindell et al. (1999) following prior work of Dutt et al. (2024) and observe moderate to strong agreement scores for all three criteria: soundness (0.983), informativeness (0.763), and relevance (0.697).

We present a detailed breakdown of the mean Likert scores and the corresponding measure of IRR agreement for the three different categories of rationales in different dimensions in Table 13. We observe that the intention rationale has the lowest score on both informativeness and relevance. However, the rationale that exhibits the highest disagreement is the presuppositions on the relevance metric.

We inspect the disagreement cases between annotators and present some instances in Table 12. We showcase examples of disagreement for both informativeness and relevance. Since the IRR agreement were the lowest for (i) INTs on the informativeness metric and (ii) the presuppositions on the relevance metric, we have more instances of those categories in our Table.

We observe that annotator B was more critical of the annotation framework; they honed in on specific cues surface cues to illustrate why the rationale is relevant. For example, the phrase "significant cultural and social event" provides an additional subtext in the first instance. Likewise, terms such as "Clinton believes" or "feels understood" expresses additional emotions that were absent from the utterance. On the other hand, annotator A judged the rationale as relevant if it introduced new information. They were also a bit more relaxed in critiquing the informativeness score, rating the rationale to be highly informative if it was able to capture the essence of the utterance. However, annotator B rated the two intentions in Table 12 with a score of two, because the rationales had omitted

Mean Likert Scores			
Metric	INT	HR	PreSup
Soundness	3.00	2.85	3.00
Informativeness	2.62	2.72	2.93
Relevance	2.43	2.67	2.72
IRR Score			
Metric	INT	HR	PreSup
Soundness	1.00	0.95	1.00
Informativeness	0.70	0.80	0.82
Relevance	0.78	0.86	0.51

Table 13: Annotation results for the different types of rationales based on different criterion.

specific information such as the "high-end price of the bike" in the former case and because the term "personal experience" was an overgeneralization of the responder's experience for the latter.

C.2 Flowchart for Scoring Rationales

We present the flowchart for annotating rationales according to soundness, informativeness, and relevance.

Step 1: Read the dialogue history, utterance and the rationale; start with judging the Speaker Intention rationale. Perform Steps 2-4 for the Speaker Intention rationale and then reiterate for Hearer Reaction and Presuppositions.

Step 2: Check for Soundness criteria if the generated rationale encapsulates the meaning of the rationale category. When checking for Speaker Intention rationales, see if it is about the speaker's beliefs, goals, objectives, outcomes. When checking for Hearer Reaction see if it is about the belief of the hearer or their interpretation. When checking for Presuppositions see if it reflects the general world view or the assumptions shared by the participants.

- If the rationale is ascribing the correct perspective, we assign a 3 to Soundness.
- If the perspective appears to be ambiguous, we assign 2 for Soundness.
- If the perspective is blatantly incorrect, for example the Hearer Reaction actually reflects the speaker's intentions we assign 1 to Soundness.
- If Soundness is 1 all criteria should be assigned 1, since it does not make sense to evaluate a wrong rationale.

Step 3: We now check whether the rationale is Informative or not, i.e. whether the information present in the rationale is accurate.

- If all the details have been carried over from the utterance, with an appropriate level of generalization assign a 3 to Informativeness.
- If the generalization has omitted some information/details that are important to the meaning of the utterance, assign a 2 for Informativeness.
- If the rationale hallucinates information, i.e. presents information that cannot be inferred from the current dialogue context, or is otherwise just wrong, assign a 1 for Informativeness.

Note that Informativeness and Relevance are always 1 when the Soundness is 1.

Step 4: We finally check for Relevance.

- If the utterance has a subtext and the rationale has identified a subtext not overtly stated in the utterance text, assign a 3 for Relevance.
- If the rationale includes information that appears earlier in the dialogue history whether it is subtext or not, but is not in the particular utterance, assign a 3 for Relevance.
- If the utterance lacks subtext, but the rationale presents an expression or action not found in the utterance, such as expressing agreement or an opinion, assign a 3 for Relevance.
- If the utterance lacks subtext and the rationale simply summarizes the details of the given utterance without adding anything new at all, assign a 2 for Relevance.
- If the utterance has an underlying subtext but that is not captured by the rationale, or an incorrect subtext is present, assign a 1 for Relevance.

D Additional Experiments and Results

D.1 Supervised Full Fine-tuning Setup

Indomain Results: We present additional results of our supervised instruction-tuning experiments

Generator	Rationale	P4G	CaSiNo	res_CB	PROP	EMH	IMP_HATE
GPT-4o	UTT	69.70 +/- 2.42	71.22 +/- 1.70	66.77 +/- 1.02	82.38 +/- 1.21	**90.91 +/- 0.13**	62.68 +/- 0.79
	+ INT	69.36 +/- 1.45	**72.35 +/- 0.50**	**70.91 +/- 0.71**	**84.66 +/- 1.07**	89.35 +/- 1.35	**67.91 +/- 1.49**
	+ HR	**70.54 +/- 1.70**	71.71 +/- 0.84	68.80 +/- 0.97	82.88 +/- 1.69	90.26 +/- 0.32	65.08 +/- 0.34
	+ PreSup	68.12 +/- 2.30	71.81 +/- 1.39	69.69 +/- 1.51	80.11 +/- 2.86	89.37 +/- 0.16	62.88 +/- 2.55
GPT-3.5-turbo	UTT	69.70 +/- 2.42	71.22 +/- 1.70	66.77 +/- 1.02	82.38 +/- 1.21	**90.91 +/- 0.13**	62.68 +/- 0.79
	+ INT	67.64 +/- 3.16	**72.35 +/- 0.38**	**71.22 +/- 3.03**	81.52 +/- 1.47	90.01 +/- 1.12	62.82 +/- 0.62
	+ HR	68.90 +/- 1.54	71.95 +/- 2.67	70.87 +/- 1.17	**83.61 +/- 2.00**	89.18 +/- 0.73	**64.16 +/- 0.97**
	+ PreSup	**72.21 +/- 0.25**	70.43 +/- 1.27	69.28 +/- 1.45	78.61 +/- 2.97	90.00 +/- 0.96	59.85 +/- 0.52

Table 14: Performance of FLAN-T5 model in an in-domain setting across six tasks. The baseline includes only the utterance (UTT), which we compare against the three kinds of rationales, i.e. intentions (INT), hearer-reactions (HR), and presuppositions (PreSup). We represent the mean and standard deviation across three runs.

Figure 10: Impact of rationales on cross-task performance for instruction-tuned models across the six datasets for different fewshot settings using the GPT-4o generated rationales.

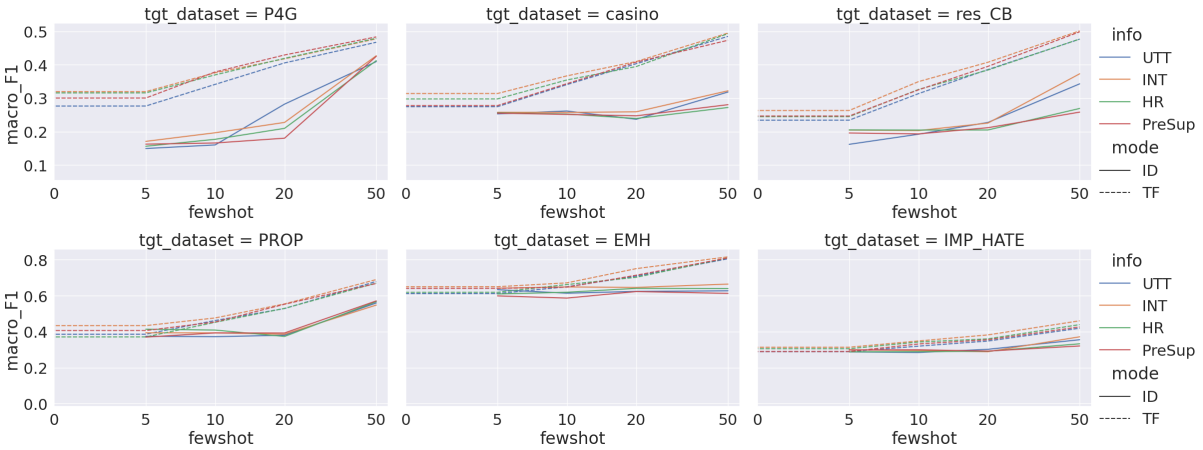


Figure 11: Impact of rationales on cross-task performance for instruction-tuned models across the six datasets for different fewshot settings using the GPT-3.5-turbo generated rationales.

