Evaluating 'Graphical Perception' with Multimodal LLMs

Kenichi Maeda*
University of Massachusetts Boston

Mahsa Geshvadi†
University of Massachusetts Boston

Daniel Haehn‡
University of Massachusetts Boston

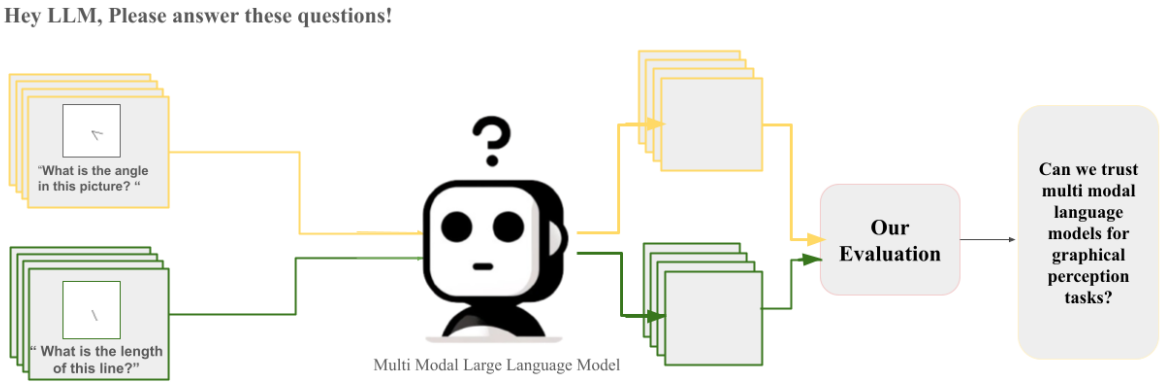

Figure 1: Evaluation Process of MM-LLMs for Graphical Perception Tasks.

Abstract

Large language models (LLMs) have gained significant attention in recent years due to their performance across various fields. Multimodal-LLMs (MM-LLMs) have particularly enabled us to process image-related queries effectively. Our study investigates the capabilities of such models in graphical perception tasks by reproducing the experiments of Cleveland and McGill (1984) and comparing against recent results with convolutional neural networks (CNNs). We evaluate the graphical perception capabilities of ChatGPT4-Vision, LLaVA, and a custom fine-tuned version of LLaVA. The results indicate that MM-LLMs do not outperform CNNs or humans in all graphical perception tasks. However, we propose that fine-tuning MM-LLMs with graphical perception stimuli could enhance their capabilities beyond their general-purpose applications, suggesting a potential pathway for improving their performance.

Index Terms: Machine Perception, Graphical Perception,Multimodal Large Language Models

1 Introduction

MM-LLMs have emerged as powerful tools across various domains, including but not limited to object detection and scientific chart question answering, often surpassing human performance in tasks equivalent to their capabilities [1]. This advancement has sparked interest among researchers in systematically evaluating these models, assessing their performance across different areas, their vulnerability to security issues [13], and potential leakage [8]. Specifically, previous research has focused on their performance in recognition [10] and visual relation problems [7].

*e-mail: kenichi.maeda001@umb.edu
†e-mail: mahsa.geshvadi001@umb.edu
‡e-mail: daniel.haehn@umb.edu

Visualization analysis, particularly the computational analysis of graphs, charts, and visual encodings, has seen increasing interest. This interest is driven by applications such as information extraction and classification, visual question answering, and even design analysis and generation. While using MM-LLMs in these tasks presents opportunities, it also introduces challenges. The computational analysis of visualizations is a more complex task than object detection or classification for LLMs, requiring the identification, estimation, and relation of visual marks to extract information. For instance, the human ability to generalize an understanding of length to a previously unseen chart design or to estimate the ratios between lengths is a capability that, despite the limitations of CNNs, remains a challenge for MM-LLMs.

Our research methodology involves a comprehensive investigation into the abilities of MM-LLMs for visualization analysis. We evaluate the performance of three MM-LLMs, namely GPT-4, LLaVA, and a fine-tuned version of LLaVA, on visualization tasks. To ensure the validity of our findings, we draw on the graphical perception settings of Cleveland and McGill. These settings include nine reasoning tasks, such as position relative to a scale, length, angle, area, and shading density, and measure human graphical perception performance on bar and pie chart quantity estimation. We replicate these settings with our chosen MM-LLMs and compare their performance to human graphical perception and CNNs.

