Edge AI-Enhanced Traffic Monitoring and Anomaly Detection Using Multimodal Large Language Models

Ryan Peruski[1], Abhilasha Saroj[2], Wenjun Zhou[3], Seddik Djouadi[1], and Charles Cao[1]

[1] Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996;
e-mail: yhg461@vols.utk.edu, mdjouadi@utk.edu, cao@utk.edu
[2] Oak Ridge National Laboratory, Oak Ridge, TN 37830; e-mail: sarojaj@ornl.gov
[3]Department of Statistics, Operations and Management Science, University of Tennessee, Knoxville, TN 37996; e-mail: wzhou4@utk.edu

ABSTRACT

This paper addresses the challenge of traffic monitoring and incident detection in remote areas, utilizing multimodal large language models (LLMs) deployed on edge AI devices. The key novelty of the LLM is to convert real-time video streams into descriptive texts, enabling low-bandwidth transmissions and reliable detection of anomalies and incidents in environments of intermittent connectivity. The model is developed based on fine-tuning open-source LLMs and extending it with multi-modal capabilities to analyze video frames. Our work also involves deploying this model on edge devices such as Nvidia IGX Orin and is planned to be tested in realistic environments as future work. The methodology includes dataset curation, iterative model fine-tuning and compression, and hardware-based optimization. This approach aims to enhance traffic safety and response speed in remote areas, marking a significant advancement in the application of AI for traffic monitoring and safety management.

INTRODUCTION

The advent of Generative Artificial Intelligence (GenAI) technologies (Vaswani 2017), particularly Large Language Models (LLMs) (Achiam et al. 2023), has opened new frontiers in intelligent transportation systems and smart mobility. In this work, we use cutting-edge multimodal LLMs to address critical challenges in traffic monitoring and anomaly detection, especially in remote and inaccessible areas.

Remote regions, such as rural roads and sprawling wilderness routes, present unique challenges for timely incident detection and response. Traditional surveillance systems face limitations due to intermittent network connectivity, high bandwidth requirements for real-time video transmission, and constraints on continuous operation. These challenges create significant gaps in monitoring and incident response, potentially delaying emergency services and compromising safety in areas with low population density.

The deployment of fine-tuned multimodal LLMs on edge computing devices at high-risk intersections presents a transformative solution for rapid incident response in transportation systems. By automatically analyzing and interpreting traffic incidents in real-time, these models can significantly reduce the critical time between accident occurrence and emergency response deployment. This capability is particularly valuable at busy intersections with historically high accident rates, where immediate and accurate incident classification can streamline communication with appropriate authorities, whether it be emergency medical services, traffic management teams, or law enforcement. Edge-computing implementation ensures rapid processing without manual operations, enabling instant decision-making even in areas with limited connectivity, thereby potentially saving crucial minutes in emergency situations where every second counts.

Our research vision centers on developing a cutting-edge multimodal LLM tailored specifically for intelligent traffic monitoring and anomaly detection on edge computing devices. It has the following key contributions:

- **GenAI for Mobility**: We leverage generative deep learning techniques, specifically multimodal LLMs, to analyze and interpret complex traffic scenarios from video feeds.
- **Infrastructure Sensing**: Our approach involves deploying sophisticated AI models on edge devices, contributing to the development of advanced urban sensing infrastructure and edge computing capabilities.
- **Human Dynamics Analysis**: By converting real-time events into descriptive summaries, our system enhances the modeling and analysis of traffic flow and human movement patterns in remote areas.
- **Cyberinfrastructure**: Our research contributes to the development of AI-powered platforms that facilitate the acquisition, management, and analysis of mobility data, even in challenging environments with limited connectivity.

Our methodology encompasses a comprehensive approach to address these challenges:

- Dataset curation, combining existing traffic incident datasets with newly collected data from remote areas.

- Selection and fine-tuning of a foundational multimodal LLM, adapting it to understand and interpret complex traffic scenarios.
- Iterative fine-tuning to continuously improve the model's performance on specific tasks related to traffic monitoring and incident detection.
- Advanced compression and quantization techniques for deploying these sophisticated models on edge devices, balancing performance and computational efficiency.

