Can AI agents understand spoken conversations about data visualizations in online meetings?

Rizul Sharma*
DaVInCi Laboratory
University of Cincinnati

Tianyu Jiang†
University of Cincinnati

Seokki Lee‡
University of Cincinnati

Jillian Aurisano§
DaVInCi Laboratory
University of Cincinnati

ABSTRACT

In this short paper, we present work evaluating an AI agent's understanding of spoken conversations about data visualizations in an online meeting scenario. There is growing interest in the development of AI-assistants that support meetings, such as by providing assistance with tasks or summarizing a discussion. The quality of this support depends on a model that understands the conversational dialogue. To evaluate this understanding, we introduce a dual-axis testing framework for diagnosing the AI agent's comprehension of spoken conversations about data. Using this framework, we designed a series of tests to evaluate understanding of a novel corpus of 72 spoken conversational dialogues about data visualizations. We examine diverse pipelines and model architectures, LLM vs VLM, and diverse input formats for visualizations (the chart image, its underlying source code, or a hybrid of both) to see how this affects model performance on our tests. Using our evaluation methods, we found that text-only input modalities achieved the best performance (96%) in understanding discussions of visualizations in online meetings.

Index Terms: AI agents, data visualization, online meetings, multi-modal understanding, evaluation framework.

1 INTRODUCTION

Data-driven projects are often collaborative and interdisciplinary. To manage these collaborations, teams periodically meet to present preliminary results, build a common understanding of the project, integrate varied perspectives and make decisions. These meetings often occur virtually, utilizing video conferencing platforms such as Zoom or Teams. Such platforms are increasingly integrating AI-agents that utilize large-language models (LLMs) to support and summarize the discussion. For these agents to be successful, they must demonstrate a deep and accurate comprehension of what meeting participants are expressing to each other. For discussions about data, a critical challenge is ensuring that these agents understand comments about data visualizations presented in the meeting. Visualizations of data are vital tools for communicating data in meetings, providing a concrete reference point for the conversation. During a discussion, meeting participants might identify a trend or outlier in the visualization ("What is that peak in the middle?"), ask questions about data provenance ("Is this a recent sample or the entire dataset? How was it sampled?"), or express feedback about how the data is represented visually ("The colors are hard to distinguish in that cluster."). To understand these comments, AI-agents need to integrate content from multiple sources (transcript, visualization, data properties) and potentially multiple modalities (unstructured text, code, images, or a hybrid of all).

There is prior work evaluating AI agents (LLMs or VLMs) in visualization literacy benchmarks [4, 7, 12]. In addition, researchers have evaluated language models for Chart Question Answering [3]. These assess how well these models understand questions about data visualizations. However, these evaluations do not address LLM's understanding of visualizations in relation to conversational dialogues, i.e., online meeting scenarios. There has been work to evaluate different natural language techniques for conversations about data visualizations in meeting scenarios [9, 10, 14]. However, these approaches focused on classifying utterances and responding to visualization requests, instead of directly assessing language-model understanding of the discussion.

A critical challenge is determining the most effective way to provide necessary context about a data visualization to AI agents. Relying solely on a chart image risks factual errors from imperfect visual interpretation [6], while relying solely on textual context, like source code, may lead to missed visual nuances [15]. Although researchers are interested in unifying these modalities for modern AI agents, there is insufficient evidence comparing these input strategies, especially for our target context of online meetings involving data visualizations.

In this paper, we consider 1) how to measure an AI agent's capacity to demonstrate an understanding of spoken conversations about data visualizations and 2) what input modalities, pipelines, and model architectures best promote this understanding. We contribute:

- A corpus of 72 conversational discussions about data visualizations in an online meeting scenario.
- A benchmark of 318 questions about the discussions in the corpus.
- A dual-axis evaluation framework for analyzing models' performance on the above benchmark.
- Results from a four-way comparative evaluation of input formats for visualizations (Image, Code(Text), Hybrid) and model architecture (LLM vs. VLM), using our evaluation benchmark.

