Evaluating the Semantic Profiling Abilities of LLMs for Natural Language Utterances in Data Visualization

Hannah K. Bako[*] University of Maryland

Arshnoor Bhutani[†] University of Maryland

Xinyi Liu[‡] University of Texas at Austin

Kwesi A. Cobbina[§] University of Maryland

Zhicheng Liu[¶] University of Mayland

Abstract

Automatically generating data visualizations in response to human utterances on datasets necessitates a deep semantic understanding of the utterance, including implicit and explicit references to data attributes, visualization tasks, and necessary data preparation steps. Natural Language Interfaces (NLIs) for data visualization have explored ways to infer such information, yet challenges persist due to inherent uncertainty in human speech. Recent advances in Large Language Models (LLMs) provide an avenue to address these challenges, but their ability to extract the relevant semantic information remains unexplored. In this study, we evaluate four publicly available LLMs (GPT-4, Gemini-Pro, Llama3, and Mixtral), investigating their ability to comprehend utterances even in the presence of uncertainty and identify the relevant data context and visual tasks. Our findings reveal that LLMs are sensitive to uncertainties in utterances. Despite this sensitivity, they are able to extract the relevant data context. However, LLMs struggle with inferring visualization tasks. Based on these results, we highlight future research directions on using LLMs for visualization generation. Our supplementary materials have been shared on GitHub: `https://github.com/hdi-umd/Semantic_Profiling_LLM_Evaluation`.

Index Terms: Human-centered computing—Visualization—Empirical studies in visualization;

1 Introduction

Designing an effective data visualization requires multiple considerations, such as identifying relevant data attributes, preparing the dataset in the right format through data wrangling and transformation, identifying analytical tasks or communication goals, and choosing appropriate visual encoding strategies. Over the years, visualization researchers have primarily focused on different ways to automatically identify appropriate visual encodings [32, 17, 34], but have largely overlooked important aspects such as automating task identification and data preparation. Only recently have researchers started to address these overlooked issues [33, 20, 29].

Among these efforts, natural language interfaces (NLI) have emerged as a popular interaction paradigm for visualization generation. To users, it is easier to articulate their visualization intents through natural language than using programming constructs or complex graphical user interfaces; to system builders, natural language utterances provide valuable information on user intent that could be hard to capture. However, natural language utterances can be difficult to handle due to uncertainties such as ambiguities [7] and under-specification [23]. Furthermore, it is necessary to address issues such as data preparation and task identification in visualization systems with natural language interfaces.

Large Language Models (LLMs) hold great promise for creating natural language interfaces tailored to data visualization, due to their ability to interpret and generate textual data. While a few tools have utilized them for visualization generation [27, 9, 30, 5], they tend to focus on low-level applications of LLMs, such as generating code for data transformations [30] or simply integrating them as part of a pipeline [5]. It is still unclear how well LLMs perform at extracting information crucial to visualization generation from utterances without human interference.

In this work, we embark on an evaluation of the capabilities of LLMs in the **semantic profiling** of natural language utterances for the purpose of data visualization generation. In line with other work, we use the term "utterance" to refer to questions or instructions people use to elicit responses from an NLI or LLM [24]. By semantic profiling, we do not evaluate visualizations generated by LLMs but instead focus on the following dimensions: 1) *clarity analysis*, which determines if an utterance is ambiguous, underspecified, or asking for missing data, 2) *data attribute and transformation identification*, which identifies relevant data columns and any necessary transforms to prepare the data into a usable format, and 3) *task classification*, which seeks to uncover user intent.

To support our research goal, we collated a corpus of 500 data-related utterances based on an evaluation of two NL datasets (NLV-Corpus [24] and Quda [6]). We analyzed utterances with the following annotations: 1) uncertainties such as ambiguities and missing data references, 2) required data attributes and data transformations, and 3) visualization tasks. We then present a systematic analysis of the capabilities of four publicly available LLMs (GPT-4, Llamma3, Mixtral, and Gemini) across the three dimensions of semantic profiling. Our results show that LLMs make inferences at a different level of abstraction than humans, causing them to be hyper-sensitive to uncertainties in utterances. We also find that LLMs perform reasonably at identifying the relevant data columns and data transformations expressed in utterances but are not able to properly infer visualization tasks. We highlight our observations on the current strengths and challenges of LLMs and present a discussion on considerations for using LLMs in visualization generation.

