Novelty Assessment Report

Paper: Adversarially Pretrained Transformers may be Universally Robust In-Context Learners
PDF URL: https://openreview.net/pdf?id=11eHIPnWDx
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-05

Abstract

Adversarial training is one of the most effective adversarial defenses, but it incurs a high computational cost. In this study, we present the first theoretical analysis suggesting that adversarially pretrained transformers can serve as universally robust foundation models, models that can robustly adapt to diverse downstream tasks with only lightweight tuning. Specifically, we demonstrate that single-layer linear transformers, after adversarial pretraining across a variety of classification tasks, can robustly generalize to unseen classification tasks through in-context learning from clean demonstrations (i.e., without requiring additional adversarial training or examples). This universal robustness stems from the model's ability to adaptively focus on robust features within given tasks. We also show the two open challenges for attaining robustness: accuracy-robustness trade-off and sample-hungry training. This study initiates the discussion on the utility of universally robust foundation models. While their training is expensive, the investment would prove worthwhile as downstream tasks can enjoy free adversarial robustness.

Core Task Landscape

This paper addresses: **Adversarial Robustness of Pretrained Transformers through In-Context Learning**

A total of **40 papers** were analyzed and organized into a taxonomy with **14 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Theoretical Foundations of In-Context Learning**
- **Adversarial Attacks on In-Context Learning**
- **Defense Mechanisms and Robustness Enhancement**
- **Empirical Robustness Evaluation**
- **Domain-Specific Applications**
- **Reliability and Trustworthiness Frameworks**

Complete Taxonomy Tree

- Adversarial Robustness of Pretrained Transformers through In-Context Learning Survey Taxonomy
- Theoretical Foundations of In-Context Learning
  - Learning Dynamics and Generalization Theory (3 papers)
  - [1] Trained transformers learn linear models in-context (Zhang Ruiqi, 2024) View paper
  - [18] Provable test-time adaptivity and distributional robustness of in-context learning (Ma Tian-Yi, 2025) View paper
  - [35] On the Role of Depth and Looping for In-Context Learning with Task Diversity (Gatmiry, 2024) View paper
  - Adversarial Robustness Theory ★ (5 papers)
  - [0] Adversarially Pretrained Transformers may be Universally Robust In-Context Learners (Anon et al., 2026) View paper
  - [3] Adversarial robustness of in-context learning in transformers for linear regression (U Anwar, 2024) View paper
  - [6] On the robustness of transformers against context hijacking for linear classification (Li Tianle, 2025) View paper
  - [9] Understanding In-Context Learning of Linear Models in Transformers Through an Adversarial Lens (Anwar, 2024) View paper
  - [23] Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers under In-Context Regression (Weiyi He, 2025) View paper
- Adversarial Attacks on In-Context Learning
  - Data Poisoning and Demonstration Attacks (5 papers)
  - [4] Data poisoning for in-context learning (He Peng-fei, 2025) View paper
  - [15] Demonstration attack against in-context learning for code intelligence (Ge, 2024) View paper
  - [16] ICLShield: Exploring and Mitigating In-Context Learning Backdoor Attacks (Ren Zhi-yao, 2025) View paper
  - [31] Can In-Context Reinforcement Learning Recover From Reward Poisoning Attacks? (Paulius Sasnauskas, 2025) View paper
  - [34] Context is the Key: Backdoor Attacks for In-Context Learning with Vision Transformers (Abad, 2024) View paper
  - Prompt-Based Hijacking Attacks (3 papers)
  - [8] Hijacking large language models via adversarial in-context learning (Yao, 2023) View paper
  - [11] Adversarial Attacks against Neural Ranking Models via In-Context Learning (Arabzadeh, 2025) View paper
  - [38] BAM-ICL: Causal Hijacking In-Context Learning with Budgeted Adversarial Manipulation (R Chu, n.d.) View paper
  - Jailbreak and Adversarial Prompt Attacks (1 papers)
  - [14] "Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence (Fu ShaoPeng, 2025) View paper
- Defense Mechanisms and Robustness Enhancement
  - Adversarial Training and Pretraining Strategies (4 papers)
  - [17] In-Context Interaction Learning: Boosting Adversarial Robustness for Model Security (Jiahao Zhao, 2025) View paper