This methodological framework ensures that our system can operate effectively in resource-constrained environments while maintaining high accuracy in detecting and reporting critical traffic events. By doing so, we aim to enhance traffic safety and response speed in remote areas, marking a significant advancement in the application of AI for traffic monitoring and safety management.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work. Section 3 details our methodology, including dataset curation, model selection, fine-tuning processes, and optimization techniques for edge deployment. In Section 4, we present our preliminary results and discuss the performance of our system in detecting and classifying various traffic incidents. We also analyze the challenges encountered and propose potential solutions. Finally, Section 5 concludes the paper.

RELATED WORK

Recent advancements in multimodal LLMs have shown promising results in visual understanding tasks (Liu et al. 2023). These models can process both text and image inputs, allowing for a more comprehensive analysis of visual scenes. However, their application to traffic monitoring, especially in resource-constrained edge environments, remains largely unexplored.

On the other hand, edge computing in transportation systems has gained significant attention in recent years (Shi et al. 2016). These approaches bring computation closer to the data source, reducing latency and bandwidth requirements. However, they often lack the reasoning capabilities required for complex traffic scenarios, which our proposed multimodal LLM aims to address.

For testing the effectiveness of traffic understanding, we utilize the Car Accident Detection and Prediction (CADP) dataset (Shah et al. 2018). The dataset contains 230 videos, each with at least one accident captured from fixed traffic camera views, and 1,416 segments of traffic accidents. Additionally, 205 segments with HD quality are annotated with spatio-temporal data for object detection, tracking, and collision detection. The CADP dataset is particularly relevant as it focuses on traffic accidents captured from fixed third-person views, which aligns with our goal of monitoring remote areas using stationary edge devices. Therefore, we choose this dataset for fine-tuning our selected models.

METHODOLOGY

Dataset Curation

We consider various traffic scenes, including incidents and a range of traffic conditions typically encountered in remote areas. To efficiently annotate the video frames, we utilize existing LLMs (GPT, Gemini, Claude) for initial text descriptions. This will be followed by careful human refinement for accuracy, particularly for critical events like accidents or wildlife encounters.

Our focus is on developing a comprehensive taxonomy for remote traffic incidents, ensuring the dataset captures a broad spectrum of anomalies and potential disruptions. To further enhance the quality and robustness of the dataset, we implement an iterative feedback loop, through retraining open-source LLM models with the evolving data. This process will place emphasis on examples where the model exhibits low confidence, ensuring continuous improvement. Data curation strictly adheres to ethical standards, prioritizing privacy by removing personally identifiable information.

Foundational Model Selection

Our work builds upon the strengths of cutting-edge open-source multimodal LLMs and chooses the most suitable existing model as the basis for fine-tuning. Our selection process prioritizes models that demonstrate strong multimodal understanding of images and text, proven transfer learning capabilities, and potential for computational efficiency. We carefully evaluate various candidate models from Meta, Google, among others. We choose LLaVA, which is based on the Llama series, for later experiments.

Iterative Fine-tuning

The iterative fine-tuning stage is the most computationally expensive of the visual language representation learning process. It involves multiple rounds of training the model on a curated dataset to enhance the large language model's (LLM's) ability to interpret complex traffic images accurately. The fine-tuning process is inspired by BLIP-2 (Li et al. 2023), but with added temporal cross-attention in the Q-Former design.

Specifically, the LLM is first subjected to a round of training on the prepared dataset. During this training, the LLM learns to map the visual features extracted from the image (using the Q-Former) to the corresponding textual descriptions. After the initial training, the model's performance is evaluated on a separate validation set. This helps identify areas where the LLM struggles to accurately interpret the visual content. Based on the evaluation results, the LLM is fine-tuned further on a subset of the training data specifically chosen to address the identified shortcomings. These steps are repeated iteratively. The LLM is continuously evaluated and finetuned on carefully selected subsets of the training data to progressively improve its performance on the task of interpreting complex traffic images. The iterative finetuning process continues until a pre-defined stopping criterion is met. This criterion could be based on the LLM achieving a desired level of accuracy on the validation set, or after a fixed number of iterations.

Compression and Quantization

In this step, we reduce the model size through compression and quantization, so that the resulting model can fit on resource-constrained edge computing devices. We aim to achieve a good trade-off between cost and accuracy of the resulting model.

