EXPLORING THE ROBUSTNESS OF IN-CONTEXT LEARNING WITH NOISY LABELS

Chen Cheng[1*] **Xinzhi Yu**[2*] **Haodong Wen**[3*] **Jingsong Sun**[3]
Guanzhang Yue[4] **Yihao Zhang**[4] **Zeming Wei**[4†]
[1]ShanghaiTech University [2]Fudan University [3]Xi'an Jiaotong University [4]Peking University

ABSTRACT

Recently, the mysterious In-Context Learning (ICL) ability exhibited by Transformer architectures, especially in large language models (LLMs), has sparked significant research interest. However, the resilience of Transformers' in-context learning capabilities in the presence of noisy samples, prevalent in both training corpora and prompt demonstrations, remains underexplored. In this paper, inspired by prior research that studies ICL ability using simple function classes, we take a closer look at this problem by investigating the robustness of Transformers against noisy labels. Specifically, we first conduct a thorough evaluation and analysis of the robustness of Transformers against noisy labels during in-context learning and show that they exhibit notable resilience against diverse types of noise in demonstration labels. Furthermore, we delve deeper into this problem by exploring whether introducing noise into the training set, akin to a form of data augmentation, enhances such robustness during inference, and find that such noise can indeed improve the robustness of ICL. Overall, our fruitful analysis and findings provide a comprehensive understanding of the resilience of Transformer models against label noises during ICL and provide valuable insights into the research on Transformers in natural language processing. Our code is available at https://github.com/InezYu0928/in-context-learning.

1 INTRODUCTION

In recent years, Large Language Models (LLMs) have achieved significant success across various tasks in real-world applications. Transformer (Vaswani et al., 2017), as the typical backbone architecture, also emerges an intriguing ability known as In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2023a) that the model can learn a new task with only a few input-output pairs demonstrated during inference *without* modifying any model parameters. This mysterious property of transformers has attracted great research interests in understanding this ability, among which one representative perspective is investigating ICL through simple function classes (Garg et al., 2023).

As large language models like Vicuna (Zheng et al., 2023) and Llama2 (Touvron et al., 2023) were trained over documents collected in the real world without sufficient supervision, such documents might contain noisy information, which could affect the ability of these models during in-context inference. Moreover, recent works (Wei et al., 2023c; Pawelczyk et al., 2023) also showed that the output of ICL might be significantly manipulated by the label of demonstrations in the prompts, uncovering safety concerns about this prompt form. Motivated by these observations, in this work, we attempt to characterize and understand such ICL ability of transformers **with noisy labels**. Noisy label learning (Natarajan et al., 2013) has a wide breadth of literature in modern machine learning research (Xiao et al., 2015; Wang et al., 2018; Wu et al., 2023), as dataset collection may be costly and noisy (Deng et al., 2009). Further discussion on related work can be found in Appendix A. Therefore, the noisy label scenario in terms of ICL, is more practical and deserves a good understanding for advancing the safety and alignment of large language models, as discussed above.

[*]Equal contribution.
[†]Corresponding author: Zeming Wei (weizeming@stu.pku.edu.cn).

Given the complexity of language distributions and tasks, directly formulating and analyzing ICL's robustness under noisy label conditions is challenging. To address this challenge, a series of work (Garg et al., 2023; Bhattamishra et al., 2024) suggest studying transformers' ICL capability via **simple function classes**. Specifically, they propose to assess the ICL ability of Transformers by learning a simple function class in context. Following this research paradigm, our work also investigates the noisy label robustness of ICL using simple function classes, offering a more interpretable and straightforward framework.

In this work, we conduct a comprehensive evaluation of the robustness against noisy labels of ICL and also explore how label noise that exists in the training phase can affect it. Our findings suggest that transformers are indeed fairly robust to label noises in the in-context demonstrations, and it's also possible to further enhance such robustness by adding noises in the training set as data augmentation. Our fruitful and comprehensive experiments provide interesting findings and conclusions for understanding the robustness of in-context learning with noisy labels.

2 Preliminaries

Following the research paradigm of studying the ICL ability of Transformers through a set of simple function classes (Garg et al., 2023; Bhattamishra et al., 2024), we explore the Transformers' ICL robustness against noisy labels. In this section, we take the **noisy linear regression** task as an example to demonstrate the overall training-inference pipeline, and other function classes evaluated can also be similarly formulated. The overall pipeline involves the following two parts.

Transformer training with in-context demonstrations. In this part, we build a synthetic dataset to train a transformer, aiming for it to acquire ICL capability to learn linear regression functions from in-context demonstrations. Specifically, we consider the class of linear functions $F = \{f_w | f_w(x) = w^T x, w \in \mathbb{R}^d\}$, and each training sample in the dataset can be formulated as $P = (x_1, f_w(x_1), \cdots, x_k, f_w(x_k))$, with input samples and function parameter w both drawn from an isotropic Gaussian distribution, *i.e.* $x_i, w \sim N(0, I_d)$. For noisy-label learning, Gaussian noise $\epsilon_i \sim N(0, \sigma^2_{train})$ is added to the demonstrations, resulting in $f_w(x_i) + \epsilon_i$.

Transformer inference with in-context demonstrations. After training the transformer model $M_\theta(\cdot)$ with parameters θ, ICL inference is conducted by prompting $P = (x_1, f_w(x_1), \cdots, x_k, f_w(x_k), x_q)$, where $(x_i, f_w(x_i)$ represent **in-context demonstrations** and x_q is the **query input**. For in-context inference, a ground-truth function f_w from F is randomly sampled, with w and x_i both drawn from $N(0, I_d)$, resulting in $f_w(x) = w^T x$. In the noisy-linear regression, noise $\epsilon_i \sim N(0, \sigma^2_{test})$ is added independently to each demonstration $f_w(x_i) + \epsilon_i$. However, during evaluation, the ground truth of the query x_q is kept as $w^T x_q$ without adding noise.

3 Comprehensive evaluation of noisy ICL inference

In this section, we focus on the inference stage of ICL with various settings. In our main evaluation, we examine the robustness of the Transformer model with different distributions of label noises in the linear regression task, and conduct further investigations on the **non-*i.i.d.* noise setting** (in Appendix C), **different function complexity** (in Appendix D), and **different input dimensions** (in Appendix E). Following (Garg et al., 2023), we construct a dataset with 10,000 prompts, each containing 100 demonstrations with default dimension $d = 20$ as the default, and train GPT-2 models (Radford et al., 2019) consisting of 12 layers and 8 attention heads and a 256-dimensional embedding for evaluation. All results are averaged over 5 independent experiments.