- [25] Adversarially Robust Decision Transformer (Ilija Bogunovic, 2024) View paper
- [26] Efficiently Robust In-Context Reinforcement Learning with Adversarial Generalization and Adaptation (J Dong, 2025) View paper
- [40] UNDERSTANDING AND IMPROVING CONTINUOUS AD (TRAINING, n.d.) View paper
- Test-Time Adaptation and Prompt Engineering (3 papers)
- [10] Improving black-box robustness with in-context rewriting (O'Brien, 2024) View paper
- [20] Context-aware Prompt Tuning: Advancing In-Context Learning with Adversarial Methods (Blau, 2024) View paper
- [27] Prompt Optimization via Adversarial In-Context Learning (Brown, 2024) View paper
- Backdoor Detection and Mitigation (1 papers)
- [36] Robustness through Forgetting: Unlearning Adversarial Contexts in Transformers (P Muna-McQuay, n.d.) View paper
- Architectural and Representational Enhancements (2 papers)
- [28] Differential Transformer (YE TianZhu, 2024) View paper
- [30] Contextual Vision Transformers for Robust Representation Learning (Bao, 2023) View paper

- Empirical Robustness Evaluation
  - Retrieval-Augmented Systems Robustness (4 papers)
  - [2] Enhancing robustness of retrieval-augmented language models with in-context learning (Park Seong Il, 2024) View paper
  - [5] Evaluating the adversarial robustness of retrieval-based in-context learning for large kanguage models (SCL Yu, 2024) View paper
  - [12] Evaluating and Safeguarding the Adversarial Robustness of Retrieval-Based In-Context Learning (Yu Simon, 2024) View paper
  - General In-Context Learning Robustness (3 papers)
  - [7] Exploring the robustness of in-context learning with noisy labels (Chen Cheng, 2025) View paper
  - [37] On the Robustness of In-Context Learning with Noisy Labels: Train, Inference, and Beyond (C Cheng, n.d.) View paper
  - [39] ACloser LOOK AT THE ROBUSTNESS OF IN-CONTEXT LEARNING WITH NOISY LABELS (C Cheng, n.d.) View paper
- Domain-Specific Applications
  - Structured Prediction and Code Intelligence (2 papers)
  - [13] Towards robust in-context learning for machine translation with large language models (S Zhu, 2024) View paper
  - [21] Solid-SQL: Enhanced Schema-linking based In-context Learning for Robust Text-to-SQL (Liu, 2024) View paper
  - Reinforcement Learning and Decision-Making (2 papers)
  - [22] Transformers Can Perform Distributionally-robust Optimisation through In-context Learning (T Kim, 2024) View paper
  - [29] Training Adaptive and Sample-Efficient Autonomous Agents (Sridhar, 2025) View paper
- Reliability and Trustworthiness Frameworks (3 papers)
  - [19] Combining large language models and or/ms to make smarter decisions (Segev Wasserkrug, 2024) View paper
  - [24] Securing reliability: A brief overview on enhancing in-context learning for foundation models (Huang Yunpeng, 2024) View paper
  - [33] Fine-tuning Does Not Remove Language Model Capabilities (Kotha, 2024) View paper

Narrative

Core task: adversarial robustness of pretrained transformers through in-context learning. The field has organized itself around six main branches that collectively address how transformers learn from demonstrations and how that learning can be attacked or defended. Theoretical Foundations of In-Context Learning explores the mathematical underpinnings, examining how transformers implement algorithms like linear regression (Transformers Learn Linear Models[1]) and the role of architectural choices such as positional encodings (Positional Encoding Complexity[23]). Adversarial Attacks on In-Context Learning investigates vulnerabilities ranging from context hijacking (Context Hijacking Robustness[6], Hijacking via Adversarial ICL[8]) to data poisoning (Data Poisoning ICL[4]) and retrieval manipulation (Neural Ranking Attacks[11]). Defense Mechanisms and Robustness Enhancement develops protective strategies including robust retrieval methods (Robust Retrieval Augmented Learning[2], Safeguarding Retrieval ICL[12]) and specialized shields (ICLShield[16]). Empirical Robustness Evaluation systematically tests model behavior under adversarial conditions (Retrieval Robustness Evaluation[5]), while Domain-Specific Applications adapt these insights to translation (Robust Translation ICL[13]), reinforcement learning (Robust ICL Reinforcement[26]), and other tasks. Reliability and Trustworthiness Frameworks address broader concerns about securing foundation models (Securing Foundation Models[24]) and ensuring safe deployment.

