DeliData: A dataset for deliberation in multi-party problem solving

Anonymous ACL submission

Abstract

Group deliberation enables people to collaborate and solve problems, however it is understudied due to a lack of resources. To this end, we introduce the first publicly available dataset containing collaborative conversations on solving a cognitive task, consisting of 500 group dialogues and 14k utterances. In 64% of these conversations, the group members are able to find a better solution than they had identified individually. Furthermore, we propose a novel annotation schema that captures deliberation cues and release 50 dialogues annotated with it. Finally, we use the proposed dataset to develop and evaluate two methods for generating deliberation utterances. The data collection platform, dataset and annotated corpus will be made publicly available.

1 Introduction

Group deliberation occurs in a variety of contexts, such as hiring panels, study groups, and scientific project meetings. It is traditionally explored in the field of psychology, where researchers examine the conditions under which a group can make better decisions. Mercier and Sperber (2011) discuss how a group can outperform even the most knowledgeable individual within it – *the assembly bonus effect*. This was also demonstrated by (Navajas et al., 2018) who showed that small focus groups can outperform the wisdom of the crowd.

In order to study what makes deliberations successful and learn how to intervene to this effect, we need a dataset that contains discussions where groups collaborate to solve a task. Furthermore, the task should be such that the decisions made can be objectively measured as correct or incorrect. Most existing datasets are between two interlocutors (Budzianowski et al., 2018; Dinan et al., 2019; Anderson et al., 1991), thus not containing group discussions. Focusing on group datasets, one could consider negotiation dialogues (Afantenos et al., 2012), which while multi-party are adversarial in nature, therefore not containing collaboration. Publicly available datasets containing collaborative group discussions are WikiDisputes (De Kock and Vlachos, 2021) and AMI (Carletta et al., 2005), but neither contains an objective measure of success, thus making it impossible to evaluate how well did the conversation go. Niculae and Danescu-Niculescu-Mizil (2016) collected a group dataset containing collaborative problem-solving conversations with an objective measurement of success but their dataset is not publicly available.

In this work, we present the first publicly available dataset for group deliberation, containing a quantitative measure of task performance: **DeliData – Deliberation Dataset**. An example conversation is shown in Figure 1, with a group deliberating to solve the Wason card selection task (Wason, 1968), a well-studied task in cognitive psychology. In the example, the group engages in various deliberation strategies: a participant is moderating the conversation by prompting the group for a response (utterance 1), whereas in utterance 4 a participant suggests exploring a different solution. Overall, the group starts with the common, but wrong, solution (utterances 2 and 3) and converges on the correct solution (utterances 6 and 9).

The DeliData corpus contains 500 group dialogues, together with a measure of task performance before and after the group discussion. Given these measures, we show that after discussing the solution, 64% of the groups perform better at the Wason task, compared to their solo performances. Moreover, in 43.8% of the groups who had a correct answer as their final solution, none of the participants had solved the task correctly by themselves, thus demonstrating how people can solve the task better through deliberation. In our analysis, we also show, that groups of 3 or more people solve the task better than conversations with 2 participants.

To aid future analysis and dialogue system de-

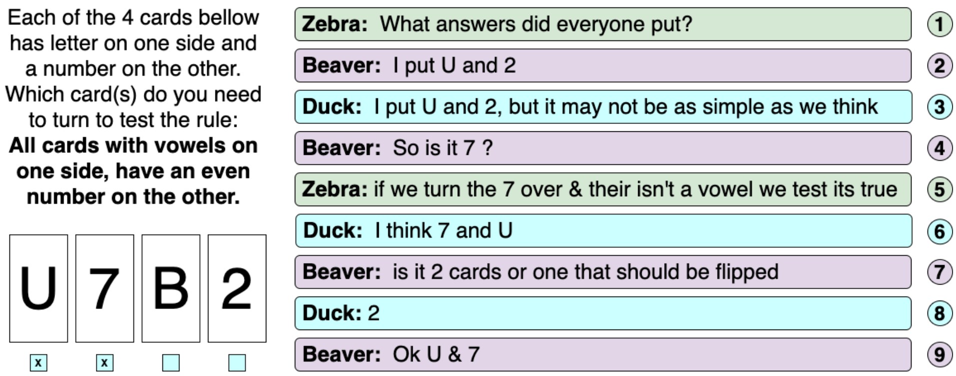

Figure 1: Abridged conversation from our dataset between 3 people solving the Wason card selection task

velopment we propose an annotation schema that captures conversational dynamics and deliberation cues in collaborative conversations, and release an annotated corpus with 50 dialogues using it. Finally, we experiment with generating utterances that *probe* the conversation by asking questions, using both retrieval and generative approaches.

2 Related Work

Niculae and Danescu-Niculescu-Mizil (2016) investigated group collaboration in the context of playing a game attempting to geo-locate a photo on the map. In their experimental setup, they evaluate each participant individually, after that they initiate a group discussion and finally ask the group to make a decision together. Unfortunately, their dataset is not publicly available, and thus cannot be used in future studies. Likewise, Kim et al. (2021) investigates how groups of people collaborate in solving a task together, as well as how can dialogue system can be incorporated within the discussion. Unfortunately, their dataset contains only 12 discussions, making it too small for any reasonable analysis or dialogue systems training, and similarly to (Niculae and Danescu-Niculescu-Mizil, 2016), their dataset is also not publicly available.

Wikipedia is a popular source of collaborative conversations. Hua et al. (2018) collect 91M discussions from Wikipedia, together with the discussed edits. It is the largest dataset that captures group collaboration, but it is not supported by an annotated corpus. This is partly addressed by Al-Khatib et al. (2018), who annotate 200k discussion turns from Wikipedia in 33 dimensions based on discourse acts, argumentative relations and semantic frames. However, unlike the conversations of Niculae and Danescu-Niculescu-Mizil (2016) and the work presented in this paper, it is impossible to know whether the participants in a conversation on Wikipedia reached a better decision, which renders assessing constructiveness more difficult because there is no objectively correct answer.

