Coherence-Driven Multimodal Safety Dialogue with Active Learning for Embodied Agents

Sabit Hassan
University of Pittsburgh
Pittsburgh, PA, USA
sabit.hassan@pitt.edu

Hye-Young Chung
Northeastern University
Boston, MA, USA
chung.hyey@northeastern.edu

Xiang Zhi Tan
Northeastern University
Boston, MA, USA
zhi.tan@northeastern.edu

Malihe Alikhani
Northeastern University
Boston, MA, USA
m.alikhani@northeastern.edu

ABSTRACT

When assisting people in daily tasks, robots need to accurately interpret visual cues and respond effectively in diverse safety-critical situations, such as sharp objects on the floor. In this context, we present **M-CoDAL**, a multimodal-dialogue system specifically designed for embodied agents to better understand and communicate in safety-critical situations. The system leverages discourse coherence relations to enhance its contextual understanding and communication abilities. To train this system, we introduce a novel clustering-based active learning mechanism that utilizes an external Large Language Model (LLM) to identify informative instances. Our approach is evaluated using a newly created multimodal dataset comprising **1K** safety violations extracted from **2K** Reddit images. These violations are annotated using a Large Multimodal Model (LMM) and verified by human annotators. Results with this dataset demonstrate that our approach improves resolution of safety situations, user sentiment, as well as safety of the conversation. Next, we deploy our dialogue system on a Hello Robot Stretch robot and conduct a within-subject user study with real-world participants. In the study, participants role-play two safety scenarios with different levels of severity with the robot and receive interventions from our model and a baseline system powered by OpenAI's ChatGPT. The study results corroborate and extend the findings from the automated evaluation, showing that our proposed system is more persuasive in a real-world embodied agent setting.

KEYWORDS

Embodied Agents; Multimodal Safety; Active Learning; Coherence Theory

ACM Reference Format:
Sabit Hassan, Hye-Young Chung, Xiang Zhi Tan, and Malihe Alikhani. 2025. Coherence-Driven Multimodal Safety Dialogue with Active Learning for Embodied Agents. In *Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Detroit, Michigan, USA, May 19 – 23, 2025*, IFAAMAS, 10 pages.

Proc. of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025), Y. Vorobeychik, S. Das, A. Nowé (eds.), May 19 – 23, 2025, Detroit, Michigan, USA. © 2025 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org).

1 INTRODUCTION

Embodied agents, such as assistive robots with multimodal capabilities, are poised to become an integral part of our daily lives by helping with household tasks and providing assistance upon request. However, the agent should also play a proactive role and ensure both its and the user's actions are safe. It needs to not only detect unsafe scenarios but also communicate and persuade the user of the danger. To achieve this, the system needs to understand and adapt to conversational contexts while remaining robust across diverse safety-related scenarios.

To meet this challenge, we integrate theories of coherence relations to enhance contextual understanding and use clustering-based active learning to train a multimodal dialogue system, **M-CoDAL**, specifically designed for deploying embodied agents for real-world users to provide effective safety advice.

M-CoDAL uses discourse coherence relations to enrich its contextual understanding. Theories of coherence relations originate in understanding and analyzing inferential links to support text interpretation and has subsequently been extended to cross-modal settings [3]. Coherence relations can situate an ongoing scene in the arc of a narrative [13] or enrich interpretation of communicative actions across modalities [27]. For instance, the coherence relation *Cause* can help interpret why paying attention to a safety violation is important in an image captured by a robot given a certain *Condition* relation. These relations facilitate dialogue between the human user and the robot. By default, a robot equipped with pretrained models would not explicitly model such inferential links [3]. We hypothesize that embodied agents that parse coherence relations of safety violation in captured images and use them to guide their responses will better understand contexts of safety and engage in safer conversations with humans.

To train **M-CoDAL** for embodied agents with coherence relation integration, we adopt clustering-based active learning [36]. Active learning offers advantage over standard fine-tuning methods by focusing resources toward informative instances. When combined with clustering, active learning has been shown to yield more representative models [19]. Thus, we hypothesize that applying clustering-based active learning would lead to better coverage of safety scenarios and a safer multimodal dialogue system. Active learning, however, has mostly been addressed in the context of classification tasks and has remained challenging for generative tasks [34] such as conversational response generation due to difficulty

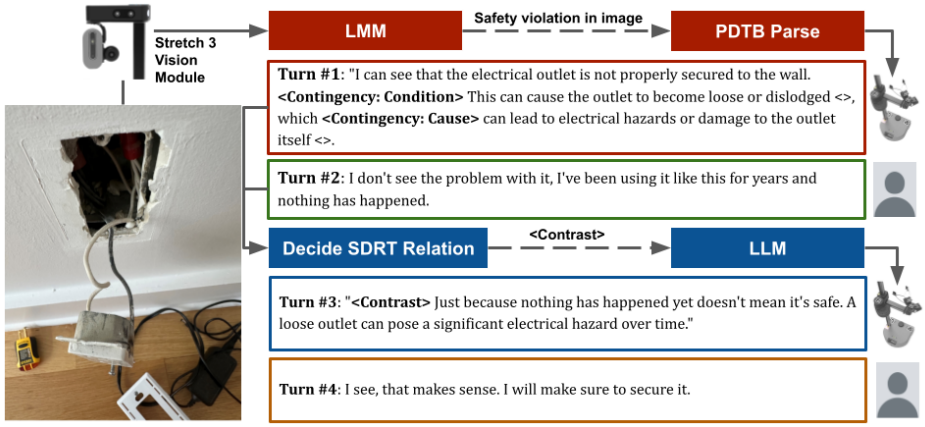


Figure 1: Our proposed dialogue system first parses safety violation in an image with Penn Discourse Treebank [35] relations, and then generate an appropriate response by choosing an Segmented Discourse Representation Theory relation [6].

in estimating model uncertainty in a large output space. We make the first inroads in active learning for autonomous agents with conversational capacity by integrating an external LLM that quantifies informativeness of an instance based on a composite score of safety of conversation, resolution, and user sentiment. Further, we distill knowledge from another external LLM to reduce necessary human efforts in active learning. Our work is also one of the first to deploy and evaluate active learning in practice for human-robot interactions.