Noise Distributions. In our experiments we consider the following label noise distributions: (1) *Gaussian distribution*: $w^T x_i + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$; (2) *Uniform distribution*: $w^T x_i + \epsilon_i$, where $\epsilon_i \sim U(-\sqrt{3\sigma}, \sqrt{3\sigma})$; (3) *Exponential distribution*: $w^T x_i + \epsilon_i$, where $\epsilon_i \sim Exp(\frac{1}{\sigma})$; (4) *Poisson distribution*: $w^T x_i + \epsilon_i$, where $\epsilon_i \sim P(\sigma^2)$; (5) *Multiplicative distribution*: $w^T x_i \times (1 + \epsilon_i)$, where $\epsilon_i \sim N(0, \sigma^2)$. (6) *Salt&Pepper distribution*: $w^T x_i + \epsilon_i$, where ϵ_i is either salt value or pepper value with probabilities $p/2$ each, and $\sigma = \sqrt{2}p$.

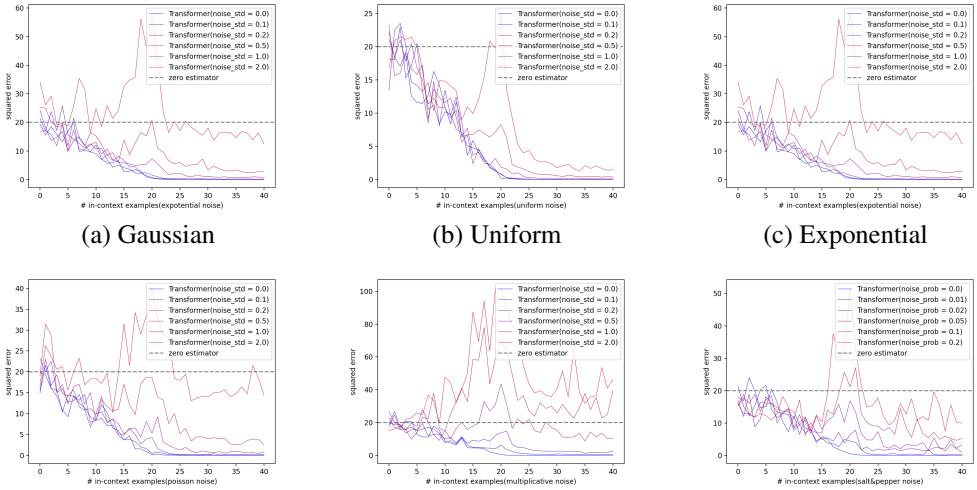
Robustness evaluation metrics. To evaluate the robustness of the transformer against label noises, following (Garg et al., 2023) we compare its performance with a few simple baselines including (a)

the least squares estimator, which computes the minimum-norm linear fit to the in-context examples (x_i, y_i); (b) *K-Nearest Neighbors*, involving the averaging of y_i values for the n nearest neighbors of x_{query}; (c) *the average of the values* $y_i x_i$ to estimate w and calculate the inner product of this estimate with x_{query}. Note that the least squares offer an optimal estimator for this problem and a lower bound for the best achievable error, while the other two baselines offer a consistent, computationally simpler estimator. Furthermore, we consider the normalized squared error

$$L_P = \frac{(M(P) - w^T x_{query})^2}{d},$$

where P is the input prompt and M is the trained model. Considering the presence of label noise, we define the model achieves the *satisfactory accuracy* if $L_P \leq 0.5$, thereby establishing a threshold for the noise level under which the Transformer model maintains robustness and providing a criterion to compare between ICL and the simple baselines.

Robustness analysis. As discussed above, we employ the loss functions introduced above to evaluate the trained Transformer models and the baselines across different noise types and varying levels of label noise. We plot the comparison of the transformer and other baselines under these settings in Figure 4 in Appendix B, including $\sigma_{test} \in \{0.2, 0.5, 1.0\}$, and also compare the robustness of the Transformers against different noise magnitude σ_{test} of a specific noise type in Figure 1.



(a) Gaussian (b) Uniform (c) Exponential

(d) Poisson (e) Multiplicative (f) S&P

Figure 1: Robustness Comparison under Different Noise Types and magnitudes. Each figure represents a noise and each line represents a σ_{test}. The X-axis represents the number of in-context examples.

From these results, we can directly draw the following observations:

i) Inadequate in-context examples. When the number of in-context examples falls below the input dimension ($d = 20$), the model's loss closely aligns with that of the least squares estimator.

ii) Near input-dimension number of examples. When the number of in-context examples nears the input dimension ($d = 20$), significant errors arise in the least-square estimator, whereas the ICL model's accuracy continues to improve.

iii) Sufficient in-context examples. Once the number of in-context examples exceeds $d = 20$, the ICL model's performance rapidly and significantly improves.

Furthermore, we can find that under **symmetric** noises (Gaussian, Uniform, Exponential, Poisson), transformers exhibit better performance in comparison to other baselines. Besides, when confronted with higher noise levels, the ICL model exhibits a challenge in ignoring label noise. This observation aligns with intuition, as inference in noisy linear regression mirrors the least squares solution with appropriately structured L_2 regularization. Additionally, upon introducing **Multiplicative** and **Salt&Pepper** noise, we observe that the transformer model does not perform well (worse

than baselines and zero estimators). This discrepancy suggests that transformer models might not generalize well to noise that deviates significantly from the **symmetric or discrete distributions** typically encountered in natural data.

Finally, we also find that there may exist a distinct threshold for each noise type, beyond which the transformer model's performance cannot outperform baselines. Once the noise level, denoted as σ, surpasses this threshold, the noise perceptibly impacts the model, rendering it non-negligible in its influence on performance. Through extensive evaluations, we can estimate such thresholds for different noise distributions and summarize them in Table 1.

Table 1: Estimated robustness threshold of label noise σ to perform comparably with the case of no noise.

