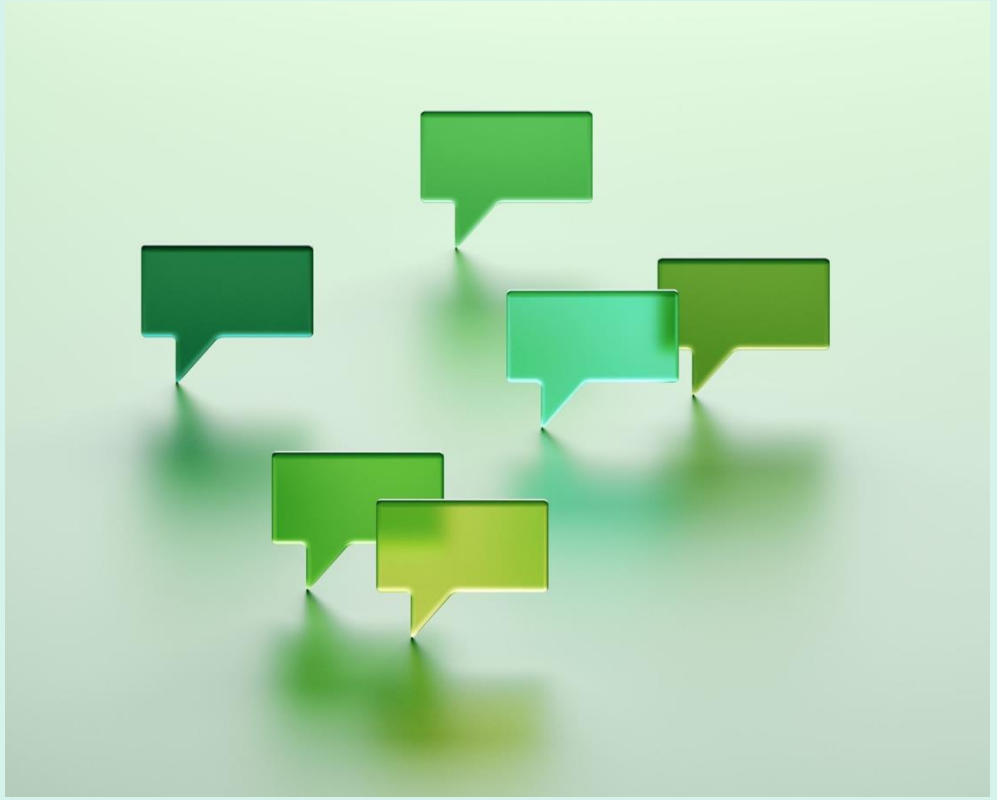
ClarVis: A Dataset of Clarification Requests and Grounding in Collaborative Data Visualization Dialogues