We found that the text-only LLM pipeline performed the best, achieving a nearly perfect accuracy of 95.9%, followed by the text-only VLM pipeline, achieving 72.4% accuracy. The hybrid pipeline performed the worst with an accuracy

*e-mail: sharmarz@mail.uc.edu
†e-mail: jiangt2@ucmail.uc.edu
‡e-mail: lee5sk@ucmail.uc.edu
§e-mail: aurisajm@ucmail.uc.edu

score of 68%, while the image-only pipeline scored 70%. We discuss the implications of these findings for the development of AI-support tools for online meetings about data.

2 Online Meeting Data Corpus

First, we needed to build a corpus of conversations about data visualizations in an online meeting scenario. This corpus is used to test AI-agent understanding.

Recruitment and Protocol: We recruited participants who had some experience working with data. These participants were recruited from the University mailing lists. This study took place over Zoom, with two participants and one researcher. The researcher showed the participants a set of slides depicting 8 data visualizations. These visualizations included bar charts, line charts, histograms, box-plots, and scatter plots. These charts visualized a movie dataset [1], which is originally sampled from the MovieLens dataset [5] utilizing the TMDB Open API. We selected this dataset because we believed it to be familiar to a general audience. During the session, the researcher introduced each visualization. Meeting participants were encouraged to discuss the data and visualization freely. Each session was recorded and transcribed using Zoom. Following the sessions, we divided each transcript into 8 sub-parts for each of the visualizations discussed.

Corpus description: We recruited 18 participants (16M, 2F) to 9 paired mock-meeting sessions. Participant ages ranged from 21 to 35 years old, from backgrounds such as computer science, business administration, biostatistics, and finance. Sessions ranged from 17 to 55 minutes, with transcripts of 2228 to 7343 words. This produced 72 (8*9) transcript-chart pairs. We conducted two pilot sessions with 4 participants to refine our study protocol.

3 Benchmark questions

Our goal is to develop a method that assesses how well AI-pipelines showcase an understanding of spoken conversational discussions about data visualizations. Two researchers iteratively designed a set of 318 evaluation questions in total for the 72 transcript-chart pairs. These questions were designed to cover significant discussion points in each conversation. They also referenced content present in the transcript and visualization. These questions were posed in a multiple-choice format to allow for automated scoring of pipeline performance, mirroring prior language-model evaluation approaches [4, 12].

4 Dual Axis Evaluation Framework

To provide a more granular score of model performance, we labeled our questions based on 1) level of complexity and 2) topic. These two labels formed our dual-axis framework for diagnosing LLM comprehension. The distribution of the 318 benchmark questions on the two axes is available in the supplementary materials (Section 9).

4.1 Level of complexity

Questions were sorted into 3 categories of increasing complexity. These labels are inspired by Bloom's Taxonomy, which is a hierarchical framework for categorizing learning objectives from simpler tasks, such as basic recall of factual information, up to more complex activities like synthesis and evaluation of information [2]. This taxonomy is often used by instructors to design assessments that measure student learning at different levels of complexity. We selected Bloom's taxonomy as a reference for our assessment because we wished to measure our pipeline's understanding in the same way you might assess the understanding of a student. Our three complexity labels are as follows: **Level 1: Factual Recall**: Does the model showcase a capacity to report the basic facts of what was said and what happened? Example questions in this category may involve asking the model to recall the insights or points directly expressed by participants in the dialogue. **Level 2: Data Interpretation**: Can the model identify the trends, patterns, and outliers in the data alluded to or implied by participants in the discussion? Example questions in this category require connecting opinions or observations to the underlying pattern in the data. **Level 2: Participant State Analysis**: Can the model go beyond the recall of statements to gauge opinions and understand reactions? Example questions in this category require inferring opinions or beliefs from participant statements. **Level 3: Causal and Process Analysis**: Can the model explain the underlying reasons for participant statements about their opinions? Example questions ask the model to identify hypotheses and theories behind participant statements.

Although Bloom's taxonomy features six levels [2], we selected three based on applicability to our tasks. Our 'Level 1: Factual recall' label directly maps to the first level, "remembering". Our 'Level 2: Data Interpretation, Participant State Analysis' labels encompass the cognitive skills of "understanding" and "applying" concepts to the data. Finally, our 'Level 3: Causal and Process Analysis' label corresponds to "analyzing", as it requires the model to deconstruct participant arguments to understand causal theories. The higher levels of the taxonomy, "evaluating" and "creating", are more aligned with generative tasks, which we will pursue in future work.