Distribution	**Gaussian**	**Uniform**	**Exponential**	**Poisson**	**Multiplicative**	**S&P**
Threshold σ	0.45	1.10	0.39	0.43	0.23	0.22

4 INFLUENCE OF TRAINING WITH NOISY LABELS ON ROBUSTNESS

In this section, we further explore how adding label noises in the training dataset influences the robustness of the Transformer during in-context inference. Since the real-world training corpus for large-scale Transformers (*e.g.*, LLMs) may be noisy and lack supervision, we attempt to seek insights into whether such noises may be unexpectedly helpful in improving the robustness of the model during ICL by this evaluation framework. Intriguingly, we find that adding such noises to the training labels can indeed enhance the desired robustness.

Experiment Set-up. To construct a noisy training set, instead of vanilla demonstrations $\{x_i, f_w(x_i)\}$, we add Gaussian noises to the labels as $\{x_i, f_w(x_i) + \epsilon_i\}$ with $\epsilon_i \sim N(0, \sigma^2)$. We still adopt the GPT-2 (Radford et al., 2019) model architecture and compare $\sigma_{\text{train}} \in \{0.2, 0.4, 0.6, 0.8, 1\}$. In this experiment, we further take the model size into consideration and compare the following three model sizes: *standard* (8 heads, 12 layers, embedding size = 256, which is adopted in the experiments above), *small* (4 heads, 6 layers, embedding size = 128), and *tiny* (2 heads, 3 layers, embedding size = 64).

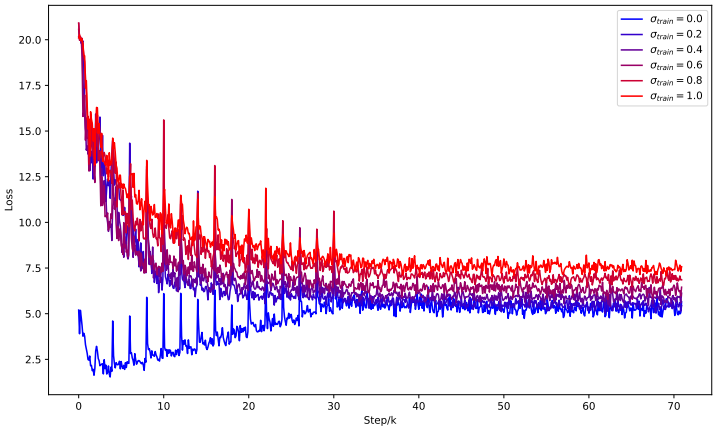


Figure 2: The training loss for a standard-sized transformer with Gaussian noise of $\sigma_{train} \in \{0.0, 0.2, 0.4, 0.6, 0.8, 1.0\}$

Training convergence with noisy labels. For each σ, we train the model with corresponding label noise distribution and plot the training loss curve in Figure 2 with each line representing a model. Similar to (Garg et al., 2023), we used curriculum learning during training where we started from a low-dimensional space and gradually increased to 20. Note that the model associated with $\sigma_{train} = 0.0$ represents the case of no-label noise, which is exactly the standard training setting. By inspecting the convergence situation, it is clear that for small label noises ($\sigma < 0.6$), the training is well converged and the final loss is very close to the standard training. Even in the case of large noise

variance ($\sigma = 0.6, 0.8, 1$), the training can still converge to a certain extent. The main difference between training with noise and without noise is at the beginning where the loss remains high even with the adoption of curriculum learning. However, curriculum learning still works well since all the models nearly converge at step 30k.

Robustness analysis with noisy label training. We further conduct in-context inference with different inference Gaussian noise $\sigma_{test} \in \{0.2, 0.4, 0.6, 0.8, 1\}$ (the same as what we used to train our models). Still, we focus on the class of noisy linear functions $F = \{f_w | f_w(x) = w^T x + \epsilon, w \in R^d\}$, in $d = 20$ dimensions. To ensure that the experimental results are convincing and reproducible, we repeated the experiment 1280 times and took an average value to plot the figure. Moreover, for the sake of fairness, we use the same set of prompts for different models in the same setting during inference. The results are Figure 3.

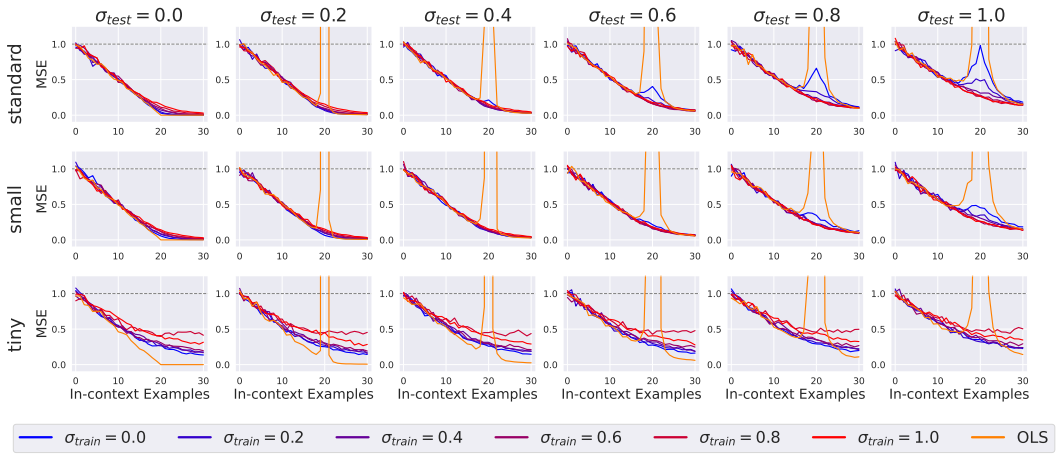


Figure 3: Effects of Gaussian Noise Magnitude and Model Size on Performance.

Robustness improved with training label noises Intriguingly, we find that adding label noises in the training dataset can indeed improve the robustness of Transformers during ICL inference. Specifically, taking the case of standard model size as an example (the first row), though the performance of different models with small test noise $\sigma_{test} < 0.5$ is small, when the test noise increases $\sigma_{test} > 0.5$ the models trained with noises (red lines) outperforms the models trained without noise and with small noise ($\sigma_{train} < 0.2$, blue lines) significantly. Similar phenomena are also observed in smaller model sizes, though the improvement of robustness exhibited in large test noises is relatively decreased.

