Visualizing Large Language Models: A Brief Survey

Adrian M.P. Braşoveanu*†, Arno Scharl*‡, Lyndon J.B. Nixon *†, and Răzvan Andonie§¶
* Modul University Vienna GmbH, Am Kahlenberg 1, 1190, Vienna, Austria,
email: {adrian.brasoveanu,lyndon.nixon,arno.scharl}@modul.ac.at
† Modul Technology GmbH, Am Kahlenberg 1, 1190, Vienna, Austria, email: {brasoveanu,nixon}@modul.ac.at
‡ webLyzard technology gmbh, Liechtensteinstrasse 41/26, 1090, Vienna, Austria, email: scharl@weblyzard.com
§ Central Washington University, 400 E University Way, Ellensburg, WA 98926, USA, email: razvan.andonie@cwu.edu
¶ Transylvania University, Bulevardul Eroilor 29, Braşov 500036, Romania

***Abstract*—This paper explores the current landscape of visualizing large language models (LLMs). The main objective was threefold. Firstly, we investigate how we can visualize LLM-specific techniques such as prompt engineering, instruction tuning, or guidance. Secondly, LLM causality, interpretability, and explainability are examined through visualization. And finally, we showcase the role of visualization in illuminating the integration of multiple modalities. We are interested in discovering the papers that present visualization systems instead of those that use visualization to showcase a part of their work. Our survey aims to synthesize the state-of-the-art in LLM visualization, offering a compact resource for exploring future research avenues.**

***Index Terms*—Large Language Models, Prompt Engineering, Visualization of Neural Networks, Natural Language Processing (NLP), Explainable AI (XAI)**

I. Introduction

LLMs appeared as a natural extension of the classic language models and dominated the news cycle during the last two years. The fast-paced development was reminiscent of the early days of the Internet, but it swiftly impacted people's lives, especially in academia and industry. It is rare for a new technology to trigger such quick reactions. The gold rush that followed to develop more AI applications should not be surprising. Visualization happens to be at the core of these applications in many cases. It can take the form of a simple old-fashioned chat, a map, or a dashboard, and if needed, it can help us incorporate natural language or other modalities. Due to its unifying role, visualization helps us navigate the new AI world.

The fast adoption of AI legislation in the European Union and other parts of the world suggests that the current AI wave might continue longer than the previous waves. However, recent studies explain that developing new AI systems is a balancing act, as it is rather difficult to assess if they are trustworthy and if the associated risks are acceptable, regardless of the number of scenarios examined [1]. Increasing safety, transparency and accountability is now one of the most important goals when publishing new LLMs, as the EU directive requires risk assessment, low energy consumption, data governance policies (e.g., assessment of bias and fairness, privacy), use of public benchmarks, extensive documentation, and increased security. All the operations need to be performed under human supervision. However, due to their high dimensionality, large training datasets, and complex reasoning strategies, it can often be difficult to discover which parts of an LLM led to a specific decision. Highlighting the information pathways between various components is one of the main functions of LLM visualizations.

The paper is organized as follows: Section II presents the motivation and the methodology of this survey; Section III showcases the various classes of LLM visualizations, the main topics, and visualization types; Section IV presents a brief discussion of these classes. The paper ends with some thoughts about the future of LLM visualizations.

II. Background and Methodology

The current generation of language models started with the release of the Transformer [2] architecture. A few years later, in 2020, the GPT-3 [3] architecture that builds on the Transformer opened the way for ChatGPT. Due to this aspect, we have selected 2020 as the starting year for our survey. Papers published before 2020 are included in the bibliography only if they are historically significant for language models or information visualization.

While large language models can be considered a rather new development, there are many papers about them and related topics. For example, a search after *large language models* on DBLP reports 5,489 articles, whereas searches for the abbreviation LLM report 1,918 articles. Searches for connected topics return significantly less results (e.g., *prompt engineering* search returns 120, *instruction tuning* returns 229, and *retrieval augmented generation* returns 116), but need to be considered. However: i) such searches only include paper titles; and ii) DBLP mostly covers computer science articles. If we consider other domains than computer science, the total number of papers will likely double (e.g., at least 14,000). If we also consider the content of the papers, then the number of papers would double at least once more, leading us to an approximate number of around 28,000. However, the number of publications is increasing fast. Due to this aspect, we have considered restrictions for including papers in this survey.

We started collecting data with well-known academic search engines like Google Scholar and DBLP. These were used to

identify the big topics and outlets in which LLM visualization papers were published. We have then moved on to the portals of the big publishing houses (e.g., SpringerNature, ACL, IEEE, ACM, Elsevier, Wiley, MDPI) to find more similar publications. While many recent articles are published directly on arXiv, we have double-checked if more recent versions were published in well-known conferences and journals. As opposed to scientific ranking engines like Clarivate's Web of Science, which helps find high-impact articles, arXiv does offer a good overview of recent research [4].

The factors that were taken into account for selecting a work include: i) the significance and relevance of the paper in connection to the major topics we identified; ii) open access (if the paper is publicly available); iii) the availability of the source code on well-known open-source repositories (e.g., GitHub); iv) the awareness created through the promotion on various media channels (e.g., LinkedIn, X/Twitter).

We decided to include only papers focused specifically on visualization systems (e.g., dashboards mainly focused on LLM visualization workflows) and visualization methods (e.g., new visualizations). For our purposes, a visualization system is considered to be a software system that leverages the principles of the grammar of graphics [5] for intuitive data exploration, analysis, communication or report generation. The well-known grammar of graphics provides a high-level framework for constructing visualizations by decomposing graphs into their basic components.

Even after implementing these restrictions, we discovered that the number of papers for some topics (e.g., prompt engineering, text-to-image generation) was high. In such cases, we have selected only the papers that proposed something novel or were considered important for one of the selected topics. We have generally avoided dataset papers, except if they were also linked to a visualization interface.

This paper mostly focuses on visualizations published in scientific conferences and journals during the described interval, as opposed to commercial tools. In rare cases, we might also mention tools focused on visualizing foundational models like BERT or RoBERTa instead of LLMs. We have only mentioned such tools or techniques when the articles were among the first to open a particular research avenue.

III. Visualizing Large Language Models

We have decided to focus the core section of our paper on the following domains: i) LLM mechanisms cover methods like prompt engineering, instruction tuning, guidance; ii) causality covers causal relation extraction and causal inference; iii) explainable AI (XAI) covers interpretability and explainability; iv) LLM evaluation covers the new domain of using visualization to evaluate LLMs; v) covers retrieval augmented generation (RAG), code generation and applications that use multiple modalities. Each subsection also includes references to various surveys that cover the respective domain. These references should give casual readers a good starting point towards accumulating more study material about a certain domain. For a general LLM survey, the reader is invited to consult [6]. A brief survey from Braşoveanu and Andonie [7] covers visualization of foundation models. Chatbot interfaces are also not covered by this survey, as there is already a survey that covers them [8].

A. LLM Mechanisms

The success of Transformers for a wide variety of tasks led to the development of a set of visualizations focused on their inner mechanisms like embeddings, attention heads or neuron activation [7]. Similarly, a large LLM visualisation category focuses on specific mechanisms like prompting or instruction tuning.