2 Related Work

Natural Language Interfaces for Visualization Generation. There has been extensive research on natural language interfaces (NLI) dating as far back as 2001 when Cox et al. proposed the use of natural language as an input medium for the generation of data visualizations [4]. Since then, a plethora of NLIs have been created [7, 26, 13, 10, 15, 20]. These NLIs use techniques, such as lexical tokenization or semantic parsing, to infer and translate representations of data attributes and tasks in utterances into visualizations. However, when users' utterances are under-specified, inferring the correct data and task representation becomes challenging. Tools such as DataTone circumvent this limitation by allowing users to

[*] e-mail: hbako@cs.umd.edu

[†] e-mail: arshnoor@terpmail.umd.edu

[‡] e-mail: xinyi.liu@utexas.edu

[§] e-mail: kcobbina@cs.umd.edu

[¶] e-mail: leozcliu@umd.edu

resolve ambiguity through GUI widgets. Similarly, Eviza [22] and Evizeon [10] provide users with the ability to interact with generated visualizations and refine designs via follow-up utterances.

Recent research has progressed towards facilitating visualization code generation based on NL input [33, 20] , generating NL explanations for visualizations [14] and recommending input utterances [25]. Together, these works demonstrate the capabilities of NLIs for visualization. However, NLIs still struggle with resolving under-specifications in utterances without human intervention.

Large Language Models for Data Visualization. Technological advances have given rise to improvements in NLIs, such as the use of BERT to translate user intent expressed in NL into a domain-specific language for visualizations [3]. More recently, we have seen an uptick in the applications of Large Language Models for visualization generation. One such tool is ChartLlama [9], which uses a fine-tuned open-source LLM trained on synthetic benchmark dataset generated from GPT-4 [21] to enhance chart generation and comprehension. Some tools develop pipelines to prompt LLM for relevant code for visualization implementations [27, 5, 18], while others use LLMs to facilitate data transformations [30].

There have also been works that evaluate the capabilities of LLMs for different visualization contexts. Li et al. evaluate prompting strategies for generating visualizations based on the nvbench dataset [16]. Vázquez also evaluates LLMs across 3 axes: the variety of generated chart types, supported libraries, and design refinement [28]. However, these evaluations do not present results for multiple LLMs and focus on the visual artifacts produced by these LLMs. Our work builds on this thread of research by evaluating the strengths and limitations of different LLMs in inferring the semantic information needed to create visualizations.

3 Collating Natural Language Utterances

To facilitate the evaluation of LLMs' capabilities for extracting relevant data and visual contexts, we need a set of data-related user utterances to provide as prompts to LLMs. These utterances need to reflect the level of uncertainty found in human speech. To this end, we sourced utterances from two publicly available corpora:

- **NLVCorpus:** This dataset presents 893 utterances collected from an online survey, where 102 respondents were asked to describe utterances they would input to an analytical system to generate a specific visualization [24].
- **Quda:** This dataset utilizes interviews with expert data analysts to generate a corpus of 920 utterances [6]. These utterances were refined and paraphrased via a crowdsourced study to generate a final dataset of 14,035 diverse utterances.

We performed a systematic examination of utterances from each dataset and filtered out utterances if they contained SQL pseudo code, e.g., *"group (region) — For each region, group by (ship status) — For each (region, ship status), calculate the sum of profit"*. For our analysis, we were interested in examining how well LLMs infer the necessary aspects of the semantic profile and not explicit visualization descriptions. Consequently, we also filtered out utterances that specified visualization types or mapping of data to visual elements, e.g., *"give me a scatterplot of imdb rating as x axis and rotten tomatoes rating as y axis"*.

