How Well Do MLLMs Judge Network Visualizations? Local vs. Global Properties

Markus Wallinger* Technical University Munich

Henry Förster* Technical University Munich

Hlib Kholodkov* Technical University Munich

Stephen G. Kobourov* Technical University Munich

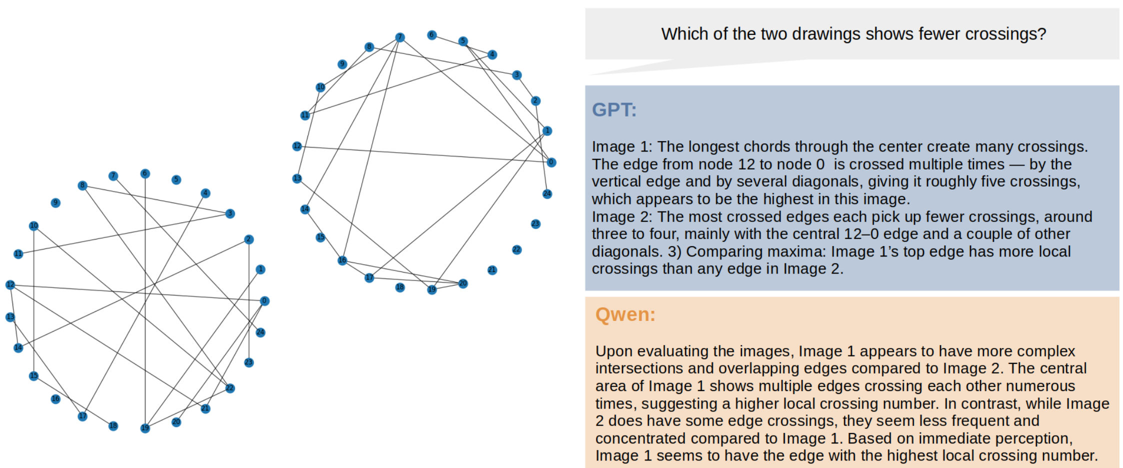


Figure 1: Illustrative example of a trial on perception of global crossings. Both MLLM models answer correctly. Additionally, we show the reasoning output of the models.

Abstract

Multimodal large language models (MLLMs) are increasingly used in visualization research, yet their perceptual capacities for network visualizations remain poorly understood. As more and more decisions are made by AI models, and visualizations often occur in articles and papers, it is, however, important to understand how MLLMs interpret complex visualizations. While prior work has shown that MLLMs can assess global layout properties, such as stress, it remains unclear whether they can also perceive local properties effectively. Our results show that ChatGPT-5 and Qwen-2.5 can accurately judge *edge crossings*, *node distribution*, and *node degree* in network layouts, both globally and locally, with accuracy above chance. However, both models demonstrate higher accuracy in judging a global property compared to its local counterpart. Additionally, we evaluate the confidence and textual reasoning output of the models. Neither is well aligned with the given answers.

Index Terms: Evaluation, Large language models, Explainable AI

1 Introduction

Recent advances in the reasoning capacity of multimodal large language models (MLLMs) have made them attractive tools in visualization research. They have been used for visualization design recommendation [14], design generation [4], or design critiquing [11]. With the addition of vision as an input channel, a natural question that has recently been investigated is how *visualization-literate* [2, 7] these models are.

Graph drawing aesthetics have long been recognized as fundamental to the readability of network visualizations. Empirical studies have demonstrated that edge crossings strongly reduce task performance [13], that node distribution influences how well patterns can be perceived [3], and that identifying high-degree nodes is central to many network visualization tasks [10]. Consequently, evaluation of graph layouts has traditionally relied on aesthetic heuristics and controlled user studies [8, 9]. However, user studies are expensive and difficult to scale, while heuristic measures may not always align with perceptual judgments.

Recent work suggests that MLLMs could provide scalable alternatives. Miller et al. [12] showed that models can perceive *stress* in network visualizations at levels comparable to trained human participants. Stress, however, is fundamentally a *global* property, as it is defined as the sum of a measure over *all* node-pairs. Miller et al. conjecture that the MLLMs use a mix of (global) proxy properties such as crossings, node distribution, and edge lengths. In contrast, many visual properties and visualization tasks rely on finer, localized properties, e.g., which edge is crossed most or what node has the highest degree. Whether MLLMs can perceive such *local* properties of network layouts remains an open question. At the same time, MLLMs are increasingly proposed as evaluators and assistants in visualization. Moreover, there are also use cases where only the visual information is available. Thus, if they are to be trusted in perceptual processing of input images, we must investigate their capabilities and limitations. Similarly, if MLLMs can judge local or global properties quickly and accurately, this knowledge can potentially be transferred to support task-solving capabilities of humans.

In this paper we extend the study of MLLMs for network visualizations to simultaneously investigate global and local perception of three properties: *crossings*, *node distribution*, and *node degree*. We evaluate two representative MLLMs, GPT-5 and Qwen2.5, comparing their accuracy, correlation with property differences, and self-reported confidence. Our results show that MLLMs can judge all three properties substantially more accurately than if it was merely guessing.However, in most cases the performance on local properties is below the global counterparts. Additionally, the confidence and textual reasoning is not well-aligned with the answers.

*e-mail: firstname.lastname@tum.de

2 Research Question and Justification

The experiment by Miller et al. [12] demonstrates that MLLMs are capable of perceiving *stress* in network visualizations. Informally, stress measures how well the Euclidean distances in the drawing correspond to the graph-theoretical distances in the underlying abstract graph considering all node-pairs. A notable observation is that stress is a global property of the drawing. Given the way MLLMs process images, a natural question is whether they can handle local properties.

> *Can MLLMs also perceive local properties in network visualization layouts, in addition to global properties?*

In our study, we focus on three properties: *edge crossings*, *node distribution*, and *node degree*. Edge crossings are among the most widely studied aesthetics for evaluating graph layout algorithms [1]. Similarly, node distribution is an important aesthetic property affecting readability [3]. In contrast, node degree is not a property of the drawing but rather of the underlying graph. Still, identifying high-degree nodes or obtaining an overview of the degree distribution are typical tasks in network visualization [10].