Distillation: This method uses a larger model to train a smaller model, which can enhance memory and computing efficiency by reducing the number of parameters. In this research direction, we will use the more powerful model as the teacher model and train a smaller student model with proactive confidence estimation. This confidence parameter will be used to guide the triage and model selection decision-making process in the field, with a focus on leveraging the right model given the computing capability of edge computing hardware, as well as load balancing between a small triage model and sending more critical cases/signals to a larger model.

Quantization: This method converts full precision parameters (e.g., 16-bit float) into discrete values. However, since this quantization step is done after training, the conversion can result in performance degradation as the quantized values are not optimized in a larger context. Alternatively, Quantization-Aware Training (QAT) attempts to train a model with quantization in mind. Recent work includes binarized networks for convolutional neural networks, binarized Transformers for machine translation and BERT, as well as promising results on QAT in LLMs (BitNet) (Wang et al. 2023). QAT provides a good balance between model compression and has not been done on an image-to-text LLM especially for multi-frame image-based event detection.

EVALUATION

Our evaluation focuses on two key aspects: (1) the accuracy of multimodal Large Language Models (LLMs) in traffic incident detection, and (2) the latency performance of these models when deployed on edge devices.

Dataset Description

Our experiments utilize a subset of the CADP (CCTV traffic camera based Accident Analysis Dataset Platform) dataset (Shah et al. 2018). The original dataset consists of numerous traffic video sequences captured as frame sequences of varying lengths. For our study, we extracted and manually annotated 200 sequences with ground truth labels for training and inference purposes.

A notable characteristic of our dataset subset is its high concentration of accident scenarios, with approximately 97% of sequences containing traffic incidents. While this distribution may not be representative of real-world traffic patterns, it allows us to focus our analysis on accident severity assessment rather than binary accident detection. The dataset predominantly features T-bone collisions and rear-end collisions, aligning with accident type distributions reported by the National Highway Traffic Safety Administration (NHTSA) in 2017, where these collision types, along with angled collisions, were identified as the most frequent accident categories.

This dataset composition enables us to specifically address the challenge of accident severity classification, though we acknowledge that future work should incorporate a more balanced dataset for accident detection tasks. A potential extension of this research could involve developing or implementing a preliminary model for accident detection before routing cases to our severity assessment model.

Accuracy Evaluation

We conducted experiments using a subset of 100 traffic video sequences from the CADP dataset, evaluating three models: OpenAI's GPT-4 Vision, LLaVA, and VILA (both based on the LLama series). Each sequence contains multiple frames capturing potential traffic incidents.

GPT-4 Vision Analysis

Initial experiments with GPT-4 Vision focused on basic accident identification using binary (yes/no) questions. Figure 1 shows the model's performance in detecting accidents.

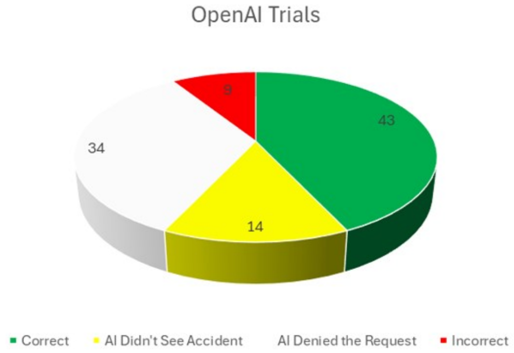


Figure 1. GPT-4 Vision performance in accident detection, showing 43% correct identifications but 34% denied requests due to safety limitations

The results revealed significant limitations: GPT-4 Vision correctly identified accidents in only 43% of cases, while in 34% of cases it denied the request due to safety limitations or content restrictions. Given these constraints, we did not proceed with more detailed accident analysis using GPT-4 Vision.

LLaVA Analysis

We next evaluated LLaVA's basic ability to identify accidents in traffic scenes. Figure 2 shows the model's overall performance in this initial trial.

