ClarVis: A Dataset of Clarification Requests and Grounding in Collaborative Data Visualization Dialogues

Abari Bhattacharya and Barbara Di Eugenio
University of Illinois Chicago
{abhatt62, bdieugen}@uic.edu

Abstract

Clarification Requests (CRs) play a crucial role in human communication. However, existing datasets are often limited to single-turn clarifications or simulated tasks. We present a novel task-oriented dialogue dataset, **ClarVis**, of CRs collected from real-time multi-user collaborative data-exploration sessions, where users analyze data together while interacting with a visualization-generating conversational assistant. This dataset fills important gaps in current CR datasets by identifying naturally occurring CRs grounded by real-world modalities like hearing, vision, and actions in the physical environment. Our dataset includes 6.3K utterances across collaborative tasks, where each CR is annotated with a grounding modality —Auditory (A), Visual (V), or Kinesthetic (K)—that captures the context the clarification pertains to. Further, we establish benchmark tasks for CR identification and grounding-modality classification, and evaluate them with traditional machine learning models as well as instruction-tuned large language models. The results highlight both the learnability and the difficulty of these tasks, and position ClarVis as a useful resource for studying clarifications in collaborative dialogue settings.

1 Introduction

The ability to ask for and respond to clarification is fundamental to dialogue. Clarification requests (CRs) help speakers repair misunderstandings, resolve uncertainty, and maintain common ground while working toward a shared goal (Clark and Brennan, 1991; Clark, 1996; Traum, 1994). Prior work has often treated CRs as repair phenomena targeting problems in hearing, reference, meaning, or action coordination (Schegloff et al., 1977; Schlangen, 2004a; Purver, 2004; Ginzburg, 2012). In collaborative settings, however, CRs do more than repair isolated misunderstandings: they help participants stay aligned as the task, shared context, and next actions evolve.

Despite sustained interest in CRs, existing resources are less suited to studying clarification in collaborative, grounded interaction. Many available datasets come from forum-based or information-seeking settings and capture useful question–answer exchanges, but they are typically asynchronous, single-turn, or not tied to a dynamically changing shared task context (Aliannejadi et al., 2019, 2021; Kumar and Black, 2020; Qu et al., 2018; Penha et al., 2019; Zamani et al., 2020). More broadly, Benotti and Blackburn (2021a) argue that collaborative grounding is central to dialogue, and that mistakes and recovery from them are a key part of how meaning is negotiated in interaction. In related work, Benotti and Blackburn (2021b) operationalize this perspective by analyzing CRs as a modality-grounded phenomenon, with failures of understanding arising in socioperceptual, auditory, visual, or kinesthetic modalities. They demonstrate the feasibility of such labeling on the SCARE corpus (Stoia et al., 2008). What remains unclear is how such CR behavior appears in naturally occurring collaborative interaction around an evolving shared task environment.

To address this gap, inspired by Benotti and Blackburn (2021b), **(i) we introduce ClarVis, a dataset in which CRs and their grounding modality are studied in two-user collaborative data-exploration dialogues.** Figure 1 shows examples of CRs and their corresponding modalities from the dataset. In these interactions, partners converse freely while performing exploratory analysis, while a conversational assistant (CA) dynamically generates data visualizations in response. In this unscripted setting, CRs are not isolated repairs, but part of how teammates negotiate understanding and coordinate actions as the shared visual state changes over time. **(ii) We also present two prediction tasks over this resource —CR identification and grounding-modality classification —and report baseline**

Proceedings of the 27th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 79–90
August 2–5, 2026.

U1: Oh, so the same time range.
U1: I think that's the...
U2: [CR—*A*] What's the same?

S: Line Chart of COVID cases vs. Date Colored by County_type
U1: Oh, the same thing.
U1: [CR—*V*] Is that June or January?

U1: See this one..this map of cardiovascular disease..
U2: Oh, I can't bring it to the front...
U2: [CR—*K*] Can I click on this? (while trying to move the generated map)

Figure 1: Examples of clarification requests from ClarVis, illustrating the three grounding modalities used in our annotation scheme: Auditory (A), Visual (V), and Kinesthetic (K). U1 and U2 are the two users collaborating on data analysis task and S is the Conversational Assistant generating the visualizations.

results from traditional and instruction-tuned language models to examine how far these distinctions are computationally recoverable from dialogue text in this setting. We release ClarVis, including the de-identified dialogue transcripts, system-generated visualization captions, CR and grounding-modality annotations, and train/validation/test splits used in the prediction tasks, at `https://github.com/uic-nlp-lab/ClarVis`.

2 Related Work

CRs play an important role in dialogue by helping speakers repair misunderstandings and maintain common ground (Schegloff et al., 1977; Clark, 1996; Ginzburg, 2012). Computational dialogue research has likewise treated CRs as part of grounding and the negotiation of shared understanding (Gabsdil, 2003; Purver et al., 2003; Schlangen, 2004b; Rieser and Moore, 2005b). Corpus-based studies also show that CRs, while important, are relatively infrequent in dialogue. For example, Purver et al. (2001) report that CRs make up just under **4%** of turns in the demographic portion of the British National Corpus (**343** cases) and just under **3%** across all domains (**374** cases). Rieser and Moore (2005a) identify **98** CRs in **2,098** turns (**4.6%**), and Rodríguez and Schlangen (2004) report **230** CRs in **3,962** turns (**5.8%**). For the SCARE corpus, Benotti and Blackburn (2021b) report a CR rate of about **6.5–6.8%**. Prior work has also shown that CRs often appear as fragments or reprises whose interpretation depends heavily on discourse context rather than explicit question forms (Purver, 2004; Fernández, 2006). This makes CR detection difficult, especially in settings where interpretation depends on a shared task context.