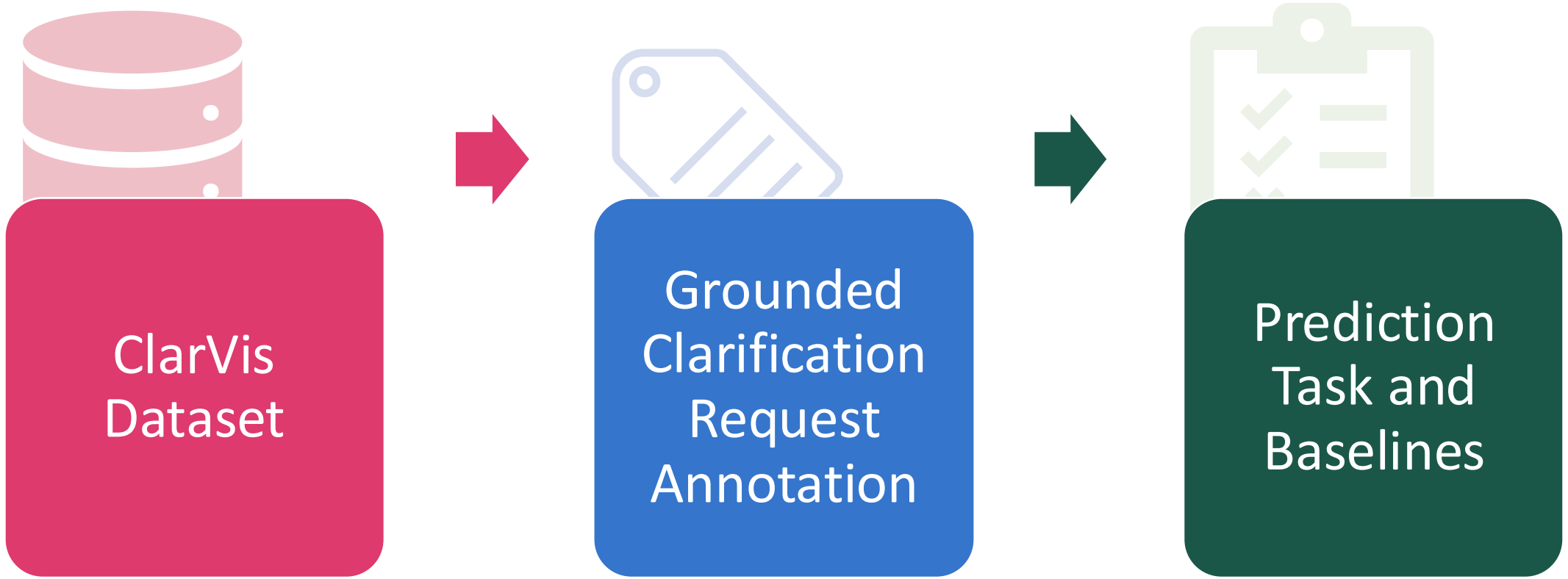
Abari Bhattacharya, Barbara Di Eugenio

University of Illinois Chicago

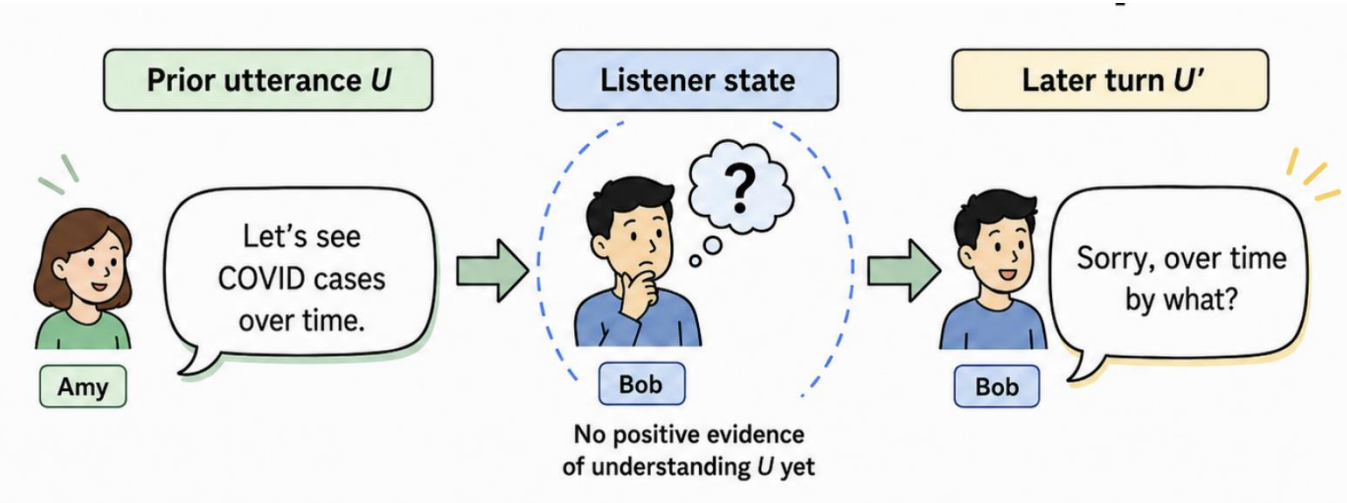
Natural Language Processing Laboratory



Roadmap


ClarVis Dataset
Grounded Clarification Request Annotation
Prediction Task and Baselines

What is a Clarification Request (CR)?



CRs help collaborators:

- repair misunderstandings
- resolve uncertainty
- maintain common ground
- stay aligned during collaborative tasks

A CR is a follow-up turn that shows something in the prior dialogue still needs to be negotiated.