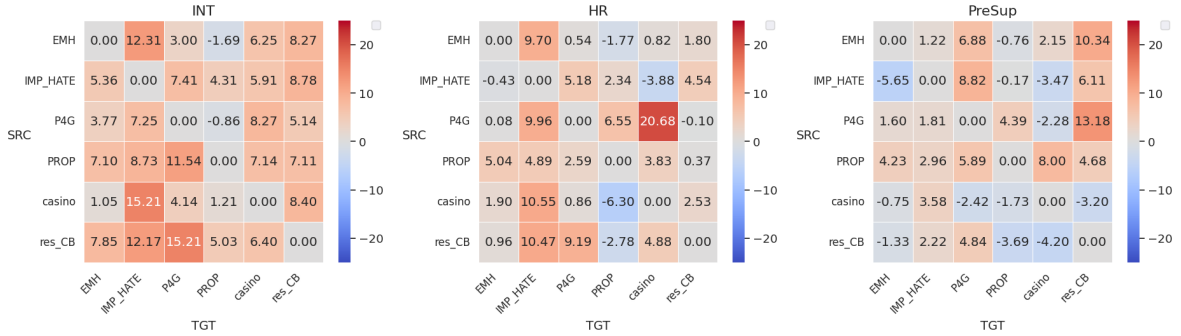


Figure 12: Relative change in performance measured in terms of F1 score over the baseline when incorporating the GPT-4o generated rationales for different source and target pairs for the cross-task transfer setting.

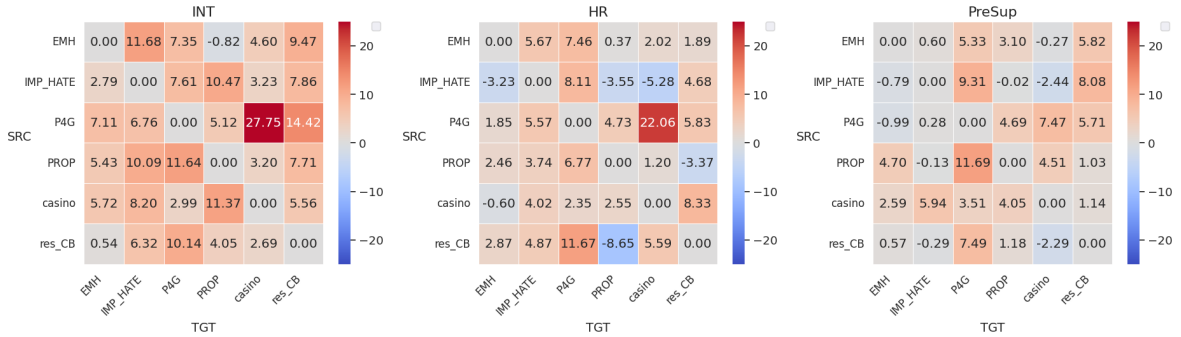


Figure 13: Relative change in performance measured in terms of F1 score over the baseline when incorporating the GPT-3.5-turbo generated rationales for different source and target pairs for the cross-task transfer setting.

in this section. Table 14 showcases the impact of adding rationales i.e. the intentions, hearer reactions, and presuppositions, generated by GPT-4o and GPT-3.5-turbo LLMs on the six datasets using the FLAN-T5 model. We see that apart from the EMH dataset, adding in the rationale improves performance for a majority of the cases.

Cross Task Results: We present the results of our cross task transfer experiments using the FLAN-T5-base model augmented with rationales generated by GPT-4o and GPT-3.5-turbo in Figures 10 and 11 respectively. We observe significant gains over the utterance (or the baseline case) when rationales are added for different datasets and few-shot settings.

We also inspect which category of rationales are the most effective for a given source and target pair in Figures 12 and 13 respectively by the net relative improvement in F1 score across different few-shot settings. We observe that for the intentions rationales, transfer almost always yields a positive relative improvement for any source and target pair, showcasing their effectiveness across different tasks. After intentions, we observe that the hearer reactions have the most impact followed by presuppositions.

Table 15: Spearmann's rank correlation between model lists for the source and target.

Dataset	Instances	Rationales
P4G	-0.04	0.46
CaSiNo	-0.07	0.00
res_CB	0.01	0.06
PROP	0.15	0.18
EMH	-0.02	-0.15
IMP_HATE	0.07	0.28

Table 16: Performance of PEFT-based LLama model for different datasets when augmented with rationales corresponding to intentions, hearer reactions, and presuppositions. We present the mean performance and standard deviation across three seeds.

Rationale	P4G	res_CB	IMP_HATE
UTT	69.4 +/- 1.5	71.5 +/- 2.6	66.5 +/- 0.6
+ INT	71.2 +/- 1.6	71.3 +/- 1.9	66.0 +/- 2.0
+ HR	**72.6 +/- 1.8**	**72.8 +/- 1.8**	68.1 +/- 1.1
+ PreSup	66.6 +/- 2.7	68.7 +/- 2.0	**68.6 +/- 1.4**

Figure 14: Impact of rationales on both in-domain and cross-task performance for PEFT-based LLama models across the three datasets for different few-shot settings. We use the rationales generated by GPT-4o

D.2 PEFT-based Fine-tuning Setup

We also explore the impact of adding rationales in a PEFT-based fine-tuning setup. We fine-tune a pre-trained LLama-3-8B-it (AI@Meta, 2024) with 4-bit double quantization and low-rank adapter (LoRA) to ensure efficient fine-tuning (Hu et al., 2021; Dettmers et al., 2024).

Due to the limited compute budget and the large number of experiments (360 in-domain and 1440 cross-task transfer runs) for a single SFT model, we experiment on only three out of six datasets, i.e. P4G, res_CB, and IMP_HATE. We chose these three datasets because they had the lowest performance in the in-domain setting.

We present the in-domain results in Table 16 and the cross-task transfer performance in Figure 14. We observe trends similar to our instruction-tuned results, i.e. rationales aid dialogue understanding and generalization for PEFT based models.

D.3 Performance against SOTA baselines

We compare the performance of our baseline, i.e. FLAN-T5 in the in-domain setting without any rationale information, against the previous reported SOTA performance (which were mostly trained on BERT based models) on all datasets as reported in their original paper. It is evident from Table 17 that our FLAN-T5 serves as a competitive baseline and achieves higher performance (in terms of macro F1 score) on all six tasks.

D.4 ICL Results

We note the effect of adding rationales for different in-context learning settings. We experiment with LLama-3-8B-it and Gemma-2-9B-it as the ICL LLM, and prompt them for different few-shot settings, i.e. 0-shot, 2-shot, and 5-shot. We present these results in Table 18 where we observe that adding rationales generally yielded higher performance over the baseline (i.e. using only the UTT). We observe that performance mostly plateaus at the 2-shot setting.

Table 17: Performance of our FLAN-T5 model against previous SOTA performance.

Dataset	FLAN-T5	Repoted SOTA
P4G	**69.7**	59.6
res_CB	66.8	66.2
CaSiNo	**71.2**	68.3
PROP	**83.4**	72.1
EMH	**90.9**	69.9
IMP_HATE	**62.7**	58.6

We also explore the impact of adding rationales generated by different LLMs, i.e. GPT-4o, GPT-3.5-turbo, and Gemma-2-27B-it, in Table 19 and note similar performance in all three cases, highlighting that the rationales generated by open-source models aid downstream task performance similar to proprietary models.