The work most closely related to ours is Benotti and Blackburn (2021b), who annotate the SCARE corpus (Stoia et al., 2008; Benotti, 2009) with CRs and their grounding modalities. Here, CRs are analyzed as grounded in *Socioperceptual*, *Auditory*, *Visual*, and *Kinesthetic* modalities within a situated instruction-following task. ClarVis builds on this view of CR as a grounded phenomenon, but studies it in a different setting: two-user collaborating on an exploratory data analysis with a visualization-generating assistant. Please note that only a subset of the SCARE corpus is publicly available, so we do not perform cross-dataset experiments and instead treat it as a conceptual point of reference.

Another related resource is the CoDraw-iCR corpus of Madureira and Schlangen (2023), which identifies instruction CRs in a collaborative drawing task. Compared with CoDraw, our setting is less centered on object and spatial description, but more on data visualization interpretation, variable comparison, and system-mediated analytic actions. In addition, ClarVis explicitly annotates the grounding modality of each CR as *Auditory*, *Visual*, or *Kinesthetic*, making it possible to study how different perceptual and operational sources of uncertainty shape CR behavior in collaborative tasks.

Beyond these collaborative, grounded-task resources, large-scale CR datasets have mainly been developed for information-seeking and question-answering settings. Datasets such as Qulac, ClariQ, MIMICS, ClarQ, MSDialog, and MANtIS (Aliannejadi et al., 2019, 2021; Microsoft Research, 2020; Kumar and Black, 2020; Qu et al., 2018; Penha et al., 2019) typically treat CRs as query refinement or the resolution of underspecified information needs in search, support, or forum-style interaction. These datasets have been valuable for CR detection and generation, but they are not designed around a shared, evolving task environment in which CRs are shaped by human-human or human-system coordination, interpretation, and action. As noted by Rahmani et al. (2023), this line of work still lacks datasets in which CRs are tied to a shared environment or collaborative task. ClarVis complements this by capturing naturally occurring CRs in collaborative interaction, where CRs depend not only on language, but also on the evolving state of the shared visual workspace and the users' ongoing coordination with one another and with the system.

3 Data Annotation and Description

The dialogue data used in this work were collected during the user study reported in Bhattacharya et al. (2024) (CC BY-NC 4.0). A related study also examined reference detection and resolution in the same corpus (Bhattacharya et al., 2023). Here, we use these collaborative dialogues to study CRs and their grounding. In the original study, pairs of participants performed open-ended exploratory data analysis tasks over a COVID-19 dataset (Tiwari et al., 2021) using a conversational assistant that generated visualizations from natural-language utterances. The system retrieved data from a structured database and produced maps, heat maps, bar charts, and line charts on a shared display, but it did not provide spoken or textual explanations. Participants were free to talk to one another and to the system as they explored the data, interpreted visualizations, and refined their analytic goals. An image of the study setup can be found in Bhattacharya et al. (2024). Although the dataset and interface were fixed, the interaction itself was unscripted —participants decided how to collaborate, what to ask, and how to pursue the tasks. This makes the corpus useful for studying naturally occurring clarification behavior in a collaborative setting. The study produced 29 dialogue transcripts containing more than 6.3K user utterances, which form the basis of the dataset presented here.

3.1 Annotation Guidelines and Inter-annotator Agreement

Our annotation follows the CR framework of Benotti and Blackburn (2021b), treating CRs as functional dialogue acts rather than only surface-form questions, that is, CRs may include fragments, reprises, or confirmation checks. We adopt their definition that **"a clarification request U' is grounded in modality m if it follows an utterance U and indicates lack of positive evidence of understanding U in m"**. We use three grounding modalities from their scheme —**Visual**, **Auditory**, and **Kinesthetic** —and omit the socioperceptual category, as our dialogues concern analytic collaboration with a visualization system.

In applying this framework to collaborative data visualization, we treat CRs as user-initiated attempts to repair or check grounding in the ongoing conversational and task context. Thus, the source of uncertainty may be a prior partner utterance, a system-generated visualization or an ongoing user/system action. The annotated CRs initiated by the users are directed to the other participant, to the conversational assistant, or to both. The conversational assistant itself did not generate CRs for the users during the interaction. We also do not label all information-seeking requests as CRs. Utterances that introduce a new analytic goal, request a new visualization, or plan the next step without signaling uncertainty about prior or currently available context are labeled **Rest**.

In this setting, **Auditory** CRs involve lack of understanding in the hearing or speech channel, such as mishearing, inattentiveness, or automatic speech recognition errors (e.g., "Did you hear me?"; "Can you repeat that?"). **Visual** CRs concern lack of understanding of what is shown on the screen and are grounded in the visual display or its interpretation (e.g., "What does this chart show?", "Is that all counties?", "Which bar is rural?", "Is this the most recent map?"). Finally, **Kinesthetic** CRs involve lack of understanding with respect to action, operation, system capability, or interaction flow in the task environment, including both system actions and user actions such as whether a visualization exists, whether an operation is supported, or how to interact with the interface (e.g., "Is there a visualization for poverty rate?", "Does it support filtering by county type?", "Should I click here or type it?", "Do we have that data?").

Two-stage annotation. Two annotators (graduate-level researchers with prior dialogue annotation experience) labeled the data in two sequential stages. **Stage 1 involves Clarification request identification.** Here, each utterance was labeled as **CR** or **Rest (Non-CR)** using the operational criteria above, focusing on conversational function rather than surface form alone. Next, **Stage 2 concerns Grounding modality labeling.** Utterances labeled as CR in Stage 1 were assigned a modality (**A**, **V**, or **K**) using the same scheme. Modality decisions were guided by a compact set of disambiguation cues —heard or misheard content was mapped to *A*, references to display content or visible state were mapped to *V*, and action-, operation-, or capability-related cases were mapped to *K*. When multiple cues co-occurred and no single cue was clearly primary, annotators applied the precedence rule from Benotti and Blackburn (2021b) —Kinesthetic cues take priority over Visual cues, and Visual cues take priority over Auditory cues (**K** $>$ **V** $>$ **A**). This priority rule

was applied only for genuinely tied cases, after annotators first attempted to identify the primary grounding source from context.

