Propositional Extraction from Collaborative Naturalistic Dialogues

Videep Venkatesha
Colorado State University
Fort Collins, CO, USA
videep.venkatesha@colostate.edu

Abhijnan Nath
Colorado State University
Fort Collins, CO, USA
abhijnan.nath@colostate.edu

Ibrahim Khebour
Colorado State University
Fort Collins, CO, USA
ibrahim.khebour@colostate.edu

Avyakta Chelle
Colorado State University
Fort Collins, CO, USA
avyakta.chelle@colostate.edu

Mariah Bradford
Colorado State University
Fort Collins, CO, USA
mbrad@rams.colostate.edu

Jingxuan Tu
Brandeis University
Waltham, MA, USA
jxtu@brandeis.edu

Hannah VanderHoeven
Colorado State University
Fort Collins, CO, USA
hannah.vanderhoeven@colostate.edu

Brady Bhalla
California Institute of Technology[*]
Pasdena, CA, USA
bbhalla@caltech.edu

Austin Youngren
Colorado State University
Fort Collins, CO, USA
austin.youngren@colostate.edu

Jack Fitzgerald
Colorado State University
Fort Collins, CO, USA
jack.fitzgerald@colostate.edu

James Pustejovsky
Brandeis University
Waltham, MA, USA
jamesp@brandeis.edu

Nathaniel Blanchard
Colorado State University
Fort Collins, CO, USA
nathaniel.blanchard@colostate.edu

Nikhil Krishnaswamy
Colorado State University
Fort Collins, CO, USA
nkrishna@colostate.edu

In the realm of collaborative learning, extracting the beliefs shared within a group is a critical capability to navigate complex tasks. Inherent in this problem is the fact that in naturalistic collaborative discourse, the same propositional content may be expressed in radically different ways. This difficulty

[*] This work performed under a CalTech Summer Undergraduate Research Fellowship (SURF) program at Colorado State University.

is exacerbated when speech overlaps and other communicative modalities are used, as would be the case in a co-situated collaborative task. In this paper, we conduct a comparative methodological analysis of extraction techniques for task-relevant propositions from natural speech dialogues in a challenging shared task setting where participants collaboratively determine the weights of five blocks using only a balance scale. We encode utterances and candidate propositions through language models and compare a cross-encoder method, adapted from coreference research, to a vector similarity baseline. Our cross-encoder approach outperforms both a cosine similarity baseline and zero-shot inference by both the GPT-4 and LLaMA 2 language models, and we establish a novel baseline on this challenging task on two collaborative task datasets—the Weights Task and DeliData—showing the generalizability of our approach. Furthermore, we explore the use of state of the art large language models for data augmentation to enhance performance, extend our examination to transcripts generated by Google's Automatic Speech Recognition system to assess the potential for automating the propositional extraction process in real-time, and introduce a framework for live propositional extraction from natural speech and multimodal signals. This study not only demonstrates the feasibility of detecting collaboration-relevant content in unstructured interactions but also lays the groundwork for employing AI to enhance collaborative problem-solving in classrooms, and other collaborative settings, such as the workforce. Our code may be found at: `https://github.com/csu-signal/PropositionExtraction`.

Keywords: collaborative problem solving, propositional extraction, natural speech, natural language processing, dialogue analysis

1. Introduction

For computer-assisted education, an important capability of automated systems is the ability to extract the meaning from student sentences or utterances to determine what they know, infer, or understand in the course of a task, activity, or assignment. In a naturalistic situated dialogue, like a small group in a classroom, information exchange is likely to consist of overlapping utterances with references grounded in the situational context, such as to objects in the scene or actions taken. Therefore, unlike in idealized scenarios such as strict turn-taking dialogues or written texts, it may be difficult to determine the exact semantic or propositional content that is being expressed by a single utterance.

An added challenge for educationally-grounded AI tasks such as knowledge tracing (Piech et al., 2015) is that the same semantics or proposition may be expressed in natural speech in radically different ways—there are likely to be incomplete sentences, repetition or restatement, filler words or disfluencies—and extracting relevant meaning despite such noise is crucial if an automated system is to make correct inferences about what students know or understand about their activity.

The propositional content that students assert is critical to tracking the collaborative process as students share their understanding and build consensus or common ground (Sun et al., 2020; Khebour et al., 2024). For example, an automated agent for collaborative problem solving support would need to track surfaced propositions as a measure of task progression. Additionally, students in collaborative settings achieve better learning outcomes when they engage in *leading* the discussion, which involves making new claims and not simply reiterating previously-stated information (Webb et al., 2021). The ability to extract propositional content from dialogue provides a way for an agent to determine whether a claim was already stated within the group. This would provide a necessary feature to determine whether a student is helping to lead the task

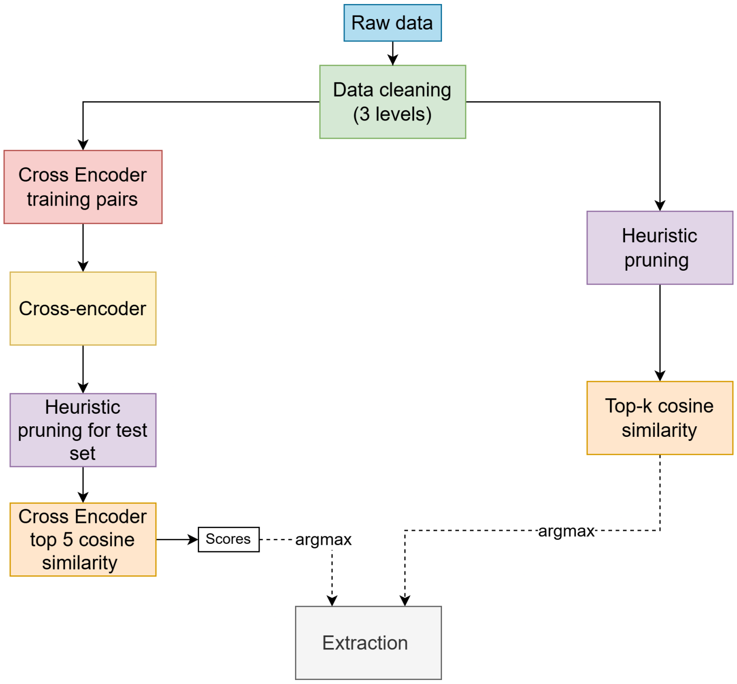


Figure 1: Schematic overview of the two methods used for propositional extraction. The process begins with Raw Data, which undergoes Data Cleaning across three levels to filter out irrelevant utterances. In the first method, a Cross-Encoder is trained on utterance-proposition pairs, followed by Heuristic Pruning for the test set, and outputs the top-5 candidate propositions using cosine similarity obtained from the trained Cross Encoder. In the second method, Top-k Cosine Similarity is directly applied to the heuristically pruned candidate propositions. The final Extraction step selects the best proposition using argmax over the similarity scores. Dashed lines indicate the selection process for the final proposition, while color coding differentiates key components of the pipeline.

forward, thereby enabling better prediction of learning outcomes from mined data.

This paper extends the contributions of Venkatesha et al. (2024), which appeared at EDM 2024. We take the transcribed utterances of a shared collaborative task, which are annotated with ground truth task-relevant propositions that are expressed therein, and use cosine similarity and cross-encoder methods to extract the propositions from the utterance text. Fig. 1 shows a schematic overview of our approach. We also extend our methods to utterances automatically segmented and transcribed by Google Cloud Platform's Automated Speech Recognition, showing how our propositional extraction methods may be incorporated into an automated system with a relatively low level of degradation due to automated transcription. Further, we explore the use of synthetic data augmentation using large language models (LLMs) to improve the robustness of our models and investigate the zero-shot inference capabilities of GPT-4 and LLaMA 2. Our results show the utility of methods adapted from coreference research in the field of natural language processing on this challenging task, and we introduce a framework for live proposition extraction, aiming to enable real-time analysis of collaborative interactions. To guide this work, we focus on several key research questions: **1)** How accurately can task-relevant propositions be extracted from naturalistic dialogues? **2)** How do different extraction methods compare in performance, and how does automated transcription affect this accuracy? **3)** Can synthetic data and zero-shot approaches enhance performance? **4)** Finally, is real-time extraction feasible in collaborative settings, and how might multimodal integration improve it?

Our novel contributions are grouped into two key areas:

- **Advancing Propositional Extraction Methods:** We establish a novel, challenging task of propositional extraction from natural speech during collaborative interactions. We compare cosine similarity and cross-encoder methods, using multiple language models and different levels of data cleaning, establishing new baselines and theoretical upper bounds for this task. Additionally, we explore synthetic data generation techniques to enhance model performance and assess zero-shot inference capabilities using large language models such as GPT-4 and LLaMA 2.
- **Practical Implementation and Real-Time Feasibility:** We assess the impact of automated speech transcription on extraction performance, demonstrating a relatively low level of performance degradation compared to manually transcribed utterances. We also introduce a framework for live propositional extraction, aiming to enable real-time analysis of collaborative interactions with a focus on practical deployment scenarios, and present a proposal for evaluation.

2. Background and Related Work

Collaborative tasks concern the construction and maintenance of a shared conception of the problem at hand (Roschelle and Teasley, 1995), involving mutual engagement and coordinated effort to solve the problem together. Within such a framework, especially one centered around shared synchronous tasks, quantity of specific propositions discussed has been shown to be a significant predictor of learning gains (Gijlers and de Jong, 2009). Therefore, propositional extraction serves an important role in automated analysis of shared task data in an educational context, or for an automated system to make inferences about construction of shared knowledge in real time.

Propositional Extraction Prior work on propositional extraction from natural language has primarily been conducted from written texts in domains such as question answering, where early methods relied on approaches such as semantic memory (Dennis, 2004). Classical machine learning approaches like support vector machines have been applied to opinion mining to find “propositional opinions,” or sentence fragments that contain the object of an assertion, incorporating word and feature-level knowledge from resources like WordNet, FrameNet, and PropBank (Bethard et al., 2004). Linguistic features have even been used to extract “ideas” from transcribed speech in the clinical domain, as a technique to predict Alzheimer’s disease and other types of cognitive decline (Chand et al., 2012). These early works not only show the utility of propositional extraction in various domains, but also demonstrate the relative sparsity of study on this topic. With the advent of neural network methods for text processing, these have been applied to NLP problems like propositional extraction from argumentation and rhetoric (Jo et al., 2019; Jo et al., 2020). These approaches include reported speech, as may appear in documents such as news articles. To the best of our knowledge, we are the first to attempt a similar task on transcribed naturalistic speech data from a collaborative task setting reminiscent of small group work in classrooms. Successfully extracting the propositional content expressed by an utterance is critical to modeling the epistemic positioning of the speaker toward the proposition. In a group context, these two components are required to track the shared and divergent beliefs of the group as they pertain to a task, as a method of modeling task progress. Khebour et al. (2024) treats this as a problem of "common ground tracking" and we demonstrate how our propositional extraction methods fit into a downstream task such as this in Sec. 8.

PAIRWISE REPRESENTATION LEARNING All of the aforementioned approaches frame the problem as one of establishing a mutual relationship between a piece of text from a dataset and another piece of text from a library of candidates, be they ideas, opinions, or propositional information more generally. Pairwise representational learning techniques have long been popular in the deep learning community for learning such relationships between two pieces of text. While some previous works modeled these relationships for text-generation tasks like abstractive document summarization (Nallapati et al., 2016), machine-comprehension (Hermann et al., 2015), or document-reconstruction (Li et al., 2015), others have also explored pairwise learning to compute similarity metrics between pairs of documents (Ahmed et al., 2023; Reimers and Gurevych, 2019; Zhang et al., 2020) as well as for masked language modeling (Devlin et al., 2019). More recently, for clustering-related tasks like coreference resolution, a "cross-encoding" framework has been used to learn pairwise features of possible coreferent mentions (Ahmed et al., 2023; Caciularu et al., 2021; Cattan et al., 2021; Held et al., 2021; Yu et al., 2022; Zeng et al., 2020). These works, originally inspired by Humeau et al. (2020), learn high-level semantic features of a mention (e.g., of an entity or event) within a sentence in the context of another mention-containing sentence and compute the coreference probabilities of such pairs before clustering mentions that refer to the same entity. We adopt this "cross-encoding" technique for both our candidate proposition generation procedure, as well as for calculating the probability of a given utterance referring to a candidate proposition.

CROSS-ENCODERS According to discourse coherence theory, in a dialogue between two or more participants, the content of the discussion is essentially a subset of the common knowledge, beliefs, and common intention (goal) that each participant has at any given point. As such, certain processing decisions like identifying referring expressions or detecting common propositional content between utterances can be made locally within the "attentional state" of the discourse. Following discourse coherence theory (Gentner, 1978; Grosz and Sidner, 1986), a human reader of a text or listener of a dialogue will focus attention on only a small subset of the total possible complement of events and entities. For instance, in a collaborative problem-solving setting, the words in an utterance that any participant uses to describe a specific sub-task within the overall task are constrained by "discourse segment purpose" or their common intention at that specific point in the dialogue. This constraint in the appearance of utterances to maintain coherence in the collaborative problem-solving dialogue allows us to map an utterance to a proposition by focusing only on the *local* elements in the utterance/proposition pairs.