Besides, we can also observe that for smaller test noises and the natural case ($\sigma_{test} = 0$), the performance of models trained with large training noises is similar to the vanilla cases for larger model sizes (standard and small), but decreases notably for tiny model size, showing the importance on the model capacity in learning with noisy labels. The conclusion is consistent with training with LLMs, where training with noisy corpus can still help the model generalize well and be robust against noisy labels during ICL (Min et al., 2022) given the large size of the model.

5 Conclusion

In this paper, we present a comprehensive study on the robustness of Transformers' in-context learning ability against label noises. First, we conduct extensive experiments covering different noise types, magnitudes, function complexities, and input dimensions to show that Transformers retain their ICL efficiency against various label noises, along with sufficient analysis and discussions. Furthermore, we discover that introducing similar noises into the training set enables good convergence and potentially enhances the robustness of transformer models. Our findings contribute to a deeper understanding of Transformers' ICL adaptability to label noise and provide new insights into ICL research in real-world models.

ACKNOWLEDGEMENT

This work was sponsored by the Beijing Natural Science Foundation's Undergraduate Initiating Research Program (Grant No. QY23041) and the National Natural Science Foundation of China (Grant No. 62172019).

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. 9

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 9

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023. 9

Satwik Bhattamishra, Arkil Patel, Phil Blunsom, and Varun Kanade. Understanding in-context learning in transformers and LLMs by learning to learn discrete functions. In *ICLR*, 2024. 2, 9

Rishi Bommasani et al. On the opportunities and risks of foundation models, 2022. 9

Tom Brown et al. Language models are few-shot learners, 2020. 1, 9

Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods, 2017a. 9

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017b. 9

Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*, 2023. 9

Huanran Chen, Yichi Zhang, Yinpeng Dong, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*, 2023. 9

Filipe R. Cordeiro and Gustavo Carneiro. A survey on deep learning with noisy labels: How to train your model when you cannot trust on the annotations?, 2020. 9

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009. 1, 9

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. A survey on in-context learning, 2023a. 1, 9

Yinpeng Dong, Huanran Chen, Jiawei Chen, Zhengwei Fang, Xiao Yang, Yichi Zhang, Yu Tian, Hang Su, and Jun Zhu. How robust is google's bard to adversarial image attacks? *arXiv preprint arXiv:2309.11751*, 2023b. 9

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes, 2023. 1, 2, 4, 9

Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. Towards robustness to label noise in text classification via noise modeling. In *CIKM*. ACM, October 2021. 9

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 9

Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images, 2018. 9

Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey, 2021. 10

Neel Jain, Avi Schwarzschild, Yuxin Wen, Gowthami Somepalli, John Kirchenbauer, Ping yeh Chiang, Micah Goldblum, Aniruddha Saha, Jonas Geiping, and Tom Goldstein. Baseline defenses for adversarial attacks against aligned language models, 2023. 9

Lu Jiang, Zhengyuan Zhou, Thomas Leung, Li-Jia Li, and Li Fei-Fei. Mentornet: Learning data-driven curriculum for very deep neural networks on corrupted labels. In *International conference on machine learning*, pages 2304–2313. PMLR, 2018. 9

Ang Li, Yifei Wang, Yiwen Guo, and Yisen Wang. Adversarial examples are not real features. *arXiv preprint arXiv:2310.18936*, 2023. 10

Haoyang Liu, Maheep Chaudhary, and Haohan Wang. Towards trustworthy and aligned machine learning: A data-centric survey with causality perspectives. *arXiv preprint arXiv:2307.16851*, 2023. 9

Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack, 2022. 9

Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning?, 2023. 9

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 9

Naresh Manwani and Senior Member Ieee P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43:1146–1151, 2011. 9

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022. 5, 9

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013. 1

Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach, 2017. 9

Martin Pawelczyk, Seth Neel, and Himabindu Lakkaraju. In-context unlearning: Language models as few shot unlearners, 2023. 1

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 2, 4

Tilman Räuker, Anson Ho, Stephen Casper, and Dylan Hadfield-Menell. Toward transparent ai: A survey on interpreting the inner structures of deep neural networks, 2023. 10

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 2022. 9

Hugo Touvron et al. Llama 2: Open foundation and fine-tuned chat models, 2023. 1

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 10

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 1, 9

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *ICML*, 2023. 9

Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models, 2023. 9

Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. *arXiv preprint arXiv:2010.04055*, 2020. 10

Yifei Wang, Yuyang Wu, Zeming Wei, Stefanie Jegelka, and Yisen Wang. How language models defend themselves against jailbreak: Theoretical understanding of self-correction through in-context learning. *arXiv preprint*, 2024. 9

Yisen Wang, Weiyang Liu, Xingjun Ma, James Bailey, Hongyuan Zha, Le Song, and Shu-Tao Xia. Iterative learning with open-set noisy labels. In *CVPR*, 2018. 1, 9

Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*, 2019. 9

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023a. 9

Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *CVPR*, 2023b. 9

Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023c. 1, 9

Zeming Wei, Jingyu Zhu, and Yihao Zhang. Sharpness-aware minimization alone can improve adversarial robustness. In *ICML Workshop on New Frontiers in Adversarial Machine Learning*, 2023d. 9

Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning, 2023. 9

Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. Noisywikihow: A benchmark for learning with real-world noisy labels in natural language processing, 2023. 1

Tong Xiao, Tian Xia, Yi Yang, Chang Huang, and Xiaogang Wang. Learning from massive noisy labeled data for image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2691–2699, 2015. 1

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022. 9

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 9

Yihao Zhang and Zeming Wei. Boosting jailbreak attack with momentum. In *ICLR 2024 Workshop on Reliable and Responsible Foundation Models*, 2024. 9

Yihao Zhang, Hangzhou He, Jingyu Zhu, Huanran Chen, Yifei Wang, and Zeming Wei. On the duality between sharpness-aware minimization and adversarial training. *arXiv preprint arXiv:2402.15152*, 2024a. 9

Yihao Zhang, Zeming Wei, Jun Sun, and Meng Sun. Towards general conceptual model editing via adversarial representation engineering. *arXiv preprint*, 2024b. 9

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023. 1

Dawei Zhu, Michael A. Hedderich, Fangzhou Zhai, David Ifeoluwa Adelani, and Dietrich Klakow. Is bert robust to label noise? a study on learning with noisy labels in text classification, 2022. 9

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to ai transparency, 2023a. 10

Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023b. 9

A ADDITIONAL RELATED WORK

A.1 UNDERSTANDING IN-CONTEXT LEARNING

The mysterious ability of in-context learning (Brown et al., 2020; Dong et al., 2023a), which typically occurs in attention-based model architectures like transformers (Vaswani et al., 2017), has attracted significant research interest in understanding its underlying mechanism and designing better learning algorithms (Lu et al., 2023; Min et al., 2022). Without modifying model parameters, these models can conduct various downstream tasks with a few input-output pairs as demonstrations included before the test input.