3 Experimental Setup

To summarize our experimental procedure, we generate node-link diagrams in which both local and global properties are systematically varied. For each trial, we present a pair of such drawings as images to an MLLM and ask it to judge the designated property, once according to its local definition and once according to its global definition. In addition to the answer, we also collect the model's textual reasoning and its self-reported confidence score. The datasets, images, and code to prompt and evaluate the experiment can be found in the supplemental material on OSF [1].

3.1 Local and Global Properties

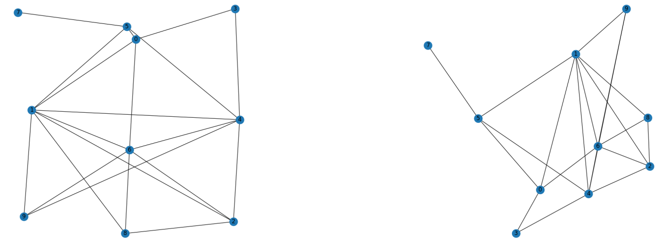
We conducted our experiment on three properties: *crossings*, *node distribution*, and *node degree*. The first two properties, crossings and node distribution, are characteristics of the network layout. The latter, node degree, is determined by the abstract graph structure itself and is independent of the particular drawing. By considering both drawing-dependent and structure-dependent properties, we aim to capture a more comprehensive picture of the factors that influence graph readability and interpretation. Mathematical definitions of the properties are in the supplemental material.

Crossings. Edge crossings in node-link diagrams are widely recognized as one of the key aesthetics to optimize. Studies have demonstrated that crossings can significantly decrease task performance [13]. In straight-line drawings, an edge crossing is defined as a point where two edges that share no common endpoint intersect. The *local crossing number* of a drawing is the maximum number of times any single edge is crossed. The *global crossing number* is the sum of all edge crossings in the drawing. In our experiment, we treat the local crossing number as the local property and the global crossing number as the global property.

Node Distribution. We consider the following definition of node distribution in our experiment. As the local property, we use the minimum Euclidean distance between any pair of nodes in the drawing. For the global property, we use the average of all pairwise Euclidean distances between nodes. This measure reflects the overall spread of the layout and serves as an indicator of how evenly the nodes are distributed in the drawing.

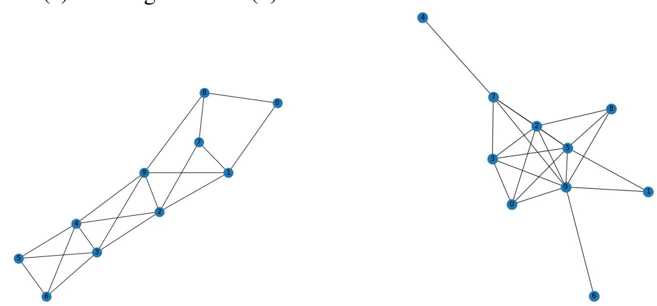
Node Degree. Lastly, we consider *node degree* as the third property in our experiment. Unlike crossings and node distribution, which are aesthetic properties of the drawing, node degree is a structural property of the underlying graph and directly relates to common network analysis tasks. As the local property, we use the maximum degree to capture the most connected node. As the global property, we measure the uniformity of the degree distribution as the standard deviation from the mean. Although this global definition differs in form from that of node distribution, it reflects the discrete nature of degrees: locally focused on hubs, and globally on the overall connectivity. This corresponds to typical tasks in network visualization where analysts either identify key nodes or assess structural heterogeneity.

[1] https://osf.io/adkt5/



(a) (b)

Figure 2: A possible image pair of a trial for node distribution. The drawing in (a) has a more uniform node distribution but a close node pair. In (b), the node distribution is less uniform, but the image does not have a pair of nodes as close as (a). The correct answer would be (a) for the global and (b) for the local trial.



(a) (b)

Figure 3: A possible image pair of a trial for node degree. The drawing in (a) has nearly uniform node degree distribution. In (b), the degree distribution is heterogeneous with several high-degree nodes. The correct answer is (b) for both local and global trial.

3.2 Dataset and Drawings

In the experiment, we consider three graph sizes, $|V| \in \{10, 25, 50\}$, and apply the following procedure to generate the layouts. For each property, we predefine both a lower and an upper bound for its local and global values. Within each range, we sample five evenly spaced values, which serve as target levels. The combination of one local and one global value then defines a specific target condition for a drawing. For each condition, we utilize our layout generators to generate a drawing that satisfies both local and global property constraints. Note that not all combinations are feasible.

All graphs are drawn as straight-line drawings. Each node is uniquely labeled with an integer identifier to facilitate unambiguous reference. The resulting layout is rendered as an image of size 800×800 pixels. Nodes are depicted as disks with a diameter of 30 pixels. Figure 1 shows an example of a pair of images. Our controlled, synthetic instances enable precise local/global manipulation but inevitably abstract away real-world regularities; we target external validity in follow-up work.

Crossings. For each graph instance, we specify two targets: the desired total number of crossings and the desired maximum number of crossings of a single edge. To construct such a drawing, we begin by arranging all nodes in a circular layout. We then add an edge between two nodes that are farthest apart along the circle. This edge partitions the node set into two groups. By subsequently adding edges between nodes belonging to different groups, we can ensure that these edges cross the partition edge. We exploit this observation to control the local crossing number and to fulfill the target for the most-crossed edge.

After meeting the local crossing requirement, we continue by inserting edges within the groups. Each time an edge is added, we recompute the number of crossings in the drawing. If either the local or the global crossing number exceeds the respective target value by a significant margin, the edge is removed. This iterative process is repeated until the drawing reaches the specified crossing targets. Figure 1 shows an illustrative example.

