Exploring Graph Structure Comprehension Ability of Multimodal Large Language Models: Case Studies

Zhiqiang Zhong
Aarhus University, Denmark
zzhong@cs.au.dk

Davide Mottin
Aarhus University, Denmark
davide@cs.au.dk

Abstract

LLMs have shown remarkable capabilities in processing various data structures, including graphs. While previous research has focused on developing textual encoding methods for graph representation, the emergence of multimodal LLMs presents a new frontier for graph comprehension. These advanced models, capable of processing both text and images, offer potential improvements in graph understanding by incorporating visual representations alongside traditional textual data. This study investigates the impact of graph visualisations on LLM performance across a range of benchmark tasks at node, edge, and graph levels. Our experiments compare the effectiveness of multimodal (*textual & visual*) approaches against purely textual graph representations. The results provide valuable insights into both the potential and limitations of leveraging visual graph modalities to enhance LLMs' graph structure comprehension abilities. [1]

1 Introduction

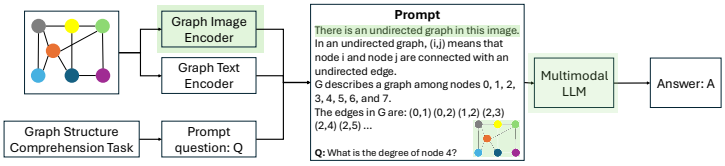

Figure 1: Overview of our framework (GAI$^+$) for graph structure comprehension using multimodal LLMs. The newly added components, compared to [1], are highlighted in green for clarity.

Recently, Large Language Models (LLMs) have revolutionised natural language processing and have been increasingly applied to diverse tasks beyond text generation and comprehension [2, 3]. One area of growing interest is the application of LLMs to *graph*-structured data, which is prevalent in numerous domains, *e.g.*, social network analysis and bioinformatics [4–6].

Conventionally, researchers have focused on developing textual encoding functions to represent graphs in a format digestible by LLMs [1, 5, 7]. These methods have shown promise, enabling LLMs to perform various graph-related tasks with increasing accuracy. While this approach has shown promise, it faces inherent limitations in capturing the full complexity of graph structures, particularly in preserving spatial relationships and global structural properties [1].

The recent emergence of multimodal LLMs marks a significant milestone in AI development [2, 3]. These advanced models, capable of processing both textual and visual information, open new avenues for enhancing machine comprehension of complex data structures. In the context of graph structure comprehension, this multimodal capability presents an exciting opportunity: *the potential to leverage visual representations of graphs alongside their textual descriptions*.

[1] The dataset and code are available at https://github.com/zhiqiangzhongddu/GaI-LoG-2024.

Z. Zhong and D. Mottin, Exploring Graph Structure Comprehension Ability of Multimodal Large Language Models: Case Studies (Extended Abstract). Presented at the Third Learning on Graphs Conference (LoG 2024), Virtual Event, November 26–29, 2024.

This research aims to explore the potential of multimodal LLMs in graph comprehension tasks. We hypothesise that by leveraging both *textual and visual* representations of graphs, these models can achieve superior performance compared to their text-only representations. Our study focuses on a comprehensive set of benchmark tasks at the node, edge, and graph levels, providing a multifaceted evaluation of multimodal approaches in graph analysis. Particularly, based on the designed framework as shown in Figure 1, we seek to address two research questions: *(i)* How does incorporating visual graph representations affect LLM performance on various graph-related tasks compared to purely textual representations? *(ii)* What are the limitations of current multimodal LLMs in processing graph visualisations, and how might these be addressed in future research?

2 Exploring Graph Structure Comprehension Ability of Multimodal LLMs

Our empirical studies follow the GraphQA benchmark settings [1]. Figure 1 provides an overview of our framework, comprehending Graph as Image (GAI$^+$). Its simplified version, GAI, indicates the only graph vision modality is included. We detail each component of our methodology below.