However, since linguistic constraints or rule-based heuristics used to determine this attentional state can be narrow in their scope or domain-specific, most previous works have modeled the attentional state using neural networks (Chai and Strube, 2022; Held et al., 2021; Jeon and Strube, 2020). The neural model creates a latent representation or high-dimensional *embedding* of discourse-relevant entities *in context* and a variety of similarity methods (such as nearest neighbors or neural attention mechanisms (Vaswani et al., 2017)) may be used to determine which entities are subjects of the current attentional state given a context. These models are typically built on top of pre-trained transformer-based language models (LMs) (Vaswani et al., 2017) like RoBERTa or Longformer (Beltagy et al., 2020; Liu et al., 2020) that are known to capture rich semantic features through their contextualized representations of tokens and sequences. Apart from computationally modeling the innate structural coherence in a discourse, these architectures can also generate potential referents by demarcating the attentional state within a dialogue, through context.

These works have focused on various natural language understanding tasks, including coreference resolution. Our task is adjacent to coreference resolution since we have to map a set of utterances to their corresponding propositions in a collaborative dialogue. As such, we take inspiration from the pairwise scorer/cross-encoder architecture commonly used as a pairwise representation learning framework in cross-document coreference resolution (Ahmed et al., 2023; Caciularu et al., 2021; Cattan et al., 2021; Nath et al., 2024; Nath et al., 2023; Yu et al., 2022; Zeng et al., 2020). This method forces a classifier to learn a combined representation of one mention (represented by a trigger word) in the context of the other, both of which are encoded within their respective sentences. This learning strategy is an effective way to generate similarity scores between pairs of event or entity mentions due to the contextualized learning framework.

TOWARD REAL-WORLD USE OF AI As AI performance has increased, more works have begun to investigate how various automated preprocessing methods impact downstream task performance, since imperfect data is inevitable in a real-world application (Castillon et al., 2022). Some of these efforts intentionally examine performance given imperfectly preprocessed data (Blanchard et al., 2018) while others have explicitly explored how various preprocessing techniques degrade performance (Terpstra et al., 2023). In particular, various recent works have explored how imperfect data corresponds with performance in small-group contexts (Donnelly et al., 2016; Donnelly et al., 2017; Blanchard et al., 2016; Bixler et al., 2015; Bradford et al., 2022). These works have shown that the influence of automated but imperfect tools, e.g., automatic speech recognition (ASR) for transcription, do degrade performance but not catastrophically so. In this work, we also examine how data imperfection degrades performance on a novel task: propositional extraction from natural dialogue during a collaborative group task.

While prior work has successfully applied propositional extraction in domains such as argumentation, and rhetoric, to the best of our knowledge, our study is the first to tackle this task in the context of naturalistic, collaborative dialogue involving multimodal signals in real-time. Our approach uniquely focuses on extracting propositions from overlapping, co-situated speech, demonstrating its applicability to collaborative educational tasks, where tracking shared knowledge is critical. We also extend cross-encoder methodology, commonly applied in coreference resolution, to propositional extraction in natural speech dialogues.

3. DATASETS

3.1. WEIGHTS TASK DATASET

The Weights Task (Khebour et al., 2024) is a situated collaborative problem-solving (CPS) task wherein groups of three work together to deduce the weights of differently colored blocks using a balance scale. There are a total of 10 groups, resulting in approximately three hours of

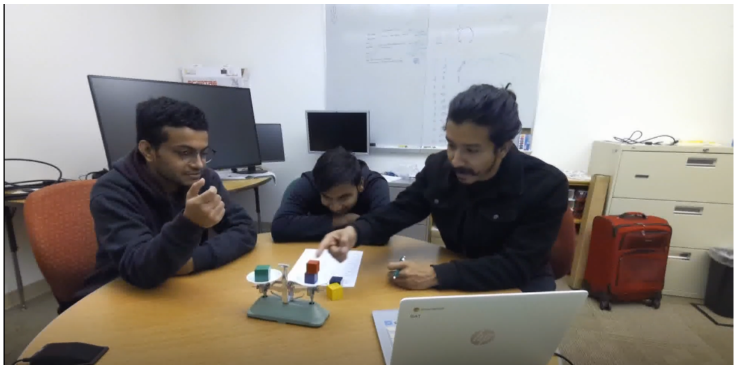


Figure 2: Example still from the Weights Task being performed. The utterance associated with this frame is “i guess green block is like twenty and red block, blue block is like ten and ten”. This utterance expresses the proposition $green = 20 \wedge red = 10 \wedge blue = 10$.

audiovisual data. Participants consented to the release of their likenesses for research purposes. The study protocol and release of A/V data were approved by the Colorado State University institutional review board.[1] In this work we focus on Phase 1 of the task, where the group has five blocks of different colors ($C = \{\text{red, yellow, green, blue, purple}\}$) whose weights follow an instance of the Fibonacci sequence ($W_n = \{10g, 10g, 20g, 30g, 50g\}$). At the start of the task, the group is told that the red block weighs 10 grams.[2]

The Weights Task Dataset (WTD) contains speech transcribed manually by humans (hereafter referred to as “Oracle” transcriptions) as well as speech transcribed automatically by Google Cloud Platform’s Automatic Speech Recognizer (Google ASR). The Oracle and Google transcription processes also *segmented* the speech into utterances—a single person’s continuous speech, delimited by silence. In the dataset, there are a total of 2,140 utterances that contain transcribed speech according to Oracle segmentation, and 1,500 utterances containing transcribed speech according to Google segmentation. Fig. 2 shows still frame of a group performing the task. Due to the overlapping nature of speech in this setting, utterance segmentation leads to many sentence fragments and overlaps, as well as mistranscription by the automated system, which leads to challenges in extracting the intended meaning behind any given utterance. An additional challenge to meaningful information extraction from the linguistic channel is that due to the multimodal nature of the task, a complete interpretation of an utterance may require recourse to another modality. For example, someone may say “this one” while pointing to a specific block. The pointing makes it clear which block is being referred to but without access to the video showing where the person is pointing, the language alone is ambiguous. The above factors enumerate just some of the challenges to extracting propositions expressed through dialogue in this setting.

The propositions themselves are annotated in the context of the *common ground* that evolves between group members as the task proceeds, that is, the set of propositions Φ each individual comes to believe as factual and that the group must agree upon, implicitly or explicitly, to arrive at the goal (Pacuit, 2017). In the case of the Weights Task, the participants must all arrive at the

[1] The dataset and consent documents associated with the original study protocol are publicly available at https://zenodo.org/records/10252341.

[2] Although a gram is a unit of mass, the colloquial dialogue in the dataset uses “mass” and “weight” interchangeably.

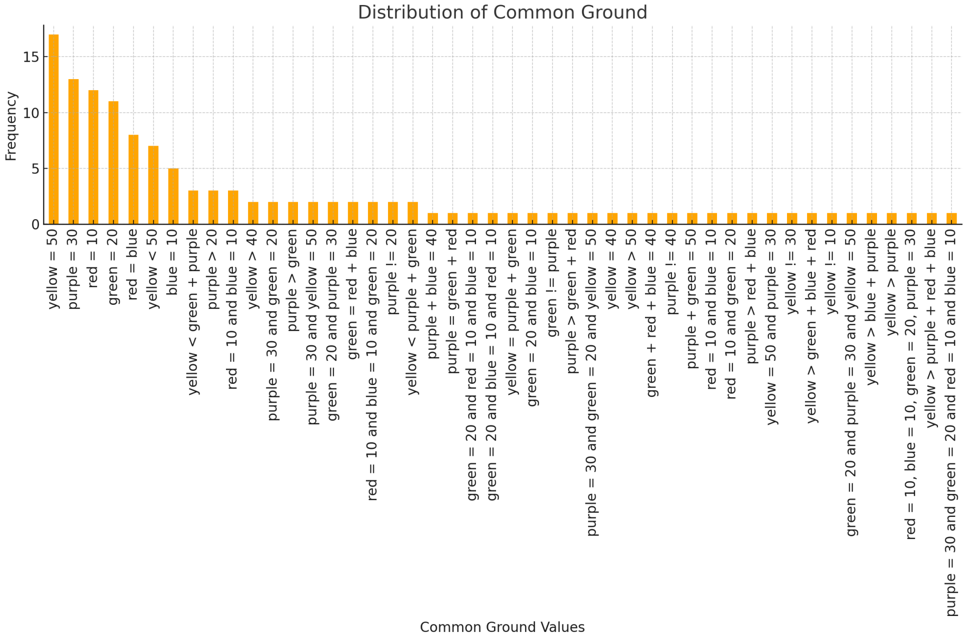


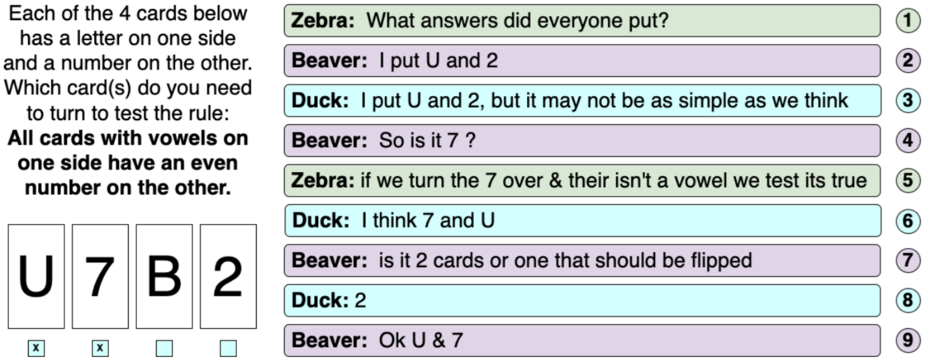
Figure 3: This figure illustrates the frequency of all 47 unique propositions expressed in the Weights Task Dataset. The horizontal axis lists the common ground propositions, such as weight assignments (e.g., *yellow = 50*), while the vertical axis represents their frequency across the dataset. The proposition *yellow = 50* is the most frequently expressed, appearing 17 times, followed by other key propositions such as *purple = 30* and *green = 20*. Less frequent propositions include combinations of weights and logical relations. This distribution highlights the diversity and repetition of propositions as participants collaboratively deduce the weights of colored blocks.

correct assignments of weight $w \in W$ to color $c \in C$ to solve the task. The WTD is annotated with the propositions that are asserted, evidenced, or agreed upon as the task unfolds, based upon the multiple modal channels and prior context. Our goal is to recover those propositions from the transcribed speech.

Within the dataset, there are 127 utterances which describe 46 unique propositions that were expressed during the task, with *yellow = 50* as the most frequent, appearing 17 times. The average frequency of each unique value is approximately 2.76. Proposition breakdowns show 107 instances containing a single proposition, 13 instances with two propositions, and 7 instances with more than two propositions. Among the operators, "$=$" is used exclusively in 95 instances, while "$\neq$" does not appear exclusively. Additional comparisons include "$>$" (14 instances) and "$<$" (12 instances), with 6 cases involving multiple operators. Notably, the top propositions occurring in the task are *yellow = 50*, *purple = 30*, *red = 10*, *green = 20*, *red = blue*, and *blue = 10*. These represent the correct weight assignments for each block, with other propositions involving various comparisons to derive the correct solution. A visualization of the distribution is provided in Fig. 3.

3.2. DeliData

DeliData (Karadzhov et al., 2023) is a dataset designed for examining group deliberation in multi-party problem-solving tasks, using the Wason card selection task as the focal activity.


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.**

Zebra: What answers did everyone put? 1

Beaver: I put U and 2 2

Duck: I put U and 2, but it may not be as simple as we think 3

Beaver: So is it 7 ? 4

Zebra: if we turn the 7 over & their isn't a vowel we test its true 5

Duck: I think 7 and U 6

Beaver: is it 2 cards or one that should be flipped 7

Duck: 2 8

Beaver: Ok U & 7 9

Figure 4: Abridged conversation from DeliData, illustrating the collaborative reasoning process of three participants solving the Wason card selection task. The task involves determining which cards to flip to test the rule that "All cards with vowels on one side have an even number on the other." Dialogue excerpts showcase how participants propose, evaluate, and revise their answers during the task. Reproduced from Karadzhov et al. (2023).

The Wason card selection task is a well-established task used in psychology research to explore reasoning processes (Evans, 2016). In this task, participants are presented with four cards, each with a number on one side and a letter on the other. The task is to determine which cards need to be turned over to test a rule, such as "All cards with vowels on one side have an even number on the other." This task is designed to reveal common reasoning errors, such as confirmation bias, where individuals might select cards that could confirm the rule rather than those that could potentially disprove it. Fig. 4 shows an example conversation.

This dataset comprises 500 group dialogues, totaling approximately 14,000 utterances. The corpus is annotated with deliberation cues, focusing on how participants propose, evaluate, and revise solutions in a collaborative setting. The groups consist of up to five participants, and the dialogues occur in an online chat format.

4. Methods

In this section, we outline the data preprocessing steps, followed by a detailed description of the methodologies employed for extracting propositional content from the datasets.

4.1. Proposition Enumeration

Weights Task Propositional content in the Weights Task takes the form of a relation between a block and a weight value (e.g., $red = 10$), between two blocks (e.g., $red = blue$), or between one block and a combination of other blocks (e.g., $red < blue + green$). To generate all possible candidate propositions in the domain, we employed a systematic process that combined the five block colors $(red, blue, green, purple, yellow)$, five potential weights $(10, 20, 30, 40, and 50)$, and four relations $(=, \neq, <, >)$ into all possible combinations that fit the aforementioned formats. "Conjunctive" propositions (e.g., *green* $>$ *20 and yellow* $<$ *50*) were also allowed, up to a length of three conjuncts (the maximum that ever appeared in the actual dataset). We normalized all candidate propositions for the symmetric property of equality (e.g., so that $red = blue$ is the same as $blue = red$), and dropped the resulting duplicates. The result was 5,005 total candidate propositions that *could* be expressed in the Weights Task domain.

