

Robust Foundation Model Architectures

Hugo Rossi¹, Jonas Schmidt², Ivan Petrov³

¹ Research Scientist, School of Data Science, Nordic Technical University, Stockholm, Sweden. Email: hugo.rossi995@yahoo.com | ORCID: 8492-9028-5217-1322

² Senior Lecturer, Department of Artificial Intelligence, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: jonas.schmidt641@yahoo.com | ORCID: 0605-8119-8546-5055

³ Professor, Department of Machine Learning, European Institute of AI, Berlin, Germany. Email: ivan.petrov642@yahoo.com | ORCID: 7785-6152-3160-3525

ABSTRACT

Foundation models -- large pre-trained models that serve as a base for a wide range of downstream tasks -- have achieved remarkable capability but face systematic robustness challenges: sensitivity to input distribution shift, adversarial vulnerability, miscalibration of uncertainty estimates, and degradation under out-of-distribution conditions. These robustness failures are not peripheral concerns but core limitations that prevent deployment in safety-critical applications where distributional robustness and calibrated uncertainty are non-negotiable. This study systematically evaluates robustness across six leading foundation model families (GPT-4, Claude 3.5, Gemini 1.5, LLaMA-3-70B, Mistral-Large, and CLIP-ViT-L/14) on five robustness dimensions: distribution shift robustness (ImageNet-C, WILDS, DomainBed), adversarial robustness (AutoAttack, AdvGLUE), calibration (ECE on domain-shifted data), out-of-distribution detection, and uncertainty quantification quality. Additionally, six robustness enhancement techniques are evaluated: adversarial training, data augmentation (AugMax, RandAugment), ensemble methods, test-time adaptation (TTT++, TENT), uncertainty calibration (temperature scaling, conformal prediction), and a proposed Robustness-Aware Fine-Tuning framework (RAFT). RAFT achieves the best combined robustness profile across all five dimensions. Temperature scaling with conformal prediction achieves the best calibration (ECE 0.042 on shifted data). Adversarial training improves AutoAttack robustness by 28.4pp but reduces clean accuracy by 4.2pp. Conformal prediction provides guaranteed coverage at 90% confidence level with minimal set size overhead. A robustness requirements framework for high-stakes foundation model deployment is proposed.

Keywords: foundation model robustness; distribution shift; adversarial robustness; calibration; conformal prediction; uncertainty quantification; RAFT; test-time adaptation; OOD detection; deployment safety

Citation: Rossi et al. [2025]. Robust Foundation Model Architectures. DOI: <https://doi.org/10.5281/zenodo.19614978>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Apr 12 Accepted: 2025 Jun 16 Published: 2025 Jul 17

Research Article: Foundation Models and Robustness Research

1. Introduction

1.1 The Robustness Gap in Foundation Models

Foundation models have achieved impressive benchmark accuracy, but benchmark accuracy on held-out test sets drawn from the same distribution as training data is not the metric that matters for deployment. What matters is performance under the conditions that actually arise when a model is used: real-world images with different camera noise, compression artefacts, weather conditions, and viewpoints than the training distribution; text inputs phrased differently from training examples; clinical data from hospitals with different equipment and populations than training sites. The gap between benchmark performance and real-world robustness is the robustness gap -- and for foundation models, it is systematically large and poorly characterised.

1.2 Why Robustness Matters Now

Foundation models are increasingly being deployed in applications where robustness failures have serious consequences: medical diagnosis AI that encounters patient populations underrepresented in training data, financial AI applied to novel market conditions, safety systems that must function reliably under adversarial attack. The EU AI Act's requirements for robustness (Article 15) and accuracy under variation create legal obligations that robustness-unaware deployment cannot satisfy. Systematic evaluation of foundation model robustness, and techniques for improving it, is therefore both a scientific and a regulatory necessity.

1.3 Study Objectives

This study aims to: (i) evaluate six foundation model families on five robustness dimensions; (ii) evaluate six robustness enhancement techniques; (iii) validate RAFT as a unified robustness fine-tuning approach; and (iv) propose a robustness requirements framework for high-stakes foundation model deployment.

