Computers & Graphics 130 (2025) 104246

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CGVC2024

# Foundation model assisted visual analytics: Opportunities and Challenges

Maeve Hutchinson *, Radu Jianu, Aidan Slingsby, Pranava Madhyastha

*Department of Computer Science, City St George's, University of London, Northampton Square, London, EC1V 0HB, United Kingdom*

ARTICLE INFO

*Keywords:*
Visual analytics
Visualisation
Foundation models
Large Language Models
Multimodality

ABSTRACT

We explore the integration of foundation models, such as large language models (LLMs) and multimodal LLMs (MLLMs), into visual analytics (VA) systems through intuitive natural language interactions. We survey current research directions in this emerging field, examining how foundation models have already been integrated into key visualisation-related processes in VA: visual mapping, the creation of data visualisations; visualisation observation, the process of generating a finding through visualisation; and visualisation manipulation, changing the viewport or highlighting areas of interest within a visualisation. We also highlight new possibilities that foundation models bring to VA, in particular, the opportunities to use MLLMs to interpret visualisations directly, to integrate multimodal interactions, and to provide guidance to users. We finally conclude with a vision of future VA systems as collaborative partners in analysis and address the prominent challenges in realising this vision through foundation models. Our discussions in this paper aim to guide future researchers working on foundation model assisted VA systems and help them navigate common obstacles when developing these systems.

## 1. Introduction

Visual analytics (VA) emphasises an analytical partnership between the computer and the human analyst, combining computational methods with interactive visualisation in an iterative process. This human–computer collaboration is vital for uncovering insights into data through VA. However, effectively leveraging such systems requires expertise across analytical techniques, data visualisation principles, and domain-specific knowledge. This creates a high barrier to entry which can put powerful VA tools out of reach for many users. Moreover, analysing large, multi-faceted datasets typically involves iterative processes where visualisations and computational analyses have to be repeatedly configured and refined to probe new perspectives, hypotheses, and insights. The interactions needed to make this happen can often add significant overhead to the analysis process.

The emergence of foundation models, such as large language models (LLMs) and multimodal LLMs (MLLMs), presents an increasingly viable solution to alleviate these limitations and support analysts within VA systems. With their vast knowledge bases and advanced natural language processing capabilities, these models can facilitate mode intuitive and expressive communication between user and system. VA tools have the potential to evolve from passive tools to collaborative partners in analysis, capable of adapting to users' needs, thus reducing the technical and cognitive barriers required to engage with VA.

Recent work has explored the intersection of VA and natural language processing (NLP), recognising the potential benefits of integrating language-based interfaces with visualisation tools. Shen et al. [1] review Visualisation-Oriented Natural Language Interfaces (V-NLIs), systems that support natural language (NL) input to produce visualisations. Similarly, Voigt et al. [2] focus on the use of NL in visualisation, covering systems that use NL either as an input or output modality. Hoque and Islam [3] survey natural language generation for visualisation, outlining key tasks, challenges, and future directions. Wang et al. [4] and Wu et al. [5] explored the use of machine learning and AI techniques in visualisation. Ye et al. [6] examine how generative AI techniques apply across different stages of the visualisation pipeline. Yang et al. [7] survey the bidirectional interplay between foundation models and visualisations, highlighting how each can enhance the other.

In contrast to these existing reviews, in this paper, we explore the integration of foundation models into VA, specifically for visualisation. We examine their current applications, future opportunities, and potential challenges. We discuss their applications across three key visualisation-related VA processes: visual mapping, visualisation observation, and visualisation manipulation.

The remainder of this paper is structured as follows: Section 2 provides the necessary background for this paper; Sections 3 to 5 explore the current state-of-the-art, highlighting examples of foundation

* Corresponding author.
*E-mail addresses:* maeve.hutchinson@citystgeorges.ac.uk (M. Hutchinson), radu.jianu@citystgeorges.ac.uk (R. Jianu), a.slingsby@citystgeorges.ac.uk (A. Slingsby), pranava.madhyastha@citystgeorges.ac.uk (P. Madhyastha).

https://doi.org/10.1016/j.cag.2025.104246
Received 24 January 2025; Received in revised form 11 April 2025; Accepted 4 May 2025
Available online 21 May 2025

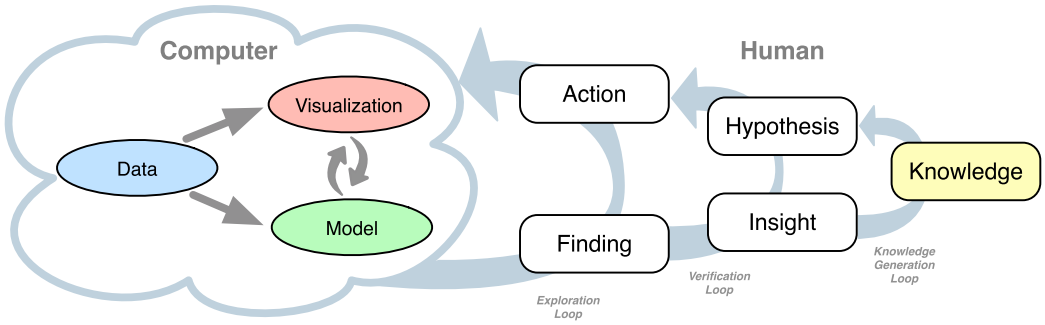



**Fig. 1.** Sacha et al.'s knowledge generation model for visual analytics. Image from Sacha et al. [15].

model-assisted VA systems and future opportunities; finally, Section 6 concludes with a vision for foundation model-assisted VA systems as collaborative analytical partners, outlining key directions for future research and addressing challenges that must be overcome to realise this vision.

## 2. Background

In this section, we present the relevant background necessary for the discussions in subsequent sections. We begin by defining foundation models and highlighting their capabilities. We then outline the key VA framework we use, defining the processes that form the focus of this paper.

### 2.1. Foundation models

Foundation Models are deep learning models trained on a large amount of broad data and can be applied to a wide range of tasks [8]. The advent of foundation models marked a shift from task-specific models towards more flexible, general-purpose machine learning models.

A Large Language Models (LLM) is a type of foundation model developed to process and generate natural language [8]. LLMs are trained on massive corpora of text data, allowing them to learn complex linguistic patterns, and thus perform a wide range of natural language processing tasks, including text generation, translation, question answering, and summarisation. Examples of prominent LLMs include OpenAI's GPT series [9] and Meta's LLaMA [10], all of which have demonstrated remarkable capabilities across a range of natural language processing tasks.

Multimodal Large Language Models (MLLMs) are LLM-based models capable of processing and generating multimodal information [11]. In this context, multimodal information refers to multiple data modalities, such as images, text, and audio. In this work, we are mostly concerned with MLLMs that can interpret images alongside text. Examples include OpenAI's GPT-4V, which extends GPT-4's capabilities to process visual inputs, enabling tasks such as image captioning and visual question answering [12], and Google's Gemini, designed to process and generate integrated text-image outputs [13]. These models are typically trained on datasets dominated by natural images and textual descriptions [11]. This training allows them to effectively align textual and visual modalities. However, this focus on natural images presents challenges when applying MLLMs to visualisations, which differ significantly in structure and intent.

### 2.2. Visual analytics

Visual Analytics is broadly defined as "the science of analytical reasoning facilitated by visual interfaces" [14]. This field combines data analytics with interactive information visualisation to support complex decision-making processes.

In this paper, we adopt Sacha et al.'s knowledge generation model of VA as a central framework [15] (Fig. 1). The model emphasises that knowledge is generated through an iterative process, modelling it as interlinked loops of exploration, verification, and knowledge generation. They separate parts of their model into "human" and "computer" components. We focus on the interactions at this human–computer interface. Fig. 2 shows a more detailed part of the model at the interface of the human and computer components, including action paths, which describe individual tasks taken by a human interacting with a VA system, and cognition paths, observations made by a human that result in a finding.

This model provides a structured lens to examine how foundation models can enhance VA. Specifically, we use the model to identify and address challenges across the three key visualisation-related processes in Fig. 2: visual mapping, the creation of data visualisations; visualisation observation, the process of generating a finding through visualisation; and visualisation manipulation, changing the viewport or highlighting areas of interest within a visualisation.

