Position Paper: Compressed Models Should Be Robust and Publicly Shared: A Call for Responsible Model Optimization

Rahma Fourati
REGIM-Lab.: REsearch Groups
in Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS),
BP 1173, Sfax, 3038, Tunisia.
Université de Jendouba,
Faculté des Sciences Juridiques, Economiques
et de Gestion de Jendouba, 8189, Jendouba, Tunisie.
rahma.fourati@ieee.org

Jihene Tmamna
REGIM-Lab.: REsearch Groups
in Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS),
BP 1173, Sfax, 3038, Tunisia.
jihen.tmamna@enis.tn

Fadoua Drira
REGIM-Lab.: REsearch Groups
in Intelligent Machines, University of Sfax,
National Engineering School of Sfax (ENIS),
BP 1173, Sfax, 3038, Tunisia.
fadoua.drira@ieee.org

Javier Sanchez-Medina
Innovation Center for the Information Society,
University of Las Palmas de Gran Canaria, Spain.
javier.sanchez@ulpgc.es

***Abstract*—This position paper argues that compressed deep learning models obtained through pruning, quantization, or knowledge distillation should be both robustly evaluated and publicly shared to support responsible model optimization. While model compression has emerged as a vital strategy for enabling efficient deployment and reducing carbon footprint, the machine learning community often shares only the techniques, not the optimized models themselves. Furthermore, these compressed models are rarely evaluated for robustness in terms of uncertainty, adversarial resilience, and sensitivity factors critical for safe deployment. We call for a shift in research culture: robustness evaluation must precede sharing, and sharing should be standard practice for models trained on public benchmarks. This approach will promote reproducibility, democratize access, and align machine learning research with environmental and ethical goals.**

***Index Terms*—Uncertainty Quantification,Adversarial Robustness,Responsible AI, Model Compression**

I. Introduction

This position paper argues that compressed deep learning models, particularly those optimized via pruning, quantization, or knowledge distillation, should be both robust and publicly shared.

Modern machine learning (ML) research has witnessed remarkable advances in model compression, driven by the need to deploy deep neural networks (DNNs) efficiently on resource-constrained devices and in energy-sensitive environments. Techniques such as pruning [1]–[4], quantization [5]–[7], and knowledge distillation [8], [9] have become standard tools in the ML optimization toolkit. These methods promise reduced inference latency, memory footprint, and carbon emissions, aligning with the goals of Green AI [10].

However, the community often overlooks two critical aspects. First, the resulting compressed models are seldom released, even when based on public datasets such as ImageNet. Second, these models are rarely evaluated beyond accuracy, leaving open concerns about robustness to adversarial attacks, calibration, uncertainty estimation, and out-of-distribution generalization [11]–[13].

While compression facilitates deployment in resource-constrained environments, it may degrade robustness. For example, quantized models can exhibit increased sensitivity to noise, while pruning can shift decision boundaries in ways that expose vulnerabilities. Thus, **compressed models must undergo systematic robustness evaluations** before being shared. These include:

- **Uncertainty quantification**, to ensure models provide calibrated confidence in predictions [14], [15].
- **Adversarial robustness**, to assess whether compression introduces vulnerabilities [16], [17].
- **Sensitivity analysis**, to study how input or internal perturbations affect output [18].
- **Signal-preserving pruning**, which maintains generalization by amplifying relevant features and removing noise [19]–[21].

We contend that benchmark compression papers (e.g., com-

pressing ResNet on ImageNet) should be accompanied by:

1) Open-source release of compressed model checkpoints.
2) Robustness evaluations on uncertainty, adversarial risk, and sensitivity.
3) Diagnostic tools and reproducible evaluation pipelines.

This shift in practice would improve transparency, trustworthiness, and real-world adoption of efficient models.

II. Background and Motivation

The rapid evolution and widespread adoption of large language models (LLMs) have transformed the AI landscape in just a few years. These models now contain hundreds of billions of parameters and perform complex tasks spanning natural language understanding, generation, reasoning, and multi-modal processing. This explosive growth in scale and capability significantly increases computational costs and energy consumption, making it impractical for many researchers and practitioners to train or deploy such models without effective compression. Consequently, there is an urgent need not only to compress these models efficiently but also to ensure their robustness and transparency. Failure to do so risks proliferating brittle or opaque systems that are difficult to reproduce, evaluate, or trust, undermining the broader benefits of LLM advances.