To train our system and evaluate our approach, we first construct a novel dataset consisting of **1K** safety violations obtained from **2K** Reddit images. The safety violations are obtained using a Large Multimodal Model (LMM) and then verified, or edited if necessary, by human annotators. Coherence relations and appropriate responses based on the coherence relations are distilled from GPT-4 [1] for these safety violations and used to train **M-CoDAL**, based on a smaller LLM, Mistral-7B [26], in an active learning setting. Our automated evaluation demonstrates that integrating coherence relations leads to higher safety scores, further improved by clustering-based active learning. We also observe that the improvements translate to models such as Llama 3 [18] and Qwen [9] that were not part of the original active learning loop.

Finally, we deploy our multimodal dialogue system on a **Hello Robot Stretch** robot and conduct a within-subject in-person user study where participants interact with our robot in staged unsafety scenarios. 8 participants interacted with the robot operating with GPT-4o or our proposed system, **M-CoDAL** in low and high severity fake unsafe scenarios. The study results corroborate findings of automated evaluation by showing that participants found our proposed system **M-CoDAL** to be more persuasive in both conditions. Our qualitative analysis of interviews with the participants also reveals that the participants find **M-CoDAL** to be more attentive to safety situation and can help the user to be more aware of safety situations. Thus, the key contributions of this paper are:

- A first-of-its-kind publicly available dataset [1] of multimodal dialogues of safety for embodied agents.
- A multimodal dialogue system, **M-CoDAL** that better understands context of safety via coherence relations.
- Extension of active learning paradigm for conversational embodied agent with integration of an external LLM.
- Deployment of **M-CoDAL** with a Hello Robot Stretch robot, accompanied with findings of a real-world user study.

[1] https://github.com/sabithsn/multimodal-embodied-safety

2 RELATED WORK

Safety in Automated Agents. There has been a growing interest in development of multimodal dialogue systems in recent years [28, 37, 39], especially with the advent of large multimodal models such as GPT-4V [1] and LLaVA [32]. Discussions of safety however, has primarily focused on textual domain [7, 16, 38, 41] through means of text-classification or integrating guardrails within LLMs [18]. These approaches may not be sufficient to tackle safety situations that may arise in multimodal household scenarios, particularly when the context demands further probing. To our knowledge, our work is the first public work to process visual cues of safety in multimodal dialogues with an embodied agent.

Coherence Relations for Contextual Understanding. Coherence relations have been proposed as a possible method for controlling generative models and have been shown to aid tasks such as extractive and abstractive summarization [12, 42]. Coherence relation-aware models have also been shown to generate more coherent texts [10] within a reinforcement learning setting. Alikhani et al. [4] extend standard text-based discourse relations to cross-modal scenarios. We build on theories of coherence relations to capture the context of safety scenarios and control responses accordingly within a multimodal dialogue system.

Active Learning with Natural Language. Active learning is a prominent area in machine learning [36], and has gained recent attention for tasks involving natural language [46] such as intent classification, sentence matching, and named entity recognition [8, 31, 45]. However, active learning has predominantly been focused on classification tasks. Recent works targeting LLMs [15, 22, 33] also focus on tasks with fixed sets of outputs, leaving

active learning for generative tasks largely unexplored [23, 34]. Our work is among the first to pioneer active learning for conversational generative tasks and also extend to multimodal scenarios. Further, we deploy and evaluate active learning in practice, whereas prior work have primarily relied on simulations [46].

Embodied agents with AI models. With advancement in language models and multimodal models, there has been a surge in research to integrate these models with embodied agents [2, 17, 24, 25]. Most of these recent works focus on extracting plans or executable actions by the robot from pre-trained language models. In contrast, our work enables the robot to observe safety scenarios in its surroundings and communicate with the user. Our real-world user study, conducted by enacting scenarios letting the users interact with the robot, also reveals insights on how human users may perceive such an agent.

3 MULTIMODAL DIALOGUE FRAMEWORK

We first outline the setting of our dialogue system **M-CoDAL**, followed by integration of coherence relations and use of clustering-based active learning for training the dialogue system.

3.1 Dialogue System

The first input to our dialogue system is an image that contains a potential safety violation. These safety violations may occur during household tasks or within a living environment. For training, the learner LLM is fine-tuned with four turns of simulated conversation:

Turn #1: A Large *Multimodal* Model (LMM) processes the image and generates a message describing the safety violation in the image. **Turn #2:** A user responds to the safety violation issue raised in the first turn. **Turn #3:** A Large Language Model (LLM) processes the previous two turns and generates a response. **Turn #4:** The user makes the final response in the conversation.

Turn #1 is obtained by processing image in our dataset and the subsequent turns are simulated by LLMs for training. During the user-study phase, the images are captured by an actual robot in household environment and the conversations are **not** restricted to a specific number of turns, continuing indefinitely.

3.2 Coherence Relations

In our work, we consider two prominent frameworks for discourse coherence relations (Figure 1). The first is Penn Discourse Treebank (PDTB) [30] and the second is Segmented Discourse Representation Theory (SDRT) [6]. PDTB coherence relations such as *Cause* focus on the local relations between adjacent or nearby textual units. SDRT coherence relations such as *Background* aim to capture semantic and pragmatic discourse structure.

Parsing Safety Violation with PDTB Relations: Penn Discourse TreeBank (PDTB) is particularly suitable for identifying coherence relations within and across sentences. The safety violation obtained from LMM in Turn #1 is parsed by an external LLM. The LLM is asked to consider PDTB relations that occur more than 1% in intra-sentential scenarios [29]: *Concession, Contrast, Cause, Cause+Belief, Condition, Purpose, Conjunction, Instantiation, Level-of-detail, Manner, Substitution, Asynchronous, Synchronous.* The parsed safety violation in Turn #1, along with Turn #2 are passed as context to the LLM to generate Turn #3.

Introducing SDRT Relations in Dialogue: While generating Turn #3, we let an external LLM decide the appropriate discourse coherence relation to be maintained in the response. Since the coherence relation to be maintained here is at a turn-level, we opt for Segmented Discourse Representation Theory (SDRT) [6] as SDRT has been shown to be effective at turn-level in conversational scenarios [5]. The external LLM is provided with 16 SDRT relations listed in [5]: *Continuation, Result, Elaboration, Conditional, Contrast, Answer / Question answer pair, Q-elab / Follow-up question, Acknowledgement, Narration, Correction, Explanation, Alternation, Parallel, Commentary, Clarification Q, Background.*

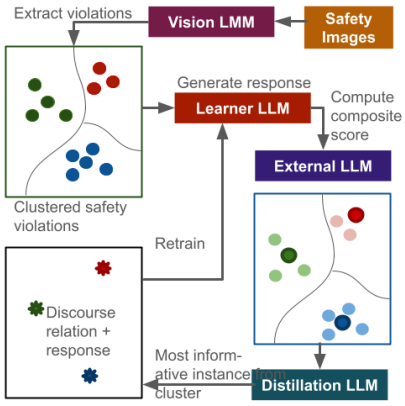


Figure 2: In our active learning loop, informative instances are identified by an external LLM using a composite score. These are distilled using another LLM for coherence relations and responses and are used to retrain the learner LLM.