VA systems have evolved significantly over the past decade, shifting from passive tools to active participants in the analytical process. To facilitate this transformation, researchers have explored the development of mixed-initiative, guiding, and adaptive VA systems. Mixed-initiative systems [16] are characterised by collaborative interaction between the user and the system, where both parties actively contribute towards a common analytical goal. Adaptive systems are designed to continuously update their knowledge and behaviour throughout the analysis process [17]. This adaptivity allows the system to respond to the user's actions, preferences, and evolving understanding of the data.

Guidance is a process that "aims to actively resolve a knowledge gap encountered by users" [18]. Initially, only system-to-user guidance was considered, but later work acknowledged that guidance can be a mixed-initiative process, including guidance from the user to the system [19]. Sperrle et al. [20] further developed this idea, proposing a model of co-adaptive guidance that emphasises the continuous adaptation of both user and system behaviour and knowledge throughout the VA process as they guide each other.

In parallel, research has explored the integration of multimodal interactions into VA systems. Multimodal systems leverage multiple input and output modalities, such as touch, gesture, gaze, and natural language, to enable more natural and intuitive interaction between users and the system [21]. By providing users with a variety of interaction modalities, these systems aim to create a richer and more engaging analytical experience, facilitating the exchange of knowledge between the user and the system.

Building on the concept of richer, multimodal interaction, the use of natural language (NL) as an input modality in VA systems has seen significant research. Early work on integrating NL into VA systems focused on using classical NLP pipelines, which often struggled with the ambiguity and noise inherent in human language [22]. However, the emergence of LLMs has revolutionised the field of NLP, offering a more robust and flexible approach to language understanding and generation.

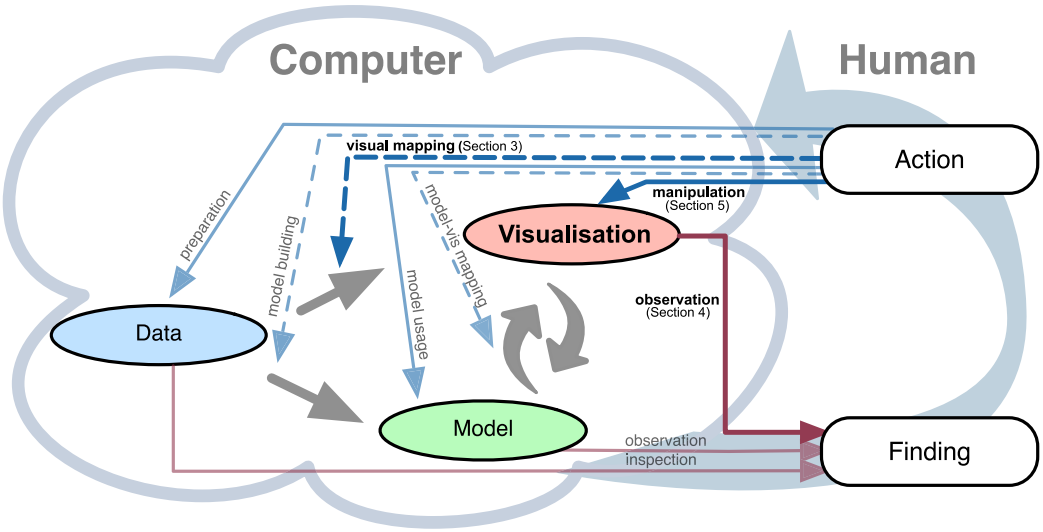



**Fig. 2.** A detailed part of the knowledge generation model for visual analytics at the human–computer interface. Solid blue arrows represent actions leading directly to analytic elements, while dashed blue arrows indicate actions leading to their mappings. Red arrows show processes where system responses are observed by the user to generate findings. This figure also highlights the three key processes explored in subsequent sections: visual mapping (Section 3), visualisation observation (Section 4), and visualisation manipulation (Section 5).
*Source:* Image adapted from Sacha et al. [15].

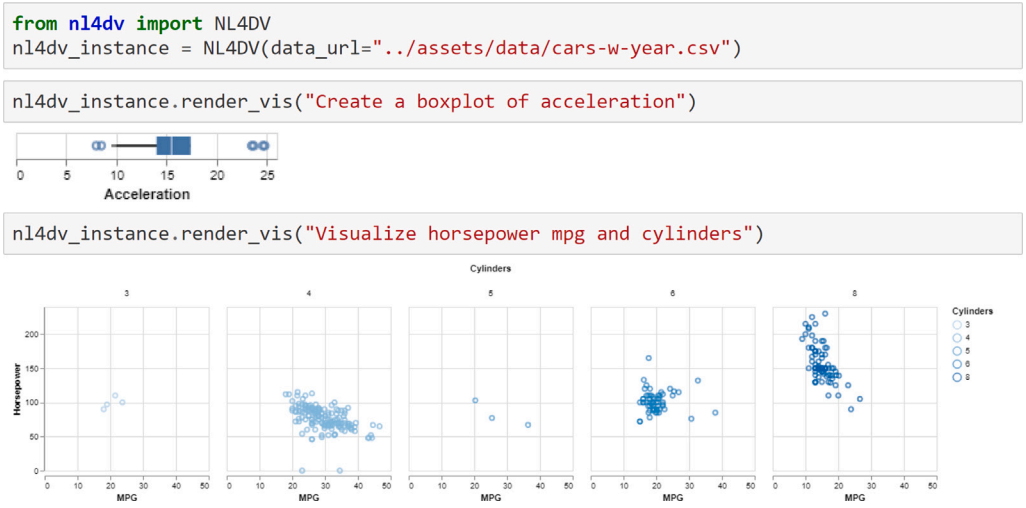



**Fig. 3.** A demonstration of the rule-based V-NLI NL4DV integrated within a Jupyter Notebook, enabling users to generate visualisations by specifying natural language instructions in Python. Image from Narechania et al. [23].

Recent research in VA has explored the use of LLMs to support parts of the analytical process, leveraging their capabilities in understanding and generating NL. Some VA research has also examined the capabilities of MLLMs in understanding images alongside text. Section 3 illustrates how foundation models have been used so far to enhance VA systems, discusses the limitations of these efforts, and motivates future research.

In the next sections we discuss current work across three key processes from the knowledge generation model for VA [15] (Fig. 2): visual mapping, visualisation observation, and visualisation manipulation. In each section, we highlight the potential opportunities and challenges associated with leveraging foundation models in VA.

## 3. Foundation models for visual mapping

Visual mapping actions are those that create data visualisations [15]. The creation of visualisations is central to VA, but requires the user both understand what they want to achieve and have the technical expertise to configure it. Foundation models have demonstrated the potential to lower these barriers by facilitating visualisation creation through natural language interactions, as discussed in the following section.

### 3.1. Existing systems

There has been significant research into NL as an input modality for visualisation creation before LLMs. One category is Visualisation-Oriented Natural Language Interfaces (V-NLIs) that given data, go directly from NL queries to generating corresponding visualisations [23–25, interalia] (Fig. 3). Shen et al. provide a comprehensive review of V-NLI systems [1]. Another related area is that of visualisation recommendation systems and search interfaces that allow users to input NL queries and output ranked recommendations [26,27, interalia] (Fig. 4). These older rule-based systems are typically categorised based on the specific tasks they are designed to handle, as they require customised development for each task. Foundation models have the potential to unify these systems, as they possess capabilities across multiple tasks.