A particularly active tension exists between understanding in-context learning as implicit optimization versus studying its failure modes under adversarial pressure. Works like Adversarial Robustness Linear Regression[3] and Linear Models Adversarial Lens[9] bridge theory and robustness by analyzing how adversarial perturbations affect the linear models that transformers approximate during in-context learning. Adversarially Pretrained Transformers[0] sits squarely within this theoretical robustness cluster, examining how pretraining strategies can build inherent resilience into the learning process itself. Compared to Adversarial Robustness Linear Regression[3], which focuses on the mathematical properties of robust regression in the ICL setting, the original work emphasizes pretraining as a proactive defense mechanism. Meanwhile, Linear Models Adversarial Lens[9] provides complementary analysis of how adversarial examples interact with the implicit models learned in-context, offering a diagnostic perspective that complements the constructive approach of adversarial pretraining.

Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

1. Adversarial robustness of in-context learning in transformers for linear regression

Authors: U Anwar, J Von Oswald, L Kirsch, D Krueger | **Year/Venue**: 2024 | **URL**: View paper

Abstract

â¦ robustness of these transformers to a hijacking attack, which is structured as follows: The user feeds into the transformer a set of in-context training â¦ , done either at pretraining stage or at â¦

Relationship Analysis

Both papers belong to the Adversarial Robustness Theory category, providing theoretical and empirical analysis of adversarial robustness in transformers performing in-context learning. The original paper focuses on adversarial pretraining of single-layer linear transformers across multiple classification tasks to achieve universal robustness through adaptive feature selection, while the candidate paper investigates hijacking attacks on transformers performing linear regression in-context and demonstrates that adversarial training (pretraining or finetuning) can improve robustness to such attacks. The key difference is that the original paper emphasizes universal robustness across diverse downstream tasks without additional adversarial training, whereas the candidate paper focuses on robustness within a single task domain (linear regression) and analyzes attack transferability between different models and algorithms.

2. On the robustness of transformers against context hijacking for linear classification

Authors: Li Tianle, Zhang Chen-yang, Chen Xing-wu, Cao Yuan, Zou, et al. (6 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Transformer-based Large Language Models (LLMs) have demonstrated powerful in-context learning capabilities. However, their predictions can be disrupted by factually correct context, a phenomenon known as context hijacking, revealing a significant robustness issue. To understand this phenomenon theoretically, we explore an in-context linear classification problem based on recent advances in linear transformers. In our setup, context tokens are designed as factually correct query-answer pairs, whe...

Relationship Analysis

Both papers belong to the Adversarial Robustness Theory category, providing theoretical analysis of adversarial robustness in transformers through in-context learning. The original paper focuses on adversarially pretrained transformers achieving universal robustness across diverse downstream tasks through clean demonstrations, analyzing robust vs. non-robust feature extraction. The candidate paper specifically examines robustness against context hijacking attacks using factually correct but misleading context examples, analyzing how model depth affects robustness through the lens of multi-step gradient descent optimization.

---

3. Understanding In-Context Learning of Linear Models in Transformers Through an Adversarial Lens

Authors: Anwar, Usman, von Oswald, Johannes, Usman Anwar, et al. (15 authors total) | **Year/Venue**: 2024 • Trans. Mach. Learn. Res. | **URL**: View paper

Abstract

In this work, we make two contributions towards understanding of in-context learning of linear models by transformers. First, we investigate the adversarial robustness of in-context learning in transformers to hijacking attacks -- a type of adversarial attacks in which the adversary's goal is to manipulate the prompt to force the transformer to generate a specific output. We show that both linear transformers and transformers with GPT-2 architectures are vulnerable to such hijacking attacks. How...

Relationship Analysis

Both papers belong to the Adversarial Robustness Theory category, providing theoretical analysis of adversarial robustness in in-context learning transformers. They overlap in examining adversarial vulnerabilities of transformers performing in-context learning on linear models and investigating adversarial training as a defense mechanism. However, the original paper focuses on theoretical guarantees for universal robustness across unseen tasks through adversarial pretraining, while the candidate paper emphasizes empirical analysis of hijacking attacks, attack transferability across models, and comparative vulnerabilities between transformers and classical learning algorithms.