This selection process was first applied to the NLVCorpus dataset, which yielded a total of 134 utterances across 3 unique datasets. We then applied the same inclusion criteria to a subset of the Quda dataset to produce the remaining 309 utterances across 32 datasets. We also included 54 utterances across 2 datasets collected from a classroom activity conducted in an undergraduate level data visualization class at a US-based University . Our final corpus consists of 500 diverse utterances across 37 unique datasets.

4 Generating Ground Truths and LLM Responses

4.1 Manually Annotating Utterances

Three of the authors performed manual annotation of utterances in our corpus. The lead annotator has 5 years of visualization research experience, while the remaining two annotators have at least 2 years of experience creating visualizations. To annotate our corpus of utterances, the lead author drafted an initial codebook from an evaluation of relevant taxonomies for visual tasks and data transformations [2, 19]. Five random utterances were then selected from the corpus, and three of the authors independently examined and annotated them. The authors met in a subsequent meeting to discuss their codes. The codebook was then updated based on this discussion. The three authors manually annotated the remaining 495 utterances over the course of 12 weeks, holding weekly meetings to discuss and resolve conflicts. Here, we describe these annotations.

Uncertainties. We labeled utterances that could lead to multiple interpretations or couldn't be answered with the provided dataset as *uncertain*. We annotated ambiguities and under-specification by highlighting confusing words, explaining their lack of clarity, and suggesting resolutions. For instance, the utterance *"In what manner are good air quality records dispersed throughout the monitored region ?"* was labeled ambiguous because the reference dataset had air quality readings generated at different times for each region. Therefore, the good air quality readings could be split into different time periods (per hour of the day, per date) or even aggregated across the entire dataset. We provided a resolution to calculate summary statistics and generate yearly trends for good air quality.

While annotating the 500 utterances in our corpus, we found 18 utterances that requested information unavailable in the dataset. For instance, on the dataset showing life expectancy by states in the US, one of the utterances asked *"show me the GDP ranking of European countries"*. This dataset did not contain any information about any countries. As such, it is not possible to answer such a question. Since these utterances were obtained from other studies, it is unclear how these utterances came to be. While we did not provide annotations for the relevant data and visual context for these utterances, we still chose to include them when prompting LLMs as we are still interested in evaluating their ability to identify and resolve such uncertainties in utterances.

Data Attributes and Transformations. For each utterance, we identified the relevant data column[s] needed to correctly answer the utterance. Some utterances require data transformations to generate a new data table that can be used to answer the question. We initially captured the operations needed to transform the data table, such as fold, unstack, and group. However, to properly assess the accuracy of these operations, we need to evaluate the actual data tables they generate. As such, we opted to capture the relevant pandas code that would be used to perform data transformations. Using the previous example utterance on the air quality dataset, the data transformation needed to generate the relevant data table was $res = df.groupby(['Generated','Station']).apply(lambda\, x : x[x['Air\_Quality'].lower() == 'good'])$

Visualization Tasks. The visual task[s] were classified based on the inferred intent of the utterance. The taxonomy for these tasks was adopted from published works by Amar et al. [2] and Munzner [19] and include: `Retrieve Value, Filter, Compute Derived Value, Find Extremum, Sort, Determine Range, Characterize Distribution, Find Anomalies, Cluster, Correlate, summarize, Compare, Dependency, Similarity, and Trend.`

4.2 Generating LLM Outputs

We evaluated two proprietary and two open-source LLMs.

Proprietary LLMs. We evaluated OpenAI's GPT4-Turbo ֍ [21] and Google's Gemini-Pro ✦ [8]. GPT4-Turbo has a training data cutoff of December 2023 and Gemini-Pro's training data cutoff is described as "early 2023" [1]. We utilized the Application Programming Interfaces (APIs) for both of these models to generate responses for the 500 utterances in our corpus.

Open Source LLMs. We evaluated two open-source LLMs, Llama3 ○, and Mixtral M, on the Llama factory code base [35]. Llama3 [1] has 70 billion parameters and a context length of 8,000 tokens, with a knowledge cutoff of December 2023. Mixtral-8x7B-Instruct [11] is configured with 46.7 billion parameters and similarly has a knowledge cutoff in December 2023.