E Experimental Details and Hyper-Parameter Tuning

We present the hyperparameters for our experiments in Table 20. We carry out the experiments over 3 seeds on a A6000 GPU with early stopping with patience of 5 over the validation set for all experiments. We implement the entire experiments in Python, with help of the Pytorch library and use the pre-trained models as specified in Huggingface under the agreed upon license agreements. We explicitly specify the software libraries and their corresponding versions in Table 21

Our experimental suite comprises encompasses 6 datasets in the indomain setting for the FLAN-T5 models for 5 few-shot settings (5, 10, 20, 50, and all) across 3 seeds and for 9 cases, corresponding to the 3 types of rationales individually (INT, HR, PreSup) and combined (i.e. ALL), for each of the two LLMs (GPT-3.5-turbo and GPT-4o) and the baseline (UTT). Furthermore, for a model pre-

Table 18: Performance for in-context learning models for different datasets and few-shot settings aggregated over different rationale categories generated by different LLMs, i.e. GPT-4o, GPT-3.5-turbo, and Gemma-2-27B-it.

		Gemma-2-9B-it						LLama-3-8B-it					
Rationale	#fshot	P4G	casino	res_CB	PROP	EMH	HATE	P4G	casino	res_CB	PROP	EMH	HATE
UTT	0	30.23	36.87	33.92	46.36	51.33	35.16	20.55	30.08	26.44	44.28	**67.72**	32.05
+ INT	0	**32.95**	38.08	**35.31**	**49.23**	54.47	**36.46**	**21.23**	30.62	28.46	**46.63**	67.4	32.54
+ HR	0	28.27	36.27	32.7	43.69	55.26	35.02	20.64	30.5	**28.82**	46.11	65.24	32.18
+ PreSup	0	30.41	38.34	35.15	43.34	50.16	35.17	20.1	**30.99**	26.98	43.94	67.25	32.33
+ ALL	0	32.78	**40.4**	34.23	46.73	**55.75**	35.89	21.06	29.36	27.43	44.89	67.57	**32.73**
UTT	2	36.24	38.69	39.67	46.72	60.82	35	**24.64**	30.96	29.17	41.85	64.73	**30.74**
+ INT	2	37.85	**39.75**	**45.01**	49.06	66.39	**37.43**	21.87	**33.17**	**33.92**	**44.07**	65.61	30.58
+ HR	2	37.86	38.89	39.56	47.87	61.31	33.63	22.5	30.54	30.98	41.75	64.37	29.75
+ PreSup	2	36.21	37.61	41.58	48.24	58.7	36.31	24.11	30.93	30.4	41.82	61.59	29.35
+ ALL	2	**38.48**	39.4	43.38	**49.69**	**66.77**	36.61	21.72	31.1	30.56	42.32	**66**	29.91
UTT	5	37.72	39.33	38.23	46.2	60.51	35.66	20.59	29.12	27.91	41.81	**66.58**	29.58
+ INT	5	37.3	**39.96**	**43.23**	49.8	63.42	37.19	19.52	29.41	**32.64**	**43.44**	64.87	29.57
+ HR	5	**38.02**	39.57	38.67	48.91	61.4	35.16	**20.81**	**29.82**	31.42	44	65.42	29.57
+ PreSup	5	36.11	37.6	39.39	46.86	64.55	34.65	20.36	29.25	32.29	43.29	63.75	29.57
+ ALL	5	36.34	36.9	40.67	**53.26**	**66.03**	**37.36**	19.31	29.38	29.88	43.43	63.54	29.57

Table 19: Performance for in-context learning models for different datasets and few-shot settings aggregated over different few-shot settings.

		Gemma-2-9B-it						LLama-3-8B-it					
Rationale	LLM	P4G	casino	res_CB	PROP	EMH	HATE	P4G	casino	res_CB	PROP	EMH	HATE
UTT	-	34.73	38.3	37.27	46.42	57.55	35.27	21.93	30.05	27.84	42.65	66.34	30.79
+ INT	gpt-4o	35.31	**40.28**	**41.88**	**48.91**	62.35	**37.57**	21.56	**31.55**	30.89	**45.16**	**66.88**	**30.98**
+ HR	gpt-4o	32.91	37.33	37.92	44.73	61.62	35.56	**22.30**	30.03	31.16	43.33	66.54	30.69
+ PreSup	gpt-4o	35.13	37.89	39.48	47.65	58.79	36.63	21.3	31.1	**31.52**	44.23	65.38	30.62
+ ALL	gpt-4o	**36.75**	39.52	39.18	47.19	**62.50**	36.93	21.11	29.93	29.77	43.5	66.91	30.51
+ INT	gpt-3.5	**36.48**	**39.74**	**41.06**	**51.84**	62.52	**36.65**	19.92	**31.26**	**32.89**	**44.87**	64.46	31.18
+ HR	gpt-3.5	35.96	35.72	35.85	48.08	57.23	34.45	19.82	30.38	30.43	43.43	65	30.32
+ PreSup	gpt-3.5	32.88	38.5	38.87	45.25	56.59	34.7	20.8	29.45	29.2	41.66	63.18	30.48
+ ALL	gpt-3.5	34.84	37.9	39.3	50.78	**63.75**	36.01	19.1	29.48	28.27	43.72	64.99	**31.27**
+ INT	Gemma	**36.32**	37.78	**40.61**	47.35	59.42	36.85	21.14	30.39	**31.24**	44.1	**66.54**	30.52
+ HR	Gemma	35.28	**41.68**	37.16	47.67	59.12	33.8	21.83	30.45	29.63	**45.1**	63.48	30.49
+ PreSup	Gemma	34.72	37.16	37.78	45.54	58.03	34.8	**22.48**	**30.61**	28.96	43.16	64.04	30.16
+ ALL	Gemma	36.02	39.29	39.81	**51.71**	**62.31**	**36.91**	21.88	30.43	29.83	43.41	65.21	30.44

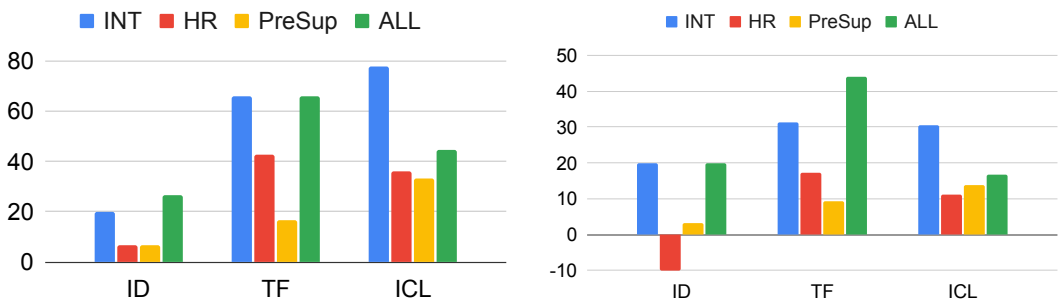


Figure 15: Proportion of cases adding rationales improve performance overall (left) and significantly (right) for different settings

Table 20: Hyperparameters used for fine-tuning the FLAN-T5-base model for all the experiments.

