Intelligent Infrastructure Monitoring: Applications of VLMs and LLMs in Structural Health Monitoring

Yingchao Zhang[1[0000-0001-6225-4493]] and Cheng Liu[1,2[0000-0003-4174-2046]]

[1] Department of Systems Engineering, City University of Hong Kong, Kowloon, Hong Kong SAR, China
[2] Centre for Intelligent Multidimensional Data Analysis, City University of Hong Kong, Kowloon, Hong Kong SAR, China
cliu647@cityu.edu.hk

Abstract. The rapid development of artificial intelligence has brought transformative opportunities to the field of structural health monitoring (SHM) for civil infrastructure. Recent advancements in vision-language models (VLMs) and large language models (LLMs) have demonstrated remarkable capabilities across various tasks, including image classification, object detection, semantic segmentation, instance segmentation, and question-answering. These technologies enable comprehensive, efficient, and intelligent analysis of structural conditions, facilitating early detection of potential issues in complex infrastructures. This work explores the integration of these cutting-edge models into SHM workflows, showcasing their ability to process multimodal data (e.g., images, sensor data, and textual descriptions) and provide actionable insights. By leveraging the strengths of VLMs and LLMs, such as natural language understanding and advanced visual feature extraction, we propose novel applications for automated damage detection, anomaly assessment, and real-time monitoring. Preliminary results highlight the potential of these models to enhance decision-making processes and reduce human intervention in infrastructure maintenance. This study will provide an overview of state-of-the-art AI methodologies, discuss their strengths and limitations in the context of SHM, and outline future research directions for applying these technologies to improve the safety and resilience of modern civil infrastructure.

Keywords: Structural health monitoring, Vision language models, Large language models, Damage detection.

1 Introduction

As the process of globalization continues to deepen and the level of urbanization continues to increase, road transport infrastructure plays a crucial role in the development of modern society. The significance of road networks lies in their role as a vital conduit for economic activities across diverse regions. The state of these networks exerts a direct influence on the economic development of nations, the quality of life of their populations, and the public's safety. However, in the process of long-term use, pavement

inevitably develops various kinds of defects, such as cracks, potholes, rutting, loosening and surface settlement. These defects not only affect the comfort of road use, but also seriously threaten driving safety. These challenges place high demands on road management and maintenance authorities for timely maintenance. However, road maintenance is a significant financial undertaking, and a considerable number of regions worldwide continue to grapple with the challenge of inadequate pavement defects management [1].

The development of pavement defects is an incremental process, initiated by microscopic cracks and subsequently exacerbated by the combined effect of environmental factors and traffic loads. Early detection and remediation of the defect can result in substantial cost savings in repair expenditures and prolong the lifespan of the highway. This "prevention is more important than cure" concept in the field of road management has formed a consensus, and the key to the realization of this concept is to establish an efficient and accurate pavement defects detection system.

Traditional pavement defects detection mainly relies on manual inspection, and this method has many limitations. Firstly, manual inspection is inefficient and difficult to apply to a vast road network. Secondly, the detection results are greatly influenced by the experience and subjective judgment of the inspectors, resulting in a lack of consistency in the assessment results. Furthermore, there is a safety risk associated with manual inspection in busy highway sections. Lastly, the manual recording and data management methods make it difficult to accumulate and analyze historical data, which is detrimental to the study of the development law of pavement defects. With the expansion of highway network scale and the improvement of maintenance standards, these problems become more and more prominent, and the traditional inspection methods can no longer meet the needs of modern road management.

In recent years, due to the rapid advancements in computer vision, artificial intelligence, and sensor technology, automated pavement defects detection technology has undergone substantial progress. Road inspection vehicles based on laser scanning, high-resolution image acquisition and three-dimensional reconstruction have been put into use in some developed countries and regions [2]. However, the existing technology still faces many challenges. On the one hand, the accuracy of defects identification in complex environments needs to be improved, such as the accuracy of defect identification under different lighting conditions, weather conditions, and pavement backgrounds [3]. On the other hand, the high cost of high-precision inspection equipment makes this hinders its widespread adoption, particularly in resource-constrained regions. Additionally, most systems are deficient in real-time capabilities, impeding the ability to make immediate decisions. The data processing capacity is also constrained, which complicates the analysis and mining of substantial pavement inspection data [4].