When CR signals missing in understanding → source of gap?

Grounded Clarification

- Gap in understanding concerns modality *m* through which listener is trying to understand prior utterance [Benotti & Blackburn '21]

- Subsequent turn → Clarification grounded in modality m
  If it is not preceded by positive evidence of understanding the prior utterance in modality m.

- **CR grounding modality ladder:**
  - Socioperception
  - Auditory
  - Vision
  - Kinesthetic action

Their framework classifies CRs according to modality in which understanding has not yet been established.

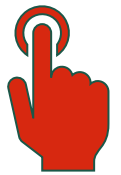
Grounding Modalities and ClarVis

ClarVis builds on Benotti & Blackburn's view of CRs as real-world modality-grounded phenomenon

- studies this in collaborative data visualization dialogues.
- uses three grounding modalities from their scheme

Auditory Modality
Lack of understanding in hearing or speech channel

Visual Modality
Lack of understanding in what can be seen in the physical environment



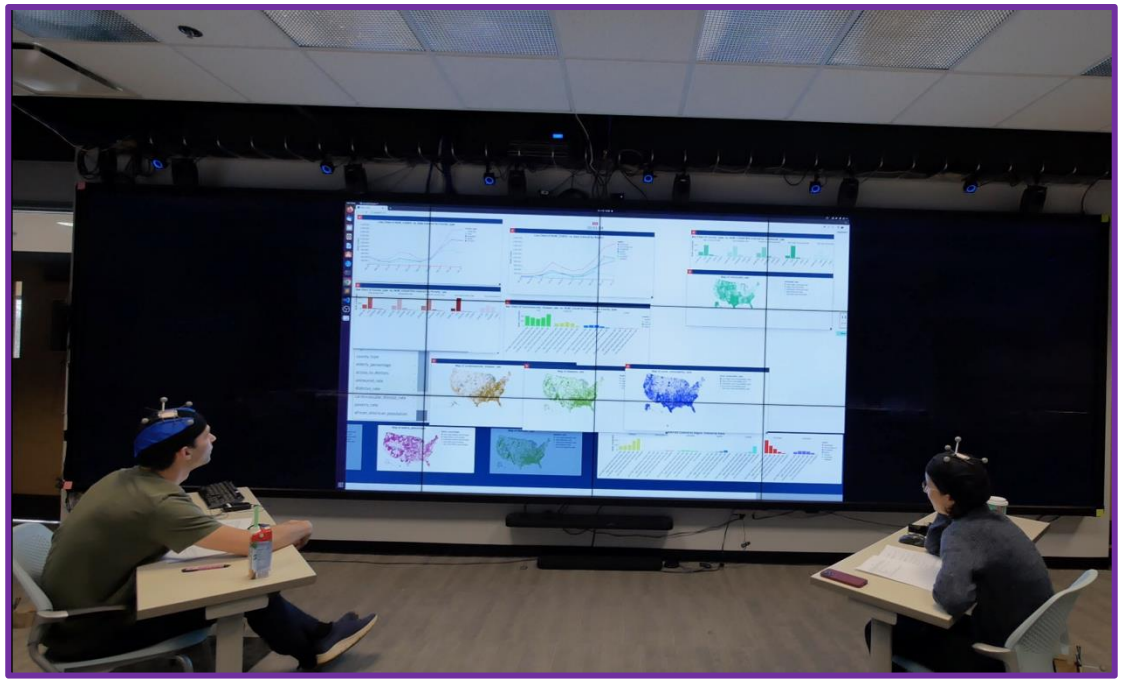
Kinesthetic Modality
Lack of understanding concerning action, operation, system capability, or interaction flow

But where do these collaborative visualization dialogues come from?

Origins of ClarVis: Source User Study [Bhattacharya et al. (2024)]

- Real-time two-user collaborative Exploratory Data Analysis (EDA) session with conversational assistant.
- Unscripted
- Task-oriented dialogues

- Pairs of users explored COVID-19 data together to complete open ended EDA tasks
- Conversational assistant generated data visualizations on large display.
- Visualizations generated: maps, heat maps, bar charts, and line charts.