Related to constructive conversations is the research on negotiation dialogues which have been explored in the context of games (Keizer et al., 2017; Cuayáhuitl et al., 2015) and trading (He et al., 2018; Lewis et al., 2017). However, even though negotiation dialogue research often deals with multiparty conversations (Cuayáhuitl et al., 2015), such systems are by nature adversarial, rather than constructive.

Multiparty conversations are also the focus of Carletta et al. (2005), who created a multi-modal corpus of business meetings containing audio, video, transcriptions and auxiliary materials provided to the participants. However, they did not explore deliberation strategies, nor tried to measure the productivity of the group. Using parts of this dataset, the CALO project (Tur et al., 2010) proposed a toolkit to assist group meetings, such as dialogue act segmentation, action item recognition and others, but no attempt to assess constructiveness was made. Finally, de Bayser et al. (2019) evaluated turn prediction in the context of group dialogues. They evaluate their system on 3 datasets: one is proprietary, one is artificially created by combining 1-to-1 dialogues from Budzianowski et al. (2018), the third dataset consists of transcripts of a popular TV show, which while containing true multi-party dialogues they are not collaborative.

3 Experimental Setup

In our experiments with the Wason card selection task (Wason, 1968), participants are presented with 4 cards with a number or a letter on them. They have to answer the following question "*Which* *cards should you turn to test the rule:* **All cards with vowels on one side have an even number on the other.**?". Most people initially select the vowel and the even number (i.e. selecting the two cards mentioned in the question), which is incorrect, demonstrating *confirmation bias* (Mercier and Sperber, 2011). The correct answer is to turn the vowel, to check for an even number on the other side, and to turn the odd number, to verify there isn't a vowel on the other side.

We calculate **task performance** in two ways. First, we consider a **coarse-grained** (binary) scoring of the task - **Correct - 1** if the vowel and odd number are selected, **Incorrect - 0** otherwise. Recognising that the coarse-grained scoring may needlessly penalise answers that are close to the correct one, we also devised an alternative **fine-grained** scoring. We grant 0.25 points for (i) turning a vowel or an odd number, and (ii) for **not** turning the even number or the consonant. Therefore, if the participant submitted a correct solution, their score would be 1, if they are off by one card - 0.75 and so on. We also calculate **performance gain**, by subtracting the average of the solo solutions from the average of the group performance. For example, if the average score of participants' solo submissions was 0.5 and improved to 0.75 after the discussion, the group performance gain would be $0.75 - 0.5 = 0.25$. We collect the data using the following protocol (full participant instructions available at Appendix A.1):

1. **Solo Phase**. Each of the participants in the group is presented with the same 4 cards and submits a solution to the task.
2. **Group Phase**. Following the solo phase solution submission, participants gain access to a chatbox to share their solutions and discuss. We encourage them to do so for at least 5 minutes but no longer than 7 minutes without enforcing these time limits; thus there are cases with very short and very long conversations.
3. **Revised Submission**. After discussing their solutions, the participants are asked to revise their initial card selection and submit again.

We posted our data collection on the crowdsourcing platform Mechanical Turk with the following job specification:

1. Everyone who completes the task is paid \$2.00 (approx. £1.60). Participants are given a bonus of \$1.00 (£0.80) if they return the right answer. As the average time for participation is about 8 minutes, each participant is paid £12/hour (or £18/hour if they solve the task correctly). This is between 35% and 102% above UK's National Living Wage [1].
2. No personal information is collected and the participants are asked not to share anything that may reveal personal details.
3. We recruited only adult participants from countries where English is a primary language, and they complete a simple reading comprehension test. The only language used in our dataset is English.

Participants are informed that we are investigating how people collaborate in solving a cognitive task and that we will be saving chat transcripts. This experimental protocol was approved by the ethics committee of the authors' institution.

The data collection is performed using a web application we call *DialogueDen*, which we open-source together with this study. The design of the platform allows us to record solo and group selections and the state of the game in key points of the experiment. This data can be used to identify when a participant reached the correct decision, even if they don't express it explicitly in the chat. Moreover, we integrated a number of features to *DialogueDen* that are specific for the data collection on Mechanical Turk, addressing various issues that arise when collecting group conversations in an unsupervised manner. These are part of the code release and are presented in detail in Appendix A.2.

4 DeliData dataset

Using the experimental protocol above we initially conducted a pilot study, where we collected 18 group dialogues, with 53 volunteers from a university psychology department, who didn't have prior knowledge of the task. After that, we ran a larger scale data collection on Mechanical Turk which is often used for data collection in behavioural research and often produces similar results to in-lab experiments (Crump et al., 2013). This data collection was not moderated in any way, making it an in-the-wild data collection. We ensure the quality and anonymity of the data from MTurk by manually checking each conversation. We excluded a total of 160 conversations that were too short, of poor quality or with too few actively engaged par-

[1] £8.91/hour as of 01/04/2021, based on https://www.gov.uk/government/publications/the-national-minimum-wage-in-2021

	Pilot	Mturk	Total
Number of Dialogues	18	482	500
Total Participants	53	1526	1579
Total number of utterances	705	13298	14003
AVG utterances	39.2	27.6	28
AVG utterance length	8.19	8.62	8.59
AVG unique tokens	78.1	67.6	68
AVG number of participants	2.94	3.17	3.16
Solo Performance (fine-grained)	0.59	0.59	0.59
Group performance (fine-grained)	0.81	0.71	0.72
Solo Performance (coarse-grained)	0.19	0.11	0.11
Group performance (coarse-grained)	0.57	0.32	0.33
AVG group agreement	0.92	0.83	0.83

Table 1: Corpus statistics for pilot and MTurk data.

ticipants. Thus, we release 482 dialogues that are of comparable quality to our in-lab pilot.