4.2.1 Prompt Design

We explored different prompting strategies (One-shot vs. Few-shot) to elicit responses from LLMs. We decided to use a few-shot prompting as it is more suited for complex tasks and allows the model to learn requirements from provided examples [31]. The prompt provided to each model contained similar instructions to those used by our human annotators in Sec.4.1. For the data transformation code, we instructed the LLMs not to include code for plots or complex analyses. We provided three utterance-dataset-output samples, which were not part of our evaluation corpus. These sample utterances included the corresponding ground truth annotations to help the model gain an understanding of the expected output. We also include the first 10 rows of the dataset to provide an overview of the input data schema. Due to space considerations, the full prompt has been provided in supplementary materials [2].

4.2.2 Challenges Retrieving Responses.

We expected to receive a total of 2000 LLM responses (500 per LLM). However, we encountered some issues eliciting responses from the LLMs. Some of our queries using the APIs of proprietary models returned null responses (֎: 9, ✦: 2). For the open-source models, 42 of the responses did not return the JSON annotations and instead returned a text-based answer to the utterance (○:20, M:22). Both models also occasionally failed to correctly format the JSON responses correctly, wrapping keys with '/,' '@,' or '<.' Wrongly formatted JSON responses were resolved manually. The final set contains 1947 valid annotations from the LLMs (֎: 491, ✦: 498, ○: 481, M: 477).

5 Analysis and Results

We analyzed the LLMs responses across three dimensions of semantic profiling: *clarity analysis* (i.e., comprehension of utterances in the presence of uncertainty), proper identification of the *relevant data context*, and proper inference of the *visualization task*.

5.1 Identifying uncertainty

Summary Statistics: Of the 500 utterances in our corpus, the human annotations found uncertainty in 96 of the utterances. A total of 813 uncertainties were found across all LLMs (֎: 268, ✦: 192, ○: 180, M: 173). Of these 813 uncertainties, only 25.1% (n=204) overlapped with human annotations (֎: 74, ✦: 46, ○: 44, M: 40).

Differences in uncertainties classified by LLMs and human annotators. We observe that all LLMs identified a higher proportion of uncertainty in the utterances than those identified by the human annotators (see Fig.1). When we examine some of these uncertainties identified by the LLMs, we find that they describe uncertainty on how to perform analysis or missing context for data column values. For instance, for the utterance *"Can we conclude that higher happiness comes from higher freedom?"*, GPT-4֎ returned the following ambiguity: *"The query does not specify if the analysis should consider other factors that might influence happiness, or if it should be isolated to just happiness and freedom."* To

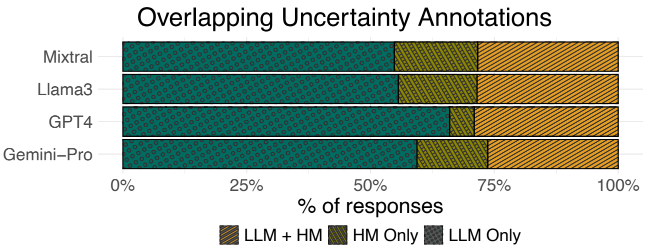


Figure 1: Overview of the overlap in uncertainty annotations between the LLMs and Human (HM) annotations.

the human annotators, this was simply a case of showing the correlation between the two attributes; hence, there was no uncertainty annotation for this utterance. Similarly, for the utterance *"Compare the number of tall buildings in Hong Kong with Taiwan"*, Gemini-Pro✦ classified this as uncertain because *"It is unclear what metric should be used to quantify the tallness of a building. Should the number of stories be used or the height in meters or feet?"*. Our human annotators inferred that the height of the building would be the measure used to answer this utterance.

Uncertainties not found by LLMs. Of the 96 utterances for which human annotators found uncertainty, some were not identified by LLMs (֎: 14, ✦: 32, ○: 34, M: 35). A majority of these uncertainties were as a result of either missing or conflicting data being referenced in the utterance. An example is the utterance "*How can the population of Ashley be illustrated to show the distribution across five years*?" Our annotations labeled this as uncertain because the dataset only contains information from 2000 to 2002, so it is impossible to answer this using the dataset. None of the LLMs labeled this utterance as uncertain.