However, given the complexity of natural languages, a popular research paradigm of understanding the ICL ability of Transformers is characterizing their learnability through a set of simple function classes, like linear functions (Garg et al., 2023) and discrete functions (Bhattamishra et al., 2024), which are more interpretable and easier to formulate. There are also other works that investigate such ability of ICL through simple functions and propose various understandings, *e.g.* interprets the inference with in-context demonstrations as implicit gradient decent (Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Wang et al., 2024). Specifically, they showed that transformers can learn the specified task through these demonstrations with implicit optimization in the hidden spaces of transformers. Besides, there are also interpretations of ICL through other perspectives, like Bayes inference (Xie et al., 2022) and PAC-learnability (Wies et al., 2023).

A.2 NOISY LABEL LEARNING

Modern deep learning methods commonly face the presence of noisy labels in the training data, since data collection and annotation may be costly and biased (Deng et al., 2009). Such noises in labels may be symmetric, asymmetric, and even from the open set that is not contained in the training classes (Cordeiro and Carneiro, 2020). To tackle this issue, numerous efforts have been made to robustify the training process against such noises. Typical approaches include estimating noise transition matrix (Patrini et al., 2017), designing robust loss functions (Manwani and Sastry, 2011; Wang et al., 2019), sample weighting (Guo et al., 2018; Wang et al., 2018) and selection (Jiang et al., 2018).

Though broadly explored in this literature, learning with noisy labels is still an open problem in modern machine learning research (Song et al., 2022). Moreover, there are also concurrent threads toward studying the robustness against label noise for the text modular (Garg et al., 2021; Zhu et al., 2022). Despite broad explorations, few works have investigated the robustness of in-context learning with noisy labels. Though studying the ICL ability of transformers through noisy linear regression, (Garg et al., 2023) only studies the standard Gaussian distribution noise, which is only one specific noise type and magnitude in our work. In this paper, we take steps further to systematically investigate the robustness of ICL against various label noise settings, including both the train and inference phases, and various types and magnitudes of noises.

A.3 LANGUAGE MODEL SAFETY AND ALIGNMENT

With the milestone success of the fast-paced development of large language models (LLMs), concerns regarding their potential for harmful generation and malicious usage have emerged (Bommasani et al., 2022; Chen and Shu, 2023; Liu et al., 2023; Zhang et al., 2024b), which are typically referred to as the jailbreaking issue (Zou et al., 2023b; Wei et al., 2023a; Dong et al., 2023b; Zhang and Wei, 2024). Such risks further extend to the in-context learning scenario, as recent work (Wang et al., 2023; Wei et al., 2023c) showed, it is possible to manipulate the safety and alignment of language models by maliciously inducing noisy labels in the demonstrations. Therefore, it is of significance to study the robustness of in-context learning with label noises.

Moreover, different from defense against conventional adversarial noises (Goodfellow et al., 2014; Carlini and Wagner, 2017b; Liu et al., 2022; Chen et al., 2023; Wei et al., 2023d) among which adversarial training methods (Madry et al., 2017; Zhang et al., 2019; Wei et al., 2023b; Zhang et al., 2024a) that adding adversarial noises in the training loops are shown to be the most effective among various defense methods (Carlini and Wagner, 2017a; Athalye et al., 2018), it is recently shown that AT may not be an effective defense (Jain et al., 2023) for LLMs. Such difference further underscores

the challenge of the interpretability of internal mechanism (Islam et al., 2021; Räuker et al., 2023; Zou et al., 2023a) and adversarial examples (Tsipras et al., 2018; Wang et al., 2020; Li et al., 2023) in the context of LLMs. In this work, we explore a similar problem: whether adding label noises in the training demonstrations can enhance the corresponding robustness during inference.