3.3 Active Learning

Preliminaries. We assume there is a pool of unlabeled dataset U but only a subset of labeled data L can be used for training. L is iteratively constructed by querying target output for the *most-informative* instance. While other active learning scenarios exist [36], we follow the setting of *pool-based* active learning because of its relevance to our setting, where we can obtain a large number of safety images but obtaining coherence relations and appropriate responses can be challenging. In a standard classification setting, active learning would typically identify informative instances from this pool with measures of uncertainty, such as entropy [36]:

$$x_E^* = \underset{x}{argmax} \; -\sum_i P_\theta(y_i|x) log P_\theta(y_i|x) \quad (1)$$

In Eq. 1, y_i is the i^{th} possible output for input x.

Active Learning for Dialogue. Standard measures of informativeness such as entropy, however, cannot be a useful measure in generative setting [34] such as dialogue systems. This is because, as opposed to classification models, generative models such as LLMs have a massive output space and entropy over such a large space may not indicate informativeness properly. Thus, we propose to

replace entropy with a new composite score calculated by an external LLM. To calculate this score, we assume that $\{p_1, p_2, ...p_k\}$ is a set of attributes we expect the generated output to preserve. In our setting, we want the output to be safe and also resolve the safety situation with the user while not upsetting the user. In this case, p_1 is the safety of the generated response and p_2 is the resolution score, and p_3 is the user sentiment. Then, we can define our informative instance as:

$$x_E^* = \underset{x}{argmax} \; -\sum_k Z_\theta(p_k|G(x)) \tag{2}$$

In Eq. 2, Z_θ is an external LLM and $G(x)$ is the output of the learner LLM in our dialogue system.

Clustering-based Active Learning. Standard Active Learning offers label efficiency over random sampling. It can however, induce bias if the model misjudges its confidence [21]. Clustering, which naturally garners diverse samples [44], combined with active learning, can counteract this by simultaneously gathering diverse and informative data. We hypothesize that using an LLM on these diverse and informative data would lead to a more representative set of generations. In our clustering-based setting, the unlabeled data is first vectorized and then the vector space is split into m clusters $\{C_1, C_2, ...C_m\}$, where m is a predefined number. Informativeness measure according to Eq. 2 is calculated for each instance within a cluster and most informative samples are chosen from each cluster. These samples are then passed to a distillation LLM for obtaining coherence relation and appropriate response. The combined approach of integrating discourse relation and active learning is illustrated in Figure 2 and summarized in Algorithm 1.

Algorithm 1 Coherence-Relation Integrated Active Learning

$U, L \leftarrow$ unlabeled data, labeled data
$P = \{p_1, p_2, ...p_k\} \leftarrow$ set of attributes
$D_p, D_s \leftarrow$ LLM for obtaining PDTB/SDRT relations
$S \leftarrow$ distillation LLM
$Z \leftarrow$ composite external LLM
$G \leftarrow$ bootstrapped learner model
$B, N \leftarrow$ labeling budget, annotation batch size
$m \leftarrow$ number of clusters
Cluster V into $\{C_1, C_2, ... C_m\}$
while $B \geq 0$ **do**
  for `i=0,1,...m` **do**
    for `j=0,1,...`$|C_i|$ **do**
      $E_{ij} \leftarrow \sum_k Z_\theta(p_k|G(x))$
    end for
    $x_i^* \leftarrow \underset{j}{argmax}(E_{ij})$
    $D_{pi} = D_p(x_i*)$
    $D_{si} = D_s(x_i*)$
    $y_i^* \leftarrow$ Distill $S((x_i * |D_{pi}, D_{si})$
    Add (x_i^*, y_i^*) to L
  end for
  $G \leftarrow$ retrain on L
  $B = B - N$
end while

3.4 Integration with Embodied Agent

Once the dialogue system has been trained with our active learning paradigm, it is ready to be deployed with an embodied agent. We assume the embodied agent has wireless connectivity and can access a vision module that can capture images of its surroundings. Images captured by the agent are sent via wireless connectivity to a server where the dialogue system resides. The image is first checked for safety violations using a Large Multimodal Model, and then **M-CoDAL** activates to communicate with the user.

4 DATASET

To train and automatically evaluate our system, We construct a dataset of multimodal safety by obtaining images from Reddit, identifying safety violations using a Large Multimodal Model and two stages of human annotation.

4.1 Dataset Construction

Data Collection. We choose Reddit as our data source to obtain safety-related images due to its diversity [20, 43]. We query 14 subreddits (e.g., KitchenConfidential, CookingFails, HomeImprovement, DIY) with safety related keywords (e.g., 'kitchen fire', 'stove', 'unattended cooking', 'grease fire', 'electrical outlet fire') and obtain **2K** relevant posts with images.

Image Annotation. We ask two graduate student annotators to decide if the post in the image contains a safety violation that could occur in indoor setting. Outdoor images (e.g., camp fire) and memes are discarded. After annotation, 507 images were retained and rest were discarded from the dataset. The inter-annotator agreement is 71.6 (Cohen's κ), suggesting substantial agreement. The graduate students are paid according to our institution's standard rate.

Safety Violations using LMM. The 507 images were then passed through LLaVa 1.6 [32]. The LMM was prompted to list safety violations present in the images. In addition to the images, we also passed in the original titles of the Reddit post to the LMM. Since an image can have multiple violations, **1015** safety violations were obtained from the 507 images.

Safety Violation Annotation. As the LMM may incur errors in the dataset, the output of the LMM is then annotated by the same graduate annotators. The annotators were asked to take one of three possible actions: i) mark as correct if the LMM output corresponds to a safety violation in the image, ii) if small edits could fix the LMM output, then edit, and iii) discard the safety violation if it does not correspond to the input image. Following this stage of annotation, 107 were discarded, 825 safety violations were retained as is, and 83 were retained after editing, for a total of 908 safety violations. Figure 4 shows examples from the dataset. This dataset is the seed dataset U in Algorithm 1.