Hyperparameter	**Value**
SFT- Instruction Tuned Setup	
Max sequence length	1024
Learning rate	$2e^{-5}$
Batch size	8
Num. epochs	10
Optimizer	Adam
Patience	5
Seeds	3
Model	`FLAN-T5-base`
SFT- PEFT Setup	
Quantization	4-bit double
Precision training	bf16
LoRA reduction factor	64
LoRA dropout	0.05
LoRA Alpha	32
Batch size	4
Weight decay	0.01
Learning Rate	$2e^{-5}$
Max sequence length	1024
Num. epochs	10
Patience	5
Seeds	3
LLM	`LLama-3-8B-it`
ICL	
Temperature	0.9
Fewshot examples	[0, 2, 5]
Batch size	8
GPUs	A6000 *2

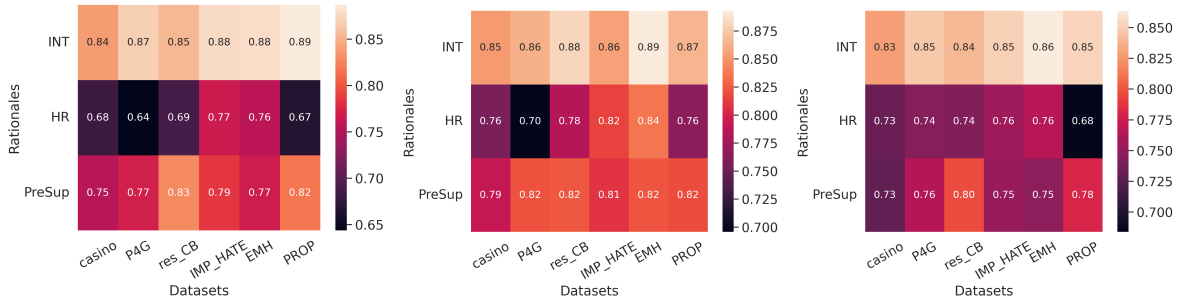
Table 21: Versions of Library used in our work.

SFT + ICL setup	
Libraries	**Version**
Python	3.9.12
torch	1.12.1+cu113
transformers	4.40.2
numpy	1.24.2
sklearn	1.2.2
PEFT setup	
Libraries	**Version**
sentence-transformers	2.7.0
flash_attn	2.7.4.post1
huggingface-hub	0.30.2
numpy	2.2.4
transformers	4.51.3
peft	0.10.0
bitsandbytes	0.45.5
accelerate	1.5.2
evaluate	0.4.3
scikit-learn	1.6.1
tokenizers	0.21.1
torch	2.5.1

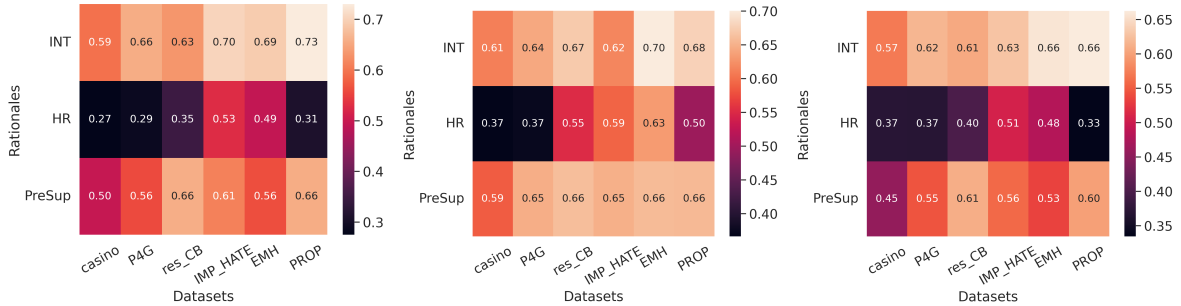
trained on a given source task, we further fine-tune it for 4 k-shot settings (5, 10, 20, and 50) for each of the 5 different target tasks. This results in a massive experimental suite of 810 in-domain experiments and 4050 cross-task experiments.

For our in-context learning setting, we experiment with instruct-tuned versions of two open-sourced models, i.e. LLama-3-8B and the Gemma-9B. To account for prompt sensitivity, the prompts used for inference were first validated on the development split for each of the 6 datasets. We use rationales generated by both proprietary and open-sourced LLMs, i.e. GPT-4o, GPT-3.5-turbo, and Gemma-2-27B-it. Our experiments thus comprise 5 different kinds of rationales (i.e. None, INT, HR, PreSup, and ALL), 2 LLMs for doing ICL, 3 LLMs that generate the rationales, 3 few-shot settings, for the 6 datasets resulting in an additional 540 experiments.

The total cost of the OpenAI credits during the course of our experiments to generate the rationales was approximately USD 265 USD, with the cost of the GPT-4o model being approximately 10 times as costly as the GPT-3.5-turbo version.



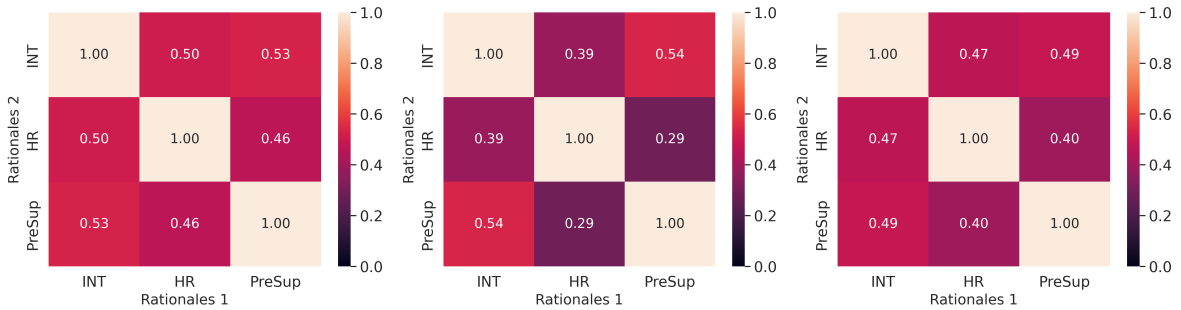
(a) GPT-4o and GPT-3.5-turbo (b) GPT-4o and Gemma-2-27B-it (c) GPT-3.5-turbo and Gemma-2-27B-it



(d) GPT-4o and GPT-3.5-turbo (e) GPT-4o and Gemma-2-27B-it (f) GPT-3.5-turbo and Gemma-2-27B-it

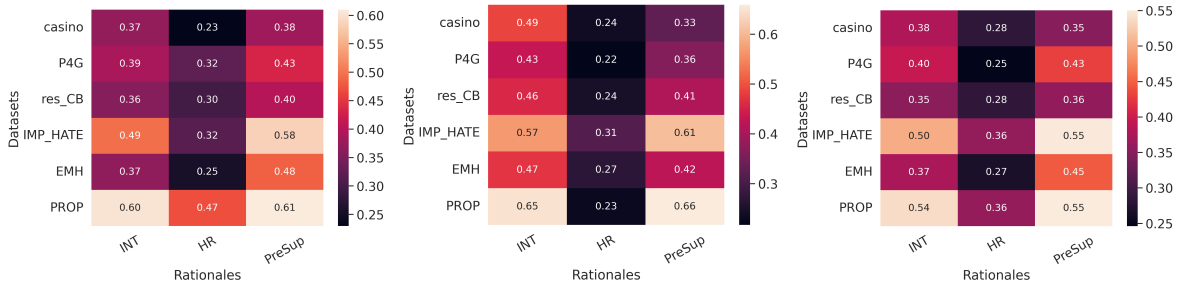
Figure 16: Cosine similarities between rationales generated by three LLMs, i.e. GPT-4o, GPT-3.5-turbo and Gemma-2-27B-it, across different datasets and rationale categories. The figures displayed on the left and right correspond to the models Mistral and MPNET, respectively.

(a) LLM: GPT-4o (b) LLM: GPT-3.5-turbo (c) LLM: Gemma-2-27B-it



(d) LLM: GPT-4o (e) LLM: GPT-3.5-turbo (f) LLM: Gemma-2-27B-it

Figure 17: Cosine similarities between different categories of rationales corresponding to intentions, hearer reactions, and presuppositions as generated by three LLMs, GPT-4o and GPT-3.5-turbo, and Gemma-2-27B-it, and evaluated by the sentence transformers, i.e. Mistral (top 3) and MPNET (bottom 3).



(a) GPT-4o (b) GPT-3.5-turbo (c) Gemma-2-27B-it

Figure 18: Cosine similarities between the original utterance and the rationales generated by different LLMs and evaluated by the sentence transformers MPNET.

.

F Characteristics of the Generated Rationales

F.1 Rationale Similarity

We measure the similarity of the generated rationales across three fronts:

(i) How similar are the three different categories of rationales to each other?

(ii) How similar are the rationales generated by different LLMs for the same rationale category?

(iii) How similar is a generated rationale to its corresponding utterance?