Inter-annotator agreement. Before the main annotation, the two annotators jointly reviewed a detailed guideline document adapted from the clarification request taxonomy of Benotti and Blackburn (2021b). They then completed a pilot round on two dialogue transcripts containing 447 utterances to calibrate decisions, identify recurring ambiguity patterns, and refine the decision rules and examples for both CR identification and modality grounding. Disagreements from this pilot phase were discussed collaboratively and used to clarify the annotation guidelines.

After this calibration stage, we randomly selected two dialogues from the 29-dialogue corpus, yielding 565 utterances. This dialogue-level sampling strategy was used to preserve contextual coherence, since the interpretation of clarification requests often depends on the surrounding dialogue. These 565 utterances were independently labeled by the annotators. An inter-annotator agreement (IAA) was computed on these independent labels —Cohen's κ was 0.89 for Stage 1 (CR vs. Rest (Non-CR)) and 0.64 for Stage 2 (Grounding Modality: A/V/K), indicating high agreement for CR identification and moderate agreement for modality assignment.

The disagreements analysis after this round of annotations helped finalize a compact set of operational decision rules for difficult cases. For Stage 1, an utterance was labeled as CR only if it expressed a clear information gap and answer-seeking intent directed to the partner or the system. Rhetorical, strategic, or self-directed planning utterances were labeled Rest (Non-CR). For Stage 2, modality decisions were resolved using the disambiguation cues described above, with the precedence ladder ($K > V > A$) applied only when multiple cues remained ambiguous. Table 1 presents examples of disagreement cases and the rule-level rationale used to resolve them. Finally, the remaining disagreements were resolved through adjudication discussion to produce the final gold labels used in the analysis. This process helped ensure that the labels grounded in the operational definitions adapted from Benotti and Blackburn (2021b) were clear, replicable, and provided a stable basis for further analysis.

Utterance (excerpt)	**Stage**	**A1 / A2**	**Decision → Final label**
Do you have a graph or a map of access to doctors and COVID-19?	CR-ID	Rest/ CR	New visualization request → Rest
What other factors do we have?	CR-ID	CR / Rest	Task planning/ Data exploration discussion → Rest
Which one do you want?	GMOD	*K* /*V*	Action coordination, priority rule applies (K>V>A)→*K*
It is like one factor, right?	GMOD	*K* /*A*	Hearing, no action focus →*A*

Table 1: Annotation disagreements between annotators A1 and A2 and resolutions across Stage-1 (CR identification) and Stage-2 (modality assignment).

3.2 Dataset Statistics

The final corpus contains **29 dialogues**. The study design included **30 dialogues** (**15 groups (of 2 users)** × **2 tasks**), but since Group 10 Task 1 was not available, **29 dialogues** were finally obtained. Across these dialogues, the corpus contains **6,927 total turns**, with an average of **220 turns per dialogue**, including **6,377 user utterances** and **550 system-generated visualizations**. Among the user utterances, **445** were annotated as Clarification Requests (CRs), accounting for **6.4%** of all turns and **7.0%** of user utterances. This proportion is consistent with prior corpus-based work showing that CRs typically form a small portion of dialogue turns as discussed in Section 2.

Figure 3 shows the distribution of grounding modalities across all group and task combinations. Of the **445 CRs**, **220 (49.3%)** were labeled as Visual, **125 (28.2%)** as Auditory, and **100 (22.5%)** as Kinesthetic. Visual CRs dominate, reflecting the assistant's purely NLU-driven design where the assistant responds only through generated visual outputs. Kinesthetic CRs, though less frequent, capture an important coordination layer—users negotiating actions, tool use, or interaction flow to maintain joint attention on the shared workspace. Auditory CRs, in contrast, mark transient communication breakdowns such as hearing repairs between the interacting users. Together, these patterns demonstrate that CRs are not an isolated phenomenon but a mechanism for achieving shared understanding

in collaborative tasks.

3.3 Quantifying context

To estimate how much prior dialogue context may be useful for interpreting the CRs, we computed a context-distance measure for the CRs identified in a sample of 440 utterances, distinct from the subsets used for IAA. The sample contained 34 annotated CRs. For each CR, we manually inspected up to ten preceding turns in the same dialogue and identified the nearest plausible prior source of uncertainty. This source could be an utterance by a *different speaker* (the co-user) that was *not itself a CR*, so that the measure approximated the earlier context giving rise to the clarification rather than another clarification in the same repair sequence. The distances were grouped by Grounding Modality —*A*, *V*, and *K* —as shown in Figure 2. Of the 34 CRs, 15 were grounded within one turn and 10 within two turns, so 73.5% fell within the previous two turns. Visually grounded CRs occasionally extended up to six turns, whereas auditory and kinesthetic CRs typically occurred within one or two turns. This suggests that short dialogue windows are generally sufficient for CR identification, with slightly larger context ranges potentially useful for visually grounded clarifications. This measure should be interpreted as a descriptive approximation rather than a definitive annotation of the source of misunderstanding.

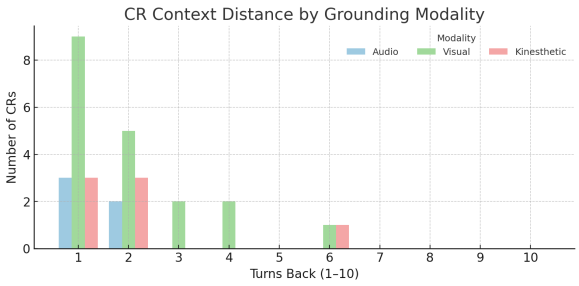

Figure 2: Grouped bar chart of context distance for CRs split by grounding modality

4 Tasks and Experimental Setup

To assess the usefulness of ClarVis for automatic modeling, we define two text classification tasks corresponding to its two annotation layers —CR identification and grounding-modality classification. These tasks serve as proof-of-concept evaluation settings for examining what can be recovered from utterance text and limited dialogue context. In doing so, they help assess both the learnability of the annotation scheme and the potential of the dataset to support conversational systems that must detect CRs and interpret their grounding modality. The resulting performance patterns also provide an idea of which grounding modalities are recoverable from text alone and which may require richer contextual or multimodal information.