---

4. Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers under In-Context Regression

Authors: Weiyi He, Yue Xing | **Year/Venue**: 2025 | **URL**: View paper

Abstract

Positional encoding (PE) is a core architectural component of Transformers, yet its impact on the Transformer's generalization and robustness remains unclear. In this work, we provide the first generalization analysis for a single-layer Transformer under in-context regression that explicitly accounts for a completely trainable PE module. Our result shows that PE systematically enlarges the generalization gap. Extending to the adversarial setting, we derive the adversarial Rademacher generalizati...

Relationship Analysis

Both papers belong to the Adversarial Robustness Theory category, providing theoretical analysis of adversarial robustness in transformers using in-context learning. They overlap in analyzing how transformers can achieve robustness through in-context learning mechanisms, with both employing theoretical frameworks involving robust and non-robust features. However, the original paper focuses on adversarial pretraining enabling universal robustness across diverse downstream tasks without additional adversarial training, while the candidate paper specifically examines the impact of positional encoding on generalization and adversarial robustness through Rademacher complexity analysis in the in-context regression setting.

Contributions Analysis

Overall novelty summary. ```json { "paragraphs": ["The paper proposes that adversarially pretrained transformers can serve as universally robust foundation models, enabling robust adaptation to downstream tasks through in-context learning without additional adversarial training. It resides in the 'Adversarial Robustness Theory' leaf under 'Theoretical Foundations of In-Context Learning', which contains five papers total. This leaf focuses specifically on theoretical analysis of robustness mechanisms and defense properties in in-context learning, representing a moderately populated research direction within a taxonomy of forty papers across the broader field of adversarial robustness in pretrained transformers.",

"The paper's leaf sits alongside 'Learning Dynamics and Generalization Theory', which examines non-adversarial properties of in-context learning such as algorithm implementation and distributional generalization. Neighboring branches include 'Defense Mechanisms and Robustness Enhancement', particularly 'Adversarial Training and Pretraining Strategies', which contains four papers on training-time defenses. The taxonomy's scope note clarifies that this leaf focuses on theoretical analysis rather than empirical evaluation or attack methods, positioning the work at the intersection of foundational theory and proactive defense design through pretraining strategies.",

"Among thirty candidates examined, contribution analysis reveals mixed novelty signals. The core theoretical analysis of universally robust pretrained transformers examined ten candidates with zero refutations, suggesting this framing may be relatively unexplored. However, the condition for robust adaptation based on robust versus non-robust features examined ten candidates and found one refutable match, indicating some overlap with existing frameworks. The identification of accuracy-robustness trade-offs and sample complexity challenges examined ten candidates with no refutations, though these are well-known phenomena in adversarial learning. The limited search scope means substantial relevant work may exist outside the top-thirty semantic matches examined.",

"Based on the examined literature, the universal robustness framing for pretrained transformers appears less explored than the underlying trade-offs and feature frameworks. The analysis covers top-thirty semantic matches plus citation expansion, providing reasonable coverage of closely related theoretical work but not exhaustive field-wide search. The taxonomy structure suggests this theoretical robustness direction, while moderately populated, remains less saturated than empirical attack-defense cycles or domain-specific applications."] } ```

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: Theoretical analysis of universally robust adversarially pretrained transformers

Description: The authors provide the first theoretical support showing that single-layer linear transformers, after adversarial pretraining on multiple classification tasks, can robustly generalize to unseen tasks through in-context learning from clean demonstrations alone, without requiring additional adversarial training or examples.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Adversarially robust hypothesis transfer learning

URL: View paper

Brief Assessment

Robust Hypothesis Transfer[44] focuses on hypothesis transfer learning with adversarial training for general hypothesis classes (RKHS, neural networks), not specifically on transformers' in-context learning capabilities or their universal robustness across multiple unseen tasks without additional training.

2. Synthetic-to-Real Transfer Learning for Chromatin-Sensitive PWS Microscopy

URL: View paper

Brief Assessment

Synthetic-to-Real Chromatin[50] addresses synthetic-to-real transfer learning for microscopy image segmentation, not adversarial pretraining or robust in-context learning in transformers. The domains are entirely distinct.

3. Learning Adversarially Fair and Transferable Representations

URL: View paper

Brief Assessment

Fair Transferable Representations[42] focuses on adversarial training for fair representation learning across demographic groups, not on robust generalization to unseen tasks through in-context learning. The candidate addresses fairness in predictions rather than adversarial robustness to input perturbations.

4. Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data

URL: View paper

Brief Assessment

Wasserstein Adversarial Transfer[47] focuses on deep transfer learning for fault diagnosis using Wasserstein distance to minimize domain distribution differences, not on theoretical analysis of transformer robustness or in-context learning capabilities.

5. Learning Robust Rewards with Adversarial Inverse Reinforcement Learning

URL: View paper

Brief Assessment

Adversarial Inverse Reinforcement[45] focuses on learning reward functions in reinforcement learning settings through adversarial training, not on transformer architectures or in-context learning capabilities across multiple classification tasks.

6. On adversarial training without perturbing all examples

URL: View paper

Brief Assessment

Training Without Perturbing All[43] focuses on adversarial training efficiency by attacking only subsets of training examples during standard supervised learning, not on theoretical analysis of transformer pretraining for robust in-context learning across multiple tasks.

7. Adversarial robustness in transfer learning models

URL: View paper

Brief Assessment

Transfer Learning Robustness[46] focuses on transfer learning pipelines and adversarial pretraining effects on downstream task robustness, but does not address in-context learning or the specific theoretical framework of transformers adapting to unseen tasks through clean demonstrations alone without additional adversarial training.

8. Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach

URL: View paper

Brief Assessment

Fake Content Detection Transfer[49] focuses on domain adaptive transfer learning for fake content detection using adversarial training to learn domain-invariant linguistic features. This is fundamentally different from the original paper's theoretical analysis of adversarially pretrained transformers achieving universal robustness through in-context learning across classification tasks.

9. CARD: Robustness-Preserving Transfer Learning for Network Intrusion Detection via Contrastive Adversarial Representation Distillation

URL: View paper

Brief Assessment

CARD[48] focuses on transfer learning for network intrusion detection systems using contrastive adversarial representation distillation, not on theoretical analysis of in-context learning in transformers across multiple classification tasks.

10. Adversarially robust transfer learning

URL: View paper

Brief Assessment

Robust Transfer Learning[41] focuses on transferring robustness from adversarially trained CNNs to new tasks via feature extraction and fine-tuning, not on theoretical analysis of transformers' in-context learning capabilities or universal robustness across multiple tasks without additional training.

Contribution 2: Condition for robust adaptation based on robust and non-robust features framework

Description: The authors derive theoretical conditions under which adversarially pretrained transformers achieve universal robustness by demonstrating that these models adaptively prioritize robust features over non-robust features in downstream tasks, using the conceptual framework of robust versus non-robust features.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Adversarial Robustness through Disentangled Representations

URL: View paper

Prior Art Analysis

Disentangled Representations Robustness[59] demonstrates prior work on the robust versus non-robust features framework in adversarial training. The candidate paper explicitly adopts the conceptual framework of robust and non-robust features from Ilyas et al. (2019) and Tsipras et al. (2019), proposing a method to disentangle these features for adversarial robustness. The candidate shows that by separating robust (class-specific) from non-robust (class-irrelevant) representations, models can achieve better robustness, which directly addresses the same conceptual framework that the original paper claims to derive theoretical conditions for.

Evidence

Evidence 1 - **Rationale**: Both papers explicitly build upon the same conceptual framework of robust versus non-robust features from Ilyas et al. (2019), demonstrating that this framework was already established in prior work before the original paper's submission. - **Original**: our analysis builds upon the conceptual framework of robust features (class-discriminative and human-interpretable) and non-robust features (human-imperceptible yet predictive) (ilyas et al., 2019; tsipras et al., 2019). based on this framework, we show that adversarially pretrained single-layer lin... - **Candidate**: the seminal work (ilyas et al. 2019) suggests that the existence of adversarial examples is a natural consequence of the non-robust (predictive, yet brittle) representation set, which exists in the natural representation distribution independently along with the robust representation set. the model'...

Evidence 2 - **Rationale**: The candidate paper demonstrates prior work that uses the robust/non-robust features framework to derive conditions for robustness, showing that models can focus on robust features by disentangling them from non-robust ones—the same conceptual approach the original paper claims as novel. - **Original**: based on the framework of robust and non-robust features, we derive the condition for successful robust adaptation. moreover, we show that the universal robustness arises from the model's adaptive focus on robust features within given (unseen) tasks. - **Candidate**: in this paper, we postulate that the robust and non-robust representations are two basic ingredients entangled in the integral representation. the robust representation is specified to the classification task (i.e., class-specific), in contrast, the non-robust representation is not capable of classifica...