B Complete visualizations for main evaluation in Section 3

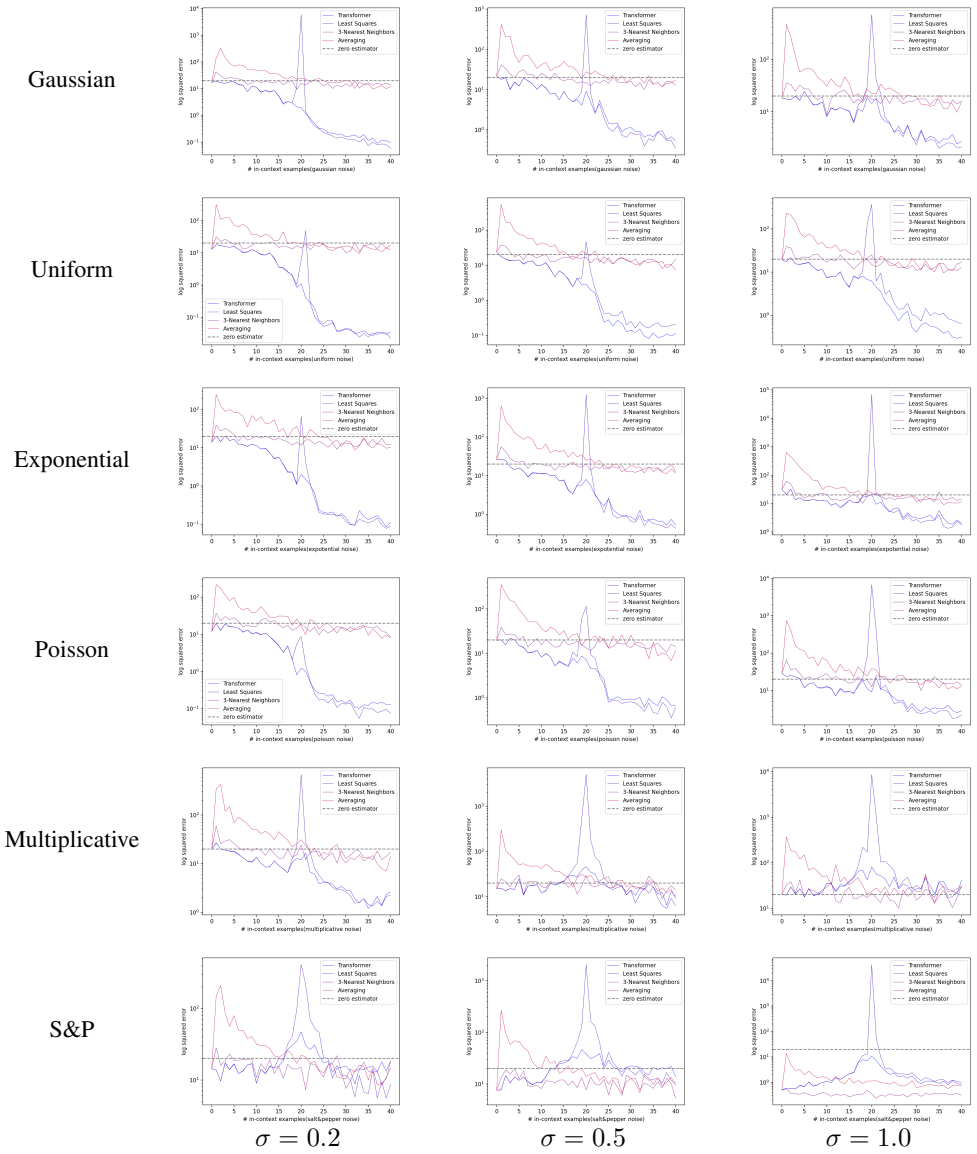


Figure 4: Robustness Comparison under Different Noise Types and magnitudes. Each figure represents a noise and each line represents a σ_{test}. The X-axis represents the number of in-context examples.

C Non-i.i.d. noises

We further extend our experiments to the *non-i.i.d.* noises scenario. Specifically, we replace the *i.i.d.* label noises with some outlier demonstrations, which is a particular case of *non-i.i.d.* noises, and analyze how the transformers respond to such *non-i.i.d* noise in the demonstrations.

Experiment set-up. In this experiment, we still focus on the class of noisy linear functions $F = \{f_w | f_w(x) = w^T x + \mathbb{1}_{outlier} * \epsilon, w \in R^d\}$ in $d = 20$ dimensions, where noises are only injected into demonstrations of **outliers**. Similar to the construction of the noisy demonstrations, we sample $x_1, \cdots, x_k, x_{query}$ from the isotropic Gaussian distribution $N(0, I_d)$. Then, we randomly select m outliers $\mathbb{S} = \{i_1, \cdots, i_m\}$, and generate *i.i.d.* Poisson noise $\{\epsilon_{i_1}, \cdots, \epsilon_{i_m}\}$ with noise magnitude σ. Subsequently, each $y_i = w^T x_i + \mathbb{1}_{i \in S} * \epsilon_i$ is computed. Finally, the prompt P is constructed as $P = (x_1, y_1, x_2, y_2, \cdots, x_k, y_k, x_{query})$.

Robustness analysis. We present the results with small multiples in Figure 5. Overall, the transformer model shows superior performance than other baselines, which is similar to the results of *i.i.d.* noises presented above and shows its consistent robustness against such noises, which we elaborate on below.

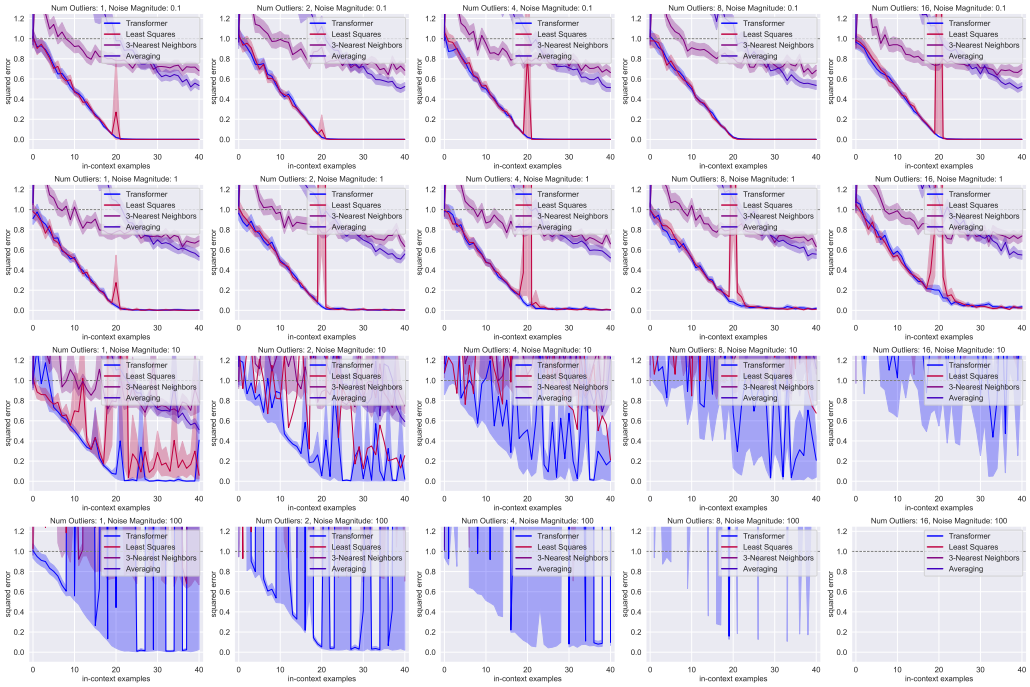


Figure 5: In-context learning on prompts with outliers. We evaluate the trained model on prompts with outliers in two cases. (1) On the rows, we explored the effect of the number of outliers ({1,2,4,8,16}), and (2) on the columns, we explored the effect of the magnitude of outliers ({0.1,1,10,100}). (Error averaged over 1280 prompts. 90% confidence intervals over 1000 bootstrap trials.)