Clarification Request Identification (CR-ID). This is a binary task which predicts whether an utterance is a *CR* or *Non-CR/Rest*. Therefore, it tests whether CRs in collaborative data-visualization dialogue can be distinguished from surrounding utterances such as visualization requests, acknowledgments, system responses, and other task-oriented discussion.

Grounding Modality Classification (GMOD). Given a CR, this three-way classification predicts its primary grounding modality: *A* (Auditory/verbal), *V* (Visual/appearance), or *K* (Kinesthetic/action or interface-based). It tests whether the modality distinctions that are encoded in the annotation scheme are recoverable from text alone. Because CRs in ClarVis depend on shared visual context and user-user/user-system interaction, this task should be interpreted as a text-based approximation of a broader grounded classification problem. Further, to better understand GMOD difficulty, we evaluate it under two settings—**Gold-CR**, where modality classification is performed only on gold CRs, and **CR→GMOD**, where the system first performs CR-ID and assigns a modality label only when *CR* is predicted.

Both of these tasks are evaluated under three context settings (Section 3.3): **Base**, using only the current utterance; **+Prev1**, which prepends the immediately preceding turn; and **+Prev2**, which prepends the two preceding turns in chronological order. Context segments are joined with a newline delimiter (`\n`), allowing direct comparison of how incremental dialogue context affects CR identification and modality grounding.

4.1 Baseline Models

To provide proof-of-concept baselines for ClarVis, we evaluate a range of model families. These experiments are not intended to establish a comprehensive state-of-the-art benchmark, but to examine what aspects of the CR annotations are recoverable from text alone and to provide initial reference points for CR modeling in a collaborative task environment.

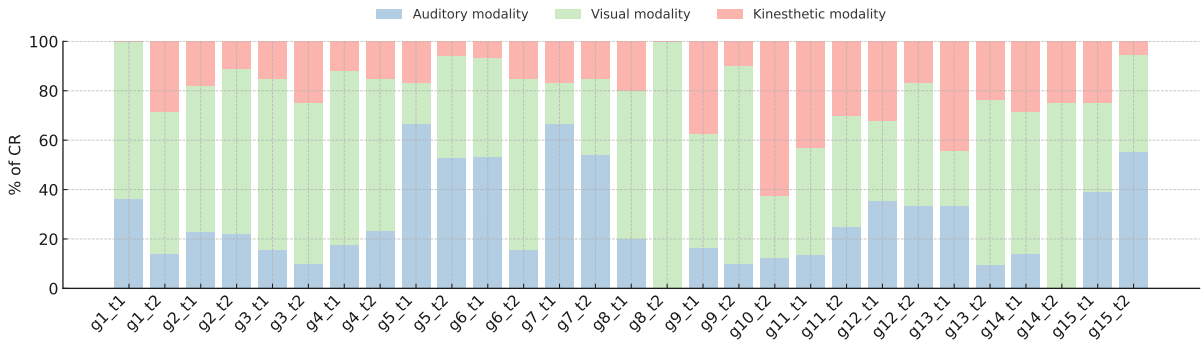

Figure 3: Distribution of grounding modalities (auditory, visual, kinesthetic) across all group–task dialogues. Bars are ordered sequentially by group and task.

The **Majority baseline** predicts the most frequent class in the training set, while the **Random** baseline samples labels according to the empirical class distribution, providing a chance-sensitive reference under class imbalance. The **TF-IDF + Logistic Regression** baseline captures lexical and shallow syntactic regularities without contextual embeddings, and helps quantify how much signal is recoverable from surface word features alone. We also experimented with **DistilBERT** (Sanh et al., 2020), which provides context-sensitive embeddings and serves as a supervised transformer baseline. These models were fine-tuned on the training split. Moreover, we used instruction-tuned **Llama-3.2-3B-Instruct** as a lightweight LLM baseline for both tasks in zero-shot and few-shot settings. All prompts were derived directly from the annotation guidelines so that the task definitions presented to the model remained closely aligned with the human coding criteria. In Figure 4, we present a shortened version of the prompts; the full prompt templates are provided in Appendix A.

For the GMOD task, we used a two-step prompting strategy. In Step 1, CRs grounded in *K* (procedure- or capability-related CR) are distinguished from non-action cases. If non-action is predicted, then in Step 2 the remaining CRs are classified as *V* (visual/display-oriented) or *A* (hearing- or wording-oriented). This hierarchical design follows the functional ordering described by Benotti and Blackburn (2021b), where action-oriented CR (*K*) takes precedence over visual (*V*) and auditory (*A*) repair.

In the few-shot setting, each prompt was preceded by an example block containing plain (Utterance, Label) pairs sampled from the training data. For CR-ID, the prompt included 5 CR and 5 Rest examples. For GMOD, Step 1 included 4 *K* and 4 *Non-Action* examples, while Step 2 included 3 *V* and 3 *A* examples. Both zero-shot and few-shot variants were evaluated across the 3 context settings described earlier.

In addition, we applied a small set of rule-based consistency checks before invoking the LLM prompts. Strong lexical cues identified from the training data such as *"do you mean..."* or *"what does . . . mean"* for *A*, procedural expressions such as *"how do/can I"*, *"click"*, or *"filter"* for *K*, and visual terms such as *"where"*, *"which"*, or *"map"* for *V* were used to assign a modality label under the same priority ordering. Cases that did not match these cues were passed to the two-step Llama prompt. These checks were intended to reduce obvious mismatches without changing the grounding-modality definitions used in the prompts.

4.2 Implementation Details

In all experiments, dialogue-level splits were used to avoid overlap between training, validation, and test dialogues, ensuring that no conversational context, vocabulary, or participant style is leaked. The final dataset split proportions were approximately 68% **training,** 12% **validation, and** 20% **test**, yielding 29 dialogues in total.