The rise of deep learning techniques has brought new opportunities for pavement damage detection. Convolutional Neural Networks (CNN), and the latest visual Transformer model [5] have demonstrated excellent performance in the field of image recognition, which provides a powerful tool for automatic pavement damage detection. Studies have shown that deep learning-based methods have been able to achieve an accuracy rate of over 90% in typical defects recognition tasks such as cracks and potholes, significantly surpassing traditional computer vision methods [6]. At the same time, the

improvement of edge computing and mobile device computing power allows lightweight deep learning models to be deployed on mobile platforms, creating conditions for real-time, low-cost pavement damage detection [5]. The development of multi-sensor fusion technology, such as combining RGB images, depth information, thermal imaging and other multimodal data, further improves the robustness of detection and can adapt to more complex and changing environmental conditions [7]. However, the existing CNN-based pavement defects detection technology still faces many challenges. CNN and other models mainly focus on visual feature extraction and lack in-depth understanding of the semantic information of the defect, which leads to easy misdetection in complex backgrounds. In addition, the traditional deep learning models usually require a large amount of labeled data for training, and the acquisition of high-quality labeled data for pavement defects is costly. Finally, existing models tend to be optimized for specific types of diseases and lack the general ability to deal with diversified defects [8].

The development of Vision-Language Models (VLMs) has introduced a novel approach to addressing these challenges, representing a significant advancement in the field of Artificial Intelligence (AI) [9]. VLMs leverage advanced technologies to establish a profound correlation between visual content and textual descriptions by concurrently processing visual and linguistic data. In contrast to conventional, purely visual models, VLMs possess the capability to comprehend and generate natural language descriptions associated with images, thereby offering a novel technical framework for the detection of pavement defects. VLMs acquire rich visual-linguistic knowledge through pre-training, which enables efficient transfer learning with very little labeled data and significantly reduces data annotation costs. In addition, VLMs can incorporate expert knowledge in the form of text into the model in conjunction with language comprehension capabilities to enhance semantic understanding of pavement defects. Recent research has demonstrated the remarkable potential of VLMs in the domain of roadway damage detection. Zhang and Liu [10] employed the Contrastive Language–Image Pre-training (CLIP) model [11] for crack classification and found that it demonstrated better generalization ability than a specially trained CNN model on unseen datasets. Liang, et al. [12] used the CLIP to enhance the crack segmentation performance of neural networks, which greatly improved the accuracy of crack segmentation.

However, there are fewer applications of these models in the field of structural health monitoring. This study aims to further explore the strengths and weaknesses of these models in the field of structural health detection. This study utilizes the latest Qwen vision language model developed by Alibaba [13] to explore its applications in scenarios such as pavement damage classification, defects analysis and so on.

The following of this paper is organized as follows: Section II introduces the Qwen and Llama models; Section III describes the experimental design and evaluation methods; Section IV presents the experimental results and makes a detailed analysis and comparison; finally, Section V summarizes the whole paper and looks forward to the future research direction.

2 Methodology

VLMs represent a significant advancement in the domain of Artificial Intelligence, offering a transformative approach to multimodal understanding and interaction. These models fundamentally change the manner in which machines interpret and analyze complex information across diverse domains through the seamless integration of visual perception and natural language processing capabilities. VLMs possess the capacity to concurrently process visual inputs, such as images, videos, and other visual data, as well as text, thereby facilitating cross-modal comprehension, inference, and generation. This capability provides a robust foundation for a range of application scenarios.

Currently common VLMs include GPT-4o [14], Llama [15], and Qwen-VL [13]. These models can understand what is in the image and are also able to continue complex reasoning to answer questions. And Qwen2.5-VL, as the latest flagship VLMs series developed by Alibaba's Qwen team, demonstrates significant technological advances, and thus was chosen by us as the base model for this study. Llama [15] was chosen as our comparison model.

2.1 Qwen-2.5-VL Model

Currently common VLMs include GPT-4o [14], Llama [15], and Qwen-VL [13]. These models can understand what is in the image and are also able to continue complex reasoning to answer questions. And Qwen2.5-VL, as the latest flagship VLMs series developed by Alibaba's Qwen team, demonstrates significant technological advances, and thus was chosen by us as the base model for this study.