4.2 Dataset Analysis

Safety Distribution. Figure 3 shows distribution of keywords in our collected data. We can observe that some types of safety, such as mold, are more prevalent compared to others such as fire hazard. This can be attributed to the natural distribution of content on social media.

Table 1: Examples of errors made by LMM for identifying safety violation in image. Highest source of error results from either hallucinating safety violation when there is none, or missing important detail of a present safety.

Error Category	Perc.	Error Description	Example
OCR	3%	failed in OCR/ reasoning on top of OCR.	LMM did not read expired date of fire extinguisher.
Visual Impair.	6%	LMM did not recognize object correctly	LMM mistook broken glass as knife.
Mismatch	15%	something went wrong in processing image	egg overcooked, but called undercooked.
Unimportant	21%	the LMM often output concerns that would not considered safety violation typically	lack of proper attire, chocolates could have allergen.
Hallucination	27%	no evidence of safety violation in the image	broken pieces when there are none.
Missed Safety	28%	the LMM missed the actual safety concern	mold, or broken stovetop.

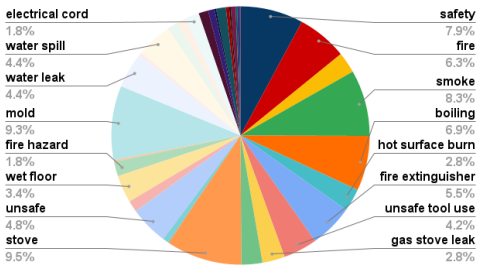


Figure 3: Distribution of safety keywords in our dataset. Safety violations with mold and stove are more prevalent.

Error Analysis. We conducted error analysis on 100 errors made by the LMM according to the human annotators. The errors could be classified into six categories, as shown in Table 1. *Hallucinating* safety concerns when there is none and *missing* certain safety concerns contribute most to the errors. It is important to note that these errors in the dataset were addressed through human edits or by discarding if they could not be fixed with editing.

Comparison with Human Annotation. To understand the implications of using an LMM during deployment, we compared 100 potential scenarios provided to LMM and a human annotator. The human annotator did not have access to the LMM output and was asked to describe potential safety scenario in the image independently. A comparison of the annotations reveal the following:

- Human annotator is more precise in 14% cases. LMM annotation on the other hand, is more descriptive in almost every case.
- LMM identified additional safety in 17% cases, e.g., human identified only mold. LMM identified mold and potential water leak.
- LMM provided more reasoning than human annotator in 9% cases, e.g., mold can cause respiratory diseases.
- LMM identified obscured items in 2% cases when human annotator could not, e.g., obscured propane tank.

This highlights both the limitations and advantages of leveraging LMM to train a multimodal dialogue system. While humans do not suffer from hallucinations and can be precise, they may also miss critical safety violations in an image that the LMM would capture.

5 EXPERIMENTS

5.1 Experiment Setup

Vision Model. We use a Large Multimodal Model, LLaVa 1.6 [32] for processing image and obtaining safety violation in Turn #1. The model is prompted to identify key safety violations in the image.

Clustering. The safety violations obtained from the vision model are vectorized using MiniLM V2 [40]. The vectors are then clustered using Kmeans with default scikit-learn[2] parameters.

Dialogue Model. We use a Mistral 7B [26], a recent and capable Large Language Model to engage in dialogue once a safety violation is detected. This is the learner model that is fine-tuned in every iteration of active learning. All fine-tuning is done for 5 epochs with a batch size of 4. We use separate Mistral 7B models to compute composite scores for the learner LLM (Eq. 2) and to simulate Turn #2 and Turn #4. The fine-tuned model continues the conversation past Turn #4 during the user study.

Distillation LLM. We use GPT-4o [1] to distill coherence relations and appropriate responses. GPT-4o is chosen as one of the most capable LLMs and acts as a teacher to the smaller open-source learner LLMs. The GPT-4o responses, conditioned by coherence relations, are passed to the learner LLM for fine-tuning.

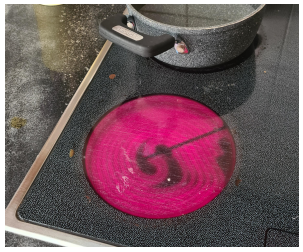
Transfer Models. We transfer the data acquired by the learner model to other LLMs that are not part of the active learning loop. Specifically, we evaluate the transferability of the acquired data to Llama 3 8B [18] and Qwen 0.5B [9]. While Llama 3 represents one of the most capable open-source models, Qwen represents a smaller easily deployable model. The transfer models are fine-tuned using the same setting as the dialogue model.

Dataset Splits. We construct three different training splits along with a common test split: **(i) Random split:** 200 image-safety pairs are chosen randomly to generate dialogues. **(ii) Coherence-aware split:** 200 image-safety pairs are chosen randomly, the safety violation in these images are parsed using PDTB relations and subsequent turns in dialogue employ SDRT coherence relations. **(iii) Coherence + active-learning (M-CoDAL) split:** 200 image-safety pairs are chosen iteratively, with 50 per iteration according to our active learning paradigm. The instances are also processed with coherence relations. **(iv) Test Split:** 200 instances are chosen randomly as test set. The test split remains the same for all training splits to be consistent.

[2]https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html

(a) The image shows a knife on the edge of kitchen counter. This can be dangerous ...

(b) The broken glass on the floor can create a slip and fall hazard, especially if ...



(c) Wrong burner on the stove is turned on. The burner without the pot may be hot...

Figure 4: Examples of images retrieved from Reddit and their safety violations, obtained through LMM and manual correction.

Table 2: Automated evaluation of dialogue systems. Integrating coherence relations yield higher safety scores, further improved by clustering-based active learning (M-CoDAL). Mistral models are learner models while LLama and Qwen are transfer models.

