Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Doctoral Consortium Track

Toward Trustworthy Foundation Models: Mitigating OOD Failure in Adaptation and Deployment

Ping Song
Certer for Research Training in Artificial Intelligence (CRT in AI), Data Science Institute,
University of Galway, Ireland
P.Song1@universityofgalway.ie

Abstract

Foundation models' versatile generalization has driven their broad adoption, yet their deployment is constrained by critical gaps in trustworthiness. First, we conduct a survey to present a comprehensive analysis of what constitute trustworthy foundation model and examined approaches for each requirement from a lifecycle perspective. We identify out-of-distribution (OOD) robustness as the primary technical bottleneck, where performance collapse under distribution shifts invalidates other pillars like fairness and accountability. The nature of this vulnerability is two-fold: first, a reliance on spurious background cues during inference, and second, dimensional collapse in the representation space during fine-tuning. This research aims to mitigate these failures by developing mechanistic insights and algorithmic solutions during foundation model adaptation and deployment phases of the lifecycle.

1 Introduction and Motivation

As foundation models increasingly drive high-stakes decisions in areas like healthcare and finance, ensuring their trustworthiness and reliability has become essential for successful real-world deployment. However, the broad nature of trustworthiness makes it difficult to address as a singular technical challenge. To navigate this complexity, my first research phase involved a comprehensive survey of trustworthy foundation models. We identified key trustworthiness requirements, discussed and examined trustworthiness criteria of foundation model throughout their entire lifecycle.

Crucially, our landscape analysis identified OOD robustness as the primary technical bottleneck for trustworthy deployment. While foundation models generalize well within known domains, their performance often collapses under distribution shifts, a vulnerability that undermines other pillars, as a model cannot be considered 'fair' or 'accountable' if its reliability is tied to specific training distributions. Consequently, this research investigates the mechanistic causes of OOD failure during the adaptation and deployment stages. We hypothesize that the observed trade-off between in-distribution performance and OOD robustness during fine-tuning stems from dimensional collapse, where models favor low-dimensional subspaces of spurious features [Garrido *et al.*, 2023; Ghanooni *et al.*, 2026]. Furthermore, we explore the representation geometry of Low-Rank Adaptation (LoRA), hypothesizing that its rank-constrained updates act as a geometric anchor to the pre-trained manifold, providing a mechanistic explanation for the stability-plasticity trade-off, where models 'forget less' pre-trained knowledge but 'learn less' from target tasks [Biderman *et al.*, 2024].

The research questions are: **RQ1**: What are the fundamental requirements for trustworthy foundation models, and how do these core pillars impact the broader landscape of requirements throughout the lifecycle? **RQ2**: How can background-aware test-time interventions improve the OOD robustness of black-box foundation models? **RQ3**: How to prevent dimensional collapse to enhance OOD robustness for VLMs fine-tuning? **RQ4**: How does the low-rank constraint of LoRA anchor model updates & impact the pre-trained representation manifold, and how does this geometric restriction influence OOD performance?

2 Contributions

2.1 A Lifecycle Perspective on Trustworthy Foundation Models (RQ1)

To address the conceptual ambiguity of "trustworthiness" in foundation model, we conducted a survey research and developed a holistic framework specifically for foundation models. We formalized a three-phase lifecycle, including Data Preparation, Model Construction, and System Operation, to structure the analysis of trust-related risks and mitigations. Through this lifecycle lens, we identified and defined eight critical dimensions of trustworthiness: Robustness, Generalization, Understandability, Transparency, Fairness, Privacy, Stewardship, and Security. Our work examined technical approaches for each dimension, such as data augmentation for robustness and reinforcement learning with human feedback (RLHF) for fairness, across all lifecycle stages [Song *et al.*, 2026].

Unlike existing literature that often focuses on a single aspect (e.g., only fairness or robustness) or examines trust from a static perspective (e.g., only training phase), this research provides two primary innovations: 1. Integrated Lifecycle

Perspective: We shifted from a point-in-time evaluation to a lifecycle-oriented analysis by formalizing the lifecycle for foundation models. We examine approaches for each aspects across the lifecycle, and identify challenges, gaps, and future direction. 2. Cross-Pillar Dependencies: By targeting the unique complexities of foundation models, such as emergence and homogenization, our work identifies how vulnerabilities in one pillar (like OOD robustness) can fundamentally undermine others (like fairness or accountability).

This foundational phase provided the empirical justification for our subsequent technical deep-dives into OOD robustness, identifying it as the primary technical bottleneck for safe real-world deployment.