Any given proposition might be expressed in multiple ways. For instance, in the data "purple block's thirty," "purple one thirty," "let's go thirty purple block's thirty," and "teeter teeter purple block's less forty greater twenty purple block's likely thirty" all appear as ways of expressing the proposition $purple = 30$, despite the fact that they may contain extra words or even mentions of additional blocks or weights not contained within the proposition actually expressed. We therefore modeled propositional extraction as a type of *coreference* problem, where the goal is not to determine whether two entity mentions refer to the same thing (Lappin and Leass, 1994), but rather to determine if two utterances mention both the same entity (block) and the same property (weight or relation).

DELIDATA Propositional content in DeliData takes the form of a structured set-member or attributive relation between a card and a property (e.g., $is(E, Vowel)$ for "[the] E [card] is a [member of set] Vowel"), or between a card and a hypothetical property on the hidden side (e.g., $has(E, Even)$ for "E has an Even number [on the other side]"). To generate all possible candidate propositions in the domain, we systematically combined the possible cards (A–Z and 1–9), the relevant properties ($Even$, Odd, $Vowel$, $Consonant$), and the defined relations ("is", "is not", "has", "does not have") into all possible combinations that align with the Wason selection task's structure. These propositions were normalized to account for symmetric properties and redundancies, resulting in a comprehensive set of propositions that could potentially be expressed in the dataset.

Given that any single proposition might be expressed in multiple ways within the dialogue, as in the WTD, we formalized the proposition in a similar manner. For example, different participants might express the same idea using varying phrases such as "E could have Even," "E would show Even," or "E must have an Even." Despite differences in phrasing or additional words, these expressions all map to the same underlying proposition $has(E, Even)$.

Unlike the WTD, which has the propositions incorporated into its common ground annotations, DeliData does not contain explicit annotations of propositions expressed. Therefore, we had 2 annotators perform this task, following a strict guideline to ensure consistency across the dataset. Utterances were annotated with the proposition expressed (if any) following the form $< card >< relation >< property >$, and the annotations were made with a focus on distinguishing between visible aspects of the card (e.g., "is Vowel") and speculative or hypothesized aspects (e.g., "has Even"). This structured annotation allowed us to capture the reasoning process of the participants. Cohen's κ for this task was 0.955, indicating high annotator agreement (Cohen, 1960), for a total of 255 utterances annotated over 100 groups. Utterances that described the multiple properties of the same card were removed since it expressed ambiguity and did not have an assertion. Examples of these include the utterance "I see, yeah, it doesn't matter if 4 is a vowel or not" expresses both $has(4, Vowel)$ and $\neg has(4, Vowel)$. Within those 100 groups, the occurrences of each operator are as follows: *is* appears 85 times, *is not* occurs 23 times, *has* appears 158 times, and *does not have* occurs 21 times. Proposition breakdowns show 197 instances with singular propositions and 57 instances with double propositions. These statistics cover 100 groups within the dataset that is being used for this task.

To generate the set of candidate propositions, first, we defined the allowable properties for each card type (letters and digits) based on the relation being asserted and according to the rules governing what can be on opposite sides of a card. For example, when the relation is "is" or "is not," a card showing a letter can only have the properties $Vowel$ or $Consonant$, while a card showing a digit can only have the properties $Even$ or Odd. Conversely, when the relation is "has" or "does not have," the properties apply to the hidden side of the card, with digits having $Vowel$ or $Consonant$ as properties and letters having $Even$ or Odd.

To avoid generating contradictory or redundant propositions, we ensured that each card was mentioned only once in any given set of propositions, which prevented conflicting assertions about the same card.

The final step involved generating all valid combinations of one or two propositions that adhered to the defined rules and consistency checks. This process produced a comprehensive set of candidate propositions that could theoretically be expressed during the Wason task. The result was a total of 38,362 propositions that could be used in the DeliData domain.

4.2. Annotation and Preprocessing of the Weights Task

Because of the multimodal nature of the Weights Task and the prevalent use of demonstratives, we enriched the transcribed utterances using a "dense paraphrasing" method inspired by Tu et al. (2022; 2023), that rewrites a textual expression to reduce ambiguity and make explicit the underlying semantics. We isolated the utterances containing at least one pronoun from a predefined set of {"it", "they", "them", "this", "that", "these", "those"}, performed a partial assignment of blocks referenced by those pronouns based on actions that overlapped the utterances, and had annotators identify the blocks denoted by the remaining pronouns, if any, while referring to the video (see Fig. 5). This annotation was performed separately for the Oracle and Google transcriptions. Utterances were dually annotated, resulting in an average Cohen's $\kappa = 0.89$ over the Oracle transcriptions and $\kappa = 0.87$ over the Google transcriptions. A gold standard was then generated through adjudication by an expert. The original utterances were then replaced with the dense paraphrased versions. High agreement scores and accuracy metrics demonstrate the reliability and effectiveness of the annotation process. This procedure *decontextualizes* the utterances from their multimodal dependencies, allowing us to evaluate the utterance as though it were text only.

4.2.1. Data Cleaning of the Weights Task

Filtration of the WTD is motivated by the fact that many utterances, even after dense paraphrasing, still do not mention a specific object or weight, meaning that extracting an object-weight or object-object relation from the utterance alone is infeasible. Our filtration steps follow steps used in existing coreference research (Ahmed et al., 2023). The decision to follow this methodology was made at the outset before any experimental results were available. We adopted three levels of data cleaning for WTD.

Level 1: The first level of cleaning consisted of removing all instances where neither color nor weight was mentioned in the transcript. An example of an utterance removed at this step would be "i mean it's not gonna go anywhere i guess it's just oh."

Level 2: The second level of cleaning involved removing all utterances where the mentioned colors and weights did not match the annotated proposition. For example, in an utterance "yeah red block, blue block should be twenty as well", "yeah" is actually an acceptance of a previously asserted proposition (in this case $green = 20$), and $red + blue = 20$, the mention of which is in the utterance, is not a valid propositional form in the task domain as the left hand side must

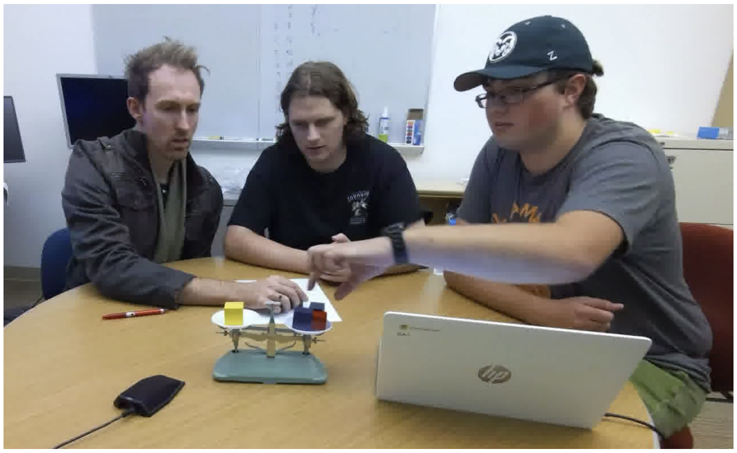


Figure 5: The image depicts participants in the Weights Task discussing potential solutions while interacting with the blocks and the balance scale. This setup emphasizes the importance of multimodal context (e.g., gestures and object interactions) in interpreting verbal utterances. For example, the original utterance is "we can replace one of [these] with the twenty." With reference to the video, an annotator can see the rightmost participant reaching for the red and blue blocks, so the dense paraphrased utterance is "we can replace one of *red block, blue block* with the twenty."

be a single block (in this case, the truth of $red + blue = 20$ is implicit in two other (valid) propositions $red = 10$ and $blue = 10$).

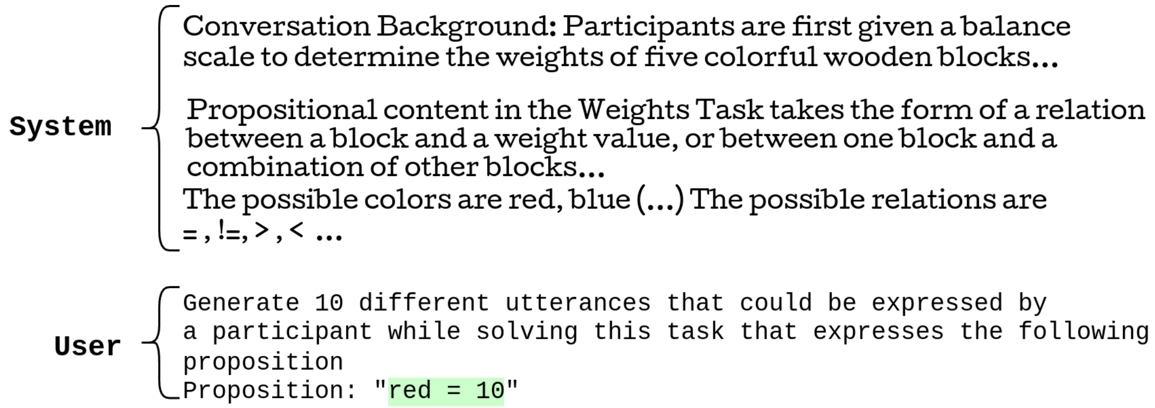
Level 3: The final level of cleaning removed all instances that do not mention a color, but only a weight. For instance, the utterance "well the top is a ten" is annotated as $blue = 10$, but with only the text, even a human would struggle to identify the correct proposition. The dataset annotators, meanwhile, had access to the video and could see that the top block referred to is blue, but as we focus only on transcriptions of natural speech, this information is not available to our method. By removing such ambiguous utterances, Level 3 cleaning results in a cleaner dataset where all remaining utterances explicitly mention both a color and a weight, making it easier for an automated system to extract propositions accurately.

Since the DeliData task does not include a multimodal component with the task, and has been already pre-processed, references to cards are already unambiguous, and so no additional cleaning was done on this dataset.

4.2.2. Data Augmentation of the Weights Task

The propositional extractor from Venkatesha et al. (2024) was limited by the sparsity of the utterances that actually expressed a proposition, totaling 47 unique propositions compared to the 5,005 possible propositions in the domain. For example, while $yellow + purple + green > red$ is a possible proposition according to the combinatorial process described in Sec. 4.1, it is unlikely to actually be expressed during task performance, as the combination of yellow, purple, and green blocks so clearly outweighs the red block that groups typically do not attempt it. In contrast, $green + purple = yellow$ is much more likely but may appear only once within a group, if at all.

To address this sparsity and improve generalization of the cross-encoder method, we explored a data-augmentation procedure inspired by prior work (Kulhánek et al., 2021; Pellicer et al., 2023; Ravi et al., 2023; Nath et al., 2024). Specifically, we prompted GPT-4 to generate 10 additional utterances for each of the 47 unique propositions present in the actual data, resulting in 470 new instances. These were combined with the 127 existing proposition instances, bringing the total coverage to 597 instances. Each generated utterance was subsequently human-validated for correctness before model training. The augmented data was used for training purposes only. The prompt used for GPT-4 is given in Fig. 6, followed by the specific proposition for which supplementary utterances were generated in the example case.



System

Conversation Background: Participants are first given a balance scale to determine the weights of five colorful wooden blocks...

Propositional content in the Weights Task takes the form of a relation between a block and a weight value, or between one block and a combination of other blocks...
The possible colors are red, blue (...) The possible relations are =, !=, >, < ...

User

```
Generate 10 different utterances that could be expressed by
a participant while solving this task that expresses the following
proposition
Proposition: "red = 10"
```

Figure 6: Synthetic data generation prompt used to augment the dataset for the Weights Task. The system defines the task context and possible relations, while the user prompt specifies generating 10 unique utterances expressing a target proposition (e.g., *red* = *10*). This approach expands the dataset while maintaining linguistic diversity and task relevance.

While we performed data augmentation on the WTD for reasons of sparsity, with 100 of the 500 DeliData groups, we accumulated sufficient training data for the cross-encoder and no additional data was required for training purposes.

4.3. Cross-Encoder

Above, and in Sec. 2, we motivated propositional extraction as a type of coreference problem. Therefore, we use a cross-encoder neural network that is common in natural lanugage processing (NLP) approaches to coreference. The cross-encoder learns a paired "contextualized" representation for an utterance proposition pair. Unlike previous coreference approaches mentioned in Sec. 2, which focus on the specific trigger word within a sentence, we encode the *entire* utterance in the context of the proposition to generate a combined representation for an utterance/proposition pair. This is for two reasons. Firstly, in our framework both the transcript and the candidate proposition can contain more than one color mention, which serves as a trigger indicating a block. For instance, consider "so *purple* block, *blue* block should be forty right there" (utterance) and $purple + blue = 40$ (candidate proposition). Encoding the utterance once for each specific color-trigger using a language model could drastically increase computational cost without any additional benefits of contextualization. This could also likely break down higher-level semantic signals that can otherwise be encoded with a wider context-window or the entire sentence. Secondly, under certain lenient pruning strategies, some transcripts may not contain any color at all. E.g., "... so you know twenty plus ten thirty probably ..." with a candidate proposition $red = 10 \wedge green = 20 \wedge purple = 30$. In such cases, full sentential context may capture more subtle semantic signals that are crucial for this task.