4.2 Topics

To form the second dimension of our dual-axis framework, we designed a set of topic tags to pinpoint the visualization elements or discussion points referenced in the question. The tags are designed for the questions about participant statements related to **Chart Layout:** Overall visualization, such as bars being too big or too small; **Axis & Labels:** X-axis, Y-axis, or their labels; **Visual Encoding:** How data is encoded, such as color scales; **Data Provenance:** The origin, sample size, time period, or accuracy of the data; **Data Pattern:** Identified trends, outliers, and patterns in the data; **Data Point:** Focusing on a single specific data point under discussion; **Participant Hypothesis:** A specific explanation or reason, such as a participant explaining a trend or pattern in the data.

5 Comparative Pipeline Design

We used this evaluation method to compare four different pipelines, which utilized different models and visualization input formats: (1) a Vision Language Model (VLM) which is given a visualization image as input, (2) an Large Language Model (LLM) which is given the code for the visualization, (3) a VLM given the code for the visualization (tested to isolate architecture from input format), and (4) a VLM analyzing a Hybrid of both image and text (code) for the visualization input. We selected llava-1.5-7b-hf [11] for the VLM and Mixtral-8x7B-Instruct-v0.1 [8] for the LLM, as both fit our testing requirements for the two different model architectures and allowed us to test our core questions related to image, text,

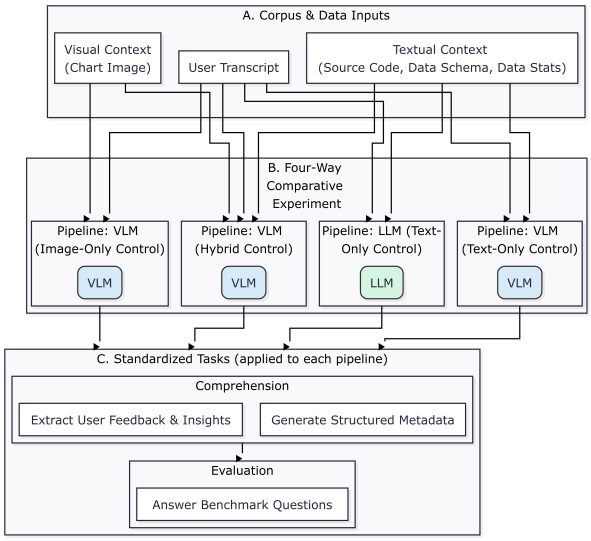

Figure 1: High-level flow for pipeline operations: (A) Inputs (B) Four Pipeline Modes (C) Standardized Comprehension Tasks

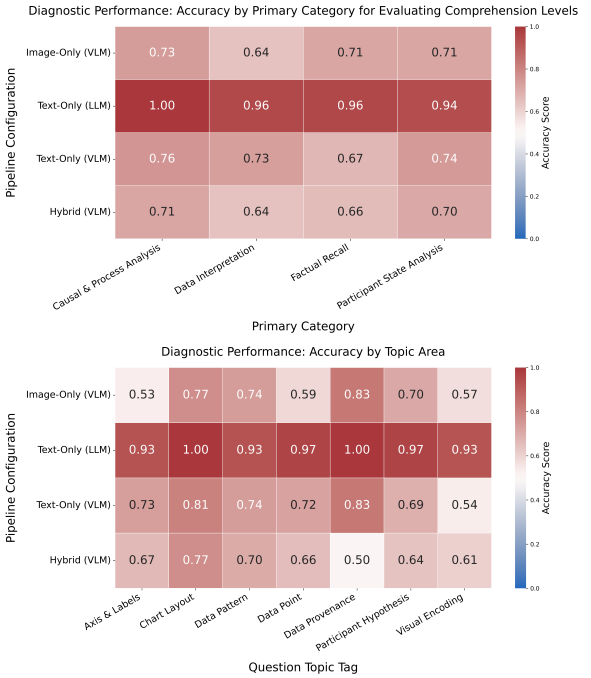

Figure 2: (top) Heatmap - Performance on Complexity Levels. (bottom) Heatmap - Performance on Topic Tags

and hybrid inputs without the confounding factor of extreme latency or memory-related failures.