Node Distribution. We begin by generating a graph with a fixed edge density. An initial layout is then computed using a standard force-directed algorithm, which provides a visually balanced starting configuration. From this layout, we apply simulated annealing to iteratively reposition individual nodes. The annealing process allows us to explore alternative layouts while gradually converging toward configurations that satisfy the predefined target values for both the local and the global properties; see Figure 2.

Node Degree. We use the following procedure to generate node-link diagrams for the degree-based condition. First, we predefine the mean node degree $\mu = 4$ and the standard deviation $\sigma = 1.5$. For the global property, we sample node degrees from a normal distribution with a mean of μ and a standard deviation of σ. To control the variability of the degree distribution, the standard deviation is scaled by a target factor. In this way, we can generate instances ranging from uniform degree distributions to more heterogeneous, normally distributed ones. For the local property, after sampling the degree sequence, we select a single node and set its degree explicitly to $\mu + 1.5k$, where k is chosen from the predefined local property range. This allows us to control the maximum degree in the graph independently of the global distribution.

Once the target degree sequence has been defined, we apply the Havel–Hakimi algorithm [5, 6] to construct a simple graph that realizes this sequence. Finally, we compute a layout of the resulting graph using a force-directed algorithm. Figure 3 shows an example.

3.3 Prompts

Our prompts approximate a perceptual judgment, simulating how a human would assess visual differences between two network drawings rather than computing exact graph-theoretic values. The MLLM is instructed to evaluate each image with respect to a designated property, then compare them; see Appendix E for details. It must decide which drawing better exhibits the property (e.g., higher global/local crossing number, more uniform node distribution, smaller minimum node distance, more homogeneous degree distribution, or higher maximum degree), or answer "None" if equivalent. Responses follow a fixed JSON format with three fields: step-by-step justification (*Reason*), categorical outcome (*Answer*), and self-reported confidence on a discrete scale (*Confidence*).

3.4 Experimental Procedure

For our experiment, we employ GPT-5 (model: `gpt-5-2025-08-07`) as a closed-source model and Qwen2.5 (model: `qwen2.5-vl-72b-instruct`) as an open-source alternative for the evaluation system. For each property, we first construct a dataset consisting of pairs of drawings sampled from the full set of generated instances. We define the target difference between the local and global properties of the drawings in the pair. We then create a dataset consisting of 5 local $\times$ 5 global $\times$ 3 sizes $\times$ 2 samples $=$ 150 pairs. This number is a trade-off between covering the span of the defined local and global ranges and the cost of running the experiment.

Table 1: Overall accuracy for crossings (XR), node distribution (ND), and node degree (DEG) for local and global conditions.

Model	XR		ND		DEG	
	Local	Global	Local	Global	Local	Global
GPT-5	0.69	0.74	0.72	0.86	0.68	0.74
Qwen2.5	0.68	0.81	0.39	0.54	0.68	0.61

Table 2: Spearman correlation (ρ) between correctness and property differences for GPT-5 and Qwen2.5 across local and global trials. crossings (XR), node distribution (ND), node degree (DEG).

Model	Trial	XR	ND	DEG
GPT-5	Global	0.49	0.49	0.48
GPT-5	Local	0.28	0.16	0.55
Qwen2.5	Global	0.21	0.03	0.33
Qwen2.5	Local	0.19	0.06	0.52

Finally, we show several training examples with the correct answer and a brief rationale. We fix a single standardized, few-shot prompt per condition to limit degrees of freedom and avoid overfitting via prompt engineering. This choice prioritizes measuring intrinsic perceptual capacity under a reasonable baseline, but it also implies sensitivity to prompting and leaves room for the development of improved prompting strategies in future work. Full protocol details are in the supplemental materials.

4 Evaluation

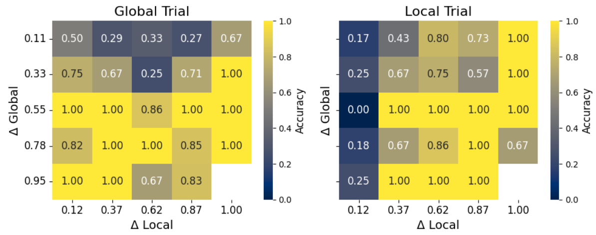
In the evaluation, we focus on three data points that we collected for each trial. Firstly, we evaluate the accuracy and correlation with differences in local and global properties. Afterwards, we briefly discuss the textual reasoning and confidence in the answer.

4.1 Accuracy

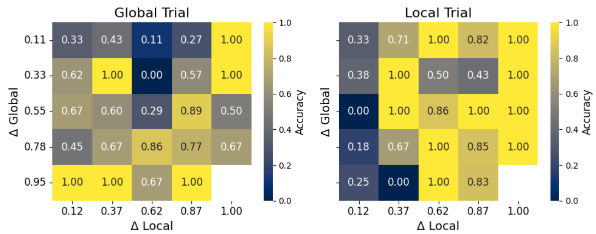
Table 1 summarizes the overall accuracies across the three tested properties, while Table 2 reports the correlations between correctness and the magnitude of local and global differences. The accompanying heatmaps provide additional insight into how accuracy varies across the full range of property values. Figure 4 shows the heatmap for node degree and the other heatmaps can be found in the supplemental material. Taken together, this suggests that both GPT-5 and Qwen2.5 operate above chance in all conditions, but their strengths and weaknesses diverge depending on whether the property is local or global, and on how strongly the drawings differ.

Crossings. Both models perform reliably on edge crossings, which confirms it is a salient visual cue for MLLMs. On global trials, Qwen2.5 achieves its best performance overall (0.81), slightly outperforming GPT-5 (0.74). On local trials, the models are close (0.69 vs. 0.68). The heatmaps show a clear trend: as either the global number of crossings or the local maximum crossing count increases, both models become more accurate. This is reflected in the correlation results (Table 2), where GPT-5 benefits strongly from both global and local differences ($\rho = 0.49$ and 0.28, respectively). Qwen2.5 shows weaker but still positive correlations with both axes ($\rho \approx 0.21$ global, $\rho \approx 0.19$ local). In other words, both models exploit it when it is visually pronounced, but GPT-5 is more consistent in leveraging both local and global cues simultaneously.