To address these challenges, model compression techniques such as pruning, quantization, distillation, and low-rank factorization have been extensively studied. These methods aim to reduce the model size and complexity, thereby enabling deployment on resource-constrained devices and accelerating inference.

Among these techniques, *pruning* removes redundant or less important parameters; *quantization* reduces the precision of model weights and activations; *distillation* transfers knowledge from a large teacher model to a smaller student model; and *low-rank factorization* approximates large weight matrices with products of smaller, lower-rank matrices, effectively reducing the number of parameters and operations.

A crucial benefit of model compression is the significant reduction in computational complexity, which directly translates to lower energy consumption and carbon emissions associated with training and deploying DNNs. This reduction aligns with the goals of *Green AI*, which advocates for the development of efficient and environmentally sustainable artificial intelligence systems.

Despite these advantages, current research often focuses predominantly on the methodological improvements of compression techniques, while overlooking two critical aspects. First, the *public sharing of optimized, compressed models* is rarely practiced, hindering reproducibility and broad accessibility. Second, comprehensive *robustness evaluation* of compressed models including uncertainty quantification, adversarial robustness, and sensitivity analysis remains limited, raising concerns about the reliability and safety of deploying such models in real-world applications.

This position paper calls for a paradigm shift: compressed models should not only reduce complexity and environmental impact but also be rigorously evaluated for robustness and shared publicly to promote transparency, reproducibility, and sustainable AI practices.

III. Model Compression Techniques and Their Impact

Model compression encompasses a variety of techniques designed to reduce the size and computational demands of DNNs while attempting to preserve their predictive performance. Among the most prominent approaches are pruning, quantization, distillation, and low-rank factorization.

Pruning removes redundant or less salient parameters from a trained network, often based on magnitude thresholds or learned importance scores. This results in sparse models with fewer active connections, reducing both memory footprint and inference time.

Quantization involves lowering the numerical precision of weights and activations, for example, from 32-bit floating-point to 8-bit integer or even binary formats. This reduces the model size and enables more efficient hardware utilization, particularly on edge devices.

Distillation transfers knowledge from a large, complex teacher model to a smaller student model by training the latter to mimic the teacher's outputs. This process can yield compact models that retain much of the teacher's accuracy while requiring fewer resources.

Low-rank factorization approximates large weight matrices by decomposing them into products of smaller, lower-rank matrices. This reduces the number of parameters and matrix operations, lowering both storage and computation costs.

These compression methods contribute significantly to reducing model complexity, which directly impacts energy consumption and carbon footprint during both training and inference phases. For example, pruning and quantization can decrease the number of multiply-accumulate operations, which are computationally expensive and energy-intensive. Similarly, low-rank factorization reduces the dimensionality of weight matrices, streamlining computations.

By lowering computational demands, compressed models promote *Green AI* a movement emphasizing the environmental sustainability of AI research and deployment. However, this environmental benefit should not come at the cost of compromised model robustness and reliability.

Despite advances in compression methods, the AI community often overlooks the importance of thorough robustness assessments and the open sharing of compressed models trained on standard benchmarks. Without these practices, reproducibility suffers, and the trustworthiness of compressed models in safety-critical applications remains uncertain.

In the following sections, we discuss the need for comprehensive robustness evaluations of compressed models and argue for a community-wide commitment to publicly releasing these models to foster reproducibility, transparency, and sustainable AI innovation.

IV. Robustness Evaluation of Compressed Models

While model compression reduces complexity and environmental impact, it can also introduce vulnerabilities that affect the robustness and reliability of neural networks. Compression techniques often alter the internal representations and decision boundaries of models, which may degrade their performance under distribution shifts, adversarial attacks, or uncertainty [22]–[24].

This issue is especially critical for *large language models* (LLMs) and other models tackling complex ML tasks such as reasoning, multi-modal understanding, and long-context generation. Due to their massive size and intricate architectures, compression via pruning, quantization, distillation, and low-rank factorization is essential for practical deployment. However, these techniques can have nuanced impacts on robustness that demand careful evaluation.

A. Uncertainty Quantification in Complex Tasks

Complex ML tasks often require models to express calibrated uncertainty, especially in domains like healthcare diagnosis, legal document analysis, or scientific discovery. Compression can degrade uncertainty quantification by affecting the model's internal representations and confidence calibration [25]. For LLMs handling multi-step reasoning or generating diverse outputs, poorly quantified uncertainty can lead to misleading or harmful decisions.

