A CLOSER LOOK AT 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) prowess 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 study ICL ability using simple function classes, we delve deeper into this issue by investigating the robustness of ICL Transformers against noisy labels. Specifically, we conduct a thorough assessment demonstrating that Transformers exhibit notable resilience against diverse types of noise in demonstration labels, surpassing prior simplistic observations. Furthermore, we explore whether introducing noise into the training set, akin to a form of data augmentation, enhances such robustness during inference. Our comprehensive findings provide valuable insights into the resilience of Transformer models against label noise, thereby laying a foundational framework for further advancements in this domain.

1 INTRODUCTION

In recent years, Large Language Models (LLMs) have significantly succeeded in various real-world applications. The Transformer architecture (Vaswani et al., 2017) has revealed an intriguing ability, known as In-Context Learning (ICL) (Brown et al., 2020; Dong et al., 2023a), enabling it to learn new tasks from a few input-output pairs during inference without altering any model parameters. This mysterious property has sparked considerable research interest, particularly in investigating ICL through simple function classes (Garg et al., 2023).

Large language models, such as Vicuna (Zheng et al., 2023) and Llama2 (Touvron et al., 2023), trained on unsupervised real-world documents, may contain noisy information that affects their in-context inference capability. Furthermore, recent studies (Wei et al., 2023b; Pawelczyk et al., 2023) have shown that ICL outputs can be significantly influenced by the labels in the demonstration prompts, raising safety concerns. Inspired by these findings, this study aims to explore and understand the ICL capability of transformers with noisy labels.

Noisy label learning (Natarajan et al., 2013) is extensively covered in modern machine learning literature (Xiao et al., 2015; Wang et al., 2018; Wu et al., 2023), given the often costly and noisy dataset collection process (Deng et al., 2009). However, despite numerous studies on language models' noisy label settings (Garg et al., 2021; Zhu et al., 2022), few have systematically addressed these issues specifically for ICL. Further discussion on related work can be found in Appendix A.

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, Garg et al. (2023) suggest studying transformers' ICL capability via simple function classes, specifically assessing their ability to learn a function class in-context. Following this research approach, our work also

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

investigates the noisy label robustness of ICL using simple function classes, offering a more interpretable and straightforward framework.

It's worth noting that while Garg et al. (2023) discussed this robustness, they only considered a specific label noise distribution during the ICL inference phase. In contrast, our study conducts a more comprehensive evaluation of this robustness, including how label noise during the training phase influences it. Our findings indicate that transformers are quite robust to label noise in in-context demonstrations, and this robustness can be further enhanced by incorporating noise into the training set as a form of data augmentation.

Overall, our fruitful and comprehensive experiments provide interesting findings and conclusions for understanding the robustness of in-context learning with noisy labels.

2 Background and Preliminaries

Following the approach of Garg et al. (2023), we explore ICL's properties through specific simple function classes, focusing on the noisy linear regression task using linear functions as the primary evaluation task. The overall pipeline involves: (1) training a transformer on a synthetic dataset of prompts sampled from this function class; (2) evaluating the transformer's ICL performance by sampling a function from the class, as detailed below.

Transformer training with in-context demonstrations. In this approach, 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. We consider the class of linear functions $F = f_w | f_w(x) = w^T x, w \in \mathbb{R}^d$. Each training sample for the transformer is 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 $x_i, w \sim N(0, I_d)$. For noisy-label learning, Gaussian noise $\epsilon_i \sim N(0, \sigma^2_{train})$ is added to the demonstrations, resulting in $f_w(x_i) + \epsilon_i$. Following Garg et al. (2023), we set $d = 20$ and construct a dataset with 10,000 prompts, each containing 100 demonstrations.

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, resulting in $f_w(x_i) + \epsilon_i$. However, for calculating the loss, the ground truth of the query x_q is kept as $w^T x_q$ without adding noise.