Prompt engineering methods create input queries or instructions that guide the model towards generating the desired output. While prompts were a common technique used in a variety of domains from creative writing to psychology and management, even before the first LLMs were developed, they have only recently started to be widely used together with techniques like one-shot and few-shot learning [3]. For multimodal use cases, visual prompts can also provide cues about what kind of images or videos we want to generate. When prompts are directed towards a specific output (e.g., prompts used to guide an LLM towards producing brief abstractive summaries), the process is called guiding [9]. Sometimes, the term guidance is also used more generally, for example, when we want to point towards the strategies for directing model output. Instruction tuning refers to fine-tuning LLMs on diverse instructions to improve their generalization capabilities.

An overview of the most essential visualization systems that belong to this category is presented in Table I. Although some of these mechanisms are rather new, some surveys are already available for prompt engineering [10], instruction tuning [11], visual instruction tuning [12].

The main idea in exploring prompt engineering is to track different variants of the same prompt to observe which one leads to the best results [13]. Prompt variation is typically represented through a deck of cards, shopping carts, or template cards. Each prompt is represented through a card, and various stylistic elements (e.g., colour highlights) or statistics are added to identify the best-performing prompt quickly. Tools like Promptaid [16] go a step further and explore topics like perturbation and testing. Besides the cards, image embeddings are also quite common if such an interface is designed for exploring variation in text-to-image generation [16].

Many papers are dedicated to creating new reasoning strategies based on prompts. Chain-of-Thought (CoT) reasoning [23], the originator of this trend, was a simple strategy for step-by-step problem-solving. Since it was a sequential strategy, it was followed by parallel strategies (e.g., Tree-of-Thought), graph strategies (e.g., Graph-of-Thought), and various other multimodal strategies. The most effective visualization from this category is probably the simple code block highlighting produced through the Chain of Code strategy [18]. The basic idea is that a strategy based on writing code can easily outperform CoT. The benchmarking visualizes each code block

TABLE I
VISUALIZING LLM MECHANISMS

Systems and Methods	Category	Topic	Chart Type
PromptIDE [13] and Promptaid [14]	Prompt engineering	Prompt variation, perturbation and testing	Control panel; scatter plot; template cards; shopping cart; perturbation plots; confusion matrix
PromptMagician [15] and PromptTHis [16]	Prompt engineering	Prompt variation for text-to-image (PrompTHis)	Control panel; image variant graph; colored tables; graphs ; image embeddings; hierarchical clustering
ChartGPT [17]	Instruction tuning	Chart dataset generation with instruction tuning	Line chart; stacked bar chart; pie chart
CoC [18]	Prompting strategy	Chain of Code	Code block highlighting; bar charts; line charts
VPT [19] and Color-based PT[20]	Visual prompt tuning	Color-based prompt tuning	t-SNE; Strategy visualization; Thumbnails; fill-in-the-blank
LLaVA [21]	Visual instruction tuning	Large Language and Vision Assistant	Dashboard; bar charts; text heatmaps
DiffusionDB [22]	Prompt benchmarking	Prompt database for text-to-image models	Interactive circle packing (clustering); contour plots; thumbnails

with a different colour depending on the used evaluator (e.g., language model versus Python evaluator).

Visual instruction tuning is especially popular for designing multimodal LLMs. While in many cases, the papers refer to the LLMs rather than the visualizations, some papers are accompanied by demos. LLaVA's interface [21] is rather simple (e.g., a dashboard which contains video thumbnails, an upload area, and text heatmaps), and even though its demo is often offline, it is still widely copied.

B. Causality

The advent of LLMs has also led to a new trend of causal LLMs. Some of the articles included in Table II can also be included in this trend, especially [24]. Many articles showcase questionnaires through which people are asked about their opinions on the LLM responses.

Most modern articles on causality build on Judea Pearl's theoretical foundation (e.g., [38]). More recently, new ideas about causality started to develop. One important trend comes from NLP. A recent survey [39] provides definitions, formalization, and guidelines on using causality within estimation, prediction, and interpretation frameworks. The paper mainly focuses on two problems: methods to estimate causal effects extracted from texts and improving the reliability of NLP methods using causal formalism. Another trend appeared at the intersection of IT and biomedical sciences. Andreas Holzinger and his colleagues [40] focused on multi-modal causal analysis and developed the concept of causability, proposing that one of the goals when designing new visualisations should be the idea that they can be used for causal analysis, similar to how we design new interfaces to be accessible.

Both causal relation extraction and causal inference are covered here.

Causal relation extraction is focused on extracting pairs (cause and effect) from structured or unstructured data. It is an area that is particularly important for medicine and statistics. The interplay between knowledge graphs and foundation models or LLMs constitutes the most interesting area in which causal relation extraction plays an important role. SciKGraph [26] visualizations allow users to cluster and track the dynamic evolution of a scientific field through the knowledge stored in scientific knowledge graphs. The SpEAR model [27] produces small knowledge graphs linked to the ontologically-grounded WordNet word senses [41]. The possibility of building etnographic causal models from these concepts distinguishes this tool. While not a grammar or graphics tool, it is great for building explainable graphs through simple means like coloured nodes and connectors.

Event Causality Identification (ECI), the detection of causal links between events from single sentences (SECI) or across multiple sentences (DECI), is discussed in the ERGO paper [25]. The relations are visualized with solid lines, whereas coreference is presented with interrupted arcs. This method can also be used for LLMs, even though the paper that introduced them mostly focuses on foundation models (e.g., Graph Transformers).

Causal inference determines whether there is a causal relationship between two variables. Casual inference is one of the core functionalities of LLMs. Due to this, the field has exploded since the launch of ChatGPT. Counterfactuals, hypothetical scenarios through which to carefully examine what would happen if a causal relationship did not exist, are one of the most common tools used in causal inference. LLM Analyzer [29] uses interactive tables and small visualizations to analyze counterfactuals.

Almost all articles summarised in Table II were published last year and are still only present in arXiv, suggesting that the field is growing rapidly. Most of these articles currently focus on causal graphs, text heatmaps and classic visualisations (e.g., bar charts, line charts, radar charts). Some of the articles deserve a special mention, in particular: (i) the work on causal geospatial reasoning [37], as it includes the largest number of visualisations; and (ii) the work on evaluating the security of LLMs using causal analysis [34], as it showcases new types

TABLE II
VISUALIZING CAUSALITY.