Outliers of different magnitudes and equal quantity. Analyzing different columns in Figure 5 where the noise of varying magnitudes was introduced to a specified number of prompt outputs, we can see that the trained model exhibits greater robustness and stability compared to baseline models, particularly as the noise magnitude escalates. Notably, the lower bound of the 0.9 confidence interval closely aligns with the actual data when both the noise magnitude and the number of outliers are minimal, suggesting that the model maintains high prediction accuracy under these conditions.

Different quantities of outliers with uniform magnitude. On the other hand, from different rows in Figure 5 where the sub-figures involve different numbers of outliers and a consistent magnitude within the prompts, we can see an increase in the trained model's robustness and stability as the number of outliers rises.

D Different function complexities

In this part, to evaluate the transformer's ICL ability in terms of capturing more complex patterns within input-output mappings, which may be closer to real-world data, we broaden our exploration beyond linear relationships by exploring the task of **polynomial regression** ranging from degrees

of 2 to 5. We compare the transformer's robustness with both the least-squares method and a 3-NN model.

Experiment Set-up. We continue to adopt the GPT-2 model architecture. To construct a noisy training set, instead of vanilla demonstrations $\{x_i, f_w(x_i)\}$, we introduce Gaussian noise to the labels as $\{x_i, f_w^{(n)}(x_i) + \epsilon_i\}$, where

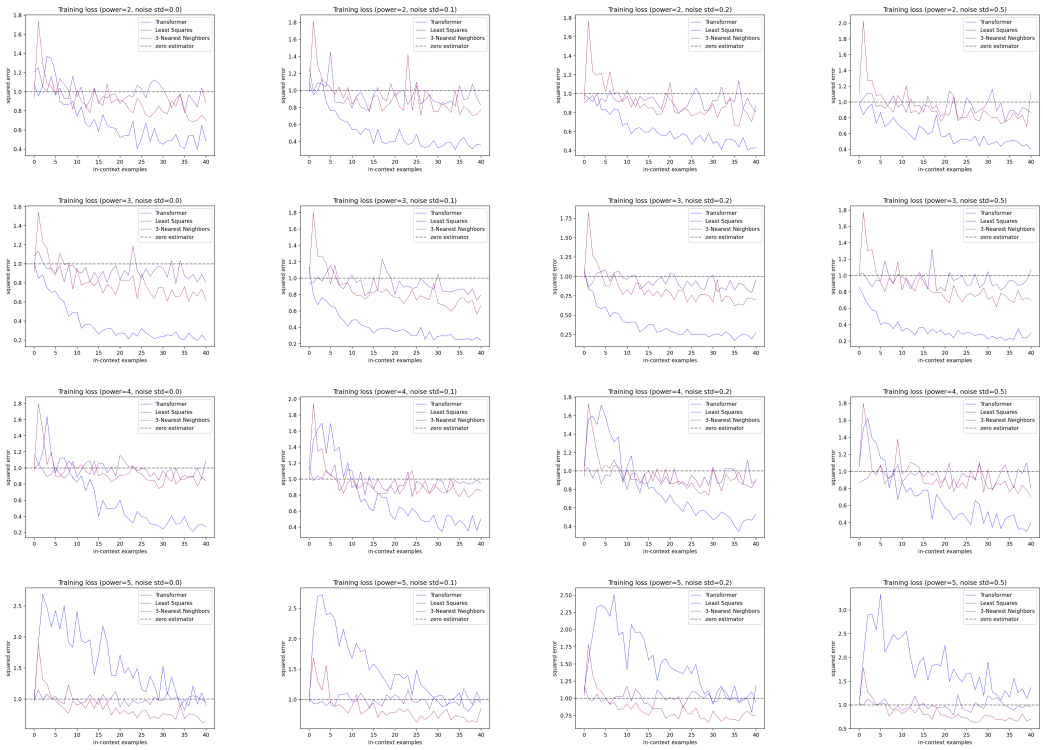
$$f_w^{(n)}(x_i) = \sum_{i=1}^{n} w_i \cdot x_i^n, \tag{1}$$

and a_i is sampled from $N(0, 1)$. Here, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, and we consider $\sigma \in \{0.1, 0.2, 0.5\}$.

Baselines. We introduce two baseline methods for evaluating the transformer's ability to model polynomial objective functions:

(1) Least Squares Model (LSM): The Ordinary Least Squares (OLS) regression constructs polynomial features up to the specified degree n for each input sample and computes the coefficients $\mathbf{w}$ using the least squares optimization $\mathbf{w} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\mathbf{y}$ where $\mathbf{X}$ is the design matrix consisting of polynomial features, $\mathbf{y}$ is the vector of labels, and $\mathbf{w}$ are the coefficients.

(2) k-Nearest Neighbors Model (k-NN): We extend the k-Nearest Neighbors algorithm applied in the experiments above with polynomial features. For each test point x_{test}, the predicted output $\hat{y}$ is computed by averaging the labels of the 3 nearest neighbors.



$\sigma_{\text{test}} = 0.0$ $\sigma_{\text{test}} = 0.1$ $\sigma_{\text{test}} = 0.2$ $\sigma_{\text{test}} = 0.5$

Figure 6: **noisy ICL performance** comparison for models trained with different polynomial degrees. Each figure represents an inference noise level σ_{test}, and each line represents a polynomial degree. The X-axis represents the number of in-context examples.

Robustness analysis. We evaluate the Transformer model's performance on polynomial functions with degrees g from 2 to 5 and plot the results in Figure D. We found that while the Transformer consistently outperformed baseline models for $g = 2$ and $g = 3$, its performance varied as polynomial complexity increased. For objectives of $g = 4$, the Transformer initially lagged but eventually surpassed baselines significantly. However, for objectives of $g = 5$, its efficacy declined compared to the baselines.