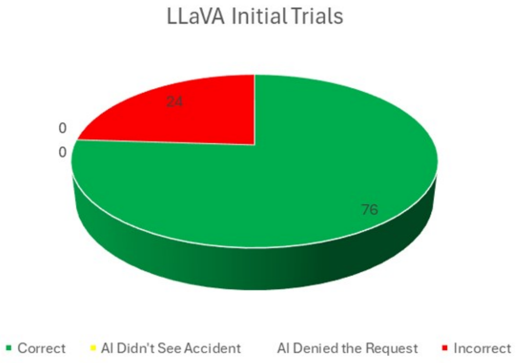
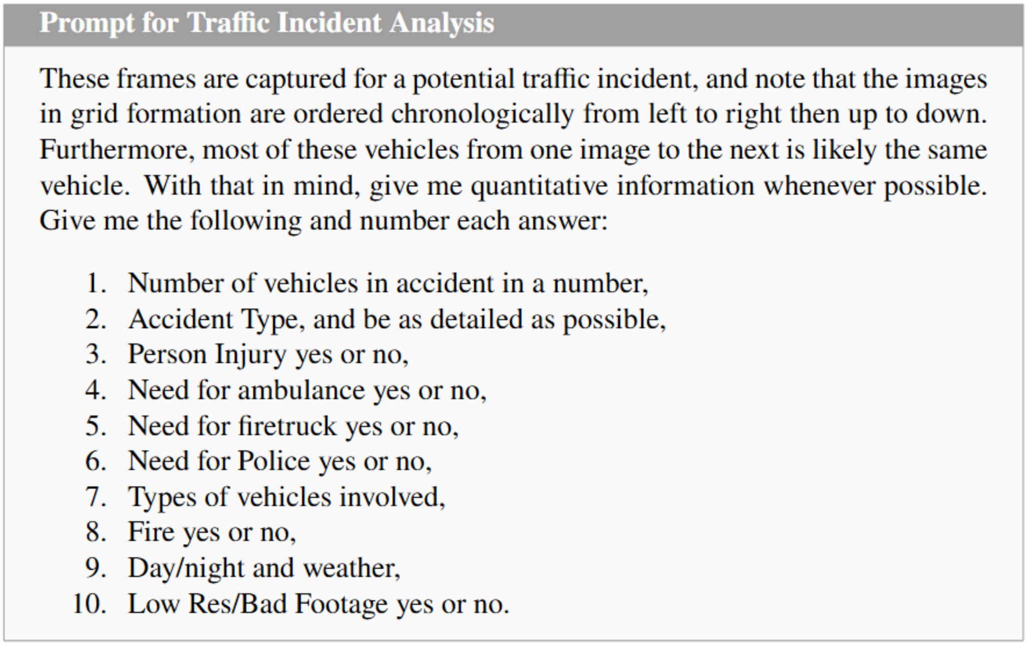


Figure 2. LLaVA's performance in initial trials showing 76% correct identifications with no denied requests, demonstrating strong reliability in basic accident detection

LLaVA achieved correct responses in 76% of cases, with incorrect responses in 24% of cases. Notably, there were no instances where LLaVA failed to see an accident or denied the request, showing robust reliability in basic accident detection tasks.

For detailed analysis, we developed a comprehensive prompt template:



Prompt for Traffic Incident Analysis

These frames are captured for a potential traffic incident, and note that the images in grid formation are ordered chronologically from left to right then up to down. Furthermore, most of these vehicles from one image to the next is likely the same vehicle. With that in mind, give me quantitative information whenever possible. Give me the following and number each answer:

1. Number of vehicles in accident in a number,
2. Accident Type, and be as detailed as possible,
3. Person Injury yes or no,
4. Need for ambulance yes or no,
5. Need for firetruck yes or no,
6. Need for Police yes or no,
7. Types of vehicles involved,
8. Fire yes or no,
9. Day/night and weather,
10. Low Res/Bad Footage yes or no.

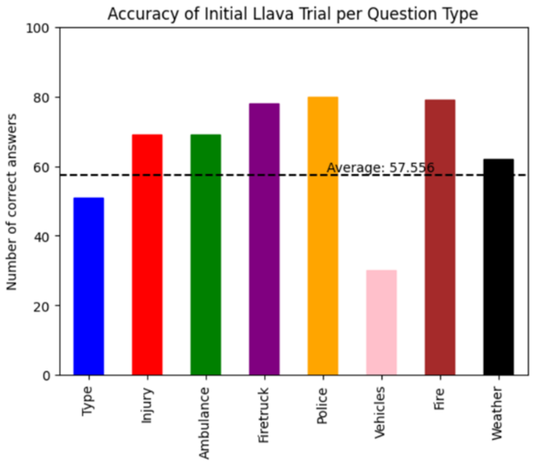


Figure 3. LLaVA's accuracy across different question types, showing strong performance in police presence (80%) and fire detection (78%), but lower accuracy in vehicle type identification (30%)

Using this detailed prompt, LLaVA showed varying performance across different question types, achieving an average accuracy of 57.55%. As shown in Figure 3, the model performed particularly well in identifying police presence (80%) and fire-related incidents (78%), but struggled with accident type classification (50%) and vehicle type identification (30%).