Qwen2.5-VL incorporates a windowed attention mechanism, reducing the computational complexity of its visual encoder from quadratic to linear, thereby significantly enhancing inference efficiency. The model also employs dynamic resolution processing, allowing it to process images of varying sizes in their original dimensions while representing object positions using actual pixel coordinates instead of normalized coordinates. For video processing, it introduces dynamic frame rate sampling and absolute time coding, enabling more accurate comprehension of temporal information without incurring additional computational overhead.

In addition, the pre-training data scale was expanded from 1.2 trillion to 4 trillion tokens, incorporating more diverse content. To ensure data quality and relevance, the team developed an image-text data scoring system. Additionally, a comprehensive document parsing system was designed to unify various elements, such as tables, charts, and formulas, into an HTML format for streamlined processing. The multilingual OCR dataset was expanded to improve the model's multilingual comprehension capabilities. Furthermore, a specialized subtitle dataset was constructed for long videos, enhancing the model's ability to process extended video content effectively.

In this paper, the processing flow using Qwen-2.5-VL is shown in Fig. 1. Initially, the user prompts the system to describe the category and characteristics of defects in the image using a cue word. Subsequently, the algorithm processes input images or videos containing structural damage, which may include various types of defect manifestations, such as structural anomalies like cracks, as illustrated in Fig. 1. The input

visual data and user prompts are processed by an Encoder structure, which transforms the raw data into standardized feature representations. The Encoder outputs feature vectors that are organized into sequences, where green blocks denote text features and blue blocks represent image features. This multimodal feature fusion approach effectively integrates diverse information sources. The resulting feature sequences are then input into the Qwen-2.5-VL vision language model, which leverages its pre-trained vision-language comprehension capabilities to perform an in-depth analysis of the structural damage presented in the input.

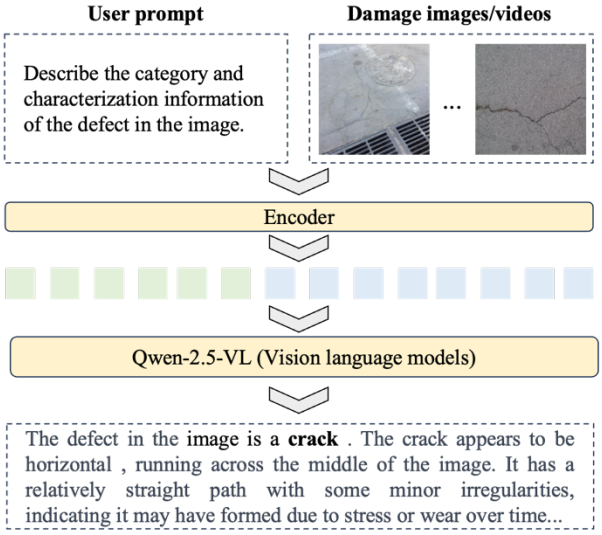

Fig. 1. Flowchart of Structural Health Inspection Algorithm based on Qwen-2.5-VL

The model ultimately generates a detailed defect description, including defect type identification (e.g., "*crack*") and a comprehensive characterization encompassing the defect's orientation ("*horizontal*"), location ("*running across the middle of the image*"), morphological features ("*relatively straight path with some minor irregularities*"), and possible causes of formation ("*may have formed due to stress or wear over time*"). This structured output equips engineers with comprehensive defect assessment information, facilitating informed maintenance decision-making.

2.2 Llama-3.2-11B-Vision

As a comparative model of Qwen-2.5-VL, Llama-3.2-11B-Vision is the latest generation of multimodal vision language models introduced by Meta AI [15], extended based on the Llama-3.2 architecture with 11 billion parameter scales. In terms of architectural design, Llama-3.2-11B-Vision utilizes the classical combination of a vision encoder and a language decoder, mapping visual features to the language model's representation space via a projection layer.

Llama-3.2-11B-Vision leverages Meta's large-scale multimodal dataset. However, the size and diversity of its training data differ significantly from those of Qwen-2.5-

VL, which is trained on a dataset comprising 4 trillion tokens. In contrast, the training dataset for Llama-3.2-11B-Vision is comparatively smaller. Furthermore, Qwen-2.5-VL places greater emphasis on data processing in specialized domains, such as document comprehension and diagram parsing, and has developed a specialized document full-parsing dataset. These efforts provide Qwen-2.5-VL with clear advantages in handling structured documents.