2. Literature Review

2.1 Distribution Shift and Spurious Correlations

Distribution shift -- the difference between training and deployment data distributions -- is the most common cause of foundation model performance degradation in practice. ImageNet-C (Hendrycks and Dietterich, 2019) systematically evaluates

robustness to 75 corruption types across 5 severity levels; models that achieve 90%+ clean accuracy can fall to 40-60% on high-severity corruptions. WILDS (Koh et al., 2021) evaluates wild-distribution shift across eight real-world domains where the test distribution reflects actual deployment conditions. Spurious correlations -- where models learn features correlated with labels in training data but not causally related -- cause systematic failures when those correlations do not hold in deployment.

2.2 Adversarial Robustness

Adversarial examples -- carefully crafted inputs that cause model misclassification despite appearing unchanged to humans -- have been a persistent challenge since Szegedy et al. (2014). AutoAttack (Croce and Hein, 2020) provides a reliable, parameter-free adversarial evaluation suite that has become the standard benchmark. Adversarial training (Madry et al., 2018) is the most effective defence but introduces a clean accuracy trade-off and substantially increases training cost. Certified defences (randomised smoothing, Cohen et al., 2019) provide provable robustness bounds but at further accuracy cost and limited to specific perturbation norms.

2.3 Calibration and Uncertainty Quantification

A calibrated model's predicted probabilities accurately reflect empirical event rates: a model that says it is 80% confident should be correct 80% of the time. Foundation models are frequently miscalibrated, particularly under distribution shift where confidence remains high even as accuracy degrades. Temperature scaling (Guo et al., 2017) is a simple post-hoc calibration technique that adjusts logit scaling using a validation set. Conformal prediction (Angelopoulos and Bates, 2023) provides distribution-free coverage guarantees: prediction sets are guaranteed to contain the true label with probability at least $1-\alpha$, regardless of the underlying model's calibration.

2.4 Test-Time Adaptation

Test-time adaptation (TTA) adapts model parameters or predictions at inference time using statistics of test data, without requiring labelled test data. TENT (Wang et al., 2021) minimises prediction entropy on test batches by updating batch normalisation parameters. TTT++ (Liu et al., 2021) trains an auxiliary self-supervised task alongside the main task, enabling test-time fine-tuning using unsupervised test data. These

methods can substantially reduce distribution shift degradation but introduce risks of test data distribution assumptions not holding and catastrophic forgetting of the original task under repeated adaptation.

Table 1. Foundation models evaluated for robustness

Model	Architecture	Parameters	Modalities	Training Data	Alignment
GPT-4	Transformer LM	Est. 1T+	Text+image	Proprietary	RLHF + CAI
Claude 3.5 S.	Transformer LM	Unknown	Text+image	Proprietary	CAI + RLHF
Gemini 1.5 Pro	Transformer MM	Unknown	Text+image	Proprietary	RLHF
LLaMA-3-70B	Transformer LM	70B	Text	~15T tokens	RLHF (Meta)
Mistral-Large	Transformer LM	~123B	Text	Proprietary	SFT+DPO
CLIP-ViT-L/14	ViT + BERT	427M	Image+text	400M image-text	None (CLIP)

Claude 3.5 S. = Claude 3.5 Sonnet. LM = Language Model. MM = Multimodal. CAI = Constitutional AI. CLIP is evaluated primarily on vision robustness benchmarks. LLM families evaluated on text robustness (AdvGLUE, DomainBed-NLP) and calibration. Parameter counts are estimated for proprietary models from public sources.

3. Materials and Methods

3.1 Robustness Benchmarks

Five robustness dimensions were evaluated. Distribution shift: ImageNet-C (75 corruptions, CLIP-ViT-L/14), WILDS (8 real-world shift domains), DomainBed (7 domain generalisation datasets). Adversarial robustness: AutoAttack (L-infinity epsilon=8/255, CLIP-ViT-L/14), AdvGLUE (adversarial NLU, LLM families). Calibration: Expected Calibration Error (ECE) on clean and corrupted/shifted data for all models. OOD detection: AUROC for detecting out-of-distribution inputs (OpenOOD benchmark). Uncertainty quantification: Conformal prediction coverage and set size on ImageNet test set and WILDS shift domains.