We encoded processed utterances as vector representations in two language models: **BERT-**`base-uncased` (Devlin et al., 2019), and **RoBERTa-**`base` (Liu et al., 2020). Before encoding, each dataset entailed slightly different preprocessing of the text. For WTD, stop words were filtered out according to a standard list augmented with words that occurred in five or fewer bigrams over all the transcriptions, and are not number words, color words, or (in)equality relation words. For DeliData, stop words were filtered out according to a standard list, with the exception of the set {'a', 'd', 'i', 'm', 'o', 's', 't', 'y', 'not', 'is', 'has', 'on'}. These characters or words were crucial to the context of the DeliData task, because they could refer to individual cards or relations between elements of a card. To retrieve the encoded vectors, we summed over the last four encoder layers of each model and took the average of the `[CLS]` (classification token, used to aggregate information from the entire sequence) or `<bos>` (beginning-of-sequence token, used for sequence initialization) token vector and all individual token vectors in the utterance. These vectors were used for propositional extraction by comparison using cosine similarity, and for training the cross-encoder architecture.

For an utterance/proposition pair (u_i, p_j), we construct an overall representation of the pair using the language model encoder. This representation consists of four individual parts, following modern standard practice in coreference established by Caciularu et al. (2021). We first surround u_i and p_j individually with special tokens `<m>` and `</m>` that are added to the language model tokenizer vocabulary and acquire learned representations during the training process. The first part of this overall representation is V_{CLS}, the pooled representation (`[CLS]`/`<bos>`) token of the last encoder hidden state. This representation is often used as a classification token in NLP tasks. Then, we encode u_i and p_j individually in the *context* of each other (that is, u_i when preceding p_j and p_j when following u_i)[3]. These comprise the second and the third components of the overall representation: V_{u_i} and V_{p_j}. We then encode the element-wise, or Hadamard, product of these two representations ($V_{u_i} \odot V_{p_j}$) to provide further cross-attention based signals. These four individual representations are then concatenated into a unified representation ($[V_{CLS}, V_{u_i}, V_{p_j}, V_{u_i} \odot V_{p_j}]$), which is fed into a multi-layer perceptron (MLP) to get similarity scores between the utterance and proposition (Eq. 1). The MLP is a two-layer neural network (768 and 128 neurons) that takes in the concatenated representation (768×4 = 3072 dimensions) and outputs a scalar, or after a sigmoid operation, the probability of an utterance referring to a proposition.

$$Score(u_i, p_j) = MLP([V_{CLS}, V_{u_i}, V_{p_j}, V_{u_i} \odot V_{p_j}]) \tag{1}$$

The candidate proposition with the highest score is retrieved, or the scores can be used to compute a *ranking* of candidate propositions, for metrics like top-k accuracy. Fig. 7 shows a schematic overview of the cross-encoder architecture.

4.3.1. Cross-Encoder Training

The parameters of the MLP are learned along with the parameters of the pretrained language model. Motivated by Ahmed et al. (2023), we use a symmetric cross-encoding framework that

[3] The positional encoder of transformer models cause the resulting representations to be different despite the input order being the same. This avoids positional bias observed in transformer models and allows for a more unbiased loss computation.

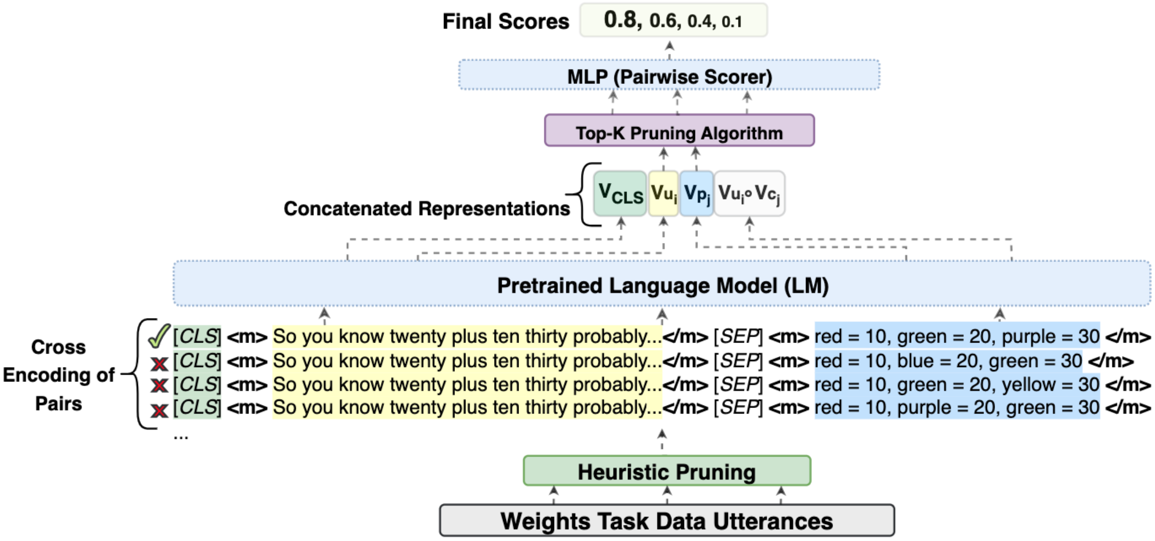

Final Scores
0.8, 0.6, 0.4, 0.1
MLP (Pairwise Scorer)
Top-K Pruning Algorithm
Concatenated Representations
V_{CLS} V_{u_i} V_{p_j} $V_{u_i} \circ V_{c_j}$
Pretrained Language Model (LM)
Cross Encoding of Pairs
✓ [CLS] <m> So you know twenty plus ten thirty probably...</m> [SEP] <m> red = 10, green = 20, purple = 30 </m>
✗ [CLS] <m> So you know twenty plus ten thirty probably...</m> [SEP] <m> red = 10, blue = 20, green = 30 </m>
✗ [CLS] <m> So you know twenty plus ten thirty probably...</m> [SEP] <m> red = 10, green = 20, yellow = 30 </m>
✗ [CLS] <m> So you know twenty plus ten thirty probably...</m> [SEP] <m> red = 10, purple = 20, green = 30 </m>
...
Heuristic Pruning
Weights Task Data Utterances

Figure 7: Schematic overview of the cross-encoder architecture, using example Weights Task data.

minimizes the mean of the Binary Cross Entropy (BCE). More specifically, an utterance (u_i) and a proposition (p_j) are encoded bidirectionally, by interchanging their sequential positions in the input text ((u_i, p_j) and (p_j, u_i)), to avoid positional bias affecting loss computation observed in transformer-based models (Hofstätter et al., 2021). This results in a different unified representation in each direction and we minimize the average of the BCE loss over the encodings in both directions. Mathematically,

$$\mathcal{L}_{\text{BCE}(\theta,\phi)} = -\frac{1}{m}\sum_{i=1}^{m}\left(y_i \cdot \log \hat{y}_i + (1 - y_i) \cdot \log\left(1 - \hat{y}_i\right)\right) \tag{2}$$

where y and $\hat{y}$ are the true and predicted probabilities for an utterance-proposition encoding in one of the directions in a sample batch of size m. θ and ϕ are the parameters of the MLP and the pretrained LM, respectively. We train using a batch size of 20 for 12 epochs, with a learning rate of $1e - 6$ on the LM parameters and $1e - 4$ on the MLP pairwise scorer. With augmented data, the same training procedure is followed but the augmented data is added to the training set for each group-wise fold.

4.4. Experiments

We investigated two methods for extracting propositional content from utterances: a *cosine similarity* baseline, and a *cross-encoder* adapted from entity and event coreference research in the field of NLP. These were both evaluated over the Oracle transcriptions of utterances, and the Google automatic transcriptions, and using various levels of data cleaning to explore performance of the different methods in settings that range from more idealized to more realistic. Below we describe the methodology for cleaning the data and training the cross-encoder.

4.4.1. Cross-Encoder

Heuristic pruning of candidate propositions As mentioned, our data suffers from an imbalance between negative and positive samples, in that the vast majority of candidate propositions are not matches for a given utterance. This phenomenon is also present in common event coreference datasets, which results in a training dataset that is severely imbalanced toward negative pairs if not handled (Ahmed et al., 2023). In our case, it is usually quite obvious when a candidate proposition is not a possible match for an utterance because the candidate does not contain the object or weight value mentioned in the utterance. Therefore, we employ a heuristic pruning strategy on both datasets.

For the Weights Task Dataset, heuristic pruning operates at two levels. 1) We compare all propositions that include both the color and weight mentioned in the utterance (e.g., candidate matches for an utterance containing "red" and "ten" would include $red = 10$, $red \neq 10$, $red < 10$, etc.) 2) If the list of candidates is still empty, as might be the case for utterances such as, e.g., "it's fifty!", we then enlarge the search space by getting all the propositions that contain any of the colors or weights mentioned in the utterance. This process is similar to the lemma-based heuristic pruning used for training a cross-encoder for cross-document event coreference by Ahmed et al. (2023).

DeliData has a similar negative-positive imbalance in the training data. We therefore employ a similar two-step pruning here. 1) We compare all propositions that include the same Card and the Property mentioned in the utterance (e.g., candidate matches for an utterance containing "Z" and "Odd" would include $has(Z, Odd)$ and $\neg has(Z, Odd)$). This essentially involves extracting a set of entities from the utterance and retaining only candidate propositions which contain the equivalent set. 2) If this list is empty, it might be that the utterance is expressing multiple propositions, e.g., "I thought the 2 needed to be turned over since it did not say that all even number cards only have vowels" is expressing the proposition $is(2, Even)$. However, the utterance contains the set of entities $\{2, Even, Vowel\}$ and no propositions contain all of these elements. In this case, in this pruning step, all individual propositions that can be expressed using elements of the utterance are retained as candidates. In this example, $is(2, Even)$, $\neg is(2, Even)$, $has(2, Vowel)$, $\neg has(2, Vowel)$, would be considered candidates.

TRAINING DATA CONSTRUCTION After filtering the candidate propositions with heuristic pruning, to create the training dataset for the cross-encoder, we pair an utterance with its annotated correct proposition as a positive pair and choose four random propositions from the filtered candidate propositions and pair them with the utterance as negative pairs. For example, the WTD utterance "ok so the red has ten" would be a positive match with $red = 10$ and a negative match with only three other candidates generated after pruning. This results in a more balanced ratio of negative to positive candidate propositions for a given utterance, which is beneficial for training. The random selection from the filtered propositions ensures a diverse and robust set of negative samples. We pick only four random negative samples because a significant number of annotated propositions are of the form appropriate for the dataset, e.g., <color, relation, weight>, which means that after the first level of heuristic pruning, certain transcripts would have only four possible candidate propositions, viz. $< color > \{=, \neq, <, >\} < weight >$.

TESTING METHODOLOGY We perform a rotating leave-one-group-out experiment where cross-encoder training is performed over 9 of 10 groups in the WTD, and 99 of 100 groups in DeliData, with the remaining group reserved for the test set. The test group is then rotated through.

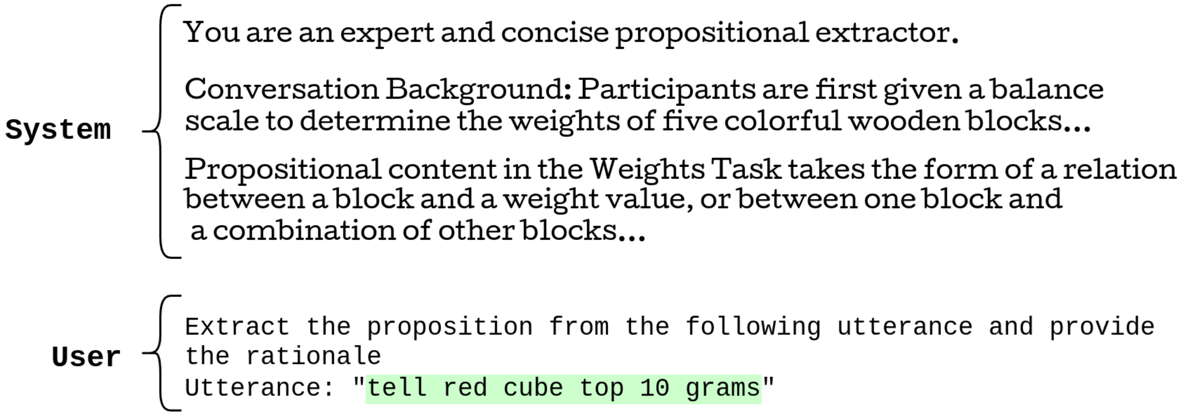
For testing, we use the same pruning methodology as described above for each dataset, but where necessary, further prune the candidate utterance-proposition pairs from the test set using a top-k pruning strategy, for which we use the previously trained cross-encoder. Specifically, we compute the cosine similarities between the embeddings of an utterance and the remaining candidate propositions, while interchanging their mutual positions. For instance, if (u_i, p_j) represents an utterance-proposition pair, we encode both $[V_{u_i}, V_{p_j}]$ and $[V_{p_j}, V_{u_i}]$ to retain their positional information. Since the cross-encoder has been trained to minimize the mean of the bidirectional BCE loss, the latent representations of positive pairs likely point in similar directions in the embedding space vis-à-vis the negative pairs. As such, a top-k pruning strategy allows us to generate the most similar candidate propositions for a particular utterance and remove mismatches which are more obvious. This helps the system's precision by minimizing the loss of pairs during pruning. We use $k = 5$ to ensure approximate consistency with the training set, which has a 1:4 ratio of positive to negative samples. We then score these leftover pairs using our trained cross-encoder. For each utterance, we consider the extracted proposition to be the one with the highest score as given by the cross-encoder since need a ranking system to choose a proposition for the evaluation metrics.