Our findings have practical implications for the use of LLMs in visualization analysis. While these models can regress most tasks, their answers are often random and do not match human baselines. However, we also discovered that fine-tuning LLMs can significantly improve performance, offering a promising approach for using them in graphical perception tasks. This insight into the potential of fine-tuning MM-LLMs could lead to improved performance and more accurate results in visualization analysis.

Additionally, we have implemented a framework to evaluate MM-LLMs systematically. This framework can be used for other tasks and with other LLMs, offering easy integration.

Our main contributions include:

1. Evaluating three MM-LLMs on elementary perceptual tasks

and position angle tasks introduced by Cleveland and McGill.

2. Fine-tuning LLaVA for these low-level visual tasks and comparing it with the general-purpose version.

3. Developing an open source framework for systematically evaluating MM-LLMs https://github.com/kenichi-maeda/LLMP/tree/main

2 Background and Research Problem

Graphical Perception. Cleveland and McGill's study [3] explores how people interpret graphical data, establishing a foundational framework for effective data visualization. They introduced the concept of "graphical perception" as the visual decoding of information from graphs, proposing that the effectiveness of a graph depends on the elementary perceptual tasks it invokes, such as position, length, and angle. Through experiments, they demonstrated that accuracy in interpreting data varies with the type of graphical task—judgments based on position are typically more accurate than those based on color or volume. Their findings led to practical guidelines for designing graphs that prioritize higher-order perceptual tasks, advocating for alternatives like dot charts over traditional forms like pie charts and bar graphs, thereby significantly influencing graphical design in statistics and related fields. Haehn et al. [5] evaluated these concepts using convolutional neural networks (CNNs), investigating how these models perform on graphical perception tasks initially designed for human assessment. They replicated Cleveland and McGill's experiments, applying CNNs to tasks like estimating bar chart lengths and pie segment angles, with mixed results. While CNNs matched human accuracy in specific scenarios, they generally did not model human graphical perception effectively, highlighting the challenges of using CNNs for data visualization analysis. This comparison tests the adaptability of CNNs to complex perceptual tasks and explores the boundaries of machine perception in graphical contexts.

Large Language Model Evaluation Recent studies have significantly contributed to evaluating and understanding LLMs in specific tasks, particularly in programming. Xu et al. [12] comprehensively evaluated LLMs tailored for programming tasks, comparing models like Codex, GPT-J, GPT-Neo, GPT-NeoX, and CodeParrot across various programming languages. They highlighted the limitations of Codex's non-open-source nature. They introduced PolyCoder, an open-source model that outperforms Codex in C programming, advocating democratizing advanced machine learning models for code generation.

Chang et al.[2] provided a detailed survey on evaluating LLMs, emphasizing the importance of robust evaluation metrics that assess task-specific performance and consider broader societal impacts. Their work underscores the evolving nature of LLMs and the need for standardized evaluation protocols to ensure a deeper understanding of LLM performance across diverse applications.

These studies collectively highlight the importance of evaluating LLMs in specific tasks, the need for open-source models to facilitate broader research and application opportunities, and the necessity of developing robust evaluation metrics and protocols to understand and utilize LLMs effectively.

3 Methodology

We evaluate MM-LLMs graphical perception capabilities and compare them against CNNs and human baselines in 11 total tasks across 3 experiments:

E1. We utilize Cleveland and McGill's elementary perceptual tasks to directly estimate quantities for visual marks (position, length, direction, angle, area, volume, curvature, and shading). (Section 4)

E2. We reproduce Cleveland and McGill's position-angle experiment that compares pie charts to bar charts. (Section 5)

3.1 Framework

To systematically evaluate LLMs for graphical perception tasks, such as those proposed by Cleveland and McGll, we developed a robust framework (Figure ef fig:framework) designed for assessing various scenarios and tasks across different LLMs. The framework is designed to simplify the integration of new LLMs. Importantly, the evaluation framework is not rigid, but adaptable to incorporate new evaluation metrics as needed, ensuring its future-proof nature. The framework accepts queries, generates different stimuli, and outputs evaluation metrics and visualizations.