Compression-aware uncertainty methods must be developed, including efficient Bayesian approximations or ensembles adapted to compressed architectures [26]. Moreover, uncertainty estimates should be evaluated not only on standard benchmarks but also on out-of-distribution or rare event scenarios, which are common in complex applications.

B. Adversarial Robustness in Generative and Reasoning Tasks

Adversarial robustness is particularly challenging for models performing generation, reasoning, or multi-modal fusion. For instance, LLMs are vulnerable to adversarial prompt attacks and data poisoning that can manipulate generated content or reasoning chains [27]–[29].

In multi-modal models (e.g., vision-language transformers), compression can distort the alignment between modalities, increasing susceptibility to cross-modal adversarial perturbations. Robustness evaluation frameworks must incorporate domain-specific adversarial tests, including semantic-preserving perturbations, prompt injection attacks, and subtle input manipulations that affect reasoning or factual accuracy.

C. Sensitivity and Stability in Long-Context and Multi-Step Reasoning

LLMs often rely on long-context attention and multi-step reasoning mechanisms, which require stable internal dynamics. Compression techniques that reduce parameter redundancy or apply low-rank factorization can alter these dynamics, potentially causing instability or brittleness in outputs [30].

Sensitivity analysis should be extended to assess how compressed models handle small perturbations in input prompts, intermediate reasoning steps, or memory representations. Instabilities may manifest as hallucinations, inconsistent answers, or degraded coherence in generated sequences, which are critical failure modes in complex tasks.

D. Evaluation Protocols and Benchmarks for Complex ML Tasks

Existing benchmarks for compressed models mostly focus on classification accuracy or simple NLP tasks. To properly evaluate robustness for complex tasks, new benchmarks should be developed that include:

- **Reasoning robustness:** Tests measuring stability and correctness of multi-step logical inference under compression.
- **Generation fidelity:** Evaluations of output quality, coherence, and factuality when generating long documents or conversations.
- **Multi-modal alignment:** Assessments of robustness to perturbations in multi-modal inputs after compression.
- **Uncertainty under distribution shifts:** Evaluations on out-of-domain or adversarially shifted inputs, especially in safety-critical domains.
- **Adversarial resistance:** Comprehensive adversarial attack suites customized for generative and multi-step tasks.

Developing such evaluation suites will require community collaboration to define relevant metrics, datasets, and protocols reflecting real-world complexities.

E. Key Insights

Robustness evaluation of compressed models is a multifaceted challenge, particularly for LLMs and models addressing complex ML tasks. To ensure safe, reliable, and environmentally sustainable deployment [31], future research must focus on:

- Designing compression techniques that preserve robustness in reasoning, generation, and multi-modal fusion.
- Developing scalable uncertainty quantification and adversarial defense methods tailored for compressed architectures.
- Establishing comprehensive benchmarks reflecting the complexity and diversity of real-world tasks.

Failing to address these robustness aspects risks undermining the benefits of model compression by producing brittle or unsafe AI systems despite their reduced computational footprint.

V. The Importance of Public Sharing of Compressed Models

Transparency and reproducibility are fundamental principles in scientific research, including ML. While many studies propose novel compression techniques, the resulting compressed models are often not publicly released. This lack of sharing impedes the reproducibility of results, slows progress, and limits the real-world applicability of compressed models.

Publicly sharing compressed models trained on standard benchmarks such as ImageNet or CIFAR-10 allows researchers and practitioners to:

- **Validate and reproduce** published results, ensuring that compression methods truly deliver on promised efficiency gains without sacrificing accuracy or robustness.
- **Benchmark and compare** different compression techniques under consistent conditions, fostering fair and transparent evaluation.
- **Accelerate research** by enabling others to build upon optimized models without repeating costly training and compression procedures.
- **Facilitate deployment** of efficient models in practical applications, including resource-constrained environments like mobile devices and edge computing.
- **Promote sustainable AI** by reducing redundant computation and energy consumption associated with retraining large models.

Moreover, sharing compressed models aligns with the broader movement toward Open Science and Green AI, encouraging responsible and sustainable innovation. It also supports the development of robust and trustworthy AI systems by enabling extensive robustness evaluations by the community.

Despite these clear benefits, there remain technical, legal, and cultural challenges to sharing compressed models, such as concerns over intellectual property, privacy, and the lack of standardized formats. Addressing these barriers requires community-driven initiatives, including developing repositories, defining metadata standards, and incentivizing open sharing through publication policies and funding agency requirements.

