Integrating LLM, VLM, and Text-to-Image Models for Enhanced Information Graphics: A Methodology for Accurate and Visually Engaging Visualizations

Chao-Ting Chen[1] and **Hen-Hsen Huang**[2]

[1]Department of Computer Science, National Chengchi University, Taipei, Taiwan
[2]Institute of Information Science, Academia Sinica, Taipei, Taiwan
chaotingchen10@gmail.com, hhhuang@iis.sinica.edu.tw

Abstract

This study presents an innovative approach to the creation of information graphics, where the accuracy of content and aesthetic appeal are of paramount importance. Traditional methods often struggle to balance these two aspects, particularly in complex visualizations like phylogenetic trees. Our methodology integrates the strengths of Large Language Models (LLMs), Vision Language Models (VLMs), and advanced text-to-image models to address this challenge. Initially, an LLM plans the layout and structure, employing Mermaid—a JavaScript-based tool that uses Markdown-like scripts for diagramming—to establish a precise and structured foundation. This structured script is crucial for ensuring data accuracy in the graphical representation. Following this, text-to-image models are employed to enhance the vector graphic generated by Mermaid, adding rich visual elements and enhancing overall aesthetic appeal. The integration of text-to-image models is a key innovation, enabling the creation of graphics that are not only informative but also visually captivating. Finally, a VLM performs quality control, ensuring that the visual enhancements align with the informational accuracy. This comprehensive approach effectively combines the accuracy of structured data representation, the creative potential of text-to-image models, and the validation capabilities of VLMs. The result is a new standard in information graphic creation, suitable for diverse applications ranging from education to scientific communication, where both information integrity and visual engagement are essential.

1 Introduction

In the realm of artificial intelligence, advancements in natural language processing and image generation have paved the way for innovative applications across various domains [Zhang *et al.*, 2023]. Large language models (LLM) [Schulman *et al.*, 2022; Achiam *et al.*, 2023] have demonstrated remarkable proficiency in understanding and generating human-like text, while text-to-image models such as DALL-E [Shi *et al.*, 2020] and Midjourney[1] have captivated audiences with its ability to generate diverse and creative images from textual descriptions. However, a significant challenge remains in bridging the gap between textual and visual representations, particularly in the creation of information graphics [Huang *et al.*, 2024; Dibia, 2023].

In the evolving landscape of data visualization and information communication, the challenge of accurately representing complex data structures while maintaining visual appeal is paramount. Information graphics, such as phylogenetic trees, play a crucial role in various fields, including biology, education, and data science. However, the existing approach of using text-to-image models often leads to a compromise between the factual correctness of the content and its aesthetic representation.

Figure 1 displays a phylogenetic tree created with DALL-E 3. While aesthetically pleasing, the content it depicts lacks accuracy and meaningfulness. Additionally, a common limitation of text-to-image models is evident here: the textual information within the figure is indecipherable, reflecting a well-known shortcoming in text-to-image generation. User control over output correctness remains challenging.

This research introduces an innovative methodology that synergizes LLMs, Vision Language Models (VLMs), and advanced text-to-image models to overcome this challenge, aiming to produce information graphics that are both visually engaging and accurate in content. Our approach begins with the use of an LLM for initial planning and layout of the information graphic. The LLM's role is to understand the intricacies of the data and conceptualize a structure that best represents it. This planning phase is critical as it sets the foundation for the accuracy and comprehensibility of the final graphic. To translate this plan into a tangible layout, we utilize Mermaid, a JavaScript-based diagramming tool. Mermaid's ability to render Markdown-inspired text definitions into dynamic diagrams makes it an ideal choice for creating structured, accurate representations of complex data relationships, such as those found in phylogenetic trees.

Once the structural foundation is laid out by the LLM and rendered by Mermaid, the methodology introduces the use of text-to-image models. These models serve to enhance the vector graphics created by Mermaid, adding a layer of vi-

[1]https://www.midjourney.com/home

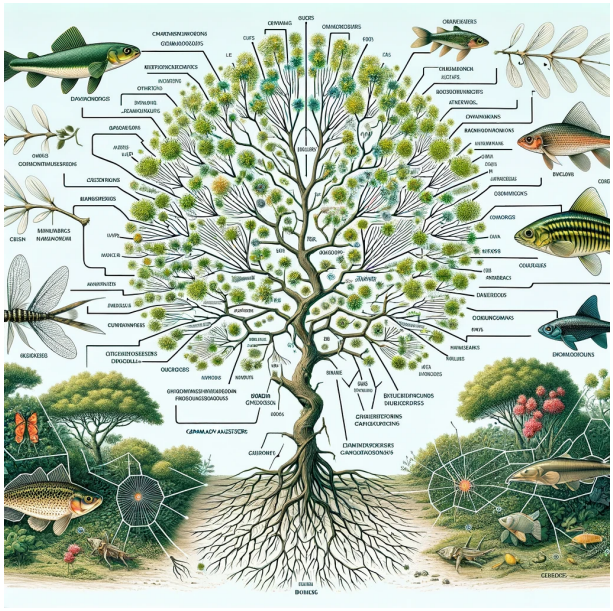

Figure 1: A phylogenetic tree generated by using DALL-E 3. Although it looks fancy, the information it depicts is non-sense and meaningless.

sual sophistication and appeal. This step is where the creative potential of AI comes into play, transforming a structurally sound diagram into a visually captivating piece of information art. The integration of text-to-image models is a key innovation of our approach, enabling us to bridge the gap between informative and aesthetically pleasing graphics.