- 15 groups of 2 participants per session
- 2 timed open ended EDA tasks
- 29 dialogue transcripts, 6.3K+ user utterances.

What Does Grounding Modality Look Like in Dialogue?

Auditory

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

Uncertainty about what was said or meant.

Visual

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?

Uncertainty about interpreting what is seen in the physical environment.

Kinesthetic

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)

Uncertainty about action or interaction.

ClarVis Annotation

Surface form was not enough. Annotators used dialogue function and context

Case 1
"What other factors do we have?"

Final label: Rest
Why? Task planning/Data exploration/discussion

Case 2
"Which one do you want?"

Final label: Kinesthetic
Why? Action/coordination
K/V ambiguity→ K by precedence

Precedence ladder [*] for grounding modality used only when no single cue was clearly primary

K > V > A

Action/Operation/ Capability | Display/Visible state | Hearing repair

Agreement
Sample of 565 utterances

κ = 0.89 CR vs Rest
κ = 0.64 A/V/K

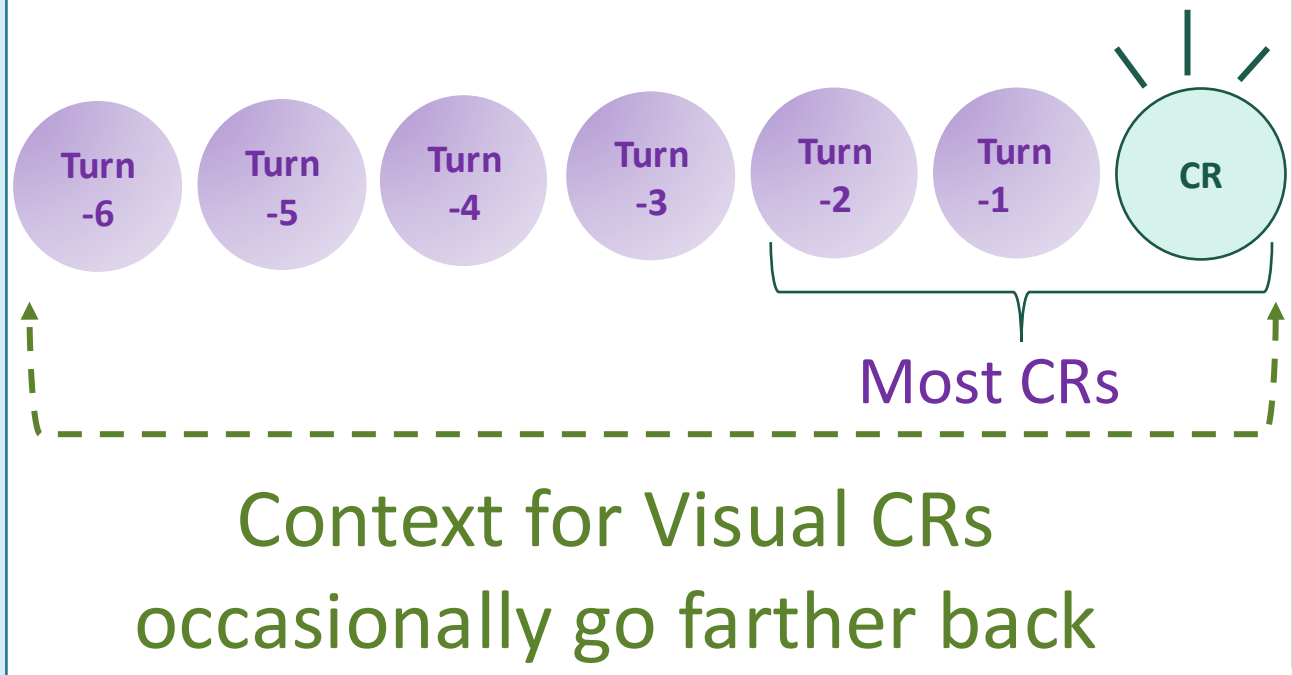
Final labels resolved through discussion

*Benotti & Blackburn (NAACL, 2021)

Most Clarifications are Grounded in Recent Context

Analysis

- 34 CRs from sample of 440 utterances
- For each CR manually inspected up to 10 preceding turns
- Identified nearest plausible prior source of uncertainty