Summarised statistics of the two subsets are presented in Table 1. While the two subsets differ in terms of absolute performance, the improvement from solo to group performance is substantial in both data collections for both coarse- and fine-grained metrics, in agreement with results from psychology research on offline deliberation (Mercier and Sperber, 2011), and thus validating our data collection approach using MTurk. Another difference is that the average number of utterances per dialogue is lower on MTurk, which we attribute to the psychology student volunteers being more dedicated than crowd workers.

In Table 2 we compare three multi-party dialogue datasets: StreetCrowd (Niculae and Danescu-Niculescu-Mizil, 2016), Settlers of Catan (SoC) (Afantenos et al., 2012), and ours. Of these three, only two are collaborative - ours and StreetCrowd, as SoC is among players competing against each other. Ours is the only one containing collaborative group conversations available for research. Moreover, while it contains fewer dialogues than StreetCrowd, these are 2.5 times longer in terms of utterances, thus more likely to exhibit collaborative strategies spanning over multiple utterances.

5 Annotating deliberation cues

5.1 Annotation Schema

In order to annotate the conversations collected we first considered using the annotation schema previously proposed for discourse parsing (Zhang et al., 2017), Wikipedia discussions (Al-Khatib et al., 2018). While both of these schemata capture some discussion markers (such as Agreement or Argumentation), they fail to identify which utterances are helping the group in terms of deliberation. In terms of collaborative discussions, the MapTask schema by Carletta et al. (1996) annotates conversations between two participants, who play a game together. However, they did not annotate reasoning utterances, limiting their annotation to basic interactions such as question and answer utterances.

Property	StreetCrowd	SoC	DeliData
dialogues	1,450	32	500
utterances	17,545	2,512	14,003
utterances per dialogue	12.1	78.5	28
utterance length	5.33	N/A	8.59
pub. available	No	No	Yes
collaborative	Yes	No	Yes

Table 2: Multiparty dialogue corpora comparison

To address this, we propose an annotation schema that contains 3 levels of annotation, each focusing on different aspects of deliberation. Figure 2 gives the overview of the schema, and we describe it in detail in the remainder of this section.

At the top level of the schema, we are interested in identifying **probing deliberation**, i.e. any utterance that provokes discussion, deliberation or argumentation *without* introducing novel information (*Hey, @Cat what do you think was the solution?*). We also recognise that most utterances in a conversation are not probing, but are inherently useful for the conversations. We label these utterances as **non-probing deliberation**, and they include all discussions that are concerned with the task's solution and participants' reasoning (*I think the answer is A, because we have to check each vowel for sure*). Finally, we include a **None** label that covers all utterances that are not related to the previous two categories. These utterances often include familiarities (*Greetings fellas*) or hesitation cues (*hmm...*). After distinguishing between probing and non-probing deliberation, we classify each utterance into 5 roles at the second level:

- **Moderation** (exclusive to probing deliberation): Moderation utterances are not concerned directly with the task at hand, but rather with *how* participants converse about it (*Let's discuss our initial solutions*).
- **Reasoning**: Utterances focusing on argumentation and can be both probing (*Why did you*

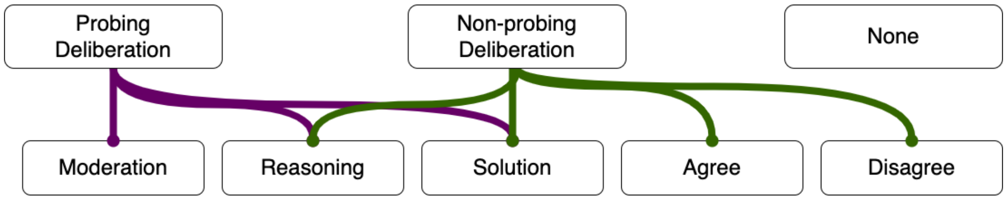

Figure 2: Hierarchical annotation structure

think it wasn't 8?) and non-probing (*I think it would be 7 to test if it would be incorrect*).

- **Solution**: Utterances that are managing the solution of the task. Can be both probing (*Are we going for A and 4?*) or non-probing (*I think the answer is 7 and A*).
- **Agree** and **Disagree** (exclusive to non-probing-deliberation): Utterances expressing agreement or disagreement with a previous argument or solution.

An important caveat with **Reasoning** is that it takes a priority over other labels.

Some of the utterances may carry additional information beyond what is captured by their type and role, i.e. the first two levels of the annotation. Therefore, we introduce a set of **additional labels** that mark specific phenomena in the conversation, which we defined as follows:

- **specific_addressee**: Utterances explicitly addressing specific participant(s) (*@Llama what do you think?*)
- **complete_solution** and **partial_solution**: Utterances advocating for either a complete task solution (*Let's turn A and 7*), or a partial one (*one of the cards is A*).
- **solution_summary**: Utterances that recall previous solutions to prompt for an agreement (*So, do we all agree on A and 5?*).
- **consider_opposite** - utterance suggesting an opposite solution. (*maybe not L?*)

5.2 Annotated dataset

Using the annotation schema introduced in this section we annotated 50 dialogues and a total of 1696 utterances from the dataset presented in section 4. We performed an annotation agreement study between 3 annotators on 41 of the dialogues using Cohen's kappa (Cohen, 1960). We obtained an inter-annotator agreement of 0.75 on the first level, 0.71 on the second level, and an average agreement of 0.53 on the additional labels.

The label distribution for the first two levels is presented in Table 3. Overall, the number of

	Probing	Non-probing deliberation	Total
Moderation	89	0	89
Reasoning	59	453	512
Solution	66	305	371
Agree	0	265	265
Disagree	0	9	9
Total	214	1032	1246

Table 3: Frequencies for the labels in the top two levels of the annotation schema

Additional Label	Count	Prevalence
specific_addressee	55	4.4 %
complete_solution	258	20.7 %
partial_solution	79	6.3 %
solution_summary	40	3.2 %
consider_opposite	11	0.9 %

Table 4: Label distribution the additional labels

Reasoning and **Solution** utterances are substantial, confirming that the subjects in our data collection engaged in substantial discussions about the solutions and their reasoning. The corpus also contains 214 **Probing** utterances, which are similarly distributed between **Moderation**, **Reasoning**, and **Solution**, thus suggesting that the strategies chosen for annotation are commonly used. Finally, 450 utterances were annotated as non-deliberative (**"None"**), and are excluded from the table.

