“Any Other Thoughts, Hedgehog?” Linking Deliberation Chains in Collaborative Dialogues

Abhijnan Nath, Videep Venkatesha, Mariah Bradford, Avyakta Chelle,
Austin Youngren, Carlos Mabrey, Nathaniel Blanchard, Nikhil Krishnaswamy
Situated Grounding and Natural Language (SIGNAL) Lab,
Department of Computer Science, Colorado State University, Fort Collins, CO USA
firstname.lastname@colostate.edu

Abstract

Question-asking in collaborative dialogue has long been established as key to knowledge construction, both in internal and collaborative problem solving. In this work, we examine *probing questions* in collaborative dialogues: questions that explicitly elicit responses from the speaker’s interlocutors. Specifically, we focus on modeling the causal relations that lead directly from utterances earlier in the dialogue to the emergence of the probing question. We model these relations using a novel graph-based framework of *deliberation chains*, and reframe the problem of constructing such chains as a coreference-style clustering problem. Our framework jointly models probing and causal utterances and the links between them, and we evaluate on two challenging collaborative task datasets: the Weights Task and DeliData. Our results demonstrate the effectiveness of our theoretically-grounded approach compared to both baselines and stronger coreference approaches, and establish a standard of performance in this novel task.

1 Introduction

Recent breakthroughs in generative AI have raised the possibility of systems that follow and interact with multiparty dialogue. Inherent in group dialogues are utterance sequences that deliberate on the same information. Modeling these is particularly challenging; while such utterances have a linear order and overlapping information, they may be distantly separated in time and the same information may be expressed very differently.

In this paper, we construct **deliberation chains** in dialogue: turn sequences that surface pieces of evidence or questions under discussion that culminate in a “probing utterance,” or explicit elicitation of input that does not introduce new information. We model deliberation chains as *probing interventions* that are preceded somewhere in the discourse by a number of *causal interventions*, each of which contribute directly to the eventual emergence of the utterance that serves as the probing intervention. Without the causal counterpart(s), the probe would not arise in the discourse (at least not in that specific form or at that specific time).

Both probe and cause are linked to effective group performance (Karadzhov et al., 2023). Tracking them requires an evolving understanding of collaborative dynamics and enables disagreement detection, prompting for deeper insights, or analysis of deliberation’s influence on individual learning/understanding (Hunter et al., 2018; Atwell et al., 2024; Khebour et al., 2024b).

Our novel approach takes inspiration from discourse coherence theory and joint modeling frameworks traditionally applied to coreference resolution. The ability to link probing interventions to their preceding causes in the dialogue is a critical prerequisite for AI systems to support deliberative/collaborative reasoning, which is of interest in domains like education and workforce development. Our novel contributions are:

- A novel task of automatically constructing “deliberation chains” of probing questions in a dialogue and with their causal utterances;
- A formal graphical framework for deliberation chains derived from formal semantics of situated conversation (Hunter et al., 2018);
- A unique adaptation of methods from coreference resolution to this new task;
- Baseline evaluation on two challenging collaborative dialogue datasets—DeliData and the Weights Task Dataset—and a novel method of jointly modeling probing and causal interventions and the links between them.

Our code may be found at: https://github.com/csu-signal/ProbingDelibration

Findings of the Association for Computational Linguistics: EMNLP 2024, pages 5297–5314
November 12-16, 2024 ©2024 Association for Computational Linguistics

2 Related Work

Collaborative Dynamics Andrews-Todd and Forsyth (2020), OECD (2017), and Sun et al. (2020) have all identified the need for teams to construct shared knowledge to function, often through asking questions. Hesse et al. (2015) also points out that collaboration requires teammates to initiate interaction. Further, Fusaroli et al. (2017) identified conversational repair as a necessary mechanism in forming common ground for a group. Graesser et al. (2018) describes the need for team members to externalize their knowledge. Karadzhov et al. (2022) explores how deliberation may lead to team members changing their minds, which is critical for building group common ground (Stalnaker, 1978).

Joint Modeling in Coreference Resolution In the well-studied problem of coreference resolution, many works (Lee et al., 2017; Zhang et al., 2018; Cattan et al., 2021; Yu et al., 2022) have proposed various joint modeling frameworks and cross-encoding architectures that optimize on coreference link assignments and building mention clusters, including ideas to make such methods more scalable (Ahmed et al., 2023; Held et al., 2021) and generalizable (Bugert et al., 2021). In contrast to such methods that often operate on a "span"-level and require exhaustive cross-computations (Thirukovalluru et al., 2021), we generate deliberation chains using utterances as distinct discourse units.

Free-Text Rationales With the advent of instruction-tuned generative LLMs like InstructGPT (Ouyang et al., 2022), recent works (Ahmed et al., 2024; Wang et al., 2024; Radhakrishnan et al., 2023; Zhao et al., 2023) have leveraged their Chain-of-Thought (COT)-style reasoning capacities for various NLP tasks like argument extraction, question-answering as well as coreference annotations, often guiding the LLM's reasoning process using Free-Text Rationales (FTRs) that explicate reasoning steps toward a decision (Wiegreffe et al., 2021; West et al., 2022; Nath et al., 2024). Our work uses such FTRs to guide the automatic annotation of interventions in collaborative task datasets.

3 Problem Formulation

Segmented Discourse Representation Theory (SDRT) posits that interpreting an utterance involves supplementing its semantics with pragmatic content based on the demands of *discourse coherence* (Asher and Lascarides, 2003). The relation between utterances and prior content required for a full interpretation gives rise to structures which in collaborative dialogues represent the evolution of information that propels such dialogues towards task-completion (Karadzhov et al., 2023). Let us define the relevant structures below:

Definition 1. Based on Hunter et al. (2018), let $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2, \lambda)$ be a **deliberation graph** in a collaborative dialogue, that in turn comprises sets of individual deliberation chains. $\mathcal{G}$ is characterized as a weakly-connected, weighted, acyclic graph. Here, $\mathcal{V}$ represents vertices for probing ($\mathcal{P}$) and causal ($\mathcal{C}$) interventions[1]; edges $\mathcal{E}_1$ denotes connectivity between vertices; weights $\mathcal{E}_2$ indicate causal influence from $\mathcal{C}$ to $\mathcal{P}$, thereby establishing a total order; and λ is a directed path induction function over $\mathcal{E}_2$ and a vertex $v \in \mathcal{V}$ that emits the root intervention $\mathcal{C}$ and terminal intervention $\mathcal{P}$ in $\mathcal{G}$, implicit in the discourse's linear order.

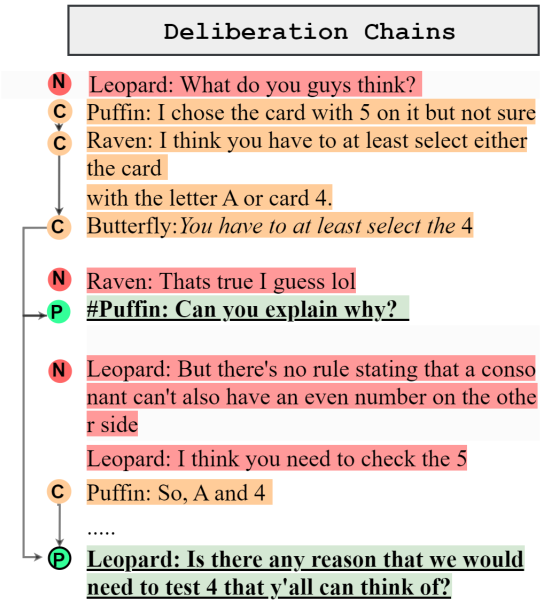
Definition 2. Given a deliberation graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}_1, \mathcal{E}_2, \lambda)$, a **deliberation chain** (or *intervention cluster*[2]) is a subgraph $\mathcal{G}' = (\mathcal{V}', \mathcal{E}_1', \mathcal{E}_2', \lambda)$ of $\mathcal{G}$, such that $\{\mathcal{P}_{\hat{i}}, \mathcal{C}_{\hat{j}}\} \subseteq \mathcal{V}'$, where $\hat{j} = \min\{j \mid \lambda(\mathcal{C}_j) \in \mathcal{V}'\}$ and $\hat{i} = \max\{i \mid \lambda(\mathcal{P}_i) \in \mathcal{V}'\}$ indicate the initial and final occurrences respectively in the traversal of $\mathcal{G}$ from $\mathcal{C}_{\hat{j}}$ to $\mathcal{P}_{\hat{i}}$. See Fig. 4.

We formulate deliberation chain construction as a coreference resolution-style clustering problem (Ng and Cardie, 2002; Lee et al., 2012), over a dialogue, D, with N utterances, that the system must cluster into probing interventions and their linked causes, such that each cluster forms a unique deliberation chain. Given the elements of a cluster, λ reconstructs the chain by enforcing transitive closure over the within-cluster links given the temporal order inherent in the discourse, under Definition 1 above. This formulation motivates our joint modeling approach, which is detailed in Sec. 5.

In Fig. 1, we provide a detailed example of a deliberation chain from our dataset. The causal interventions (*e.g., "You have to at least select either the letter A or card 4."*) and probing questions

[1] Past probing interventions ($\mathcal{P}_{<i}$) likely influence current and future ones ($\mathcal{P}_i$), ensuring weak connectivity, and any $\mathcal{P}$ cannot be the cause of its own $\mathcal{C}$, thereby guaranteeing acyclicity. This structure reflects the linear progression typical in turn-based dialogues. Potential non-linearities in multimodal contexts (Hunter et al., 2018) largely do not affect the acyclic structure because multimodal channels tend to overlap rather than invert the linear order of dialogue entirely (Alahverdzhieva et al., 2017).