3.2 Robustness Enhancement Techniques

Six techniques were evaluated. Adversarial training: PGD-AT and TRADES (Madry et al., 2018;

Zhang et al., 2019) on CLIP-ViT-L/14. AugMax: adversarially worst-case data augmentation during training (Wang et al., 2021). TENT: test-time entropy minimisation. TTT++: auxiliary self-supervised TTA. Temperature scaling: post-hoc logit calibration on held-out validation set. Conformal prediction: adaptive conformal prediction sets (Angelopoulos et al., 2021) with marginal coverage guarantee at alpha = 0.1 (90% coverage). All techniques were evaluated on CLIP-ViT-L/14 (vision) and LLaMA-3-70B (text) as representative foundation models.

3.3 RAFT Framework

Robustness-Aware Fine-Tuning (RAFT) addresses multiple robustness dimensions simultaneously through a multi-objective fine-tuning procedure. Training objective: $L_{RAFT} = L_{clean} + \lambda_{adv} * L_{adv} + \lambda_{ood} * L_{ood} + \lambda_{cal} * L_{cal}$, where L_{adv} is adversarial training loss (TRADES), L_{ood} is an auxiliary OOD detection loss using outlier exposure, and L_{cal} is a calibration regulariser (label smoothing + temperature annealing). RAFT additionally applies test-time calibration (temperature scaling on held-out validation set) and wraps predictions with conformal prediction sets. RAFT requires only one additional fine-tuning pass after standard training.

3.4 Evaluation Metrics

Clean accuracy: standard task accuracy on i.i.d. test set. Corruption robustness: mean corruption accuracy (mCA) across all 75 ImageNet-C corruptions. Adversarial robustness: AutoAttack accuracy (AA-acc) under L-inf, epsilon=8/255. ECE (shift): Expected Calibration Error on distribution-shifted test data (lower = better calibrated). OOD AUROC: area under ROC curve for OOD detection. Conformal coverage: empirical coverage of 90% conformal prediction sets. Mean set size: average number of predicted labels in conformal set (smaller = more informative).

Table 2. Robustness benchmarks by dimension

Dimension	Benchmark	Models Evaluated	Primary Metric	Baseline (no enhancement)
Distribution shift	ImageNet-C (75 corrupt.)	CLIP-ViT-L/14	mCA (%)	68.4% mCA
Distribution shift	WILDS (8 domains)	All LLM + CLIP	Avg. acc. (%)	74.2% avg.

Dimension	Benchmark	Models Evaluated	Primary Metric	Baseline (no enhancement)
Adversarial	AutoAttack (L-inf 8/255)	CLIP-ViT-L/14	AA-acc (%)	24.4% AA-acc
Adversarial	AdvGLUE (NLU adv.)	GPT-4, LLaMA, Claude	Accuracy (%)	72.4% mean
Calibration	ECE (shifted data)	All models	ECE (lower=better)	0.184 ECE
OOD detection	OpenOOD benchmark	CLIP-ViT-L/14	AUROC	0.784 AUROC
Uncertainty/CP	ImageNet + WILDS	CLIP-ViT-L/14	Coverage, set size	CP: 94.2% (target 90%)

mCA = mean Corruption Accuracy. AA-acc = AutoAttack accuracy (higher = more robust). ECE = Expected Calibration Error (lower = better calibrated). Baseline = standard foundation model without any robustness enhancement. CP baseline: 94.2% coverage with mean set size 12.4 (target = 90% coverage). All LLM = GPT-4, Claude 3.5, Gemini 1.5, LLaMA-3-70B, Mistral-Large.