73.5% Of sampled CRs grounded within previous 2 turns

ClarVis Dataset At A Glance

The Corpus

29 dialogues

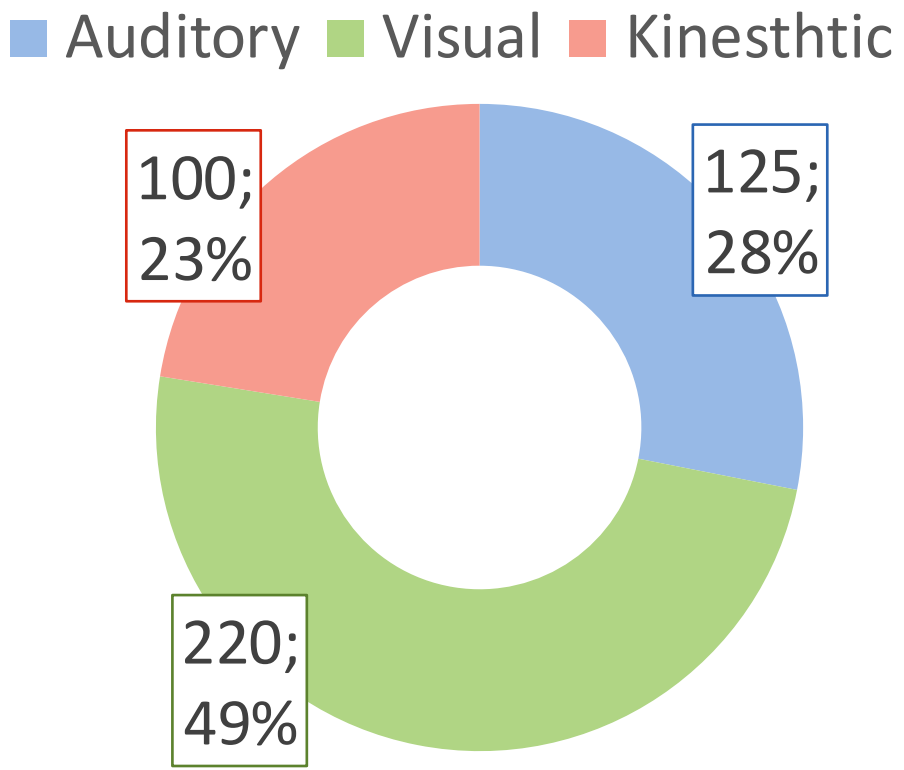
6,927 total turns (all user and system interactions)

- **6,377 user utterances**
- **550 system generated visualizations**

445 Clarification Requests (CRs)
6.4% of all turns, 7.0% of all user utterances

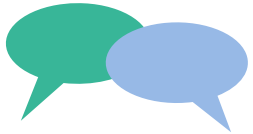
★ CRs are infrequent but **essential**

Grounding Modality Distribution (445 CRs)



Recovering Clarifications from Dialogue Text

Task 1: CR Identification (CR-ID)



Input: User utterance + optional dialogue context

Output: CR vs Rest

Baseline Models

TF-IDF + Logistic Regression

DistilBERT

Llama-3.2-3B-Instruct (Zero Shot)

Llama-3.2-3B-Instruct (Few Shot)

Context Setting

Base:
Current utterance only

+ Prev1 :
Previous turn + current utterance

+ Prev2 :
Previous two turns + current utterance

CR-ID Results

Best Results

★ Best CR-ID: DistilBERT +Prev1

Base	+Prev 1	+Prev 2
0.65	0.66	0.61

Main Takeaways

- CR-ID is learnable from dialogue text.
- One-turn context gives the strongest CR-ID result.
- Llama zero-shot showed high recall but low precision, suggesting that CR-like surface cues → easy to trigger but not enough for accurate CR detection.