In Table 4 we present the distribution of additional labels. In column **Count** we show the total number of occurrences of each of these labels, while in **Prevalence** we show how often this label occurs in *all* utterances, including those without annotation for an additional label. The most prevalent label is **complete_solution**, appearing in about 20% of the utterances. While the other additional labels occur less in the conversation (around 5% or less), they might be useful for dialogue analysis.

6 Analysis and Experiments

6.1 Two-party and multi-party conversations

While in our dataset two-party and multi-party (3 or more participants) conversations have similar statistics, there are notable differences that we highlight in this section. In Figure 3, we present histograms comparing three conversational statistics - the total number of messages, number of unique tokens and participation balance, represented by entropy. First, dialogues between two interlocutors have mostly between 10 and 25 utterances, while group discussions in DeliData are uniformly represented in a larger range, between 20 and 40 utterances, with a long tail of conversations longer than 50 utterances. This naturally occurs, as multiparty discussions, contain more arguments and exchange of ideas. Likewise, participants in these discussions tend to use a larger vocabulary of words, as shown on the histograms of the unique tokens.

In this analysis, we also look at how balanced are the conversations, i.e. whether all of the participants contributed equally. We calculate the participation entropy similarly to Niculae and Danescu-Niculescu-Mizil (2016), where the entropy is maximised if everyone participated equally, and approaches 0 if there is a large imbalance. In our dataset, the balance for two-party conversation is better, where 40 % of the discussions are almost uniformly balanced, while in the multi-party discussions, it is often the case that one of the participants is driving the discussion. This is not surprising, as in one-to-one conversations if one of the participants asks a question, it is customary that the other participant answers. Such is not the case for multi-party discussions, where some of the participants may decide to have a more passive role.

Besides conversation statistics, we analyse the difference in task performance. Verifying for the initial conditions first, the solo performance of both types of groups is comparable - 0.597 and 0.585. On the other hand, the collective performance of these groups was 0.694 for two-party conversations and 0.724 for multi-party, thus the performance gain is 0.096 and 0.139 respectively. Therefore, we argue that it is the multi-party (as opposed to two-party) discussion that led to an improved conversational performance.

6.2 Propagation of correct solutions

Analysing our data we found out that there is 0.36 Kendall's Tau B correlation Kendall (1938) between group consensus and performance gain. An investigation of how correct solutions propagate through the conversations showed that 21.2% of conversations started and finished with the same

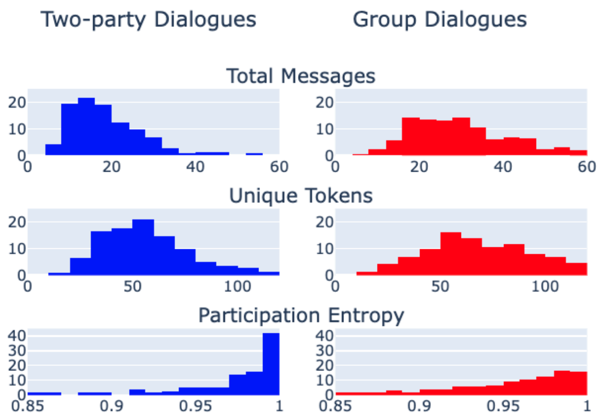

Figure 3: Comparison between conversational statistics of two-party dialogues(left) and group dialogues (right). Each of the histograms is showing percentage of dialogues on the y-axis.

amount of correct submissions, thus the participants didn't convince anyone of the correctness of their response. In 35% of the discussions where a single participant had answered correctly in their solo submission, they convinced at least one more participant in the group phase. However the reverse also happened - in 4% of all dialogues, the group convinced a participant with the correct answer to change it, which is considerably rarer than changing to the correct solution. Finally, in 43.8% of the groups in which at least one participant submitted a correct response after the conversation, no participant had submitted a correct solution in their solo phase. This supports the *group is better than the sum of its parts* hypothesis, suggesting that deliberation offers more than just facilitating the spread of a correct solution among group members, and is consistent with the findings of Moshman and Geil (1998) and Schulz-Hardt et al. (2006), who show that deliberation plays a bigger role in task success, compared to individual participants' ability.

Furthermore, we present an analysis of different solution propagation patterns based on the annotation schema. We compared the groups where at least one of the participants had the correct solution in their solo phase, to the groups which reach the correct solution without anyone knowing the solution in their solo phase (referred to as DELI). The DELI subset contains a higher percentage of probing (17.3% vs 14.4%), and reasoning (43.8% vs 37.8%) utterances, suggesting that the participants are actively engaging in deliberation to get to the correct solution. Naturally, the DELI subset contains fewer utterances that propose a solution

(30.4% vs 35.7%), as participants are more engaged with the reasoning behind the solution, opposed to the solution itself. These findings are suggestive of the rich source of information about the dynamics of deliberation present in the data.

6.3 Predicting conversation success

In order to analyse the factors that make a conversation constructive as well as showcase possible applications of the DeliData corpus, we perform a series of modelling experiments, where we predict the constructiveness of a conversation.