[2] We will use *deliberation chain* for the ordered sequence of interventions in a dialogue, and *intervention cluster* for the clusters output by our system. Both denote a chain of sequential interventions linked by transitive closure, similar to *entity clusters* in coreference literature.



Leopard: What do you guys think?
Puffin: I chose the card with 5 on it but not sure
Raven: I think you have to at least select either
the card
with the letter A or card 4.
Butterfly:*You have to at least select the* 4

Raven: Thats true I guess lol
#Puffin: Can you explain why?

Leopard: But there's no rule stating that a conso
nant can't also have an even number on the othe
r side
Leopard: I think you need to check the 5
Puffin: So, A and 4
.....
**Leopard: Is there any reason that we would
need to test 4 that y'all can think of?**

Figure 1: Example of a deliberation chain, showing the flow of interventions and their causal relationships within a collaborative task. This example is adapted from our model's output on the DeliData corpus.

(*e.g., "Can you explain why?"*) form a structured sequence, where probing interventions are linked to their causal antecedents. This transitive closure forms the deliberation chain, which reflects how participants navigate the problem-solving process.

4 Dataset Annotation

We evaluate intervention clustering on two recent, challenging collaborative dialogue datasets: DeliData and the Weights Task Dataset.

4.1 DeliData

The DeliData corpus (Karadzhov et al., 2023) is a publicly-available dataset intended for studying group deliberation in multiparty problem-solving. It comprises 500 group dialogues, totaling 14,003 utterances, centered around the Wason card selection task, a well-established cognitive puzzle (Wason, 1968). Each group contains 5 participants, who are presented with 4 cards that have a number or a letter on them. They must collectively decide which cards to turn over to test the rule, "All cards with vowels on one side have an even number on the other?" The dataset includes both the dialogues themselves, which denote cards by the symbols on them (letters or numbers), and a measure of decision correctness (task performance) before and after the group discussion, and is annotated with deliberation cues, argumentation structures, and other conversational dynamics. DeliData splits consist of 300, 100, and 100 randomly-chosen groups for training, development, and testing, respectively.

4.2 Weights Task Dataset

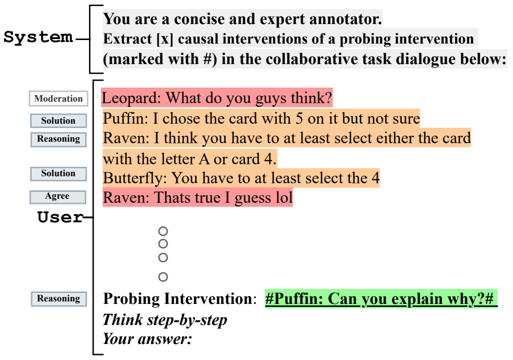
The Weights Task Dataset (WTD) (Khebour et al., 2024a) is an anonymized publicly-available dataset intended for studying small group collaboration. It comprises 10 videos, where groups of three participants must use a balance scale to identify the weights of differently-colored weighted blocks and the pattern that describes the weights. The task unfolds in 3 stages, where users solve the problem with the scale, without the scale, and with inferred knowledge of the pattern in weights. The dataset includes multiple annotations, including human gold-standard transcriptions of the participants' dialogues. Utterances reference blocks by color and deduced candidate weights, and can be used to identify probing questions and their potential causal interventions. WTD splits consist of 7, 1, and 2 randomly-chosen groups for training, development, and testing, respectively.

4.3 Data Augmentation of WTD

The WTD is a multimodal dataset, but as the focus of this paper is establishing this novel task, our current study does not incorporate non-verbal cues. Instead, we employ *dense paraphrasing* (Tu et al., 2023) as an augmentation technique to explicitly define which blocks are being referred to in the situated dialogue, so that probing and causal interventions can be modeled using just a textual signal. The WTD annotations include dense paraphrased utterances for the first stage but not the second two. We followed the procedure from Khebour et al. (2024b) to dense paraphrase the remainder of the dataset (e.g., replacing "those" with "red block and blue block" in cases where the video makes clear that those blocks are the intended denotata). Utterances were dually annotated (Cohen's $\kappa = 0.69$) and adjudicated by an expert.

4.4 GPT Annotations of Deliberation Chains

Like coreference cluster annotation, which often requires exhaustive cross-comparisons across tokens (Bugert et al., 2020), human annotation of deliberation chains is time-consuming and expensive. Therefore, to create "gold" chains for fair



You are a concise and expert annotator.
Extract [x] causal interventions of a probing intervention (marked with #) in the collaborative task dialogue below:

Leopard: What do you guys think?
Puffin: I chose the card with 5 on it but not sure
Raven: I think you have to at least select either the card with the letter A or card 4.
Butterfly: You have to at least select the 4
Raven: Thats true I guess lol

Probing Intervention: #Puffin: Can you explain why?#
Think step-by-step
Your answer:

Figure 2: Prompting framework for GPT to select causal interventions given a probing intervention and a dialogue history (example from DeliData). Ground-truth labels for probing and causal interventions are marked in green and brown, respectively.

comparison, we draw on work in LLM-augmented annotations with Chain-of-Thought (COT) reasoning (Radhakrishnan et al., 2023; Wei et al., 2023; Nath et al., 2024) for "soft" gold labels.

We apply a two-pronged strategy. (1) We sequentially prompt `GPT-3.5-turbo-0125` using an argument-extraction framework (Ahmed et al., 2024) (see Fig. 2) to extract causal interventions for all probing interventions in the data[3] with prior dialogue history[4] and a system-based task-description to guide its reasoning. We also explicitly ask the LLM to generate free-text rationales (FTRs) corresponding to every causal intervention extracted, to augment its reasoning (Kunz et al., 2022; Ravi et al., 2023). (2) We do an extensive human evaluation of these LLM-generated annotations to validate quality of extracted clusters. FTRs were used as an additional reference for human evaluators to validate the GPT's annotations and their alignment with human reasoning. This evaluation demonstrated high acceptability of GPT labels and reasoning to humans (see 4.5 for details).

Since deliberation graphs are weakly-connected, in each iteration we also apply a labeling algorithm (see Algorithm 1 in the appendix) to assign the correct preceding cluster for newly appearing interventions in the loop. Table 1 provides cluster-level details of the two datasets.

4.5 Human Evaluation of GPT-Annotated Labels

We conducted a human evaluation to assess the quality of the GPT-generated annotations on a random representative subset of 25 samples from both DeliData and WTD test sets. These samples were evaluated across several dimensions: relevance, presence in sequence, information sufficiency, acceptability, and rationale overlap.

The annotators consistently agreed that the annotated utterances were indeed causal to the probing utterance, as indicated by high agreement on the first two questions concerning *Relevance to Context* and *Presence in Sequences*. These are the most critical aspects of the evaluation, and the high level of agreement demonstrates that the core annotations were valid. The annotators' answers to questions concerning rationale alignment, however, showed more variability, as expected and seen in Fig. 3. While annotators may agree that an utterance is causal, they may align less with the specifics of the rationale behind why it is causal. This variation is natural and does not impact the overall validity of the annotations.

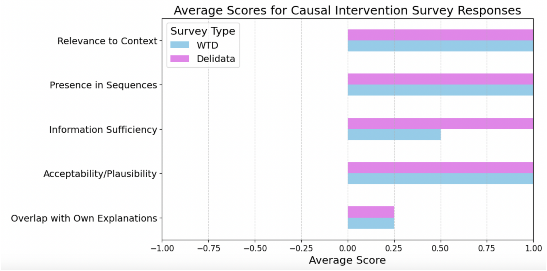


Figure 3: Average Scores for Causal Intervention Survey Responses.

We calculated Krippendorff's alpha to measure inter-annotator agreement. Each unique qualitative response was mapped to distinct numerical categories (e.g., Yes, No, Not enough information, Enough information) to capture the differences between responses more effectively. This calculation resulted in Krippendorff's alpha values of 0.88 for DeliData and 0.92 for WTD, indicating strong agreement between annotators on these samples.

Further details on the evaluation process, can be found in Appendix G.

5 Joint Learning of Deliberation Chains

To automatically cluster interventions that form a deliberation chain $\mathcal{G}'$, a model must learn to assign, for each possible $\mathcal{P}_i$, the most suitable antecedent

[3] For the WTD, which does not already contain probing labels, we use the dense paraphrased utterances to extract probing labels before this step. See Appendix D.

[4] A probing intervention can cause another probing statement within a dialogue (Sukmadewi, 2014; Behr et al., 2012). As such, we do not omit probing labels from the previous utterances given as context.

	DeliData			WTD		
	Train	Dev	Test	Train	Dev	Test
# Probing	1005	317	358	115	15	64
# Causal	1975	637	700	247	37	134
# Total	2980	954	1058	362	52	198
Min Chain Length	2	2	2	2	3	2
Max Chain Length	21	15	14	14	12	13
Mean Chain Length	5.4	5.5	5.1	5.1	5.2	4.7
# Clusters	552	174	206	71	10	42
Avg. Dialogue Length	33	35.3	34.9	222	204	326
Tokens to Probing	227	241	214	293	367	276

Table 1: Descriptive statistics of true (gold) intervention clusters/deliberation chains in DeliData and Weights Task datasets across different splits. Note that there are no singletons in either dataset. "Length" refers to utterances. Last row denotes mean number of tokens from start of a dialogue until a probing intervention.

utterance $\mathcal{C}_j$, that forms a correct link in the chain. Prior works in coreference resolution (Lee et al., 2017; Zhang et al., 2018) typically addressed such assignments using joint-learning frameworks that exhaustively score antecedent "spans" and thereby produce coreference chains. Our approach implicitly produces the correct chain since interventions in a dialogue follow a linear order assuming transitivity across links.