2.2 Background-Aware Robustness in NLP (RQ2)

In real-world deployments, foundation models often fail on OOD data due to a reliance on spurious cues that are correlated with labels in the training set but fail in new domains. As many models are deployed as frozen black-boxes, test-time augmentation (TTA) via In-Context Rewriting (ICR) [O'Brien *et al.*, 2024] has emerged as a practical alternative to retraining. These methods prompt an LLM to rewrite OOD inputs to statistically resemble a set of in-distribution exemplars. While effective at masking distribution shifts, these methods often employ unconstrained rewriting, risking unintended semantic shifts.

We propose Background-Aware TTA (BA-TTA), which decomposes domain shift into semantic and background (style/syntax) axes. We identify that OOD performance degradation is frequently driven by models' reliance on spurious background cues. By utilizing a semantic-constrained alignment strategy, we prompt LLMs to adjust only specific background characteristics (e.g., tone, register) while strictly preserving the original meaning. By narrowing the transformation scope, this targeted approach allows smaller, 7B-parameter models to generate high-fidelity transformations typically requiring much larger architectures, making BA-TTA viable for resource-constrained or latency-sensitive edge deployments. Our experiments demonstrate that explicitly correcting background shift achieves superior robustness over generic LLM-based rewriting. (Under review at TMLR [1]).

3 Future Work

RQ3: Building on RQ2, RQ3 investigates mitigating reliance on spurious cues during the fine-tuning process. We hypothesize that the ID-OOD trade-off stems from dimensional collapse, a phenomenon where models favor low-dimensional subspaces of spurious features over diverse, robust ones. While studies like RankMe [Garrido *et al.*, 2023] and Spectral Regularization [Ghanooni *et al.*, 2026] address this in uni-modal pre-training, our work addresses a critical gap: the behavior of multi-modal (VLM) representations during fine-tuning. We investigate whether preserving the representation's effective rank can decouple ID performance from OOD vulnerability. Our evaluation follows a two-stage approach: 1. Diagnostic Phase: We quantify dimensional collapse by measuring the singular value distribution of VLM feature matrices before and after fine-tuning on ImageNet-1K, correlating these metrics with performance on OOD benchmarks like ImageNet-R and ObjectNet. 2. Mitigation Phase: We are developing spectral constraints for multi-modal fine-tuning. By introducing a regularizer that promotes a uniform eigenspectrum, we aim to prevent "gradient dominance" by spurious features and ensure the model retains its original foundational representational capacity.

RQ4: To answer RQ4, we investigate the underlying learning dynamics of Low-Rank Adaptation (LoRA) through the lens of representation geometry. While empirical studies suggest that LoRA "forgets less" of pre-trained knowledge but "learns less" of target tasks [Biderman *et al.*, 2024] compared to full fine-tuning, the mechanistic cause remains poorly understood. We hypothesize that this stability-plasticity trade-off is a direct result of LoRA's rank-constrained updates, which act as a geometric anchor to the pre-trained representation manifold. By restricting the degrees of freedom during adaptation, LoRA prevents the model's representation trajectories from diverging into high-dimensional "spurious subspaces" that often lead to OOD failure.

[1] https://openreview.net/forum?id=xptPQVCy5X

Acknowledgments

This publication was conducted with the financial support of the Research Ireland Centre for Research Training in Artificial Intelligence under Grant No.18/CRT/6223, and also by a grant from Insight Research Ireland Centre for Data Analytics under Grant Number SFI/12/RC/2289_P2.

References

[Biderman *et al.*, 2024] Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, et al. Lora learns less and forgets less. *arXiv preprint arXiv:2405.09673*, 2024.

[Garrido *et al.*, 2023] Quentin Garrido, Randall Balestriero, Laurent Najman, and Yann Lecun. Rankme: Assessing the downstream performance of pretrained self-supervised representations by their rank. In *International conference on machine learning*, pages 10929–10974. PMLR, 2023.

[Ghanooni *et al.*, 2026] Naghmeh Ghanooni, Waleed Mustafa, Dennis Wagner, Sophie Fellenz, Anthony Lin, and Marius Kloft. Mitigating spurious features in contrastive learning with spectral regularization. *Advances in Neural Information Processing Systems*, 38:107856–107888, 2026.

[O'Brien *et al.*, 2024] Kyle O'Brien, Nathan Ng, Isha Puri, Jorge Mendez, Hamid Palangi, Yoon Kim, Marzyeh Ghassemi, and Thomas Hartvigsen. Improving black-box robustness with in-context rewriting. *arXiv preprint arXiv:2402.08225*, 2024.

[Song *et al.*, 2026] Ping Song, Adegboyega Ojo, and Edward Curry. Trustworthy requirements for foundation models—a comprehensive survey and roadmap. *Engineering Applications of Artificial Intelligence*, 163:113111, 2026.