Systems and Methods	Category	Topic	Chart Type
ERGO [25]	Causal relation extraction	Event causality identification	Arcs between causes and effects
SciKGraph [26]	Causal relation extraction	Clustering of scientific knowledge fields	Clusters relation graph; cluster comparison; network visualizations
SpEAR [27]	Causal relation extraction	Knowledge graph visualization	Colored graph traversals with nodes and relations
PolyJuice [28] and LLM Analyzer [29]	Causal inference	Counterfactual generation	Textual heatmaps; bar charts; error charts; dashboard; interactive tables
Study by Kiciman et al. [30]	Causal inference	Probing Causal Reasoning	Text heatmaps
ILS-CLS [31] and study by Long et al. [32]	Causal inference	Causal discovery	Small multiples; bar charts with error bars; causal graphs
CaCo-CoT [33]	Causal inference	Faithful knowledge reasoning	Text heatmaps
Casper [34]	Causal inference	Causal analysis for evaluating security	Bar charts; line charts; causal effect charts; scatter plots
MoCa [35]	Causal inference	Alignment between humans - LLMs	Text heatmaps; radar chart
CaRing [36]	Neuro-symbolic AI	Neuro-symbolic causal integration	Causal graphs
Study by Chen et al. [37]	Causal Inference	Causal geo-spatial reasoning	Heatmaps; geomaps; line charts; bar charts; histograms;

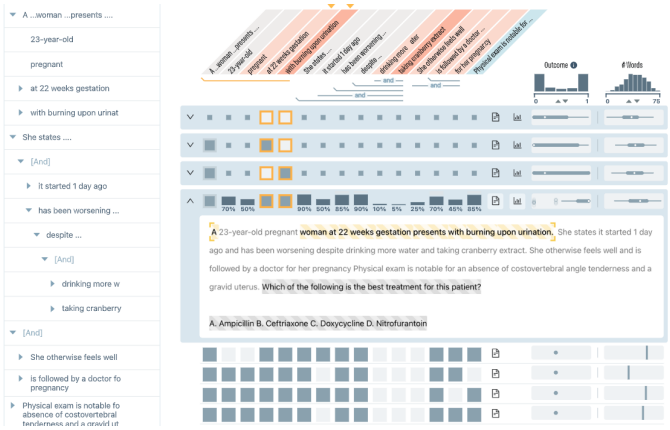

Fig. 1. LLM Analyzer's tabular interface for analyzing counterfactuals. Reproduced from [29].

of visualisations, including one that illustrates the concept of tracing causal effects within LLMs (e.g., on normal generation, single layer and single neurons).

Causal inference is essential for the problem of alignment between humans and machines. Two articles deserve to be mentioned here: one about alignment on moral and causal task [35] and the second on neuro-symbolic AI [36].

C. Explainable AI

As already mentioned in the introduction, the introduction of the EU's AI Act has turned LLM explainability into one of the most important topics in LLM research, as all LLM decisions need to be accompanied by an explanation. Two large-scale surveys on the topic of LLM explainability are available: one offers an overview of the topic in[42], and the second is focused on the issue of LLM trustworthiness [43].

Some of the models discussed in the previous subsection can also be integrated into this subsection. Still, due to causality playing a major role in their development, they were placed in the respective subsection. Similar judgments were made about models that appear in the following two subsections.

Cito et al. [54] discuss using counterfactual models to explain model predictions. The visualisations are focused on changes in code and testing.

Natural language to visualization is the trend closest to the original idea of the grammar of graphics, as it allows us to generate various visualizations based on a specification. A recent survey about the trend is available in Shen et al. [55], whereas some general ideas about these types of visualization generation environments are described in Shen et al. [56]. Two interesting tools perform the opposite operations: NL2VIS generates visualizations from prompts based on Vega-Lite, whereas VL2NL generates visualization specifications for the Vega-Lite format. Vega [57], or Vega-Lite [58] are libraries designed by Jeffrey Heer's group for automated visualization generation. The NLDV toolkit [49] belongs to this trend and provides interactive guidelines for generating specifications for visualisations from natural language queries. The paper describes a Python package that takes a table as an input and returns a list of Vega-Lite visualisation specifications as a JSON document. Vega-Lite is an interactive graphics grammar built under the supervision of Jeffrey Heer, one of the creators of D3 visualisation library [59]. The NLDV toolkit is particularly impressive, as it parses queries and performs implicit and explicit attribute inference (e.g., inference through both attribute names and values), explicit and implicit task inference, and visualisation generation.

The Anthropic team has been constantly publishing foundational work on mechanistic interpretability. The recent work on LLM feature activation [47] is included here as it showcases how to use visualization to identify and correct LLM features related to bias, deception, manipulation or even criminal content.

Other similar tools that can be categorized under visualization recommendation include Chat2VIS [53] which is

TABLE III
VISUALIZING XAI

Systems and Methods	Category	Topic	Chart Type
FAIR [44] and FedJudge [45]	Legal intelligence	Legal intelligence and inference	Bar charts; t-SNE plots; text heatmaps
Theory-of-Mind (ToM) [46]	Explainability	Explaining social reasoning	ToM templates; causal graphs; text heatmaps, error bar charts
LEGO [24]	Explainability	Agentic causal explanation generation	Text heatmaps; bar charts
Claude 3 Sonnet [47]	Interpretability	Extraction of interpretable features	Density and conditional distribution plots; text heatmapts; scatter and bubble charts; parallel coordinates
Boundless DAS method [48]	Interpretability	Scaling Alpaca interpretability	Token heatmaps; line charts
NL4DV [49] and NL2VIS [50]	Natural language to visualization	Visualization generation from natural language	Vega-Lite visualizations; classic charts
VL2NL [51]	Natural language benchmarking	Visualization generation from prompts	Generation of natural language datasets
LLM4Vis [52] and [53]	Visualization recommendation	Few-shot prompting for visualization recommendation	Text heatmaps; classic charts

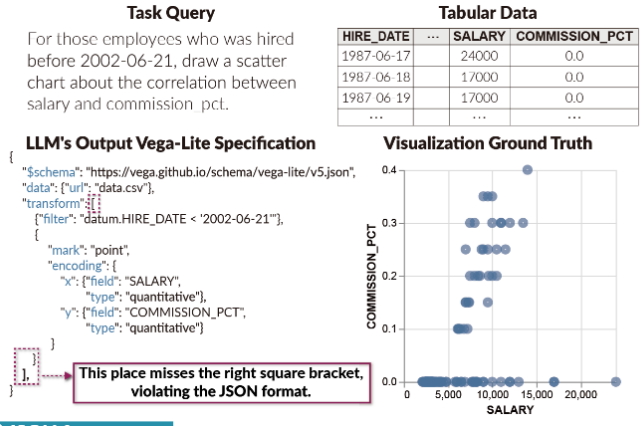

Fig. 2. Visualizing benchmarking errors -nvBench JSON errors. Reproduced from [50].

focused on the problem of fine-tuning data visualisations using ChatGPT and CodeLLama, an instruction-tuned version of LLama focused on code generation, and LLM4Vis [52] which is focused on few-shot prompting techniques for ChatGPT. Visualization recommenders can also be included in the applications subsection. However, we decided to include them here since these tools also have built-in interpretability and explainability, as every recommended visualisation comes with an explanation of why it was selected in the first place.

D. Visual LLM Evaluation

Traditionally, the evaluation process involved computing some metrics, analyzing the errors, improving the process, and repeating it repeatedly until the desired results were reached. If explainability was required, some feature importance or hyperparameter charts were added in some cases. However, LLM evaluations need to consider many factors (e.g., language, domain, training data, speed, corectness, etc) and cover the entire lifecycle of the models, therefore visualizations are often front and centre when designing LLM evaluation tools. Two large-scale surveys about this topic are available: [60] and [61].

We are interested in showcasing a rather new development related to evaluating LLMs, namely their direct usage in evaluations and visualizations' role in exploring this particular aspect. We are, therefore, interested in cases in which LLMs are both the evaluated systems and the evaluators.