5.2 Identifying Relevant Data Context

For each data column identified in LLM-generated responses, we examined if they were also identified by human annotators. We defined three levels of agreement between LLMs and human annotations: 1) *total agreement*, where LLMs identify all relevant data columns; 2) *partial agreement*, where LLMs identify some of the data columns; and 3) *total disagreement*, where LLMs identify none of the data columns.

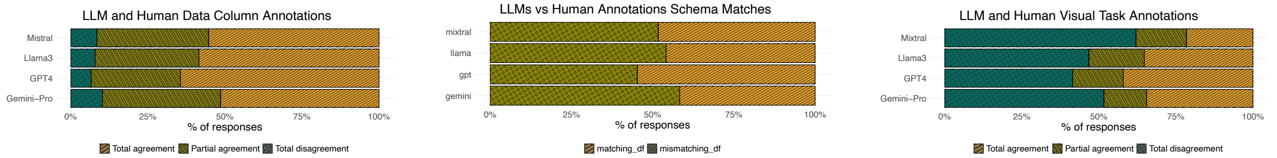
Summary Statistics. Of the 1947 responses returned by LLMs, we filtered out 53 responses that were related to the utterances for which our human annotators did not generate codes for data columns (see Sec. 4.1). We also eliminated an additional 13 responses where the LLMs did not generate data column values, bringing the total responses evaluated for data columns to 1881.

LLMs are able to correctly infer relevant data columns for most utterances. As shown in Fig. 2a , 57.5% of the valid annotations generated by LLMs had a total agreement with the human annotations (֎:312, ✦:241, ○:273, M:255). 34.24% had partial agreement(֎:140, ✦:180, ○:157, M:167) between LLMs and human annotations, while 8.29% had complete disagreement in the relevant data columns identified (֎:32, ✦:48, ○:37, M:39). We observed that 43.6% of these complete disagreement cases had uncertainties identified by either human annotators or LLMs.

5.2.1 Data transformations

For each response generated by an LLM, we executed both the LLM-produced and human-annotated transformations, extracted the resulting data tables from both executions and compared their underlying data schemas (i.e., attribute types) to verify the accuracy of the transformations presented by LLMs. For example, for the utterance *"What is the relationship, if any, between wind and pressure?"*, both the data transforms provided by Llama3○ and human annotations returned a data table with the following schema

[1] According to Google AI documentation
[2] Supplementary Materials



(a) Agreement between LLM and human annotations for relevant data columns.

(b) Data schema matches between data tables returned by LLMs generated code and human annotations.

(c) Agreement between LLM and human annotations for visualization tasks.

Figure 2: Overview of overlapping annotations between LLMs and humans for data attributes, transformations and visual tasks.

$\{wind : int, pressure : int\}$. Since the data tables have the same number and types of attributes, this is a positive match.

While evaluating the data transformations, we found 31 instances where the code for data transformations violated instructions on not returning code for visualization plots or performing complex analyses which were excluded from our analyses (֍:1, ✦:0, ○:15, M:15). Furthermore, we found that 385 of the transformations raised errors of various kinds (֍:59, ✦:96, ○:119, M:111) or returned raw values and not data tables (֍:66, ✦:90, ○:57, M:52). Since the human annotation prioritized data tables as the output of data transformations, we exclude such responses in our analyses.

Data transformations produced by LLMs do not always match those generated by human annotators. The final set for our analysis on data transformation is 1238 responses (֍:360, ✦:290, ○:292, M:296). 48.1% of these responses produced data tables with schemas that match those produced by the human annotations (see Fig. 2b). For the remaining 51.9% where the data did not match what was produced by the code annotated by humans, our evaluation focuses on matches between data schemas. As such, we cannot verify if the resulting data tables provide meaningful answers to the utterance or if they were the result of incorrect data transformations.