We use cosine distance between the sentential representations as the metric for quantifying similarity. We explore two models to generate these representations, i.e., the popular MPNET model of (Reimers and Gurevych, 2019) for its simplicity and the instruction-tuned version of Mistral-7B (Wang et al., 2023) for its superior performance on the MTEB leaderboard (Muennighoff et al., 2023). We present the similarity scores across different LLMs, different rationale categories, and between the utterance and the rationale in Figures 16, 17, and 18 respectively.

We observe similar trends in the scores regardless of the model used to generate the representations, i.e., MPNET and Mistral. The rationales generated by GPT-4o and GPT-3.5-turbo vary considerably in their similarity scores depending on their category; those corresponding to the speaker's intentions (INT) are the most similar, followed by presuppositions (PreSup), while the hearer reactions (HR) are highly dissimilar. Furthermore, we note a low similarity between rationales corresponding to different categories (the weakest scores occur between PreSup and HR) and between the rationale and the original utterance. Overall, these results highlight that the categories capture perspectives distinct from each other and the original utterance.

F.2 Instance-wise Performance

We investigate several factors that could predict the performance of rationales on an instance-wise basis. The covariates observed, i.e. the factors include (i) the length of the rationale, (ii) the length of the preceding dialogue history, (iii) the similarity between the rationale and the utterance, (iv) the similarity between the rationale and the label description being classified, (v) the readability score measured using the Flesch'as readability ease (Farr et al., 1951; Kincaid, 1975), (vi) the valence, arousal, and dominance scores measured via the VAD NRC lexicon (Mohammad, 2018), and (vii) scores corresponding emotional intensity, emotional polarity and empathy (Wu et al., 2024). The correlation between each of the factors and instance-wise task performance is highlighted in Table 22.

G Generalization Characteristics

We inspect the factors that characterize generalizability over the different experimental settings using the different rationales. We perform a multivariate ANOVA analysis with the absolute performance difference over the baseline as the dependent variable. The independent variables chosen were the rationale category, the LLM used to generate the rationales, the choice of the source and target dataset [6], and the few-shot setting; we also consider the effects of pairwise interaction of each of these variables. We note the F-statistic and their corresponding p-value for the indomain, cross-task and incontext-learning setting respectively in Tables 23, 24, and 25 in the Appendix F.

[6] For the indomain setting we consider only the target dataset

Table 22: Correlation of different rationale characteristiscs with classification accuracy. We explore intentions, hearer reactions, and presuppositions for in-domain, cross-task transfer, and in-context learning settings.

	In-domain			Cross-task Transfer			In-context Learning		
Factor	INT	HR	PreSup	INT	HR	PreSup	INT	HR	PreSup
#RAT Length	-0.07	-0.05	-0.06	-0.09	-0.07	-0.07	-0.15	-0.15	-0.13
# Dial Length	0.05	0.05	0.05	0.09	0.09	0.09	0.08	0.10	0.10
LBL Sim	-0.06	-0.06	-0.04	-0.07	-0.07	-0.04	-0.08	-0.11	-0.05
UTT Sim	-0.02	0.02	-0.02	-0.02	0.01	-0.02	-0.09	-0.01	-0.07
Valence	0.02	0.06	0.04	0.07	0.07	0.04	0.08	0.13	0.07
Arousal	-0.01	-0.01	0.00	-0.04	-0.03	-0.01	-0.08	-0.06	-0.04
Dominance	0.01	0.05	0.03	0.04	0.05	0.01	-0.01	0.10	0.01
Emo Intensity	-0.01	-0.04	-0.02	-0.05	-0.05	-0.02	-0.09	-0.14	-0.07
Emo Polarity	-0.01	-0.04	-0.02	-0.05	-0.05	-0.02	-0.09	-0.14	-0.07
Empathy	-0.01	-0.04	-0.02	-0.05	-0.05	-0.02	-0.09	-0.14	-0.07
Flesch's Readability	0.02	0.03	0.02	0.00	0.04	0.00	0.03	0.08	0.02

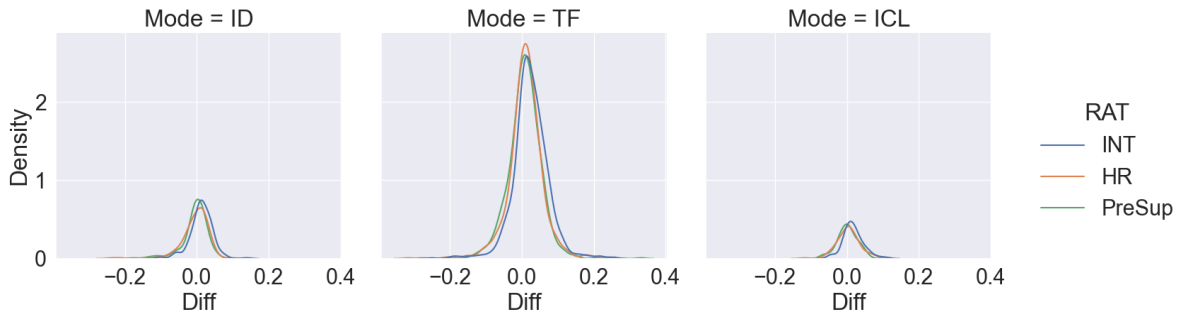


Figure 19: Distribution of the net performance difference across the three different settings, i.e. in-domain (ID), cross-task transfer (TF), and in-context learning (ICL) for the three rationales, i.e. intentions (INT), hearer reactions (HR), and presuppositions (PreSup).

Table 23: The F-statistics and corresponding p-value for the multi-variate ANOVA analysis to investigate the factors that characterize the performance difference in an indomain setting for SFT setup.

Category	F-statistic	p-value
C(LLM)	0.363057	5.47E-01
C(RAT)	21.073603	**1.69E-09**
C(Dataset)	5.252105	**1.05E-04**
C(fewshot)	11.699875	**4.50E-09**
C(Dataset):C(LLM)	1.642512	1.47E-01
C(RAT):C(Dataset)	2.680245	**3.36E-03**
C(LLM):C(RAT)	3.627177	**2.73E-02**
C(fewshot):C(LLM)	0.566543	6.87E-01
C(RAT):C(fewshot)	4.213318	**6.76E-05**
C(fewshot):C(Dataset)	10.810497	**4.69E-28**

For the in-domain setting, the performance change hinges on the rationale category, the number of few-shot examples, the target dataset, and also their pairwise interactions. We also observe mild significant pairwise effects between the LLM and the rationale category. A similar trend emerges during cross-task transfer; the rationale category, the target dataset, and the number of few-shot examples play a significant effect in influencing performance. However, the choice of the source dataset is significant only when we consider its pairwise interaction with the other covariates. The story differs slightly for the ICL setup; the choice of the dataset, the rationale, and the LLM (but not the few-shot setting) significantly impact performance.