An early attempt on using visualization for evaluating language models focused on context-sensitive visualization methods of the most influential word combinations for a classifier [68]. It leads to heatmaps that include more relevant information on the classification and more accurately highlight the most important words from the input text. The method uses a dependency parser, a foundational model (BERT), and the leave-n-out technique. Further papers ([69] and [70]) investigate how to evaluate explanations and visualizations resulting from NLP models for classification.

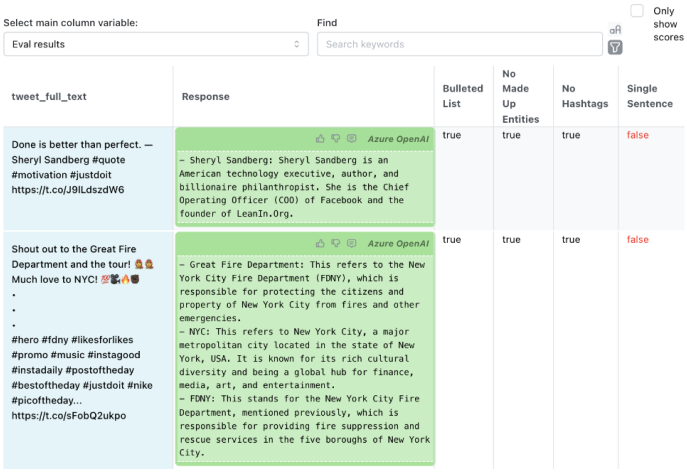

Fig. 3. EvalGen annotated LLM NER evaluation interface. Reproduced from [67]

A major challenge is comparing the performances of different visualizations of LLMs. Accuracy cannot be used to

TABLE IV
VISUAL EVALUATION OF LLMS.

Systems and Methods	Category	Topic	Chart Type
Study by Zhang et al.[62]	SQL optimization	Text-to-SQL, SQL debugging and optimization	Scatter plot; word cloud; bar charts
MatPlotAgent [63]	Scientific visualization	Evaluating Agent-based scientific visualization	Tabular views; text heatmaps; matplotlib visualizations
EvaLLM [64]	AI-generated visualizations	Conceptual evaluation of AI-generated visualization	Scatter plot; bar charts
Causal Auditor [65]	Causal evaluation	Augmented LLM causality evaluation	Causal graphs; causal diagrams (confounder charts; debate charts)
LLM Comparator [66]	Visual evaluation	Side by side evaluations	Tabular views; visualization summaries
EvalGen [67]	Evaluating Alignment	Evaluating LLM-assisted alignment	Dashboard; tabular views; confusion matrix

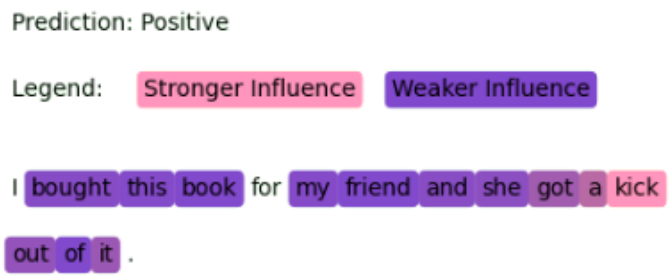

Fig. 4. Text heatmap for a positive product review from the Amazon Reviews dataset. Reproduced from [70]

evaluate visualisation quality, but more rigorous criteria are needed to measure the usefulness of the extracted knowledge for explaining the models. The LLM Comparator [66] continues a tradition specific to NLP systems, that of side-by-side comparisons of the annotations produced by different systems (in our case, LLMs). EvalGen [67] provides annotated tabular interfaces to evaluate LLM results. Matplotagent evaluates agentic visualizations built with the MatPlotLib Python library [63]. Two other articles ([71] and [64]) are focused on conceptual evaluations rather than visualization systems.

Perhaps due to the need to compare results, tabular views and scatter plots are some of the most common visualizations for dashboards in this category. Additionally, visualization summaries or matplotlib visualizations are included depending on the requirements.

E. LLM Applications

The number of applications is rather large, so in this section, we generally focus on multimedia applications. Some of the modalities we look at include text, image, video, audio, speech, sensor data, and sometimes haptics. Applications are usually focused on dashboards as they provide a unified interface that seamlessly integrates diverse data sources, facilitating comprehensive insights and correlations across different modalities in a concise and accessible format.

Multimodal LLMs (e.g., [83] and [84]) integrate and process information from multiple types of data sources, such as text, images, and sound, to enhance analysis or interaction. In contrast, cross-modal systems focus on translating or linking information between different types of data modalities, such as converting visual data into textual descriptions.

Multimedia systems have a long tradition of using visualization for benchmarking, one of the best-known recent systems used for such tasks being Grad-CAM [85], a system that generates explanations using gradient-based localization (e.g., detect which objects from an image are the most significant contributors to the model's prediction). Dashboards are also typically used for multimodal or cross-modal visualizations. They provide a unified interface that seamlessly integrates diverse data sources, facilitating comprehensive insights and correlations across different modalities in a concise and accessible format. They enable the automated tracking of various stories, points of view or formats.

Typical interfaces offer easy options for navigating the video frames for spatio-temporal reasoning, one or multiple text areas for subtitles or linguistic relation extraction (e.g., see [86]) or inference, and visualisations (e.g., embeddings [87]). Visualisations themselves might also include text areas with highlighted words, and they can be as simple as highlighting correctly or incorrectly retrieved images [87].

No-code and low-code platforms are a relatively new development. JarviX [78] presents an easy interface to optimize and analyze tabular data. LLMs can also help new programmers visually generate code, low-code platforms providing workflows and colour highlighting whenever needed [79].

Retrieval-Augmented Text Generation (RAG) combines information retrieval with generative AI techniques like summarization, recommendation, or text generation [88]. Some recent RAG applications include a tool for recognizing PDF structure called ChatDoc [80], a traffic prediction application called RealGen [81] which uses animation and RAG to predict various traffic-related scenarios, and an application that tracks large collection of documents in context [82]. RealGen [81] deserves a special mention here for its use of animation.

These applications offer a preview of what is possible by integrating visualizations with LLMs. From now on, we expect interfaces blending multiple modalities and simulators that use more animation to play a significant role.

TABLE V
LLM APPLICATIONS.

Systems and Methods	Category	Topic	Chart Type
ChartLlama [72] and VizAbility [73]	Multimodal LLMs	Chart generation, accessibility, and understanding	Classic visualizations (line, bar, cart, etc.)
HiLM-D [74]	Multimodal	Autonomous driving	Text heatmaps; object recognition
CMCL algorithm [75]	Cross-modal	Causal Structure and representation learning	Line charts; weighted matrix; interventions (medical pictures)
DeVADG [76]	Cross-modal	Domain generalization via confounder disentanglement	Causal graphs; bar charts; t-SNE; videos with text heatmaps
VLCI [77]	Cross-modal	Visual-linguistic intervention for generating radiology reports	Medical dashboards; images; model charts; text heatmaps
JarviX [78]	No code or low-code	No code tabular data analysis	Tabular views; basic charts
LowCodeLLM [79]	No code or low-code	Low-code visual programming	Workflow diagrams; text heatmaps; block code highlighting
ChatDoc [80]	Retrieval-Augmented Generation	PDF structure recognition	PDF visualization; code editor
RealGen [81]	Retrieval-Augmented Generation	Traffic control	Traffic visualizations (e.g., car crash prediction)
HINTs [82]	Retrieval-Augmented Generation	Sense making of large document collections	Cluster views; heatmaps (document and chatbot views)

IV. DISCUSSION

The number of papers published about LLMs is rather high for a new discipline. This is a testament to the rapid impact of this new research field. On the other hand, many papers are published directly on arXiv, a well-known preprint server that does not use the classic peer review process scientific journals and conferences tend to use. Without arXiv reviews, it is only possible to evaluate the quality of a paper once it is accepted by a proper scientific outlet or if the code itself is also available open-source.