Standard joint-learning frameworks for coreference resolution typically operate at the *span*-level. For our task, where the entire deliberative utterance forms a distinct discourse unit (Hunter et al., 2018), this is an incompatible approach. As such, we propose a joint-learning framework that models the task as a conditional probability distribution $Pr(P, C, L \mid D)$, partitioned into multinomial probabilities, assuming that utterance spans are conditionally independent given the dialogue D. Mathematically,

$$Pr(P,C,L \mid D) = \prod_{i=1}^{N}\prod_{j=1}^{N} Pr(p_i \mid D)Pr(c_j \mid D)Pr(l_{ij} \mid D), \tag{1}$$

where P refers to the probability of an utterance being a Probing intervention, C refers to the probability of an utterance being a Causal intervention, and L refers to the probability of a Link between the two utterances. p_i, c_j and l_{ij} are treated as random variables denoting the probabilities of an utterance being probing, being causal, and of the link between the two interventions, respectively; N denotes the number of individual utterances within a dialogue D.

5.1 Model

Intervention Pair Representation As the right-hand side of Eq. 1 represents causal dynamics as probabilities of links between pairs of utterances in the discourse, we draw on a cross-encoding strategy from coreference research (Humeau et al., 2020; Cattan et al., 2021; Ahmed et al., 2023) to score pairs of utterances. Since some dialogues, especially in the Weights Task Dataset, can reach up to ~200 utterances, we use the Longformer model (Beltagy et al., 2020) as the base encoder. To construct an expressive representation for a pair of interventions ($\mathcal{P}_i$, $\mathcal{C}_j$), we first demarcate their start and end with special tokens (<m> and </m>). For context around a probing intervention, we also concatenate the k previous utterances[5] along with participant name or number as given in the dataset. We extract the [CLS] token representation of this concatenated input, the cross-attentional context of $\mathcal{P}_i$ and $\mathcal{C}_j$, as well as their Hadamard product, $\mathcal{P}_i \odot \mathcal{C}_j$. This results in a combined vector representation for pair ($\mathcal{P}_i$, $\mathcal{C}_j$):

$$V(\mathcal{P}_i, \mathcal{C}_j) = [V_{CLS}, V_{\mathcal{P}_i}, V_{\mathcal{C}_j}, V_{\mathcal{P}_i} \odot V_{\mathcal{C}_j}] \tag{2}$$

Next, to maximize the log-likelihood in our joint-learning framework (Eq. 1), we generate three sets of scores from specific segments of Eq. 2 using three feed-forward neural networks (FFNN): (1) a linking score $l_{ij} = \text{FFNN}_l(V(\mathcal{P}_i, \mathcal{C}_j))$, the probability of a pair of utterances forming a true link; and (2) two intervention scores, $s_i = \text{FFNN}_p(V_{\mathcal{P}_i})$ and $s_j = \text{FFNN}_c(V_{\mathcal{C}_j})$ of the candidate and the antecedent, respectively, being valid interventions.

Thus, the model picks up on two types of learning signals: correctly assigning a true antecedent to a candidate intervention while also learning what constitutes a valid intervention. We directly optimize the model with $\mathcal{L}_{\text{joint}}$:

$$\mathcal{L}_{\text{joint}} = \alpha_{\text{p}}\mathcal{L}_{\text{probing}} + \alpha_{\text{c}}\mathcal{L}_{\text{causal}} + \alpha_{\text{l}}\mathcal{L}_{\text{link}} \tag{3}$$

that consists of a weighted-combination of three separate loss terms. $\mathcal{L}_{\text{probing}}$ and $\mathcal{L}_{\text{causal}}$ are each defined as:

$$\mathcal{L}_{[\text{probing,causal}]}(*) = -\sum_{*=1}^{N} y_* \log(\sigma(s_*)) \tag{4}$$

where $*$ corresponds to i and j in $\mathcal{L}_{\text{probing}}$ and $\mathcal{L}_{\text{causal}}$, respectively, σ is the sigmoid function, and

[5] Setting $k = 10$ and max sequence length (probing intervention with preceding utterances) to 512 was empirically found to cross-encode both utterances in a pair, on average, without losing expressive tokens or incurring inordinate compute cost. See Table 1 for more details.

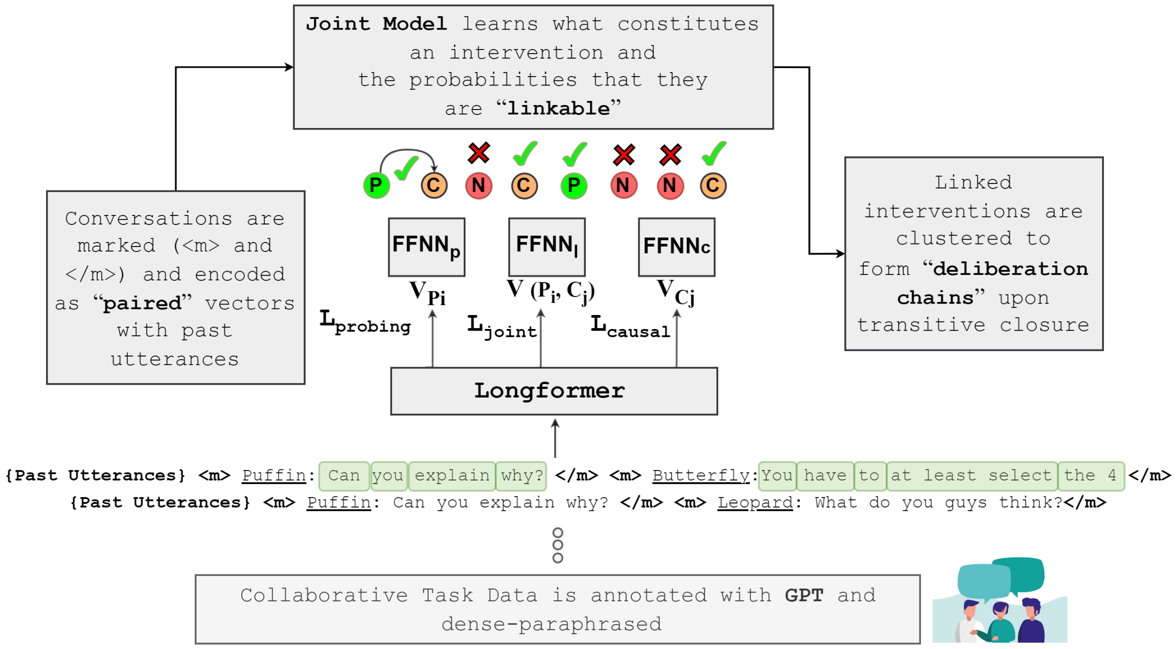

Joint Model learns what constitutes an intervention and the probabilities that they are "linkable"
P
C
N
C
P
N
N
C
FFNN$_p$
FFNN$_l$
FFNNc
V_{Pi}
$V (P_i, C_j)$
V_{Cj}
$L_{probing}$
L_{joint}
L_{causal}
Longformer
Conversations are marked (<m> and </m>) and encoded as "paired" vectors with past utterances
Linked interventions are clustered to form "deliberation chains" upon transitive closure
{Past Utterances} <m> Puffin: Can you explain why? </m> <m> Butterfly: You have to at least select the 4 </m>
{Past Utterances} <m> Puffin: Can you explain why? </m> <m> Leopard: What do you guys think?</m>
Collaborative Task Data is annotated with GPT and dense-paraphrased

Figure 4: Our joint-learning framework for *deliberation chains*, learning to assign correct antecedent utterances for every valid intervention using a "probing" score, a "causal" score, and a "linking" score. Pairs of utterances are encoded with global attention (in green between <m> and </m>), further contextualized by past utterances.

y is the predicted output. The final term is $\mathcal{L}_{\text{link}}$:

$$\mathcal{L}_{\text{link}}(i,j) = -\sum_{i=1}^{N}\sum_{j=1}^{N} y_{ij}\log(l_{ij}) + (1-y_{ij})\log(1-l_{ij}) \tag{5}$$

See Fig. 4 for further details. α_{p}, α_{c} and α_{l} are learned regularization parameters tuned on the development sets of our data. Following Nath et al. (2024), we fixed $\alpha_{\text{l}} = 1$ and α_{c}, $\alpha_{\text{p}} = 0.01$ after initial experiments.

Training Pair Generation For training a pairwise scorer model, an efficient pair generation process is crucial. A naive way to implement Eq. 1 compares each utterance u_i to the set of all its preceding antecedents $U(i) = \{\epsilon,\ u_1, \ldots, u_{i-1}\}$ to generate pairwise scores.[6] This results in $\sim O(N^2)$ complexity for a dialogue of N utterances. Discourse-coherence theory (Grosz and Sidner, 1986; Held et al., 2021) suggests that the most pertinent information to a specific utterance remain *within* an "attentional state", i.e., the point of focus of participants within a dialogue. As such, given a dialogue of N utterances, for each target u_i, we define a window W of previous utterances considered for training. Because of the long tail of true negative samples (non-links), this value is tuned over the dev split of each dataset to make the ratio of positive to negative samples more balanced (cf. Ahmed et al. (2023) for optimal training.). Given a true intervention cluster after annotation and labeling, all pairs within it are considered positive pairs. Negatives comprise all other pairs under consideration (which may be limited by window W).