3 Comprehensive Robustness Evaluation of Noisy ICL Inference

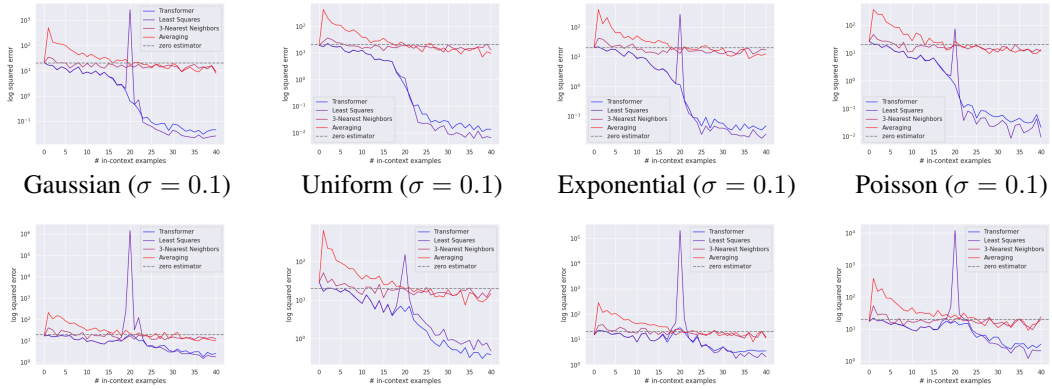
In this section, we focus on the inference stage of ICL with different distributions of label noises, including Gaussian, Uniform, Exponential, and Poisson distributions. We follow the same setting as Garg et al. (2023) and deploy their released pre-trained GPT-2 model (Radford et al., 2019) consists of 12 layers, 8 attention heads, and a 256-dimensional embedding for evaluation. However, unlike Garg et al. (2023) that only consider Gaussian noise with $\sigma_{test} = 1$, our evaluation is more comprehensive and substantial.

Noise Distributions. In this experiment, we consider the following noise distributions: (1) **Gaussian distribution**, $\epsilon_i \sim N(0, \sigma^2)$; (2) **Uniform distribution**, $\epsilon_i \sim U(-\sqrt{3\sigma}, \sqrt{3\sigma})$, where $U(\cdot, \cdot)$ represents the uniform distribution; (3) **Exponential distribution**, $\epsilon_i = \hat{\epsilon}_i - \frac{1}{\sigma}$ and $\hat{\epsilon}_i \sim Exp(\frac{1}{\sigma})$. Note that by standardizing the noise, we produce noise that has zero means and captures the moral of the exponential distribution $Exp(\cdot)$; (4) **Poisson distribution**, $\epsilon_i = \hat{\epsilon}_i - \sigma^2$ and $\hat{\epsilon}_i \sim P(\sigma^2)$. Similarly, by standardizing the noise, we produce noise that has zero mean and captures the moral of the Poisson distribution $P(\cdot)$. We also consider the case of non-i.i.d. noise which is detailed in Appendix C.2.

Robustness evaluation. To evaluate the robustness of the transformer against label noises, we compare its performance with other simple learning algorithms like (Garg et al., 2023) and also define a metric named *satisfactory accuracy*, indicating a desirable performance in this task. More details about the baselines and the metric are illustrated in Appendix B.

Experiment results. As discussed above, we employ the loss functions introduced above to evaluate the transformer model and baseline models' robustness across different noise types and varying levels of label noise. We plot the representative comparison of the transformer and other baselines under these settings in Figure 1, including $\sigma_{test} \in \{0.1, 1\}$, and leave the comprehensive visualization in Figure 3. In addition, we report some examples of test error and the corresponding satisfactory accuracy under label noise in Table 2 in Appendix C.

Figure 1: Robustness Comparison under Different Noise Types and magnitudes.