3 Experiments

3.1 Evaluation metrics

As this study emphasizes the quality of text generation by various VLMs, we employed metrics from the field of natural language processing to evaluate their performance. Rouge-1, Rouge-2, Rouge-L, BLEU, Meteor, and Bert-score were used as our evaluation metrics.

The Rouge family of metrics (Rouge-1, Rouge-2, Rouge-L) serves as the primary method for evaluating the quality of text summarization and generation. Rouge-1 and Rouge-2 assess the overlap between the candidate text and the reference text at the levels of single words (unigrams) and two consecutive words (bigrams), respectively. In contrast, Rouge-L focuses on the longest common subsequence, which captures more flexible word order matches and is more tolerant of variations in sentence structure. These metrics are widely used to evaluate content coverage in summarization, question-answering, and text generation systems. The BLEU metric is primarily used in machine translation evaluation to calculate the n-gram exact match between a candidate translation and a reference translation. It balances accuracy and completeness by penalizing overly short translations and is considered a classic evaluation metric in the field of machine translation. In recent years, BLEU has also been applied to the evaluation of other text generation tasks. Meteor, on the other hand, is a more comprehensive evaluation metric that not only accounts for exact word matching but also incorporates synonym, stemming, and paraphrase matching while considering word order. It balances precision and recall by calculating the harmonic mean, offering a fairer score for semantically similar yet differently worded expressions. Bert-score represents a new generation of evaluation metrics based on pre-trained language models like BERT, leveraging contextualized word embeddings to measure the semantic similarity between a candidate text and a reference text.

These metrics serve as the foundation for the evaluation of VLMs in this study.

3.2 Experimental platforms

Table 1 provides a detailed overview of the hardware and software configuration utilized in this study. Regarding hardware, the computing system is equipped with an Intel Xeon processor as CPU and an NVIDIA A100-40GB high-performance GPU. The system includes 72GB of RAM, offering ample memory for large-scale model execution and data processing. In terms of the software environment, the system operates on the Ubuntu 22.04 operating system and utilizes Python 3.12 as the primary programming

language. For deep learning applications, the system is configured with CUDA 11.8 to maximize GPU computing capabilities and employs PyTorch 2.1.2 as the deep learning framework. Additionally, the system integrates the Transformers 4.49.0 library to implement and deploy various models based on the Transformer structure.

Table 1. Software and hardware versions

Content	Version	Content	Version
CPU	Intel Xeon Processor (Skylake, IBRS)	Cuda	11.8
GPU	Nvidia A100-40GB	RAM	72GB
OS	Ubuntu 22.04	PyTorch	2.1.2
Python	3.12	Transformers	4.49.0

3.3 Dataset

In this study, 134 images of pavement damage were selected, and their corresponding text descriptions were generated in the form of manual annotation. An example is shown in Fig. 2. In this study, we directly compare the outputs of the Qwen or Llama model with manually labeled results to calculate various evaluation metrics.

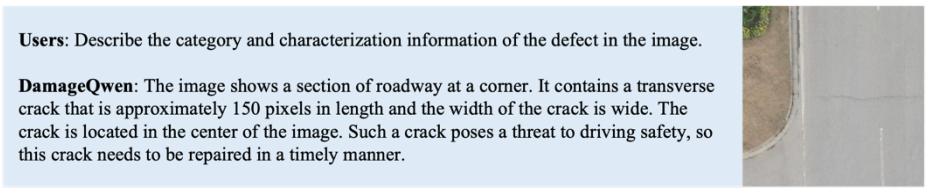

Fig. 2. An example of the dataset

4 Results and Discussions

4.1 Results of different models

Table 2 compares the performance of two VLMs, Qwen-2.5-VL and Llama-3.2-11B-Vision, on the pavement defect detection task. Based on multidimensional evaluation metrics, Qwen-2.5-VL demonstrates superior performance across all six measures. In the Rouge-series metrics, which rely on n-gram matching, Qwen-2.5-VL achieves a Rouge-1 score of 0.3845, significantly surpassing Llama-3.2-11B-Vision's score of 0.3382, reflecting a 13.7% relative advantage in word-level matching. Similarly, Qwen-2.5-VL outperforms Llama-3.2-11B-Vision on the Rouge-2 and Rouge-L metrics, with scores of 0.1674 and 0.2428 compared to 0.1352 and 0.2144, representing relative improvements of 23.8% and 13.2%, respectively.