Node distribution. GPT-5 is clearly more successful, particularly on the global trials (0.86 vs. 0.54 for Qwen2.5). Its heatmap reveals a gradient along the global axis. This sensitivity is confirmed by the correlation between correctness and global differences ($\rho = 0.49$). By contrast, Qwen2.5 's heatmap is much flatter, with no consistent improvement as global variation increases ($\rho \approx 0.03$). Interestingly, neither model benefits from larger local



(a) GPT-5



(b) Qwen2.5

Figure 4: Heatmap showing the accuracy of the global (left) and local (right) trials for node degrees. Each cell represents combination of local and global property values. Note that we bin and each cell represents a range of values.

differences in node spacing; in fact, both sometimes show declining accuracy in such cases. This suggests that while average pairwise distance is a usable global cue, the minimum-distance signal may be too subtle or unstable for MLLMs to exploit reliably.

Node degree. The node degree condition highlights the strongest contrast between the models. GPT-5 excels at global trials (0.86), demonstrating sensitivity to the uniformity of the degree distribution. The heatmap shows accuracy rising along the global axis, consistent with its positive correlation with global differences ($\rho = 0.48$). However, GPT-5 is noticeably worse in the local setting (0.72), indicating difficulty in consistently identifying the highest-degree node. Qwen2.5 shows the opposite trend: it performs better locally (0.68 vs. 0.61 globally), and its heatmap reveals accuracy rising along the local axis, in line with its strong correlation with local differences ($\rho = 0.52$). In other words, GPT-5 is more attuned to the global balance of connectivity, while Qwen2.5 is comparatively stronger at detecting hubs. This division of strengths underscores that the two models prioritize different cues when reasoning.

Summary. Across properties, GPT-5 demonstrates a general bias toward global structure, while Qwen2.5 is more responsive to localized signals, such as highly crossed edges or individual hubs. The heatmaps make this distinction particularly visible, showing how accuracy improves along different axes for each model. The correlation analysis (Table 2) reinforces this picture: correctness tracks the intended property differences for crossings and degree, but less so for node distribution, where only GPT-5 shows systematic global sensitivity. Overall, the results suggest that while both models are capable of above-chance perceptual judgments, their strategies diverge substantially in terms of what they attend to. We emphasize that absolute accuracies are time-sensitive as model families evolve; the qualitative pattern (global > local; GPT-5 global bias; Qwen-2.5 local advantage) is the more stable takeaway.

4.2 Reasoning

For each trial, we also collected the textual reasoning that accompanied the model's answer. Although we did not conduct a systematic thematic analysis of these responses, we observed several instances where the textual explanation did not align with the actual content of the image pairs. Hand-picked examples of such inconsistencies are provided in the supplemental material. For instance, in trials targeting the local crossing property, the model sometimes incorrectly claimed that a particular edge had more crossings, even though this was not the case in the drawing. Similarly, for global crossings, models occasionally described areas with "many crossings" despite these not being present in the drawing. These observations suggest that both GPT-5 and Qwen2.5 may rely on predefined textual templates or general heuristics for what the correct answer should look like, without fully aligning these explanations with the visual evidence. Therefore, textual reasoning should be interpreted with caution, and we recommend requiring evidence references (e.g., node IDs, marked edges) for downstream use.

4.3 Confidence

The tables of the confidence analysis can be found in the supplemental material. For GPT-5, confidence correlates moderately with accuracy for node distribution, but not for crossings or degree. For Qwen2.5, confidence scores are almost uniformly high (close to 4 on a 5-point scale) regardless of correctness, with only local crossings showing a significant correlation. This systematic overconfidence means that even when answers are incorrect, Qwen2.5 still reports high confidence. Consequently, we advise against using raw confidence for gating decisions without separate calibration.

5 Conclusion

We present an experiment that evaluates the perception of local and global properties in network visualizations by MLLMs. Our results show that GPT-5 and Qwen2.5 can accurately judge edge crossings, node distribution, and node degree. However, the results also indicate that global properties can be perceived more accurately than local properties. This is an interesting result, as intuitively local properties should be easier to perceive and compare than global properties (e.g., highest degree node vs. degree distribution). One possible explanation lies in how vision transformers process images: they segment an input drawing into patches and then aggregate information across the entire image. Such mechanisms may naturally favor capturing broad global patterns (e.g., overall density or distribution) rather than fine-grained local features that require precise identification of single nodes or edges. This suggests that current MLLMs, while powerful, may be biased toward global statistics rather than localized details, which has direct implications for their use in visualization tasks that require fine perceptual discrimination. Lastly, similar to Miller et al. [12], we find evidence that confidence and textual reasoning must be carefully considered, as they are only poorly aligned with the actual answers. Taken together, these results provide immediately usable guidance for deploying MLLMs as perceptual aides in network visualization, which favor global judgments, scaffold local ones, and treat confidence/rationales with caution.

Limitations. (i) *Prompt sensitivity:* We used fixed, few-shot prompts, which may change results with different prompting strategies. (ii) *Synthetic data:* Controlled instances ease ground-truthing but under-represent real-world heterogeneity (e.g., community structure, noise, labels). (iii) *No human baseline:* Without a matched human study, absolute performance is hard to contextualize. (iv) *Model drift:* Results are time-sensitive as MLLMs evolve. (v) *Scope:* We tested three properties and straight-line drawings; other aesthetics (symmetry, angular resolution) are out of scope.

Future Work. Add a human baseline with matched tasks and power analysis; test on real-world networks (e.g., social, biological) and alternative drawing styles; probe prompt variants (concise vs. chain-of-thought; region-first vs. global-first), tool-assisted evidence (node/edge highlighting), and calibration methods; explore fine-tuning or synthetic augmentation specifically targeting local perception (hubs, local crossings).