Labeling Grounding Modality

Task 2: Grounding Modality (GMOD)Classification

Input: A CR+ optional dialogue context
Output: Auditory/Visual/Kinesthetic grounding modality

Evaluation Setting

Gold-CR: Classify A/V/K for gold CRs only
CR→GMOD: predict CR then classify A/V/K

Task 2: Grounding Modality (GMOD)Classification

TF-IDF + Logistic Regression
DistilBERT
Llama-3.2-3B-Instruct (Zero Shot)
Llama-3.2-3B-Instruct (Few Shot)

Context Setting

Base:
Current utterance only

+ Prev1 :
Previous turn + current utterance

+ Prev2 :
Previous two turns + current utterance

GMOD Results

GMOD: Gold-CR

★ **Best overall:** TF-IDF + Logistic Regression (**Macro F1 = 0.52** , Base)

- **Context affected modalities differently:** +Prev2 improved A and V, but K dropped.
- Visual strongest, Auditory hardest, Kinesthetic intermediate.

A	V	K
(+Prev2)	(+Prev2))	(Base)
0.34	0.79	0.55

GMOD: CR→GMOD pipeline

★ **Best overall :** Llama-3.2-3B-Instruct, zero-shot (**Macro F1 ≈ 0.50** across contexts)

- High CR recall passes many true CRs forward;
- low CR precision also sends non-CRs into A/V/K classification.
- Adding context helped A and K, while V was strongest without added context.

A	V	K
(+Prev2)	(Base)	(+Prev2)
0.30	0.61	0.27

🔍 Error Analysis

The modality-level results explained by **three recurring error patterns**

Cross-modal ambiguity

Visual reference + intended action

Example:
"Can we see that region?"

Issue:
V/K Overlap

Underspecified reference

Reference to visible content, target unclear from text alone

Example: "This one?" / "Where exactly?"

Issue: Visual-auditory confusion

Sparse auditory cues

Minimal and rarest grounding modality

Example: "What?" / "Sorry?"

Issue: little textual evidence → unstable precision/recall

Takeaways

ClarVis **captures CRs as part of collaborative interaction**:
shaped by language, the shared visual workspace, and users' ongoing coordination.

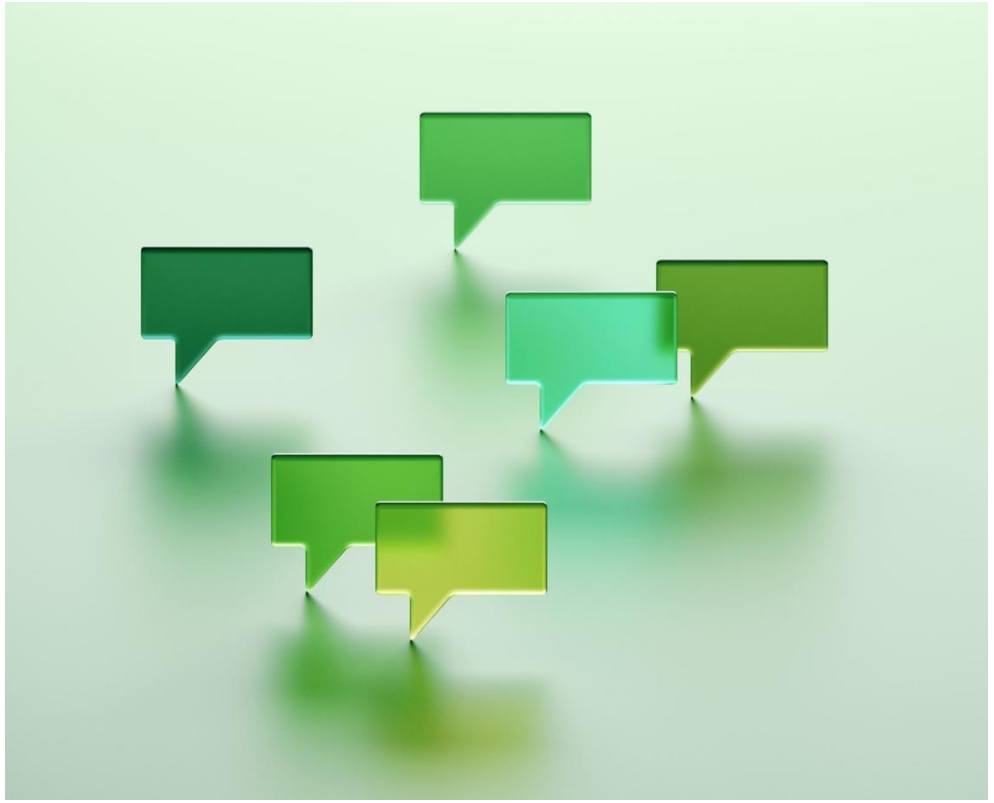
ClarVis makes grounding source of CRs explicit through A/V/K labels:
auditory uptake, **visual interpretation**, and **action/interaction flow**.