5.3 Inferring Visualization Tasks

Similar to the analysis for data columns, we identify three levels of agreement between human and LLM annotations for visual tasks.

Summary Statistics. Of the 1947 responses returned by LLMs, visualization tasks were identified in 1940 responses (֍:490, ✦:494, ○:479, M:477).

Higher proportion of disagreements between human annotations and LLMs for visual task classifications. We observed the highest level of disagreement between LLMs and human annotations in the visual task classifications. 50.4% of the visual tasks were in total disagreement, as seen in Fig. 2c (֍:205, ✦:253, ○:224, M:296). There was total agreement in 33.43% of the responses (֍:208, ✦:169, ○:169, M:103) while the remaining 16.17% had partial agreement for the visual task (֍:81, ✦:68, ○:86, M:78). When we examine a portion of the cases with total disagreement, we observe that some of the issues are a result of conflicting interpretations. For instance, for the utterance *"What is the main factor depending on different status (wind, time, pressure, etc)?"* Gemini✦ classified this as "correlation" whereas the human annotations classified the utterance as "dependency" since correlation cannot be calculated between categorical and numerical attributes. We also see instances where LLMs mix data transformations with visual tasks, e.g., for the utterance *"What was the average budget for each content rating and creative type, as multiple column charts?"* MixtralM classified the utterance as "aggregation, categorization & relationship".

6 Discussion And Future Work

We evaluated the capabilities of four publicly available LLMs in semantic profiling of natural language utterances for data visualization. Our results pose interesting insights for future research.

Using uncertainties to facilitate deeper data exploration and analysis. Our findings show that LLMs found a higher number of uncertainties in utterances compared to our human annotators. It is possible that humans and LLMs identify uncertainties at different levels of abstraction, as humans are able to interpret context more deeply and make better inferences. One such instance of this difference can be seen in the inference in the "tallest building" example provided in Sec.5.1. As a result, LLMs might be more sensitive to uncertainties in utterances. These results amplify the need for LLMs to express their intrinsic uncertainty in responses to allow humans to make informed judgments on how to resolve such uncertainties [12]. Furthermore, the sensitivity of LLMs to uncertainties can be leveraged to pose questions to analysts and help them think deeply about their analysis questions or approach. Facilitating such interactions in NLIs is an interesting research direction.

Improving programming-based responses to utterances. We observed that LLMs are also capable of inferring the appropriate data columns and transformations for over half of the utterances. Yet, for many of the data transformations, we found a number of issues within the code returned by LLMs. This issue is known and tools circumvent this by prompting for multiple code scripts and filtering out erroneous scripts [5, 30]. While these erroneous responses can improve via feedback and fine-tuning prompts, there is a need for further research on how to improve the generation of relevant code for visualization contexts.

Improving visualization task inference to facilitate exploration. We also found that LLMs struggle to correctly infer appropriate visualization tasks from utterances. Nevertheless, there is a need to investigate ways to improve LLMs' ability to infer visualization tasks properly. This is important as these tasks often inform visualization design choices, such as using bar charts for comparison or violin plots to characterize distributions [2, 19, 20]. Proper inference of visual contexts can also facilitate a breadth-wise exploration of data similar to the Voyager system [32]. For instance, if a user is working on the movies dataset and an LLM can infer they are trying to *find anomalies* in the IMDB ratings, it can recommend potentially interesting utterances based on the relevant tasks, such as *comparing* IMDB ratings across creative tasks or finding *correlations* between IMBD and Rotten Tomato ratings.

7 Conclusion

We evaluated the capabilities of four publicly available LLMs (GPT-4֍, Gemini✦, Llama3○ and MixtralM) at correctly inferring the semantic profiles of natural language utterances for data visualization generation. Our findings reveal important strengths of LLMs at identifying uncertainties in utterances and inferring relevant data columns. We also highlight the current limitations of LLMs for generating data transformation code and inferring visualization tasks. Based on our findings, we present future research directions on the use of LLMs for visualization generation.