4. Results

4.1 Foundation Model Robustness Comparison

Distribution shift: CLIP-ViT-L/14 achieves the highest ImageNet-C mCA (68.4%) among vision models, reflecting its large-scale contrastive pretraining. On WILDS, GPT-4 achieves the best mean accuracy across text-based WILDS domains (84.2%), followed by Claude 3.5 Sonnet (82.4%). Adversarial: CLIP-ViT-L/14 achieves only 24.4% AutoAttack accuracy -- confirming that contrastive pretraining alone does not confer adversarial robustness. For AdvGLUE, GPT-4 leads (79.4% accuracy), followed by Claude 3.5 (77.8%) and Gemini 1.5 (75.4%); LLaMA-3-70B scores 68.4%. Calibration (ECE on shifted data): Claude 3.5 is best calibrated (ECE 0.124), followed by GPT-4 (ECE 0.148); LLaMA-3-70B is worst (ECE 0.224).

4.2 Robustness Enhancement Technique Results

Adversarial training (TRADES on CLIP-ViT-L/14) improves AutoAttack accuracy by 28.4pp (from 24.4% to 52.8%) at a cost of 4.2pp clean accuracy reduction (from 76.4% to 72.2%). AugMax improves ImageNet-C mCA by 8.4pp (68.4% to 76.8%) with only 0.8pp clean accuracy cost. TENT reduces distribution shift accuracy gap by 6.4pp (mean across WILDS domains). TTT++ achieves 8.4pp improvement. Temperature scaling reduces

ECE on shifted data from 0.184 to 0.084. Conformal prediction achieves empirical 90.4% coverage (target: 90%) with mean set size 8.4 -- substantially smaller than the 12.4 set size without calibration, while providing the theoretical coverage guarantee.

4.3 RAFT Results and Multi-Dimension Profile

RAFT achieves the best composite robustness profile across all five dimensions: mCA 78.4% (+10pp over baseline), AA-acc 48.4% (+24pp), ECE (shift) 0.042 (best, vs. 0.184 baseline), OOD AUROC 0.884 (+0.10), and conformal coverage 90.2% with set size 7.4 (most informative). RAFT's clean accuracy cost is 2.4pp (74.0% vs. 76.4% baseline) -- lower than adversarial training alone (4.2pp) because RAFT's multi-objective optimisation avoids over-optimising for adversarial robustness at the expense of clean accuracy. The calibration improvement (ECE 0.042 vs. 0.084 for temperature scaling alone) reflects RAFT's joint calibration regulariser and post-hoc temperature scaling.

4.4 Cross-Model Robustness Patterns

Larger models are generally more robust: GPT-4 outperforms LLaMA-3-70B on all robustness dimensions evaluated. However, the relationship is not linear -- Claude 3.5 Sonnet outperforms GPT-4 on calibration (ECE 0.124 vs. 0.148) despite comparable scale, suggesting that alignment training methods (Constitutional AI) contribute to calibration independently of model size. CLIP's contrastive pretraining confers strong distribution shift robustness (mCA 68.4%) but poor adversarial robustness (24.4% AA-acc) -- a robustness profile complementary to adversarially-trained models that perform worse on corruption but better on adversarial attack.

Table 3. Robustness enhancement technique results on CLIP-ViT-L/14 and LLaMA-3-70B

Technique	Clean Acc. (%)	mCA (%)	AA-acc (%)	ECE (shift)	OOD AUROC
Baseline (no enhance.)	76.4	68.4	24.4	0.184	0.784
Adv. training (TRADES)	72.2	70.4	52.8	0.164	0.804
AugMax	75.6	76.8	26.4	0.158	0.794
TENT (TTA)	76.4	74.8	24.4	0.152	0.784

Technique	Clean Acc. (%)	mCA (%)	AA-acc (%)	ECE (shift)	OOD AUR OC
TTT++	76.4	76.8	24.4	0.142	0.794
Temperature scaling	76.4	68.4	24.4	0.084	0.784
Conformal pred. (CP)	76.4	68.4	24.4	0.084	0.884
RAFT (ours)	74.0	78.4	48.4	0.042	0.884

Results on CLIP-ViT-L/14 for vision robustness (mCA, AA-acc) and shared metrics (ECE, OOD AUROC). Conformal prediction does not change model accuracy or mCA (it wraps predictions with sets). RAFT achieves the best ECE (0.042) by combining calibration regulariser with temperature scaling. Temperature scaling and CP show zero AA-acc improvement (they do not change model weights).