4.4.2. Cosine Similarity

For a given utterance's vector representation, we compute the cosine similarities between the embeddings of all candidate propositions and the utterance embeddings. We then sort these cosine similarities, retrieving the proposition(s) with the most similar embeddings to the utterance embedding. We use the same pruning strategies mentioned in Sec. 4.3 to be consistent. Because cosine similarity calculations only require the utterances to be encoded through a pre-trained model, and no training of a separate model, we simply compare the encodings of utterances to those of propositions without the need for a leave-one-group-out split.

4.4.3. Zero-Shot Baselines

In addition to the cross-encoder and cosine similarity methods, we also establish zero-shot baselines using GPT-4 and LLaMA2-13B, inspired by prior works (Yang et al., 2022). The goal of these baselines is to assess the feasibility of extracting propositions directly from utterances without any explicit training or fine-tuning. The zero-shot approach involves prompting the language models with an utterance and instructing them to identify the underlying proposition. The prompt structure is designed to provide the models with context about the task and the expected format of the output. The model is also asked to produce a rationale for its decision, we leads the model to perform chain-of-thought -style reasoning, which has been shown to elicit improved reasoning in LLMs and guard against erroneous outputs (Wei et al., 2022; Nath et al., 2024; Nath et al., 2024). The use of this technique provides extra guidance to the LLM, resulting in zero-shot baselines that are not artificially low. The specific prompts used for GPT-4 and LLaMA2-13B are shown in Fig. 8. We do not report LLaMA 2-13B zero-shot performance on DeliData, as the model was unable to extract any coherent, properly-formed propositions from the provided utterances given the prompt.



System

You are an expert and concise propositional extractor.

Conversation Background: Participants are first given a balance scale to determine the weights of five colorful wooden blocks...

Propositional content in the Weights Task takes the form of a relation between a block and a weight value, or between one block and a combination of other blocks...

User

```
Extract the proposition from the following utterance and provide
the rationale
Utterance: "tell red cube top 10 grams"
```

Figure 8: Prompt used to establish zero-shot baselines for propositional extraction. The system prompt specifies the task context and defines the structure of propositional content, while the user prompt provides an utterance (e.g., "tell red cube top 10 grams") for which the system must extract the corresponding proposition and rationale. This approach evaluates the model's ability to generalize without prior task-specific training.

4.5. Metrics

To evaluate our system's performance, we use 3 common metrics for retreival tasks: Intersection Over Union (IOU), top-1 accuracy, top-3 accuracy.

IOU measures how well we extract components of propositions, even if the entire proposition is not retrieved perfectly. It calculates the overlap between predicted and true proposition. For example, if the true proposition is $red = 10 \wedge blue = 20 \wedge green = 10$ and we extract proposition $red = 10 \wedge blue = 30$, we consider the cardinality of the intersection of the two sets ($\{red = 10\}$) over their union ($\{red = 10, blue = 20, green = 10, blue = 30\}$). This assesses partial matches where some, but not all, of the correct propositional content is retrieved. In the example, the IOU score would be $\frac{1}{4}$ or 0.25. This is because only one element ($red = 10$) matches out of a total of four unique elements across both propositions.

Top-1 accuracy is stricter; it only counts if we extract the exact proposition. For example, if the true proposition is $red = 10$, the only way to attain a score of 1 is if the prediction is also $red = 10$.

Top-3 accuracy also requires an exact match, but it counts if the correct proposition is among the top three extractions. For instance, if the true proposition is $red = 10$, a set of top three predictions $red \neq 10$, $red > 10$, and $red = 10$, would get a top-3 accuracy score of 1 since the correct proposition is present within the top 3. Top-3 accuracy is not reported for zero-shot baselines since GPT-4 and LLaMA 2 extract only one proposition from each utterance.

5. Results

For the Weights Task Dataset, we report all results across the three different levels of data cleaning discussed in Sec. 4.2.1. Results include the cross-encoder with and without augmented training, the cosine similarity method, and zero shot.[4] First, we established the performance of our methods on a "best case" baseline. Then, we explored how performance was modulated by the level of data cleaning and automatic speech recognition (ASR). Across these data variants, our cross-encoder outperformed the other methods. We then explored how these methods, specifically the cross-encoder, the cosine similarity, and zero shot methods, generalized to a new domain: the DeliData dataset. All results are presented in Tables 1 - 7 while detailing extraction method, and the performance metrics. '-aug' refers to the cross encoder model trained on augmented data.

Best-case Cross-Encoder First, we evaluated our methods on Level 3 (the most rigorous level of data cleaning) with Oracle transcriptions (Table 1). The performance on this data establishes a "best-case" baseline for propositional extraction on the Weight Task, where the transcriptions are manually transcribed and optimally cleaned, removing utterances that contain no color but only a weight value.

Impact of data cleaning In Tables 2–6, we evaluated our methods on increasingly challenging data conditions. Specifically, we first reduced how clean the data was (as detailed in Sec. 4.2.1) to Levels 2 (Table 3) and Level 1 (Table 5), where Level 1 is the most difficult condition.

Impact of automatic speech recognition Then, we explored how ASR transcriptions impacted performance across Level 3 (Table 2), Level 2 (Table 4), and Level 1 (Table 6). As expected, all methods exhibited a drop in performance as data became more realistic (and thus more difficult); however, our cross-encoder-based method consistently outperformed the other methods.

[4] We do not reproduce the Longformer results reported in Venkatesha et al. (2024) because the Longformer-based cross-encoder substantially underperformed the cross-encoders using BERT and RoBERTa, so we focus on those other models here.

Generalization to DeliData Finally, we present the DeliData results in Table 7. Unlike the WTD, DeliData preprocessing does not involve varying levels of cleaning or different transcription methods. Instead it provides a more straightforward evaluation scenario where the dialogue is solely in text form and inherently cleaner, making it approximately comparable of the WTD results in the condition reported in Table 1, with maximal cleaning and Oracle transcription.

Table 1: Propositional extraction performance on the Weights Task dataset with *Level 3 cleaning* using *Oracle* transcriptions. The columns represent IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric

	IOU	**Acc.**	**Top-3**
BERT	0.664	0.640	0.773
BERT-aug	**0.771**	**0.762**	**0.905**
RoBERTa	0.683	0.671	0.829
RoBERTa-aug	0.753	0.747	0.867
BERT-cosine	0.570	0.547	0.747
RoBERTa-cosine	0.337	0.307	0.520
GPT-4	0.659	0.546	–
LLaMA 2	0.643	0.513	–

Table 2: Propositional extraction performance on the Weights Task dataset with *Level 3 cleaning* using *automatic* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	0.635	0.607	0.787
BERT-aug	**0.651**	**0.633**	**0.817**
RoBERTa	0.645	0.607	0.738
RoBERTa-aug	0.648	0.617	0.800
BERT-cosine	0.281	0.262	0.344
RoBERTa-cosine	0.057	0.049	0.147
GPT-4	0.483	0.417	–
LLAMA 2	0.463	0.416	–

6. Model Selection and Statistical Analysis

We selected **Oracle Level 3 RoBERTa** as the best-performing model on the Weights Task Dataset (WTD) based on its superior results across the tested conditions, *assuming no data augmentation* as augmenting data on the WTD did not significantly impact performance. To validate this selection, we performed paired t-tests, comparing Oracle Level 3 RoBERTa IOU scores with each other condition. For each group, the IOU scores were calculated and compared across all models and configurations. The results, as shown in Table 8, reveal that *data cleaning significantly impacts model performance*, with Level 3 cleaning yielding significantly better results than Levels 1 and 2 (Table 8). This supports the conclusion that data cleaning at this level enhances the model's effectiveness. The cross-encoder demonstrates statistically significant improvements over the cosine-based approaches, as seen in the comparisons between RoBERTa Level 3 cross encoder model and and Cosine RoBERTa/BERT.

Furthermore, *there is no statistically significant difference between Google and Oracle transcription systems*, as demonstrated in the non-significant comparisons (Table 8). Similarly, *training with augmented data did not result in statistically significant improvements* over standard training. Both findings suggest that additional complexities, such as alternative transcription systems or data augmentation, do not meaningfully impact performance for this task. We also observed that our model's performance did not exhibit statistically significant differences compared to GPT and LLAMA 2-13B. Given that these models are state of the art models and trained on vast datasets, achieving comparable results without significant performance differences is a promising indication that our cross-encoder approach is an effec-

Table 3: Propositional extraction performance on the Weights Task dataset with *Level 2 cleaning* using *Oracle* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	0.596	0.562	0.730
BERT-aug	**0.639**	**0.607**	**0.831**
RoBERTa	0.585	0.573	0.789
RoBERTa-aug	0.599	0.573	0.753
BERT-cosine	0.505	0.472	0.651
RoBERTa-cosine	0.284	0.258	0.461
GPT-4	0.599	0.472	–
LLaMA 2	0.520	0.460	–

Table 4: Propositional extraction performance on the Weights Task dataset with *Level 2 cleaning* using *automatic* transcriptions. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	0.537	0.526	0.737
BERT-aug	**0.563**	**0.547**	**0.747**
RoBERTa	0.530	0.500	0.697
RoBERTa-aug	0.498	0.480	0.680
BERT-cosine	0.232	0.210	0.276
RoBERTa-cosine	0.052	0.039	0.118
GPT-4	0.391	0.333	–
LLaMA 2	0.418	0.373	–

Table 5: Propositional extraction performance on the Weights Task dataset with Level 1 cleaning using Oracle transcriptions. Columns include IOU (Intersection Over Union), Acc. (Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	0.526	0.496	0.609
BERT-aug	**0.561**	**0.539**	**0.678**
RoBERTa	0.448	0.426	0.557
RoBERTa-aug	0.501	0.474	0.649
BERT-cosine	0.229	0.200	0.356
RoBERTa-cosine	0.419	0.347	0.514
GPT-4	0.453	0.374	–
LLaMA 2	0.336	0.304	–

Table 6: Propositional extraction performance on the Weights Task dataset with Level 1 cleaning using automatic transcriptions. Columns include IOU (Intersection Over Union), Acc. (Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	0.353	0.309	0.427
BERT-aug	0.389	0.345	**0.527**
RoBERTa	0.383	0.336	0.464
RoBERTa-aug	**0.422**	**0.373**	0.473
BERT-cosine	0.036	0.027	0.081
RoBERTa-cosine	0.164	0.114	0.198
GPT-4	0.298	0.261	–
LLaMA 2	0.336	0.304	–

tive alternative with far fewer parameters. This suggests that our method holds substantial potential in practical applications, providing competitive accuracy without the need for large generative models.

Table 7: Propositional extraction performance on the *DeliData dataset*. Columns include IOU (Intersection Over Union), Acc. (Top-1 Accuracy), and Top-3 Accuracy. Bolded values indicate the best performance for each metric.

	IOU	**Acc.**	**Top-3**
BERT	**0.707**	**0.634**	**0.773**
RoBERTa	0.675	0.605	**0.773**
BERT-cosine (+ pruning)	0.499	0.436	0.668
BERT-cosine (− pruning)	0.175	0.102	0.130
RoBERTa-cosine (+ pruning)	0.413	0.344	0.680
RoBERTa-cosine (− pruning)	0.228	0.090	0.165
GPT-4	0.545	0.433	–

Table 8: Paired t-test results (Significant, $p < 0.05$) comparing Oracle Level 3 RoBERTa with other models

Comparison	t-statistic	p-value
Oracle Level 2 RoBERTA	2.35	0.043*
Oracle Level 1 RoBERTA	6.22	<0.001*
Oracle Level 3 BERT-cosine	-1.14	0.027*
Oracle Level 3 RoBERTa-cosine	-1.14	<0.001*
Oracle Level 3 RoBERTa-aug	-1.14	0.285
Google Level 3 RoBERTa	1.01	0.337
GPT-4	0.07	0.941
LLaMA 2	0.06	0.948

* Indicates statistical significance at $p < 0.05$.

7. Discussion

Comparison of data cleaning strategies As expected, with increased levels of data cleaning on the Weight Task, we see a trend of improving performance across all extraction strategies, language models, and transcription methods. The progressive removal of noise, such as incomplete or ambiguous utterances as discussed in Sec. 4.2.1, directly enhances the accuracy and IOU of propositional extraction. This trend is consistent across different language models (BERT, RoBERTa) and holds true whether using manually segmented Oracle transcriptions or automatically generated ASR transcriptions. However, it is important to note that this increase in performance comes at a trade-off. As we apply more rigorous cleaning criteria, the number of usable utterances decreases significantly. With fewer samples to evaluate, the models may become overly tuned to the cleaner dataset. Furthermore, in real-world applications where such extensive cleaning might not be feasible, the performance gains seen under these ideal conditions might not fully translate.

Comparison of extraction methods The cross-encoder consistently outperforms all other baselines across all three metrics. Comparing the extraction methods across Tables 1– 6 shows that the cross-encoder outperforms the cosine baseline by at least 0.2 IOU on average. On the other hand, with a metric that does not reward partial selection, like traditional or Top-1 accuracy, the cross-encoder outperforms the cosine baseline by at least 40%, on average, although the absolute scores are typically lower than the more lenient IOU metric.