Gaussian ($\sigma = 0.1$) Uniform ($\sigma = 0.1$) Exponential ($\sigma = 0.1$) Poisson ($\sigma = 0.1$)

Gaussian ($\sigma = 1$) Uniform ($\sigma = 1$) Exponential ($\sigma = 1$) Poisson ($\sigma = 1$)

Robustness analysis. As shown in these figures and tables, we can find that under most error levels, loss functions associated with symmetrical distributions, transformers exhibit better performance in comparison to other baselines. However, when confronted with higher noise levels, the ICL model exhibits a challenge in ignoring label noise. Drawing from these discoveries, we deduce the presence of 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. From the experiment results we 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.

Noise Distribution	**Gaussian**	**Uniform**	**Exponential**	**Poisson**
Threshold σ	0.45	1.10	0.39	0.43

Further observations. Beyond the preceding analysis, we present several significant observations from our experimental results:

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.

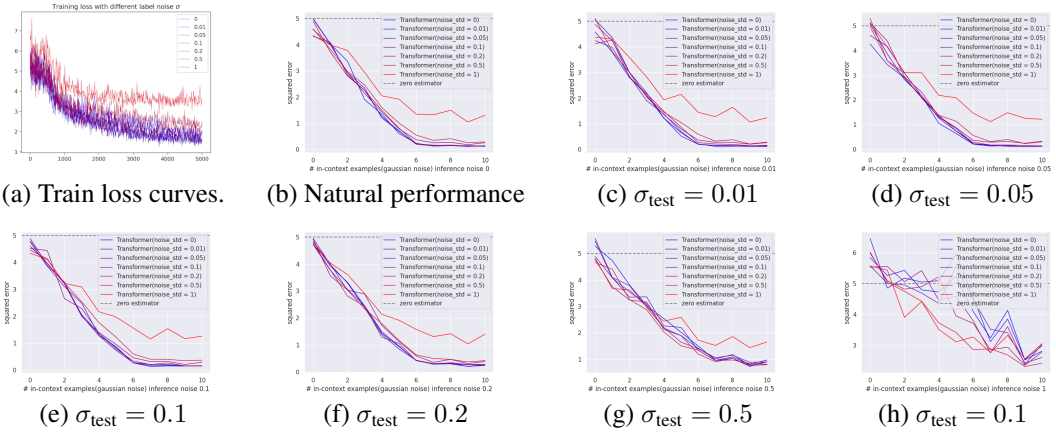
ii) Near input-dimension number of examples. When the number of in-context examples nears the input dimension ($d = 20$), significant errors arise in the least-square estimator, whereas the ICL model's accuracy continues to improve.

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

4 How training with noisy demonstrations affect robustness

Experiment Set-up. Following Garg et al. (2023), we still adopt the GPT-2 (Radford et al., 2019) model architecture. 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 $\epsilon_i \sim N(0, \sigma^2)$. In this experiment, we compare $\sigma_{train} \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1\}$. In the following, we first show that the training of transformers can still converge well with moderate label noises, then further compare their performance with training without label noises.

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(a) with each line representing a model. Note that the model associated with $\sigma = 0$ represents the case of no-label noise, which is exactly the standard training setting. It is clear that for small label noises ($\sigma < 0.5$), the training is well converged and the final loss is very close to the standard training. Even in the case of large noise variance ($\sigma = 0.5, 1$), the training can still converge to a certain extent.



(a) Train loss curves. (b) Natural performance (c) $\sigma_{\text{test}} = 0.01$ (d) $\sigma_{\text{test}} = 0.05$

(e) $\sigma_{\text{test}} = 0.1$ (f) $\sigma_{\text{test}} = 0.2$ (g) $\sigma_{\text{test}} = 0.5$ (h) $\sigma_{\text{test}} = 0.1$

Figure 2: **(a)Train loss curve, (b) natural ICL performance, (c-h) noisy ICL performance** comparison for models trained with different σ_{train}. 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 with noisy label training. We further conduct in-context inference with different inference Gaussian noise $\sigma_{test} \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1\}$ (the same as what we used to train our models). To ensure that the experimental results are convincing and reproducible, we repeated the experiment 10 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 during inference.