Utterances were lowercased and stripped of extraneous punctuation and whitespace; no stemming or lemmatization was applied. Where applicable, all text inputs were tokenized using the model-specific tokenizers provided by the HuggingFace Transformers library. The TF-IDF Logistic Regression baseline was implemented with scikit-learn. TF-IDF features were extracted with unigrams and bigrams, term frequency scaling, and L2 normalization; Logistic Regression used an L_2 penalty with the lbfgs solver and default regularization ($C = 1.0$).

Prompt summary used in Llama-3.2-3B-Instruct experiments

1. Clarification Request Identification (CR vs Rest). Classify each utterance as **CR** or **Rest**. Label **CR** if the utterance seeks missing information, meaning, or confirmation; asks a clarification question (e.g., *what, which, where*); asks about procedure/capability/interaction (e.g., *how do I..., can it show...*); or repeats a term for repair. Label **Rest** for commands, acknowledgments, or unrelated statements. If ambiguous, choose **Rest**. Optional *Prev1*/*Prev2* may be included as read-only context.
Output: one token: **CR** or **Rest**.
Few-shot runs prepend k=5 labeled examples.

2. Grounding Modality Classification. Given a clarification request, classify its grounding modality using a two-step hierarchy with lightweight cue-based routing.
Step 1: assign **K** if the utterance expresses uncertainty about how or where to act, UI controls, procedure, or system capability (e.g., *how do I...*, *where do I...*, *can it/can you...*, *click, tap, select, apply, filter, sort, open, close, undo, redo, type, upload, download, export*); otherwise assign **Non-Action**.
Step 2: if **Non-Action**, assign *V* for uncertainty about what is visible or where something appears (e.g., *chart, view, legend, axes, labels, highlight, filter state, this, that, which, where, show, see*) and assign **A** for hearing, wording, or meaning uncertainty (e.g., *misheard, definition, do you mean...?, echo repair*).
Output: Step 1 returns **K** or **Non-Action**; Step 2 returns **V** or **A**, yielding a final label in {**A**, **V**, **K**}.
Few-shot runs prepend 4 **K** and 4 **Non-Action** examples in Step 1 and 3 **A** and 3 **V** examples in Step 2.

Figure 4: Shortened prompts for CR identification and grounding modality classification with Llama-3.2-3B-Instruct. CR identification uses a binary CR-vs-Rest prompt, while grounding modality classification uses a two-step hierarchy: first identifying Kinesthetic CRs, and then distinguishing Visual from Auditory CRs for the remaining cases. The two-step modality structure follows the hierarchy presented by Benotti and Blackburn (2021b).

For supervised transformer baselines using DistilBERT (Sanh et al., 2020) we fine-tuned each model variant for three epochs using a learning rate of 2×10^{-5}, batch sizes of 16 (training) and 32 (evaluation), and AdamW optimization with linear warmup. Each run used a fixed random seed for reproducibility. No parameter freezing was applied, and early stopping was governed by validation F1. All models were trained and evaluated on a single T4 GPU in Google Colab using the HuggingFace transformer library.

All experiments involving LLMs were implemented using the open-source meta-Llama/Llama-3.2-3B-Instruct model via HuggingFace (https://huggingface.co/meta-Llama/Llama-3.2-3B-Instruct) in evaluation mode without fine-tuning. Inference was run on Google Colab (A100 GPU) using deterministic greedy decoding with $temperature = 0.0, top\_p = 1.0, max\_new\_tokens = 8$, and EOS tokens aligned to PAD. Outputs were normalized by regex-matching the first valid token in {CR, Rest} or {K, *V*, *A*}.

5 Results and Error Analysis

We evaluated all models on CR-ID and GMOD tasks to examine how well the annotated distinctions in ClarVis can be recovered from utterance text and limited dialogue context. Overall, the results show learnable signal for both tasks, but performance varies by task setting, context, and modality. The contrast between CR-ID and GMOD, and between Gold-CR and CR→GMOD, helps show which distinctions are more accessible from text alone and which appear to require richer contextual or grounded information.

Model/Context	CR (P/R/F1)	Rest (P/R/F1)
Majority	0.00/0.00/0.00	0.94/1.00/0.97
Random	0.02/0.03/0.02	—
TF-IDF + Logistic Regression		
No Context	0.28/0.45/0.35	0.97/0.93/0.95
+Prev1	0.19/0.33/0.24	0.96/0.91/0.93
+Prev2	0.19/0.29/0.23	0.96/0.92/0.94
DistilBERT		
No Context	0.70/0.60/0.65	0.98/0.98/0.98
+Prev1	**0.75 / 0.59 / 0.66**	0.98/0.99/0.98
+Prev2	0.66/0.56/0.61	0.97/0.98/0.98
Llama-3.2-3B-Instruct — Zero-shot		
No Context	0.17 / 0.95 / 0.29	0.99 / 0.72 / 0.83
+Prev1	*0.22 / 0.82 / 0.35*	0.99 / 0.83 / 0.90
+Prev2	0.20 / 0.78 / 0.32	0.98 / 0.81 / 0.89
Llama-3.2-3B-Instruct — Few-shot		
No Context	0.09 / 0.97 / 0.17	1.00 / 0.42 / 0.59
+Prev1	0.08 / 1.00 / 0.15	1.00 / 0.31 / 0.48
+Prev2	0.09 / 0.97 / 0.16	1.00 / 0.37 / 0.53

Table 2: **CR-ID results with per-class Macro Precision/Recall/F1**. The highest overall score is highlighted in bold.