To be more precise, the framework generates stimuli and ground truth for different tasks, sends queries to different LLMs, parses the answers, and then returns the results based on various evaluation metrics. We use metrics as explained in section 3.5. The purpose of evaluating LLMs is to understand their performance on graphical perception abilities. To gain a deeper understanding, our framework leverages visualization to provide the best results and compare different LLMs.

3.2 Multi modal Large language models

Introducing GPT-4, a powerful MM-LLM that is designed to process both text and image inputs. While it may have some limitations in certain real-world scenarios, it showcases exceptional performance on professional and academic benchmarks. In fact, it often matches or even surpasses human performance in a variety of tasks, setting a new standard in large language models.

LLaVA Large Language and Vision Assistant is an open-source MM-LLM. It represents the first end-to-end trained large multimodal model (LMM) that achieves impressive chat capabilities. We chose LLaVA for two primary reasons: 1) It is open-source, allowing us to fine-tune it for our specific use case. 2) It has been tuned with vision-language instruction data, and its performance in multimodal tasks usually surpasses that of language-only tasks [9].

Fine Tuned LLaVA. LLaVA is an MM-LLM that can be fine-tuned for specific tasks. We fine-tuned LLaVA with our stimuli to evaluate and compare if the fine-tuned version can better answer graphical perception questions regarding our visual marks. The number of training stimuli for fine-tuning is 2100.

3.3 Stimuli

We utilized the stimuli generator developed by Haehn et al [5], which generates stimuli for Cleveland and McGill's elementary perceptual tasks. Each image is a 100x100 binary image corresponding to one of the elementary tasks. Each image includes a ground truth, which we used for our evaluation.

3.4 Prompt Engineering Method

Prompt engineering is a technique that involves enhancing a large pre-trained model with task-specific hints, known as prompts, to adapt the model to new tasks [4]. The greatest limitation of prompt engineering is the difficulty of designing a prompt for a particular type of task and the lack of automated methods [11].

Gu et al. [4] comprehensively surveyed cutting-edge research in the prompt engineering of pre-trained Vision-Language Models (VLMs). We followed their Multimodal-Text Prompting Methods. They introduced two main categories for prompting: Hard and Soft. Hard prompts refer to human-interpretable natural language instructions used for prompting, while soft prompts are continuous vector representations optimized directly in the model's embedding space. Hard prompting is better suited for our case and has four sub-categories: Task Instruction Prompting, In-context Learning, Retrieval-based Prompting, and Chain-of-Thought Prompting. We performed an experiment for comparing prompts, and based on the results, we selected the Task Instruction Prompting method.

Our study's structure begins with a description of what the image contains for all experiments. MM-LLMs are then asked to estimate,

for example, the length of the angle and report it in a range. For instance, one example question is: "This image contains a simple line drawing that forms an acute angle. Please estimate the angle. Please respond with a possible range not larger than 10 degrees and report just the numbers." This is particularly noteworthy because we observed significantly improved results when we guide these MM-LLMs with queries; otherwise, they, especially GPT-4, occasionally fail to provide answers. In this study, we refer to prompts as queries since we use APIs for our experiment and we send queries.

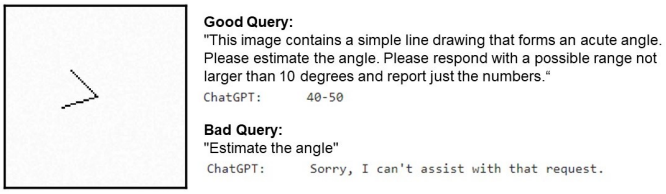

Figure 2: **Good query vs bad query**. MM-LMMs perform better with detailed guidance.

3.5 Measures and Analysis

Following Cleveland and McGill's methodology in their pioneering 1984 study, we utilize the mid-Mean Logistic Absolute Error (MLAE) as a metric to quantify the precision of graphical perception accuracy. This metric directly compares our evaluation, CNN evaluations by Haehn et al., and human performance baselines. The MLAE is computed as follows:

$$MLAE = \log_2(|predictedpercent - truepercent| + 0.125) \quad (1)$$

Our analysis also uses standard error metrics such as Mean Squared Error (MSE) and Mean Absolute Error (MAE) to comprehensively compare percentage errors across different models.