Figure 2(b-h) illustrates that with smaller inference noises $\sigma(\leq 0.2)$, model performance is relatively stable, indicating that training with label noise does not necessarily improve robustness compared to standard training. However, for larger noise levels $\sigma_{\text{test}} = 0.5, 1$, models trained with these noises slightly surpass the baselines in handling such noise, as depicted in Figure 2(g)(h). Specifically, in subfigure (g) for $\sigma_{test} = 0.5$, the model trained with $\sigma_{train} = 0.2, 0.5$ demonstrates significantly improved robustness over the baseline. Meanwhile, for $\sigma_{test} = 1$ in subfigure (h), the model trained with $\sigma_{train} = 0.5, 1$ shows markedly enhanced robustness compared to others, highlighting how incorporating label noise as data augmentation can bolster this robustness.

5 Conclusion and Future Work

This paper presents a comprehensive study on transformers' in-context learning with label noises, using simple function classes. Our experiments and analysis show that transformers retain their in-context learning efficiency amidst various types and levels of label noise. Furthermore, we discover that introducing similar noises into the training set enables good convergence and potentially enhances the robustness of transformer models. Our findings enhance understanding of transformers' adaptability to label noise, raising several intriguing research questions for further investigation. In

the next version, we intend to expand our experimental analysis to include additional tasks, model sizes, and input dimensions.

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

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the opportunities and risks of foundation models, 2022. 9

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 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*, pages 248–255, 2009. doi: 10.1109/CVPR.2009.5206848. 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, 3, 4, 9, 12

Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. Towards robustness to label noise in text classification via noise modeling. In *CIKM*. ACM, October 2021. doi: 10.1145/3459637.3482204. URL http://dx.doi.org/10.1145/3459637.3482204. 1, 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

Hao Huang, Ziyan Chen, Huanran Chen, Yongtao Wang, and Kevin Zhang. T-sea: Transfer-based self-ensemble attack on object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20514–20523, 2023. 9

Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey, 2021. 9

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. 9

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. URL https://api.semanticscholar.org/CorpusID:391854. 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. 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. 9

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, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein,

Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 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. 9

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, 2023a. 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. 9

Xindi Wang, Yufei Wang, Can Xu, Xiubo Geng, Bowen Zhang, Chongyang Tao, Frank Rudzicz, Robert E. Mercer, and Daxin Jiang. Investigating the learning behaviour of in-context learning: A comparison with supervised learning, 2023b. 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, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023b. 1, 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

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. 1, 9

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

A ADDITIONAL RELATED WORK

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. One popular thread on understanding this ability interprets the inference with in-context demonstrations as implicit gradient decent (Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023). 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 in-context learning through various perspectives, like Bayes inference (Xie et al., 2022) and PAC-learnability (Wies et al., 2023).

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. The closest work to ours is (Wang et al., 2023b), which compared the robustness of in-context learning and supervised learning with the text classification tasks, and found that in-context learning is more robust than supervised learning. In addition, by studying the ICL ability of transformers through noisy linear regression with standard Gaussian distribution, (Garg et al., 2023) also showed that transformers outperform toy baselines in this specific setting. In this work, we take steps further to systematically investigate the robustness of in-context learning against various label noise settings, including both the train and inference phases, and various types and magnitudes of noises.

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), which are typically referred to as the jailbreaking issue (Zou et al., 2023; Wei et al., 2023a; Dong et al., 2023b). Such risks further extend to the in-context learning scenario, as recent work (Wang et al., 2023a; Wei et al., 2023b) 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; Huang et al., 2023; Chen et al., 2023) among which adversarial training (AT) (Madry et al., 2017; Zhang et al., 2019) 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) 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 EVALUATION DETAILS

To evaluate the robustness of transformers, similar to Garg et al. (2023) we also compare its performance with a few simple baselines. These include (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}, and (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 offers 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, to assess model robustness, we employ the normalized squared error

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

. 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 ICL model maintains robustness and providing a criterion to compare between ICL and baseline models.