Given the size of our dataset and the potential instability of neural models, herein we use a simple decision tree classifier (Pedregosa et al., 2011) with a maximum depth of 7 and minimum samples per leaf set to 5 and use leave-one-out cross-validation (LOOCV). As the dataset is imbalanced (318 conversations with performance gain and 182 without), we evaluate our models using the area under the ROC curve. For these experiments we considered 4 types of features:(i) interaction (SC Interaction) and (ii) linguistic (SC Linguistic) features, borrowed from StreetCrowd (Niculae and Danescu-Niculescu-Mizil, 2016), (iii) participation dynamics (i.e. whether one of the participants dominated the conversation), and finally (iv) conversational statistics (number of messages, tokens, etc.). Full experimental details can be found in Appendix A.3 and the code will be made publicly available. As shown in Table 5, the interaction features from StreetCrowd don't transfer well in our setup, if used alone, achieving performance that is below the baseline. On the other hand, SC Linguistic features together with participation features, achieve fair stand-alone performance. Finally, without feature combinations, conversational statistics are the best predictor of conversational performance. Interestingly, the best performance from feature combinations is achieved by using the interaction features from StreetCrowd, the participation dynamics and the conversational statistics. Both SC Interaction and Participation Dynamics, model how participants interact with each other, providing a glimpse into group collaboration. These results suggest that conversational dynamics are a strong addition to traditional feature-based approaches for dialogue classification. On table 5 we also report model stability, which is the consistency of the selected features in the first two levels of the decision tree. While SC Interaction and Participation Dynamics by themselves are not as stable as other feature sets, the best combination achieves perfect stability, by producing consistent decision trees in every split of the LOOCV.

Features	AUC	Stability
[0] Majority Baseline	0.5	
[1] SC Interaction	0.49	0.848
[2] SC Linguistic	0.57	0.975
[3] Participation Dynamics	0.61	0.886
[4] Statistics	0.65	0.997
Best [1] + [3] + [4]	0.68	1

Table 5: Predicting conversation performance

6.4 Generating Probing Utterances

We conclude by developing and evaluating two methods for generating probing utterances. We consider two different approaches - a retrieval-based approach and a generative approach with language models. The task setup is: given the previous dialogue utterances and the **Role** of a probing utterance (i.e. Probing-Moderation, Probing-Reasoning, Probing-Solution), generate the most appropriate utterance to continue the dialogue. For these experiments, we consider the 50 annotated dialogues using the annotation schema of Section 5 as we assume the Role of the utterance to be generated given, and split them into a training set of 30 dialogues and a test set of 20. In our experiments we compare 4 candidate responses:

- **Original.** We take the utterance by the human participant from the original dataset.
- **Random.** We sample from the training data a random utterance that has the same **Role** as the one we need to generate. This is a strong baseline, as sampling for the same role often yields a contextually adequate utterance (albeit not necessarily the best).
- **Retrieval.** We find the most similar utterance with the same Role in our training dataset. To calculate similarity we encode the context of the probing utterance using a pretrained DialoGPT model
- **Generative** We use a pretrained DialoGPT to generate the next utterance based on the current conversation context.

For every method (except for the original) we replaced with placeholders both the mentions of participants and solutions. Once we generate an utterance, if it has a mention of a participant or a solution, we use a simple rule-based system to select appropriate substitution from the context. We

Context	but if we are trying to verify then maybe we select them all
Original	how else could you know?
Random	Why did you press V
Retrieval	How many cards do you think at minimum we need to flip to confirm the rule
Generative	I think he means that the list of possible candidates is a list that will be evaluated in the upcoming days.

Table 6: Utterances generated by different methods

Method	BLEU-4	Similarity	BERT Score
Retrieval	0.39	0.56	0.83
Random	0.35	0.55	0.83
Generative	0.09	0.42	0.79

Table 7: Automatic evaluation of Probing generation

Original	Retrieval	Random	Generative	
-	0.5	0.46	0.28	**Original**
0.5	-	0.48	0.29	**Retrieval**
0.54	0.52	-	0.27	**Random**
0.72	0.71	0.73	-	**Generative**

Table 8: The table reports pairwise preferences in columns over rows, i.e. the first column reports the preference of the Original text vs the other 3 methods.

show an abridged example from our experiments in Table 6 (additional examples in Appendix C). We evaluate the three generated candidate responses using both automatic and human evaluation.

First we applied three commonly used measures for evaluating NLG applications - BLEU 4 (Papineni et al., 2002), sentence similarity using RoBERTa (Liu et al., 2019), and BERTScore (Zhang et al., 2019). As none of our NLG methods is trained to generate the *same* utterance as the Original, we do not expect that any of the candidate responses will achieve strong results, but automatic measures for NLG evaluation can be a good proxy for the quality of generated responses. On Table 7, we present the results where we compare to the Original response. The **Retrieval** approach has the best overall performance, with BLEU-4 score of 0.39 compared to 0.35 and 0.09. If we consider just the Similarity and BertScore measures, the **Retrieval** and **Random** approaches have similar performance. On the other hand, **Generative** performs consistently worse on all measures.

We also perform a human evaluation study, where we asked people to rate the generated responses. We recruited 28 workers from Prolific using comparable worker qualifications and payment level as on MechanicalTurk. We gave crowd workers the following instructions: "Please rank the 4 candidate responses from 1 (for the best response) to 4 (for the worst). You can give the same rank for responses you consider equally good/bad by placing them in the same box.". We asked each of the crowd workers to rank 10 sets of candidate responses, which resulted in 280 annotations of 89 probing cases. First, we compared the average ranks of each of the NLG methods. The Original and the Retrieval approaches had similar ranks - 2.12 and 2.15, while the Random candidate was ranked on average at 2.23. Finally, the generative approach performed the worst, being ranked on average at 3.02. To gain a more fine-grained understanding on which method is preferable, we calculated the pairwise preferences (adjusted for ties), presented in Table 8, which showed similar results, with the Original and Retrieval being considered equal, followed closely by Random, and Generative a distant fourth.

Qualitative analysis showed that the responses of the Retrieval are coherent despite the simple representation of dialogue context. Also, we found that, while large-scale pre-trained language models can be adequate in responding to general queries, they fail to produce good responses where more advanced vocabulary and reasoning are required.