During training, the model is forced to learn discourse-relevant signals from the positive pairs drawn from true intervention clusters. Applying Longformer's *global attention* to *all* tokens in the pair (Fig. 4) allows us to encode relevant global features within W. Utterances in the preceding neighborhood W typically display lexical overlap for items with similar semantic roles, or task-specific phrases.[7] When such pairs are sourced from separate intervention clusters that occur within W, they naturally form difficult samples for encoder-only LLMs like Longformer due to misleading lexical overlap (Ravi et al., 2023; Ahmed et al., 2023).

[6] Our training method is generalizable to all utterances, since the ground truth label on any candidate can be causal, probing, or *neither* (a non-intervention dummy variable, ϵ). Generated pairs may have true labels that are any combination of probing and causal, since two causal interventions may be linked to the same probing intervention, or two probing interventions may share a cause, which results in these pairs themselves being linked under transitive closure. This follows standard practice in pairwise approaches to coreference across long documents.

[7] For instance, in the Weights Task, neighboring utterances contain overlapping arguments like "red block" when the group is solving a particular subtask relevant to that block.

Inference We evaluate two inference strategies. For our naive approach, we relax W and generate all candidate antecedent utterances within D, score them using using the intervention scores (mean of s_i and s_j), and only keep the remaining pairs based on a threshold τ (details in Appendix B). This reduces cross-comparisons in building the intervention clusters as we only use the pairwise scorer to score the remaining utterances. We also consider all scores generated without relaxing W. While the naive approach tests the system's recall under a long-tail of true-negatives, this method enforces a more balanced distribution, resulting in a "soft" upper-bound on model precision. Pairwise scoring generates an adjacency matrix of links between utterances. Inducing transitivity between links using a connected-components based clustering method with a threshold of 0.5 generates the final intervention clusters. Under temporal ordering, these expand to deliberation chains within a dialogue.

6 Experiments

We evaluated our joint modeling method against 3 similarity baselines and two cross-encoder methods adapted from coreference research.

6.1 Similarity Baselines

For simple similarity baselines, we assessed:

- Simple token overlap between utterances. This may indicate correspondence between a probing intervention and its cause(s), as the utterances may share terms. To assess lexical similarity between utterance pairs, we calculated the Levenshtein distance ratio (0–100) between the two strings.
- The overlap of salient *entities* within the utterances. We computed an Intersection over Union (IoU) of entity counts score based on categorical features derived from task-relevant categories referenced in each utterance (i.e., vowels, consonants, even and odd numbers in DeliData, and colors and weights in WTD).
- Cosine similarities between embeddings of the two utterances, extracted from BERT-`base-uncased`, following the intuition that probing utterances should share some *semantic*, not just token or entity similarity (Jawahar et al., 2019) with their causal counterparts.

For each, we set a threshold value for each dataset, equal to the average of the relevant metric over the dev set. If the relevant metric for a test pair exceeded this calculated threshold for the dataset, we linked that pair. More details on these baselines and the threshold values are given in Appendix C.

6.2 Cross-Encoder Baselines

For trainable baselines, we specifically chose recent coreference resolution frameworks that operate on an "utterance" level (instead of a span-level) for a valid comparison (see Sec. 5). For fairness, we used the base encoders from these frameworks as well as with their cross-encoding strategies, but not their fine-tuned weights, since fine-tuning on a separate task can likely tilt the model out of distribution (Kumar et al., 2022).

We used Caciularu et al. (2021)'s Cross-Document Language Model (CDLM). with a context length of 1,024 preceding tokens along with their cross-encoding setup.[8] We also employed Ahmed et al. (2023)'s "bidirectional" BCE loss-based learning method. This generates a mean of the BCE losses over the forward pass of utterances paired in both directions: (u_i, u_j and u_j, u_i). Like our joint modeling approach, the context window here is 512 tokens.

6.3 Joint Modeling Hyperparameters

For joint modeling, we use the Adam (Kingma and Ba, 2014) optimizer with batch size 24, with learning rates of $1e{-}6$ for the encoder fine-tuning, $1e{-}4$ for the pairwise scorers, and $1e-5$ for the intervention scorers. Each training epoch on an NVIDIA A100 took ∼20 and ∼40 minutes for DeliData and WTD, respectively. Each model was evaluated after a single training run for 16 epochs after robust hyperparameter tuning on the validation sets.

7 Results

We evaluate against coreference methodology using cluster metrics computed using the CoVal coreference scorer (Moosavi et al., 2019), specifically B^3 and CoNLL F_1 metrics, as presented for both datasets in Table 2.[9] We also present zero-shot results from LLaMA 2-7B-chat. The prompting framework for this is given in Appendix E. Appendix F presents results according to other

[8] CDLM (`https://huggingface.co/biu-nlp/cdlm`) is trained on documents with overlapping information and is suitable for handling long inputs, which are both traits of our dialogues (Sec. 5.1). For compute reasons, we truncate pairs at a maximum sequence length of 1,024 tokens after tokenization since the token-length of utterance pairs in training is ∼220 tokens for both datasets, on average.

[9] Since we are using the gold intervention labels for our experiments, using B^3 is more reliable compared to other metrics (Moosavi and Strube, 2016; Held et al., 2021).

	DeliData				WTD			
	B^3			CoNLL	B^3			CoNLL
	R	P	F_1	F_1	R	P	F_1	F_1
Lexical Overlap	26.6	81.3	40.0	28.6	41.6	50.0	45.4	36.6
Entity Overlap	34.9	71.7	46.9	40.6	27.2	70.0	39.2	26.7
BERT-Cosine	98.6	49.9	66.3	69.2	**100.0**	7.1	13.2	35.3
LongContext	84.7	60.7	70.7	68.2	72.1	23.8	35.8	45.5
Bidirectional	90.8	59.2	71.7	70.9	64.5	31.5	42.4	44.3
LLaMA 2-7B-chat	**99.9**	49.7	66.4	69.7	**100.0**	7.1	13.2	35.3
— Ours (Joint - W)	92.3	60.5	73.1	73.6	54.4	**75.0**	63.0	50.3
— Ours (Joint + W)	87.8	**72.6**	**79.5**	**76.4**	67.9	61.7	**64.7**	**58.1**

Table 2: B^3 and CoNLL F_1 metrics on DeliData and WTD test set results. "LongContext" denotes Caciularu et al. (2021)'s coreference methodology applied to deliberation chain clustering. "Bidirectional" denotes Ahmed et al. (2023)'s methodology.

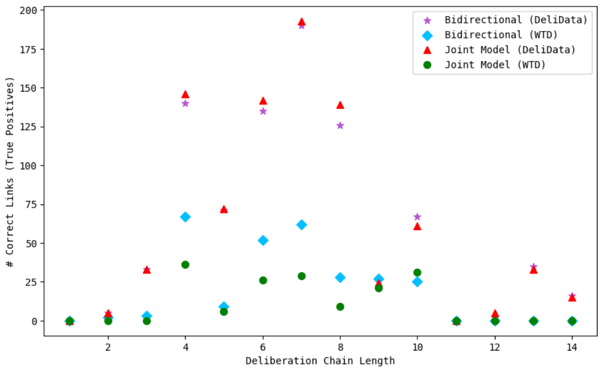


Figure 5: Cluster-level distribution of correctly assigned intervention links for the best-performing cross-encoder baseline compared to Joint - W on both datasets.

common coreference metrics. Results empirically demonstrate the strength of our theoretically-grounded method on this challenging task and data.

The multimodal nature of the WTD likely makes it more challenging than DeliData due to cues that may be missed in even the dense paraphrased language. The use of the windowed approach results in a small performance improvement due to the exclusion of false positive links outside W. The BERT-Cosine and LLaMA 2 zero-shot baselines perform extremely similarly (returning identical metric values on WTD) and achieve perfect or near-perfect recall. This is likely due to these methods returning a very high proportion of false positive links and transitive closure subsequently clustering (almost) all interventions in a dialogue.

8 Discussion

Quantitative Analysis Fig. 5 shows the count of correct links between interventions assigned by the bidirectional baseline and our (non-windowed) joint model for each cluster size.[10] DeliData has much longer chains on average with a wider distribution at every chain-size, and the joint model consistently links more pairs correctly in frequent medium cluster sizes, while at larger cluster sizes joint modeling and the bidirectional baseline are competitive. The joint model may be learning a more global representation of deliberation chains, as from a discourse-coherence perspective mid-sized chains better reflect the true distribution in collaborative dialogues.

In WTD, the distribution of cluster sizes is narrower, while dialogues are much longer. Our joint model links interventions more conservatively, but also more correctly, than the bidirectional model. This is most evident in the joint model's ~45-point increase in B^3 precision compared to the bidirectional baseline. The latter's aggressive linking, while boosting recall for mid-sized clusters, does so at a higher cost to precision. This suggests that when the cluster distribution is skewed (smaller chains, longer dialogues), the joint model is better at avoiding impure clusters.

Qualitative Analysis Table 3 presents two test pairs from each dataset that our non-windowed model classified successfully that all other methods did not. Utterances are numbered and labeled as "C" (causal) or "P" (probing). For space reasons, examples given at the link level, not the full cluster level, but we include some other utterances that are clustered into the same chain, as well as the FTR generated by GPT during the COT-guided intervention labeling process, for context and to illustrate the kind of information our method is able to leverage for its decisions that others cannot:

[10] Only the non-windowed model results in a full comparison to all other baselines because Joint + W does not consider all gold pairs.