---

2. ProFeAT: Projected Feature Adversarial Training for Self-Supervised Learning of Robust Representations

URL: View paper

Brief Assessment

ProFeAT[53] focuses on self-supervised adversarial training for image classification using distillation and projection heads, not on theoretical conditions for transformer adaptation or the robust/non-robust feature framework in the context of in-context learning.

---

3. Minimizing adversarial training samples for robust image classifiers: analysis and adversarial example generator design

URL: View paper

Brief Assessment

Minimizing Training Samples[57] focuses on minimizing adversarial training samples for image classifiers and adjusting non-robust features, not on theoretical conditions for universal robustness in transformers through in-context learning.

---

4. Distilling robust and non-robust features in adversarial examples by information bottleneck

URL: View paper

Brief Assessment

Information Bottleneck Features[54] focuses on distilling robust and non-robust features in adversarial examples using information bottleneck techniques for feature-level analysis, not on deriving theoretical conditions for universal robustness in transformers through in-context learning as the original paper does.

---

5. Adversarial feature alignment: Balancing robustness and accuracy in deep learning via adversarial training

URL: View paper

Brief Assessment

Adversarial Feature Alignment[52] focuses on feature space alignment through contrastive learning to balance accuracy and robustness in adversarial training, rather than analyzing conditions for robust adaptation in transformers using the robust vs. non-robust features framework for in-context learning.

---

6. Exploring robust features for improving adversarial robustness

URL: View paper

Brief Assessment

Robust Features Exploration[51] focuses on disentangling robust and non-robust features in adversarial training for image classification, not on theoretical conditions for universal robustness in transformers through in-context learning.

---

7. Feature purification: How adversarial training performs robust deep learning

URL: View paper

Brief Assessment

Feature Purification[58] focuses on adversarial training of two-layer ReLU networks on sparse coding data, analyzing how dense mixtures accumulate during clean training. The ORIGINAL paper studies adversarially pretrained transformers achieving universal robustness through in-context learning across diverse tasks without additional training—a fundamentally different setting and contribution.

---

8. Learning More Robust Features with Adversarial Training

URL: View paper

Brief Assessment

Robust Features Adversarial Training[56] focuses on improving feature robustness through adversarial training on MNIST/CIFAR-10, not on theoretical conditions for universal robustness in transformers or in-context learning adaptation across diverse tasks.

---

9. Evidence-Based Multi-Feature Fusion for Adversarial Robustness

URL: View paper

Brief Assessment

Evidence-Based Multi-Feature Fusion[55] focuses on quantifying feature trustworthiness and fusing multi-block features in DNNs for adversarial defense, not on deriving theoretical conditions for adversarially pretrained transformers' universal robustness through in-context learning using the robust vs. non-robust features framework.

---

10. Few-Shot Anomaly Detection with Adversarial Loss for Robust Feature Representations

URL: View paper

Brief Assessment

Few-Shot Anomaly Detection[60] focuses on anomaly detection using adversarial training for feature robustness in few-shot scenarios, not on theoretical conditions for universal robustness in transformers or the robust/non-robust features framework for in-context learning.

Contribution 3: Identification of accuracy-robustness trade-off and sample-hungry in-context learning as open problems

Description: The authors formally show that adversarially pretrained single-layer linear transformers exhibit two persistent challenges: lower clean accuracy compared to standard models and the requirement for more in-context demonstrations to achieve comparable performance.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Context - Enhanced Meta-Reinforcement Learning with Data-Reused Adaptation for Urban Autonomous Driving

URL: View paper

Brief Assessment

Context-Enhanced Meta-RL Driving[69] focuses on meta-reinforcement learning for autonomous driving with context-enhanced state representations and data reuse. It does not address adversarial training, in-context learning, or accuracy-robustness trade-offs in transformers, which are the core focus of the original paper's contribution.

---

2. Which examples to annotate for in-context learning? towards effective and efficient selection

URL: View paper

Brief Assessment

Effective Example Selection[67] focuses on active learning for selecting which examples to annotate for in-context learning in a low-resource setting, not on adversarial robustness or the accuracy-robustness trade-off in adversarially pretrained models.