Acknowledgments

Thanks to the Human-Data Interaction Group for their feedback and support. This work was supported by NSF grant IIS-2239130.

References

[1] AI@Meta. Llama 3 model card. https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md, 2024. 3

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *Proceedings of the 2005 IEEE Symposium on Information Visualization*, p. 15. IEEE Computer Society, USA, 2005. doi: 10.1109/INFOVIS.2005.24 2, 4

[3] Q. Chen, S. Pailoor, C. Barnaby, A. Criswell, C. Wang, G. Durrett, and I. Dillig. Type-directed synthesis of visualizations from natural language queries. *Proc. ACM Program. Lang.*, 6(OOPSLA2), article no. 144, 28 pages, 2022. doi: 10.1145/3563307 2

[4] K. Cox, R. E. Grinter, S. L. Hibino, L. J. Jagadeesan, and D. Mantilla. A multi-modal natural language interface to an information visualization environment. *International Journal of Speech Technology*, 4:297–314, 2001. doi: 10.1023/A:1011368926479 1

[5] V. Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In D. Bollegala, R. Huang, and A. Ritter, eds., *Proceedings of the 61st Annual Meeting of the ACL*, pp. 113–126. ACL, Toronto, Canada, 2023. doi: 10.18653/v1/2023.acl-demo.11 1, 2, 4

[6] S. Fu, K. Xiong, X. Ge, S. Tang, W. Chen, and Y. Wu. Quda: Natural language queries for visual data analytics, 2020. doi: 10.48550/arXiv.2005.03257 1, 2

[7] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios. Datatone: Managing ambiguity in natural language interfaces for data visualization. In *Proc.*, UIST '15, 12 pages, p. 489–500. ACM, New York, NY, USA, 2015. doi: 10.1145/2807442.2807478 1

[8] Google. Gemini api docs and reference — google ai for developers. https://ai.google.dev/gemini-api/docs. 2

[9] 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. doi: 10.48550/arXiv.2311.16483 1, 2

[10] E. Hoque, V. Setlur, M. Tory, and I. Dykeman. Applying pragmatics principles for interaction with visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):309–318, 2018. doi: 10.1109/TVCG.2017.2744684 1, 2

[11] 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. doi: 10.48550/arXiv.2401.04088 3

[12] Z. Jiang, J. Araki, H. Ding, and G. Neubig. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the ACL*, 9:962–977, 2021. doi: 10.1162/tacl_a_00407 4

[13] J.-F. Kassel and M. Rohs. Valletto: A multimodal interface for ubiquitous visual analytics. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, 6 pages, p. 1–6. ACM, New York, NY, USA, 2018. doi: 10.1145/3170427.3188445 1

[14] D. H. Kim, E. Hoque, and M. Agrawala. Answering questions about charts and generating visual explanations. In *Proc.*, CHI '20, 13 pages, p. 1–13. ACM, New York, NY, USA, 2020. doi: 10.1145/3313831.3376467 2

[15] A. Kumar, J. Aurisano, B. Di Eugenio, A. Johnson, A. Gonzalez, and J. Leigh. Towards a dialogue system that supports rich visualizations of data. In R. Fernandez, W. Minker, G. Carenini, R. Higashinaka, R. Artstein, and A. Gainer, eds., *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 304–309. ACL, Los Angeles, 2016. doi: 10.18653/v1/W16-3639 1

[16] G. Li, X. Wang, G. Aodeng, S. Zheng, Y. Zhang, C. Ou, S. Wang, and C. H. Liu. Visualization generation with large language models: An evaluation, 2024. doi: 10.48550/arXiv.2401.11255 2

[17] J. Mackinlay, P. Hanrahan, and C. Stolte. Show me: Automatic presentation for visual analysis. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1137–1144, 2007. doi: 10.1109/TVCG.2007.70594 1

[18] P. Maddigan and T. Susnjak. Chat2vis: Fine-tuning data visualisations using multilingual natural language text and pre-trained large language models, 2023. doi: 10.48550/arXiv.2303.14292 2

[19] T. Munzner. *Visualization analysis and design*. CRC press, 2014. doi: 10.1201/b17511 2, 4