Graph Generation. To systematically evaluate the graph comprehension capabilities of multimodal LLMs, we generated a diverse set of graphs using the Erdős–Rényi (ER) model [8]. Our dataset comprises 500 graphs, each containing between 5 and 20 nodes. This range allows us to assess the models' performance across varying graph complexities. Figure 2 illustrates two example graphs.

Graph Text Encoder. While Fatemi et al. [1] propose several text encoding functions to represent graphs, we focused on two specific methods: adjacency and incident encoding. This choice was motivated by the need to visualise graphs as images, where complicated textual representations might be challenging to depict within a constrained visual space. These encoding methods provide a balance between informational content and visual clarity. Examples see Appendix B.

Graph Visualiser. The graph visualiser component generates visual representations of the structural graphs. While there can be numerous variations in visual aspects such as background colours, layouts, and node shapes, we opted for a standardised approach using Matplotlib [9] with default settings. This decision ensures consistency across our visual graph representations. All graphs are plotted to a fixed size to maintain uniformity. We acknowledge that different visualisation techniques could influence results, and we identify this as an area for future investigation.

Prompt Construction. We adopted prompt designs from [1], which include: Zero-shot prompting (ZERO-SHOT), Few-shot in-context learning (FEW-SHOT), Chain-of-thought (CoT), Zero-shot CoT prompting (ZERO-COT) and Bag prompting (COT-BAG). For scenarios where a visual graph representation is available, we augmented the prompts by prepending the sentence: "There is an undirected graph in this image." (as illustrated in Appendix B). This modification ensures that the LLM is aware of the presence of visual information. Our study encompasses a comprehensive set of graph structure comprehension tasks, including *Node* tasks: node degree, connected nodes; *Edge* tasks: edge existence, shortest path; and *Graph* tasks: node count, edge count, cycle check, and triangle counting. This diverse set of tasks allows us to evaluate the models' performance across various aspects of graph comprehension.

LLMs. Our study focuses on LLMs in a black-box setup, where the model parameters are fixed, and the system only consumes and produces text. This setting reflects the most common scenario for practical LLM usage. We selected two state-of-the-art multimodal LLMs for our main experiments: GPT-4 [3], GPT-4o [3]. To extend the vision of this study, another open-source multimodal LLM, LLaVA-NeXT-7B [10], is also included in the additional experiments Appendix C. These models represent the current pinnacle of multimodal language models, capable of processing both text and image inputs.

3 Results and Discussions

Our main results are summarised in Tables 1 and 2. We discuss our findings in detail below:

Superior performance of multimodal LLMs. An impressive observation from our results is the markedly superior performance of GPT-4o and GPT-4 compared to the PaLM model. In several tasks, these newer models demonstrate near-perfect accuracy, correctly answering questions about graph

