RADAR - Retrieval-Augmented Data Analysis and Representation

Allen Daniel Sunny

September 2024

1 Abstract

The transfer and visualization of information are critical to modern decision-making and data-driven processes. While traditional visualization tools like PowerBI and Tableau have standardized data representation, they are often constrained in flexibility and adaptability. To overcome these limitations, we introduce RADAR, a Retrieval-Augmented Generation (RAG)-based Large Language Model (LLM) designed to enhance the visualization process. By dynamically retrieving and integrating relevant external information, RADAR enables more flexible, real-time, and accurate data visualizations, making it highly adaptable for evolving information needs.

2 Introduction

Information transfer has been a cornerstone of progress in the modern world, shaping how we communicate, innovate, and make decisions. As technology has advanced, continuous efforts have been made to improve the speed and accuracy of transferring information. Studies such as [2] and [21] highlight the significant role that modern methods of information transfer have

played in enhancing these processes. In the early days, free-form visualizations using simple tools like Paint were common, allowing for creativity but lacking structure. Over time, the field has evolved with the introduction of standardized platforms like PowerBI and Tableau, which offer more consistent and streamlined ways to represent data. However, these tools come with inherent limitations, as their visualizations are often restricted by the tools' built-in functionalities, resulting in a lack of flexibility and adaptability [12].

This is where Natural Language to Visualization (NL2VIS) systems fill the gap. NL2VIS offers a solution by enabling users to generate visualizations from natural language inputs, making the creation of data-driven graphics more intuitive, flexible, and accessible. [24] This technology bridges the gap between understanding data and effectively presenting it, removing the technical barriers that many users face with traditional tools.[15]

Creating a chart is not just about visualizing data; it's about building a clear and meaningful representation that can be easily understood by others. Historically, this task has required the collaboration of data analysts and skilled graphic designers, with the latter crafting visually appealing and informative infographics. This synergy is essential for translating complex data into digestible insights. By moving beyond basic visualizations to sophisticated infographics, we can significantly enhance the transfer and communication of information, making it more impactful and accessible to a wider audience. [4]

3 Literature Review

Large Language Models (LLMs) have seen remarkable evolution, especially in their ability to grasp and engage with creative concepts. Early LLMs were primarily designed for straightforward tasks, but innovations in techniques like chain-of-thought prompt engineering [22] have enabled these models to process complex reasoning sequences, allowing them to tackle a wide range of creative and problem-solving challenges. As the adoption of LLMs has surged, with models like OpenAI's GPT [3] leading the way, they have become specialized in areas

such as coding, content creation, and creative problem solving. The release of Llama 3.5 [6] has further broadened the capabilities of these systems, enabling them to manage multiple tasks with greater creativity and efficiency. Additionally, models specifically designed for coding, like CodeLLama, have made the transition from natural language to machine code even more seamless.

The early history of natural language processing (NLP) in relation to NL2VIS (Natural language to Visual) can be divided into three main phases. [5], [13]

The first phase involved rule-based or symbolic systems, where the primary idea was to translate linguistic rules into visual representations, which could then generate graphs. Early systems, such as Articulate [18], DataTone [8], Eviza [16], and DeepEye [11], each had their own methodology for mapping natural language to visualizations. However, these approaches were often too complex and required developers with expertise in the specific system. A small improvement came with the integration of semantic parsers, such as NLTK [1], Automatic Named Entity Recognition (NER), and the Stanford Treebank [14], which introduced features like part-of-speech tagging, thematic analysis, and entity recognition. While some deep learning techniques were introduced during this phase, the rule-based nature of these systems made them brittle, with many hand-crafted rules and a lack of flexibility. [7]

The second phase aimed to address these limitations through the use of machine learning-based systems. New approaches aimed to achieve greater robustness flexibility and adeptness as compared to the first phase of models. [19] Approaches like ADVISor [10], ncNet [12], and RGVisNet [17] utilized machine learning techniques. Many of these platforms were underpinned by early versions of LLMs, such as BERT or BART, to improve the generation of visualizations. This phase laid the groundwork for the next major advancement in the field. The third and current phase involves systems powered entirely by large language models (LLMs), which automatically generate visualization rules based on input.[20] Examples in-

clude LIDA[5] and Chat2Vis [13], which represent the state of the art in NL2VIS by leveraging LLMs for dynamic and flexible visualization generation, eliminating the need for complex rule-based systems. The current systems also have problems due to the compute power and size of the models. Other issues include AI hallucination and a lack of model explainability, all issues that hamper the widespread adoption of these systems. [5]

One solution to the issues of traditional LLM powered systems is the RAG method. Retrieval-Augmented Generation (RAG) is a more recent approach that enhances the capabilities of large language models (LLMs) by supplying them with relevant external information during the prompting process.[23] One of the key benefits of RAG is that it generally does not require the LLM itself to be retrained, which saves time and computational resources. Additionally, RAG allows for continuous updates, meaning that new information can be easily incorporated into the retrieval process without needing to modify the core model. This makes it highly adaptable and efficient for use cases where real-time or constantly evolving information is needed.[9]. RAG based methods also constrain LLM outputs making it far more certain to return logical answers and bring down hallucination.