For each pipeline, the AI agent performs a series of initial steps to extract information from its input. This includes extracting user insights/feedback from the transcripts, generating structured metadata from the visualization images (for the VLM-Image and VLM-Hybrid pipelines). We then pose the questions to test each pipeline's comprehension of the user discussions, as shown in Figure 1.

Analysis methods: Quantitative analysis is performed by calculating the accuracy of each pipeline's answers to the benchmark questions, broken down by complexity and topic, as discussed previously. This approach enables us to analyze specific strengths and weaknesses. For instance, we can isolate and examine all Level 2 (Interpretive) errors that fall under the 'Data Provenance' topic tag to understand why a model struggles with that specific task.

6 Findings

The Text-only LLM pipeline performed the best, achieving a nearly perfect accuracy of 95.9%, followed by the Text-only VLM pipeline, achieving 72.4% accuracy. This shows that even with the same input modality, the architecture of the pipeline, in this case, an LLM versus VLM, can significantly affect its performance. The Hybrid pipeline performed the worst with an accuracy score of 68%, while the Image-only pipeline scored 70% on our evaluation. This was an interesting observation since the Hybrid pipeline had more contextual knowledge input compared to the other pipelines.

Text-Only (LLM): The Gold Standard This pipeline is the top performer with a score greater than or equal to 90% in almost every category for every comprehension level (Figure 2 t,b). The only areas where it scored less than 90% were Axis & Labels for Data Interpretation (67%), Data Pattern for Factual Recall (81%), and Visual Encoding for Participant State Analysis (80%). These results suggest that with structured text input (source code, schema), the LLM can robustly handle all levels of comprehension, from factual recall to analysis.

Text-Only (VLM): The Architectural Contender As discussed earlier, this pipeline was designed to isolate the model architecture as a variable. Although it performs reasonably well, its accuracy of 72.4% is significantly lower and more inconsistent than the 95.9% accuracy of the Text-Only (LLM) pipeline, despite receiving the exact same text-only input. For example, it struggles with Causal & Process Analysis on Visual Encoding (20%) and general Factual Recall. This suggests that the architecture of the underlying model has a tangible impact on reasoning ability. The LLM is demonstrably better at text-based reasoning tasks than the VLM, even when the VLM is not "distracted" by an image in this case.

Hybrid (VLM): The Confused Performer This pipeline's performance indicates challenges in fusing inputs (text & image). Although it performs well on some tasks (e.g., Causal & Process Analysis on Axis & Labels and on Chart Layout), it fails on others. For example, it achieved zero accuracy in Data Interpretation on Data Provenance questions, on which the Text-Only VLM scored perfectly, which implies that adding the chart image caused a performance collapse. The model may have exhibited a knowledge conflict, ignoring the correct textual information in favor of incorrect visual information.

Image-Only (VLM): The Brittle Visionary This pipeline demonstrates the unreliability of relying solely on visual interpretation for complex analysis. The scores vary significantly. It performs perfectly on some high-level tasks (for example, Causal & Process Analysis on Chart Layout - 100%), but fails on others (for example, Causal & Process Analysis on Axis &

Labels - 0%). It struggles with tasks requiring precise data extraction, such as Data Interpretation on Data Point (40%) and Axis & Labels (33%). This suggests that while it might grasp the general layout, it cannot be trusted for detailed, factual analysis and fine-grained visual perception.

7 Discussion

Our findings suggest that AI agents perform well in our benchmark when given structured textual inputs, such as visualization code and data schemas. Given a 96% performance of the text-only LLM pipeline on our benchmark, there are strong reasons to believe that AI-agents with an LLM can be explored as support tools for meetings about data. However, these findings do not imply limiting inputs to text, but encourage researchers to explore better strategies for combining modalities.

Implications: These findings encourage us to explore AI's potential collaboration in meetings. Future studies should assess how the presence and interventions of an AI agent impact the nature of human-to-human conversation; the aim should be to make AI a helpful collaborator instead of an unwelcome intruder. Our findings also highlight a challenge for the design of AI-agents for meetings. Typical tools for online meetings, such as Slideware like PowerPoint, and video conferencing systems like Zoom & Teams, primarily have access to visualizations in the form of images, either embedded in the slide or within the video feed. However, we found that image-only information about visualizations may be insufficient for achieving model understanding of the discussion. Meeting tools may need to embed visualization code and other metadata in text form to enable effective AI support.