---

3. Are sample-efficient nlp models more robust?

URL: View paper

Brief Assessment

Sample-Efficient NLP Robustness[63] examines sample efficiency and robustness in NLP through fine-tuning approaches, not in-context learning. The original paper studies adversarially pretrained transformers and their in-context learning capabilities, which is a fundamentally different setting.

---

4. A theoretical design of concept sets: improving the predictability of concept bottleneck models

URL: View paper

Brief Assessment

Concept Sets Design[68] focuses on concept-bottleneck models for interpretable machine learning, addressing sample efficiency and robustness through concept set design. This is fundamentally different from the original paper's focus on adversarially pretrained transformers and in-context learning challenges.

---

5. Probing the decision boundaries of in-context learning in large language models

URL: View paper

Brief Assessment

Decision Boundaries Probing[66] focuses on visualizing and analyzing decision boundaries in binary classification tasks to understand in-context learning behavior, not on adversarial robustness or sample efficiency challenges in adversarially pretrained models.

---

6. A Review on Machine Learning Applications in Localization in 5G and Beyond Wireless Communications

URL: View paper

Brief Assessment

Machine Learning Localization 5G[70] focuses on localization techniques in wireless communications, not on adversarial training, in-context learning, or accuracy-robustness trade-offs in transformers. The domains are entirely different.

---

7. Using natural language explanations to improve robustness of in-context learning

URL: View paper

Brief Assessment

Natural Language Explanations Robustness[62] focuses on improving robustness of in-context learning through natural language explanations for NLI and paraphrasing tasks, not on analyzing adversarially pretrained transformers or formally characterizing accuracy-robustness trade-offs in the context of adversarial pretraining.

---

8. Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning

URL: View paper

Brief Assessment

Lifelong In-Context Meta-RL[65] focuses on meta-reinforcement learning with distributional shift and lifelong in-context learning, not on adversarial training or accuracy-robustness trade-offs in transformers.

---

9. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning

URL: View paper

Brief Assessment

Few-Shot Parameter-Efficient[61] focuses on parameter-efficient fine-tuning methods for few-shot learning in NLP tasks, not on adversarial robustness or in-context learning challenges in adversarially pretrained transformers.