Prompt	Encoding	Edge Existence	Node degree	Node count	Edge count	Connected nodes	Cycle check	Triangle counting	Shortest path
ZERO-SHOT	GraphQA [1]	49.0*	25.0*	24.2*	15.0*	53.8*	82.0*	1.5*	11.5*
	Adjacency	96.2	75.8	**100.0**	67.4	76.8	96.2	33.0	69.0
	GAI$_{\text{ADJ}}$	74.6	55.8	93.4	21.2	35.8	97.0	26.4	53.0
	GAI$^{+}_{\text{ADJ}}$	96.4	70.8	**100.0**	65.8	76.4	**98.8**	30.6	63.2
	Incident	97.0	84.4	**100.0**	54.2	89.4	93.8	26.2	68.6
	GAI$_{\text{INC}}$	77.2	50.8	92.2	22.2	36.6	96.6	25.2	56.8
	GAI$^{+}_{\text{INC}}$	99.2	78.6	**100.0**	55.4	88.0	98.4	26.6	69.2
ZERO-COT	GraphQA [1]	41.4*	26.6*	19.4*	12.2*	35.2*	46.2*	12.7*	33.6*
	Adjacency	92.2	76.6	95.4	73.4	82.6	96.6	**33.8**	71.6
	GAI$_{\text{ADJ}}$	60.2	46.4	94.8	24.2	35.4	97.2	25.4	52.4
	GAI$^{+}_{\text{ADJ}}$	90.8	69.0	99.6	66.8	79.2	98.4	31.4	70.8
	Incident	97.2	72.6	97.6	54.6	89.2	90.6	28.6	73.6
	GAI$_{\text{INC}}$	62.2	47.6	93.8	24.8	35.0	96.2	24.6	52.8
	GAI$^{+}_{\text{INC}}$	98.2	72.6	**100.0**	62.0	86.6	97.0	26.0	74.2
FEW-SHOT	GraphQA [1]	42.8*	33.6*	51.2*	18.6*	36.6*	47.8*	3.0*	22.7*
	Adjacency	93.0	69.6	**100.0**	67.8	82.4	93.2	29.2	67.8
	GAI$_{\text{ADJ}}$	84.0	49.4	94.0	19.6	32.0	96.8	24.8	62.4
	GAI$^{+}_{\text{ADJ}}$	96.4	70.0	99.4	64.4	80.6	94.6	27.0	68.6
	Incident	**99.4**	**94.0**	**100.0**	30.2	90.4	94.2	24.0	75.2
	GAI$_{\text{INC}}$	83.4	49.6	92.8	20.2	34.2	96.4	24.4	60.0
	GAI$^{+}_{\text{INC}}$	98.6	90.8	98.2	48.0	89.4	96.6	27.0	75.6
COT	GraphQA [1]	46.6*	75.0*	57.6*	25.2*	30.2*	62.6*	8.1*	38.6*
	Adjacency	92.2	70.2	**100.0**	67.8	84.8	93.4	28.6	70.0
	GAI$_{\text{ADJ}}$	84.2	47.4	92.6	16.2	30.8	96.6	24.6	61.4
	GAI$^{+}_{\text{ADJ}}$	95.0	71.8	99.8	63.8	80.4	95.8	27.4	69.0
	Incident	98.4	92.2	99.8	27.0	**90.2**	95.4	24.4	**76.4**
	GAI$_{\text{INC}}$	84.4	48.4	94.0	18.8	31.8	97.0	24.2	60.6
	GAI$^{+}_{\text{INC}}$	98.4	89.8	98.8	36.0	89.2	97.2	25.6	74.8
COT-BAG	GraphQA [1]	45.8*	75.2*	51.2*	25.0*	41.0*	63.0*	8.1*	40.4*
	Adjacency	94.0	71.2	**100.0**	**70.4**	83.6	92.6	27.0	68.2
	GAI$_{\text{ADJ}}$	86.6	48.8	93.4	17.6	31.0	96.6	25.4	60.6
	GAI$^{+}_{\text{ADJ}}$	96.0	66.0	99.8	65.6	79.6	93.6	27.0	67.6
	Incident	98.8	90.6	99.8	22.0	**90.2**	93.4	24.2	74.2
	GAI$_{\text{INC}}$	83.8	49.6	93.4	17.0	31.8	97.0	24.2	60.4
	GAI$^{+}_{\text{INC}}$	99.0	90.2	98.8	23.0	89.0	95.6	23.6	75.4

Table 1: Comparison of various graph encoder functions based on their accuracy on different graph tasks using GPT-4o. * indicates the results reported in [1] based on PaLM [2]. The results where GAI^{+} makes improvements are highlighted in blue. The results where GAI outperforms the corresponding baseline are highlighted in gray.

structures for almost all test cases. This substantial improvement indicates that recent advancements in multimodal LLMs have significantly enhanced their graph structure comprehension abilities.

Impact of graph visualisation. Our results show that incorporating graph visualisations can enhance LLMs' graph comprehension, though this effect is not uniform across all tasks. The impact of visual input varies depending on: *(i)* The complexity of the graph structure. *(ii)* The specific nature of the task (*e.g.*, local *vs.* global graph properties). For instance, tasks involving global properties (*e.g.*, cycle detection) seem to benefit more from visual input compared to local tasks (*e.g.*, node degree).