Limitations: Our current corpus is focused on participants with a range of data visualization experience examining charts from a tabular dataset. It is possible that we would find different performance results given a corpus of conversations between experts in various domains with spatial, temporal, or graph-structured data in different visual templates. In addition, our evaluation approach required human-generated benchmark questions, which may be difficult to extend and scale for more domains. Finally, given the strong performance of the LLM on our benchmark, future benchmarks may need to be more challenging and based on more complex and domain-specific corpora, so that we can distinguish the performance improvements of future model pipelines.

8 Conclusion

We contribute a novel corpus of 72 conversational discussions about visualizations in an online meeting setting. We developed a benchmark of 318 questions to evaluate the AI-pipelines' understanding of these discussions. We created a dual-axis framework & labeled these questions based on complexity and topic. Finally, we compared four pipelines with different model architectures (VLM vs LLM) and input modalities (text, image, and hybrid), using our benchmark and corpus. We found that text-only LLM achieved 96% accuracy on our benchmark. Future environments for online meetings may need to consider ways to integrate text-based information about visualizations to design AI-driven support tools.

9 Supplemental Materials

To promote future benchmarking and reproducibility, materials from this paper are published at GitHub [13], including: (1) code for the pipelines, (2) study visualizations, (3) the corpus of discussions, and (4) the benchmark questions and labels.

References

[1] R. Banik. The Movies Dataset. Kaggle, 2017. Version 5. Retrieved from https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset/version/5.

[2] B. S. Bloom, M. D. Engelhart, E. J. Furst, W. H. Hill, D. R. Krathwohl, et al. *Taxonomy of educational objectives: The classification of educational goals. Handbook 1: Cognitive domain*. Longman New York, 1956.

[3] Y. Dai, S. C. Han, and W. Liu. Msg-chart: Multimodal scene graph for chartqa. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, p. 3709–3713. ACM, Oct. 2024. doi: 10.1145/3627673.3679967

[4] Y. Han, C. Zhang, X. Chen, X. Yang, Z. Wang, G. Yu, B. Fu, and H. Zhang. Chartllama: A multimodal llm for chart understanding and generation, 2023.

[5] F. M. Harper and J. A. Konstan. The movielens datasets: History and context. *ACM Trans. Interact. Intell. Syst.*, 5(4), Dec. 2015. doi: 10.1145/2827872

[6] T. Hiippala. *A multimodal perspective on data visualization*, pp. 277–294. Amsterdam University Press, 2020.

[7] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do llms have visualization literacy? an evaluation on modified visualizations to test generalization in data interpretation, 2025.

[8] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts, 2024. Used mistralai/Mixtral-8x7B-Instruct-v0.1 model from: https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1.

[9] A. Kumar, B. Di Eugenio, J. Aurisano, and A. Johnson. Augmenting small data to classify contextualized dialogue acts for exploratory visualization. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 590–599, 2020.

[10] A. Kumar, B. Di Eugenio, J. Aurisano, A. Johnson, A. Alsaiari, N. Flowers, A. Gonzalez, and J. Leigh. Towards multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification. In *The 21st Workshop on the Semantics and Pragmatics of Dialogue (SemDial 2017–SaarDial)(August 2017)*, vol. 48, 2017.

[11] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning, 2024. Used the llava-hf/llava-1.5-7b-hf model from Hugging Face: https://huggingface.co/llava-hf/llava-1.5-7b-hf.

[12] S. Pandey and A. Ottley. Benchmarking visual language models on standardized visualization literacy tests. In *Computer Graphics Forum*, p. e70137. Wiley Online Library, 2025.

[13] R. Sharma. AI-multimodal-meeting-assistant: A Multimodal Approach to Insight-Driven Meeting Summarization, July 2025. https://github.com/anij715/AI-multimodal-meeting-assistant.

[14] R. S. Tabalba Jr, C. J. Lee, G. Tran, N. Kirshenbaum, and J. Leigh. A pragmatics-based approach to proactive digital assistants for data exploration. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–14, 2025.

[15] C. Xu, Y. Wang, L. Wei, L. Sun, and W. Huang. Improved iterative refinement for chart-to-code generation via structured instruction, 2025.