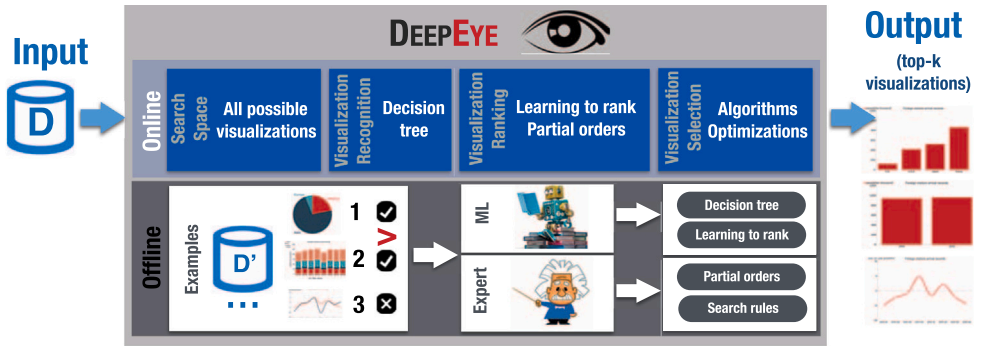
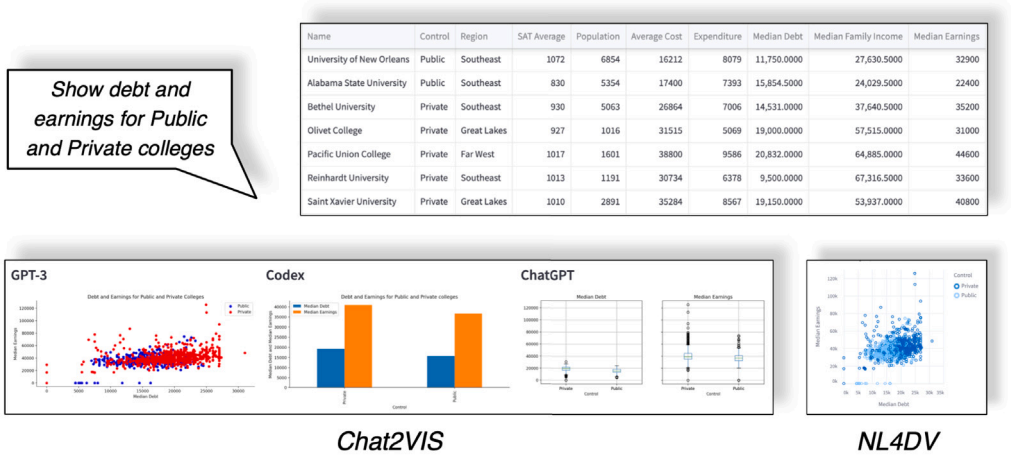



**Fig. 4.** Overview of DEEPEYE, a rule-based visualisation recommendation system. Image from Luo et al. [26].




| Name | Control | Region | SAT Average | Population | Average Cost | Expenditure | Median Debt | Median Family Income | Median Earnings |
|---|---|---|---|---|---|---|---|---|---|
| University of New Orleans | Public | Southeast | 1072 | 6854 | 16212 | 8079 | 11,750.0000 | 27,630.5000 | 32900 |
| Alabama State University | Public | Southeast | 830 | 5354 | 17400 | 7393 | 15,854.5000 | 24,029.5000 | 22400 |
| Bethel University | Private | Southeast | 930 | 5063 | 26864 | 7006 | 14,531.0000 | 37,640.5000 | 35200 |
| Olivet College | Private | Great Lakes | 927 | 1016 | 31515 | 5069 | 19,000.0000 | 57,515.0000 | 31000 |
| Pacific Union College | Private | Far West | 1017 | 1601 | 38800 | 9586 | 20,832.0000 | 64,885.0000 | 44600 |
| Reinhardt University | Private | Southeast | 1013 | 1191 | 30734 | 6378 | 9,500.0000 | 67,316.5000 | 33600 |
| Saint Xavier University | Private | Great Lakes | 1010 | 2891 | 35284 | 8567 | 19,150.0000 | 53,937.0000 | 40800 |

**Fig. 5.** Visualisations generated with Chat2Vis, an LLM-based system, compared with rule-based NL4DV. Image from Maddigan et al. [28].

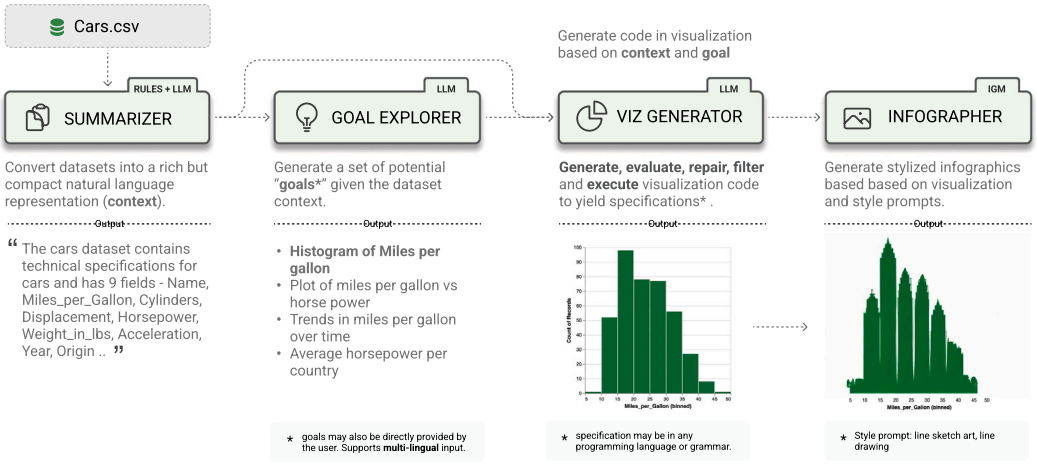
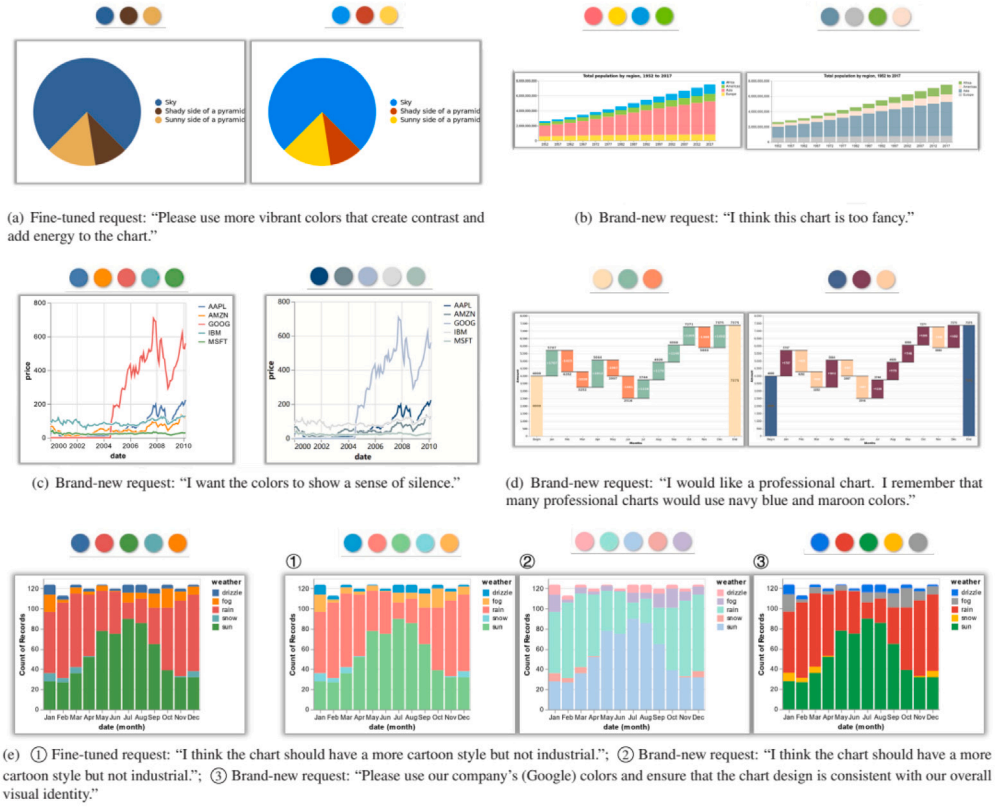



**Fig. 6.** LIDA system outputs, showcasing its four core modules: data summarisation, goal exploration, visualisation generation, and infographic generation. Image from Dibia [29]).

While we typically think of a visualisation as an image, most successful work in visualisation creation to date relies on generating specifications to map data to visual attributes. Despite advances in foundation models, visualisation grammars or code continue to play a crucial role as a bridge between NL inputs and the rendering of visualisations. For example, Chat2Vis [28] demonstrates GPT-3, Codex, and ChatGPT to generate visualisations from user queries, first generating Python code based on the query, which is then used to produce the visualisation (Fig. 5). LIDA [29] works similarly, generating Python code to produce visualisations from user queries (Fig. 6). ChartGPT [30] produces Vega-Lite specifications from user queries in a structured way, generating answers to sub-tasks defining the filter, mark, encoding, and sort, and combining them to form a visualisation.