Table 2 summarizes the CR-ID results. CRs are distinguishable from surrounding dialogue, although the task remains challenging. DistilBERT yields the most balanced performance, with the best CR F1 (0.66) under +Prev1 context, consistent with our earlier finding (3.3) that many CRs depend on the immediately preceding turn. By contrast, Logistic Regression does not benefit from added context,

Model/Context	Overall P/R/F1	A_{F1}	V_{F1}	K_{F1}
TF-IDF + Logistic Regression				
Gold-CR — No Context	0.54 / 0.52 / 0.52	0.30	0.73	**0.55**
+Prev1	0.50 / 0.46 / 0.45	0.26	0.71	0.40
+Prev2	0.54 / 0.50 / 0.48	0.34	**0.79**	0.31
CR→GMOD — No Context	0.50 / 0.13 / 0.21	0.21	0.35	**0.27**
+Prev1	0.53 / 0.11 / 0.17	0.10	0.43	0.16
+Prev2	0.24 / 0.07 / 0.11	0.10	0.32	0.00
DistilBERT				
Gold-CR — No Context	0.34 / 0.42 / 0.37	**0.43**	0.67	0.00
+Prev1	0.25 / 0.35 / 0.28	0.17	0.66	0.00
+Prev2	0.28 / 0.35 / 0.26	0.11	0.68	0.00
CR→GMOD — No Context	0.35 / 0.09 / 0.15	**0.30**	0.28	0.00
+Prev1	0.42 / 0.09 / 0.14	0.22	0.35	0.00
+Prev2	0.28 / 0.07 / 0.10	0.11	0.31	0.00
Llama-3.2-3B-Instruct — Zero-shot				
Gold-CR (No Context)	0.40 / 0.39 / 0.39	0.28	0.59	0.31
+Prev1	0.39 / 0.38 / 0.38	0.28	0.58	0.29
+Prev2	0.40 / 0.40 / 0.40	0.28	0.58	0.33
CR→GMOD (No Context)	0.55 / 0.46 / 0.50	0.28	0.61	0.26
+Prev1	0.55 / 0.46 / 0.50	0.29	0.56	0.26
+Prev2	0.57 / 0.45 / 0.50	**0.30**	0.56	**0.27**
Llama-3.2-3B-Instruct — Few-shot				
Gold-CR (No Context)	0.41 / 0.41 / 0.40	0.29	0.60	0.31
+Prev1	0.36 / 0.36 / 0.36	0.26	0.54	0.28
+Prev2	0.37 / 0.36 / 0.36	0.26	0.53	0.30
CR→GMOD (No Context)	0.55 / 0.47 / 0.50	**0.29**	**0.63**	0.26
+Prev1	0.54 / 0.44 / 0.48	0.27	0.50	0.25
+Prev2	0.55 / 0.43 / 0.47	0.28	0.48	0.25

Table 3: **GMOD results.** Gold-CR evaluates modality prediction on gold clarification requests only; CR→GMOD evaluates a two-stage pipeline where modality prediction depends on correct CR identification. Overall macro Precision/Recall/F1 and per-modality F1 for *A*, *V*, and *K* are shown. Bold marks the best F1 in each column; color marks the best value within each model block by setting.

suggesting that shallow lexical features are unable to use local dialogue history.

The instruction-tuned Llama models show a different pattern—CR recall is very high, but precision remains low across zero-shot and few-shot settings. This suggests that the model tends to overpredict CRs based on surface patterns such as *'how'*, *'what'*, and *'can'*. In zero-shot runs, +Prev1 yields the best CR F1, whereas +Prev2 slightly reduces performance. Overall, the CR-ID results show that ClarVis contains usable signal for CR detection, while also illustrating the difficulty of separating CRs from other task-oriented questions using text alone.

Table 3 shows GMOD results under two settings. Under Gold-CR, where modality is predicted only for gold CRs, TF-IDF + Logistic Regression achieves the highest overall F1 (0.52), driven mainly by strong performance on the *V* modality. DistilBERT also performs well on the *V* modality, but is less stable across other modalities, particularly for *K*. The Llama models achieve lower overall Gold-CR performance, but reproduce the same broad CR grounding modality ordering.

Across models, we observe that *V* is consistently the easiest grounding modality to predict, while *A* remains the hardest. Visual CRs often contain explicit referential cues such as *map*, *chart*, and *region*, making their grounding more recoverable from text. By contrast, *A* grounded CRs are typically short and have context-poor phrasing, for example *What?* or *Sorry?*. *K* modality grounded CRs show intermediate difficulty, but vary more by model family, suggesting that action-oriented grounding is not captured equally well by lexical, contextual, and instruction-based approaches.

In the CR→GMOD setting, similar tendencies remain, but the scores must be interpreted together with CR-ID behavior. Because modality is predicted only after CR identification, the pipeline results reflect both modality difficulty and upstream CR detection behavior. Here, the Llama models obtain comparatively stronger overall scores than the non-LLM baselines, but these should be read together with their CR-ID tendency to overpredict CRs, especially in the few-shot setting. Thus, the pipeline scores reflect not only modality classification, but also different upstream selection behaviors across models.

Overall, the results support two conclusions. First, the distinctions encoded in ClarVis are at least partially recoverable from text and short dialogue context, supporting the usefulness of the dataset for proof-of-concept automatic modeling. Second, recoverability is uneven across tasks and modalities: visual grounding is more accessible from text alone, whereas other distinctions appear to depend more strongly on richer contextual or grounded information.

5.1 Error Analysis

The error patterns help explain why these tasks remain difficult under a text-only setup. Rather than indicating a flaw in the dataset, the observed mistakes reflect the inherent ambiguity of CR behavior in situated collaborative dialogue, where users often rely on shared visual context, prior ac-

tions, and highly elliptical language (Schlangen, 2004a). In this sense, the errors show where text alone is sufficient and where richer grounding is likely to be necessary. A closer inspection of the misclassifications reveals the following recurring error types.

(1) **Cross-modal ambiguity:** utterances such as "Can we see that region?" or "Should we remove this filter?" express both a visual reference and an intended action, producing *V*/K overlap. Similarly, many *K*-type CRs describe actions using visually oriented wording (e.g., "Can it show poverty by region?"), leading to frequent confusion with *V*-type CRs, particularly in zero-shot runs without context.