Limitations of visual-only input. Interestingly, we found that providing only graph visualisations, without accompanying textual descriptions (similar to the settings of [11]), is insufficient for LLMs to fully comprehend graph structures. This observation highlights the complementary nature of visual and textual information in graph comprehending tasks.

Comparison with specialised graph encoding models. Our comparison with the work of [7], which uses neural networks to encode graph information for LLMs, reveals that our multimodal LLM approach outperforms these carefully trained models in graph structure comprehension tasks. This finding is significant because it suggests that: *(i)* General-purpose multimodal LLMs can compete with, and even surpass, specialised graph encoding models. *(ii)* The versatility of multimodal LLMs allows them to adapt effectively to graph comprehending tasks without task-specific training.

Challenges in graph visualisation. Figure 2 illustrates two contrasting examples of graph visualisation: a simple graph with clear visual representation and a complex graph where GAI provides incorrect responses. This comparison highlights a critical challenge in our approach: the effective visualisation of graphs for multimodal LLMs. The disparity in performance between simple and complex graphs raises several important questions: *(i)* How does graph complexity affect the model's

Prompt	Encoding	Edge Existence	Node degree	Node count	Edge count	Connected nodes	Cycle check	Triangle counting	Shortest path
ZERO-SHOT	GraphQA [1]	49.0*	25.0*	24.2*	15.0*	53.8*	82.0*	1.5*	11.5*
	Adjacency	94.2	44.2	99.4	63.2	74.8	96.0	23.6	74.4
	GAI$_{\text{ADJ}}$	72.4	43.2	82.2	20.4	27.6	95.2	23.6	50.0
	GAI$^{+}_{\text{ADJ}}$	92.0	70.0	**100.0**	61.4	74.4	**98.6**	27.8	55.6
	Incident	97.6	64.8	99.2	42.6	89.2	88.4	26.4	76.8
	GAI$_{\text{INC}}$	74.8	43.8	81.6	20.4	28.0	95.4	22.8	50.0
	GAI$^{+}_{\text{INC}}$	95.6	66.4	99.8	48.2	89.6	97.6	27.0	62.0
ZERO-COT	GraphQA [1]	41.4*	26.6*	19.4*	12.2*	35.2*	46.2*	12.7*	33.6*
	Adjacency	95.0	61.6	99.4	63.8	77.0	96.0	31.0	71.4
	GAI$_{\text{ADJ}}$	73.4	40.8	79.0	19.8	26.8	95.6	23.4	50.8
	GAI$^{+}_{\text{ADJ}}$	83.6	67.6	**100.0**	59.2	73.6	97.8	**31.2**	58.8
	Incident	98.0	76.2	99.6	39.6	88.8	88.4	26.8	76.8
	GAI$_{\text{INC}}$	74.4	42.6	76.4	19.0	24.8	96.0	21.8	49.8
	GAI$^{+}_{\text{INC}}$	93.0	71.4	**100.0**	50.2	88.6	97.4	27.8	63.2
FEW-SHOT	GraphQA [1]	42.8*	33.6*	51.2*	18.6*	36.6*	47.8*	3.0*	22.7*
	Adjacency	95.6	63.2	**100.0**	60.6	75.0	94.6	25.6	69.0
	GAI$_{\text{ADJ}}$	80.4	45.6	85.0	20.0	25.2	93.0	22.6	61.8
	GAI$^{+}_{\text{ADJ}}$	94.4	65.4	99.6	61.8	76.8	95.0	29.0	65.4
	Incident	98.2	92.6	**100.0**	24.4	90.6	90.4	27.6	74.4
	GAI$_{\text{INC}}$	80.2	45.0	82.4	21.6	26.2	92.6	23.2	61.0
	GAI$^{+}_{\text{INC}}$	97.0	91.2	98.6	30.0	90.4	94.6	27.6	69.8
COT	GraphQA [1]	46.6*	75.0*	57.6*	25.2*	30.2*	62.6*	8.1*	38.6*
	Adjacency	95.4	64.2	99.6	63.6	80.2	95.2	28.8	69.0
	GAI$_{\text{ADJ}}$	79.6	43.4	91.8	19.2	24.6	93.6	22.4	60.2
	GAI$^{+}_{\text{ADJ}}$	94.6	66.6	99.8	63.2	79.8	95.0	29.2	68.2
	Incident	**98.8**	**93.8**	99.8	26.2	90.0	90.6	26.2	74.4
	GAI$_{\text{INC}}$	80.8	44.4	91.2	18.2	25.8	93.0	24.2	61.2
	GAI$^{+}_{\text{INC}}$	97.0	93.6	97.8	28.6	90.4	95.6	26.2	71.0
COT-BAG	GraphQA [1]	45.8*	75.2*	51.2*	25.0*	41.0*	63.0*	8.1*	40.4*
	Adjacency	95.8	63.0	**100.0**	63.8	81.8	96.0	27.6	69.0
	GAI$_{\text{ADJ}}$	78.8	42.4	90.8	18.8	25.2	92.6	24.2	59.6
	GAI$^{+}_{\text{ADJ}}$	96.0	66.0	99.4	**64.0**	80.8	96.6	29.2	68.2
	Incident	98.0	93.2	**100.0**	24.4	**90.8**	89.8	25.6	**76.2**
	GAI$_{\text{INC}}$	81.0	42.8	91.4	19.6	23.4	92.8	23.6	61.6
	GAI$^{+}_{\text{INC}}$	97.4	92.0	98.8	27.2	90.0	95.6	25.6	69.4