7 Conclusion and Future work

In this work, we introduced a dataset containing conversations where a group of participants collaborate in order to solve a task. Furthermore, we proposed an annotation schema and annotated corpus that capture key elements of group deliberation, such as probing. This dataset can be analysed to test theories of the dynamics of group deliberation and develop dialogue agents that could be used to improve the outcome in numerous setups, for example debating groups, project meetings, etc., and thus a step towards addressing the call for "discourse optimization" of Vecchi et al. (2021). Such dialogue agents can roughly be decomposed into 3 major modules - determining intervention timing, intervention type (i.e. moderation, probing for reasoning) and generating a probing utterance. Given that we present an adequate approach for probing generation, we advise that future researchers focus on the first 2 modules.

8 Ethics Statement

In this work, we present a corpus containing conversations, where participants collaborate to solve a cognitive task. Details on our setup and ethical considerations are presented in Section 3 and appendices A.1 and A.2, but in this section we will reiterate the most important points.

We collected our dataset using the crowdsourcing platform MechanicalTurk and in-lab volunteers for the initial experiments. Participants gave informed consent to their participation, and we told them the purpose of the study and that the transcripts of the dialogues would be collected and used for further research. The only language used in our dataset is English. Participants were free to withdraw at any time. We asked participants not to share any personal information, and as part of quality control, we have removed any instances of such (like the city they were living in, or the institution they were studying in). We asked the participants not to use any offensive language, and as part of the quality control, we verified whether this is the case, fortunately not finding any such instances. When recruiting participants, we selected adult participants from countries where English is a primary language and where MechanicalTurk operated at the time of collection: US, Canada, UK, Ireland, Australia. Besides that, we did not put any restrictions on (nor have a record of) participants' exact age, gender, nationality, race, political leaning, education, etc.

Crowd workers were paid on average between £12/hour and £18/hour (approx. $16.46/h-$24.68/h), depending on their time of participation and whether they solved the task correctly. This is well above the UK's living wage (£8.91/hour), as well as the minimum wage in the US ($7.25) [2]. Moreover, in cases where we were unable to start the data collection (due to inactive users for example), we paid the participants for their time.

For our human evaluation experiments, we recruited participants from Prolific. We put similar qualification requirements as on MechanicalTurk, namely, minimum age of 18, fluent in English, and minimum approval rate of 90%. We paid annotators in the same pay range as on MechanicalTurk, averaging £14.25/hr (19.5$/h).

The full experimental design was approved by the ethics committee of the authors' institution. We will release the DeliData corpus under Creative Commons 4.0.

Limitations While this work aims to investigate how people collaborate in order to solve a task, we limit the scope of our dataset and experiments to the Wason Card Selection Task. Future work may be needed to evaluate whether this dataset would apply to other types of problem-solving (for example in a business setting).

[2]https://www.dol.gov/general/topic/wages/minimumwage

References

Stergos Afantenos, Nicholas Asher, Farah Benamara, Anais Cadilhac, Cedric Dégremont, Pascal Denis, Markus Guhe, Simon Keizer, et al. 2012. Modelling strategic conversation: model, annotation design and corpus.

Khalid Al-Khatib, Henning Wachsmuth, Kevin Lang, Jakob Herpel, Matthias Hagen, and Benno Stein. 2018. Modeling deliberative argumentation strategies on Wikipedia. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2545–2555, Melbourne, Australia. Association for Computational Linguistics.

Anne H. Anderson, Miles Bader, Ellen Gurman Bard, Elizabeth Boyle, Gwyneth Doherty, Simon Garrod, Stephen Isard, Jacqueline Kowtko, Jan McAllister, Jim Miller, Catherine Sotillo, Henry S. Thompson, and Regina Weinert. 1991. The hcrc map task corpus. *Language and Speech*, 34(4):351–366.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gasic. 2018. Multiwoz-a large-scale multi-domain wizard-of-oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, et al. 2005. The ami meeting corpus: A pre-announcement. In *International workshop on machine learning for multimodal interaction*, pages 28–39. Springer.

Jean Carletta, Amy Isard, Jacqueline Kowtko, and G Doherty-Sneddon. 1996. *HCRC dialogue structure coding manual*. Human Communication Research Centre.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating amazon's mechanical turk as a tool for experimental behavioral research. *PloS one*, 8(3):e57410.

Heriberto Cuayáhuitl, Simon Keizer, and Oliver Lemon. 2015. Strategic dialogue management via deep reinforcement learning. *arXiv preprint arXiv:1511.08099*.

Maira Gatti de Bayser, Paulo Cavalin, Claudio Pinhanez, and Bianca Zadrozny. 2019. Learning multi-party turn-taking models from dialogue logs. *arXiv preprint arXiv:1907.02090*.

Christine De Kock and Andreas Vlachos. 2021. I beg to differ: A study of constructive disagreement in online conversations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2017–2027.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered conversational agents. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

He He, Derek Chen, Anusha Balakrishnan, and Percy Liang. 2018. Decoupling strategy and generation in negotiation dialogues. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2333–2343.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*.

Yiqing Hua, Cristian Danescu-Niculescu-Mizil, Dario Taraborelli, Nithum Thain, Jeffery Sorensen, and Lucas Dixon. 2018. Wikiconv: A corpus of the complete conversational history of a large online collaborative community. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2818–2823.

Simon Keizer, Markus Guhe, Heriberto Cuayáhuitl, Ioannis Efstathiou, Klaus-Peter Engelbrecht, Mihai Dobre, Alex Lascarides, Oliver Lemon, et al. 2017. Evaluating persuasion strategies and deep reinforcement learning methods for negotiation dialogue agents. ACL.

M. G. Kendall. 1938. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93.

Soomin Kim, Jinsu Eun, Joseph Seering, and Joonhwan Lee. 2021. Moderator chatbot for deliberative discussion: Effects of discussion structure and discussant facilitation. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. 2017. Deal or no deal? end-to-end learning of negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453.

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

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? arguments for an argumentative theory. *Behavioral and brain sciences*, 34(2):57–74.

David Moshman and Molly Geil. 1998. Collaborative reasoning: Evidence for collective rationality. *Thinking & Reasoning*, 4(3):231–248.