VILA Model Analysis

VILA demonstrated improved overall performance with an average accuracy of 62.67%. As shown in Figure 4, it excelled in specific tasks such as fire detection (95%)

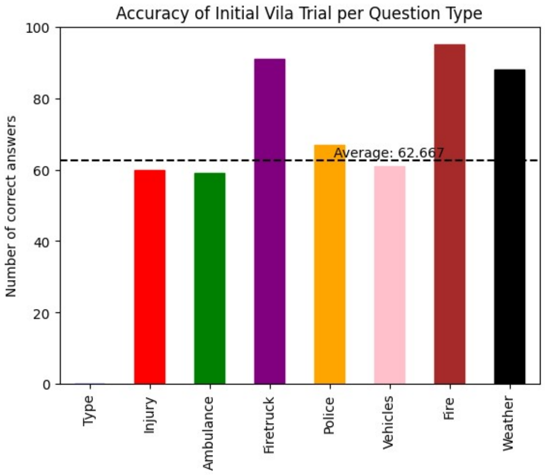


Figure 4. VILA's accuracy across different question types, demonstrating superior performance in fire detection (95%) and firetruck identification (90%), with an improved overall average of 62.67%

and firetruck identification (90%). However, like LLaVA, it struggled with accident type classification, often defaulting to generic responses like "traffic accident" or "traffic collision" without providing specific details.

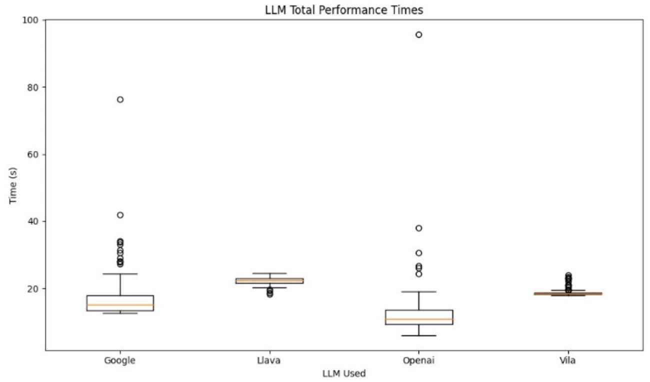


Figure 5. Response time distribution comparison between models, highlighting VILA’s consistent performance around 19 seconds compared to other models’ greater variability

Performance Analysis

Response time analysis revealed distinct patterns between the models. As shown in Figure 5, VILA demonstrated remarkable consistency with a narrow interquartile range centered around 19 seconds. LLaVA showed similar consistency but with slightly higher response times (approximately 22 seconds). In contrast, GPT-4 Vision showed greater variability, with response times occasionally reaching up to 95 seconds.

Fine-tuning Challenges

Our attempts at fine-tuning the models revealed several consistent challenges. First, models tended to overfit even with 100 training sequences, often converging on simplified, generic responses. Second, the imbalanced nature of our dataset, particularly for binary classifications like fire presence (where negative cases significantly outnumber positive ones), led to biased model responses. These challenges suggest that effective fine-tuning for traffic monitoring applications requires substantially larger and more balanced datasets, potentially beyond the scope of a single research team.

Latency Evaluation of LLaVA

In this section, we evaluate the performance of LLaVA under various configurations to assess its suitability for edge deployment. Specifically, we choose different model sizes (7B and 13B parameters) and quantization levels (4-bit and 8-

bit). The evaluation was conducted on L4 GPU, which has a more comparable TFLOPS compared to our target edge platform than high-end GPUs such as A100 or H100. Hence, the latency measurement should be more relevant for edge computing environments.