As expected, some classic visualization types are frequently used (e.g., line or bar charts). NLP visualizations use a lot of text heatmaps, which might also include arrows to showcase connections between the words. Exploratory LLM visualizations (e.g., prompt variation) use the deck of cards or similar design patterns to present multiple options on the same chart. Computer vision applications use grids to display videos or pictures, embeddings to showcase the most important topics, and text heatmaps for prompts or associated texts.

Due to their relatively recent emergence, the subsections on LLM mechanics and visual LLM evaluations are rather brief.

On the other hand, the subsection on applications can be continuously extended as new applications appear every day. RAG seems to be one of the possible futures of information retrieval, the other being the agent that provides a single answer. Low-code or no-code environments are becoming more popular every day, not only for teaching programming but also in the industry. Generating quick visualizations is particularly useful in the industry, as they can always be reused through statistics, slides, or reports. As the previous section shows, new visualization paradigms emerge in each domain. While not all these domains are yet mature enough to warrant separate surveys, they certainly present opportunities for further exploration.

V. CONCLUSION

We have selected domains with a lot of potential for future growth. This does not mean that they are necessarily the ones that will grow the most, but rather that, given what we have already seen, there are still many growth areas.

Our decision to focus on systems and methods meant that many good papers, especially those focused on annotation tools, datasets, or new models, would not have been included if visualization had been used as support, and it was not the paper's main goal. Our survey should be seen as a starting point for exploring the topics related to LLM visualization and not as a comprehensive study.

As it can be seen from our previous sections, we are only starting to use LLMs and visualizations together. The fact that most papers were published last year suggests this is a fast-moving field. New visualization domains appear daily and are quickly integrated into business workflows.

ACKNOWLEDGMENTS

Dr. Adrian M.P. Braşoveanu is partially funded by the Vienna Science and Technology Fund project (WWTF) [10.47379/ICT20096] and SDG-HUB (FFG, G.A. No. 892212). Prof. Arno Scharl is partially funded by the E.U. through the Horizon Europe project ENEXA (G.A. 101070305). Dr. Lyndon J.B. Nixon is partially funded by the E.U. through the Horizon Europe project TRANSMIXR (G.A. 101070109). Pictures of the visualizations are reproduced with permission from the authors.

REFERENCES

[1] C. Novelli, F. Casolari, A. Rotolo, M. Taddeo, and L. Floridi, "AI risk assessment: A scenario-based, proportional methodology for the AI act," *Digit. Soc.*, vol. 3, no. 1, p. 13, 2024. [Online]. Available: https://doi.org/10.1007/s44206-024-00095-1

[2] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[3] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, "Language models are few-shot learners," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020. [Online]. Available: https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html

[4] B. Markscheffel and S. Mollenhauer, "A comprehensive comparison of arxiv and the web of science (wos)," in *2021 ICoASL : 7th International Conference of Asian Special Libraries*, 2021, pp. 55–69. [Online]. Available: https://doi.org/10.22032/dbt.55570

[5] L. Wilkinson, *The Grammar of Graphics, Second Edition*, ser. Statistics and computing. Springer, 2005.

[6] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J. Nie, and J. Wen, "A survey of large language models," *CoRR*, vol. abs/2303.18223, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.18223

[7] A. M. P. Brasoveanu and R. Andonie, "Visualizing transformers for NLP: A brief survey," in *24th International Conference on Information Visualisation, IV 2020, Melbourne, Australia, September 7-11, 2020*, E. Banissi, F. Khosrow-shahi, A. Ursyn, M. W. M. Bannatyne, J. M. Pires, N. Datia, K. Nazemi, B. Kovalerchuk, J. Counsell, A. Agapiou, Z. Vrcelj, H. Chau, M. Li, G. Nagy, R. Laing, R. Francese, M. Sarfraz, F. Bouali, G. Venturini, M. Trutschl, U. Cvek, H. Müller, M. Nakayama, M. Temperini, T. D. Mascio, F. Sciarrone, V. Rossano, R. Dörner, L. Caruccio, A. Vitiello, W. Huang, M. Risi, U. Erra, R. Andonie, M. A. Ahmad, A. Figueiras, A. Cuzzocrea, and M. S. Mabakane, Eds. IEEE, 2020, pp. 270–279. [Online]. Available: https://doi.org/10.1109/IV51561.2020.00051

[8] E. Kavaz, A. Puig, and I. Rodríguez, "Chatbot-based natural language interfaces for data visualisation: A scoping review," *Applied Sciences*, vol. 13, no. 12, p. 7025, 2023. [Online]. Available: https://doi.org/10.3390/app13127025

[9] H. Kumar, I. Musabirov, M. Reza, J. Shi, A. Kuzminykh, J. J. Williams, and M. Liut, "Impact of guidance and interaction strategies for LLM use on learner performance and perception," *CoRR*, vol. abs/2310.13712, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.13712

[10] P. Sahoo, A. K. Singh, S. Saha, V. Jain, S. Mondal, and A. Chadha, "A systematic survey of prompt engineering in large language models: Techniques and applications," *CoRR*, vol. abs/2402.07927, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.07927

[11] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu, and G. Wang, "Instruction tuning for large language models: A survey," *CoRR*, vol. abs/2308.10792, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.10792

[12] J. Huang, J. Zhang, K. Jiang, H. Qiu, and S. Lu, "Visual instruction tuning towards general-purpose multimodal model: A survey," *CoRR*, vol. abs/2312.16602, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.16602

[13] H. Strobelt, A. Webson, V. Sanh, B. Hoover, J. Beyer, H. Pfister, and A. M. Rush, "Interactive and visual prompt engineering for ad-hoc task adaptation with large language models," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 1, pp. 1146–1156, 2023. [Online]. Available: https://doi.org/10.1109/TVCG.2022.3209479

[14] A. Mishra, U. Soni, A. Arunkumar, J. Huang, B. C. Kwon, and C. Bryan, "Promptaid: Prompt exploration, perturbation, testing and iteration using visual analytics for large language models," *CoRR*, vol. abs/2304.01964, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.01964

[15] Y. Feng, X. Wang, K. Wong, S. Wang, Y. Lu, M. Zhu, B. Wang, and W. Chen, "Promptmagician: Interactive prompt engineering for text-to-image creation," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 1, pp. 295–305, 2024. [Online]. Available: https://doi.org/10.1109/TVCG.2023.3327168

[16] Y. Guo, H. Shao, C. Liu, K. Xu, and X. Yuan, "Prompthis: Visualizing the process and influence of prompt editing during text-to-image creation," *arXiv preprint arXiv:2403.09615*, 2024. [Online]. Available: https://doi.org/10.1109/TVCG.2024.3408255