Table 2: Comparison of various graph encoder functions based on their accuracy on different graph tasks using GPT-4-turbo. * indicates the results reported in [1] based on PaLM [2]. The results where GAI^{+} makes improvements are highlighted in blue. The results where GAI outperforms the corresponding baseline are highlighted in gray.

Method	Edge Existence	Node degree	Node count	Edge count	Connected nodes	Cycle check	Triangle counting	Shortest path
GraphQA [1]	49.0*	75.2*	57.6*	25.2*	53.8*	82.0*	12.7*	40.4*
GCN [12]	68.0§	26.4§	74.6§	5.6§	26.4§	96.4§	20.8§	60.4§
GraphToken [7]	73.8§	**96.2**§	99.6§	42.6§	26.4§	95.6§	**34.8**§	63.8§
GAI	**99.4**	94.0	**100.0**	**70.4**	**90.8**	**98.8**	33.8	**76.4**

Table 3: Comparison of various graph encoder functions based on their accuracy on different graph tasks. * indicates the best results reported in [1] based on PaLM [2]. § indicates the results reported in [7]. The best performances are highlighted in **Bold**.

ability to extract relevant information from visualisations? *(ii)* What are the optimal ways to visually represent different types of graph structures? *(iii)* How can we balance information density and visual clarity in graph representations? These observations underscore the need for further research into graph visualisation techniques that are optimised for LLM comprehension. Future work should explore various visualisation strategies, potentially incorporating: *(i)* Sampling-based interactive or dynamic graph representations. *(ii)* Hierarchical visualisations for complex graphs. *(iii)* Novel encoding techniques that highlight relevant graph properties. More discussion see Appendix D.

4 Conclusion

This study explored the graph structure comprehension abilities of multimodal LLMs through a series of empirical evaluations. We highlight the potential of multimodal LLMs for advancing graph structure comprehension tasks and suggests promising directions for future work in improving graph visualisations and multimodal integration. Limitations and future work see Appendix D.