Finally, to ensure the integrity and coherence of the enhanced graphic, a VLM is employed for quality control. This stage is crucial as it validates the accuracy of both the content and its visual representation. The VLM assesses whether the enhancements made by the text-to-image models align with the factual data and maintains the overall clarity of the information presented.

This research aims to set a new standard in the creation of information graphics. By leveraging the unique capabilities of LLMs for structure and content planning, text-to-image models for visual enhancement, and VLMs for final validation, we present a comprehensive solution that addresses the dual needs of accuracy and aesthetic appeal in data visualization. This methodology has the potential to revolutionize the way complex information is communicated, making it more accessible and engaging for a wider audience. Our contributions are threefold as follows:

1. **Innovative AI Technology Integration**: This work pioneers the integration of Large Language Models, Vision Language Models, and text-to-image models, establishing a novel approach for creating information graphics that are both accurate and visually appealing.
2. **Enhanced Accuracy in Data Representation**: The methodology significantly improves the precision of complex data visualizations, using LLMs for detailed structure planning and Mermaid for accurate graphical layouts, setting new standards in data representation accuracy.
3. **Aesthetic Enhancement of Information Graphics**: The integration of text-to-image models elevates the visual appeal of information graphics, making complex data more accessible and engaging to a wider audience without compromising informational integrity.

2 System Overview

With our system, the user can create an information graphic with a simple prompt. Then the prompt will be processed in the following four steps to yield the final outcome.

1. **Planning with LLM**: Initially, the LLM is utilized to plan the overall layout and structure of the information graphic. This involves understanding the content requirements (like the specific family of animals for a phylogenetic tree) and determining the best way to visually represent this information. In our system, GPT-4, one of the most powerful LLM so far, is served as the LLM.
2. **Script Generation in Mermaid**: The LLM then generates a script in Mermaid, a tool that uses Markdown-inspired text to create diagrams. This script will define the structure and layout of the information graphic. Mermaid is particularly suitable for this task as it's designed to handle structured data and relationships, making it ideal for things like phylogenetic trees.
3. **Enhancement with Text-to-Image Models**: Once the basic structure is in place, text-to-image models are used to enhance and illustrate the vector graphic created by Mermaid. This step adds visual appeal, making the graphic more engaging and easier to understand for a broader audience. It can add realistic textures, colors, and other visual elements that make the diagram more visually compelling. In this work, we employ DALL-E 3 as the text-to-image model.
4. **Quality Control with VLM**: Finally, a VLM checks the outcome. This step is crucial for ensuring that the visual enhancements do not distort or misrepresent the information. The VLM can verify the accuracy of both the text and the visual representation, ensuring that the final product is both informative and aesthetically pleasing. In our system, we further employ GPT-4 as the VLM for monitoring the outcomes.

Figure 2 shows an example of a phylogenetic tree being generated by our system. This process leverages the strengths of both LLMs and VLMs. The LLM ensures the accuracy and appropriateness of the content and structure, while the VLM and text-to-image models enhance the visual presentation. The result is a more accurate, informative, and visually appealing information graphic.

This method could significantly improve the way complex information is conveyed, making it more accessible and engaging, especially for educational purposes or in scientific communication. It's a promising direction for the development of AI-assisted graphic design and data visualization.