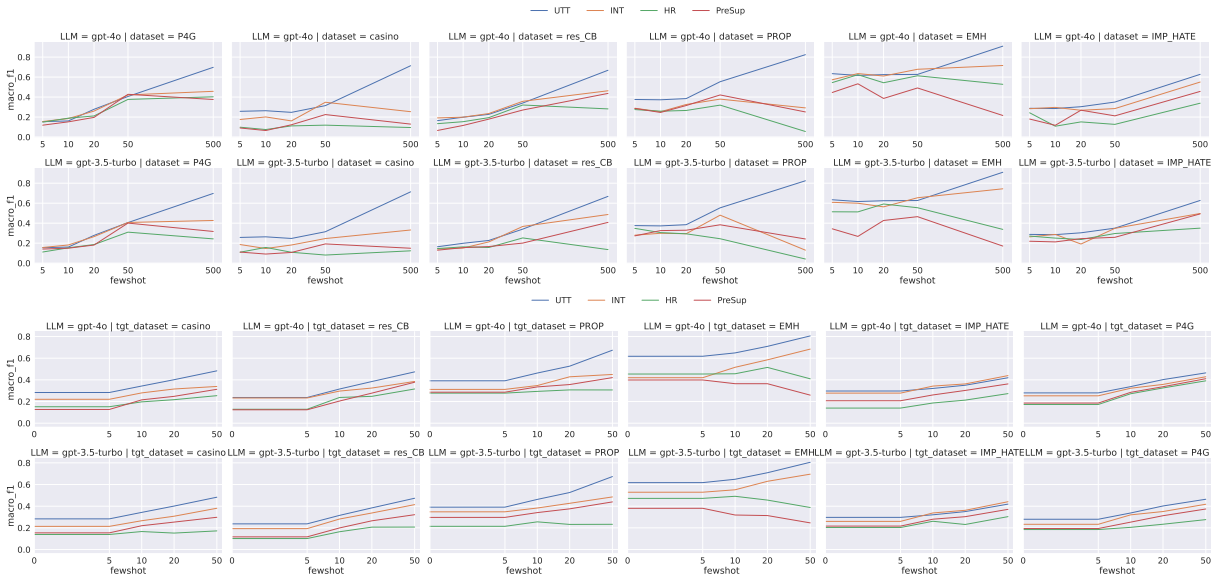


Figure 20: In-domain performance (top) and cross-task performance of models in presence of only the rationale across different few-shot cases. Note that the model was trained on BOTH the rationale and utterance.

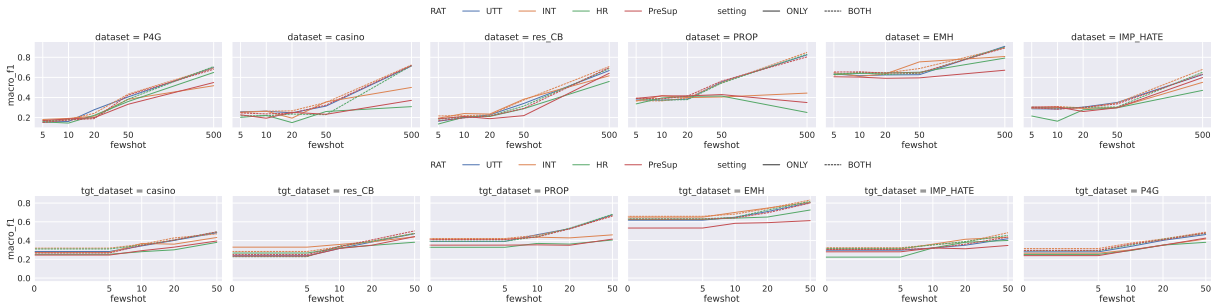


Figure 21: In-domain performance (top) and cross-task performance (below) of models using only the rationale across different few-shot cases. Note that the model was trained on ONLY the rationale.

Table 24: The F-statistics and corresponding p-value for the multi-variate ANOVA analysis to investigate the factors that characterize the performance difference in a cross-task transfer setting for SFT setup.

Category	F-statistic	p-value
C(LLM)	2.350972	1.25E-01
C(RAT)	31.459235	**3.17E-14**
C(fewshot)	2.599193	**3.45E-02**
C(src_dataset)	1.806214	1.25E-01
C(tgt_dataset)	5.282518	**3.09E-04**
C(LLM):C(RAT)	3.847212	**2.15E-02**
C(LLM):C(fewshot)	1.138982	3.36E-01
C(LLM):C(src_dataset)	2.245978	**4.73E-02**
C(LLM):C(tgt_dataset)	3.028266	**9.92E-03**
C(fewshot):C(RAT)	1.161916	3.18E-01
C(src_dataset):C(fewshot)	4.966472	**3.11E-12**
C(fewshot):C(tgt_dataset)	4.083211	**3.01E-09**
C(RAT):C(src_dataset)	2.137128	**1.90E-02**
C(RAT):C(tgt_dataset)	2.86715	**1.47E-03**
C(src_dataset):C(tgt_dataset)	3.242511	**1.52E-06**

Table 25: The F-statistics and corresponding p-value for the multi-variate ANOVA analysis to investigate the factors that characterize the performance difference in fewshot setting for in-context learning models.

Category	F-statistic	p-value
C(LLM)	5.202281	**6.10E-03**
C(RAT)	10.668473	**3.50E-05**
C(dataset)	7.535951	**1.00E-06**
C(fewshot)	0.356484	7.00E-01
C(model_name)	1.22807	2.69E-01
C(LLM):C(RAT)	1.561942	1.85E-01
C(LLM):C(dataset)	0.734409	6.92E-01
C(LLM):C(fewshot)	1.258991	2.87E-01
C(LLM):C(model_name)	0.831352	4.37E-01
C(RAT):C(dataset)	0.647286	7.72E-01
C(RAT):C(fewshot)	0.750312	5.59E-01
C(RAT):C(model_name)	2.665021	7.15E-02
C(dataset):C(fewshot)	2.14782	**2.15E-02**
C(dataset):C(model_name)	3.456222	**4.85E-03**
C(fewshot):C(model_name)	0.938185	3.93E-01

H Ablation Results

H.1 Importance of the utterance information

We carry out ablation studies to investigate the role of the utterance on task performance i.e. how does the performance vary when we omit out the utterance and evaluate the fine-tuned model using only the rationale. We explore two settings: (i) where the model is provided with both the utterance and rationale information during training, but use only the rationale during inference, (see Figures 20) and (ii) where we train and test the model with only the rationale as an augmentation (see Figure 21).

We observe a noticeable degradation in performance compared to the baseline (the model is trained only on the utterance) in the former case for both the indomain and cross-task setting; the drop progressively increases with the amount of training data, highlighting that fine-tuned models do not solely rely on the rationale to make its predictions. The latter scenario where the model is fine-tuned with only the rationales fares better, albeit still falling short of the baseline in the in-domain setting. When trained on only the rationale information, the impact of the rationale category on the task performance becomes more pronounced. We see higher gains from adding the hearer reactions to P4G, the presuppositions to IMP_HATE, and the intentions to casino, and EMH. In the cross-task setting, the performance drop is almost negligible; in fact we see marked improvements for res_CB, IMP_HATE and EMH with the intention rationales over the baseline. In short, we see that the utterance information is crucial for task performance and though rationales provides a useful augmentation, they cannot be used as a replacement or substitute for the utterance. Future work needs to inspect how to design free-text rationales that can capture all the salient aspects of the utterance (Chen et al., 2023).

H.2 Perturbation of the Rationales

We also carry out sensitivity analysis of the rationales by observing how perturbing the rationale text affects task performance. We compare different kinds of perturbations such as synonym swap using Checklist (Ribeiro et al., 2020) and WordNet, different kinds of augmentations (EmbedDA), deletions or combination of them (EDA) (Wei and Zou, 2019). We also control for the fraction of words being perturbed in the rationale text i.e. 10%, 50% and 90%. We depict the change in task performance due to perturbations in Table 22

Overall, on a macro scale, we observe that perturbations indeed decrease task performance with the deterioration becoming more pronounced as the proportion of words being perturbed increases. We also note that certain methods are more effective than others such as deletion as opposed to synonym matching or entity replacement. Such an analysis highlights that the instruct-tuned model does rely on the rationales for classification.