References

[1] Bahare Fatemi, Jonathan Halcrow, and Bryan Perozzi. Talk like a graph: Encoding graphs for large language models. In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 3, 4, 7

[2] Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. Palm 2 technical report. *CoRR*, abs/2305.10403, 2023. 1, 3, 4, 7

[3] 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. *CoRR*, abs/2303.08774, 2023. 1, 2

[4] Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Harnessing large language models as post-hoc correctors. In *Findings of the 2024 Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 14559–14574. ACL, 2019. 1

[5] Zhikai Chen, Haitao Mao, Hang Li, Wei Jin, Hongzhi Wen, Xiaochi Wei, Shuaiqiang Wang, Dawei Yin, Wenqi Fan, Hui Liu, and Jiliang Tang. Exploring the potential of large language models (llms)in learning on graphs. *SIGKDD Explor.*, 25(2):42–61, 2023. 1

[6] Zhiqiang Zhong, Kuangyu Zhou, and Davide Mottin. Benchmarking large language models for molecule prediction tasks. *CoRR*, abs/2403.05075, 2024. 1

[7] Bryan Perozzi, Bahare Fatemi, Dustin Zelle, Anton Tsitsulin, Seyed Mehran Kazemi, Rami Al-Rfou, and Jonathan Halcrow. Let your graph do the talking: Encoding structured data for llms. *CoRR*, abs/2402.05862, 2024. 1, 3, 4

[8] P ERDdS and A R&wi. On random graphs i. *Publ. math. debrecen*, 6(290-297):18, 1959. 2

[9] Sandro Tosi. *Matplotlib for Python developers*. Packt Publishing Ltd, 2009. 2

[10] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the 2014 Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296. IEEE, 2024. 2

[11] Yunxin Li, Baotian Hu, Haoyuan Shi, Wei Wang, Longyue Wang, and Min Zhang. Visiongraph: Leveraging large multimodal models for graph theory problems in visual context. In *Proceedings of the 2024 International Conference on Machine Learning (ICML)*. PMLR, 2024. 3

[12] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 2017 International Conference on Learning Representations (ICLR)*, 2017. 4, 8

[13] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7

[14] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *CoRR*, abs/2404.16821, 2024. 7

A Illustration of example graphs

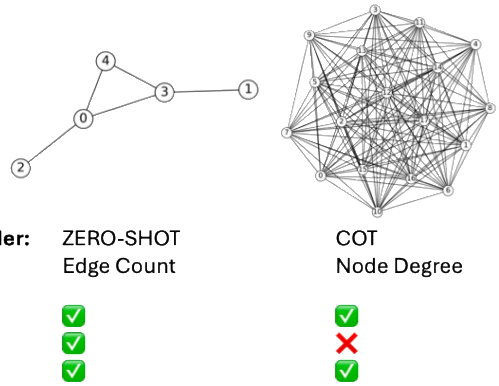

Figure 2: Illustrations of input images and the correctness of different models.

Figure 2 illustrates two example graphs from our generated datasets. Two contrasting examples of graph visualisation are a simple graph with clear visual representation and a complex graph where GAI provides incorrect responses.

B Prompts

In this work, we design the prompt as follows. Precisely, the prompt consists of two components: *Instruction*: Provides general guidance to the LLM, clarifying its role in the conversation. *Message*: Tasks the LLM to comprehend graph structure considering the given textual and visual information.

Instruction: You are an AI designed to analyse graphs and answer specific questions about the graphs. You will receive an image containing a graph and a textual description of the graph's structure. Based on the visual and textual information, you should provide accurate answers to the provided questions.

Message (Adjacency): There is an undirected graph in this image (URL to the image). In an undirected graph, (i,j) means that node i and node j are connected with an undirected edge. G describes a graph among nodes 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. The edges in G are: (0, 8) (1, 7) (2, 3) (2, 8) (2, 9) (3, 7) (3, 9) (4, 5) (4, 7) (4, 9) (4, 10) (5, 6) (6, 7) (8, 9) (9, 10).
Q: How many nodes are in this graph?
A: Let's think step by step. Answer the question using this format: The number of nodes in the graph is [].