3.6 Human Baselines

We take human baseline measurements for the position-angle (E2) and position-length (E3) experiments from Cleveland and McGill [3], which had 51 participants. For the position-length experiment, we also utilize human baseline measurements from Heer and Bostock's crowdsourced reproduction of Cleveland and McGill's experiments [6], which involved 50 participants. Each participant in both studies reviewed 10 stimuli for each condition. We used baseline measurements from Haehn et al.[5] for the experiments E1 and E2.

3.7 Convolutional Neural Networks

We took CNN baseline measurements from Haehn et al. [5]. They evaluated graphical perception tasks across four CNN architectures. We use their measure and analysis to compare with our results.

4 Experiment 1: Elementary Perceptual Tasks

Cleveland and McGill describe a set of elementary graphical perceptual tasks across ten encodings, where each encodes a quantitative variable in a graphical element or visual mark. These tasks are the low-level building blocks for information visualizations 3: estimating position on a common scale, position on non-aligned scales, length, direction (or slope), angle, area, volume, curvature, and shading (or ink density). We leave color saturation experiments for future work.

We generated 55 versions of each elementary perceptual task as 100×100 raster images and tested the ability of MM-LLMs to regress quantitative values from these images. For instance, we created multiple vertical lines of varying lengths distributed across the x-axis- and y-axis for the length estimation task. We evaluated the three models' responses to these images using the task Instruction prompting method.

4.1 Hypotheses

H1.1: Regression Capability of MM-LLMs

The MM-LLM tested will be able to regress quantitative variables from graphical elements in all tasks. We generate different visual encodings and test whether the model can return a number as an answer. For simplicity, we asked the models to return a range for their prediction, and in our analysis, we considered the midpoint of the range to compare with single value ground truth.

H1.2: Superiority of Fine-Tuned Models

We expect a fine-tuned MM-LLM to perform better on all elementary perceptual tasks than a general-purpose MM-LLM.

4.2 Results

In our study, we conducted a comprehensive analysis of the performance of three MM-LLMs by evaluating their responses to our queries in different tasks. Our findings, which are detailed in the following sections, provide insights into each model's strengths and limitations.

H1.1. We observed that both LLaVA and fine-tuned LLaVA consistently returned numerical answers to our queries, whereas GPT-4 occasionally failed to provide a numerical response, instead returning non-numerical answers such as "I cannot assist you with that." Upon re-querying with the same stimuli and question, GPT-4 was able to eventually return a numerical answer. But GPT-4 was not able to answer our queries regarding the shading experiment, and it consistently returned, 'Your input image may contain content that is not allowed by our safety system.' Therefore, we do not have any results for this experiment. This pattern was particularly evident in the position-common task, where we had to send a total of 301 requests to obtain answers for 55 stimuli, indicating a failure rate of 246 cases. However, GPT-4 was able to return numerical answers for all requests in the curvature and direction tasks. GPT-4 returned a range for all of the answers as we requested, but LLaVA and fine-tuned LLaVA returned only a number. Based on these results, we **partially accept H1.1**.

H1.2. Results show that none of the MM-LLMs performed better among all tasks. We conclude that LLaVA was not able to be 1st in any tasks, and the competition was between fine-tuned LLaVA and GPT-4. Only the volume estimation task did a better job than the fine-tuned LLaVA. In most of the tasks (5/8), custom LLaVA performed the best. So we **reject H1.2**.