[17] A. Masry, M. Shahmohammadi, M. R. Parvez, E. Hoque, and S. Joty, "Chartinstruct: Instruction tuning for chart comprehension and reasoning," *arXiv preprint arXiv:2403.09028*, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.09028

[18] C. Li, J. Liang, A. Zeng, X. Chen, K. Hausman, D. Sadigh, S. Levine, L. Fei-Fei, F. Xia, and B. Ichter, "Chain of code: Reasoning with a language model-augmented code emulator," *CoRR*, vol. abs/2312.04474, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.04474

[19] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, and L. Jiang, "Visual prompt tuning for generative transfer learning," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 2023, pp. 19 840–19 851. [Online]. Available: https://doi.org/10.1109/CVPR52729.2023.01900

[20] Y. Yao, A. Zhang, Z. Zhang, Z. Liu, T. Chua, and M. Sun, "CPT: colorful prompt tuning for pre-trained vision-language models," *CoRR*, vol. abs/2109.11797, 2021. [Online]. Available: https://arxiv.org/abs/2109.11797

[21] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, Eds., 2023. [Online]. Available: http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

[22] Z. J. Wang, E. Montoya, D. Munechika, H. Yang, B. Hoover, and D. H. Chau, "Diffusiondb: A large-scale prompt gallery dataset for text-to-image generative models," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, A. Rogers, J. L. Boyd-Graber, and N. Okazaki, Eds. Association for Computational Linguistics, 2023, pp. 893–911. [Online]. Available: https://doi.org/10.18653/v1/2023.acl-long.51

[23] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, Eds., 2022. [Online]. Available: http://papers.nips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

[24] Z. He, P. Cao, Y. Chen, K. Liu, R. Li, M. Sun, and J. Zhao, "LEGO: A multi-agent collaborative framework with role-playing and iterative feedback for causality explanation generation," in *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Association for Computational Linguistics, 2023, pp. 9142–9163. [Online]. Available: https://aclanthology.org/2023.findings-emnlp.613

[25] M. Chen, Y. Cao, K. Deng, M. Li, K. Wang, J. Shao, and Y. Zhang, "ERGO: event relational graph transformer for document-level event causality identification," in *Proceedings of the 29th International Conference on Computational Linguistics, COLING 2022, Gyeongju, Republic of Korea, October 12-17, 2022*, N. Calzolari, C. Huang, H. Kim, J. Pustejovsky, L. Wanner, K. Choi, P. Ryu, H. Chen, L. Donatelli, H. Ji, S. Kurohashi, P. Paggio, N. Xue, S. Kim, Y. Hahm, Z. He, T. K. Lee, E. Santus, F. Bond, and S. Na, Eds. International Committee on Computational Linguistics, 2022, pp. 2118–2128. [Online]. Available: https://aclanthology.org/2022.coling-1.185

[26] M. D. L. Tosi and J. C. dos Reis, "Understanding the evolution of a scientific field by clustering and visualizing knowledge graphs," *J. Inf. Sci.*, vol. 48, no. 1, pp. 71–89, 2022. [Online]. Available: https://doi.org/10.1177/0165551520937915

[27] S. E. Friedman, I. H. Magnusson, V. Sarathy, and S. Schmer-Galunder, "From unstructured text to causal knowledge graphs: A transformer-based approach," *CoRR*, vol. abs/2202.11768, 2022. [Online]. Available: https://arxiv.org/abs/2202.11768

[28] T. Wu, M. T. Ribeiro, J. Heer, and D. S. Weld, "Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Association for Computational Linguistics, 2021, pp. 6707–6723. [Online]. Available: https://doi.org/10.18653/v1/2021.acl-long.523

[29] F. Cheng, V. Zouhar, R. S. M. Chan, D. Fürst, H. Strobelt, and M. El-Assady, "Interactive analysis of llms using meaningful counterfactuals," *arXiv preprint arXiv:2405.00708*, 2024. [Online]. Available: https://arxiv.org/pdf/2405.00708

[30] E. Kiciman, R. Ness, A. Sharma, and C. Tan, "Causal reasoning and large language models: Opening a new frontier for causality," *CoRR*, vol. abs/2305.00050, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.00050

[31] T. Ban, L. Chen, D. Lyu, X. Wang, and H. Chen, "Causal structure learning supervised by large language model," *CoRR*, vol. abs/2311.11689, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.11689

[32] S. Long, T. Schuster, and A. Piché, "Can large language models build causal graphs?" *CoRR*, vol. abs/2303.05279, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.05279

[33] Z. Tang, R. Wang, W. Chen, K. Wang, Y. Liu, T. Chen, and L. Lin, "Towards causalgpt: A multi-agent approach for faithful knowledge reasoning via promoting causal consistency in llms," *CoRR*, vol. abs/2308.11914, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.11914

[34] W. Zhao, Z. Li, and J. Sun, "Causality analysis for evaluating the security of large language models," *CoRR*, vol. abs/2312.07876, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.07876

[35] A. Nie, Y. Zhang, A. Amdekar, C. Piech, T. Hashimoto, and T. Gerstenberg, "Moca: Measuring human-language model alignment on causal and moral judgment tasks," *CoRR*, vol. abs/2310.19677, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.19677

[36] S. Yang, X. Li, L. Cui, L. Bing, and W. Lam, "Neuro-symbolic integration brings causal and reliable reasoning proofs," *CoRR*, vol. abs/2311.09802, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.09802

[37] Y. Chen, Y. Gan, S. Li, L. Yao, and X. Zhao, "More than correlation: Do large language models learn causal representations of space?" *CoRR*, vol. abs/2312.16257, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.16257

[38] J. Pearl, *Causality*. Cambridge university press, 2009.

[39] A. Feder, K. A. Keith, E. Manzoor, R. Pryzant, D. Sridhar, Z. Wood-Doughty, J. Eisenstein, J. Grimmer, R. Reichart, M. E. Roberts, B. M. Stewart, V. Veitch, and D. Yang, "Causal inference in natural language processing: Estimation, prediction, interpretation and beyond," *Trans. Assoc. Comput. Linguistics*, vol. 10, pp. 1138–1158, 2022. [Online]. Available: https://transacl.org/ojs/index.php/tacl/article/view/4005

[40] A. Holzinger, A. M. Carrington, and H. Müller, "Measuring the quality of explanations: The system causability scale (SCS)," *Künstliche Intell.*, vol. 34, no. 2, pp. 193–198, 2020. [Online]. Available: https://doi.org/10.1007/s13218-020-00636-z

[41] C. Fellbaum and A. Hicks, "When wordnet met ontology," in *Ontology Makes Sense - Essays in honor of Nicola Guarino*, ser. Frontiers in Artificial Intelligence and Applications, S. Borgo, R. Ferrario, C. Masolo, and L. Vieu, Eds., vol. 316. IOS Press, 2019, pp. 136–151. [Online]. Available: https://doi.org/10.3233/978-1-61499-955-3-136

[42] H. Zhao, H. Chen, F. Yang, N. Liu, H. Deng, H. Cai, S. Wang, D. Yin, and M. Du, "Explainability for large language models: A survey," *CoRR*, vol. abs/2309.01029, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.01029