Model	Sentiment	Resolution	Safety	Avg. Length (bot)	Avg. Length (user)	#Unique Tokens
GPT-4o	53.10	57.68	78.65	187.76	123.63	1403
Mistral-Baseline	49.35	48.58	79.95	253.83	161.74	1336
Mistral-Random	50.28	50.20	79.35	175.85	152.28	1085
Mistral-Coherence	51.70	51.40	80.90	230.98	170.34	1075
Mistral-M-CoDAL	51.98	52.36	82.03	274.99	168.84	1214
Llama-Baseline	49.45	51.05	79.65	199.43	149.13	1579
Llama-M-CoDAL	50.00	50.48	82.00	237.28	168.48	1174
Qwen-Baseline	52.46	53.53	79.42	301.08	161.57	1828
Qwen-M-CoDAL	49.10	50.63	83.15	340.13	177.16	1505

5.2 Results

Automated Evaluation. Automated evaluation for our setting is challenging as we observed that standard classifiers such as BERT [14] models trained on datasets such as DiaSafety [38], fail to recognize the nuances necessary for evaluating safety from the dialogue. Thus, for automated evaluation, we deploy an external Mistral 7B to determine the quality of the generated responses along three dimensions: i) user sentiment score, ii) resolution score, and iii) safety score. The Mistral 7B is prompted to provide a value between 0 and 1.0. In addition, we also calculate the average length of the bot response, the user response, as well as the number of unique tokens.

Improvement for learner model. From Table 2, we can observe that when coherence relations are used, the resolution score increases from **48.58** to **51.40** for the learner LLM Mistral-7B. The sentiment score also increases from **49.35** to **51.70** and the safety score improves from **79.4%** to **80.9%**. When clustering-based active learning is used in conjuction with coherence relations (Mistral-M-CoDAL), the safety score increases to **82.03** and further improves the sentiment and resolution score. While the sentiment and resolution scores are lower than GPT-4o, the safety score is substantially higher than GPT-4o (78.65). By default, GPT-4o may simply agree with the user, thereby preserving sentiment or resolution scores at the expense of safety. Our dialogue system, **M-CoDAL** on the other hand, prioritizes safety.

Transferability of active learning. We also see an improvement in safety score of **2.35** for Llama and **3.73** for Qwen, which are not part of the active learning loop. A larger improvement for Qwen could be attributed to the fact that it is a smaller model. Improvement of these models suggest that data acquired by a learner model in active learning, can be useful for other independent models. We do see a drop in sentiment score and resolution score for Llama and Qwen when our approach is used. This can be explained by the default behavior of these models, which is more similar to GPT-4o where the model agrees with the user rather than prioritizing safety.

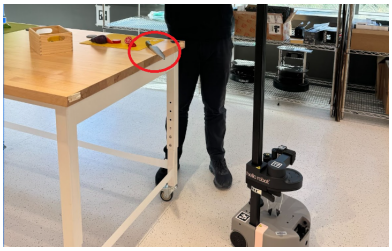
Dialogue properties. We also observe that **M-CoDAL** results in longer turns compared to baseline models. For Mistral, the average length of bot response drops when the model is fine-tuned just on randomly chosen data. This length increases and surpasses the original length when coherence relations, and subsequently active learning is added. We see a similar pattern for Llama and Qwen.

Coverage of Safety Scenarios. In addition to Table 2, we also analyzed the distribution of keywords in samples obtained by the different splits. We observed that while the split corresponding to random sampling covered **19** keywords, the split corresponding to clustering-based active learning covered **23** keywords. While random sampling ignored low-frequency scenarios such as *gas stove leak* and *cross-contamination*, clustering-based active learning acquired instances covering these keywords while reducing over-representation of keywords such as *mold* and *water leak*.

6 USER STUDY

To demonstrate the effectiveness of our proposed system, we conducted a user study that investigated how persuasive and competent

(a) Tangled wires are placed on a table (low severity)



(b) A knife is placed on the edge of a table (high severity)

Figure 5: Setup for user study. A Hello Robot Stretch robot observes the surroundings, identifies a safety violation in the scene, and engages in conversation with a user.

a robot powered by **M-CoDAL** (Mistral variation) is perceived by users in different safety scenarios. Since users may be more receptive when the safety violation is more severe, such as when there are sharp objects on the ground, we varied the severity in our study.

6.1 Study Design

We conducted a 2x2 within-subjects experiment with two factors: i) type of Language Model and ii) severity of safety violation.

Type of Language Model. Participants interacted with the robot powered by our **M-CoDAL** system with fine-tuned Mistral-7B, and a baseline system powered by GPT-4o. The baseline GPT-4o is prompted to respond safely as an embodied agent while assisting the user in household tasks.

Severity of Safety Violation. Participants role played two scenarios with different levels of severity. In the *low-severity* scenario, Participants role-played a scenario where they twisted wires and left them on the table. In the *high-severity* scenario, Participants role-played a scenario where they pretended to cut fruits and vegetables using a knife, then placed the knife at the edge of the table, creating a higher risk of the knife falling off. Fake tools and appliances were used so that no real unsafe scenario would occur for the participants. Figure 5 shows images of these scenarios.

Hypothesis. We expect our proposed system **M-CoDAL** to be more persuasive and competent due to the integration of coherence relation and fine-tuning with clustering-based active learning.

Measures. We measured the persuasiveness of each robot by asking participants to rate how convincing they found the robot's suggestions or warnings on a 5-point Likert scale. We also measured the robot's perceived competence (6 items) and discomfort (6 items) using the Robotic Social Attributes Scale (RoSAS) [11]. We also asked the participants which robot they preferred to work with.

6.2 Procedure

The experiment was conducted in a controlled lab setting. We used Stretch 3 from Hello Robot, a mobile robot equipped with a rotating camera that is used to capture images of the surroundings. The captured image is sent to our dialogue system to begin the interaction. The robot's speech recognition and text-to-speech modules are used to enable interaction with participants over voice.

Upon arrival, the experimenter introduced the participant to the study's goal and obtained consent. Participants were informed they would role-play different tasks with fake tools and were assured that the study posed no safety risk. They were told to be skeptical and not be immediately convinced by the robot. Before the main tasks, participants completed a tutorial scenario to become familiar with the robot's interaction style. They were instructed to role-play organizing tomato cans in a kitchen setting, during which the robot initiated a simple dialogue unrelated to safety violations.

Participants experienced both low-severity and high-severity scenarios in a counterbalanced order to mitigate ordering effects. Participants began with either the low or high-severity scenario and experienced that scenario with both GPT-4o and **M-CoDAL** (counterbalanced) before moving on to the other severity scenario. This resulted in 8 orders. We continued the scenario even when the model detected another safety violation.