COMPARISON OF TRANSCRIPTION METHODS As expected, using automatic transcriptions of the speech leads to a consistent degradation in performance, as automated segmentation and transcription may incorrectly conflate two overlapping utterances from different people, or as annotators leave out or insert words, where such errors are expected to be minimized by a careful human transcriber. However, this degradation can sometimes be quite small, especially at higher levels of data cleaning, when using the cross-encoder, and the BERT or RoBERTa models. For instance, when using the cross-encoder, the accuracy using BERT embeddings of Google automated transcriptions increases from 30.9% at data cleaning Level 1 (least stringent) to 60.7% at data cleaning Level 3 (most stringent), while when using cosine similarity with pruning, accuracy only increases from 14.4% to 26.2%.

COMPARISON OF LANGUAGE MODELS Using embeddings from BERT typically achieves the best performance, but the performance gap with RoBERTa embeddings is usually quite small especially for the cross-encoder. RoBERTa sometimes performs better than BERT on less clean data, which may reflect the larger and more diverse training data of RoBERTa. Both of these models significantly outperform the Longformer model from Venkatesha et al. (2024).

COMPARISON OF AUGMENTED VS. RAW TRAINING DATA Training with augmented data results in a small increase in the performance across all metrics on the WTD. The lack of a significant performance jump can be attributed to the fact that the multimodal nature of the Weights Task. A sentence like “10 10 20” is annotated as $green = 20$. This is because the participant is pointing at the green block at the time, which the human annotators can see, but the utterance does not explicitly convey that green weighs 20g. Thus, GPT-4 does not have the capacity to generate a sentence that is both similar to “10 10 20” *and* makes clear that the intended meaning is $green = 20$. While GPT-4 is capable of generating clean, syntactically correct utterances that can enrich the dataset, it struggles to replicate the nuanced, context-dependent nature of the original utterances. In the case of “10 10 20,” the crucial information—namely, the association of the utterance with the green block—is derived from visual context, something that GPT-4 cannot infer or incorporate when generating new data. This limitation suggests that while data augmentation can help increase the quantity of training data and may offer some performance benefits, it does not necessarily equip the model to better handle the complexities introduced by the multimodal nature of the task. The model’s improved performance with augmented data is therefore marginal, as it still struggles to interpret or generalize from utterances where the meaning heavily relies on non-verbal cues, such as gestures or object references.

We initially hypothesized that the limited availability of annotated multi-modal WTD data was a primary factor limiting model performance. However, the lack of significant gains from this augmentation points to nuances and intricacies involved in proposition expression that go beyond mere data quantity.

COMPARISON OF ZERO-SHOT BASELINES Zero-shot LLM performance reflects the complexity of the task. The utterances are often not clean or complete, and therefore LLaMA 2-13B and GPT-4 are often unable to extract propositions from them. However, zero-shot baselines still follow the same pattern of the cross-encoder and the cosine baselines, where a cleaner dataset results in a better performance. This is expected, as with cleaner data the large language models are provided with coherent sentences with explicit mentions of the blocks and the weights. Level 3 cleaning on the WTD results in a zero-shot IOU comparable to that of the cross-encoder. Zero-shot results on DeliData further confirm the pattern that Deli is most similar to clean versions of the WTD, and that the cleaner utterances hold explicit semantic meaning and convey the relationship between the task-relevant elements.

Our results demonstrate that a fine-tuned cross-encoder model is comparable to baselines from powerful LLMs like GPT-4 and LLaMA 2-13B in the task of propositional extraction from dialogue. While these large language models offer impressive capabilities, their deployment in real-world scenarios, especially within educational contexts, presents significant challenges. Particularly, GPT-4 employs a pay-per-use model, making it financially unsustainable for large-scale or continuous applications. Similarly, even though LLaMA 2 is an open-weight model, running a 13-billion parameter model necessitates access to substantial computation resources including high-end GPUs, which might be prohibitively expensive or unavailable for many communities. Moreover, directly sending transcribed student utterances or other private information to external LLMs, even for zero-shot inference, could pose significant privacy risks and compliance issues.

As noted in Sec. 4.4.3, LLaMA 2-13B failed to extract any coherent propositions from DeliData, highlighting the complex nature of the task and ways in which propositions may be expressed. For instance, when given the utterance, “It’s asking you to test the rule. Would it not prove that the rule is tested if the 2 turns up a vowel?”, which expresses the proposition $has(2, Vowel)$, LLaMA 2-13B returned the proposition, “A is even, B is odd, C is vowel, D is consonant,” which is completely unanchored from the entities and card properties actually contained in the utterance.

COMPARISON OF DATASETS The performance of the best cross-encoder on the DeliData dataset (0.707 IOU, using BERT) falls between the best cross-encoder’s performance on WTD with Level 2 (0.639 IOU, using BERT-sug) and Level 3 (0.771 IOU, using BERT-aug) cleaning. This is in part due to the unimodal (language-only) nature of the DeliData task and the tendency of the participants to be explicit about the elements of the cards and their properties. This may also be an effect of the text-chat nature of the dialogue, where participants explain themselves more fully to avoid ambiguity and misinterpretation by their task partners. The more explicit nature of DeliData utterances is also reflected in the zero-shot performance, which has a much higher baseline on DeliData than on all but the most rigorously cleaned version of the Weights Task data; our cross encoder method is more uniquely suited to handling the ambiguities that arise in a multimodal task like the Weights Task.

7.1. GROUP-WISE ANALYSIS

Figs. 9 shows IOU and 10 shows top-3 accuracy results from the test samples of each group, at Level 1 (most lenient) data cleaning, using BERT embeddings. The plots compare performance using Oracle (left charts) vs. Google (right charts) utterances and compare the cross-encoder to cosine similarity with heuristic candidate pruning.

We can see that cross-encoder performance on Group 7 is nearly identical regardless of which transcription method was used. This is likely because Group 7’s utterances used mainly simple propositions of the form <color> <relation> <weight>. These instances are easy to extract from the transcripts, and the automated transcripts are likely of high-fidelity.

We can see in Fig. 9 that Group 4’s IOU drops significantly when comparing cosine similarity’s performance over Oracle transcriptions vs. over Google transcriptions. While exploring the samples from this group, several issues were noted. We found eight utterances in the Oracle data and only seven in the Google data, meaning that one of the utterances was completely missed by Google ASR. This utterance happened to be very straightforward and easy for the cosine method to classify. The Oracle transcript is simply “blue ten.” Another issue, again due to the segmentation, is Google ASR may merge two utterances. This highlights a limitation of ASR models, where some additional context may needed to know when a speaker has moved to another sentence. Obviously, the main difference between using the different transcription methods is the transcripts themselves. One instance from Group 4 states “easy green block twenty cause ...” whereas Google ASR transcribed the utterance as “okay e green block red block 10 ...”. These results highlight certain issues that should be considered when deploying such an information extraction system over the outputs of an ASR system, as may be required in classroom environments.

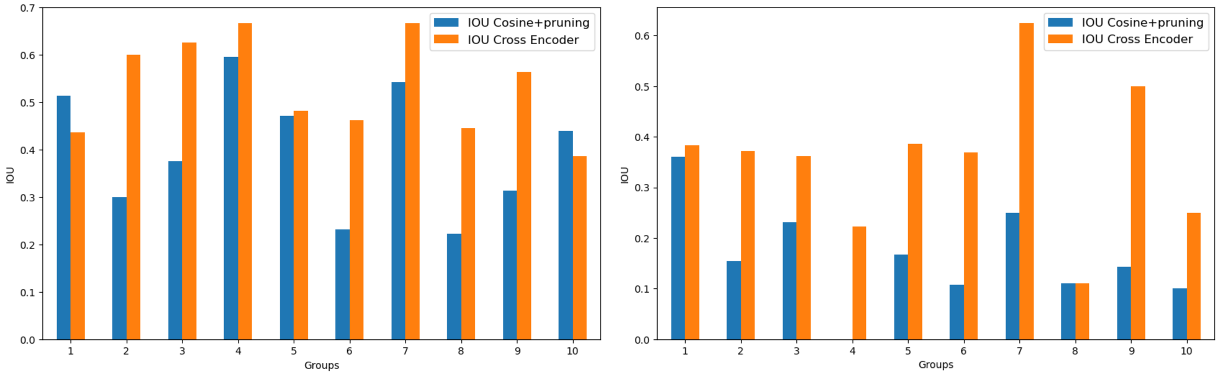


Figure 9: Group-wise *Intersection over Union (IOU)* comparison at Level 1 data cleaning using BERT embeddings. The left chart shows performance with *Oracle* transcriptions, while the right chart reflects performance with *Google ASR* transcriptions. Blue bars represent Cosine Similarity with Pruning, and orange bars represent the Cross-Encoder method across all groups.

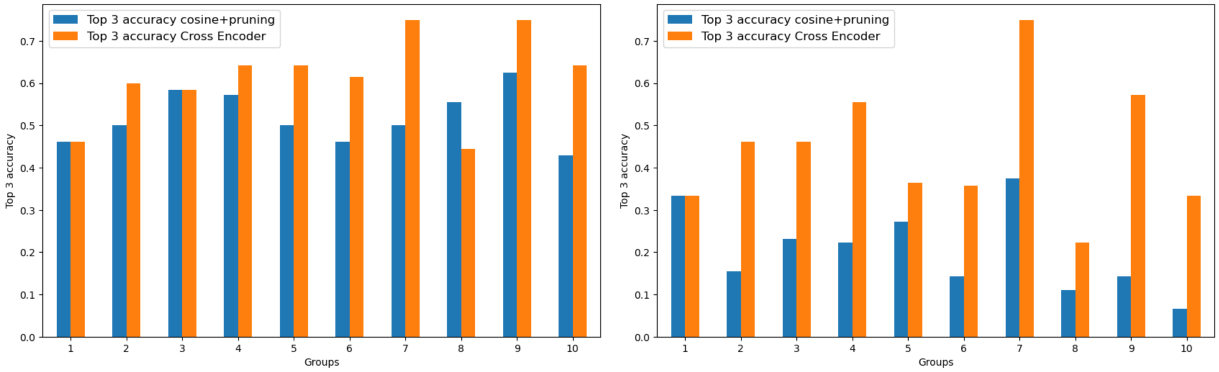


Figure 10: Group-wise *Top-3 Accuracy* comparison at Level 1 data cleaning using BERT embeddings. The left chart displays performance with *Oracle* transcriptions, and the right chart shows performance with *Google ASR* transcriptions. Blue bars represent Cosine Similarity with Pruning, and orange bars represent the Cross-Encoder method across groups.

7.2. Error Analysis

As the cross-encoder is consistently the best-performing extraction method, examining samples it gets wrong is informative. One such example is the utterance “green block one probably twenty ten ten twenty”. The correct proposition is $blue = 10 \wedge green = 20 \wedge red = 10$. The annotators have access to the video and can see that when saying “ten ten twenty,” the speaker is actually pointing to the blue block, then the red block, then the green block. This information is not available through the textual medium alone.

The nature of the DeliData task leads to some errors, mostly pertaining to misinterpreting letters. For example, the utterance “makes sense flip 6 check a vowel” mentions only the card ‘6’. However, since the word ‘a’ could also represent a card (“A”), propositions with the cards 6 and A are retained as candidates, leading to mis-retrieval errors. The same occurs with the card “I”.

ASR transcription errors appear to play a role in zero-shot extraction errors. For example, zero-shot LLMs are unable to extract a proposition from the automatically transcribed utterance “blue block seems done” (actual utterance “blue block seems 10”), because the word “done” does not provide any context to the weight of the blue block. The cross-encoder models successfully retrieve the proposition $blue = 10$ from the utterance, even with the transcription error, because it is trained with task context that links the blue block and weight 10g.

TOP-k ERRORS In order to compare our two extraction methods, we carried out a detailed analysis of candidate propositions that were ranked similarly, based on their cross-encoder scores or cosine similarities. On average, at Level 1 (most lenient) data cleaning, the cross-encoder performs comparatively better at ranking the correct propositions in the top 5. For instance, the cross-encoder ranks 8 and 21 correct propositions higher than the cosine similarity method, for Google and Oracle transcripts respectively. The cosine similarity method ranks 1 (Google) and 11 (Oracle) correct propositions higher.

On the other hand, there were at least 14 Google utterance transcripts and 37 Oracle utterance transcripts where both the extraction methods performed equivalently.

QUALITATIVE ANALYSIS On average, simpler utterances that contain a reference to only one color and/or weight are correctly retrieved by both the cross-encoder and cosine similarity. For instance, “I tell red cube ten grams” (correct proposition $red = 10$) and “green twenty” ($green = 20$). More interestingly, the cross-encoder seems to retrieve utterances with ambiguous context without a direct reference to color or with multiple colors more effectively than the cosine similarity method. For example, “Fifty I” ($yellow = 50$) and “green block twenty red block, blue block ten ten” ($blue = 10 \wedge green = 20 \wedge red = 10$). This is likely due to the cross-encoder’s cross-attention based signals that are being sourced from the entire utterance in the context of the candidate proposition. This was previously observed in (Caciularu et al., 2021) where modeling global signals in parallel with local features led to an overall increase in coreference resolution performance. Since in the actual task data “yellow” was expressed most frequently in the context of “50” and relation $=$, when only “Fifty” is expressed in an utterance, $yellow = 50$ gets the highest score. This is not possible with the cosine similarity since it is not trained on the data and there is no particular relationship between “yellow” and “fifty” in general language.