References

[1] S. D. Bartolomeo, T. Crnovrsanin, D. Saffo, E. Puerta, C. Wilson, and C. Dunne. Evaluating graph layout algorithms: A systematic review of methods and best practices. *Comput. Graph. Forum*, 43(6), 2024. doi: 10.1111/CGF.15073 2

[2] A. Bendeck and J. T. Stasko. An empirical evaluation of the GPT-4 multimodal language model on visualization literacy tasks. *IEEE Trans. Vis. Comput. Graph.*, 31(1):1105–1115, 2025. doi: 10.1109/TVCG.2024.3456155 1

[3] C.-C. Carbon, T. Mchedlidze, M. H. Raab, and H. Wächter. The power of shape: How shape of node-link diagrams impacts aesthetic appreciation and triggers interest. *i-Perception*, 9(5):2041669518796851, 2018. 1, 2

[4] V. Dibia. LIDA: A tool for automatic generation of grammar-agnostic visualizations and infographics using large language models. In D. Bollegala, R. Huang, and A. Ritter, eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2023, Toronto, Canada, July 10-12, 2023*, pp. 113–126. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.ACL-DEMO.11 1

[5] S. L. Hakimi. On realizability of a set of integers as degrees of the vertices of a linear graph. i. *Journal of the Society for Industrial and Applied Mathematics*, 10(3):496–506, 1962. doi: 10.1137/0110037 3

[6] V. Havel. Poznámka o existenci konečných grafů. *Časopis pro pěstování matematiky*, 80:477–480, 1955. 3

[7] J. Hong, C. Seto, A. Fan, and R. Maciejewski. Do LLMs have visualization literacy? An evaluation on modified visualizations to test generalization in data interpretation. *IEEE Trans. Vis. Comput. Graph.*, 2025. doi: 10.1109/TVCG.2025.3536358 1

[8] W. Huang, S. Hong, and P. Eades. Effects of crossing angles. In *IEEE VGTC Pacific Visualization Symposium 2008, PacificVis 2008, Kyoto, Japan, March 5-7, 2008*, pp. 41–46. IEEE Computer Society, 2008. doi: 10.1109/PACIFICVIS.2008.4475457 1

[9] S. G. Kobourov, T. Mchedlidze, and L. Vonessen. Gestalt principles in graph drawing. In E. D. Giacomo and A. Lubiw, eds., *Graph Drawing and Network Visualization - 23rd International Symposium, GD 2015, Los Angeles, CA, USA, September 24-26, 2015, Revised Selected Papers*, vol. 9411 of *Lecture Notes in Computer Science*, pp. 558–560. Springer, 2015. doi: 10.1007/978-3-319-27261-0_50 1

[10] B. Lee, C. Plaisant, C. S. Parr, J. Fekete, and N. Henry. Task taxonomy for graph visualization. In E. Bertini, C. Plaisant, and G. Santucci, eds., *Proceedings of the 2006 AVI Workshop on BEyond time and errors: novel evaluation methods for information visualization, BELIV 2006, Venice, Italy, May 23, 2006*, pp. 1–5. ACM Press, 2006. doi: 10.1145/1168149.1168168 1, 2

[11] L. Y. Lo and H. Qu. How good (or bad) are llms at detecting misleading visualizations? *IEEE Trans. Vis. Comput. Graph.*, 31(1):1116–1125, 2025. doi: 10.1109/TVCG.2024.3456333 1

[12] J. Miller, M. Wallinger, L. Felder, T. Brand, H. Förster, J. Zink, C. Chen, and S. G. Kobourov. Exploring mllms perception of network visualization principles. *CoRR*, abs/2506.14611, 2025. doi: 10.48550/ARXIV.2506.14611 1, 2, 4

[13] H. C. Purchase. Which aesthetic has the greatest effect on human understanding? In G. D. Battista, ed., *Graph Drawing, 5th International Symposium, GD '97, Rome, Italy, September 18-20, 1997, Proceedings*, vol. 1353 of *Lecture Notes in Computer Science*, pp. 248–261. Springer, 1997. doi: 10.1007/3-540-63938-1_67 1, 2

[14] L. Wang, S. Zhang, Y. Wang, E. Lim, and Y. Wang. Llm4vis: Explainable visualization recommendation using chatgpt. In M. Wang and I. Zitouni, eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: EMNLP 2023 - Industry Track, Singapore, December 6-10, 2023*, pp. 675–692. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-INDUSTRY.64 1

A Property Definitions

A.1 Edge Crossings

Let $G=(V,E)$ be a graph with straight-line drawing P, where each node $u \in V$ is assigned a position $p_u \in \mathbb{R}^2$. For two edges $e=(u,v)$ and $f=(x,y)$ with $\{u,v\} \cap \{x,y\} = \emptyset$, we say that e and f *cross* if the line segments (p_u, p_v) and (p_x, p_y) intersect at a point other than their endpoints.

The *local crossing number* of the drawing is defined as

$$\mathrm{cr}_{\max}(P) = \max_{e \in E} \#\{f \in E \mid f \ \text{ crosses } \ e\},$$

that is, the maximum number of crossings over all individual edges.

The *global crossing number* of the drawing is defined as

$$\mathrm{cr}_{\mathrm{tot}}(P) = \#\{\{e,f\} \subseteq E \mid e \ \text{ crosses } f\},$$

that is, the total number of pairwise edge crossings in the drawing.

A.2 Node Distribution

We define the local and global node distribution using Euclidean distances between node positions. Let V denote the set of nodes with $|V|$ nodes, and let $p_u \in \mathbb{R}^2$ denote the position of node $u \in V$ in the drawing. For two nodes $u, v \in V$, we write $\|p_u - p_v\|_2$ for their Euclidean distance.

The *local node distribution* is given by the minimum pairwise distance:

$$d_{\min}(P) = \min_{u<v \in V} \|p_u - p_v\|_2,$$

which captures the most crowded region of the drawing.

The *global node distribution* is given by the average pairwise distance:

$$d_{\mathrm{avg}}(P) = \frac{1}{\binom{|V|}{2}} \sum_{u<v \in V} \|p_u - p_v\|_2,$$

which reflects the overall spread of the node layout.