[43] Y. Liu, Y. Yao, J. Ton, X. Zhang, R. Guo, H. Cheng, Y. Klochkov, M. F. Taufiq, and H. Li, "Trustworthy llms: a survey and guideline for evaluating large language models' alignment," *CoRR*, vol. abs/2308.05374, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2308.05374

[44] M. He, N. Gu, Y. Shi, Q. Zhang, and Y. Chen, "FAIR: A causal framework for accurately inferring judgments reversals," *CoRR*, vol. abs/2306.11585, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.11585

[45] L. Yue, Q. Liu, Y. Du, W. Gao, Y. Liu, and F. Yao, "Fedjudge: Federated legal large language model," *CoRR*, vol. abs/2309.08173, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.08173

[46] K. Gandhi, J. Fränken, T. Gerstenberg, and N. D. Goodman, "Understanding social reasoning in language models with language models," *CoRR*, vol. abs/2306.15448, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2306.15448

[47] A. Templeton, T. Conerly, J. Marcus, J. Lindsey, T. Bricken, B. Chen, A. Pearce, C. Citro, E. Ameisen, A. Jones, H. Cunningham, N. L. Turner, C. McDougall, M. MacDiarmid, C. D. Freeman, T. R. Sumers, E. Rees, J. Batson, A. Jermyn, S. Carter, C. Olah, and T. Henighan, "Scaling monosemanticity: Extracting interpretable features from claude 3 sonnet," *Transformer Circuits Thread*, 2024. [Online]. Available: https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html

[48] Z. Wu, A. Geiger, C. Potts, and N. D. Goodman, "Interpretability at scale: Identifying causal mechanisms in alpaca," *CoRR*, vol. abs/2305.08809, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.08809

[49] A. Narechania, A. Srinivasan, and J. T. Stasko, "NL4DV: A toolkit for generating analytic specifications for data visualization from natural language queries," *IEEE Trans. Vis. Comput. Graph.*, vol. 27, no. 2, pp. 369–379, 2021. [Online]. Available: https://doi.org/10.1109/TVCG.2020.3030378

[50] 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," *CoRR*, vol. abs/2401.11255, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.11255

[51] H. Ko, H. Jeon, G. Park, D. H. Kim, N. W. Kim, J. Kim, and J. Seo, "Natural language dataset generation framework for visualizations powered by large language models," *CoRR*, vol. abs/2309.10245, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.10245

[52] L. Wang, S. Zhang, Y. Wang, E. Lim, and Y. Wang, "Llm4vis: Explainable visualization recommendation using chatgpt," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*, M. Wang and I. Zitouni, Eds. Association for Computational Linguistics, 2023, pp. 675–692. [Online]. Available: https://aclanthology.org/2023.emnlp-industry.64

[53] P. Maddigan and T. Susnjak, "Chat2vis: Generating data visualizations via natural language using chatgpt, codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45 181–45 193, 2023. [Online]. Available: https://doi.org/10.1109/ACCESS.2023.3274199

[54] J. Cito, I. Dillig, V. Murali, and S. Chandra, "Counterfactual explanations for models of code," in *44th IEEE/ACM International Conference on Software Engineering: Software Engineering in Practice, ICSE (SEIP) 2022, Pittsburgh, PA, USA, May 22-24, 2022*. IEEE, 2022, pp. 125–134. [Online]. Available: https://doi.org/10.1109/ICSE-SEIP55303.2022.9794112

[55] L. Shen, E. Shen, Y. Luo, X. Yang, X. Hu, X. Zhang, Z. Tai, and J. Wang, "Towards natural language interfaces for data visualization: A survey," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 6, pp. 3121–3144, 2023. [Online]. Available: https://doi.org/10.1109/TVCG.2022.3148007

[56] Y. Wang, Z. Hou, L. Shen, T. Wu, J. Wang, H. Huang, H. Zhang, and D. Zhang, "Towards natural language-based visualization authoring," *IEEE Trans. Vis. Comput. Graph.*, vol. 29, no. 1, pp. 1222–1232, 2023. [Online]. Available: https://doi.org/10.1109/TVCG.2022.3209357

[57] A. Satyanarayan, R. Russell, J. Hoffswell, and J. Heer, "Reactive vega: A streaming dataflow architecture for declarative interactive visualization," *IEEE Trans. Vis. Comput. Graph.*, vol. 22, no. 1, pp. 659–668, 2016. [Online]. Available: https://doi.org/10.1109/TVCG.2015.2467091

[58] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer, "Vega-lite: A grammar of interactive graphics," *IEEE Trans. Vis. Comput. Graph.*, vol. 23, no. 1, pp. 341–350, 2017. [Online]. Available: https://doi.org/10.1109/TVCG.2016.2599030

[59] M. Bostock, V. Ogievetsky, and J. Heer, "D^3 data-driven documents," *IEEE Trans. Vis. Comput. Graph.*, vol. 17, no. 12, pp. 2301–2309, 2011. [Online]. Available: https://doi.org/10.1109/TVCG.2011.185

[60] Y. Chang, X. Wang, J. Wang, Y. Wu, K. Zhu, H. Chen, L. Yang, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," *CoRR*, vol. abs/2307.03109, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2307.03109

[61] Z. Guo, R. Jin, C. Liu, Y. Huang, D. Shi, Supryadi, L. Yu, Y. Liu, J. Li, B. Xiong, and D. Xiong, "Evaluating large language models: A comprehensive survey," *CoRR*, vol. abs/2310.19736, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2310.19736

[62] B. Zhang, Y. Ye, G. Du, X. Hu, Z. Li, S. Yang, C. H. Liu, R. Zhao, Z. Li, and H. Mao, "Benchmarking the text-to-sql capability of large language models: A comprehensive evaluation," *arXiv preprint arXiv:2403.02951*, 2024. [Online]. Available: https://arxiv.org/pdf/2403.02951

[63] Z. Yang, Z. Zhou, S. Wang, X. Cong, X. Han, Y. Yan, Z. Liu, Z. Tan, P. Liu, D. Yu, Z. Liu, X. Shi, and M. Sun, "Matplotagent: Method and evaluation for llm-based agentic scientific data visualization," *CoRR*, vol. abs/2402.11453, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.11453

[64] L. Podo, M. Ishmal, and M. Angelini, "Vi(e)va llm! A conceptual stack for evaluating and interpreting generative ai-based visualizations," *CoRR*, vol. abs/2402.02167, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.02167

[65] Y. Zhang, B. Fitzgibbon, D. Garofolo, A. Kota, E. Papenhausen, and K. Mueller, "An explainable AI approach to large language model assisted causal model auditing and development," *CoRR*, vol. abs/2312.16211, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.16211

[66] M. Kahng, I. Tenney, M. Pushkarna, M. X. Liu, J. Wexler, E. Reif, K. Kallarackal, M. Chang, M. Terry, and L. Dixon, "LLM comparator: Visual analytics for side-by-side evaluation of large language models," *CoRR*, vol. abs/2402.10524, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.10524

[67] S. Shankar, J. D. Zamfirescu-Pereira, B. Hartmann, A. G. Parameswaran, and I. Arawjo, "Who validates the validators? aligning llm-assisted evaluation of LLM outputs with human preferences," *CoRR*, vol. abs/2404.12272, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2404.12272