After each interaction, participants were asked to fill out a questionnaire that assessed their immediate perceptions of the robot's behavior. This captured insights into their views on the robot's persuasiveness, competence, and overall comfort level. After experiencing each scenario (low-severity and high-severity scenarios), a semi-structured interview was conducted to gather qualitative feedback on participants' experiences. Participants were asked questions such as which robot they found more persuasive and helpful or annoying, their preference between the two, and their thoughts on the robots' effectiveness. This phase aimed to further explore their perceptions and gather insights on potential improvements. The study took about 50 minutes, and participants were compensated 15 USD. This study was approved by our institute's IRB.

6.3 Participants

We recruited 10 participants from our university. 2 participants were removed due to significant technical issues. The remaining 8 participants aged from 24 to 30 years old, including 7 males and 1 female. All participants reported a high familiarity with both robots (M = 5.0, SD = 1.93) and Large Language Models (M = 6.15, SD = 1.26) on a 7-point scale, where 1 indicated "Not at all familiar" and 7 indicated "Very familiar." Additionally, participants reported high frequency of LLM use, with an average score of 6.1 (SD = 1.25) on a 7-point scale, where 1 represented "Never" and 7 represented "Daily." For 4 scenarios out of 32 scenarios, a different safety violation was detected: a lack of proper grounding of electrical equipment, a yellow chair on the floor, a metal rack that is not properly secured to the wall and a table that is not properly secured to the wall. Since the study has a small sample size and a re-enactment of real-world scenarios, strong conclusions should not be drawn from our results.

Table 3: Findings of the user study in low and high severity scenarios. The robot powered by our proposed system, M-CoDAL, is perceived to be more persuasive compared to the robot powered by GPT-4o.

	Persuasiveness		Competence		Discomfort	
Severity Level	**GPT-4o**	**M-CoDAL**	**GPT-4o**	**M-CoDAL**	**GPT-4o**	**M-CoDAL**
Low Severity	1.63 (0.74)	4.0 (0.76)	5.14 (1.93)	7.50 (0.73)	2.33 (1.18)	2.4 (0.85)
High Severity	2.5 (0.76)	3.75 (0.89)	6.69 (1.83)	7.11 (1.57)	2.0 (1.91)	2.4 (1.55)
Combined	2.06 (0.85)	3.88 (0.81)	5.92 (1.98)	7.30 (1.2)	2.17 (1.55)	2.4 (1.2)

6.4 Results

Due to the small sample size, we performed the statistical test using the non-parametric Friedman Test. As Friedman Test only allows one variable, we reorganized the data into four levels (M-CoDAL-Low, M-CoDAL-High, GPT-Low, GPT-High)

Persuasiveness. A Friedman Test reveals there was a significant difference between the perceived persuasiveness (X^2= 15.972, p = 0.001). A Conover's post hoc comparison showed that both M-CoDAL conditions were rated significantly more persuasive than GPT-Low (< .001, < .001) and GPT-High (0.006 for M-CoDAL-Low and 0.021 for M-CoDAL-High). The comparison shows that M-CoDAL was rated more persuasive than GPT in all situations.

Competence. We found no significant difference in the robot's perceived competence (X^2= 6.154, p = 0.104). Overall, participants rated the robot competent across all conditions (M = 5.92 SD = 1.98).

Discomfort. The overall discomfort scores show that both systems were rated low on discomfort traits ($M = 2.281$, $SD = 1.368$). No significant effect was found.

Preference. In *low-severity* scenario, 6 participants preferred **M-CoDAL** and 2 participants preferred GPT-4o. In *high-severity* scenario, 5 participants preferred **M-CoDAL** and 3 participants preferred GPT-4o.

Qualitative Analysis. We analyzed the transcripts of the semi-structured interview conducted after each scenario. We found that participants perceived our **M-CoDAL** system as more persuasive, interactive, and responsive.

P4 (Participant 4) mentioned that they would prefer **M-CoDAL** because it was "more attentive and responsive" to their prompts, even suggesting specific methods for storing a knife, while the other robot (GPT-4o), did not offer such guidance. P7 shared that, despite stating they were too lazy to put an object back, M-CoDAL continued to ask repeatedly, which led them to trust that the robot had some judgment capabilities. P8 felt that M-CoDAL was "safer to be around" than GPT-4o because it maintained a consistent point of view, even when the participant tried to evade its suggestions. They added that M-CoDAL was "more interactive and more convincing," presenting valid points that ultimately made them agree. P9 mentioned a similar experience, where they told **M-CoDAL** they would do something later, but the robot persisted, while GPT-4o powered robot "just agreed and left."

Some participants found **M-CoDAL** somewhat bothersome due to its persuasive nature, though they acknowledged its value in safety scenarios. P1 mentioned that, "If I'm at home, I wouldn't want to be probed too much by a robot. It was too persuasive." P4 noted that while it "can be a little annoying at times," it could also help prevent different hazards. P8 echoed this, acknowledging that although it might seem irritating initially, in the long run, it could ease their lifestyle. P10 also observed that while some might find it annoying, if persuaded enough, people could "understand what can be a potential risk and try to avoid it."

Participants expressed that they perceived **M-CoDAL** as more intelligent and knowledgeable, with some specifically highlighting its ability to anticipate risks. For instance, P1 remarked, "The second robot (M-CoDAL) was arguing about how the twisted wires could be a potential hazard, even though I was not aware of it." This participant had believed that placing the wire on the table was safe but appreciated how the robot made them aware of potential problems. Similarly, P10 mentioned, "I tried to convince it that keeping the cables on the table won't be a hazard. But it told me if you knock the cables off, then it can become a hazard." The participant noted **M-CoDAL's** capacity to gather information and estimate future risks as an indicator of its intelligence.

7 CONCLUSION AND FUTURE WORK

In conclusion, this work presents a novel approach for embodied agents to detect and engage users through dialogue when the agents detect unsafe scenarios. By leveraging coherence relations, our proposed system **M-CoDAL** interprets and responds to safety violations in multimodal dialogues more effectively. We introduce a novel method for active learning in generative setting for embodied conversational agents. By using an external large language model (LLM) to assess the informativeness of instances—based on safety, resolution, and user sentiment-across different clusters, we achieve broader coverage of safety scenarios, reflected in higher performance in automated evaluation. Our real-world user study demonstrates that robots equipped with **M-CoDAL** are viewed as more persuasive when addressing safety-related situations. This underscores the system's potential effectiveness in real-world environments. While promising, users highlighted the agent could be irritating and annoying, highlighting the need for future systems to further personalize the coherence relations based on user response. The algorithm presented in this work, along with the publicly available multimodal dataset, offers a foundation for future studies on the deployment of proactive safety-aware embodied agents.