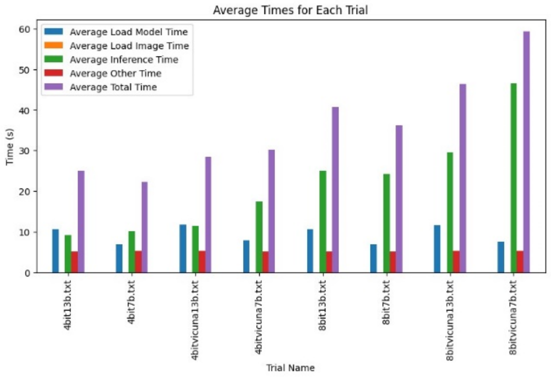


Fig. 6. Trial latency measurements for different LLaVA configurations

Figure 6 shows the average latency measurements for different LLaVA configurations, demonstrating the trade-offs between model size, quantization level, and processing speed.

Discussions

The evaluation results demonstrate the potential of multimodal LLMs, particularly LLaVA, for traffic monitoring and incident detection in remote areas. We conclude that using open-source models seem more preferable than commercial models. In our planned work, we also aim to fine-tune the open models to further improve its performance. The latency evaluation reveals trade-offs between model size, quantization level, and inference time. The 4-bit quantified 7B model shows promise for edge deployment, balancing performance and latency. Our future work will focus on fine-tuning LLaVA on diverse traffic incident datasets, optimizing edge performance through advanced compression techniques, exploring prompt engineering strategies, and conducting extensive real-world testing.

CONCLUSION

This paper has presented a comprehensive evaluation of multimodal large language models for traffic monitoring and incident detection in remote areas, with a particular focus on edge AI deployment. Our experimental results demonstrate that not only commercial models, but also open-source models like LLaVA and VILA can achieve competitive performance in analyzing traffic incidents. The key advantage of our approach lies in converting real-time video streams into descriptive texts, enabling efficient monitoring even in areas with limited connectivity.

The research also uncovered significant challenges that need to be addressed. Model fine-tuning proved particularly demanding due to data imbalance issues, especially in binary classification tasks like fire detection. Commercial models, despite their powerful capabilities, showed inconsistent behavior due to built-in safety limitations, highlighting the need for specialized solutions in traffic monitoring applications.

Looking ahead, our research points to several crucial directions for advancing this technology. Future work should focus on developing more comprehensive and balanced datasets specifically for traffic incident analysis in remote areas. This includes gathering diverse data across different weather conditions, lighting scenarios, and incident types. Additionally, we plan to investigate advanced compression techniques and quantization-aware training methods to optimize model deployment on edge devices such as the Nvidia IGX Orin platform.

ACKNOWLEDGEMENT

This paper is in part supported by USDA/NIFA under contract 2023-67021-40613, AI TENNessee Initiative Seed Fund, and by UT-Battelle, LLC, under contract DE-AC05-00OR227252 with the US Department of Energy (DOE). The publisher acknowledges the US government license to provide public access under the DOE Public Access Plan (https://www.energy.gov/doepublic-access-plan).

REFERENCES

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al. (2023). “GPT-4 technical report.” arXiv preprint arXiv:2303.08774.

Li, J., Li, D., Savarese, S., and Hoi, S. (2023). “BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.” International Conference on Machine Learning, PMLR, 19730–19742.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. (2023). “Visual Instruction Tuning.” arXiv preprint arXiv:2304.08485.

Shah, A. P., Lamare, J.-B., Nguyen-Anh, T., and Hauptmann, A. G. (2018). “CADP: A novel dataset for CCTV traffic camera based accident analysis.” 2018 15th

IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), IEEE, 1–7.

Shi, W., Cao, J., Zhang, Q., Li, Y., and Xu, L. (2016). “Edge computing: Vision and challenges.” IEEE Internet of Things Journal, 3(5), 637–646.

Vaswani, A. (2017). “Attention is all you need.” Advances in Neural Information Processing Systems.

Wang, H., Ma, S., Dong, L., Huang, S., Wang, H., Ma, L., Yang, F., Wang, R., Wu, Y., and Wei, F. (2023). “BitNet: Scaling 1-bit transformers for large language models.” arXiv preprint arXiv:2310.11453.