[68] A. Dunn, D. Inkpen, and R. Andonie, "Context-sensitive visualization of deep learning natural language processing models," in *25th International Conference Information Visualisation, IV 2021, Sydney, Australia, July 5-9, 2021*, E. Banissi, A. Ursyn, M. W. M. Bannatyne, J. M. Pires, N. Datia, M. L. Huang, W. Huang, Q. V. Nguyen, K. Nazemi, B. Kovalerchuk, M. Nakayama, J. Counsell, A. Agapiou, F. Khosrow-shahi, H. Chau, M. Li, R. Laing, F. Bouali, G. Venturini, M. Temperini, and M. Sarfraz, Eds. IEEE, 2021, pp. 170–175. [Online]. Available: https://doi.org/10.1109/IV53921.2021.00035

[69] ——, "Evaluation of deep learning context-sensitive visualization models," in *26th International Conference Information Visualisation, IV 2022, Vienna, Austria, July 19-22, 2022*, E. Banissi, A. Ursyn, M. W. M. Bannatyne, J. M. Pires, N. Datia, K. Nazemi, B. Kovalerchuk, R. Andonie, M. Nakayama, F. Sciarrone, W. Huang, Q. V. Nguyen, M. S. Mabakane, A. Rusu, M. Temperini, U. Cvek, M. Trutschl, H. Müller, H. Siirtola, W. L. Woo, R. Francese, V. Rossano, T. D. Mascio, F. Bouali, G. Venturini, S. Kernbach, D. Malandrino, R. Zaccagnino, J. J. Zhang, X. Yang, and V. Geroimenko, Eds. IEEE, 2022, pp. 359–365. [Online]. Available: https://doi.org/10.1109/IV56949.2022.00066

[70] ——, "Designing and evaluating context-sensitive visualization models for deep learning text classifiers," in *Artificial Intelligence and Visualization: Advancing Visual Knowledge Discovery*, B. Kovalerchuk, K. Nazemi, R. Andonie, N. Datia, and E. Bannissi, Eds. Cham: Springer Nature Switzerland, 2024, pp. 399–421. [Online]. Available: https://doi.org/10.1007/978-3-031-46549-9_14

[71] P.-P. Vázquez, "Are llms ready for visualization?" *arXiv preprint arXiv:2403.06158*, 2024. [Online]. Available: https://arxiv.org/pdf/2403.06158

[72] 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," *CoRR*, vol. abs/2311.16483, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2311.16483

[73] J. Gorniak, J. Ottiger, D. Wei, and N. W. Kim, "Vizability: Multimodal accessible data visualization with keyboard navigation and conversational interaction," in *Adjunct Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology, UIST 2023, San Francisco, CA, USA, 29 October 2023- 1 November 2023*, S. Follmer, J. Han, J. Steimle, and N. H. Riche, Eds. ACM, 2023, pp. 18:1–18:3. [Online]. Available: https://doi.org/10.1145/3586182.3616669

[74] X. Ding, J. Han, H. Xu, W. Zhang, and X. Li, "Hilm-d: Towards high-resolution understanding in multimodal large language models for autonomous driving," *CoRR*, vol. abs/2309.05186, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.05186

[75] H. Mao, H. Liu, J. X. Dou, and P. V. Benos, "Towards cross-modal causal structure and representation learning," in *Machine Learning for Health, ML4H 2022, 28 November 2022, New Orleans, Lousiana, USA & Virtual*, ser. Proceedings of Machine Learning Research, A. Parziale, M. Agrawal, S. Joshi, I. Y. Chen, S. Tang, L. Oala, and A. Subbaswamy, Eds., vol. 193. PMLR, 2022, pp. 120–140. [Online]. Available: https://proceedings.mlr.press/v193/mao22a.html

[76] S. Zhang, X. Feng, W. Fan, W. Fang, F. Feng, W. Ji, S. Li, L. Wang, S. Zhao, Z. Zhao, T. Chua, and F. Wu, "Video-audio domain generalization via confounder disentanglement," in *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, B. Williams, Y. Chen, and J. Neville, Eds. AAAI Press, 2023, pp. 15 322–15 330. [Online]. Available: https://doi.org/10.1609/aaai.v37i12.26787

[77] W. Chen, Y. Liu, C. Wang, G. Li, J. Zhu, and L. Lin, "Visual-linguistic causal intervention for radiology report generation," *CoRR*, vol. abs/2303.09117, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2303.09117

[78] S. Liu, S. Wang, T. Chang, W. Lin, C. Hsiung, Y. Hsieh, Y. Cheng, S. Luo, and J. Zhang, "Jarvix: A LLM no code platform for tabular data analysis and optimization," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*, M. Wang and I. Zitouni, Eds. Association for Computational Linguistics, 2023, pp. 622–630. [Online]. Available: https://aclanthology.org/2023.emnlp-industry.59

[79] Y. Cai, S. Mao, W. Wu, Z. Wang, Y. Liang, T. Ge, C. Wu, W. You, T. Song, Y. Xia, J. Tien, and N. Duan, "Low-code LLM: visual programming over llms," *CoRR*, vol. abs/2304.08103, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2304.08103

[80] D. Lin, "Revolutionizing retrieval-augmented generation with enhanced PDF structure recognition," *CoRR*, vol. abs/2401.12599, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2401.12599

[81] W. Ding, Y. Cao, D. Zhao, C. Xiao, and M. Pavone, "Realgen: Retrieval augmented generation for controllable traffic scenarios," *CoRR*, vol. abs/2312.13303, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2312.13303

[82] S. Y. Lee and K. Ma, "Hints: Sensemaking on large collections of documents with hypergraph visualization and intelligent agents," *CoRR*, vol. abs/2403.02752, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2403.02752

[83] J. Wu, W. Gan, Z. Chen, S. Wan, and P. S. Yu, "Multimodal large language models: A survey," in *IEEE International Conference on Big Data, BigData 2023, Sorrento, Italy, December 15-18, 2023*, J. He, T. Palpanas, X. Hu, A. Cuzzocrea, D. Dou, D. Slezak, W. Wang, A. Gruca, J. C. Lin, and R. Agrawal, Eds. IEEE, 2023, pp. 2247–2256. [Online]. Available: https://doi.org/10.1109/BigData59044.2023.10386743

[84] D. Caffagni, F. Cocchi, L. Barsellotti, N. Moratelli, S. Sarto, L. Baraldi, L. Baraldi, M. Cornia, and R. Cucchiara, "The (r)evolution of multimodal large language models: A survey," *CoRR*, vol. abs/2402.12451, 2024. [Online]. Available: https://doi.org/10.48550/arXiv.2402.12451

[85] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020. [Online]. Available: https://doi.org/10.1007/s11263-019-01228-7

[86] Y. Liu, G. Li, and L. Lin, "Cross-modal causal relational reasoning for event-level visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 11 624–11 641, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3284038

[87] X. Dong, X. Zhan, Y. Wei, X. Wei, Y. Wang, M. Lu, X. Cao, and X. Liang, "Entity-graph enhanced cross-modal pretraining for instance-level product retrieval," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13 117–13 133, 2023. [Online]. Available: https://doi.org/10.1109/TPAMI.2023.3291237

[88] L. Li, Y. Zhang, D. Liu, and L. Chen, "Large language models for generative recommendation: A survey and visionary discussions," *CoRR*, vol. abs/2309.01157, 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2309.01157