5 Experiment 2: Position-Angle

Cleveland and McGill measure how humans perceive the ratios of positions and angles through comparisons on bar charts and pie charts [10]. We create rasterized images following Cleveland and McGill's proposed encoding and investigate the computational perception of our networks (Fig. 1). These have five bar or pie sectors representing numbers that add up to 100, where each is greater than three and smaller than 39. One required change is in the minimal differences between the values: Cleveland and McGill create stimuli with a minimum scale difference of 0.1. However, as our networks only take 100×100 pixel images as input, we can only minimally represent a difference of 1 pixel. Cleveland and McGill ask participants to estimate the ratio of the four smaller bars or sectors to the known and marked largest bar or sector. As such, we mark the largest quantity of the five in each visualization with a single pixel dot. We ask our networks to perform multiple regression and produce four ratio estimates. Since the position of the largest element changes, we generate the targets such that the largest element

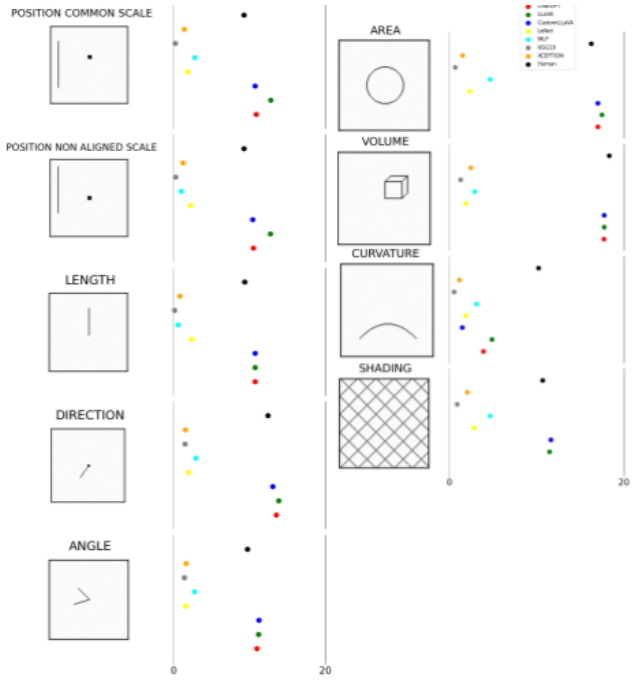

Figure 3: Comparison of our results for Experiment 1 with CNN results and human baseline. In all experiments, except for volume and curvature, LLMs did not perform better than human baselines and CNNs

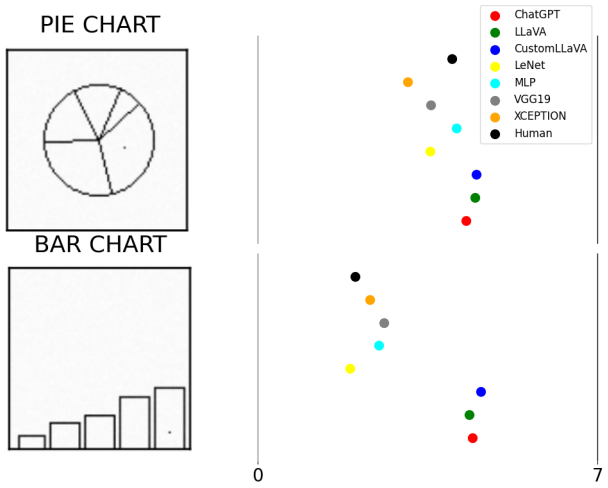

Figure 4: Comparison of our results for Experiment 2 with CNN and human baseline data. The results indicate that in both bar chart and pie chart tasks, LLMs could not outperform the CNN and human baseline

is marked with 1, and the smaller elements follow in a counter-clockwise direction for the pie chart and to the right for the bar chart. Each of the bar and pie chart visualizations has 878,520 possible permutations.

5.1 Hypotheses

H2.1: Perceptual Accuracy of Bar Charts vs. Pie Charts

Computed perceptual accuracy will be higher for bar charts than pie charts. Cleveland and McGill report that position judgments are almost twice as accurate (measured by MLAE) as angle judgments in humans. Following our ranking of elementary perceptual tasks (Table 2), we see that MM-LLMs also judge position encodings more accurately than angles. Therefore, our MM-LLMs will excel in interpreting bar charts compared to pie charts.

5.2 Results

In the position-angle experiment discussed in the paper, different models were tested for their ability to perceive ratios in bar and pie charts. The results indicated that the LLaVA model performed slightly better on bar charts (MLAE=4.3) compared to pie charts (MLAE=4.4). However, other models, including Custom LLaVA and GPT-4, showed better performance on pie charts. For instance, GPT-4 reported an MLAE of 4.2 for pie charts, outperforming its accuracy on bar charts (MLAE=4.4). This finding led to the **rejection of H2.1**. Additionally, GPT-4 frequently responded with "I can't assist" in cases involving bar charts, indicating a higher difficulty processing bar charts than pie charts for this model.