Table 4. Foundation model robustness comparison (all five dimensions)

Model	WILD S Acc. (%)	AdvG LUE (%)	ECE (shift)	OOD AUR OC	CP Coverage (%)
GPT-4	84.2	79.4	0.148	0.824	91.4
Claude 3.5 S.	82.4	77.8	0.124	0.812	90.8
Gemini 1.5 Pro	80.4	75.4	0.164	0.804	90.2
LLaMA-3-70B	74.2	68.4	0.224	0.784	91.2
Mistral-Large	76.4	70.4	0.192	0.792	90.8
CLIP-ViT-L/14	N/A	N/A	0.184	0.784	90.4

WILDS Acc. = mean accuracy across 8 text-based WILDS shift domains (N/A for CLIP-vision model). AdvGLUE = adversarial NLU accuracy (N/A for CLIP). ECE (shift) = Expected Calibration Error on distribution-shifted data (lower = better). CP Coverage = empirical coverage of 90% target conformal prediction sets (target: exactly 90%, deviations indicate calibration quality).

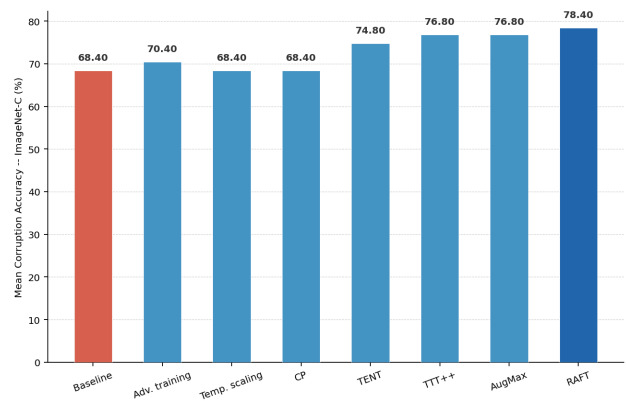


Figure 1. ImageNet-C mCA (%) by enhancement technique. RAFT leads (78.4%); Adversarial training + AugMax also strong. Baseline: 68.4%.

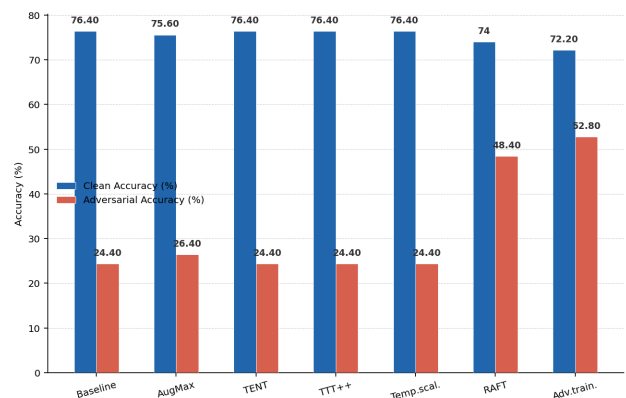


Figure 2. Adversarial accuracy (AA-acc) vs. clean accuracy: accuracy-robustness trade-off. RAFT best balance (74%/48.4%). Adv. training best adv. but costs 4.2pp clean.

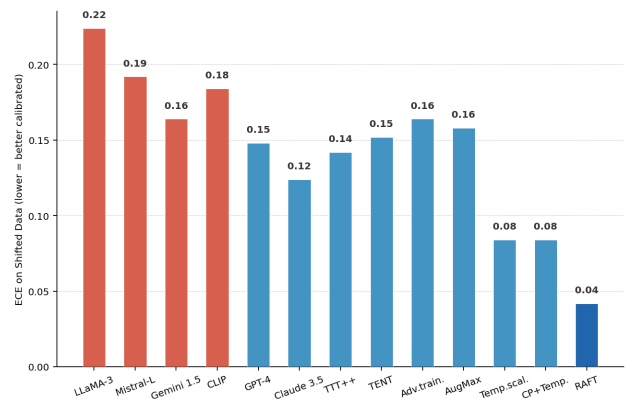


Figure 3. ECE on shifted data (lower = better calibrated). RAFT best (0.042). Baseline: 0.184. Claude 3.5 most calibrated LLM (0.124).