4 References

References

[1] arXiv:cs/0205028v1 [cs.CL] 17 may 2002.

[2] Cindy Xiong Bearfield, Lisanne van Weelden, Adam Waytz, and Steven Franconeri. Same data, diverging perspectives: The power of visualizations to elicit competing interpretations.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini

Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners.

[4] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[5] Victor Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models.

[6] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu

Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya

Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A,

Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin

Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. The llama 3 herd of models.

[7] Rania Mkhinini Gahar, Olfa Arfaoui, and Minyar Sassi Hidri. Open research issues and tools for visualization and big data analytics. 15(1):1103–1117.

[8] Tong Gao, Mira Dontcheva, Eytan Adar, Zhicheng Liu, and Karrie G. Karahalios. DataTone: Managing ambiguity in natural language interfaces for data visualization. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, pages 489–500. ACM.

[9] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey.

[10] Can Liu, Yun Han, Ruike Jiang, and Xiaoru Yuan. ADVISor: Automatic visualization answer for natural-language question on tabular data. In *2021 IEEE 14th Pacific Visualization Symposium (PacificVis)*, pages 11–20. IEEE.

[11] Yuyu Luo, Xuedi Qin, Nan Tang, and Guoliang Li. DeepEye: Towards automatic data visualization.

[12] Yuyu Luo, Nan Tang, Guoliang Li, Jiawei Tang, Chengliang Chai, and Xuedi Qin. Natural language to visualization by neural machine translation. 28(1):217–226.

[13] Paula Maddigan and Teo Susnjak. Chat2vis: Generating data visualisations via natural language using ChatGPT, codex and GPT-3 large language models.

[14] Alec Radford, Rafal Jozefowicz, and Ilya Sutskever. Learning to generate reviews and discovering sentiment.

[15] Bahador Saket, Dominik Moritz, Halden Lin, Victor Dibia, Cagatay Demiralp, and Jeffrey Heer. Beyond heuristics: Learning visualization design.

[16] Vidya Setlur, Sarah E. Battersby, Melanie Tory, Rich Gossweiler, and Angel X. Chang. Eviza: A natural language interface for visual analysis. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 365–377. ACM.

[17] Yuanfeng Song, Xuefang Zhao, Raymond Chi-Wing Wong, and Di Jiang. RGVisNet: A hybrid retrieval-generation neural framework towards automatic data visualization generation. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1646–1655. ACM.

[18] Yiwen Sun, Jason Leigh, Andrew Johnson, and Sangyoon Lee. Articulate: A semi-automated model for translating natural language queries into meaningful visualizations. In Robyn Taylor, Pierre Boulanger, Antonio Krüger, and Patrick Olivier, editors, *Smart Graphics*, volume 6133, pages 184–195. Springer Berlin Heidelberg. Series Title: Lecture Notes in Computer Science.

[19] Henrik Voigt, Monique Meuschke, Kai Lawonn, and Sina Zarrieß. Challenges in designing natural language interfaces for complex visual models.

[20] Pere-Pau Vázquez. Are LLMs ready for visualization?

[21] Maggie Walter, Raymond Lovett, Bobby Maher, Bhiamie Williamson, Jacob Prehn, Gawaian Bodkin-Andrews, and Vanessa Lee. Indigenous data sovereignty in the era of big data and open data. 56(2):143–156.

[22] Xiaosong Yuan, Chen Shen, Shaotian Yan, Xiaofeng Zhang, Liang Xie, Wenxiao Wang, Renchu Guan, Ying Wang, and Jieping Ye. Instance-adaptive zero-shot chain-of-thought prompting.

[23] Xinyang Zhao, Xuanhe Zhou, and Guoliang Li. Chat2data: An interactive data analysis system with RAG, vector databases and LLMs.

[24] Chen Zhu-Tian, Yun Wang, Qianwen Wang, Yong Wang, and Huamin Qu. Towards automated infographic design: Deep learning-based auto-extraction of extensible timeline. 26(1):917–926.