6 Discussion and Conclusion

This research has advanced our understanding of the graphical perception capabilities of MM-LLMs like GPT-4, LLaVA, and fine-tuned LLaVA. It has demonstrated that while these models show promise in certain scenarios, they do not universally surpass human performance or existing Convolutional Neural Network (CNN) benchmarks in graphical perception elementary tasks.

The study emphasizes the critical role of fine-tuning and model selection in achieving optimal performance on specific tasks. Fine-tuned LLaVA exhibited enhanced performance compared to its general-purpose counterpart, although it did not always outperform GPT-4. In all tasks, LLaVA consistently returned the same answer for the same query, indicating a limitation in its graphical perception capabilities. However, through fine-tuning, it was able to produce different and improved answers..This indicates that fine-tuning is beneficial but requires targeted approaches and extensive task-specific training data to achieve significant improvements.

Moreover, the research highlights the variability in MM-LLM performance across different visual encoding tasks, suggesting that these models have not yet matched human flexibility and intuition in graphical perception. The study also introduces a comprehensive framework for evaluating the graphical perception abilities of MM-LLMs, which could be pivotal for future research and development in this domain.

This research shows a meaningful step forward in exploring the capabilities of MM-LLMs in the field of graphical perception. It has implications for both the development of new visualization tools and the enhancement of existing models in interpreting visual data.

7 Future Work

For future research, we aim to comprehensively include and evaluate all the tasks originally proposed by Cleveland and McGill to provide a more exhaustive assessment of multi-MM-LLMs' graphical perception capabilities. Additionally, we plan to enhance the performance of the LLaVA model by fine-tuning it with a larger and more complex dataset to better understand the impacts of extensive training on the model's accuracy in graphical tasks.

Moreover, we intend to integrate newer and state-of-the-art MM-LLMs and compare their effectiveness against the current models. For example, Llama3, the latest version of LLaVA, has just been released. This will help identify the best practices and MM-LLMs that are most effective for graphical perception tasks.

REFERENCES

[1] L. Bojic, P. Kovacevic, and M. Cabarkapa. Gpt-4 surpassing human performance in linguistic pragmatics. *arXiv preprint arXiv:2312.09545*, 2023.

[2] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[3] W. S. Cleveland and R. McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984. doi: 10.1080/01621459.1984.10478080

[4] J. Gu, Z. Han, S. Chen, A. Beirami, B. He, G. Zhang, R. Liao, Y. Qin, V. Tresp, and P. Torr. A systematic survey of prompt engineering on vision-language foundation models. *arXiv preprint arXiv:2307.12980*, 2023.

[5] D. Haehn, J. Tompkin, and H. Pfister. Evaluating 'graphical perception'with cnns. *IEEE transactions on visualization and computer graphics*, 25(1):641–650, 2018.

[6] J. Heer and M. Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, p. 203–212. Association for Computing Machinery, New York, NY, USA, 2010. doi: 10.1145/1753326.1753357

[7] Z. Huang, Z. Zhang, Z.-J. Zha, Y. Lu, and B. Guo. RelationVLM: Making large vision-language models understand visual relations, 2024.

[8] S. Kim, S. Yun, H. Lee, M. Gubri, S. Yoon, and S. J. Oh. Propile: Probing privacy leakage in large language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[9] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

[10] H. Qu, Y. Cai, and J. Liu. Llms are good action recognizers. *arXiv preprint arXiv:2404.00532*, 2024.

[11] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–7, 2021.

[12] F. F. Xu, U. Alon, G. Neubig, and V. J. Hellendoorn. A systematic evaluation of large language models of code. In *Proceedings of the 6th ACM SIGPLAN International Symposium on Machine Programming*, pp. 1–10, 2022.

[13] Y. Yao, J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. A survey on large language model (llm) security and privacy: The good, the bad, and the ugly. *High-Confidence Computing*, 4(2):100211, 2024. doi: 10.1016/j.hcc.2024.100211