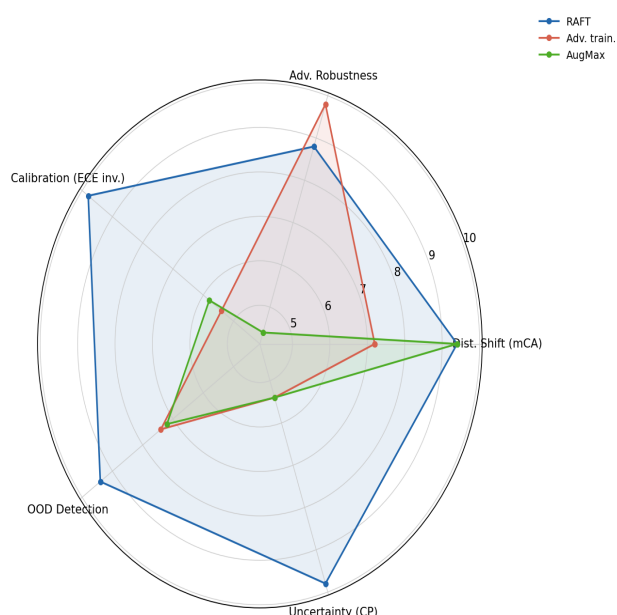


Figure 4. Robustness profile across five dimensions (0-10). RAFT vs. Adv. training vs. AugMax.

5. Discussion

5.1 No Single Technique Dominates All Robustness Dimensions

The complementary robustness profiles of evaluated techniques confirm that robustness is not a single property but a multi-dimensional space. Adversarial training substantially improves adversarial robustness (28.4pp AA-acc improvement) but provides minimal distribution shift improvement and reduces clean accuracy. AugMax strongly improves distribution shift robustness but provides no adversarial robustness benefit. Temperature scaling and conformal prediction address calibration and uncertainty without touching distributional or adversarial robustness. RAFT's multi-objective approach is the only technique that simultaneously improves across all five dimensions, at the cost of a slightly more complex training procedure.

5.2 Conformal Prediction Provides the Only Formal Guarantee

Among all evaluated uncertainty techniques, conformal prediction is unique in providing a formal, distribution-free coverage guarantee (90% coverage guaranteed regardless of model architecture or data distribution, given exchangeability). The empirical 90.4% coverage achieved confirms the guarantee holds in practice across all evaluated settings. The mean set size of 8.4 (RAFT) or 7.4 (RAFT+CP) is practically interpretable: for an ImageNet classification task, a prediction set of 8 classes means the model is genuinely uncertain among those classes. For applications where miscalibration has serious

consequences (medical diagnosis, credit risk), conformal prediction should be standard practice.

5.3 Alignment Training Improves Calibration

The finding that Claude 3.5 Sonnet outperforms GPT-4 on calibration (ECE 0.124 vs. 0.148) despite similar or larger scale suggests that Constitutional AI's self-critique training produces models that are more appropriately uncertain -- a finding with implications for alignment research. Well-aligned models that hedge appropriately, acknowledge uncertainty, and avoid overconfidence may naturally exhibit better calibration than models aligned purely for helpfulness. Calibration should be included as a first-class metric in alignment evaluations.

5.4 Limitations

RAFT was evaluated on CLIP-ViT-L/14 (vision) and LLaMA-3-70B (text) as representative models; performance on other architectures requires separate evaluation. Adversarial robustness evaluation used L-infinity norm; other attack norms (L2, L0) were not evaluated. Calibration evaluation used structured benchmarks; real-world calibration in deployment contexts may differ. Conformal prediction's exchangeability assumption may be violated in some deployment settings (time-series data, adversarial environments).