Message (Incident): There is an undirected graph in this image (URL to the image). G describes a graph among nodes 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, and 10. In this graph: Node 0 is connected to nodes 8. Node 1 is connected to nodes 7. Node 2 is connected to nodes 3, 8, 9. Node 3 is connected to nodes 2, 7, 9. Node 4 is connected to nodes 5, 7, 9, 10. Node 5 is connected to nodes 4, 6. Node 6 is connected to nodes 5, 7. Node 7 is connected to nodes 1, 3, 4, 6, 10. Node 8 is connected to nodes 0, 2, 9. Node 9 is connected to nodes 2, 3, 4, 8, 10. Node 10 is connected to nodes 4, 7, 9.
Q: How many nodes are in this graph?
A: Let's think step by step. Answer the question using this format: The number of nodes in the graph is [].

Prompt	Encoding	Edge Existence	Node degree	Node count	Edge count	Connected nodes	Cycle check	Triangle counting	Shortest path
ZERO-SHOT	GraphQA [1]	49.0*	25.0*	24.2*	15.0*	53.8*	82.0*	1.5*	11.5*
	Adjacency	66.4	32.2	95.0	17.0	16.2	93.6	7.4	22.2
	GAI_{ADJ}	65.6	11.4	1.4	3.6	0.6	89.8	5.8	3.8
	$\text{GAI}^{+}_{\text{ADJ}}$	69.2	11.6	90.0	12.8	13.0	95.6	3.8	12.0
	Incident	76.0	49.0	65.4	8.4	32.8	91.0	5.2	22.8
	GAI_{INC}	65.6	11.4	1.4	3.6	0.6	89.8	5.8	3.8
	$\text{GAI}^{+}_{\text{INC}}$	73.6	47.6	68.6	6.4	33.4	90.4	5.4	2.6
ZERO-COT	GraphQA [1]	41.4*	26.6*	19.4*	12.2*	35.2*	46.2*	12.7*	33.6*
	Adjacency	66.6	23.4	94.6	19.8	16.0	93.2	2.0	0.6
	GAI_{ADJ}	61.0	9.8	1.0	2.4	0.8	92.6	0.4	5.8
	$\text{GAI}^{+}_{\text{ADJ}}$	70.0	22.8	91.4	14.6	14.4	93.4	0.6	7.0
	Incident	75.4	50.6	67.6	8.2	30.8	90.0	1.0	0.4
	GAI_{INC}	61.0	9.8	1.0	2.4	0.8	92.6	0.4	5.8
	$\text{GAI}^{+}_{\text{INC}}$	71.2	60.4	78.4	5.4	32.4	86.8	1.8	0.4
FEW-SHOT	GraphQA [1]	42.8*	33.6*	51.2*	18.6*	36.6*	47.8*	3.0*	22.7*
	Adjacency	0.0	7.8	0.0	0.0	0.0	69.4	0.0	1.0
	GAI_{ADJ}	0.0	3.0	0.0	0.0	0.0	0.0	0.0	1.6
	$\text{GAI}^{+}_{\text{ADJ}}$	0.0	7.4	0.0	0.0	0.0	76.0	0.0	4.8
	Incident	0.0	5.8	0.0	0.0	0.0	69.4	0.0	0.4
	GAI_{INC}	0.0	3.0	0.0	0.0	0.0	0.0	0.0	2.2
	$\text{GAI}^{+}_{\text{INC}}$	0.0	6.0	0.0	0.0	0.0	78.8	0.0	0.2
COT	GraphQA [1]	46.6*	75.0*	57.6*	25.2*	30.2*	62.6*	8.1*	38.6*
	Adjacency	0.0	5.4	0.0	0.0	0.0	72.0	0.0	0.4
	GAI_{ADJ}	0.0	2.6	0.0	0.0	0.0	0.0	0.0	2.4
	$\text{GAI}^{+}_{\text{ADJ}}$	0.0	4.6	0.0	0.0	0.0	77.0	0.0	3.0
	Incident	0.0	6.8	0.0	0.0	0.0	69.8	0.0	0.6
	GAI_{INC}	0.0	3.2	0.0	0.0	0.0	0.0	0.0	4.0
	$\text{GAI}^{+}_{\text{INC}}$	0.0	6.4	0.0	0.0	0.0	77.8	0.0	0.4
COT-BAG	GraphQA [1]	45.8*	75.2*	51.2*	25.0*	41.0*	63.0*	8.1*	40.4*
	Adjacency	0.0	9.2	0.0	0.0	0.0	65.2	0.0	0.6
	GAI_{ADJ}	0.0	1.2	0.0	0.0	0.0	0.0	0.0	3.4
	$\text{GAI}^{+}_{\text{ADJ}}$	0.0	8.6	0.0	0.0	0.0	67.4	0.0	0.8
	Incident	0.0	4.4	0.0	0.0	0.0	1.6	0.0	1.6
	GAI_{INC}	0.0	2.2	0.0	0.0	0.0	0.0	0.0	3.6
	$\text{GAI}^{+}_{\text{INC}}$	0.0	5.4	0.0	0.0	0.0	66.2	0.0	0.2