Baseline results show:

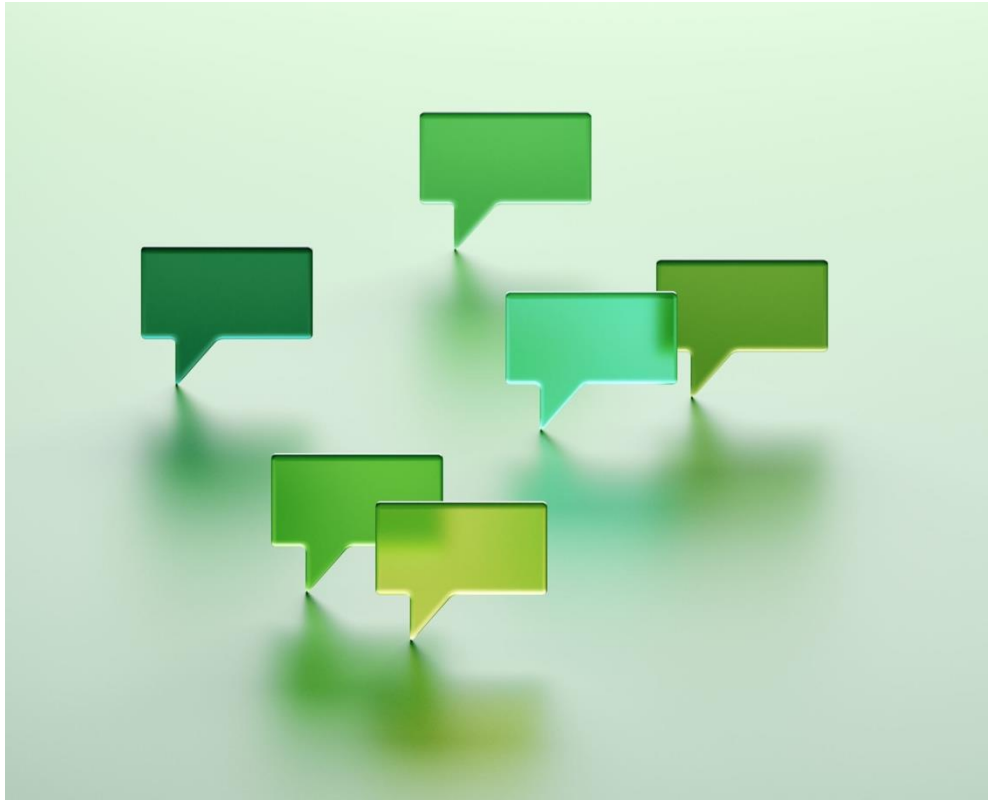
- Learnable text signals for CR-ID
- Modality-specific recoverability for GMOD
- **Grounding modality is challenging:** CRs can combine visual reference, action intent, elliptical wording, and minimal auditory repair.

Thank you!

Released with ClarVis
De-identified speaker diarized dialogue transcripts
System-generated visualization captions
CR and Grounding modality annotations
GitHub: uic-nlp-lab/ClarVis



Appendix



Clarification Requests are infrequent but essential

	Corpus / dataset	CRs	Total turns / dialogues	*CR rate*
1	British National Corpus — all domains[1]	374	11k turns 59 dialogues	*< 3%*
2	Communicator corpus[2]	98	2,098 turns 31 dialogues	*4.6%*
3	Bielefeld corpus[3]	230	3,962 turns; 22 dialogues	*5.8%*
4	SCARE corpus[4]	772	11,350 turns; 15 dialogues	*≈ 6.5–6.8%*
5	CoDraw-iCR corpus[5]	8,807 iCRs	77,502 rounds; 9,993 dialogues	*11.36%*
6	ClarVis (Ours)	445	29 Dialogues; 6,927 turns	*6.4%*

Not always explicit questions

CRs often appear as fragments or reprises → interpretation depends heavily on discourse context.[6,7]

Why this matters?