[20] A. Narechania, A. Srinivasan, and J. Stasko. Nl4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):369–379, 2021. doi: 10.1109/TVCG.2020.3030378 1, 2, 4

[21] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. doi: 10.48550/arXiv.2303.08774 2

[22] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang. Eviza: A natural language interface for visual analysis. In *Proc.*, UIST '16, 13 pages, p. 365–377. ACM, New York, NY, USA, 2016. doi: 10.1145/2984511.2984588 2

[23] V. Setlur, M. Tory, and A. Djalali. Inferencing underspecified natural language utterances in visual analysis. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, IUI '19, 12 pages, p. 40–51. ACM, New York, NY, USA, 2019. doi: 10.1145/3301275.3302270 1

[24] A. Srinivasan, N. Nyapathy, B. Lee, S. M. Drucker, and J. Stasko. Collecting and characterizing natural language utterances for specifying data visualizations. In *Proc.*, CHI '21, article no. 464, 10 pages. ACM, New York, NY, USA, 2021. doi: 10.1145/3411764.3445400 1, 2

[25] A. Srinivasan and V. Setlur. Snowy: Recommending utterances for conversational visual analysis. In *Proc.*, UIST '21, 17 pages, p. 864–880. ACM, New York, NY, USA, 2021. doi: 10.1145/3472749.3474792 2

[26] Y. Sun, J. Leigh, A. Johnson, and S. Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In *Proc., SMARTGRAPH'10, Banff, Canada*, pp. 184–195. Springer, 2010. doi: 10.1007/978-3-642-13544-6_18 1

[27] Y. Tian, W. Cui, D. Deng, X. Yi, Y. Yang, H. Zhang, and Y. Wu. Chartgpt: Leveraging llms to generate charts from abstract natural language. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–15, 2024. doi: 10.1109/TVCG.2024.3368621 1, 2

[28] P.-P. Vázquez. Are llms ready for visualization?, 2024. 2

[29] C. Wang, Y. Feng, R. Bodik, I. Dillig, A. Cheung, and A. J. Ko. Falx: Synthesis-powered visualization authoring. In *Proc.*, CHI '21, article no. 106, 15 pages. ACM, New York, NY, USA, 2021. doi: 10.1145/3411764.3445249 1

[30] C. Wang, J. Thompson, and B. Lee. Data formulator: Ai-powered concept-driven visualization authoring. *IEEE Transactions on Visualization and Computer Graphics*, 30(1):1128–1138, 2024. doi: 10.1109/TVCG.2023.3326585 1, 2, 4

[31] J. Wei, X. Wang, D. Schuurmans, M. Bosma, brian ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou. Chain of thought prompting elicits reasoning in large language models. In A. H. Oh, A. Agarwal, D. Belgrave, and K. Cho, eds., *Advances in Neural Information Processing Systems*, 2022. doi: 10.48550/arXiv.2201.11903 3

[32] K. Wongsuphasawat, D. Moritz, A. Anand, J. Mackinlay, B. Howe, and J. Heer. Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):649–658, 2016. doi: 10.1109/TVCG.2015.2467191 1, 4

[33] Z. Wu, V. Le, A. Tiwari, S. Gulwani, A. Radhakrishna, I. Radiček, G. Soares, X. Wang, Z. Li, and T. Xie. Nl2viz: natural language to visualization via constrained syntax-guided synthesis. In *Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering*, ESEC/FSE 2022, 12 pages, p. 972–983. ACM, New York, NY, USA, 2022. doi: 10.1145/3540250.3549140 1, 2

[34] J. Zhao, M. Fan, and M. Feng. Chartseer: Interactive steering exploratory visual analysis with machine intelligence. *IEEE Transactions on Visualization and Computer Graphics*, 28(3):1500–1513, 2022. doi: 10.1109/TVCG.2020.3018724 1

[35] Y. Zheng, R. Zhang, J. Zhang, Y. Ye, Z. Luo, and Y. Ma. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024. doi: 10.48550/arXiv.2403.13372 3