---

10. Enhancing in-context learning via linear probe calibration

URL: View paper

Brief Assessment

Linear Probe Calibration[64] focuses on calibrating output probabilities to improve in-context learning reliability and performance, not on adversarial robustness or accuracy-robustness trade-offs in adversarially pretrained models.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Adversarially Pretrained Transformers may be Universally Robust In-Context Learners View paper
- [1] Trained transformers learn linear models in-context View paper
- [2] Enhancing robustness of retrieval-augmented language models with in-context learning View paper
- [3] Adversarial robustness of in-context learning in transformers for linear regression View paper
- [4] Data poisoning for in-context learning View paper
- [5] Evaluating the adversarial robustness of retrieval-based in-context learning for large kanguage models View paper
- [6] On the robustness of transformers against context hijacking for linear classification View paper
- [7] Exploring the robustness of in-context learning with noisy labels View paper
- [8] Hijacking large language models via adversarial in-context learning View paper
- [9] Understanding In-Context Learning of Linear Models in Transformers Through an Adversarial Lens View paper
- [10] Improving black-box robustness with in-context rewriting View paper
- [11] Adversarial Attacks against Neural Ranking Models via In-Context Learning View paper
- [12] Evaluating and Safeguarding the Adversarial Robustness of Retrieval-Based In-Context Learning View paper
- [13] Towards robust in-context learning for machine translation with large language models View paper
- [14] "Short-length" Adversarial Training Helps LLMs Defend "Long-length" Jailbreak Attacks: Theoretical and Empirical Evidence View paper
- [15] Demonstration attack against in-context learning for code intelligence View paper
- [16] ICLShield: Exploring and Mitigating In-Context Learning Backdoor Attacks View paper
- [17] In-Context Interaction Learning: Boosting Adversarial Robustness for Model Security View paper
- [18] Provable test-time adaptivity and distributional robustness of in-context learning View paper
- [19] Combining large language models and or/ms to make smarter decisions View paper
- [20] Context-aware Prompt Tuning: Advancing In-Context Learning with Adversarial Methods View paper
- [21] Solid-SQL: Enhanced Schema-linking based In-context Learning for Robust Text-to-SQL View paper
- [22] Transformers Can Perform Distributionally-robust Optimisation through In-context Learning View paper
- [23] Impact of Positional Encoding: Clean and Adversarial Rademacher Complexity for Transformers under In-Context Regression View paper
- [24] Securing reliability: A brief overview on enhancing in-context learning for foundation models View paper
- [25] Adversarially Robust Decision Transformer View paper
- [26] Efficiently Robust In-Context Reinforcement Learning with Adversarial Generalization and Adaptation View paper
- [27] Prompt Optimization via Adversarial In-Context Learning View paper
- [28] Differential Transformer View paper
- [29] Training Adaptive and Sample-Efficient Autonomous Agents View paper
- [30] Contextual Vision Transformers for Robust Representation Learning View paper
- [31] Can In-Context Reinforcement Learning Recover From Reward Poisoning Attacks? View paper
- [32] Evaluating the Adversarial Robustness of Retrieval-Based In-Context Learning for Large Language Models View paper
- [33] Fine-tuning Does Not Remove Language Model Capabilities View paper
- [34] Context is the Key: Backdoor Attacks for In-Context Learning with Vision Transformers View paper
- [35] On the Role of Depth and Looping for In-Context Learning with Task Diversity View paper
- [36] Robustness through Forgetting: Unlearning Adversarial Contexts in Transformers View paper
- [37] On the Robustness of In-Context Learning with Noisy Labels: Train, Inference, and Beyond View paper
- [38] BAM-ICL: Causal Hijacking In-Context Learning with Budgeted Adversarial Manipulation View paper
- [39] ACloser LOOK AT THE ROBUSTNESS OF IN-CONTEXT LEARNING WITH NOISY LABELS View paper
- [40] UNDERSTANDING AND IMPROVING CONTINUOUS AD View paper
- [41] Adversarially robust transfer learning View paper
- [42] Learning Adversarially Fair and Transferable Representations View paper
- [43] On adversarial training without perturbing all examples View paper
- [44] Adversarially robust hypothesis transfer learning View paper
- [45] Learning Robust Rewards with Adversarial Inverse Reinforcement Learning View paper
- [46] Adversarial robustness in transfer learning models View paper
- [47] Wasserstein distance based deep adversarial transfer learning for intelligent fault diagnosis with unlabeled or insufficient labeled data View paper
- [48] CARD: Robustness-Preserving Transfer Learning for Network Intrusion Detection via Contrastive Adversarial Representation Distillation View paper
- [49] Augmenting fake content detection in online platforms: A domain adaptive transfer learning via adversarial training approach View paper
- [50] Synthetic-to-Real Transfer Learning for Chromatin-Sensitive PWS Microscopy View paper
- [51] Exploring robust features for improving adversarial robustness View paper
- [52] Adversarial feature alignment: Balancing robustness and accuracy in deep learning via adversarial training View paper
- [53] ProFeAT: Projected Feature Adversarial Training for Self-Supervised Learning of Robust Representations View paper
- [54] Distilling robust and non-robust features in adversarial examples by information bottleneck View paper
- [55] Evidence-Based Multi-Feature Fusion for Adversarial Robustness View paper
- [56] Learning More Robust Features with Adversarial Training View paper
- [57] Minimizing adversarial training samples for robust image classifiers: analysis and adversarial example generator design View paper
- [58] Feature purification: How adversarial training performs robust deep learning View paper

- [59] Adversarial Robustness through Disentangled Representations View paper
- [60] Few-Shot Anomaly Detection with Adversarial Loss for Robust Feature Representations View paper
- [61] Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning View paper
- [62] Using natural language explanations to improve robustness of in-context learning View paper
- [63] Are sample-efficient nlp models more robust? View paper
- [64] Enhancing in-context learning via linear probe calibration View paper
- [65] Meta-reinforcement learning robust to distributional shift via performing lifelong in-context learning View paper
- [66] Probing the decision boundaries of in-context learning in large language models View paper
- [67] Which examples to annotate for in-context learning? towards effective and efficient selection View paper
- [68] A theoretical design of concept sets: improving the predictability of concept bottleneck models View paper
- [69] Context - Enhanced Meta-Reinforcement Learning with Data-Reused Adaptation for Urban Autonomous Driving View paper
- [70] A Review on Machine Learning Applications in Localization in 5G and Beyond Wireless Communications View paper