8. TRANSITION PATH TO REAL-TIME DEPLOYMENT

The feasibility of extracting propositions in an offline setting, as demonstrated in the preceding sections, lays the groundwork for using real-time propositional extraction in classroom-like interactions. The ability to identify and track propositions as they emerge in a live dialogue opens up possibilities for immediate analysis and potential interventions in scenarios such as collaborative problem solving. As an example scenario, we use the *common ground tracking* task presented in Khebour et al. (2024). As mentioned in Sec. 2, the two core modules for common ground tracking are an epistemic positioning classifier (the focus of Khebour et al. (2024)) that operates over the outputs of a *propositional extractor*. We implement this task in a real-time multimodal context that processes live video and speech signals to determine the beliefs shared by the group as they engage in the Weights Task, and replace the cosine similarity-based propositional extraction method used in Khebour et al. (2024) with our cross-encoder-based method. Fig. 11 shows two stills from a live demonstration prototype. In this section we describe the components of such a system, which adapts insights from VanderHoeven et al. (2024) about multimodal agent design to this scenario.

REAL-TIME TRANSCRIPTION AND SEGMENTATION For live deployment, a robust ASR system, such as FasterWhisper, is essential for transcribing and segmenting speech with minimal latency. Our

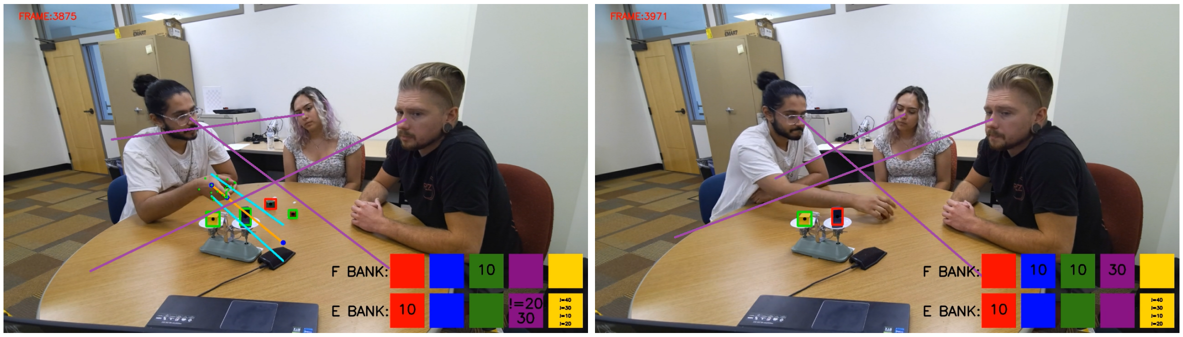


Figure 11: Frames from a live instance of propositional extraction in the Weights Task. Automatically-detected objects are outlined in red or green. Pointing directions are shown in orange (with a pointing cone or "frustum" shown in light blue). Gaze directions are shown in purple. In the left-hand frame, the participant on the left points at the purple and blue blocks and says "so purple is 30 and blue is 10" which the participant on the right affirms. The right-hand frame shows that the propositions $blue = 10$ and $purple = 30$ are successfully extracted (shown by the numbers, signifying weight values, in colored squares symbolizing the blocks), and shown to be accepted as factual by the group.

results indicate that ASR-induced degradation in the propositional extractor's performance is minimal compared to using manual transcriptions.

Multimodal Integration Nonverbal cues like gestures and gaze are vital for disambiguating references in collaborative tasks. Our approach incorporates depth cameras (e.g., Azure Kinect) for 3D body tracking and uses a FasterRCNN model for object detection. Pointing gestures are identified through a MediaPipe-based method, while gaze direction is approximated using nose orientation extracted via body rigs.

Multimodal Dense Paraphrasing Real-time paraphrasing resolves ambiguous references by integrating gesture and object detection. For instance, "this one is 10" accompanied by a pointing gesture toward a blue block is paraphrased as "blue block is 10." This process enriches utterances for more accurate propositional extraction.

Propositional Extraction Model The enriched utterances serve as input for a lightweight cross-encoder model trained on annotated task data. By fusing speech with multimodal cues, the model enables efficient real-time operation, making it suitable for use in environments like classrooms where responsiveness is critical.

Planned Evaluation Formal validation can be conducted of the system prototype in real-time multimodal dialogue contexts. Planned evaluations include:

- Substitution studies: To assess the impact of live performance of each module (e.g., gesture, gaze, ASR) by replacing automated features with ground truth
- Component interaction analysis: To understand how modules influence one another
- Group variability testing: To evaluate adaptability across different group dynamics and task structures

PRELIMINARY METRICS AND EXPECTATIONS Primary evaluation metrics will include the Dice Similarity Coefficient, a statistical measure of similarity used to evaluate the overlap between predicted and ground truth sets, for alignment between predictions and ground truth. Both live and controlled settings will be used, where controlled settings involve predefined, static test cases, while live settings introduce real-time variability, to ensure accuracy and responsiveness, helping refine the system's scalability in dynamic environments.

9. LIMITATIONS

DATASET SIZE AND LOSS DUE TO CLEANING The data preparation and cleaning procedures inevitably result in the loss of several utterances. This leads to small datasets, with the Weights Task Dataset (WTD) ranging from dozens to slightly over 100 utterances, depending on the level of data cleaning, and 255 utterances for DeliData. The reduced dataset size can impact the robustness of the analysis and limit the generalizability of the results.

IMPACT OF AUTOMATED TRANSCRIPTION ERRORS Errors in automated transcripts can adversely affect the efficacy of the candidate pruning process. For example, Google transcribes an utterance as "blue block's obviously time," when the transcribed word "time" was actually uttered as "10." Such transcription errors disrupt the pruning process for candidate propositions, as it incorrectly limits the search space to propositions mentioning "blue" without "10," thereby affecting performance.

DEPENDENCE ON HEURISTIC PRUNING Heuristic pruning of candidate propositions significantly affects performance, as seen in the results of cosine similarity with and without pruning. Pruning not only reduces the search space but also helps maintain a balanced sample distribution for training and aligns test data with the distribution of the training data. However, the pruning methodologies are task-specific and must be adapted to the nature of the propositions in each scenario, limiting the automatic generalizability of the method.

TASK DEFINITION DEPENDENCY The system assumes a well-defined task structure with a finite set of propositions that can be expressed. This reliance on predefined proposition templates means the approach is inherently limited to scenarios where possible outcomes are known a priori. Consequently, the method cannot be readily applied to tasks with open-ended goals or undefined propositional spaces.

RELIANCE ON ANNOTATED TRAINING DATA The system requires access to annotated training data for the cross-encoder model. While synthetic data augmentation using GPT-4 helps mitigate the scarcity of training samples, the generated utterances may fail to capture the nuanced linguistic variations present in collaborative settings. As a result, the quality of augmented data does not fully replicate the richness of real-world dialogues, limiting the system's ability to generalize.

NEED FOR DOMAIN EXPERTISE The approach necessitates domain expertise to define relevant propositions and validate their relevance to the task. Subject matter experts play a crucial role in ensuring that extracted propositions are meaningful and aligned with task requirements. However, this dependence on manual oversight hinders the scalability of the system to new tasks or domains without significant investment in expert resources.

PRIVACY CONCERNS IN REAL-WORLD DEPLOYMENT Participants in the Weights Task Dataset consented to the recording and analysis of their data using third-party tools such as Google ASR for research purposes. However, in real-world classroom implementations, using cloud-based services like

Google ASR raises ethical concerns regarding student privacy. To address this, local custom models would need to be developed and deployed to ensure data privacy and compliance with ethical standards.

10. Conclusions

In this paper, we have defined and explored the complex problem of automatically identifying propositional content from transcriptions of natural speech in a collaborative task. Automated propositional extraction from speech serves a number of important educational purposes. For example, tracking the assertion of propositions over time indicates how students are or are not discussing key concepts relevant to the task, which in turn indicates the construction of shared knowledge (Roschelle and Teasley, 1995).

The Weights Task data presents many challenges, from overlapping speech to incomplete sentences, and we have evaluated a suite of transformer-based language models based on two different methodological frameworks: a cosine similarity baseline vs. a cross-encoder. Our experiments present a feasible method for performing the extraction of task-relevant propositions by building upon publicly-available language models and pairwise representation learning techniques. The successful implementation of the same task on DeliData, a dataset with an entirely different domain, shows the generalizability of our methods given only an inventory of task-relevant propositions, which can be enumerated deterministically. While ground-truth annotation is needed for cross-encoder training, our success on a small amount of data demonstrates the small amount of needed annotation. We have also shown that this task is not a trivial one to be disposed of with off-the-shelf LLMs, as demonstrated by the inferior performance of GPT-4 and LLaMA 2-13B when compared to our own methods.

Our best performing methods, particularly the cross-encoding framework, show a narrow performance gap when operating over automated transcriptions when compared to human transcriptions, suggesting a feasible path forward toward fully automating such a system in a live environment. A clear application in a classroom is in a system that models the shared knowledge of a group toward the task goal, and might be a component of an AI agent who assists small groups in collaborative problem solving (CPS) (Graesser et al., 2018). In Sec. 8 we outlined how we can use our propositional extraction methods in conjunction with live processing of multimodal inputs. We use the example task of common ground tracking as a relevant use case, but due to the modular nature of the present live extraction framework, propositional extraction from real-time dialogue could also be used as a component of other similar inference task, such as multimodal CPS facet classification (Bradford et al., 2023).

Acknowledgment

This material is based in part upon work supported by the National Science Foundation (NSF) under subcontracts to Colorado State University and Brandeis University on award DRL 2019805 (Institute for Student-AI Teaming), and by Other Transaction award HR00112490377 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program. Approved for public release, distribution unlimited. Views expressed herein do not reflect the policy or position of the National Science Foundation, the Department of Defense, or the U.S. Government. All errors are the responsibility of the authors. Our thanks to the anonymous reviewers whose feedback helped improve the final copy of this paper, and to August Garibay, Shamitha Gowra, Carlos Mabrey, and Maniteja Vallala for extensive data annotation.

Use of Generative AI Software Tools

In this study, we utilized generative AI tools for specific purposes. Zero-shot inference was conducted using GPT-4 and LLaMA 2 to establish baseline comparisons for propositional extraction from collaborative dialogues. Additionally, GPT-4 was employed for data augmentation to address data sparsity in the Weights Task Dataset. Specifically, GPT-4 generated supplementary utterances for each unique proposition in the dataset, which were subsequently validated by human annotators to ensure correctness and relevance.

These AI tools were used to enhance methodological rigor and expand the dataset while maintaining alignment with the goals of the study. All other processes, analyses, and interpretations were conducted independently of these tools to preserve the integrity and originality of the research.

References

Ahmed, S. R., Nath, A., Martin, J. H., and Krishnaswamy, N. 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*. Association for Computational Linguistics, Toronto, Canada, 1569–1583.

Ahmed, S. R., Nath, A., Regan, M., Pollins, A., Krishnaswamy, N., and Martin, J. H. 2023. How good is the model in model-in-the-loop event coreference resolution annotation? In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, J. Prange and A. Friedrich, Eds. Association for Computational Linguistics, Toronto, Canada, 136–145.

Beltagy, I., Peters, M. E., and Cohan, A. 2020. Longformer: The long-document transformer. *CoRR abs/2004.05150*, 1–17.

Bethard, S., Yu, H., Thornton, A., Hatzivassiloglou, V., and Jurafsky, D. 2004. Automatic extraction of opinion propositions and their holders. In *Exploring Attitude and Affect in Text: Theories and Applications*. Papers from the 2004 AAAI Spring Symposium. AAAI, Palo Alto, California, 20–27.

Bixler, R., Blanchard, N., Garrison, L., and D'Mello, S. 2015. Automatic detection of mind wandering during reading using gaze and physiology. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ICMI '15. Association for Computing Machinery, New York, NY, USA, 299–306.

Blanchard, N., Donnelly, P., Olney, A. M., Samei, B., Ward, B., Sun, X., Kelly, S., Nystrand, M., and D'Mello, S. K. 2016. Identifying teacher questions using automatic speech recognition in classrooms. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, and A. Gainer, Eds. Association for Computational Linguistics, Los Angeles, 191–201.

Blanchard, N., Moreira, D., Bharati, A., and Scheirer, W. 2018. Getting the subtext without the text: Scalable multimodal sentiment classification from visual and acoustic modalities. In *Proceedings of Grand Challenge and Workshop on Human Multimodal Language (Challenge-HML)*, A. Zadeh, P. P. Liang, L.-P. Morency, S. Poria, E. Cambria, and S. Scherer, Eds. Association for Computational Linguistics, Melbourne, Australia, 1–10.

Bradford, M., Hansen, P., Beveridge, J. R., Krishnaswamy, N., and Blanchard, N. 2022. A deep dive into microphone hardware for recording collaborative group work. In *Proceedings of the 15th International Conference on Educational Data Mining*, A. Mitrovic and N. Bosch, Eds. International Educational Data Mining Society, Durham, United Kingdom, 588–593.

Bradford, M., Khebour, I., Blanchard, N., and Krishnaswamy, N. 2023. Automatic detection of collaborative states in small groups using multimodal features. In *Artificial Intelligence in Education*, N. Wang, G. Rebolledo-Mendez, N. Matsuda, O. C. Santos, and V. Dimitrova, Eds. Springer Nature Switzerland, Cham, 767–773.