On the BLEU metric, despite the overall low scores of both models (attributable to the differences between the defect description task and traditional translation tasks), Qwen-2.5-VL outperforms Llama-3.2-11B-Vision, achieving a score of 0.0533

compared to Llama-3.2-11B-Vision's 0.0476, reflecting a relative advantage of 12.0%. The Meteor metric further corroborates this trend, with Qwen-2.5-VL scoring 0.2595, significantly surpassing Llama-3.2-11B-Vision's 0.2231, an improvement of 16.3%. This indicates that Qwen-2.5-VL demonstrates superior performance in terms of semantic completeness and accuracy. Notably, in the Bert-score evaluation, which best reflects the depth of semantic comprehension, the two models exhibit relatively close performance. Qwen-2.5-VL slightly leads with a score of 0.8874 compared to Llama-3.2-11B-Vision's 0.8812, a marginal difference of only 0.0062. This observation suggests that, while there are significant differences in surface text matching features, the gap in deep semantic comprehension ability between the two models is relatively small. This may be attributed to their shared use of advanced Transformer architectures and large-scale pre-training strategies, which enable both models to achieve a high level of semantic representation.

Table 2. Detection results of different vision language models

Model	Rouge-1	Rouge-2	Rouge-L	BLEU	Meteor	Bert-score
Qwen-2.5-VL	0.3845	0.1674	0.2428	0.0533	0.2595	0.8874
Llama-3.2-11B-Vision	0.3382	0.1352	0.2144	0.0476	0.2231	0.8812

The comprehensive analysis reveals that Qwen-2.5-VL outperforms Llama-3.2-11B-Vision in the defect detection task, likely due to its larger parameter size, richer training data, and specific optimizations for visual content understanding.

4.2 Zero-shot detection capability for VLMs

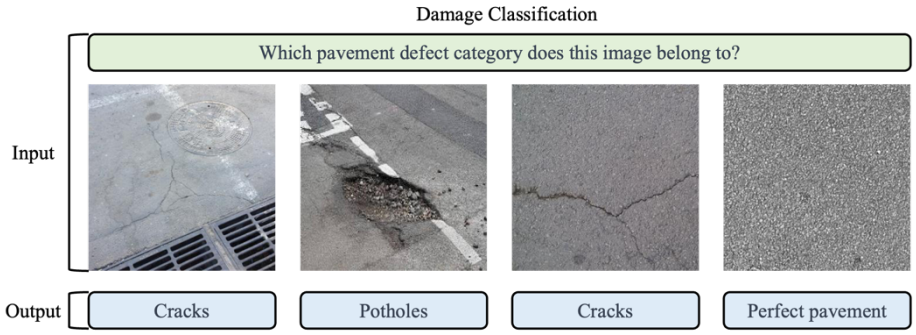

Fig. 3. Zero-shot classification results

In addition to employing metrics to assess the model's performance, we evaluated the zero-shot classification capability of Qwen-2.5-VL. The image is input directly into Qwen-2.5-VL, and subsequently, the category to which the image belongs is output. The results are shown in Fig. 3. As demonstrated by the Fig. 3, Qwen-2.5-VL is capable of accurately identifying the presence of pavement defects without requiring samples training. This indicates that the present VLMs have strong zero-shot capability.

In practice, the robust zero-shot capability significantly reduces the deployment threshold for road inspection systems. When engineers encounter new or rare types of pavement damage, there is no need to collect a large dataset or train a specialized model. Instead, the system can directly analyze and classify the damage using existing knowledge. This feature is particularly crucial for addressing seasonal variations, unique distresses caused by extreme weather, or unknown damage types associated with new pavement materials. By enabling rapid analysis, it dramatically shortens response times and allows for proactive intervention to mitigate potential safety hazards.

4.3 Discussions

Experimental results demonstrate that VLMs show significant potential in structural health monitoring. As shown in Section 4.1, Qwen-2.5-VL outperforms Llama-3.2-11B-Vision across all evaluation metrics, particularly with a 23.8% relative advantage in Rouge-2, reflecting its superior capability in describing structural details. Meanwhile, the zero-shot capability test in Section 4.2 confirms these models' flexibility in accurately classifying pavement defects without specialized training, offering valuable applications for identifying newly emerging structural damage.