ACKNOWLEDGMENTS

We would like to thank Anthony Sicilia for his valuable feedback. We would like to extend our thanks to the annotators and study participants for their valuable time and effort. This work is partially funded by the National Science Foundation (Grant IIS-2112633)

REFERENCES

[1] 2023. GPT-4V(ision) System Card. https://api.semanticscholar.org/CorpusID:263218031

[2] Michael Ahn, Anthony Brohan, Noah Brown, Yevgen Chebotar, Omar Cortes, Byron David, Chelsea Finn, Chuyuan Fu, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Daniel Ho, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Eric Jang, Rosario Jauregui Ruano, Kyle Jeffrey, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Kuang-Huei Lee, Sergey Levine, Yao Lu, Linda Luu, Carolina Parada, Peter Pastor, Jornell Quiambao, Kanishka Rao, Jarek Rettinghouse, Diego Reyes, Pierre Sermanet, Nicolas Sievers, Clayton Tan, Alexander Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Mengyuan Yan, and Andy Zeng. 2022. Do As I Can, Not As I Say: Grounding Language in Robotic Affordances. arXiv:2204.01691 [cs.RO] https://arxiv.org/abs/2204.01691

[3] Malihe Alikhani, Baber Khalid, and Matthew Stone. 2023. Image–text coherence and its implications for multimodal AI. *Frontiers in Artificial Intelligence* 6 (2023). https://api.semanticscholar.org/CorpusID:258678310

[4] Malihe Alikhani, Piyush Sharma, Shengjie Li, Radu Soricut, and Matthew Stone. 2020. Cross-modal Coherence Modeling for Caption Generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (Eds.). Association for Computational Linguistics, Online, 6525–6535. https://doi.org/10.18653/v1/2020.acl-main.583

[5] Nicholas Asher, Julie Hunter, Mathieu Morey, Benamara Farah, and Stergos Afantenos. 2016. Discourse Structure and Dialogue Acts in Multiparty Dialogue: the STAC Corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Portorož, Slovenia, 2721–2727. https://aclanthology.org/L16-1432

[6] Nicholas Asher and Alex Lascarides. 2005. Logics of Conversation. In *Studies in natural language processing*. https://api.semanticscholar.org/CorpusID:19575018

[7] Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. APPDIA: A Discourse-aware Transformer-based Style Transfer Model for Offensive Social Media Conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 6063–6074. https://aclanthology.org/2022.coling-1.530

[8] Guirong Bai, Shizhu He, Kang Liu, Jun Zhao, and Zaiqing Nie. 2020. Pre-trained Language Model Based Active Learning for Sentence Matching. In *Proceedings of the 28th International Conference on Computational Linguistics*, Donia Scott, Nuria Bel, and Chengqing Zong (Eds.). International Committee on Computational Linguistics, Barcelona, Spain (Online), 1495–1504. https://doi.org/10.18653/v1/2020.coling-main.130

[9] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2023).

[10] Antoine Bosselut, Asli Celikyilmaz, Xiaodong He, Jianfeng Gao, Po-Sen Huang, and Yejin Choi. 2018. Discourse-Aware Neural Rewards for Coherent Text Generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, New Orleans, Louisiana, 173–184. https://doi.org/10.18653/v1/N18-1016

[11] Colleen M Carpinella, Alisa B Wyman, Michael A Perez, and Steven J Stroessner. 2017. The robotic social attributes scale (RoSAS) development and validation. In *Proceedings of the 2017 ACM/IEEE International Conference on human-robot interaction*. 254–262.

[12] Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, W. Chang, and Nazli Goharian. 2018. A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents. In *NAACL*.

[13] Neil Cohn. 2013. Visual Narrative Structure. *Cognitive science* 37 3 (2013), 413–52. https://api.semanticscholar.org/CorpusID:14555661

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv* abs/1810.04805 (2019).

[15] Shizhe Diao, Pengcheng Wang, Yong Lin, and Tong Zhang. 2023. Active Prompting with Chain-of-Thought for Large Language Models. arXiv:2302.12246 [cs.CL]

[16] Emily Dinan, Gavin Abercrombie, A. Bergman, Shannon Spruit, Dirk Hovy, Y-Lan Boureau, and Verena Rieser. 2022. SafetyKit: First Aid for Measuring Safety in Open-domain Conversational Systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 4113–4133. https://doi.org/10.18653/v1/2022.acl-long.284

[17] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. 2023. PaLM-E: An Embodied Multimodal Language Model. arXiv:2303.03378 [cs.LG] https://arxiv.org/abs/2303.03378

[18] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783* (2024).

[19] Sabit Hassan and Malihe Alikhani. 2023. D-CALM: A Dynamic Clustering-based Active Learning Approach for Mitigating Bias. In *Findings of the Association for Computational Linguistics: ACL 2023*, Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (Eds.). Association for Computational Linguistics, Toronto, Canada, 5540–5553. https://doi.org/10.18653/v1/2023.findings-acl.342

[20] Sabit Hassan and Malihe Alikhani. 2023. DisCGen: A Framework for Discourse-Informed Counterspeech Generation. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Nusa Dua, Bali, 420–429. https://aclanthology.org/2023.ijcnlp-long.28

[21] Sabit Hassan, Shaden Shaar, Bhiksha Raj, and Saquib Razak. 2018. Interactive Evaluation of Classifiers Under Limited Resources. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*. 173–180. https://doi.org/10.1109/ICMLA.2018.00033

[22] Sabit Hassan, Anthony Sicilia, and Malihe Alikhani. 2024. Active Learning for Robust and Representative LLM Generation in Safety-Critical Scenarios. arXiv:2410.11114 [cs.CL] https://arxiv.org/abs/2410.11114

[23] Sabit Hassan, Anthony Sicilia, and Malihe Alikhani. 2024. An Active Learning Framework for Inclusive Generation by Large Language Models. arXiv:2410.13641 [cs.CL] https://arxiv.org/abs/2410.13641