CR detection is challenging when interpretation depends on shared task context, not just explicit question form.

[1] Purver et al. (2001) ; [2] Rieser & Moore (2005) ; [3] Rodríguez & Schlangen (2004) ; [4] Benotti & Blackburn (2021) ; [5] Madureira & Schlangen (2023); [6] Purver (2004) ; [7] Fernandez(2006)

Why ClarVis?

1. Large-scale CR datasets

Qulac, ClariQ, MIMICS, ClarQ, MSDialog, MANtIS

Typically treat CRs as **query refinement or the resolution of underspecified information needs in search, support, or forum-style interaction**.

Aliannejadi et al., 2019; Aliannejadi et al., 2021; Zamani et al.; Kumar & Black, 2020; Qu et al., 2018; Penha et al., 2019.

2. Grounded collaborative CR datasets

SCARE

CRs grounded in **Socioperceptual, Auditory, Visual, and Kinesthetic** modalities in a **situated instruction-following task**.

Benotti & Blackburn, 2021

CoDraw-iCR

Instruction CRs in a **collaborative drawing task**.

Madureira & Schlangen, 2023

What is still missing?

- Dataset of **naturally occurring CRs**
- **collaborative interaction around an evolving shared task environment**,
- where clarification is shaped by **human-human and human-system coordination, interpretation, and action**.

ClarVis extends grounded CR research to collaborative data visualization, where clarification depends not only on language, but also on the evolving shared visual workspace and users' ongoing coordination.

Llama GMOD: Two-Step Classification + Cue Routing

Two-step prompting strategy
Step 1:
K vs Non-Action
Identify procedure-, capability-, or action-related CRs.
Step 2:
If Non-Action → **V vs A**
Visual/display-oriented vs hearing/wording-oriented clarification.

Before prompting: lightweight cue-based routing
Examples of training-data cues:
- **A:** "do you mean…", "what does … mean"
- **K:** "how do/can I", "click", "filter"
- **V:** "where", "which", "map"

Priority ordering:
K > V > A

Why?
These checks were used to reduce obvious cue-label mismatches without changing the grounding-modality definitions in the prompts.

Implementation Details

Dialogue-level splits

~68% train / 12% validation / 20% test

No overlap in conversational context, vocabulary, or participant style

Splits provided in GitHub Repository

A Question is not Always a Clarification Request

Surface form was not enough. Annotators used dialogue function and context

Case 1
"Do you have a graph or a map of access to doctors and COVID-19?"

Final label: Rest
Why? New visualization request

Case 2
"What other factors do we have?"

Final label: Rest
Why? Task planning/Data exploration/discussion

Case 3
"Which one do you want?"

Final label: Kinesthetic
Why? Action coordination
K/V ambiguity→ K by precedence

Agreement
Sample of 565 utterances
κ = 0.89 CR vs Rest
κ = 0.64 A/V/K

Final labels resolved through discussion

Utterances labeled as CRs when they indicated uncertainty about understanding prior or currently available context.

Precedence ladder for grounding modality used only when no single cue was clearly primary

K > V > A

Action/Operation/ Capability

Display/Visible state

Hearing repair

Baseline Experiments and Results

CR-ID. DistilBERT

★ Best CR-ID: DistilBERT +Prev1 suggesting immediately preceding turn is most useful for detecting CRs

Base	+Prev 1	+Prev 2
0.65	0.66	0.61

GMOD: Gold-CR. TF-IDF + Logistic Regression

★ Context helps V and A modalities

Context	Macro F1	A	V	K
Base	0.52	0.30	0.73	0.55
+ Prev 1	0.45	0.26	0.71	0.40
+ Prev 2	0.48	0.34	0.79	0.31

GMOD: CR→GMOD pipeline. Llama Zero Shot

★ Llama zero-shot 0.50 macro F1 across context settings.
High CR recall passes many true CRs forward
Low precision also sends non-CRs into A/V/K classification.

Context	A	V	K
Base	0.28	0.61	0.26
+ Prev 1	0.29	0.56	0.26
+ Prev 2	0.30	0.56	0.27