Here, $P = \{p_u \mid u \in V\}$ denotes the set of node positions.

A.3 Node Degree

Let $G=(V,E)$ be a graph with node set V and edge set E. For each node $u \in V$, the *degree* $\deg(u)$ is defined as the number of edges incident to u.

The *local node degree property* is defined as the maximum degree in the graph:

$$\Delta(G) = \max_{u \in V} \deg(u).$$

The *global node degree property* is defined as the standard deviation of the degree distribution:

$$\sigma_{\deg}(G) = \sqrt{\frac{1}{|V|} \sum_{u \in V} \left(\deg(u) - \mu_{\deg}\right)^2},$$

where

$$\mu_{\deg} = \frac{1}{|V|} \sum_{u \in V} \deg(u)$$

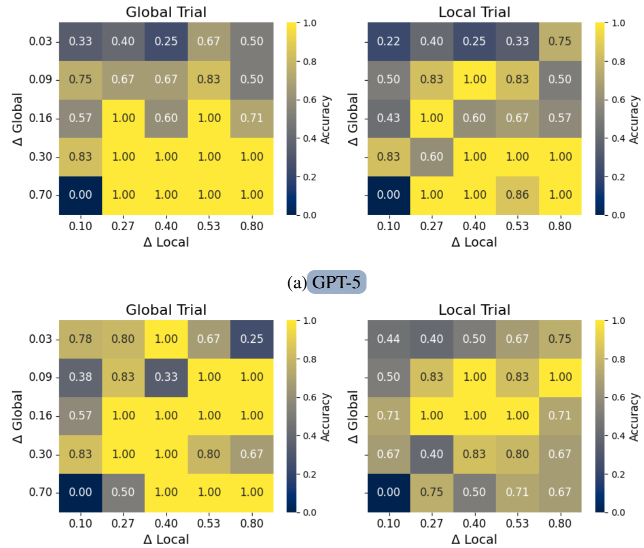
is the average node degree. This measure captures how uniformly degrees are distributed: lower values indicate a more even distribution, while higher values reflect the presence of hubs or strong degree heterogeneity.

B Accuracy

Figure 5 and Figure 6 show the heatmaps for the edge crossing and node distribution properties.

C Textual Reasoning

We will present some examples where the MLLM models do not align with the given answer.



(a) GPT-5

(b) Qwen2.5

Figure 5: Heatmap showing the accuracy of the global (left) and local (right) trials for edge crossings. Each cell represents combination of local and global property values. Note that we bin and each cell represents a range of values.

C.1 Example 1 (Global Crossing)

See Figure 7 for the corresponding pair of drawings. In this example, the model invents edges in the visualization that do not exist in the graph.

```
Image 1: Crossings are concentrated in the middle where long
      diagonals from nodes near the top (6-8) run down to the
      lower-right and intersect with the long edges from 12 and 14
       to 24. I see several distinct intersections along the 12-24
       and 14-24 lines plus a few among the downward spokes from
      6/7. Overall this looks like a moderate number of crossings
      (roughly a dozen or so).
Image 2: There is a dense fan of edges from node 0 to many nodes
      on the upper arc and another set of long edges from node 12
      to the lower-right. These sets overlap heavily: the 0-{2..9}
       edges cross the 12-{18,19,0?} and 13-24 lines, and the
      24-{21,22,23} edges also cut across several of those. The
      central band around the diameter contains many more
      intersections than in Image 1 (visually around a couple
      dozen). 3) Comparing, Image 2 has clearly more overall
      crossings than Image 1.
```

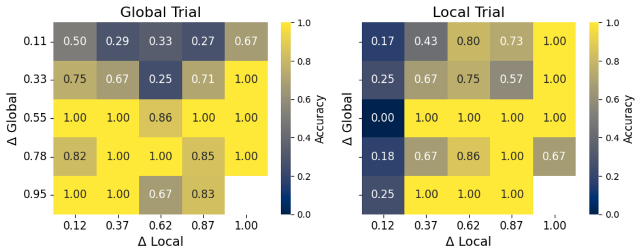
C.2 Example 2 (Local Node Degree)

See Figure 8 for the corresponding pair of drawings. In this example, Qwen2.5 identifies the overall structure correctly but invents high-degree nodes in one of the images.

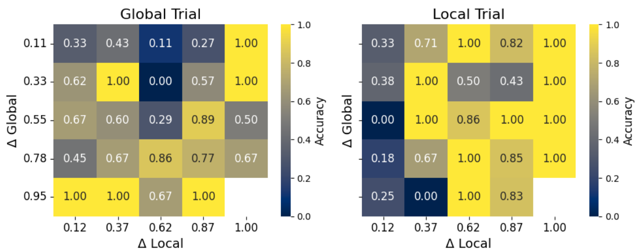
```
Upon evaluating the images, Image 1 appears to have a more linear
       structure with nodes connected in a chain-like manner,
      suggesting a relatively low degree for most nodes. In
      contrast, Image 2 exhibits a more complex structure with
      multiple nodes having several connections, indicating a
      higher degree. Based on the immediate perception, Image 2
      seems to contain nodes with a higher degree compared to
      those in Image 1.
```

C.3 Example 3 (Global Node Distribution)

See Figure 9 for the corresponding pair of drawings. In this example, GPT-5 perceives details of the drawing (spikes) that are not

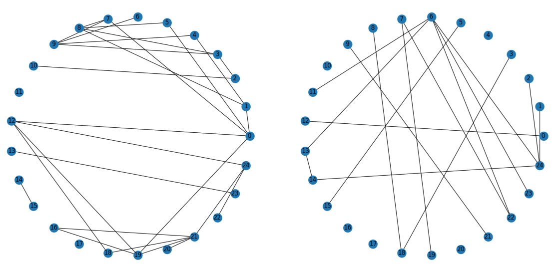


(a) GPT-5



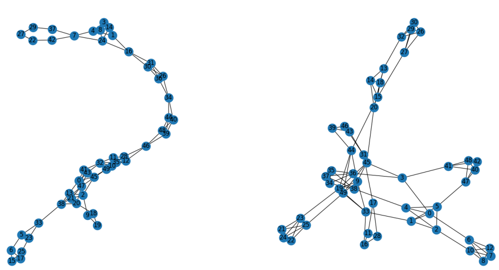
(b) Qwen2.5

Figure 6: Heatmap showing the accuracy of the global (left) and local (right) trials for node distribution. Each cell represents combination of local and global property values. Note that we bin and each cell represents a range of values.