C More experiments results in Section 3

C.1 More visualization and numerical results

In this section, we provide more experiment results discussed in Section 3.

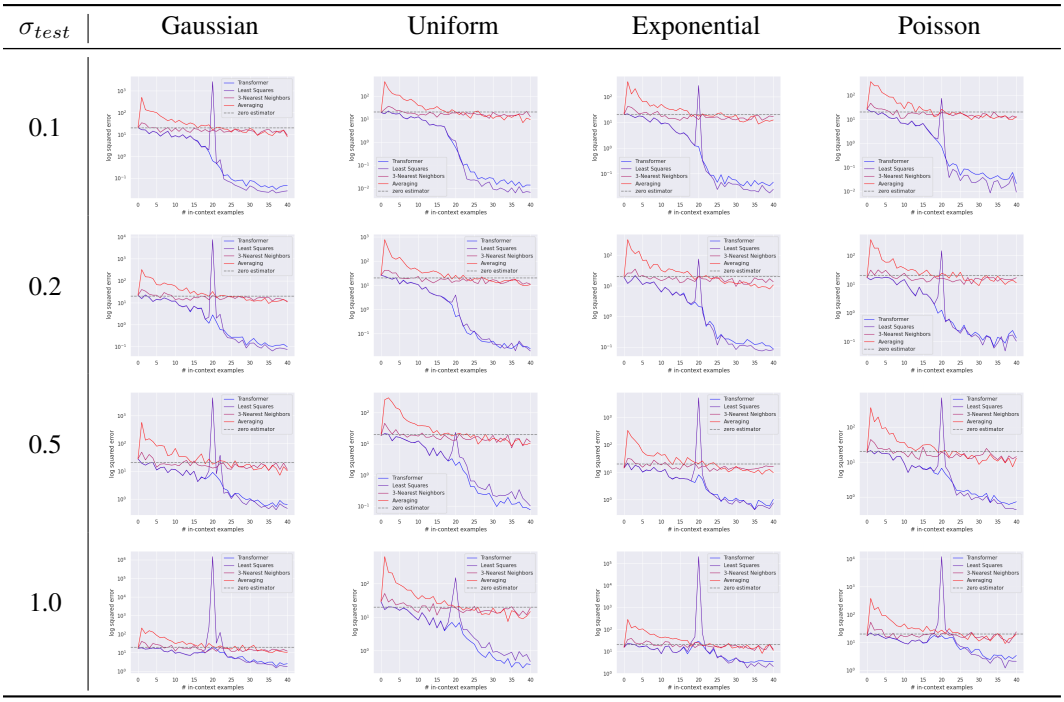


Figure 3: Complete robustness Comparison under different noise types and magnitudes.

Table 2: Average Squared Error with Different Number of ICL Examples. The results that show satisfactory accuracy are shown in **bold**.

# Examples	**Gaussian**					**Uniform**					**Exponential**					**Poisson**				
σ	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**1**	**1.1**	**1.2**	**1.3**	**1.4**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**
25	1.16	1.97	1.92	3.32	3.23	1.07	1.10	1.09	1.07	2.20	**0.42**	0.58	1.48	1.29	1.69	**0.42**	0.58	1.48	1.29	1.69
30	0.58	0.91	1.25	2.13	2.41	0.58	0.64	0.74	0.96	0.88	**0.14**	**0.20**	0.39	0.85	0.85	**0.14**	**0.20**	0.39	0.85	0.85
35	0.39	0.76	0.87	1.15	2.12	**0.39**	0.58	0.46	0.82	0.71	**0.12**	**0.27**	0.52	0.82	1.17	**0.12**	**0.27**	0.52	0.82	1.17
40	**0.39**	0.72	0.96	1.70	1.61	**0.36**	**0.35**	0.74	0.66	0.89	**0.13**	**0.17**	**0.48**	0.88	0.97	**0.13**	**0.17**	**0.48**	0.88	0.97

C.2 NON-I.I.D. NOISES

In this section, 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.