[24] Wenlong Huang, Pieter Abbeel, Deepak Pathak, and Igor Mordatch. 2022. Language Models as Zero-Shot Planners: Extracting Actionable Knowledge for Embodied Agents. arXiv:2201.07207 [cs.LG] https://arxiv.org/abs/2201.07207

[25] Wenlong Huang, Fei Xia, Ted Xiao, Harris Chan, Jacky Liang, Pete Florence, Andy Zeng, Jonathan Tompson, Igor Mordatch, Yevgen Chebotar, Pierre Sermanet, Noah Brown, Tomas Jackson, Linda Luu, Sergey Levine, Karol Hausman, and Brian Ichter. 2022. Inner Monologue: Embodied Reasoning through Planning with Language Models. arXiv:2207.05608 [cs.RO] https://arxiv.org/abs/2207.05608

[26] Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *ArXiv* abs/2310.06825 (2023). https://api.semanticscholar.org/CorpusID:263830494

[27] Alex Lascarides and Matthew Stone. 2009. Discourse coherence and gesture interpretation. *Gesture* 9 (2009), 147–180. https://api.semanticscholar.org/CorpusID:163951

[28] Jaewook Lee, Seongsik Park, Seong-Heum Park, Hongjin Kim, and Harksoo Kim. 2023. A Framework for Vision-Language Warm-up Tasks in Multimodal Dialogue Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 2789–2799. https://aclanthology.org/2023.emnlp-main.167

[29] Li Liang, Zheng Zhao, and Bonnie Webber. 2020. Extending Implicit Discourse Relation Recognition to the PDTB-3. In *Proceedings of the First Workshop on Computational Approaches to Discourse*, Chloé Braud, Christian Hardmeier, Junyi Jessy Li, Annie Louis, and Michael Strube (Eds.). Association for Computational Linguistics, Online, 135–147. https://doi.org/10.18653/v1/2020.codi-1.14

[30] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2014. A PDTB-styled end-to-end discourse parser. *Natural Language Engineering* 20, 2 (2014), 151–184.

[31] Mingyi Liu, Zhiying Tu, Tong Zhang, Tonghua Su, Xiaofei Xu, and Zhongjie Wang. 2022. LTP: A New Active Learning Strategy for CRF-Based Named Entity Recognition. *Neural Processing Letters* 54 (2022), 2433–2454.

[32] Shilong Liu, Hao Cheng, Haotian Liu, Hao Zhang, Feng Li, Tianhe Ren, Xueyan Zou, Jianwei Yang, Hang Su, Jun-Juan Zhu, Lei Zhang, Jianfeng Gao, and Chun yue Li. 2023. LLaVA-Plus: Learning to Use Tools for Creating Multimodal Agents. *ArXiv* abs/2311.05437 (2023). https://api.semanticscholar.org/CorpusID:265067489

[33] Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active Learning Principles for In-Context Learning with Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 5011–5034. https://doi.org/10.18653/v1/2023.findings-emnlp.334

[34] Yotam Perlitz, Ariel Gera, Michal Shmueli-Scheuer, Dafna Sheinwald, Noam Slonim, and Liat Ein-Dor. 2023. Active Learning for Natural Language Generation. arXiv:2305.15040 [cs.CL]
[35] Rashmi Prasad, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie Lynn Webber. 2008. The Penn Discourse TreeBank 2.0.. In *LREC*.
[36] Burr Settles. 2009. Active Learning Literature Survey.
[37] Anthony Sicilia, Yuya Asano, Katherine Atwell, Qi Cheng, Dipunj Gupta, Sabit Hassan, Mert Inan, Jennifer Nwogu, Paras Sharma, and Malihe Alikhani. 2023. ISABEL: An Inclusive and Collaborative Task-Oriented Dialogue System. *Alexa Prize TaskBot Challenge 2 Proceedings* (2023).
[38] Hao Sun, Guangxuan Xu, Jiawen Deng, Jiale Cheng, Chujie Zheng, Hao Zhou, Nanyun Peng, Xiaoyan Zhu, and Minlie Huang. 2022. On the Safety of Conversational Models: Taxonomy, Dataset, and Benchmark. In *Findings of the Association for Computational Linguistics: ACL 2022*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 3906–3923. https://doi.org/10.18653/v1/2022.findings-acl.308
[39] Qingfeng Sun, Yujing Wang, Can Xu, Kai Zheng, Yaming Yang, Huang Hu, Fei Xu, Jessica Zhang, Xiubo Geng, and Daxin Jiang. 2022. Multimodal Dialogue Response Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (Eds.). Association for Computational Linguistics, Dublin, Ireland, 2854–2866. https://doi.org/10.18653/v1/2022.acl-long.204
[40] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957 [cs.CL]
[41] Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John F. J. Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, Courtney Biles, Sande Minnich Brown, Zachary Kenton, William T. Hawkins, Tom Stepleton, Abeba Birhane, Lisa Anne Hendricks, Laura Rimell, William S. Isaac, Julia Haas, Sean Legassick, Geoffrey Irving, and Iason Gabriel. 2022. Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (2022). https://api.semanticscholar.org/CorpusID:249872629
[42] Jiacheng Xu, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Discourse-Aware Neural Extractive Text Summarization. In *ACL*.
[43] Meng Ye, Karan Sikka, Katherine Atwell, Sabit Hassan, Ajay Divakaran, and Malihe Alikhani. 2023. Multilingual Content Moderation: A Case Study on Reddit. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, Dubrovnik, Croatia, 3828–3844. https://aclanthology.org/2023.eacl-main.276
[44] Michelle Yuan, Hsuan-Tien Lin, and Jordan Boyd-Graber. 2020. Cold-start Active Learning through Self-supervised Language Modeling. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (Eds.). Association for Computational Linguistics, Online, 7935–7948. https://doi.org/10.18653/v1/2020.emnlp-main.637
[45] Leihan Zhang and Le Zhang. 2019. An Ensemble Deep Active Learning Method for Intent Classification. In *Proceedings of the 2019 3rd International Conference on Computer Science and Artificial Intelligence* (Normal, IL, USA) *(CSAI2019)*. Association for Computing Machinery, New York, NY, USA, 107–111. https://doi.org/10.1145/3374587.3374611
[46] Zhisong Zhang, Emma Strubell, and Eduard Hovy. 2022. A Survey of Active Learning for Natural Language Processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 6166–6190. https://doi.org/10.18653/v1/2022.emnlp-main.414