(a) (b)

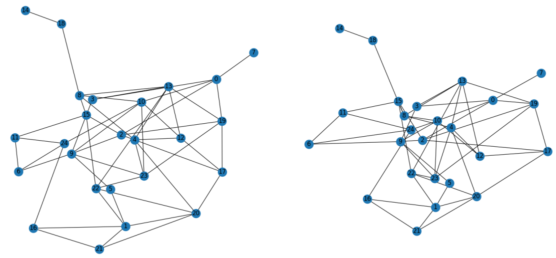
Figure 7: A pair of drawings were the textual reasoning does not align with the actual answer. GPT-5 answered "Image 2" (b) while "Image 1" (a) is the correct answer.



(a) (b)

Figure 8: A pair of drawings were the textual reasoning does not align with the actual answer. Qwen2.5 answered "Image 2" (b) while "Image 1" (a) is the correct answer.

present. Additionally, it also fails to correctly determine which image is more "clustered".



(a) (b)

Figure 9: A pair of drawings were the textual reasoning does not align with the actual answer. GPT-5 answered "Image 2" (b) while "Image 1" (a) is the correct answer.

```
Image 2 has a more uniform node distribution compared to Image 1.
     In Image 1, nodes are more clustered in the center with
     some spikes, such as node 7 and node 14. Image 2 shows a
     more even spread of nodes across the canvas, reducing
     clustering and spikes.
```

D Confidence

Table 3 shows the mean and standard deviation of the Likert scale answers.

E Prompts

E.1 Edge Crossings (Global)

```
You are participating in a study that investigates how well
     people can see the difference between drawings of networks.
     The aim is to help guide the design of visualisations of
     networks for ease of understanding. We will first introduce
     the terms and then define your task.

**Definition of 'Edge Crossing':** In this context, 'edge
     crossings' of a pair of edges in a network drawing is
     defined as if the two straight lines representing an edge
     cross, i.e. share a common point beside the endpoints.

**Your task** Your task is to decide which of the two images
     shows a network with more overall edge crossings.

Make your decision based on your immediate perception, without
     examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You
     will be shown two graph images (labeled Image 1 and Image 2)
     .

**steps**

1. Evaluate the individual edge crossings of Image 1
2. Try to quantify the total edge crossings of Image 1
3. Evaluate the individual edge crossings of Image 2
4. Try to quantify the total edge crossings of Image 2
5. Compare Image 1 against Image 2 and derive your decision.
     Answer Image 1 if you think that Image 1 has more total
     crossings. Awnser Image 2 if you think that Image 2 has more
      total crossings. If you think both are the same, answer
     with None.

## Output Format

Output a single JSON object with the following keys and data
     types:
- "Reason": string - Your detailed, step-by-step reasoning and
     justification for the answer.
```

Table 3: Mean (m) and standard deviation (sd) of Likert confidence (1-5) for GPT-5 and Qwen2.5 across local and global trials. XR = crossings, ND = node distribution, DEG = node degree.

Model	Trial	XR		ND		DEG	
		m	sd	m	sd	m	sd
GPT-5	Global	3.07	0.50	3.74	0.51	4.09	0.28
GPT-5	Local	2.88	0.45	3.40	0.53	3.99	0.18
Qwen2.5	Global	4.00	0.00	4.00	0.00	4.00	0.00
Qwen2.5	Local	3.98	0.20	4.02	0.14	4.00	0.00

Table 4: Pearson correlation (r) and p-values between correctness and Likert confidence (1-5) for GPT-5 and Qwen2.5 across local and global trials. XR = crossings, ND = node distribution, DEG = node degree. Missing values (–) indicate constant confidence values with zero variance.

Model	Trial	XR		ND		DEG	
		r	p	r	p	r	p
GPT-5	Global	0.01	0.91	0.25	$< .01$	0.00	0.78
GPT-5	Local	0.01	0.88	0.22	$< .01$	0.11	0.18
Qwen2.5	Global	–	–	–	–	–	–
Qwen2.5	Local	0.27	$< .01$	0.10	0.24	0.00	1.00

```
- "Answer": string - Your direct, concise final answer. Either
      Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of
      certainty in the answer.
```

E.2 Edge Crossings (Local)

```
You are participating in a study that investigates how well
      people can see the difference between drawings of networks.
      The aim is to help guide the design of visualisations of
      networks for ease of understanding. We will first introduce
      the terms and then define your task.

**Definition of 'Edge Crossing':** In this context, 'edge
      crossings' of a pair of edges in a network drawing is
      defined as if the two straight lines representing an edge
      cross, i.e. share a common point beside the endpoints.

**Definition of 'Local Crossing':** In this context, 'local
      crossing number' of an edge is how often an edge is crossed.

**Your task** Your task is to decide which of the two images
      shows the edge with the highest local crossing number.

Make your decision based on your immediate perception, without
      examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You
      will be shown two graph images (labeled Image 1 and Image 2)
      .

**steps**

1. Evaluate the individual local edge crossings of Image 1
2. Try to quantify the edge with the highest local crossing
      number of Image 1
3. Evaluate the individual local edge crossings of Image 2
4. Try to quantify the edge with the highest local crossing
      number of Image 2
5. Compare Image 1 against Image 2 and derive your decision.
      Answer Image 1 if you think that Image 1 has the edge with
      most local crossings. Anwser Image 2 if you think that Image
       2 has the edge with most local crossings total crossings.
      If you think both are the same, answer with None.

## Output Format

Output a single JSON object with the following keys and data
      types:
- "Reason": string - Your detailed, step-by-step reasoning and
      justification for the answer.
- "Answer": string - Your direct, concise final answer. Either
      Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of
      certainty in the answer.
```