	Dialogue	Free-Text Rationale(s)
(a)	**[C1] Emu: I picked the card with the vowel A on it, because the rule said all cards with vowels on one side will have an even number on the other** [C2] Koala: I think it is A and 2 [C3] Hamster: I agree ... **[P4] Bee: So are we ready to final submit**	[C1] *"Emu's statement directly relates to the reasoning behind choosing the card with the vowel A, which is crucial in the decision-making process."*
(b)	[C1] Narwhal: What card did you think needed to be turned? ... [C2] Guinea pig: I picked 6 and U ... **[C3] Kiwi: We need to pick one that wouldn't fit the rule to test it. Maybe?** ... **[P4] Kiwi: 7 and U?**	[C3] *"This statement hints at the strategy of testing a card that would break the rule to confirm its validity, indicating a shift in the participant's thought process during the discussion."*
(c)	[C1] Participant 2: Oh maybe I'll try holding it here ... **[C2] Participant 2: Mystery block, blue block, red block, green block, purple block, yellow block kinda feels the same** ... **[C3] Participant 1: So how about purple block, green block two, I had eh purple block, yellow block two** ... **[P4] Participant 2: Is there a better way to measure mystery block?**	[C2] *"This utterance indicates the participant's initial attempts to compare the weights of various blocks using their fingers, setting the groundwork for exploring different measurement techniques."* [C3] *"This utterance directly led to the probing question as it involved a new approach of grouping blocks on fingers to measure their weights."*

Table 3: Test samples from DeliData (a & b) and WTD (c). Bolded utterances indicate $(\mathcal{P}, \mathcal{C})$ pairs that our method (Joint - W) linked correctly and all other methods failed to. FTRs are given for the annotation of the indicated utterance as causal. These are not included in the input for inference, but are provided as indicators of the kinds of information our framework is likely to learn from the labels that were created using this COT-guided process.

(a) Our model links P4 to C1, which references the letter A, vowels, and even numbers, which are also referenced in C2, which states what the participants agree on. P4 elicits confirmation of all that aggregate information.

(b) We see in the FTR that C3 and P4 indicate a shift in Kiwi's thought process, and our system picks up the link between the causal and probing utterances made by the same participant, which others miss.

(c) In this example from the WTD, our model makes two links (between P4 and C2 and C3). Both C2 and C3 pertain to measurement techniques but this is not immediately apparent from the text. The FTR makes apparent that the GPT labels are based on the probing utterance's mention of measuring blocks.

We note that our model tends to successfully make links much further back in the dialogue history than even the longer-context CDLM model. In the examples presented, we show only the causal and probing interventions that form the response cluster, omitting utterances that are neither (indicated by ellipses). It is notable that these utterances alone still form complete exchanges.

9 Conclusion and Future Work

In this paper, we established a novel task of *deliberation chain construction* in collaborative dialogues. We developed a formal graphical model of deliberation chains grounded in discourse coherence theory, and applied coreference resolution techniques to two challenging datasets. Our joint modeling approach emerged as the best model on both datasets, setting a performance standard in this novel task.

Our joint model predicts the probability of an utterance being probing or causal and uses only prior context—a next logical step is to adapt our method to a live interaction, doing intervention *detection*, and predicting when a probing utterance will (or should) occur. This would represent a significant step forward for AI systems that can mediate real-time collaboration.

The WTD's multimodal aspect represents a rich opportunity to investigate multimodality's role in deliberation. For instance, a gesture or action might itself be a probing or causal intervention, and Asher et al. (2020) provide a compatible multimodal framework in which to pursue this.

Finally, deliberation chains construction is adaptable to interactions with particular characteristics, like argumentation (Afantenos and Asher, 2014), and computational understanding these distinctions will open new horizons in human-AI interaction.

Limitations

We used `GPT-3.5-turbo-0125` to annotate utterances as causal interventions given a corresponding probing intervention. While machine-assisted annotation is an accepted method in the field (Vossen et al., 2018; Ahmed et al., 2024) and we validated the annotations with human judgments (see Appendix G), there is always a risk that annotations provided by AI are noisy or unreliable. Therefore, one limitation that could be addressed in future work is the lack of true gold-standard human annotations for probing and causal interventions in the two datasets. Additionally, we believe future work should directly compare human vs. automated annotations of things like speech transcriptions in datasets like the WTD in order to further validate the use of automated annotations on these datasets. Finally, we noted that utterances which prompted actions—such as a participant instructing another to move a block—were marked as probing when we used GPT to annotate probing utterances in the WTD. This created a type of probing label that was not present in the DeliData, which entails no physical actions. This may be an indication of the type of noise introduced by automated labeling, but we also speculate that this type of probing is a potential avenue for investigating dialogue driven tasks that require action—future work will need to investigate the validity of such annotations.

Since we adapt coreference techniques to deliberation chain construction, we also use coreference metrics. Moosavi and Strube (2016) note the different pitfalls of all common coreference metrics—however, future work should examine the specific limitations of these metrics in the context of the task; some of these metrics may not be a good fit for this and new metrics may need to be developed. See Appendix F for a more detailed look into our current thoughts on this limitation.

Ethical Statement

Perhaps the largest risk inherent in systems that model deliberation and probing is how they are deployed. Consider, for example, a classroom context: modeling deliberation in a normative fashion may risk disadvantaging persons whose modes of collaboration are non-normative. In the worst case, a system could identify these persons as lacking engagement or having poor collaboration skills, resulting in undue punitive measures.

These models are designed to monitor and aid interaction; however, we do not believe such systems should exist in isolation—explicitly, in a classroom context, we believe such systems should augment teachers, not replace them.

Especially for multimodal use cases, like the Weights Task Dataset (WTD) (Khebour et al., 2024a), there is a risk of such technologies being used for tracking and surveillance, as modeling how individuals collaborate also involves modeling their linguistic and reasoning patterns, which may be sensitive. In this paper, we use publicly-available anonymized datasets that were collected under protocols reviewed by institutional review boards for ethical research, and were conducted with subjects who consented to the release of the data. However, collaboration modeling technology should be treated cautiously when it comes to ingesting multiple modal channels about specific people.

Finally, extending the model of deliberation to an agent that actually intervenes in dialogues could be exploited by bad actors who create bad agents, that bring dialogue to a halt through excessive introduction of “friction,” thus impeding the reasoning and productive benefits that collaboration brings.

These risks are inherent in the deployment of systems that perform the task we have developed herein as a precondition for other actions in the world, not the formal or computational model of deliberation itself.

Acknowledgements

This material is based in part upon work supported by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program, and by the National Science Foundation (NSF) under a subcontract to Colorado State University on award DRL 2019805 (Institute for Student-AI Teaming). Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of, the Department of Defense, the National Science Foundation, or the U.S. Government. All errors are the responsibility of the authors. Our thanks also go out to the anonymous reviewers whose feedback helped improve the final copy of this paper, and to Maniteja Vallala, Shamitha Gowra, Tarun Varma Buddaraju, and Sai Kiran Ganesh Kumar for extensive data annotation. We also thank Bruce Draper for the lively discussions that contributed to this work.

References

Stergos Afantenos and Nicholas Asher. 2014. Counter-argumentation and discourse: A case study. CEUR Workshop Proceedings.

Shafiuddin Rehan Ahmed, George Arthur Baker, Evi Judge, Michael Reagan, Kristin Wright-Bettner, Martha Palmer, and James H. Martin. 2024. Linear cross-document event coreference resolution with X-AMR. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 10517–10529, Torino, Italia. ELRA and ICCL.

Shafiuddin Rehan Ahmed, Abhijnan Nath, James H. Martin, and Nikhil Krishnaswamy. 2023. $2 * n$ is better than n^2: Decomposing event coreference resolution into two tractable problems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1569–1583, Toronto, Canada. Association for Computational Linguistics.

Katya Alahverdzhieva, Alex Lascarides, and Dan Flickinger. 2017. Aligning speech and co-speech gesture in a constraint-based grammar. *Journal of Language Modelling*, 5.

Jessica Andrews-Todd and Carol M. Forsyth. 2020. Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior*, 104:105759.

Nicholas Asher, Julie Hunter, and Kate Thompson. 2020. Modelling structures for situated discourse. *Dialogue & Discourse*, 11:89–121.

Nicholas Asher and Alex Lascarides. 2003. *Logics of conversation*. Cambridge University Press.

Katherine Atwell, Mert Inan, Anthony B Sicilia, and Malihe Alikhani. 2024. Combining discourse coherence with large language models for more inclusive, equitable, and robust task-oriented dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3538–3552.

Dorothée Behr, Lars Kaczmirek, Wolfgang Bandilla, and Michael Braun. 2012. Asking probing questions in web surveys: which factors have an impact on the quality of responses? *Social Science Computer Review*, 30(4):487–498.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv e-prints*, pages arXiv–2004.

Michael Bugert, Nils Reimers, Shany Barhom, Ido Dagan, and Iryna Gurevych. 2020. Breaking the subtopic barrier in cross-document event coreference resolution. In *Text2Story @ ECIR*, pages 23–29.

Michael Bugert, Nils Reimers, and Iryna Gurevych. 2021. Generalizing cross-document event coreference resolution across multiple corpora. *Computational Linguistics*, 47(3):575–614.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2648–2662.

Arie Cattan, Alon Eirew, Gabriel Stanovsky, Mandar Joshi, and Ido Dagan. 2021. Cross-document Coreference Resolution over Predicted Mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 5100–5107.

Riccardo Fusaroli, Kristian Tylén, Katrine Garly, Jakob Steensig, Morten H. Christiansen, and Mark Dingemanse. 2017. Measures and mechanisms of common ground: Backchannels, conversational repair, and interactive alignment in free and task-oriented social interactions. pages 2055–2060. Cognitive Science Society.