Joaquin Navajas, Tamara Niella, Gerry Garbulsky, Bahador Bahrami, and Mariano Sigman. 2018. Aggregated knowledge from a small number of debates outperforms the wisdom of large crowds. *Nature Human Behaviour*, 2(2):126–132.

Vlad Niculae and Cristian Danescu-Niculescu-Mizil. 2016. Conversational markers of constructive discussions. In *Proceedings of NAACL-HLT*, pages 568–578.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Stefan Schulz-Hardt, Felix C Brodbeck, Andreas Mojzisch, Rudolf Kerschreiter, and Dieter Frey. 2006. Group decision making in hidden profile situations: dissent as a facilitator for decision quality. *Journal of personality and social psychology*, 91(6):1080.

Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of Language and Social Psychology*, 29:24 – 54.

Gokhan Tur, Andreas Stolcke, Lynn Voss, Stanley Peters, Dilek Hakkani-Tur, John Dowding, Benoit Favre, Raquel Fernández, Matthew Frampton, Mike Frandsen, et al. 2010. The calo meeting assistant system. *IEEE Transactions on Audio, Speech, and Language Processing*, 18(6):1601–1611.

Eva Maria Vecchi, Neele Falk, Iman Jundi, and Gabriella Lapesa. 2021. Towards argument mining for social good: A survey. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers), pages 1338–1352, Online. Association for Computational Linguistics.

Peter C Wason. 1968. Reasoning about a rule. *Quarterly journal of experimental psychology*, 20(3):273–281.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing online discussion using coarse discourse sequences. In *11th AAAI International Conference on Web and Social Media (ICWSM)*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with BERT. *CoRR*, abs/1904.09675.

A Reproducibility Checklist

A.1 Data Collection - Participant Instructions

Participants are given the following description of the task and experiment:

1. You will be part of a small-group chat (3-5 people), where you will try solving a puzzle.
2. Finish the task by yourself
3. Participate in a group discussion (via the chat), collaborate with the other participants and try to find the best solution together. Give your best effort both in solving the task and in the group discussion.
4. You are expected to participate actively in the conversation for at least 5 minutes.
5. Based on the discussion and arguments you had, submit the revised task solution again. You can submit the same answer if you believe it's the correct one.
6. **Task:** Each of the 4 cards below has a letter on one side and a number on the other. Which card(s) do you need to turn to test the rule: All cards with vowels on one side have an even number on the other. NB: Select ONLY the card(s) required to verify the rule. Most people get this task wrong.
7. Please remember that these transcripts may be used in future research, and therefore you have the right to withdraw from this study at any given time. To do so, press the "Leave room" button above. Please ensure you do not use any offensive language or disclose any personal information which would make you identifiable to others as it's important that your anonymity is maintained. Any information which may reveal your identity will be deleted from this chat.

A.2 Data Collection: Mechanical Turk Modifications

We recognise that collecting data on Mechanical-Turk, we will face more challenging conditions compared to a controlled lab setup. Moreover, by design, MechanicalTurk is providing a platform for a single person to complete a task. As we aim at collecting group dialogues we applied to following recruitment protocol that enables synchronous data collection between multiple turkers:

1. **Room Routing**. Every crowd worker that joins our task is routed to a group that is recruiting participants or if none available - creates a new room. As we recognise, that some participants might leave after joining a room, we identified the following 3 room states:
   (a) **Recruiting**: if the room has less than 3 active participants, a new participant can join at any time
   (b) **Final Call**: After there are at least 3 people in the room, a 1-minute timer starts, which allows for up to 2 more participants to join. By allowing more than 3 people to join, we mitigate the effect of inactive or leaving participants.
   (c) **Ready to Start**: Once the **final call** timer elapses, the game is ready to start.
2. **Crowd worker requirements**. To get high-quality data collection, the crowd workers participating in our task should meet the following conditions:
   (a) Complete a simple reading comprehension test
   (b) Fluency in English, which is established by being a resident of countries where English is an official language
   (c) Have more than 95% success rate on previous crowd-sourcing tasks
   (d) Have completed at least 1000 tasks on Mechanical Turk
3. **Notifications**. Sometimes it takes a while for a group of 3 people to be ready, and, naturally, some of the participants may be inactive while waiting. To ensure that everyone is online, when the group is ready to start, there are audible notifications during key phases of the experiment, as well if someone is being inactive or not responsive during the game.
4. **Quality Control**. We perform two kinds of quality control over the collected data. Initially, we automatically exclude all conversations that either have only a single participant in them or have less than 10 messages. Then, each conversation is manually checked, to ensure that no personal information was shared.

Finally, we excluded conversations based on poor quality, i.e. when participants are not discussing the task at all. That said, participants are still getting paid if the conversation was excluded to no fault of their own.

A.3 Predicting Performance Gain

To encourage reproducibility we will describe in details how we predict performance gain.

Conversation Statistics (9 features): Number of participants in the chat, total number of messages, average number of messages per player, average number of tokens per player, total unique tokens, average unique tokens per player, participants' individual performance, diversity in participants' individual solutions, and group consensus.

Participation Dynamics (13 features). In the context of this work, we built a solution and participation tracker. Knowing the cards, presented to the participants, we track each solution proposal, as well as per participant change of solution. We do this by applying a simple rule-based system - if the message mentions one or more of the cards we save this as participant's solution proposal. Next time the same participant proposes a different solution we mark this event as a solution change.

Complimentary to the solution tracker, we also keep a record of how actively each participant engages in the discussion. We identify 4 categories of participation, based on how many messages each player issued - 0, 0-20, 40-50, 50-100 %. Thus we are able to record both more silent users, and those who participate more than the rest of the group.

That said, we extract the following features: Number of solution changes (normalised by the number of messages), The 4 categories of participation at 20/50/all messages.

StreetCrowd Features For more details, please refer to (Niculae and Danescu-Niculescu-Mizil, 2016).