The integration of foundation models into VA systems has opened up new possibilities for more intuitive and flexible visualisation design. The ability of LLMs to understand nuanced NL input allows users to communicate their preferences and requirements more expressively, enabling the system to customise the visualisation accordingly. NL2Color [31] illustrates this capability by leveraging GPT-3 to interpret user expressions and modify the colour scheme of the visualisation




(a) Fine-tuned request: "Please use more vibrant colors that create contrast and add energy to the chart."

(b) Brand-new request: "I think this chart is too fancy."

(c) Brand-new request: "I want the colors to show a sense of silence."

(d) Brand-new request: "I would like a professional chart. I remember that many professional charts would use navy blue and maroon colors."

(e) ① Fine-tuned request: "I think the chart should have a more cartoon style but not industrial."; ② Brand-new request: "I think the chart should have a more cartoon style but not industrial."; ③ Brand-new request: "Please use our company's (Google) colors and ensure that the chart design is consistent with our overall visual identity."

**Fig. 7.** Examples of visualisation colour palette refinement from a user query using an LLM by NL2Color. Image from Shi et al. [31].

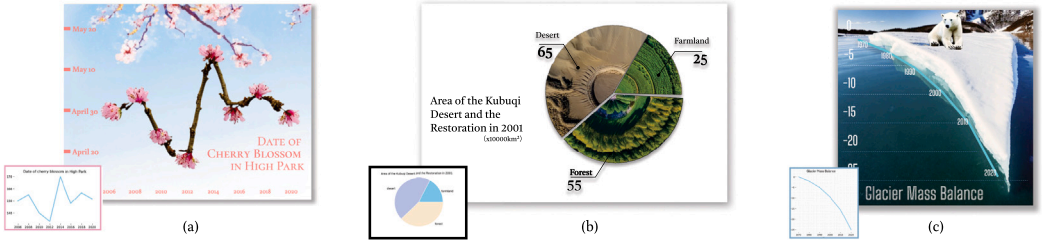



**Fig. 8.** Examples of pictorial visualisations created by ChartSpark using a text-to-image generative model. From [32].

based on the user's input (Fig. 7). This approach highlights the potential for LLMs to help create more customised visualisation designs.

Image-based generative foundation models have also been applied to the design aspects of visualisation. LIDA [29] uses a text-to-image generation model to turn visualisations into stylised infographics based on user prompts (Fig. 6). ChartSpark [32] employs a similar approach to generate pictorial visualisations (Fig. 8). These systems demonstrate the potential for leveraging generative models to create visually appealing artistic representations of data. Schetinger et al. [33] offer a comprehensive review of previous work and opportunities for text-to-image generative models in data visualisation.

### 3.2. Opportunities and challenges

Existing approaches to support visual mapping with foundation models have made significant improvements on earlier rule-based systems by enabling more flexible NL input. However, these systems are still constrained in the types of visualisations they can produce, as they mostly rely on Python or Vega-Lite as a bridge between user input and visualisation, which have limited expressiveness. There is opportunity to develop more flexible visualisation systems using foundation models. One possibility is to use a more versatile bridge, such as D3.js [34], which provides a wider range of capabilities for creating interactive visualisations.

Image-based generative foundation models provide another prospect for overcoming the limitations of code as a bridge, as they could potentially generate visualisations directly without the need for an intermediary programming language. Such models have been primarily applied to general computer vision tasks on natural images and creative applications but could be extended to visualisation-specific stimuli. By enabling models to directly generate visual representations, rather than relying on intermediate text-based specifications, multimodal models could offer a more flexible approach to visualisation, moving beyond the expressivity of particular specification languages. This could expand the range of possible visualisations and allow for more natural and expressive interactions between the user and the system. While this approach is still in its early stages, we believe it holds promise for creating highly customisable and diverse visualisations.

However, this direction also presents significant challenges, particularly in interpreting and generating interactivity, which is a crucial aspect of VA. Generating static images may limit the user's ability to explore and interact with the data effectively. Additionally, the

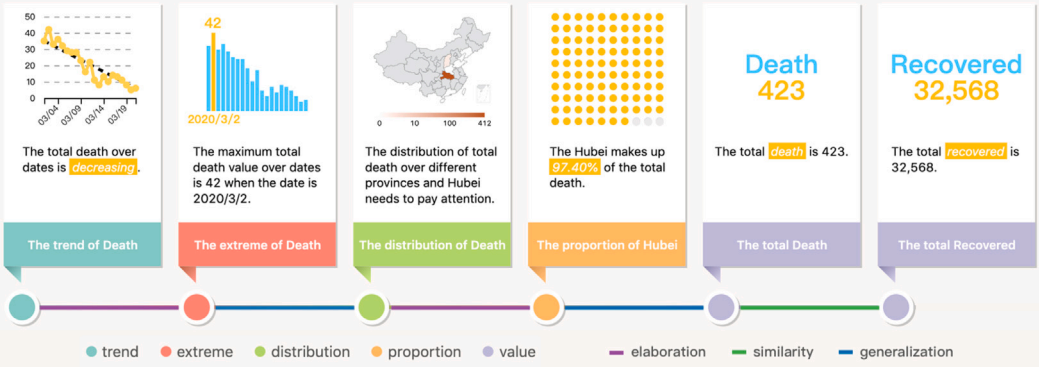



**Fig. 9.** A full-automatically generated visual data story created by Calliope, consisting of six data facts. Image from Shi et al. [35].

current state of reasoning capabilities in foundation models can make it difficult to specify exact requirements when generating images directly. These models may struggle to understand and incorporate complex constraints or design principles, leading to visualisations that may not effectively convey the intended information or insights.

## 4. Foundation models for visualisation observation

Visualisation observation is the process through which analysts make a finding about their data through visualisation [15]. Analysts can engage in this process either by examining the visualisation itself or by observing analytic input from the system. LLMs can facilitate visualisation observation, as the system is able to provide more expressive and intuitive feedback to users, as they can generate flexible human-like natural language. There are also opportunities for MLLMs to support this process as they can directly observe the visualisation image themselves, potentially aiding the user in examining the visualisation to create a finding.

### 4.1. Existing systems

Prior to LLMs, NL was already being used as an output modality to communicate findings and explanations to users in VA systems. Early research exploited rule-based NL generation techniques to communicate insights to users. Systems like Calliope [35] (Fig. 9) and Voder [36] produce data facts, NL descriptions of statistical facts about the data used. These systems typically rely on a limited set of predefined fact types and use template-based generation methods, limiting the diversity of language that they can produce. Similarly, some systems generate captions or titles for visualisations using template-based approaches [37].

There has also been research into rule-based Chart Question Answering (CQA) systems, which enable users to ask questions about charts. Kim et al. [38] developed a pipeline that interprets user questions referencing chart elements, translating them into table-based queries to retrieve answers. The system then generates visual explanations for the user input by linking the queried data to corresponding visual elements (Fig. 10). They also produced a dataset consisting of 52 charts and 629 questions relating to those visualisations. Hoque et al. provide a comprehensive review of rule-based CQA systems and datasets [39]. CQA systems have moved beyond rule-based approaches to leveraging (M)LLMs, but the corpora of visualisations and NL produced are still relevant to current and future research.

LLMs have the capability to generate more flexible and diverse NL compared to rule-based systems. Some recent systems have leveraged LLMs to generate individual facts or annotations to supplement visualisations. For example, the InkSight [40] system uses an LLM to generate annotations from user sketches on visualisations (Fig. 11). LLMs have also been used to construct entire narratives. DATATALES [41] is a prototype system that leverages an LLM to help users author data-driven articles based on a given chart and user annotations. Hoque and Islam [3] offer a comprehensive review of systems using NLG techniques in visualisation.

### 4.2. Opportunities and challenges

While these LLM-based approaches demonstrate the potential for more flexible and contextually relevant NL in VA systems, they still face limitations. LLMs struggle with analytical reasoning and may generate text that is fluent but not always accurate to the underlying data. For this reason, many LLM-based NLG approaches in VA still rely on templates, limiting their ability to fully leverage the flexibility of LLMs. For example, the InkSight [40] system still relies on a template-based approach — the statistical facts about the data are generated separately, and the LLM is only used to generate more fluent NL.