I Qualitative Analysis

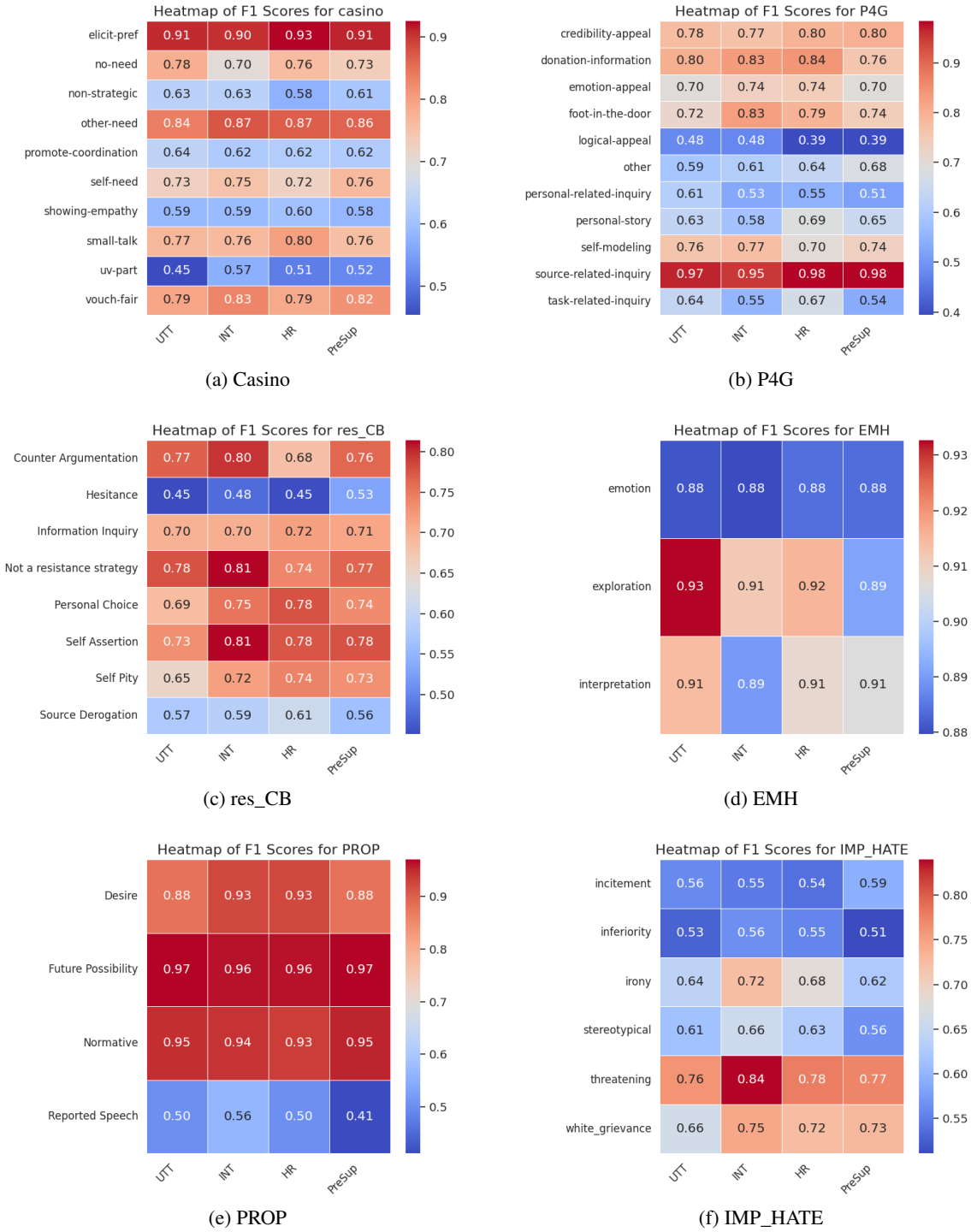
We now carry out a qualitative analysis to investigate the specific instances where including the rationales actively improves the model's predictions in an indomain setting.

We depict the fraction of cases that benefit from adding rationales in the form of a Venn Diagram in Figure 25 in the Appendix. The overlapping areas indicate the fraction of instances that benefit from more than one types of rationale; for example, 10.0% of all instances benefit from all three rationales in CaSiNo. We consider only those instances where the baseline (i.e., only the utterance text) fails to predict the label correctly a majority of times, but succeeds when the rationale is provided.

The rationale with the greatest impact on performance is dependent on the nature of the task. The hearer reaction or HR has the highest impact on P4G, possibly because it captures the thought processes of the persuadee (EE) as they are being persuaded to donate. For example, the utterance "Anything would help even small donations add up when everyone pitches in." evokes a sense of reassurance from the persuadee (EE) that any contribution is valuable and is thus recognized as a "foot-in-the-door" strategy. Presuppositions are useful for IMP_HATE, a dataset that directly references stereotypes and thus requires generic knowledge to infer the type of implicit hatred. Tasks that are centered around the outcome the speaker is invested in, i.e. strategies employed to resist persuasion (res_CB), or signaling empathy to someone in therapy (EMH) benefit mostly from intentions. Furthermore, similar tasks e.g., CaSiNo and res_CB which deal with negotiation have similar relative performance for the same rationales.

However, it should also be noted that a given rationale category does not serve as a silver bullet for all instances. We highlight some examples where model improvements were due to only one type

Figure 22: Impact of different kinds of perturbation on the rationale text for classification performance.

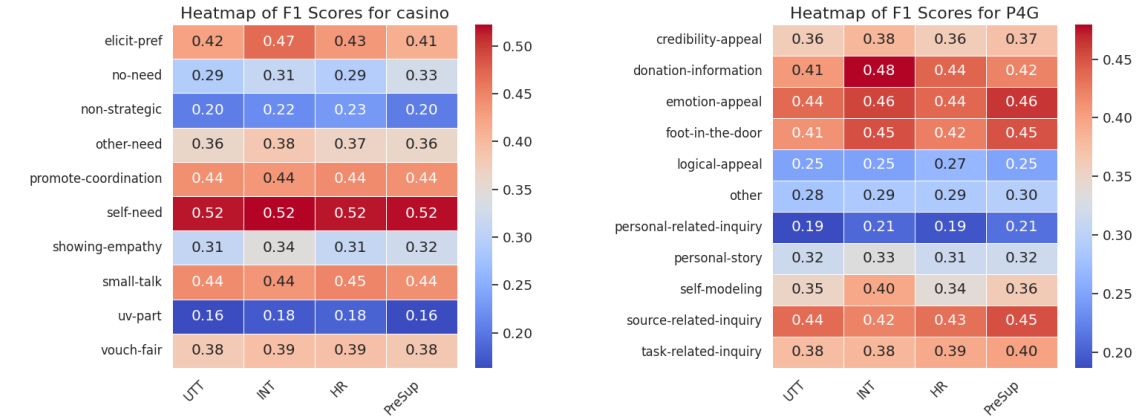


(a) Casino (b) P4G

(c) res_CB (d) EMH

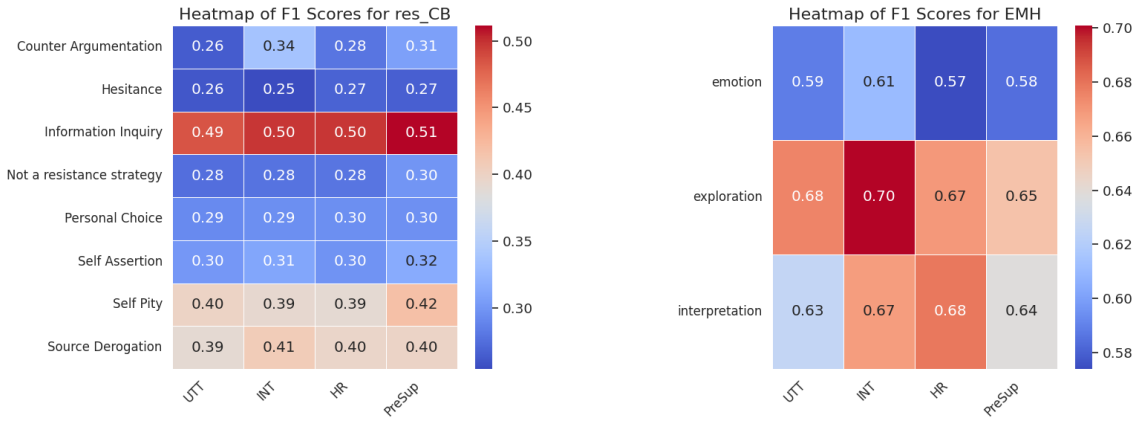
(e) PROP (f) IMP_HATE

Figure 23: Comparative performance of rationales in terms of macro F1 score across different labels for different tasks in an indomain setting.