Arthur C. Graesser, Stephen M. Fiore, Samuel Greiff, Jessica Andrews-Todd, Peter W. Foltz, and Friedrich W. Hesse. 2018. Advancing the Science of Collaborative Problem Solving. *Psychological Science in the Public Interest*, 19(2):59–92. Publisher: SAGE Publications Inc.

Barbara J Grosz and Candace L Sidner. 1986. Attention, intentions, and the structure of discourse. *Computational linguistics*, 12(3):175–204.

Barbara Jean Grosz. 1977. *The representation and use of focus in dialogue understanding*. University of California, Berkeley.

William Held, Dan Iter, and Dan Jurafsky. 2021. Focus on what matters: Applying discourse coherence theory to cross document coreference. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1406–1417, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Friedrich Hesse, Esther Care, Juergen Buder, Kai Sassenberg, and Patrick Griffin. 2015. A Framework for Teachable Collaborative Problem Solving Skills. In Patrick Griffin and Esther Care, editors, *Assessment and Teaching of 21st Century Skills: Methods and Approach*, Educational Assessment in an Information Age, pages 37–56. Springer Netherlands, Dordrecht.

Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*.

Julie Hunter, Nicholas Asher, and Alexandra Lascarides. 2018. A formal semantics for situated conversation. *Semantics and Pragmatics*, 11:1–52.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2022. What makes you change your mind? an empirical investigation in online group decision-making conversations. In *Proceedings of the 23rd Annual*

Meeting of the Special Interest Group on Discourse and Dialogue, pages 552–563.

Georgi Karadzhov, Tom Stafford, and Andreas Vlachos. 2023. Delidata: A dataset for deliberation in multiparty problem solving. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–25.

Ibrahim Khebour, Richard Brutti, Indrani Dey, Rachel Dickler, Kelsey Sikes, Kenneth Lai, Mariah Bradford, Brittany Cates, Paige Hansen, Changsoo Jung, Brett Wisniewski, Corbyn Terpstra, Leanne M Hirshfield, Sadhana Puntambekar, Nathaniel Blanchard, James Pustejovsky, and Nikhil Krishnaswamy. 2024a. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of open humanities data*, 10(1).

Ibrahim Khebour, Kenneth Lai, Mariah Bradford, Yifan Zhu, Richard A Brutti, Christopher Tam, Jingxuan Tu, Benjamin A Ibarra, Nathaniel Blanchard, Nikhil Krishnaswamy, and James Pustejovsky. 2024b. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3587–3602.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. *Preprint*, arXiv:2202.10054.

Jenny Kunz, Martin Jirenius, Oskar Holmström, and Marco Kuhlmann. 2022. Human ratings do not reflect downstream utility: A study of free-text explanations for model predictions. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 164–177, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Heeyoung Lee, Marta Recasens, Angel Chang, Mihai Surdeanu, and Dan Jurafsky. 2012. Joint entity and event coreference resolution across documents. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 489–500.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.

Nafise Sadat Moosavi, Leo Born, Massimo Poesio, and Michael Strube. 2019. Using automatically extracted minimum spans to disentangle coreference evaluation from boundary detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642.

Lampros Mouselimis. 2021. *fuzzywuzzyR: Fuzzy String Matching*. R package version 1.0.5.

Abhijnan Nath, Shadi Manafi, Avyakta Chelle, and Nikhil Krishnaswamy. 2024. Okay, let's do this! modeling event coreference with generated rationales and knowledge distillation. In *Proceedings of the Meeting of the North American Chapter of the Association for Computational Linguistics*.

Vincent Ng and Claire Cardie. 2002. Improving machine learning approaches to coreference resolution. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 104–111.

OECD. 2017. *PISA 2015 Assessment and Analytical Framework*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, et al. 2023. Question decomposition improves the faithfulness of model-generated reasoning. *arXiv preprint arXiv:2307.11768*.

Sahithya Ravi, Chris Tanner, Raymond Ng, and Vered Shwartz. 2023. What happens before and after: Multi-event commonsense in event coreference resolution. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1700–1716.

Robert C Stalnaker. 1978. Assertion. In *Pragmatics*, pages 315–332. Brill.

Titin Suryati Sukmadewi. 2014. Improving students' mathematical thinking and disposition through probing and pushing questions. *Jurnal Matematika Integratif ISSN*, 1412:6184.

Chen Sun, Valerie J. Shute, Angela Stewart, Jade Yonehiro, Nicholas Duran, and Sidney D'Mello. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education*, 143:103672.

Raghuveer Thirukovalluru, Nicholas Monath, Kumar Shridhar, Manzil Zaheer, Mrinmaya Sachan, and Andrew McCallum. 2021. Scaling within document coreference to long texts. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3921–3931.

Jingxuan Tu, Kyeongmin Rim, Eben Holderness, Bingyang Ye, and James Pustejovsky. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 39–49.

Videep Venkatesha, Abhijnan Nath, Ibrahim Khebour, Avyakta Chelle, Mariah Bradford, Jingxuan Tu, James Pustejovsky, Nathaniel Blanchard, and Nikhil Krishnaswamy. 2024. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the International Conference on Educational Data Mining*.

Piek Vossen, Filip Ilievski, Marten Postma, and Roxane Segers. 2018. Don't annotate, but validate: A data-to-text method for capturing event data. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Xiaolong Wang, Yile Wang, Yuanchi Zhang, Fuwen Luo, Peng Li, Maosong Sun, and Yang Liu. 2024. Reasoning in conversation: Solving subjective tasks through dialogue simulation for large language models. *Preprint*, arXiv:2402.17226.

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2022. Symbolic knowledge distillation: from general language models to commonsense models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4602–4625, Seattle, United States. 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.

Xiaodong Yu, Wenpeng Yin, and Dan Roth. 2022. Pairwise representation learning for event coreference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 69–78.

Rui Zhang, Cicero dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107.

Wenting Zhao, Justin T Chiu, Claire Cardie, and Alexander M Rush. 2023. Abductive commonsense reasoning exploiting mutually exclusive explanations. *arXiv preprint arXiv:2305.14618*.

A Performance of Linking vs. Turns Between Interventions

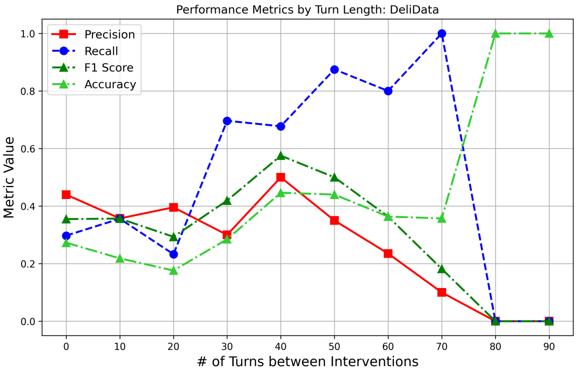


Figure 6: Ablations of our Joint + W model on DeliData by number of turns between interventions, when run on all possible intervention pairs.

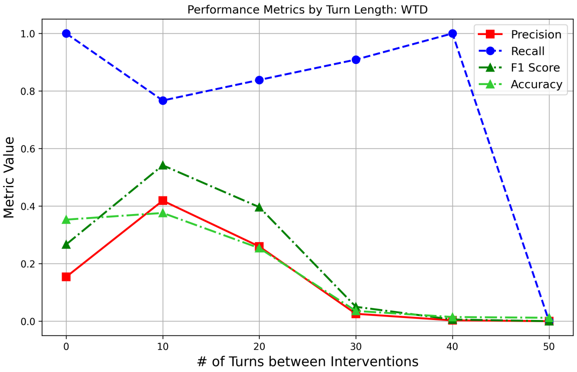


Figure 7: Ablations of our Joint + W model on the Weights Task Dataset by number of turns between interventions, when run on all possible intervention pairs.

For both training and inference for the Joint + W model, we fix the value of W, the window of previous utterances considered during pair generation, after tuning on the development set (18 for DeliData and 9 for WTD). This establishes a soft-upper limit on the model's performance since the model only learns from a balanced distribution of pairs likely representing an "attentional state" (Grosz, 1977; Held et al., 2021). To specifically evaluate this model on pairs that appear are widely separated in the discourse, we conduct ablations considering a range of turn-lengths. We bin all pairs based on their relative distance in the dialogue and report linking performance of the Joint + W model at a pair-level within the bins to provide further insights into its behavior.

As seen in Fig. 6 for DeliData, we find that overall precision dips slightly as turn-length increases and reaches a maximum within pairs binned between 30 and 40, after which it declines until all pairs are exhausted. Similar peaks for precision are also seen on WTD (Fig. 7), albeit at a shorter turn-length of 10 utterances. This suggests that our joint learning framework is likely helping the model pick up signals of the "validity" of interventions *per se*, without having to solely rely on previous context as in its encoding strategy (since we restrict k to be 10). On the other hand, the much smaller size of the WTD corpus with a narrower distribution of clusters and much longer dialogues makes it more challenging for our model to assign correct links to widely separated interventions in the discourse.

B Pruning Pairs for Naive Approach

Since our naive approach while relaxing W (Sec. 5.1) still operates within an utterance-level pairing of interventions, unlike "span"-level pruning strategies, our threshold τ directly prunes at the utterance-pair level instead of using a token-based filtering to improve recall (Cattan et al., 2021). Specifically, once all antecedents have been scored using the two intervention FFNNs (mean of s_i and s_j), we retain the top $\tau = G \times C$ highest-scoring pairs, where G represents the total number of interventions and C denotes the chain size, as detailed in Table 1. For DeliData, we use the mean chain size while for WTD, where dialogues are much longer ($\sim$200) utterances, we use the maximum chain size. Since we only prune pairs for the naive approach, this strategy lets us reduce cross-comparisons at inference while also reducing the impurity of chains by improving recall.