These LLM-based techniques generally interact with the underlying code or data to generate findings. For example, InkSight is actually observing the underlying data related to the selected part of the visualisation, not the visualisation itself. MLLMs could be leveraged to interpret visualisations directly. MLLMs have not yet been integrated into a VA system for this purpose, but there has been research examining the visualisation capabilities of these models and fine-tuning them further for visualisation tasks. Bendeck and Stasko [42] evaluate the visualisation literacy of the MLLM GPT4-V using the Visualisation Literacy Assessment Test (VLAT) [43], a set of multiple choice questions about visualisations originally designed to assess humans. They also assess the model's ability to answer questions about deceptive visualisations — for example with a truncated or inverted axis — using a custom dataset (Fig. 12). Similarly, Lo and Qu [44] benchmark the ability of MLLMs to detect misleading visualisations using a small dataset of real-world visualisations. In both papers, it was found that these models have some capabilities in visualisation tasks but still face problems in tasks requiring analytical reasoning. To try and alleviate some of the problems these models have with visualisation tasks, Zeng et al. [45] use instruction fine-tuning, which does increase the performance of models on a dataset similar to the VLAT (Fig. 13).

There are limitations with the datasets used in current research to benchmark the visualisation literacy and reasoning of MLLMs. The datasets are often contrived and do not accurately represent real-world VA tasks. Previous research has explored reverse engineering both visualisation specifications and the underlying data from real-world bitmap visualisation images using pipeline and neural network-based techniques [46,47]. Benchmarking or training MLLMs on real-world visualisations like these — such as research paper figures and illustrations in the news and popular media across a broad range of domains — could harness a large, untapped body of knowledge to link data, visualisation design, and the findings they supported.

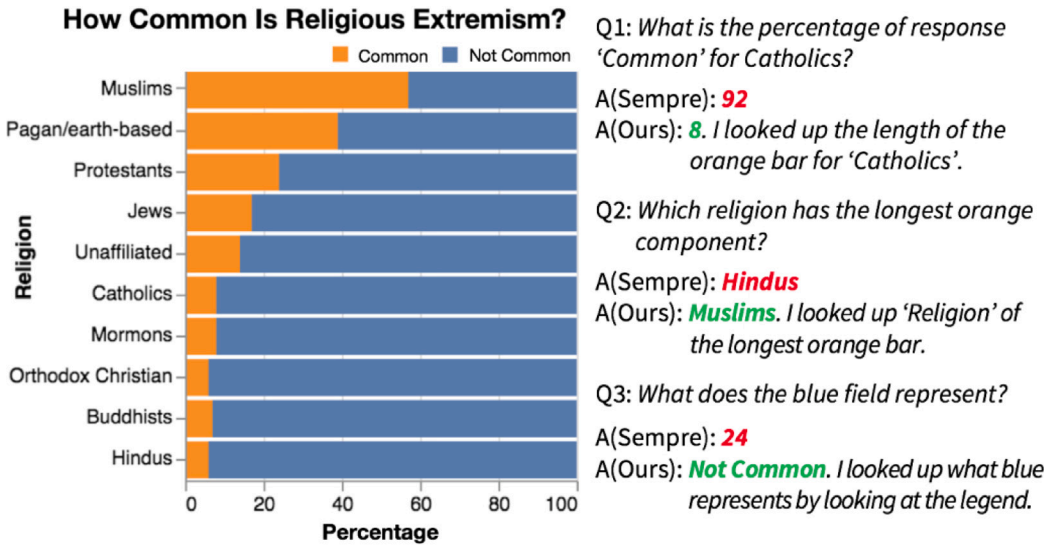



**Fig. 10.** Example questions from Kim et al.'s rule-based CQA pipeline, which interprets user questions about chart elements and generates visual explanations. Image from Kim et al. [38].

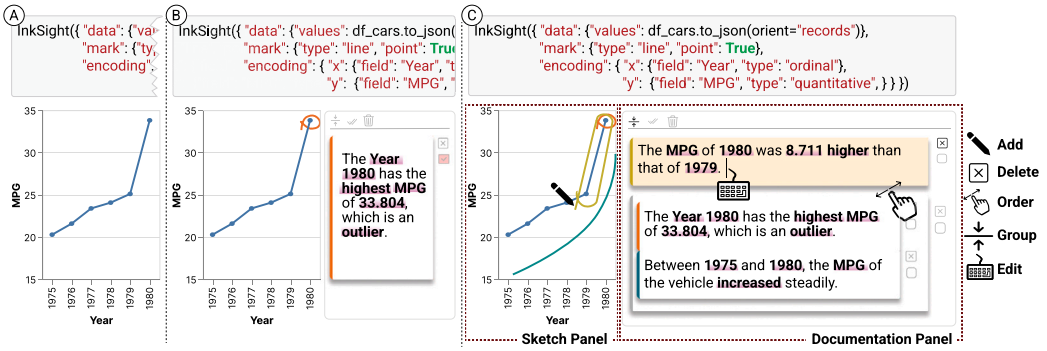



**Fig. 11.** A demonstration of the InkSight using a user-drawn sketch on a visualisation to generate corresponding natural language annotations using an LLM. Image from Lin et al. [40].

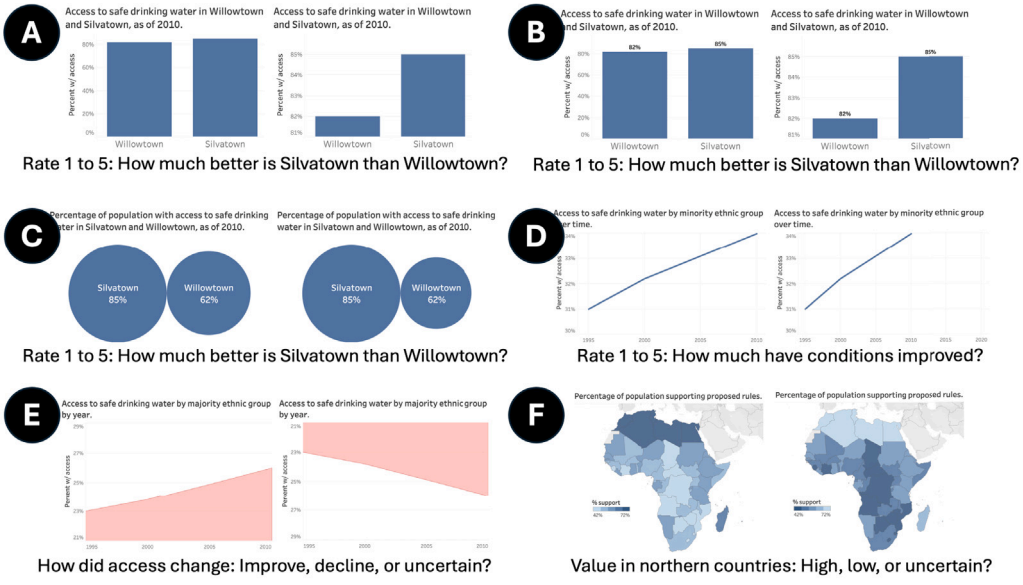



**Fig. 12.** Examples from Bendeck and Stasko's deceptive visualisation design dataset, showing the control visualisation on the left, the deceptive visualisation on the right, and a simplified version of the question asked to GPT-4V underneath. Image from Bendeck and Stasko [42].

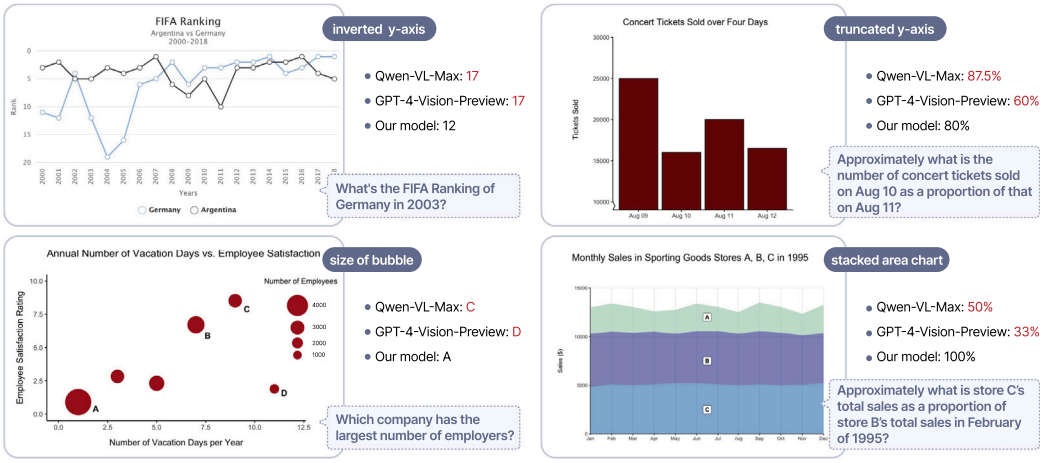



**Fig. 13.** Sample results from Zeng et al.'s fine-tuned MLLM on deceptive visualisation chart question answering compared with existing state-of the art MLLMs. Image from Zeng et al. [45].