Table 4: Comparison of various graph encoder functions based on their accuracy on different graph tasks using LLaVA-NeXT-7B. * indicates the results reported in [1] based on PaLM [2]. The results where GAI^{+} makes improvements are highlighted in blue. The results where GAI outperforms the corresponding baseline are highlighted in gray.

C Results on open-source LLMs

Our additional experimental results on open-source LLMs are summarised in Table 4. Similar to the results of Tables 1-2, it illustrates that incorporating graph visualisations can enhance LLMs' graph comprehension, though this effect is not uniform across all tasks. For instance, visual representations do not bring helpful information for tasks with higher complicity (*e.g.*, shortest path)

D Limitation and Future Work

Larger open-source multimodal LLMs. This study explored the graph structure comprehension abilities of state-of-the-art LLMs in a black-box setup. However, only small-size open-source multimodal LLMs are not included. In many practical situations, user data are too sensitive to upload to an online LLM server. In this case, adopting open-source LLMs will be an alternative solution. Therefore, we intend to explore the capabilities of large-size open-source multimodal LLMs like LLaVA-Next [13] and Intern-VL [14] in understanding graph structure.

In-depth understanding of the impact of visualisation quality. In Section 3, we have highlighted the challenges in graph visualisation, where Figure 2 illustrates two contrasting examples: a simple graph with clear visual representation and a complex graph where GAI provides incorrect responses. This performance disparity between simple and complex graphs raises future work that systematically investigates how different aspects of visualisation (e.g., layout algorithms, node/edge styles,

information density) affect comprehension. This includes developing methods to automatically adapt visualisations based on graph complexity and task requirements. For instance, *(i)* Sampling-based interactive or dynamic graph representations, *(ii)* Hierarchical visualisations for complex graphs, and *(iii)* Novel encoding techniques that highlight relevant graph properties.

Integration with Graph Neural Networks. Our current approach uses multimodal LLMs directly for graph comprehension. However, message-passing neural networks (MPNNs) [12] have established strengths in handling graph-structured data, particularly in preserving permutation invariance properties that are lost in image representations. This suggests an opportunity to combine the complementary strengths of both approaches through lightweight solutions. For example, developing adapters to align MPNN embeddings with multimodal LLM representations in a common space could enable more efficient and theoretically grounded graph processing while maintaining the flexibility and general capabilities of LLMs. However, deploying such multimodal systems in practice introduces significant challenges. These models are typically resource-intensive, requiring substantial computational power, both for training and inference, as well as increased memory to handle the alignment between modalities. Furthermore, the integration of MPNNs with multimodal LLMs involves managing synchronization between distinct data modalities and ensuring compatibility in embedding spaces, which can be technically complex and prone to inconsistencies. Such systems often demand specialised hardware and careful engineering to achieve real-time performance, which might not be feasible in many practical applications where computational resources are limited or latency requirements are stringent.