(2) **Vague or underspecified references:** CRs using pointing terms such as "this one?" or "where exactly?" lack sufficient visual grounding, leading models to confuse visual with auditory interpretations.

(3) **Sparse auditory cues:** *A*-type clarifications are rare and minimal ("What?", "Sorry?"), yielding unstable precision and recall across all models.

These errors point to several directions that future work could explore, including **context shaping** with structured cues about prior system actions or visualization state, **contrastive few-shot examples** for closely related cases such as "What does this map show?" [*V*] versus "Can it show poverty by region?" [*K*], and lightweight parameter-efficient adaptation such as **adapter-** or **LoRA-based tuning**. More broadly, future gains are likely to come not only from stronger text models, but from better incorporation of the grounded interaction context that gives rise to these CRs.

6 Conclusion

We introduced **ClarVis**, a corpus of collaborative data-exploration dialogues with a visualization-generating conversational assistant, in which dialogue utterances are annotated for whether they are CRs and, if so, for their grounding modality —Auditory, Visual, or Kinesthetic. Our experiments show that these annotations are sufficiently systematic to support automatic modeling from dialogue text and short context. Visual CRs are the most consistently captured, whereas Auditory and Kinesthetic cases remain more challenging. This modality-specific variation shows that ClarVis supports computational study of how CR grounding modalities are expressed in collaborative dialogue. Although ClarVis is grounded in collaborative data visualization, it also provides a useful setting for studying clarification in situated collaborative interactions. Future work can therefore use ClarVis to examine similar CR behavior in other domains and to evaluate models that incorporate richer contextual and grounded information.

Limitations

Although ClarVis captures genuine collaborative clarification behavior, it remains moderate in scale compared to large open-domain dialogue resources, containing 29 dialogues and 6.9K total turns. In addition, the dataset is drawn from a single task setting —collaborative data exploration with a fixed conversational assistant over a COVID-19 dataset —which may limit generalizability to other collaborative, task-oriented, or multimodal domains.

The baseline results should also be interpreted with these constraints in mind. The prediction tasks are text-based approximations of a more broadly grounded phenomenon, even though many CRs in ClarVis depend on shared visual context and interaction state. In addition, the Auditory and Kinesthetic modalities remain more challenging to model because they are less frequent and often signaled by subtler cues than Visual CRs. Future work can extend this setting by collecting data from additional domains, expanding modality coverage, and incorporating richer multimodal context such as speech, gesture, and visualization state.

References

Mohammad Aliannejadi, Julia Kiseleva, Aleksandr Chuklin, Jeff Dalton, and Mikhail Burtsev. 2021. Building and evaluating open-domain dialogue corpora with clarifying questions. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. ClariQ challenge/-dataset.

Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of SIGIR*, pages 475–484.

Luciana Benotti. 2009. Clarification potential of instructions. In *Proceedings of SemDial*.

Luciana Benotti and Patrick Blackburn. 2021a. Grounding as a collaborative process. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*,

pages 515–531, Online. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2021b. A recipe for annotating grounded clarifications. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4065–4077, Online. Association for Computational Linguistics.

Abari Bhattacharya, Barbara Di Eugenio, Veronica Grosso, Andrew Johnson, Roderick Tabalba, Nurit Kirshenbaum, Jason Leigh, and Moira Zellner. 2024. A conversational assistant for democratization of data visualization: A comparative study of two approaches of interaction. *Stat. Anal. Data Min.*, 17(6).

Abari Bhattacharya, Abhinav Kumar, Barbara Di Eugenio, Roderick Tabalba, Jillian Aurisano, Veronica Grosso, Andrew Johnson, Jason Leigh, and Moira Zellner. 2023. Reference resolution and new entities in exploratory data visualization: From controlled to unconstrained interactions with a conversational assistant. In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 370–380, Prague, Czechia. Association for Computational Linguistics.

Herbert H. Clark. 1996. *Using Language*. "Using" Linguistic Books. Cambridge University Press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In Lauren Resnick, Levine B., M. John, Stephanie Teasley, and D., editors, *Perspectives on Socially Shared Cognition*, pages 127–149. American Psychological Association.

Raquel Fernández. 2006. Clarification requests in dialogue: Data and formalisation. In *Proceedings of the 10th Workshop on the Semantics and Pragmatics of Dialogue (SemDial)*.

Malte Gabsdil. 2003. *Clarification in Spoken Dialogue Systems*. Ph.D. thesis, University of Edinburgh.

Jonathan Ginzburg. 2012. *The Interactive Stance*. Oxford Press.

Vaibhav Kumar and Alan W Black. 2020. ClarQ: A large-scale and diverse dataset for clarification question generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7296–7301, Online. Association for Computational Linguistics.

Brielen Madureira and David Schlangen. 2023. Instruction clarification requests in multimodal collaborative games. In *Proceedings of EACL*, pages 2320–2333.

Microsoft Research. 2020. Mimics dataset. https://www.microsoft.com/en-us/research/project/mimics/.

Gustavo Penha, Ayoub Bagga Balan, and Claudia Hauff. 2019. Introducing mantis: A multi-domain information seeking dialogue dataset. In *Proceedings of the 20th SIGIR Conference on Conversational Search and Information Retrieval*, pages 1–9.

Matthew Purver. 2004. The theory and use of clarification requests in dialogue. *Ph.D. thesis, King's College, University of London, Supervised by Jonathan Ginzburg*.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2001. On the means for clarification in dialogue. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*.

Matthew Purver, Jonathan Ginzburg, and Patrick Healey. 2003. On the means for clarification in dialogue. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 116–123.

Chen Qu, Liu Yang, W. Bruce Croft, Johanne R. Trippas, Yongfeng Zhang, and Minghui Qiu. 2018. Analyzing and characterizing user intent in information-seeking conversations. In *Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 989–992. Dataset: MSDialog.