## 5. Foundation models for visualisation manipulation

A visualisation manipulation action involves interacting with an existing visualisation without altering how the data is mapped or creating a new visualisation, for example, zooming, highlighting, or filtering [15]. Observation and manipulation tasks are closely related, often working together in an iterative process — a user observes a finding and then uses this to inform their manipulation of the visualisation. Some systems can also manipulate the visualisation themselves to convey an observed finding. This section explores how foundation model enhanced systems support visualisation manipulation.

### 5.1. Existing systems

Current systems highlight the coupling between observation and manipulation, with work demonstrating how foundational models can mediate this process. For example, the InkSight system [40] allows users directly manipulate a visualisation through sketching to highlight data they find interesting, and the system responds with an observation about the selected data (Fig. 11). As discussed, there are limitations in the observations created by this system as the system is actually using the underlying data selected by the sketch, rather than interpreting the visualisation itself. The manipulation is also limited as the only interaction modality is sketching. There remains the opportunity to integrate more interaction modalities, allowing the user to communicate with the system more expressively. Despite these limitations, the InkSight system demonstrates how LLMs can be leveraged for both manipulation and observation support in VA systems.

There are also tools in which it is the system that manipulates the visualisation rather than the user. The HiChart system [48] highlights parts of a visualisation based on text input, such as a sentence or paragraph from a related article (Fig. 14). The system achieves this by first reverse-engineering the Vega-Lite specifications from the visualisation image and then using the specification to highlight areas related to the provided text span. Although the system uses the deep-learning-based ChartOCR [49] to recover the specifications the system is actually manipulating the visualisation mapping, not the visualisation itself. There remains the opportunity to leverage MLLMs to manipulate visualisations directly, without the need for a bridge.

### 5.2. Opportunities and challenges

The integration of foundation models into VA systems has led to a growing focus on conversational interfaces that rely primarily on NL input. While this approach has shown promise in enabling more intuitive and accessible interactions, manipulating a visualisation through NL alone is sometimes not ideal due to the inherent ambiguity of NL. Users may struggle to articulate their intentions clearly, leading to misinterpretations or unintended outcomes. To address this, there is a significant opportunity to combine language-based input with other well-established interaction modalities, such as direct manipulation, touch, or gesture-based inputs. These modalities allow users to express themselves more precisely, offering greater control and specificity when manipulating visualisations.

Direct manipulation techniques which underlie WIMP (Windows, Icons, Menus, Pointer) interfaces have been fundamental to supporting interaction in VA systems to date and are likely to continue playing a significant role due to their effectiveness in supporting certain tasks. For example, selecting data points, zooming into specific regions of a visualisation, or adjusting parameters through sliders can be more easily accomplished through direct manipulation than with language alone. By integrating NL alongside these traditional interaction modalities, VA systems can offer a more powerful and flexible user experience that combines the best of both interaction paradigms.

As mentioned, some existing systems have already begun exploring the integration of various interaction modalities alongside natural language input in LLM-based systems, such as the InkSight system (Fig. 11) [40] which combines sketch-based input with natural language. Similarly, foundation models have demonstrated capabilities in speech recognition [50] and gesture recognition [51], opening up new possibilities for integrating these modalities more widely into VA systems. By leveraging multiple input channels simultaneously, systems can gain a more comprehensive understanding of the user's intent and level of understanding throughout the analysis process.

However, there are also limitations to consider when integrating multiple modalities alongside NL. One challenge is the complexity of processing multiple input modalities in real-time. Systems would need to be able to capture and interpret several different input modalities simultaneously, whilst also handling any conflicts or ambiguities in these different channels. Systems may need specialised hardware to capture certain types of inputs, which may not be feasible in application. Further research is required to explore how to effectively implement foundation model based multimodal interaction techniques and their impact on user experience.

There also remains a significant opportunity to enhance user visualisation manipulation actions through guidance. Existing VA systems primarily react to user manipulation in a direct and task-specific manner. In Ceneda's model, guidance in VA incorporates interaction history [18]. Foundation models have the potential to transform VA systems by leveraging this rich history — encompassing user queries, interaction behaviours, and the dataset itself — to provide tailored guidance. Systems could proactively suggest relevant action or even

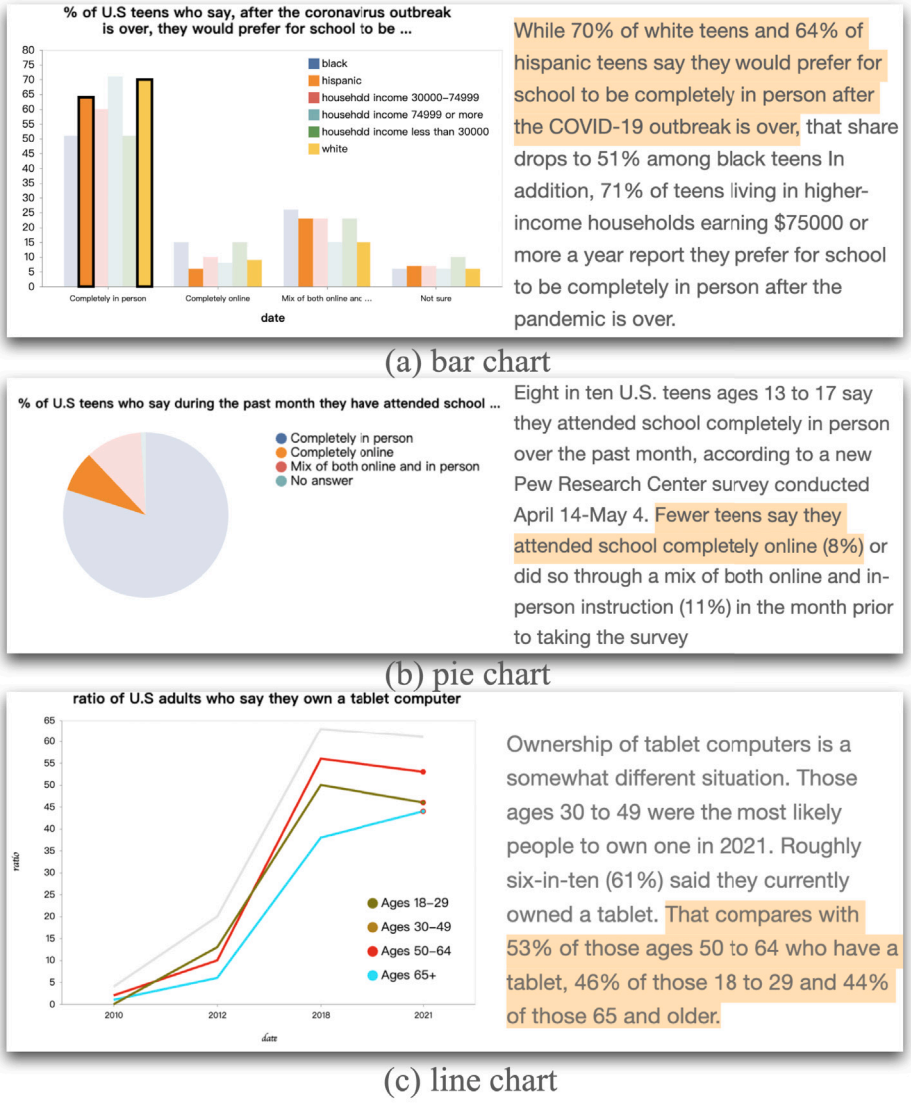



**Fig. 14.** Examples of visualisations highlighted by HiChart in relation to user-selected text. Image from Yang et al. [48].

conduct exploratory analyses on the user's behalf. They could adapt the level of guidance provided based on the user's knowledge gap and the complexity of the task at hand, rather than reactively following user instructions.

Some recent systems are beginning to move in this direction by explicitly considering users' analytical goals. For example, LEVA [52] incorporates LLMs to support users across different stages of analysis, from onboarding and exploration to summarisation, aiming to align system behaviour with the user's broader analytical intent. Similarly, PhenoFlow [53] leverages LLMs within a VA interface to reduce cognitive load and help clinicians iteratively refine their analysis goals in complex clinical data spaces. By moving beyond reactive support in this way, future VA systems could offer more intelligent, context-aware proactive guidance.