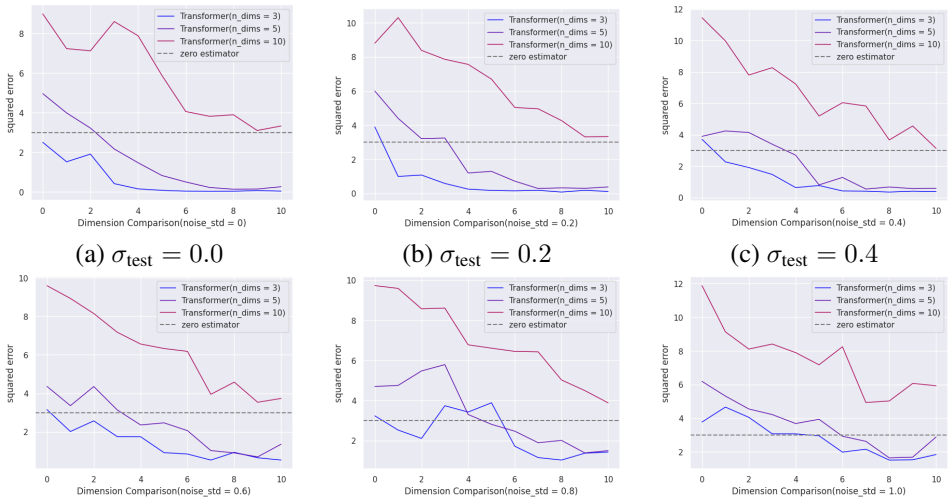
This performance disparity can be attributed to several factors. Firstly, the heightened polynomial degree engenders an escalated functional complexity, posing formidable challenges to the Transformer's capacity to model intricate higher-order nonlinear relationships. Secondly, the dearth of adequately diverse examples may exacerbate performance degradation, impeding the Transformer's ability to accurately capture and characterize complex functions due to insufficient training data. This limitation underscores the Transformer's susceptibility to "in-context" learning deficits when confronted with inadequate exemplars. Lastly, the phenomenon of overfitting may further compound performance disparities, particularly evident in the context of degree 5 objectives.

Notwithstanding these challenges, our comprehensive analysis underscores the Transformer model's commendable performance in addressing polynomial objective functions encountered in routine real-world scenarios. Its demonstrated prowess in generalization and robustness, particularly evident in outperforming baseline methods for lower-degree polynomials, underscores its potential efficacy in diverse applications and its capacity to navigate intricate nonlinear modeling tasks.

E DIFFERENT INPUT DIMENSIONS

In this section, we change the input vector dimension d to see how the robustness of the model varies with it.

Experiment Set-up. In this experiment, we still focus on the class of linear functions $F = \{f_w | f_w(x) = w^T x + \epsilon, w \in R^d\}$, where d varies in $\{3, 5, 10\}$, and $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Specifically, we train the transformer model with different input dimensions, then we add Gaussian noise $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ during evaluation to study the influence of different dimensions on the robustness of Transformers.



(a) $\sigma_{\text{test}} = 0.0$ (b) $\sigma_{\text{test}} = 0.2$ (c) $\sigma_{\text{test}} = 0.4$

(d) $\sigma_{\text{test}} = 0.6$ (e) $\sigma_{\text{test}} = 0.8$ (f) $\sigma_{\text{test}} = 1.0$

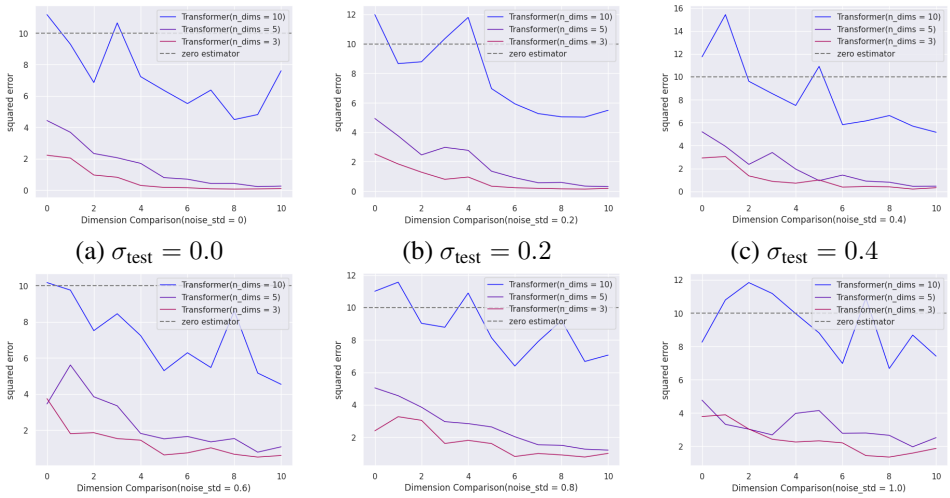
Figure 7: **ICL performance** comparison for models with **different input dimensions**. Each figure represents an inference noise level σ_{test}, and each line represents a model. The X-axis represents the number of in-context examples.

Robustness Analysis. The experiment results are shown in Figure 7, where we can see the larger the input dimension, the worse the performance of the model in in-context learning. The performance difference between the cases of $d = 3$ and $d = 5$ is relatively small, while the model performance decreases significantly when the model $d = 10$.

This may be because when the dimension increases, the domain of the function class to be learned expands greatly, which exceeds the model capacity, making the training much harder. Meanwhile, in the ICL process, the function to be learned itself becomes more complex due to the rise in the dimension of the input examples, so the model with higher dimensions learns poorly with the same number of examples.

F REVISITING INPUT DIMENSIONS WITH NOISY LABEL TRAINING

We also revisit the impact of the input dimensions on the robustness in the case of noisy label training. The overall experimental setup of this experiment is similar to that of section E, but it is worth noting that here we add Gaussian noise $\sigma \in \{0, 0.2, 0.4, 0.6, 0.8, 1\}$ during the training process in different input dimensions, and test the model with the same training noise in different dimensions in different test noise levels. The results are shown in Figure 8, where we can still observe the degradation of model performance with increasing dimensions in different training noises. Intuitively, we believe that the change rule of model performance with dimension in the case of training noise is not fundamentally different from that in the case of no training noise, and in both cases, the increase of dimension will make pre-training and ICL more difficult, which will result in the degradation of performance.



(a) $\sigma_{\text{test}} = 0.0$ (b) $\sigma_{\text{test}} = 0.2$ (c) $\sigma_{\text{test}} = 0.4$

(d) $\sigma_{\text{test}} = 0.6$ (e) $\sigma_{\text{test}} = 0.8$ (f) $\sigma_{\text{test}} = 1.0$

Figure 8: **(a-h) noisy ICL performance** comparison for models trained with different dimensions. Each figure represents an inference noise level σ_{test}, and each line represents a model trained with the same noise $\sigma = 0.6$ and different input dimension $d \in \{3, 5, 10\}$. The X-axis represents the number of in-context examples.