E.3 Node Distribution (Global)

```
You are participating in a study that investigates how well
      people can see the difference between drawings of networks.
      The aim is to help guide the design of visualisations of
      networks for ease of understanding. We will first introduce
      the terms and then define your task.

**Definition of 'Node distance':** In this context, 'node
      distance' of a pair of nodes in a network drawing is defined
       as the Euclidean straight-line distance between the two
      nodes. It does not matter if the nodes are connected or not.

**Definition of 'Uniform node distribution':** In this context, '
      Uniform node distribution' is how evenly nodes are
      distributed in the layout. You could also think of this as a
       pairwise average node distance without too much deviation.

**Your task** Your task is to decide which of the two images
      shows a network with more uniform distribution.

Make your decision based on your immediate perception, without
      examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You
      will be shown two graph images (labeled Image 1 and Image 2)
      .

**steps**

1. Evaluate the node distribution of Image 1
2. Try to quantify the node distribution of Image 1
3. Evaluate the node distribution of Image 2
4. Try to quantify the node distribution of Image 2
5. Compare Image 1 against Image 2 and derive your decision.
      Answer Image 1 if you think that Image 1 has a more uniform
      distribution. Awnser Image 2 if you think that Image 2 has a
       more uniform distribution. If you think both are the same,
      answer with None.

## Output Format

Output a single JSON object with the following keys and data
      types:
- "Reason": string - Your detailed, step-by-step reasoning and
      justification for the answer.
- "Answer": string - Your direct, concise final answer. Either
      Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of
      certainty in the answer.
```

E.4 Node Distribution (Local)

```
You are participating in a study that investigates how well
      people can see the difference between drawings of networks.
      The aim is to help guide the design of visualisations of
      networks for ease of understanding. We will first introduce
      the terms and then define your task.

**Definition of 'Node distance':** In this context, 'node
      distance' of a pair of nodes in a network drawing is defined
       as the Euclidean straight-line distance between the two
      nodes. It does not matter if the nodes are connected or not.
```

```
**Definition of 'Minimum node distance':** In this context, 'minimum node distance' is the minimum of pairwise node distances.

**Your task** Your task is to decide which of the two images shows the pair of nodes with the minimum distance.

Make your decision based on your immediate perception, without examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You will be shown two graph images (labeled Image 1 and Image 2).

**steps**

1. Evaluate the individual node distances of Image 1
2. Try to quantify the pair with the lowest node distance in Image 1
3. Try to quantify the pair with the lowest node distance in Image 2
4. Try to quantify the pair with the lowest node distance in Image 2
5. Compare Image 1 against Image 2 and derive your decision. Answer Image 1 if you think that Image 1 has the pair of nodes with the minimum distance. Anwser Image 2 if you think that Image 2 has the pair of nodes with the minimum distance. If you think both are the same, answer with None.

## Output Format

Output a single JSON object with the following keys and data types:
- "Reason": string - Your detailed, step-by-step reasoning and justification for the answer.
- "Answer": string - Your direct, concise final answer. Either Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of certainty in the answer.
```

E.5 Node Degree (Global)

```
You are participating in a study that investigates how well people can see the difference between drawings of networks. The aim is to help guide the design of visualisations of networks for ease of understanding. We will first introduce the terms and then define your task.

**Definition of 'Node Degree':** In this context, 'node degree' of a node in a network drawing is defined as how many lines connect to other nodes.

**Your task** Your task is to decide which of the two images shows a network with more homogeneous node degree distribution.

Make your decision based on your immediate perception, without examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You will be shown two graph images (labeled Image 1 and Image 2).

**steps**

1. Evaluate the individual node degrees of Image 1
2. Try to quantify the node degrees of Image 1
3. Evaluate the individual node degrees of Image 2
4. Try to quantify the node degrees of Image 2
5. Compare Image 1 against Image 2 and derive your decision. Answer Image 1 if you think that Image 1 has a more homogeneous distribution. Anser Image 2 if you think that Image 2 has a more homogeneous distribution. If you think both are the same, answer with None.

## Output Format

Output a single JSON object with the following keys and data types:
- "Reason": string - Your detailed, step-by-step reasoning and justification for the answer.
- "Answer": string - Your direct, concise final answer. Either Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of certainty in the answer.
```

E.6 Node Degree (Local)

```
You are participating in a study that investigates how well people can see the difference between drawings of networks. The aim is to help guide the design of visualisations of networks for ease of understanding. We will first introduce the terms and then define your task.

**Definition of 'Node Degree':** In this context, 'node degree' of a node in a network drawing is defined as how many lines connect to other nodes.

**Definition of 'Average Node Degree'** In this context, 'average node degree' is the average of all nodes' node degree.

**Your task** Your task is to decide which of the two images shows the node with the highest degree.

Make your decision based on your immediate perception, without examining the drawings in great detail.
Do not try to extract the graph and calculate the result. You will be shown two graph images (labeled Image 1 and Image 2).

**steps**

1. Evaluate the individual node degrees of Image 1
2. Try to quantify the highest node degree of Image 1
3. Evaluate the individual node degrees of Image 2
4. Try to quantify the highest node degree of Image 2
5. Compare Image 1 against Image 2 and derive your decision. Answer Image 1 if you think that Image 1 has the node with the highest degree. Anser Image 2 if you think that Image 2 has the node with the highest degree. If you think both are the same, answer with None.

## Output Format

Output a single JSON object with the following keys and data types:
- "Reason": string - Your detailed, step-by-step reasoning and justification for the answer.
- "Answer": string - Your direct, concise final answer. Either Image 1, Image 2, or None if both are the same.
- "Confidence": integer (between 1 and 5) - Your level of certainty in the answer.
```