Sperrle et al. [20] explore the concept of co-adaptive guidance in mixed-initiative VA systems, incorporating guidance from the user to the system. With their ability to both generate and understand language, LLMs have the capability to facilitate this two-way guidance. Through asking questions, providing options, and eliciting user feedback, LLMs can learn from users and adapt their behaviour accordingly. This bi-directional communication can lead to a more collaborative and adaptive VA experience, where both the user and the system continuously learn from each other. Systems need to strike the right balance between proactive guidance and user autonomy.

## 6. Discussion and conclusion

We have presented an exploration of how foundation models have the potential to enhance visualisation-related processes in VA: visual mapping, visualisation observation, and visualisation manipulation. This paper has examined existing systems that integrate these models, discussing their strengths and limitations. We have also identified opportunities for future advancements, particularly in enabling more guiding, adaptive, and multimodal systems that support users in their analytical tasks.

Visual mapping, visualisation observation, and visualisation manipulation do not happen in isolation but are interconnected, iterative processes that form part of VA [15]. Leveraging natural language capability could enable future VA systems to facilitate seamless transitions across these processes. Multimodal models could enhance this by allowing systems to interpret and interact with visualisations directly, as humans do. This would create a shared visual context between the user and the system, fostering collaboration in analysis. Intuitive interactions spanning text, visualisations, and other modalities facilitate rich, collaborative exchanges between users and systems.

While the integration of foundation models into VA systems offers opportunities, there are still significant challenges that need to be addressed. These models often lack alignment with domain-specific VA

expertise, leading to outputs that fail to adhere to established visualisation principles or best practices. Future research should focus on developing approaches to align model outputs with human preferences and established VA principles.

Evaluating these systems also presents difficulties. Traditional NLP metrics are effective for automatic evaluations but fail to capture the nuanced requirements of VA tasks, such as user experience [44]. Conversely, design-focused evaluations and human studies provide rich insights but lack scalability. There is a need for scalable evaluation frameworks tailored to VA that balance automatic and nuanced assessments of system performance and user experience.

Additionally, foundation models often struggle with complex analytical reasoning, which limits their ability to effectively support VA workflows [54]. To address this, future research should prioritise the development of robust benchmarking datasets to assess and enhance reasoning capabilities in VA-specific contexts. These datasets should reflect real-world tasks and challenges, providing a foundation for improving model capabilities and enabling a more realistic evaluation of their reasoning processes in context.

The integration of foundation models into VA presents a significant opportunity to transform how users engage with data. We hope that the directions and opportunities highlighted in this paper will inspire further research, advancing the development of VA systems that act as collaborative analytical partners.

## CRediT authorship contribution statement

**Maeve Hutchinson:** Writing – review & editing, Writing – original draft. **Radu Jianu:** Writing – review & editing, Writing – original draft, Supervision. **Aidan Slingsby:** Writing – review & editing, Writing – original draft, Supervision. **Pranava Madhyastha:** Writing – review & editing, Writing – original draft, Supervision.

## Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Data availability

No data was used for the research described in the article.

## References


[1] Shen L, Shen E, et al. Towards natural language interfaces for data visualization: A survey. IEEE TVCG 2022;29(6):3121–44. http://dx.doi.org/10.1109/TVCG.2022.3148007.

[2] Voigt H, Alacam O, et al. The why and the how: A survey on natural language interaction in visualization. In: Carpuat M, de Marneffe M-C, Meza Ruiz IV, editors. Proc NAACL-HLT. Seattle: ACL; 2022, p. 348–74. http://dx.doi.org/10.18653/v1/2022.naacl-main.27.

[3] Hoque E, Islam MS. Natural language generation for visualizations: State of the art, challenges and future directions. Comput Graph Forum 2025;44(1):e15266. http://dx.doi.org/10.1111/cgf.15266, URL https://onlinelibrary.wiley.com/doi/abs/10.1111/cgf.15266.

[4] Wang Q, Chen Z, Wang Y, Qu H. A survey on ML4vis: Applying machine learning advances to data visualization. IEEE Trans Vis Comput Graphics 2022;28(12):5134–53. http://dx.doi.org/10.1109/TVCG.2021.3106142.

[5] Wu A, Wang Y, et al. AI4VIS: Survey on artificial intelligence approaches for data visualization. IEEE TVCG 2021;28(12):5049–70. http://dx.doi.org/10.1109/TVCG.2021.3099002.

[6] Ye Y, Hao J, Hou Y, Wang Z, Xiao S, Luo Y, Zeng W. Generative AI for visualization: State of the art and future directions. Vis Informatics 2024;8(2):43–66. http://dx.doi.org/10.1016/j.visinf.2024.04.003, URL https://www.sciencedirect.com/science/article/pii/S2468502X24000160.

[7] Yang W, Liu M, Wang Z, Liu S. Foundation models meet visualizations: Challenges and opportunities. Comput Vis Media 2024;10(3):399–424. http://dx.doi.org/10.1007/s41095-023-0393-x, URL DOI: 10.1007/s41095-023-0393-x.

[8] Bommasani R, et al. On the opportunities and risks of foundation models. 2022, http://dx.doi.org/10.48550/arXiv.2108.07258.

[9] Brown TB, et al. Language models are few-shot learners. 2020, http://dx.doi.org/10.48550/arXiv.2005.14165.

[10] Touvron H, Lavril T, Izacard G, Martinet X, Lachaux M-A, Lacroix T, Rozière B, Goyal N, Hambro E, Azhar F, Rodriguez A, Joulin A, Grave E, Lample G. LLaMA: Open and efficient foundation language models. 2023, http://dx.doi.org/10.48550/arXiv.2302.13971.

[11] Yin S, Fu C, Zhao S, Li K, Sun X, Xu T, Chen E. A survey on multimodal large language models. 2024, http://dx.doi.org/10.48550/arXiv.2306.13549.

[12] OpenAI, Achiam J, et al. GPT-4 technical report. 2024, http://dx.doi.org/10.48550/arXiv.2303.08774.

[13] Team G, Anil R, et al. Gemini: A family of highly capable multimodal models. 2024, http://dx.doi.org/10.48550/arXiv.2312.11805.

[14] Thomas JJ, Cook KA, editors. Illuminating the path: The research and development agenda for visual analytics. Los Alamitos, Calif: IEEE Computer Soc; 2005.

[15] Sacha D, Stoffel A, Stoffel F, Kwon BC, Ellis G, Keim DA. Knowledge generation model for visual analytics. IEEE Trans Vis Comput Graphics 2014;20(12):1604–13. http://dx.doi.org/10.1109/TVCG.2014.2346481.

[16] Horvitz E. Principles of mixed-initiative user interfaces. In: Proc. CHI. Pittsburgh: ACM Press; 1999, p. 159–66. http://dx.doi.org/10.1145/302979.303030.

[17] Oppermann R, editor. Adaptive user support: Ergonomic design of manually and automatically adaptable software. USA: L. Erlbaum Associates Inc.; 1994.

[18] Ceneda D, Gschwandtner T, May T, Miksch S, Schulz H-J, Streit M, Tominski C. Characterizing guidance in visual analytics. IEEE TVCG 2016;23(1):111–20. http://dx.doi.org/10.1109/TVCG.2016.2598468.

[19] Ceneda D, Gschwandtner T, Miksch S. A review of guidance approaches in visual data analysis: A multifocal perspective. Comput Graph Forum 2019;38(3):861–79. http://dx.doi.org/10.1111/cgf.13730.

[20] Sperrle F, Jeitler A, Bernard J, Keim D, El-Assady M. Co-adaptive visual data analysis and guidance processes. Comput Graph 2021;100:93–105. http://dx.doi.org/10.1016/j.cag.2021.06.016.

[21] Lee B, Isenberg P, Riche NH, Carpendale S. Beyond mouse and keyboard: Expanding design considerations for information visualization interactions. IEEE TVCG 2012;18(12):2689–98. http://dx.doi.org/10.1109/TVCG.2012.204.

[22] Srinivasan A, Dontcheva M, Adar E, Walker S. Discovering natural language commands in multimodal interfaces. In: Proc. IUI. Marina del Ray: ACM; 2019, p. 661–72. http://dx.doi.org/10.1145/3301275.3302292.