C Further Details on Similarity Baselines

C.1 Token Similarity baseline

Simple token overlap may indicate correspondence between a probing intervention and its cause(s), as the utterances may share terms. We used the FuzzyWuzzy library (Mouselimis, 2021), to assess lexical similarity between utterances, using the Levenshtein distance ratio (0–100) between two strings. We computed the token overlap percentage for each probing question and its preceding utterances in both the dev and test sets. Using an empirically-derived threshold from the dev set, based on the average token overlap percentage, if a test pair's token overlap exceeded this threshold, we linked that pair. The computed thresholds for DeliData and WTD were 0.247 and 0.263, respectively.

C.2 Entity Similarity baseline

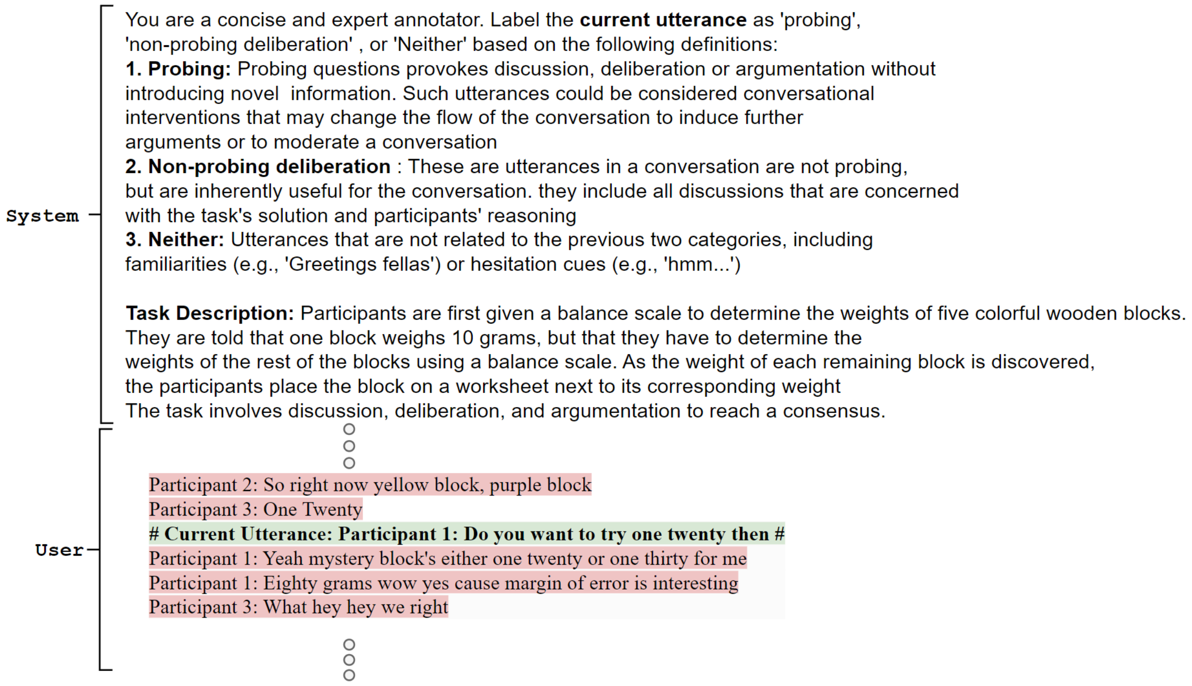
Rather than consider all tokens, which may include semantically irrelevant words, we considered an overlap of salient *entities* between utterances. We computed an Intersection over Union (IoU) of entity counts score based on categorical features derived from task-relevant categories referenced in each utterance. For the DeliData, these categories included vowels, consonants, even digits, and odd digits. For the Weights Task Data, these categories included the five block colors, and their weights, as in Venkatesha et al. (2024). Analogously to the token similarity baseline, we calculated the average IoU between probing interventions and their causal counterparts in the dev set. If a test pair's entity overlap exceeded this threshold, we linked that pair. The computed thresholds for DeliData and WTD were 0.287 and 0.173, respectively.

C.3 Cosine Similarity Baseline

Our final non-trained baseline leveraged BERT (`base-uncased`) to generate contextualized sentence embeddings for probing interventions and candidate causal counterparts. This followed the intuition that linked utterances should share some semantic similarity beyond the token or entity level. As with the previous two baselines, we calculated an empirical threshold (average cosine similarity between pairs) from the dev set, being the average of all cosine similarities. If a test pair's cosine similarity exceeded this threshold, we linked that pair. The computed thresholds for DeliData and WTD were 0.597 and 0.644, respectively.

D Further Details on Label Generation with GPT

Since the Weights Task Dataset does not contain annotations for probing utterances like DeliData does, we also used GPT to label these. Fig. 8 gives the prompting framework used for this. The probing annotations were also validated by humans



You are a concise and expert annotator. Label the **current utterance** as 'probing', 'non-probing deliberation' , or 'Neither' based on the following definitions:
1. Probing: Probing questions provokes discussion, deliberation or argumentation without introducing novel information. Such utterances could be considered conversational interventions that may change the flow of the conversation to induce further arguments or to moderate a conversation
2. Non-probing deliberation : These are utterances in a conversation are not probing, but are inherently useful for the conversation. they include all discussions that are concerned with the task's solution and participants' reasoning
3. Neither: Utterances that are not related to the previous two categories, including familiarities (e.g., 'Greetings fellas') or hesitation cues (e.g., 'hmm...')

Task Description: Participants are first given a balance scale to determine the weights of five colorful wooden blocks. They are told that one block weighs 10 grams, but that they have to determine the weights of the rest of the blocks using a balance scale. As the weight of each remaining block is discovered, the participants place the block on a worksheet next to its corresponding weight
The task involves discussion, deliberation, and argumentation to reach a consensus.

Participant 2: So right now yellow block, purple block
Participant 3: One Twenty
# Current Utterance: Participant 1: Do you want to try one twenty then #
Participant 1: Yeah mystery block's either one twenty or one thirty for me
Participant 1: Eighty grams wow yes cause margin of error is interesting
Participant 3: What hey hey we right

Figure 8: Prompting framework for labeling WTD utterances as probing or non-probing deliberation with GPT.

(see Appendix G).

Algorithm 1 provides the iterative labeling algorithm that is used during GPT label generation to assign newly-tagged interventions to the correct preceding cluster.

E Zero-Shot Prompt Design and Details

For zero-shot evaluation with LLaMA 2-7B-chat, we designed three prompts, depending on if the paired sentences were gold-labeled as causal and probing, causal and causal, or probing and probing. These prompts were designed to lead the model to a better chance at the correct conclusion, given that the relation between a causal and probing intervention is qualitatively different from that of two causals to a probing intervention that occurs elsewhere, or two probing interventions that share a cause. The prompts are given below.

LLAMA-2-7B-CHAT ZERO-SHOT PROMPT FORMAT: CAUSAL-PROBING

SYSTEM_PROMPT: Think step by step. You need to identify if one utterance in a dialogue is going to cause the other utterance to emerge later in the dialogue. Answer in one word: yes or no.
USER_PROMPT: sentence_1: {sentence_1} sentence_2: {sentence_2}

LLAMA-2-7B-CHAT ZERO-SHOT PROMPT FORMAT: CAUSAL-CAUSAL

SYSTEM_PROMPT: Think step by step. You need to identify if these two utterances in a dialogue are going to cause a probing question to emerge later in the dialogue. Answer in one word: yes or no.
USER_PROMPT: sentence_1: {sentence_1} sentence_2: {sentence_2}

LLAMA-2-7B-CHAT ZERO-SHOT PROMPT FORMAT: PROBING-PROBING

SYSTEM_PROMPT: Think step by step. You need to identify if these two utterances in a dialogue have been caused to emerge by the same preceding utterance in the dialogue. Answer in one word: yes or no.
USER_PROMPT: sentence_1: {sentence_1} sentence_2: {sentence_2}

For a small number of samples (DeliData: 21 out of 7,079, or ∼0.297%; WTD: 232 out of 10,761, or ∼2.156%), LLaMA 2 would not directly provide an answer to the question before reaching the maximum generation length. These were discarded from evaluation.

Due to profanity in a single utterance in the

WTD (“*So ten plus ten is twenty, twenty plus ten is thirty, thirty plus twenty is fifty, so mystery block’s eighty, so I was fucking right*”), LLaMA 2, which is known for its guardrails, would not process 7 pairs (out of 7,079 $\approx$ 0.099%) containing this utterance, citing offensive language or ethical or moral standards. These samples were discarded. The limitations inherent in evaluating LLMs on such a PG-13 dataset should be noted.

F Additional Results Tables

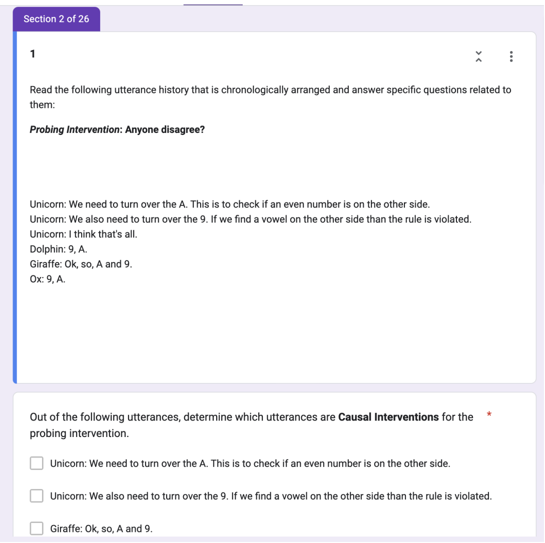
In coreference tasks, choice of metric bears heavily on the results. Tables 4 and 5 present results on our two test sets according to the MUC, B^3, $CEAF_e$, and CoNLL F_1 cluster metrics.