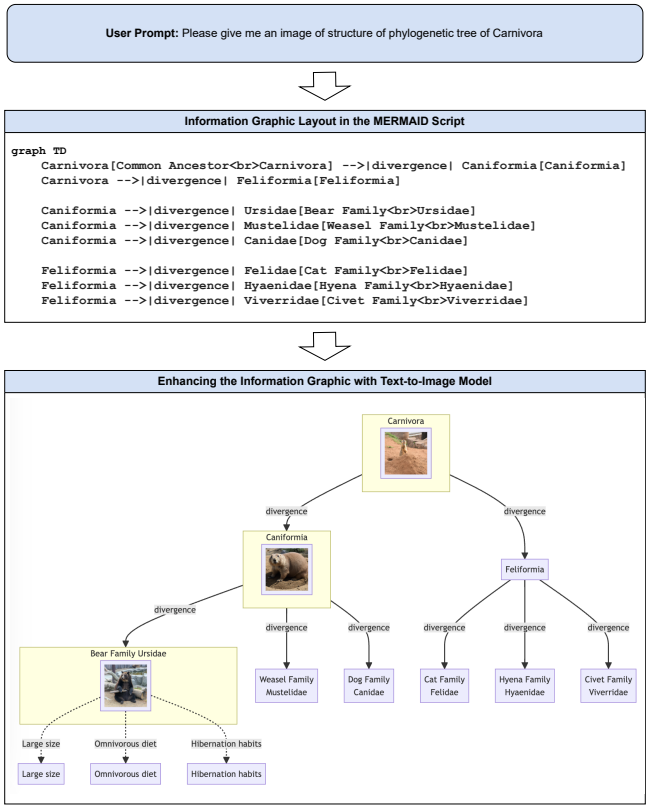

Figure 2: Example of an information graphic of the structure of phylogenetic tree of Carnivora being generated by our system. LLM generates a plan for the user's intention in the Mermaid script, and the slots are further illustrated by the text-to-image model, DALL-E 3.

3 Technical Details

Our system processes the raw Mermaid script generated by the LLM and incorporates illustrative enhancements, as detailed below.

- **Mermaid Parsing**: The parsing module is responsible for extracting and interpreting the Mermaid syntax from the LLM output. It identifies and segregates text segments intended for diagrammatic representation, focusing on the structure and semantics inherent in the Mermaid language.
- **Validation and Correction**: Once extracted, the Mermaid script is subjected to a rigorous validation process. This step checks for syntactic accuracy and logical coherence, ensuring that the diagram's structure aligns with the intended representation. Any errors or inconsistencies detected during this phase are corrected to maintain the integrity of the diagram.
- **Mermaid Rendering**: Post-validation, the Mermaid renderer translates the script into a visual diagram. This step involves converting the text-based instructions into a graphical representation, effectively bringing the script's encoded data and relationships to visual life.
- **Illustration Enhancement**: Key elements within the rendered Mermaid diagram are identified for visual augmentation. Using contextually derived prompts from the diagram, text-to-image models like DALL-E 3 are employed to generate and integrate detailed, relevant illustrations into specific sections of the diagram, thereby enhancing its informational and aesthetic value.
- **Quality Control and Assurance**: An independent GPT-4 module performs a final evaluation, focusing on the integrity of the entire process—from Mermaid generation to illustration integration. This quality control ensures the final output is not only visually appealing but also maintains accuracy and relevance to the user's original input and intent.

This multi-stage approach ensures that the final information graphic is both visually compelling and accurately represents the data or relationships intended by the user.

4 Applications and Use Cases

Our innovative approach enhances information representation, making complex data more accessible and engaging. Key applications and use cases include:

- **Educational Material Creation**: Enhances teaching by generating visually engaging diagrams and flowcharts for complex subjects.
- **Scientific Research and Publication**: Simplifies the illustration of complex scientific concepts for research publications and presentations.
- **Business Analytics and Reporting**: Transforms data into intuitive visual representations for clearer business decision-making and reporting.
- **Technical Documentation and Manuals**: Automates the creation of accurate diagrams and illustrations for user manuals and technical guides.
- **Interactive Media and Communication**: Produces tailored visuals for digital media, enhancing engagement in articles and marketing content.
- **Creative Industries**: Assists artists and designers in prototyping visual concepts rapidly, streamlining the creative process.

5 Conclusion

This work demonstrates a novel system that synergizes LLM with text-to-image models to revolutionize the generation of information graphics. By seamlessly integrating Mermaid script parsing and illustrative enhancements through AI models, we have demonstrated a powerful tool that significantly improves the accuracy and visual appeal of complex data representations. Our system[2] finds diverse applications across educational materials, scientific research, business analytics, technical documentation, interactive media, accessibility, and creative industries, proving its versatility and effectiveness.

[2]Please visit our demo site https://mermaid-gpt.vercel.app/

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Dibia, 2023] Victor Dibia. Lida: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. *arXiv preprint arXiv:2303.02927*, 2023.

[Huang *et al.*, 2024] Qirui Huang, Min Lu, Joel Lanir, Dani Lischinski, Daniel Cohen-Or, and Hui Huang. Graphimind: Llm-centric interface for information graphics design, 2024.

[Schulman *et al.*, 2022] John Schulman, Barret Zoph, Christina Kim, Jacob Hilton, Jacob Menick, Jiayi Weng, Juan Felipe Ceron Uribe, Liam Fedus, Luke Metz, Michael Pokorny, et al. Chatgpt: Optimizing language models for dialogue. *OpenAI blog*, 2022.

[Shi *et al.*, 2020] Zhan Shi, Xu Zhou, Xipeng Qiu, and Xiaodan Zhu. Improving image captioning with better use of captions, 2020.

[Zhang *et al.*, 2023] Chenshuang Zhang, Chaoning Zhang, Mengchun Zhang, and In So Kweon. Text-to-image diffusion models in generative ai: A survey, 2023.