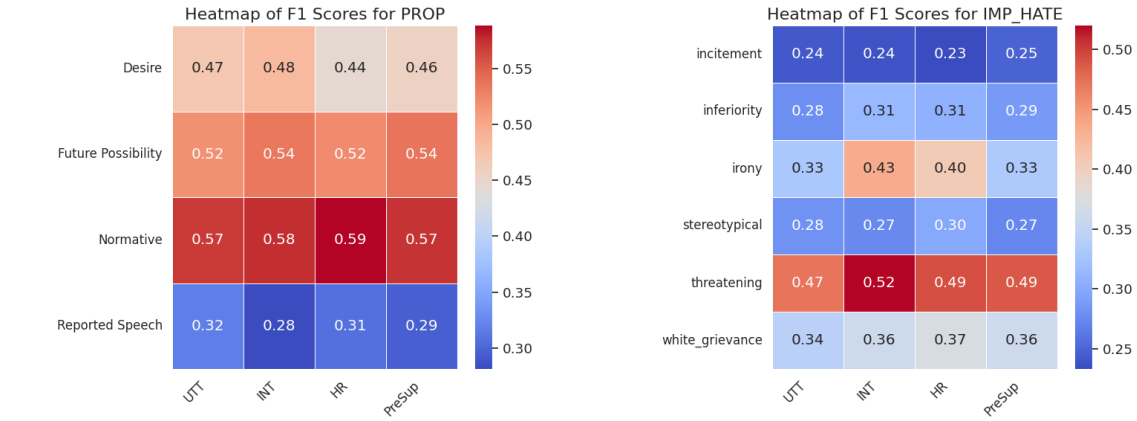
(a) Casino

(b) P4G



(c) res_CB

(d) EMH



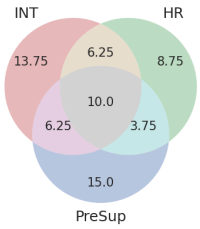
(e) PROP

(f) IMP_HATE

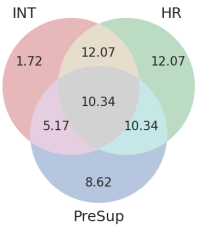
Figure 24: Comparative performance of rationales in terms of macro F1 score across different labels for the different target tasks in a cross-task setting

Table 26: We present instances across different datasets where adding the rationale information was crucial in predicting the correct label always. We compute Shapley values for each token in the rationale to observe its contribution to the model's decision; the highlighted portions correspond to high positive associations with the label.

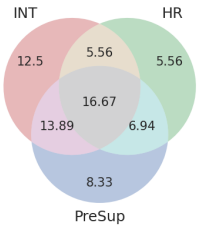
Dataset	Label	Utterance text	Rationale Text	CAT
casino	showing-empathy	Are you sure that's enough firewood for you and the baby? I know that babies can easily get very sick from dropping temperatures.	**Expressing concern about the** adequacy of **firewood for the baby**	INT
casino	vouch-fair	hey buddy I hope we both end up with a good deal:)	Expressing hope for a **mutually beneficial outcome**	INT
IMP_HATE	white_grievance	is zuma talking about seizing white land or not ? are you denying it ?	Seeking clarification and confirmation about Zuma's stance on **seizing white land**; challenging the respondent to deny it.	INT
IMP_HATE	irony	why do we have an immigration policy? should just let them have a free for all and all can pay!	Expressing frustration **and sarcasm** about the current immigration policy and suggesting an extreme alternative to highlight perceived issues.	INT
IMP_HATE	irony	mt : ""we must resist ebolaphobia. these viruses just come here for a better life, to do jobs that american viruses refuse to do.	The speaker is using **satire to comment** on the fear of Ebola (ebolaphobia) by comparing it to immigration issues suggesting that the fear is irrational and **mocking the idea that viruses have intentions** similar to human immigrants.	INT
IMP_HATE	white_grievance	but that wouldn't enable them to destroy white neighbourhoods .	There is a belief or concern that certain actions or policies could lead to **the destruction of white neighborhoods.**	PreSup
IMP_HATE	threatening	and they should have to pay every dime back, plus interest. from their own damn pockets.	People **should be held** financially accountable for their actions and **should not be able** to avoid paying back debts.	PreSup
P4G	credibility-appeal	As a donor I wanted you to see the site and the children that you would be helping	**EE is informed about the purpose of the link and the impact of their donation.**	HR
P4G	foot-in-the-door	Every little bit help.	EE feels reassured **that their small donation** is still valuable.	HR
P4G	foot-in-the-door	Every little bit help.	**Reassure the listener that any contribution is valuable.**	INT
P4G	foot-in-the-door	Your right, but I'm not asking for much.	**Minimizing** the financial impact of the donation	INT
res_CB	Source Derogation	Too be honest don't like the front bumper would be better without that black cover	**The seller might feel a need to address the buyer's concern about the bumper.**	HR
res_CB	Self Pity	at this i can only pay about 1600 could you do that	Seller **realizes the buyer's budget constraints.**	HR
res_CB	Source Derogation	Yes. What didn't your wife like about the bed?	Seller feels questioned about the **reason for selling the bed.**	HR



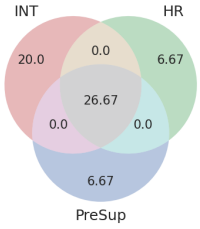
(a) CaSiNo



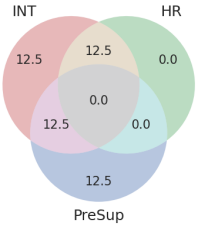
(b) P4G



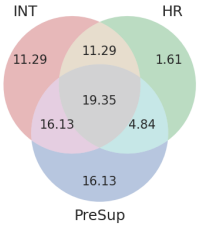
(c) res_CB



(d) EMH

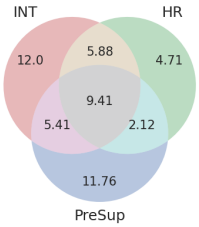


(e) PROP

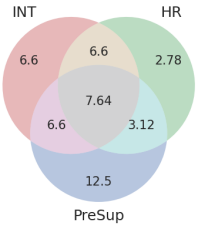


(f) IMP_HATE

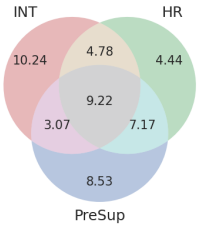
Figure 25: Venn Diagram showing the proportion of instances where including the rationales fared better than the baseline in an in domain setting.



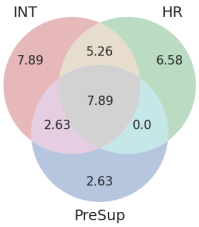
(a) CaSiNo



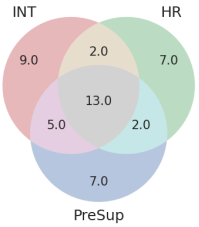
(b) P4G



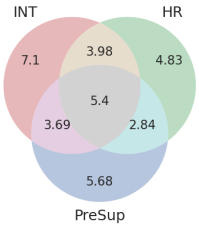
(c) res_CB



(d) EMH



(e) PROP



(f) IMP_HATE

Figure 26: Venn Diagram showing the proportion of instances where including the rationales fared better than the baseline in a 5-shot transfer setting.

of rationale in Table 26 in the Appendix and the possible reasoning for the same. While all three rationales are valid with respect to the utterance, we hypothesize that certain phrases or terms in the given generation might make it easier to predict the label category. For example, the phrase "feels questioned" in the HR hints at source derogation, which is not observed for the other rationales for the res_CB example. Likewise, the wording "how one might treat a dog" in the presupposition conveys the sense of inferiority more prominently than the generic idea of mistreatment in IMP_HATE. Since the rationales were not generated with a particular task in mind, the number of instances where the wording aligns with one of the task label's definition is also infrequent.

J Use of AI Assistants

We relied on Github Co-pilot and GPT-4o for generating the plots and figures used in our study. We also used GPT-4o and the in-built AI assistant spell-check to help tighten the abstract and catch grammatical errors.