In summary, we urge the ML community to adopt best practices for publicly releasing compressed models, thereby enhancing reproducibility, fostering collaboration, and promoting environmentally sustainable AI research and deployment.

VI. Challenges and Future Directions

Despite the clear benefits of compressed models, several challenges hinder their robust evaluation and widespread public sharing.

A. Technical Challenges

- **Standardization of Robustness Metrics:** There is no universally accepted benchmark or suite of robustness metrics tailored specifically for compressed models. Developing such standardized evaluations is critical for fair comparison and trustworthiness assessment.
- **Compression-Aware Robustness Techniques:** Existing robustness methods often focus on full-size models and may not translate effectively to compressed counterparts. Novel approaches that account for the unique characteristics of pruning, quantization, distillation, and low-rank factorization are needed.
- **Efficient Evaluation Protocols:** Robustness testing, especially adversarial robustness, can be computationally expensive. Efficient protocols are necessary to enable large-scale evaluations without negating the computational savings gained by compression.

B. Sharing and Reproducibility Barriers

- **Intellectual Property and Privacy Concerns:** Organizations may hesitate to release compressed models due to proprietary information or data privacy issues.
- **Lack of Standardized Formats and Repositories:** Unlike trained full models, compressed models may use diverse formats and frameworks, complicating sharing and reuse.
- **Incentives and Cultural Norms:** The academic and industrial culture currently prioritizes novel methods over sharing artifacts. Incentives such as badges, reproducibility tracks, and funding requirements can encourage openness.

C. Future Research Directions

- **Developing Compression-Robustness Theories:** Formalizing the relationship between compression and robustness to guide principled design of compression methods.
- **Unified Frameworks for Compression and Robustness:** Creating tools that integrate compression with robustness evaluation and certification in a seamless workflow.
- **Green AI Metrics Beyond FLOPs:** Expanding sustainability metrics to include energy consumption, carbon footprint, and hardware-specific efficiency.
- **Community Platforms for Model Sharing:** Establishing accessible, well-maintained repositories with rich metadata and standardized interfaces for compressed models.

Addressing these challenges will be essential to realize the full potential of compressed models as efficient, robust, and environmentally sustainable AI tools.

VII. Conclusion

Model compression techniques such as pruning, quantization, distillation, and low-rank factorization have emerged as vital tools to reduce the computational demands of large neural networks, significantly lowering energy consumption and carbon footprint. This aligns closely with the principles of green AI, aiming to make artificial intelligence both efficient and environmentally sustainable. By enabling the deployment of large models like LLMs on limited hardware, compression democratizes access to powerful AI technologies.

However, the benefits of compression come with substantial challenges. Compression can alter model behavior in subtle yet critical ways, potentially degrading robustness, reliability, and safety. This is especially important in complex ML tasks involving multi-step reasoning, long-context generation, and multi-modal understanding, where brittle or miscalibrated models can lead to severe downstream consequences. Moreover, the evaluation of robustness in compressed models remains underdeveloped, with a lack of standardized benchmarks that capture the unique vulnerabilities introduced by compression.

In this position paper, we emphasize the urgent need for a comprehensive research agenda that couples compression with rigorous robustness evaluation. Such evaluation must encompass uncertainty quantification to ensure calibrated confidence, adversarial robustness to defend against malicious inputs, and sensitivity analysis to maintain output stability. For large language models and other complex architectures, this requires domain-specific tests and benchmarks reflecting real-world challenges, including out-of-distribution scenarios and adversarial prompt attacks.

Equally critical is the call for transparency and openness: sharing compressed models alongside original ones is necessary to foster reproducibility, accelerate research progress, and facilitate safer AI deployment. Without access to these models, efforts to evaluate robustness and improve methods remain severely limited.

Looking forward, we envision a research ecosystem where efficiency, environmental sustainability, and trustworthiness are jointly optimized. This requires collaboration across the AI community to develop compression algorithms that inherently preserve robustness, design evaluation frameworks tailored to complex tasks, and establish community-driven benchmarks. Only through these concerted efforts can the promise of compressed models be fully realized without compromising safety or reliability.