- **Interaction Features** (6 features). These features are calculated based on the whole conversation (rather than on an individual message). First, (Niculae and Danescu-Niculescu-Mizil, 2016) include language matching on stopword, token and POS tag levels. Further, the interaction features capture agreement and disagreement markers in words.

- **Linguistic Features** (15 features). These are message level features, that capture specific linguistic phenomena: message length (and it's variation), psycholinguistic features from LIWC (Tausczik and Pennebaker, 2010), task specific jargon, and POS patterns.

Model Selection and Hyperparameter Search. Due to the relatively small size of the dataset, and the high information load of each conversation (large number of utterances), the selection of an appropriate model is a challenging endeavour. In our experiments, we found out that most models are either unable to generalise well or are very unstable in terms of performance. Models that performed poorly in either generalisation or stability were: Linear Regression, Support Vector Machine (both linear and RBF kernels), RandomForest, K-Nearest Neighbour, and a multilayer perceptron. Thus, we selected a decision tree, as it is a fairly stable model by design, and it allows us to analyse variability between different runs of the model. We performed hyperparameter search with the following parameters: Max Depth: [2, 3, 5, 7 (selected), 20, max] and Min Samples per leaf: [1, 2, 3, 5 (selected), 10]. Total number of parameter tuning runs - 30. The best model is selected based on model accuracy and stability. Due to the size of the model and the dataset, the hyperparameter search does not require any special infrastructure and the training time is negligible.

A.4 Packages used

For training and evaluation of the performance gain we used (Pedregosa et al., 2011) version 1.0.2. For general language tasks and featurisers we used NLTK (Bird et al., 2009) version 3.5, Spacy (Honnibal and Montani, 2017) version 2.3.2. For generative experiments, we used DialoGPT-large from HuggingFace's transformers version 4.11.3.

For evaluation, we used BertScore (Zhang et al., 2019) version 0.3.11, Sentence Transformers version 2.1.0.

B Example of a constructive and non-constructive conversation

User	Utterance	Is Probing	Role	Additional Labels
Alpaca	What did everybody put?	Probing	Moderation	
Leopard	I put 6 and S, how about you?	NPD	Solution	complete_solution
Alpaca	Oh, i thought we could only chose one card. I chose A	NPD	Solution	complete_solution
Alpaca	Why did you choose	Probing	Reasoning	
Tiger	I put 6 - to see if has a vowel on the other side A to see if it has an even number and 7 to see if it has a consonant	NPD	Reasoning	complete_solution
Alpaca	6 and S	NPD	Solution	complete_solution
Tiger	I mean a vowel on 7	NPD	Reasoning	partial_solution
Tiger	as if it is a vowel the rule wouldn't apply	NPD	Reasoning	partial_solution
Tiger	@Alpaca why do you think you need to turn s?	Probing	Solution	specific_addressee, partial_solution
Leopard	Okay I put 6 because I thpught we need to check if there's a vowel on the other side, and then S to make sure there's not an even number on that	NPD	Reasoning	complete_solution
Alpaca	No i would only turn A	NPD	Disagree	complete_solution
Alpaca	i would not choose 6 as the rule is not whether all even numbers have a vowel on the back, its if all vowels have an even number on the back	NPD	Reasoning	complete_solution
Leopard	Actually yeah I change my answer to A and 7	NPD	Agree	complete_solution
Tiger	Actually - do we need 6? it doesn't matter if it has a vowel or not	NPD	Solution	partial_solution
Alpaca	so definitely A...	NPD	Solution	partial_solution
Alpaca	and i think 7	NPD	Solution	partial_solution
Leopard	Don't we need to check 7 to make sure it doesn't have a vowel?	Probing	Solution	partial_solution
Alpaca	Yes, I agree	NPD	Agree	
Tiger	Definettly A and I think 7 too	NPD	Solution	complete_solution
Leopard	Okay final answer A and 7 then?	Probing	Solution	solution_summary, complete_solution
Alpaca	Do we all agree on 7 and A?	Probing	Solution	solution_summary, complete_solution
Tiger	yes	NPD	Agree	

Table 9: Constructive conversation ending in a correct solution

User	Utterance	Is Probing	Role	Additional Labels
Beaver	I think we should check all four cards.	NPD	Solution	complete_solution
Bee	I am going with the last 2	NPD	Solution	complete_solution
Narwhal	At the very least we should definitely include the 3rd card.	NPD	Solution	partial_solution
Beaver	Ok, anything else?	Probing	Moderation	
Bee	Why A?	Probing	Reasoning	
Narwhal	The rule is that all cards with a vowel on one side have an even number on the other side.	NPD	Solution	
Narwhal	Well, our third card is a vowel to start with. We do not know what is on the other side of that card. If we flip our only apparent vowel and we find an even number, that is a pretty good indication to the rule right off the start.	NPD	Reasoning	partial_solution
Beaver	ok	NPD	Agree	
Bee	makes sense	NPD	Agree	
Narwhal	None of the other cards would do us any good to flip them over because they are either an odd number or a consonant.	NPD	Reasoning	
Narwhal	So A is the way to go.	NPD	Solution	complete_solution
Bee	sounds good to me.	NPD	Agree	
Beaver	A it is,	NPD	Agree	complete_solution
Bee	Thanks for the help,			
Narwhal	Thanks for being willing to listen!			

Table 10: Non-constructive conversation

C Examples of different approaches to generating utterances

Context	
Narwhal	Hello
Dolphin	Hi
Original	Anyone have any suggestion to a solution
Random	Dolphin what did you select
Retrieval	so what we are supposed to discuss about
Generative	hey

Table 11: Example of different methods for generating Probing-Moderation utterances

Context	but it says it might be as simple as we think and it seems pretty simple to put U and 2 as that is the vowel and the even number
Original	So is it 7 ?
Random	so 2 , U , and 7
Retrieval	So you think the 2 Card ?
Generative	I concur

Table 12: Example of different methods for generating Probing-Solution utterances