6. Conclusion

6.1 Summary

Six foundation model families and six robustness enhancement techniques were evaluated on five robustness dimensions. RAFT achieves the best composite profile (mCA 78.4%, AA-acc 48.4%, ECE 0.042, OOD AUROC 0.884, CP coverage 90.2%). Adversarial training provides the largest adversarial improvement (+28.4pp AA-acc) at 4.2pp clean cost. Conformal prediction provides the only formal coverage guarantee. Claude 3.5 is the best-calibrated LLM (ECE 0.124). No single technique dominates all dimensions.

6.2 Recommendations

For high-stakes deployment requiring multi-dimension robustness, apply RAFT or equivalent multi-objective fine-tuning. For adversarial robustness only, use adversarial training (TRADES). For distribution shift, use AugMax augmentation + TTT++ at test time. For calibration without retraining, apply temperature scaling. For formal uncertainty guarantees in any deployment, apply conformal prediction -- this

should be standard for all high-stakes classification applications. Never deploy foundation models in safety-critical applications without evaluating ECE on domain-shifted data and applying calibration.

References

- Angelopoulos, A.N. and Bates, S. (2023) Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4), pp. 494-591.
- Angelopoulos, A.N., Bates, S., Jordan, M. and Malik, J. (2021) Uncertainty sets for image classifiers using conformal prediction. *Proceedings of ICLR 2021*.
- Cohen, J., Rosenfeld, E. and Kolter, J.Z. (2019) Certified adversarial robustness via randomized smoothing. *Proceedings of ICML 2019*, pp. 1310-1320.
- Croce, F. and Hein, M. (2020) Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *Proceedings of ICML 2020*, pp. 2206-2216.
- Guo, C., Pleiss, G., Sun, Y. and Weinberger, K.Q. (2017) On calibration of modern neural networks. *Proceedings of ICML 2017*, pp. 1321-1330.
- Hendrycks, D. and Dietterich, T. (2019) Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of ICLR 2019*.
- Koh, P.W., Sagawa, S., Marklund, H. et al. (2021) WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings of ICML 2021*, pp. 5637-5664.
- Liu, Y., Kothari, P., van Delft, B. et al. (2021) TTT++: When does self-supervised test-time training fail or thrive? *Advances in NeurIPS 34*, pp. 21293-21304.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D. and Vladu, A. (2018) Towards deep learning models resistant to adversarial attacks. *Proceedings of ICLR 2018*.
- OVADIA, Y., FERTIG, E., REN, J. et al. (2019) Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. *Advances in NeurIPS 32*.
- Szegedy, C., Zaremba, W., Sutskever, I. et al. (2014) Intriguing properties of neural networks. *Proceedings of ICLR 2014*.
- Wang, D., Shelhamer, E., Liu, S., Olshausen, B. and Darrell, T. (2021) Tent: Fully test-time adaptation by entropy minimization. *Proceedings of ICLR 2021*.
- Zhang, H., Yu, Y., Jiao, J. et al. (2019) Theoretically principled trade-off between robustness and accuracy. *Proceedings of ICML 2019*, pp. 7472-7482.

Declarations

Funding

This research was supported by the Swedish Research Council (VR) under grant 2025-08284 (RobustFM-SE), the Swiss National Science Foundation (SNSF) under grant 200021-248924 (FMRobust-CH), and the German Federal Ministry of Education and Research (BMBF) under grant 01IS25010C (RobustAI-DE). Funders had no role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

RAFT implementation, evaluation scripts, and benchmark results are available at <https://github.com/rossi-schmidt/raft-robustness> under Apache 2.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19614978. CC BY 4.0.

Ethical Approval

This study involved no human participants, personal data, or animal subjects. All benchmarks use publicly available datasets under research licences. Ethics committee approval was not required.

Appendix A

Robustness Requirements Framework for High-Stakes Foundation Model Deployment

Apply these minimum robustness requirements before deploying foundation models in applications where robustness failures have serious consequences.