However, VLMs still face several challenges in structural health monitoring. First, they show notable limitations in precise object detection and semantic segmentation. Second, these models encounter computational efficiency issues when processing high-resolution images common in engineering applications, making them unsuitable for real-time monitoring systems. Additionally, VLMs have limited capabilities in quantitative analysis (such as crack width and area measurements), excelling primarily in qualitative descriptions. Finally, existing models lack sufficient depth in understanding specialized structural health monitoring knowledge, struggling to accurately differentiate between surface cracks and structural cracks or predict structural evolution under specific conditions.

Therefore, future research should consider integrating VLMs with specialized engineering models to leverage the strengths of both approaches, developing more comprehensive intelligent structural health monitoring systems.

5 Conclusion

This study demonstrates the promising application of vision language models in structural health monitoring. Our experiments show Qwen-2.5-VL outperforming Llama-3.2-11B-Vision across all metrics, highlighting its effectiveness in pavement defect detection. The zero-shot capabilities of these models offer particular value for field applications by eliminating the need for extensive labeled datasets.

Despite these advantages, current VLMs face challenges including lower precision compared to specialized detection models, computational inefficiency with high-resolution images, limited quantitative assessment abilities, and insufficient domain knowledge. Future research should focus on developing hybrid approaches combining VLMs' semantic understanding with engineering models' precision, while optimizing computational efficiency for real-time monitoring. As these limitations are addressed,

VLMs hold significant potential to transform infrastructure inspection practices through more accessible and comprehensive assessment tools.

References

1. Zhang, Y., Ma, Z., Song, X., Wu, J., Liu, S., Chen, X., and Guo, X.: Road surface defects detection based on IMU sensor. IEEE Sensors Journal 22(3), 2711-2721 (2021).
2. Alrajhi, A., Roy, K., Qingge, L., and Kribs, J.: Detection of road condition defects using multiple sensors and IoT technology: A review. IEEE Open Journal of Intelligent Transportation Systems 4, 372-392 (2023).
3. Zhang, Y., and Liu, C.: Crack segmentation using discrete cosine transform in shadow environments. Automation in Construction 166, 105646 (2024).
4. Zhang, Y., Zuo, Z., Xu, X., Wu, J., Zhu, J., Zhang, H., Wang, J. and Tian, Y.: Road damage detection using UAV images based on multi-level attention mechanism. Automation in construction, 144, 104613 (2022).
5. Zhang, Y., and Liu, C.: Real-time pavement damage detection with damage shape adaptation. IEEE Transactions on Intelligent Transportation Systems 25(11), 18954 – 18963 (2024).
6. Zhang, Y., and Liu, C.: Network for robust and high-accuracy pavement crack segmentation. Automation in Construction, 162, 105375 (2024).
7. Tan, Y., Deng, T., Zhou, J., and Zhou, Z.: LiDAR-based automatic pavement distress detection and management using deep learning and BIM. Journal of Construction Engineering and Management, 150(7), 04024069 (2024).
8. Zhang, Y., and Liu, C.: Generative adversarial network based on domain adaptation for crack segmentation in shadow environments. Computer‐Aided Civil and Infrastructure Engineering. Early access (2025).
9. Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023).: Visual instruction tuning. In: Advances in neural information processing systems 36 (NeurIPS 2023), pp. 34892-34916. NIPS Foundation, USA.
10. Zhang, Y., and Liu, C.: Few-shot crack image classification using clip based on Bayesian optimization. arXiv preprint arXiv: 2503.00376 (2025).
11. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J. and Krueger, G.: Learning transferable visual models from natural language supervision. In: International conference on machine learning, pp. 8748-8763 (2021).
12. Liang, F., Li, Q., Yu, H., and Wang, W.: CrackCLIP: Adapting Vision-Language Models for Weakly Supervised Crack Segmentation. Entropy 27(2), 127 (2025).
13. Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., ... and Qiu, Z. Qwen2.5 technical report. arXiv preprint arXiv:2412.15115 (2024).
14. Hurst, A., Lerer, A., Goucher, A. P., Perelman, A., Ramesh, A., Clark, A., ... and Kivlichan, I.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024).
15. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M. A., Lacroix, T., ... and Lample, G. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023).