Our method performs well on all metrics, including restrictive ones like $CEAF_e$. We underperform some competing baselines on MUC, but this can largely be attributed to the permissiveness of the MUC metric. We observe that, given the threshold mechanism for the BERT-Cosine baseline, $\sim$40% of pairs in both test sets are labeled as positives by default. Given that the resulting false positives link to interventions that have true links to a larger chain, the transitive closure mechanism tends to link most or all utterances into a single intervention cluster. This is reflected in the 100% or near-100% recall achieved by BERT-Cosine and LLaMA 2 zero-shot in both MUC *and* B^3, and the extremely low $CEAF_e$ recall due to $CEAF_e$’s assumption that each key entity should only be mapped to a single reference entity. This indicates that while MUC especially is foundational in coreference, it may be a less useful metric in deliberation chain construction.

We currently exclude the LEA metric from our evaluation metrics for two reasons. First, we use gold intervention labels since the current work only considers link assignment to pairs in building deliberation chains and *not* intervention detection. Moreover, assigning an “importance” measure to various interventions at a linguistic level is beyond the scope of the paper. As such, for a fair evaluation between commonly used metrics, we focus on CoNLL F_1 as the average of the MUC, B^3 and $CEAF_e$ F_1 scores. By contrast, LEA specifically re-weighs evaluations to mitigate the “mention identification effect” and would apply a task-irrelevant importance measure to interventions (Moosavi and Strube, 2016). We leave determination of optimal metrics for this task to future work.

G Human Evaluation of GPT-Annotated Labels

Four evaluators (all adult English speakers) took a survey containing probing interventions and candidate causal interventions (25 sets each drawn from the WTD and DeliData corpora), the ground truth label (which was also given to GPT 3.5-turbo for generation), and the generated inner monologue FTR (see Fig. 9). They were asked to answer seven multiple choice questions for each sample, designed to explore various aspects of the dialogue explanation.



1

Read the following utterance history that is chronologically arranged and answer specific questions related to them:

***Probing Intervention*: Anyone disagree?**

Unicorn: We need to turn over the A. This is to check if an even number is on the other side.
Unicorn: We also need to turn over the 9. If we find a vowel on the other side than the rule is violated.
Unicorn: I think that's all.
Dolphin: 9, A.
Giraffe: Ok, so, A and 9.
Ox: 9, A.

Out of the following utterances, determine which utterances are **Causal Interventions** for the probing intervention. *

☐ Unicorn: We need to turn over the A. This is to check if an even number is on the other side.

☐ Unicorn: We also need to turn over the 9. If we find a vowel on the other side than the rule is violated.

☐ Giraffe: Ok, so, A and 9.

Figure 9: Causal intervention sample presented to evaluators.

The questions included:

- **Relevance to Context**: Are the Causal Intervention(s) relevant to the context? (*Yes/No*)
- **Presence in Sequences**: Are the Causal Interventions(s) present in the sequences of utterances? (*yes/no/not enough information*)
- **Information Sufficiency**: How much information do the Causal Interventions(s) have, to justify them being actual causal interventions? (*enough/not enough/more than enough/can’t say*)
- **Acceptability/Plausibility**: Are the Causal Interventions acceptable or plausible considering the context? (*yes/no/can’t say*)
- **Overlap with Own Explanations**: If you were to use your own explanations for selecting the causal interventions, how much of an

	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Lexical Overlap	18.2	56.4	27.5	26.6	81.3	40.0	43.9	11.6	18.3	28.6
Entity Overlap	38.6	64.0	48.2	34.9	71.7	46.9	48.8	18.5	26.8	40.6
BERT-Cosine	98.9	88.4	93.4	98.6	49.9	66.3	36.0	71.3	47.8	69.2
LongContext	85.9	87.5	86.7	84.7	60.7	70.7	48.9	45.6	47.2	68.2
Bidirectional	90.7	88.3	89.5	90.8	59.2	71.7	48.6	54.7	51.4	70.9
LLaMA 2-7B-chat	**99.9**	88.5	**93.8**	**99.9**	49.7	66.4	35.9	**77.1**	49.0	69.7
— Ours (Joint - W)	92.7	89.2	90.9	92.3	60.5	73.1	52.1	62.4	56.8	73.6
— Ours (Joint + W)	88.1	**91.5**	89.8	87.8	**72.6**	**79.5**	**64.4**	55.9	**59.9**	**76.4**

Table 4: DeliData test set results. "LongContext" denotes Caciularu et al. (2021)'s coreference methodology applied to deliberation chain clustering. "Bidirectional" denotes Ahmed et al. (2023)'s methodology.

	MUC			B^3			$CEAFe$			CoNLL
	R	P	F_1	R	P	F_1	R	P	F_1	F_1
Lexical Overlap	38.6	55.1	45.4	41.6	50.0	45.4	30.3	13.8	18.9	36.6
Entity Overlap	17.1	42.2	24.4	27.2	70.0	39.2	36.1	10.7	16.5	26.7
BERT-Cosine	**100.0**	80.9	**89.5**	**100.0**	7.1	13.2	1.7	**30.3**	3.3	35.3
LongContext	76.4	74.8	75.6	72.1	23.8	35.8	24.0	26.2	25.0	45.5
Bidirectional	65.7	73.0	69.2	64.5	31.5	42.4	25.4	18.2	21.2	44.3
LLaMA 2-7B-chat	**100.0**	80.9	**89.5**	**100.0**	7.1	13.2	1.7	**30.3**	3.3	35.3
— Ours (Joint - W)	50.0	**83.3**	62.5	54.4	**75.0**	63.0	45.6	17.5	25.3	50.3
— Ours (Joint + W)	67.9	81.9	74.2	67.9	61.7	**64.7**	**47.5**	28.2	**35.4**	**58.1**

Table 5: WTD test set results with all methods. "LongContext" denotes Caciularu et al. (2021)'s coreference methodology applied to deliberation chain clustering. "Bidirectional" denotes Ahmed et al. (2023)'s methodology..

overlap does your thought-pattern have with the given rationales? (*high overlap/some overlap/minimal overlap/no overlap*)

The statistics for the chain lengths of the drawn samples are as follows:

	Chain Length Statistics	
	DeliData	WTD
Min Chain Length	3	3
Max Chain Length	8	10
Mean Chain Length	5.65	5.3

Table 6: Chain length statistics of the human evaluation samples.

These chain lengths indicate that the sampled probing interventions are representative of the respective test sets, as their mean chain lengths align with the dataset averages, and the distributions are within the expected ranges.

Annotations were performed by members of the authors' research lab in the course of their normal duties. A different pair of annotators was used to assess samples from each corpus. Three annotators were male and one, female. Annotators had no prior experience in the task. The survey was determined to be Not Human Subjects Research by the institutional review board.

The survey response results shown in Fig. 3 show that the causal interventions, for both the WTD and DeliData, have positive valences when evaluated for relevance to the context, presence in sequence, and acceptability/plausibility. These positive scores show high association between causal interventions and their respective probing questions, given the context of the utterance history. DeliData information sufficiency is rated higher than WTD's, which shows that the DeliData contained more information to support the justification of classifying utterances as causal interventions. This could be a reflection of the use of different cards for the Wason Selection Task between groups in the DeliData experiment; where the WTD experiments utilized the same task items across all experiments, resulting in more repetitive phrases.

Algorithm 1 Gold Cluster Mapping via GPT

```
Require: $\mathcal{D}$: Sequence of dialogues, $P$: Probing interventions, $GPT(\cdot)$: LLM Prompting Operator
$G \leftarrow \{\}, R \leftarrow []$    ▷ Gold labels and all GPT responses
for $i = 1$ to $|P|$ do
    $ctx \leftarrow D[0 : P[i].index())$    ▷ Prior context until probing index
    $r \leftarrow GPT(P[i], ctx)$    ▷ Generate responses
    $R$.append($r$)
    if $i = 1$ then
        for $resp$ in $r$ do
            $G[resp] \leftarrow$ new unique label
        end for
        $G[P[i]] \leftarrow G[r[0]]$
    else
        $found \leftarrow$ False
        for $resp$ in $r$ do
            if $resp$ in $G$ then
                $G[P[i]] \leftarrow G[resp]$
                $found \leftarrow$ True
                break
            end if
        end for
        if not $found$ then
            if $r$ contains element from $P[1 : i - 1]$ then
                $idx \leftarrow$ index of match in $P$
                $G[P[i]] \leftarrow G[P[idx]]$
                for $resp$ in $r$ do
                    $G[resp] \leftarrow G[P[idx]]$
                end for
            else
                for $resp$ in $r$ do
                    $G[resp] \leftarrow$ new unique label
                end for
                $G[P[i]] \leftarrow G[r[0]]$
            end if
        end if
    end if
end for
return $G$
```

Our Gold Cluster Mapping Algorithm iteratively prompts an LLM (GPT) to extract causal interventions and rationales. Note that we do not show the rationales generated for each iteration of the loop for space reasons. These generated intervention clusters along with the rationales are then further validated with an exhaustive human evaluation component (see Appendix G).