We urge researchers, practitioners, and policymakers to recognize robustness and transparency as first-class priorities in the model compression paradigm. By doing so, the AI community can advance towards responsible, green, and robust artificial intelligence that benefits society at large.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Ministry of Higher Education and Scientific Research of Tunisia under grant agreement number LR11ES48, and grant number PEJC2025-D3P01.

REFERENCES

[1] S. Han, H. Mao, and W. J. Dally, "Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding," *arXiv preprint arXiv:1510.00149*, 2015.

[2] D. Blalock, J. J. G. Ortiz, J. Frankle, and J. Guttag, "What is the state of neural network pruning?" *Proceedings of Machine Learning and Systems*, 2020.

[3] J. Tmamna, E. B. Ayed, R. Fourati, A. Hussain, and M. B. Ayed, "A cnn pruning approach using constrained binary particle swarm optimization with a reduced search space for image classification," *Applied Soft Computing*, vol. 164, p. 111978, 2024.

[4] J. Tmamna, E. B. Ayed, R. Fourati, M. Gogate, T. Arslan, A. Hussain, and M. B. Ayed, "Pruning deep neural networks for green energy-efficient models: A survey," *Cognitive Computation*, vol. 16, no. 6, pp. 2931–2952, 2024.

[5] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2704–2713, 2018.

[6] A. Gholami, S. Kim, Z. Dong, Z. Yao, M. Mahoney, and K. Keutzer, "A survey of quantization methods for efficient neural network inference," *Proceedings of the IEEE*, vol. 109, no. 9, pp. 1494–1531, 2021.

[7] J. Tmamna, E. B. Ayed, R. Fourati, A. Hussain, and M. B. Ayed, "Barebones particle swarm optimization-based quantization for fast and energy efficient convolutional neural networks," *Expert Systems: The Journal of Knowledge Engineering*, vol. 41, no. 4, 2024.

[8] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," *arXiv preprint arXiv:1503.02531*, 2015.

[9] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *International Journal of Computer Vision*, vol. 129, pp. 1789–1819, 2021.

[10] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[11] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *International Conference on Machine Learning*, 2017, pp. 1321–1330.

[12] M. Hein and M. Andriushchenko, "Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 41–50, 2019.

[13] L. Carbone, F. Kratsch, and K. Kersting, "Robustness for compressed neural networks: a survey," *arXiv preprint arXiv:2301.11841*, 2023.

[14] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *NeurIPS*, 2017.

[15] Y. e. a. Ovadia, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *NeurIPS*, 2019.

[16] Y. e. a. Guo, "Sparse dnns with improved adversarial robustness," *NeurIPS*, 2018.

[17] A. e. a. Shafahi, "Adversarial training for free!" *NeurIPS*, 2019.

[18] S. e. a. Hooker, "Characterising the sensitivity of pruning methods to data ordering," in *ICLR*, 2020.

[19] T. e. a. Zhou, "Optimal subnetwork structure for robustness in pruned neural networks," *ICLR*, 2022.

[20] H. Tanaka, D. Kunin, D. Yamins, and S. Ganguli, "Pruning neural networks without any data by iteratively conserving synaptic flow," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[21] J. Frankle and M. Carbin, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," in *ICLR*, 2019.

[22] ——, "The lottery ticket hypothesis: Finding sparse, trainable neural networks," *ICLR*, 2019.

[23] B. Jacob, S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, "Quantization and training of neural networks for efficient integer-arithmetic-only inference," in *CVPR*, 2018.

[24] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPS Deep Learning Workshop*, 2015.

[25] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017.

[26] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *ICML*, 2016.

[27] R. Jia and P. Liang, "Adversarial examples for evaluating reading comprehension systems," in *EMNLP*, 2017.

[28] E. Wallace, S. Feng, N. Kandpal, M. Gardner, and S. Singh, "Universal adversarial triggers for attacking and analyzing nlp," in *EMNLP-IJCNLP*, 2019.

[29] P. V. Dantas, L. C. Cordeiro, and W. S. Junior, "A review of state-of-the-art techniques for large language model compression," *Complex & Intelligent Systems*, vol. 11, no. 9, p. 407, 2025.

[30] S. Bubeck *et al.*, "Sparks of artificial general intelligence: Early experiments with gpt-4," *arXiv preprint arXiv:2303.12712*, 2023.

[31] R. Fourati, "Position paper: The environmental cost of deep learning: A call for sustainable ai practices," in *2025 International Joint Conference on Neural Networks (IJCNN)*, 2025, pp. 1–6.