[23] Narechania A, Srinivasan A, Stasko J. NL4dv: A toolkit for generating analytic specifications for data visualization from natural language queries. IEEE TVCG 2020;27(2):369–79. http://dx.doi.org/10.1109/TVCG.2020.3030378.

[24] Gao T, Dontcheva M, Adar E, Liu Z, Karahalios KG. DataTone: Managing ambiguity in natural language interfaces for data visualization. In: Proc. UIST. New York: ACM; 2015, p. 489–500. http://dx.doi.org/10.1145/2807442.2807478.

[25] Setlur V, Battersby SE, Tory M, Gossweiler R, Chang AX. Eviza: A natural language interface for visual analysis. In: Proc. UIST. New York: ACM; 2016, p. 365–77. http://dx.doi.org/10.1145/2984511.2984588.

[26] Luo Y, Qin X, Tang N, Li G. DeepEye: Towards automatic data visualization. In: IEEE ICDE. Paris: IEEE; 2018, p. 101–12. http://dx.doi.org/10.1109/ICDE.2018.00019.

[27] Oppermann M, Kincaid R, Munzner T. VizCommender: Computing text-based similarity in visualization repositories for content-based recommendations. IEEE TVCG 2020;27(2):495–505. http://dx.doi.org/10.1109/TVCG.2020.3030387.

[28] Maddigan P, Susnjak T. Chat2VIS: Generating data visualizations via natural language using ChatGPT, codex and GPT-3 large language models. IEEE Access 2023;11:45181–93. http://dx.doi.org/10.1109/ACCESS.2023.3274199.

[29] Dibia V. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In: Bollegala D, Huang R, Ritter A, editors. Proc. ACL. Toronto: ACL; 2023, p. 113–26. http://dx.doi.org/10.18653/v1/2023.acl-demo.11.

[30] Tian Y, Cui W, Deng D, Yi X, Yang Y, Zhang H, Wu Y. Chartgpt: Leveraging LLMs to generate charts from abstract natural language. IEEE TVCG 2024;1–15. http://dx.doi.org/10.1109/TVCG.2024.3368621.

[31] Shi C, Cui W, Liu C, Zheng C, Zhang H, Luo Q, Ma X. NL2color: Refining color palettes for charts with natural language. IEEE Trans Vis Comput Graphics 2024;30(1):814–24. http://dx.doi.org/10.1109/TVCG.2023.3326522.

[32] Xiao S, Huang S, et al. Let the chart spark: Embedding semantic context into chart with text-to-image generative model. IEEE TVCG 2023;30(1):284–94. http://dx.doi.org/10.1109/TVCG.2023.3326913.

[33] Schetinger V, Di Bartolomeo S, El-Assady M, McNutt A, Miller M, Passos JPA, Adams JL. Doom or deliciousness: Challenges and opportunities for visualization in the age of generative models. Comput Graph Forum 2023;42(3):423–35. http://dx.doi.org/10.1111/cgf.14841.

[34] Bostock M, Ogievetsky V, Heer J. $D^3$ data-driven documents. IEEE Trans Vis Comput Graphics 2011;17(12):2301–9. http://dx.doi.org/10.1109/TVCG.2011.185.

[35] Shi D, Xu X, Sun F, Shi Y, Cao N. Calliope: Automatic visual data story generation from a spreadsheet. IEEE TVCG 2020;27(2):453–63. http://dx.doi.org/10.1109/TVCG.2020.3030403.

[36] Srinivasan A, et al. Augmenting visualizations with interactive data facts to facilitate interpretation and communication. IEEE TVCG 2018;25(1):672–81. http://dx.doi.org/10.1109/TVCG.2018.2865145.

[37] Hsu T-Y, Giles CL, Huang T-H. SciCap: Generating captions for scientific figures. In: EMNLP. Punta Cana: ACL; 2021, p. 3258–64. http://dx.doi.org/10.18653/v1/2021.findings-emnlp.277.

[38] Kim DH, Hoque E, Agrawala M. Answering questions about charts and generating visual explanations. In: Proc. CHI. CHI '20, New York: ACM; 2020, p. 1–13. http://dx.doi.org/10.1145/3313831.3376467.

[39] Hoque E, Kavehzadeh P, Masry A. Chart question answering: State of the art and future directions. Comput Graph Forum 2022;41(3):555–72. http://dx.doi.org/10.1111/cgf.14573.

[40] Lin Y, Li H, et al. InkSight: Leveraging sketch interaction for documenting chart findings in computational notebooks. IEEE TVCG 2023;30(1):944–54. http://dx.doi.org/10.1109/TVCG.2023.3327170.

[41] Sultanum N, Srinivasan A. DATATALES: Investigating the use of large language models for authoring data-driven articles. In: 2023 IEEE VIS. 2023, p. 231–5. http://dx.doi.org/10.1109/VIS54172.2023.00055.

[42] Bendeck A, Stasko J. An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks. IEEE Trans Vis Comput Graphics 2024;1–11. http://dx.doi.org/10.1109/TVCG.2024.3456155.

[43] Lee S, Kim S-H, Kwon BC. VLAT: Development of a visualization literacy assessment test. IEEE Trans Vis Comput Graphics 2017;23(1):551–60. http://dx.doi.org/10.1109/TVCG.2016.2598920.

[44] Lo LY-H, Qu H. How good (or bad) are LLMs at detecting misleading visualizations? IEEE Trans Vis Comput Graphics 2024;1–10. http://dx.doi.org/10.1109/TVCG.2024.3456333.

[45] Zeng X, Lin H, Ye Y, Zeng W. Advancing multimodal large language models in chart question answering with visualization-referenced instruction tuning. IEEE Trans Vis Comput Graphics 2024;1–11. http://dx.doi.org/10.1109/TVCG.2024.3456159.

[46] Poco J, Heer J. Reverse-engineering visualizations: Recovering visual encodings from chart images. Comput Graph Forum 2017;36(3):353–63. http://dx.doi.org/10.1111/cgf.13193.

[47] Jung D, Kim W, Song H, Hwang J-i, Lee B, Kim B, Seo J. ChartSense: Interactive data extraction from chart images. In: Proc. CHI. CHI '17, New York: ACM; 2017, p. 6706–17. http://dx.doi.org/10.1145/3025453.3025957.

[48] Yang C, Fan R, Tang N, Zhang M, Zhao X, Fan J, Du X. Pay ‘‘attention" to chart images for what you read on text. In: Companion of the 2023 international conference on management of data. SIGMOD '23, New York, NY, USA: Association for Computing Machinery; 2023, p. 111–4. http://dx.doi.org/10.1145/3555041.3589714.

[49] Luo J, Li Z, Wang J, Lin C-Y. Chartocr: Data extraction from charts images via a deep hybrid framework. In: 2021 IEEE winter conference on applications of computer vision (WACV). 2021, p. 1916–24. http://dx.doi.org/10.1109/WACV48630.2021.00196.

[50] Hu Y, Chen C, Yang C-HH, Li R, Zhang C, Chen P-Y, Chng E. Large language models are efficient learners of noise-robust speech recognition. 2024, http://dx.doi.org/10.48550/arXiv.2401.10446, ArXiv Preprint.

[51] Wicke P. Probing language models' gesture understanding for enhanced human-AI interaction. 2024, http://dx.doi.org/10.48550/arXiv.2401.17858.

[52] Zhao Y, Zhang Y, Zhang Y, Zhao X, Wang J, Shao Z, Turkay C, Chen S. LEVA: Using large language models to enhance visual analytics. IEEE Trans Vis Comput Graphics 2025;31(3):1830–47. http://dx.doi.org/10.1109/TVCG.2024.3368060, URL https://ieeexplore.ieee.org/document/10458347.

[53] Kim J, Lee S, Jeon H, Lee K-J, Bae H-J, Kim B, Seo J. PhenoFlow: A human-LLM driven visual analytics system for exploring large and complex stroke datasets. IEEE Trans Vis Comput Graphics 2025;31(1):470–80. http://dx.doi.org/10.1109/TVCG.2024.3456215, URL https://ieeexplore.ieee.org/abstract/document/10689638.

[54] McCoy RT, Yao S, et al. Embers of autoregression: Understanding large language models through the problem they are trained to solve. 2023, http://dx.doi.org/10.48550/arXiv.2309.13638.