Caciularu, A., Cohan, A., Beltagy, I., Peters, M., Cattan, A., and Dagan, I. 2021. CDLM: Cross-document language modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Association for Computational Linguistics, Punta Cana, Dominican Republic, 2648–2662.

Castillon, I., Venkatesha, V., VanderHoeven, H., Bradford, M., Krishnaswamy, N., and Blanchard, N. 2022. Multimodal features for group dynamic-aware agents. In *Interdisciplinary Approaches to Getting AI Experts and Education Stakeholders Talking Workshop at AIEd. International AIEd Society*. Springer Cham, Durham, UK, 1–6.

Cattan, A., Eirew, A., Stanovsky, G., Joshi, M., and Dagan, I. 2021. Cross-document coreference resolution over predicted mentions. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, Online, 5100–5107.

Chai, H. and Strube, M. 2022. Incorporating centering theory into neural coreference resolution. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, M. Carpuat, M.-C. de Marneffe, and I. V. Meza Ruiz, Eds. Association for Computational Linguistics, Seattle, United States, 2996–3002.

Chand, V., Baynes, K., Bonnici, L. M., and Farias, S. T. 2012. A rubric for extracting idea density from oral language samples. *Current Protocols in Neuroscience 58,* 1, 10–5.

Cohen, J. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement 20,* 1, 37–46.

Dennis, S. 2004. An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences 101,* suppl_1, 5206–5213.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186.

Donnelly, P. J., Blanchard, N., Olney, A. M., Kelly, S., Nystrand, M., and D'Mello, S. K. 2017. Words matter: Automatic detection of teacher questions in live classroom discourse using linguistics, acoustics, and context. In *Proceedings of the Seventh International Learning Analytics & Knowledge Conference*. LAK '17. ACM, New York, NY, USA, 218–227.

Donnelly, P. J., Blanchard, N., Samei, B., Olney, A. M., Sun, X., Ward, B., Kelly, S., Nystrand, M., and D'Mello, S. K. 2016. Multi-sensor modeling of teacher instructional segments in live classrooms. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. Association for Computing Machinery, Tokyo, Japan, 177–184.

Evans, J. S. B. 2016. Reasoning, biases and dual processes: The lasting impact of Wason (1960). *Quarterly Journal of Experimental Psychology 69,* 10, 2076–2092.

Gentner, D. 1978. Testing the psychological reality of a representational model. In *Theoretical Issues in Natural Language Processing-2*, D. L. Waltz, Ed. Association for Computational Linguistics, Las Cruces, New Mexico, 1–7.

Gijlers, H. and de Jong, T. 2009. Sharing and confronting propositions in collaborative inquiry learning. *Cognition and Instruction 27,* 3, 239–268.

Graesser, A. C., Fiore, S. M., Greiff, S., Andrews-Todd, J., Foltz, P. W., and Hesse, F. W. 2018. Advancing the science of collaborative problem solving. *Psychological Science in the Public Interest 19,* 2, 59–92.

GROSZ, B. J. AND SIDNER, C. L. 1986. Attention, intentions, and the structure of discourse. *Computational Linguistics 12,* 3, 175–204.

HELD, W., ITER, D., AND JURAFSKY, D. 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*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 1406–1417.

HERMANN, K. M., KOCISKY, T., GREFENSTETTE, E., ESPEHOLT, L., KAY, W., SULEYMAN, M., AND BLUNSOM, P. 2015. Teaching machines to read and comprehend. *Advances in Neural Information Processing Systems 28*, 1693–1701.

HOFSTÄTTER, S., LIPANI, A., ALTHAMMER, S., ZLABINGER, M., AND HANBURY, A. 2021. Mitigating the position bias of transformer models in passage re-ranking. In *Advances in Information Retrieval: 43rd European Conference on IR Research, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part I 43*. Springer, virtual event, 238–253.

HUMEAU, S., SHUSTER, K., LACHAUX, M.-A., AND WESTON, J. 2020. Poly-encoders: Architectures and pre-training strategies for fast and accurate multi-sentence scoring. In *International Conference on Learning Representations*. ICLR, Addis Ababa, Ethiopia, 1–14.

JEON, S. AND STRUBE, M. 2020. Centering-based neural coherence modeling with hierarchical discourse segments. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Online, 7458–7472.

JO, Y., VISSER, J., REED, C., AND HOVY, E. 2019. A cascade model for proposition extraction in argumentation. In *Proceedings of the 6th Workshop on Argument Mining*, B. Stein and H. Wachsmuth, Eds. Association for Computational Linguistics, Florence, Italy, 11–24.

JO, Y., VISSER, J., REED, C., AND HOVY, E. 2020. Extracting implicitly asserted propositions in argumentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, B. Webber, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Online, 24–38.

KARADZHOV, G., STAFFORD, T., AND VLACHOS, A. 2023. Delidata: A dataset for deliberation in multi-party problem solving. *Proceedings of the ACM on Human-Computer Interaction 7,* CSCW2, 1–25.

KHEBOUR, I., BRUTTI, R., DEY, I., DICKLER, R., SIKES, K., LAI, K., BRADFORD, M., CATES, B., HANSEN, P., JUNG, C., WISNIEWSKI, B., TERPSTRA, C., HIRSHFIELD, L., PUNTAMBEKAR, S., BLANCHARD, N., PUSTEJOVSKY, J., AND KRISHNASWAMY, N. 2024. When text and speech are not enough: A multimodal dataset of collaboration in a situated task. *Journal of Open Humanities Data 10,* 1 (Jan), 1–7.

KHEBOUR, I. K., LAI, K., BRADFORD, M., ZHU, Y., BRUTTI, R. A., TAM, C., TU, J., IBARRA, B. A., BLANCHARD, N., KRISHNASWAMY, N., AND PUSTEJOVSKY, J. 2024. Common ground tracking in multimodal dialogue. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. ELRA and ICCL, Torino, Italia, 3587–3602.

KULHÁNEK, J., HUDEČEK, V., NEKVINDA, T., AND DUŠEK, O. 2021. AuGPT: Auxiliary tasks and data augmentation for end-to-end dialogue with pre-trained language models. In *Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI*, A. Papangelis, P. Budzianowski, B. Liu, E. Nouri, A. Rastogi, and Y.-N. Chen, Eds. Association for Computational Linguistics, Online, 198–210.

LAPPIN, S. AND LEASS, H. J. 1994. An algorithm for pronominal anaphora resolution. *Computational Linguistics 20,* 4, 535–561.

LI, J., LUONG, T., AND JURAFSKY, D. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, C. Zong and M. Strube, Eds. Association for Computational Linguistics, Beijing, China, 1106–1115.

LIU, Y., OTT, M., GOYAL, N., DU, J., JOSHI, M., CHEN, D., LEVY, O., LEWIS, M., ZETTLEMOYER, L., AND STOYANOV, V. 2020. Roberta: A robustly optimized BERT pretraining approach. In *International Conference on Learning Representations*. ICLR, Addis Ababa, Ethiopia, 1–15.

NALLAPATI, R., ZHOU, B., DOS SANTOS, C., GULÇEHRE, Ç., AND XIANG, B. 2016. Abstractive text summarization using sequence-to-sequence RNNs and beyond. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, S. Riezler and Y. Goldberg, Eds. Association for Computational Linguistics, Berlin, Germany, 280–290.

NATH, A., MANAFI AVARI, S., CHELLE, A., AND KRISHNASWAMY, N. 2024. Okay, let‘s do this! modeling event coreference with generated rationales and knowledge distillation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, K. Duh, H. Gomez, and S. Bethard, Eds. Association for Computational Linguistics, Mexico City, Mexico, 3931–3946.

NATH, A., MANNAN, S., AND KRISHNASWAMY, N. 2023. AxomiyaBERTa: A phonologically-aware transformer model for Assamese. In *Findings of the Association for Computational Linguistics: ACL 2023*, A. Rogers, J. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, Toronto, Canada, 11629–11646.

NATH, A., VENKATESHA, V., BRADFORD, M., CHELLE, A., YOUNGREN, A. C., MABREY, C., BLANCHARD, N., AND KRISHNASWAMY, N. 2024. “any other thoughts, hedgehog?” linking deliberation chains in collaborative dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, Eds. Association for Computational Linguistics, Miami, Florida, USA, 5297–5314.

PACUIT, E. 2017. *Neighborhood Semantics for Modal Logic*, 1st ed. Springer Publishing Company, Incorporated, New York, NY.

PELLICER, L. F. A. O., FERREIRA, T. M., AND COSTA, A. H. R. 2023. Data augmentation techniques in natural language processing. *Applied Soft Computing 132*, 109803.

PIECH, C., BASSEN, J., HUANG, J., GANGULI, S., SAHAMI, M., GUIBAS, L. J., AND SOHL-DICKSTEIN, J. 2015. Deep knowledge tracing. In *Advances in Neural Information Processing Systems*, C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, Eds. Vol. 28. Curran Associates, Inc., Montreal, Canada, 505 – 513.

RAVI, S., TANNER, C., NG, R., AND SHWARTZ, V. 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*, A. Vlachos and I. Augenstein, Eds. Association for Computational Linguistics, Dubrovnik, Croatia, 1708–1724.

REIMERS, N. AND GUREVYCH, I. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Association for Computational Linguistics, Hong Kong, China, 3982–3992.

ROSCHELLE, J. AND TEASLEY, S. D. 1995. The construction of shared knowledge in collaborative problem solving. In *Computer Supported Collaborative Learning*, C. O'Malley, Ed. Springer Berlin Heidelberg, Berlin, Heidelberg, 69–97.

SUN, C., SHUTE, V. J., STEWART, A., YONEHIRO, J., DURAN, N., AND D'MELLO, S. 2020. Towards a generalized competency model of collaborative problem solving. *Computers & Education 143*, 103672.

TERPSTRA, C., KHEBOUR, I., BRADFORD, M., WISNIEWSKI, B., KRISHNASWAMY, N., AND BLANCHARD, N. 2023. How good is automatic segmentation as a multimodal discourse annotation aid? In *Proceedings of the 19th Joint ACL-ISO Workshop on Interoperable Semantics (ISA-19)*, H. Bunt, Ed. Association for Computational Linguistics, Nancy, France, 75–81.

TU, J., RIM, K., HOLDERNESS, E., YE, B., AND PUSTEJOVSKY, J. 2023. Dense paraphrasing for textual enrichment. In *Proceedings of the 15th International Conference on Computational Semantics (IWCS)*. Association for Computational Linguistics, Nancy, France, 39–49.

TU, J., RIM, K., AND PUSTEJOVSKY, J. 2022. Competence-based question generation. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 1521–1533.

VANDERHOEVEN, H., BRADFORD, M., JUNG, C., KHEBOUR, I., LAI, K., PUSTEJOVSKY, J., KRISHNASWAMY, N., AND BLANCHARD, N. 2024. Multimodal design for interactive collaborative problem-solving support. In *Human Interface and the Management of Information*, H. Mori and Y. Asahi, Eds. Springer Nature Switzerland, Cham, 60–80.

VASWANI, A., SHAZEER, N., PARMAR, N., USZKOREIT, J., JONES, L., GOMEZ, A. N., KAISER, L., AND POLOSUKHIN, I. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*. NIPS'17. Curran Associates Inc., Red Hook, NY, USA, 6000—-6010.

VENKATESHA, V., NATH, A., KHEBOUR, I., CHELLE, A., BRADFORD, M., TU, J., PUSTEJOVSKY, J., BLANCHARD, N., AND KRISHNASWAMY, N. 2024. Propositional extraction from natural speech in small group collaborative tasks. In *Proceedings of the 17th International Conference on Educational Data Mining*, B. Paaßen and C. D. Epp, Eds. International Educational Data Mining Society, Atlanta, Georgia, USA, 169–180.

WEBB, N. M., ING, M., BURNHEIMER, E., JOHNSON, N. C., FRANKE, M. L., AND ZIMMERMAN, J. 2021. Is there a right way? Productive patterns of interaction during collaborative problem solving. *Education Sciences 11,* 5, 214.

WEI, J., WANG, X., SCHUURMANS, D., BOSMA, M., XIA, F., CHI, E., LE, Q. V., ZHOU, D., ET AL. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems 35*, 24824–24837.

YANG, X., PEYNETTI, E., MEERMAN, V., AND TANNER, C. 2022. What GPT knows about who is who. In *Proceedings of the Third Workshop on Insights from Negative Results in NLP*, S. Tafreshi, J. Sedoc, A. Rogers, A. Drozd, A. Rumshisky, and A. Akula, Eds. Association for Computational Linguistics, Dublin, Ireland, 75–81.

YU, X., YIN, W., AND ROTH, D. 2022. Pairwise representation learning for event coreference. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, V. Nastase, E. Pavlick, M. T. Pilehvar, J. Camacho-Collados, and A. Raganato, Eds. Association for Computational Linguistics, Seattle, Washington, 69–78.

ZENG, Y., JIN, X., GUAN, S., GUO, J., AND CHENG, X. 2020. Event coreference resolution with their paraphrases and argument-aware embeddings. In *Proceedings of the 28th International Conference* *on Computational Linguistics*, D. Scott, N. Bel, and C. Zong, Eds. International Committee on Computational Linguistics, Barcelona, Spain (Online), 3084–3094.

ZHANG, T., KISHORE, V., WU, F., WEINBERGER, K. Q., AND ARTZI, Y. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*. ICLR, Virtual, 1–43.