Hossein A. Rahmani, Xi Wang, Yue Feng, Qiang Zhang, Emine Yilmaz, and Aldo Lipani. 2023. A survey on asking clarification questions datasets in conversational systems. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2698–2716, Toronto, Canada. Association for Computational Linguistics.

Verena Rieser and Johanna Moore. 2005a. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, pages 239–246, Ann Arbor, Michigan. Association for Computational Linguistics.

Verena Rieser and Johanna D. Moore. 2005b. Implications for generating clarification requests in task-oriented dialogues. In *Proceedings of SIGDIAL*, pages 99–106.

Kepa J. Rodríguez and David Schlangen. 2004. Form, intonation and function of clarification requests in dialogue. In *Proceedings of SemDial*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Emanuel A. Schegloff, Gail Jefferson, and Harvey Sacks. 1977. The preference for self-correction in the organization of repair in conversation. *Language*, 53(2):361–382.

David Schlangen. 2004a. Causes and strategies for requesting clarification in dialogue. In *Proceedings of the 5th SIGdial Workshop on Discourse and Dialogue at HLT-NAACL 2004*, pages 136–143, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

David Schlangen. 2004b. Causes and strategies for requesting clarification in dialogue. In *Proceedings of SIGDIAL*, pages 136–143.

Laura Stoia, Darla Magdalene Shockley, Donna K. Byron, and Eric Fosler-Lussier. 2008. SCARE: a situated corpus with annotated referring expressions. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Anuj Tiwari, Arya Dadhania, Vijay Avin, and Edson Oliveira. 2021. Using machine learning to develop a novel covid-19 vulnerability index (c19vi). *Science of The Total Environment*, 773:145650.

David R. Traum. 1994. A computational theory of grounding in natural language conversation. Technical report, USA.

Hamed Zamani, Gord Lueck, Everest Chen, Rodolfo Quispe, Flint Luu, and Nick Craswell. 2020. Mimics: A large-scale data collection for search clarification. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM)*. Collection of clarification datasets for web search.

A Llama Prompt Templates

This appendix reports the prompt templates used for the Llama-3.2-3B-Instruct experiments. The prompts below are formatting-normalized for readability, but preserve the prompt content used in the code. At runtime, placeholders such as `{utt}` were replaced with the target utterance, and optional context lines were prepended for the `+Prev1` and `+Prev2` settings. In few-shot runs, an example block was prepended before the task instruction. Model outputs were normalized by extracting the first valid label token. For CR-ID few-shot prompting, the example block contained 5 CR and 5 Rest examples. For GMOD few-shot prompting, Step 1 contained 4 *K* and 4 *Non-Action* examples, while Step 2 contained 3 *A* and 3 *V* examples. For GMOD, the prompts were used together with the lightweight cue-based routing rules described in the main text.

A.1 CR-ID Prompt

CR-ID System Prompt

You are an expert dialogue annotator. Reply with exactly one token.

CR-ID User Prompt

Classify the utterance as a Clarification Request (CR) or Non-Clarification (Rest). **Definition:** A Clarification Request (CR) explicitly seeks missing information, disambiguation, or confirmation from the dialogue partner. CRs may be full questions or short repairs (e.g., echoing a term), and may appear without a question mark. **Rules — label CR if the utterance:**

- Asks for meaning, definition, explanation, or confirmation (e.g., "is that right?", "do you mean ...?").
- Identifies or disambiguates a referent (what/which/where/this/that/"which one?").
- Requests procedure/next-step guidance or capability ("how do I...", "should I click...", "can it show ...").
- Performs repair by repeating/echoing a term to check understanding (e.g., "the rate?", "by county?").

Label Rest if the utterance:

- Is a command, instruction, or request to act (e.g., "filter by age", "click the legend", "show a bar chart").
- Is a statement, acknowledgment, or description without seeking information (e.g., "okay", "that looks fine").
- Is unrelated to the task or does not ask for information.

Ambiguity: If uncertain, choose Rest. **[Optional context lines, if available:]** Previous turn U (prev-2): `<prev2>` Previous turn U: `<prev1>` Utterance U': "`<utterance>`" Respond with exactly one token: CR or Rest.

A.2 GMOD Two-Step Prompts

GMOD Step 1 System Prompt

Reply with a single token: K or NON-ACTION.

GMOD Step 1 User Prompt: K vs Non-Action

Stage 1 — Action vs Non-Action for Clarification Request (CR). Label K if the utterance expresses uncertainty about HOW/WHERE TO ACT, UI CONTROLS, PROCEDURE, or SYSTEM CAPABILITY (e.g., how do I..., where do I..., can it/can you..., click/tap/select/apply/filter/sort/open/-close/undo/redo/type/upload/download/export). Otherwise label NON-ACTION. Utterance: "`{utt}`" Respond with exactly one token: K or NON-ACTION.

GMOD Step 2 System Prompt

Reply with a single token: V or A.

GMOD Step 2 User Prompt: V vs A

Stage 2 — Visual vs Auditory for Clarification Request (CR) (only if NON-ACTION in Stage 1). Label V for uncertainty about WHAT IS VISIBLE or WHERE something APPEARS (chart/view/legend/axes/labels/highlight/filter state; deictics this/that/which/where; show/see/zoom as display state). Label A for uncertainty about HEARING/-WORDING/MEANING (misheard, definition, 'do you mean...?', echo repair). Do NOT label A unless there is an explicit wording/hearing/meaning cue. Utterance: "`{utt}`" Respond with exactly one token: V or A.

A.3 Prompt Assembly

For few-shot runs, the following example-block format was prepended before the task instruction:

Few-Shot Example Block Format

Examples: Utterance: "`<training utterance>`" Label: `<label>` Utterance: "`<training utterance>`" Label: `<label> ...`

For context-based runs, the following lines were prepended before the task instruction when available:

Context Line Format

Previous turn U (prev-2): "`<prev2>`" Previous turn U: "`<prev1>`"