C.3 EXPERIMENT SET-UP

In this experiment, we 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 $P = (x_1, y_1, x_2, y_2, \cdots, x_k, y_k, x_{query})$.

We meticulously design the number of outliers m, since the efficacy of our results is based on having more in-context examples than outliers. When outliers otherwise outnumber in-context examples, they yield a different function, divergent from the one intended for testing on the query input.

C.4 EXPERIMENT RESULTS

We present the results with small multiples in Figure 4.Overall, the transformer model shows superior performance than other baselines, which is similar to the results of *i.i.d.* noises presented in Section 3 and shows its consistent robustness against such noises, which we elaborate on below.

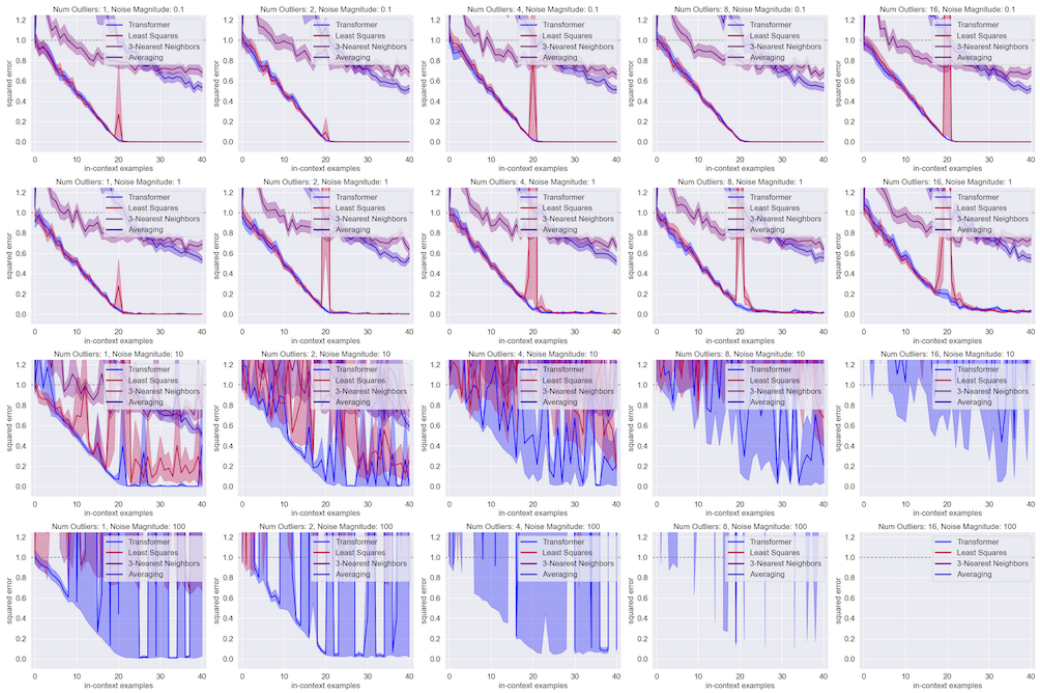


Figure 4: In-context learning on prompts with outliers. We evaluate the trained model on prompts with outliers in two cases. (1) On the rows, we explored the effect of the number of outliers on performance, and (2) on the columns, we explored the effect of the magnitude of outliers on performance.

Outliers of different magnitudes and equal quantity. Analyzing different columns in Figure 4 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 4 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. In addition, contrary to the findings in Garg et al. (2023), the error curve does not follow the pattern of the double descent error curve typically observed in ordinary least squares analysis. This suggests a different between the *i.i.d.* noisy label regression setting and the *non-i.i.d.* scenario.