On the Robustness of In-Context Learning with Noisy Labels: Train, Inference, and Beyond

Chen Cheng* **Haodong Wen*** **Xinzhi Yu*** **Zeming Wei**†

UC Berkeley

Abstract

Recently, the mysterious In-Context Learning (ICL) ability of Transformer architecture, particularly in large language models, has garnered considerable research interest for its underlying mechanisms and the development of more effective learning algorithms. Despite this, the resilience of Transformers' in-context learning capabilities in the face of noisy samples, both in training corpora and prompt demonstrations, remains underexplored. In this paper, we address this gap by examining the robustness of ICL Transformers with noisy labels, employing a linear function class for analysis. Our study extends beyond existing research, which predominantly focuses on label noise during the inference phase of ICL, by also considering the effects of such noise during the training phase. Furthermore, we explore this robustness in scenarios beyond *i.i.d.* label noises. Overall, our findings contribute novel insights into the robustness of Transformer models against label noises, laying the groundwork for further research in this area. Our code is available at `https://github.com/InezYu0928/in-context-learning-CS182`.

1 Introduction

In recent years, Large Language Models (LLMs) have achieved significant success across various tasks in real-world applications. Transformer [38], as the typical backbone architecture, also emerges an intriguing ability acknowledged as In-Context Learning (ICL) [5, 12] that the model can learn a new task with only a few input-output pairs demonstrated during inference *without* modifying any model parameters. Such mysterious property of transformers has attracted great research interests in understanding this ability from various perspectives like simple function classes [14], implicit gradient descent [39, 3].

However, as large language models like Vicuna [52] and Llama2 [36] 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 [46, 31] also showed that the output of ICL might be significantly manipulated by the label of demonstrations in the prompts. Motivated by these observations, in this work, we attempt to characterize and understand such ICL ability of transformers **with noisy labels** during both the training and inference phases in a toy model. Noisy label learning [29] has a wide breadth of literature in modern machine learning research [49, 43, 48], as dataset collection may be costly and noisy [11]. 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.

Though there is a series of works exploring the noisy label setting of language models [15, 53], few works have studied such problems systematically specified to in-context learning. In this work, unlike [14] that only consider one certain label noise distribution during the inference phase of ICL, we also study the influence of such noises during the training phase of transformers and show that the

[*] Equal contribution.

[†] Corresponding author. Email: zemingwei@berkeley.edu

Preprint. Under review.

training of transformers is fairly robust against such noises. Specifically, we find that adding label noises in the training set would not significantly affect their in-context inference ability. Besides, we also conduct systematic experiments in terms of various noise distributions and magnitudes to show that transformers are more robust against label noises than simple baselines, indicating their superior label-noise robustness during inference. In addition, we explore further problems including whether adding label noises in the demonstrations during training can enhance such robustness during the inference phase. Surprisingly, we find that adding moderate label noises in the demonstrations used in training, which can be regarded as a kind of data augmentation, **cannot** effectively enhance the robustness of transformers during inference. This result is contrary to existing viewpoints like adversarial training [26] where adding adversarial noises into training samples **can** enhance the model's robustness against such attacks [6]. This observation may correlated to the fundamental difference between the underlying mechanism of supervised learning and in-context learning. Finally, we extend our experiments to the scenarios of *non-i.i.d.* (independent and identically distribution) noises like anomaly outliers and show that the transformers are still robust in such noises.

Overall, our fruitful and comprehensive experiments provide interesting findings and conclusions for understanding the robustness of in-context learning with noisy labels. To summarize, we make the following contributions in this paper:

- We reveal that the in-context learning ability of transformers is fairly robust to the label noises in the training set in Section 4, which can not significantly impact its in-context learning performance.
- We conduct systematic experiments showing that transformers are robust to various types and magnitudes of label noises during in-context inference in Section 5. In addition, we reveal inducing the same noises in the training set cannot enhance such robustness effectively in Section 6, which is an unexpected phenomenon.
- We further extend experiments over *i.i.d.* label noises to the *non-i.i.d.* scenarios and show that transformers are still robust to such noise distributions in Section 7.

2 Related work

Understanding In-context Learning. The mysterious ability of in-context learning [5, 12], which typically occurs in attention-based model architectures like transformers [38], has attracted significant research interest in understanding its underlying mechanism and designing better learning algorithms [25, 28]. 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 [1, 39, 3]. 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 [50] and PAC-learnability [47].

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 [11]. Such noises in labels may be symmetric, asymmetric, and even from the open set that is not contained in the training classes [10]. To tackle this issue, numerous efforts have been made to robustify the training process against such noises. Typical approaches include estimating noise transition matrix [30], designing robust loss functions [27, 44], sample weighting [17, 43] and selection [21]. Though broadly explored in this literature, learning with noisy labels is still an open problem in modern machine learning research [35]. Moreover, there are also concurrent threads toward studying the robustness against label noise for the text modular [15, 53].

Despite broad explorations, few works have investigated the robustness of in-context learning with noisy labels. The closest work to ours is [42], 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, [14] 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 [4, 8, 23], which are typically referred to as the jailbreaking issue [54, 45, 13]. Such risks further extend to the in-context learning scenario, as recent work [40, 46, 32] 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 [16, 7, 24, 18, 9] among which adversarial training (AT) [26, 51] that adding adversarial noises in the training loops are shown to be the most effective among various defense methods [6, 2], it is recently shown that AT may not be an effective defense [20] for LLMs. Such difference further underscores the challenge of the interpretability of internal mechanism [19, 34] and adversarial examples [37, 41, 22] in the context of LLMs. In this work, we uncover a similar phenomenon: adding label noises in the training demonstrations may not enhance the corresponding robustness during inference, which further underscores the safety concerns in terms of label noises.

3 Preliminaries

Following an elementary research [14] that studies the in-context learning ability of transformers through a few simple function classes, we consider similar settings specified to the **noisy linear regression** task. We start with illustrating the vanilla training and inference pipeline without any label noises, and the specific settings for different experiments are detailed in the following corresponding sections.

Transformer training with in-context demonstrations. Given the class of linear functions $F = \{f_w | f_w(x) = w^T x, w \in \mathbb{R}^d\}$, each sample in the training set for the transformer can be formulated as $P = (x_1, f_w(x_1), \cdots, x_k, f_w(x_k))$, where the input samples are sampled from the isotropic Gaussian distribution $x_i \sim N(0, I_d)$, and the function parameter w is sampled from the isotropic Gaussian distribution $w \sim N(0, I_d)$. We adopt the GPT-2 [33] model architecture, which is the same as [14]. However, due to computational limitations, we apply the toy configuration provided in [14] with 5000 training epochs.[3]

Transformer inference with in-context demonstrations. After training the transformer model $M_\theta(\cdot)$ with its parameters θ by the training set above, the ICL inference of the model can be formulated as prompting $P = (x_1, f_w(x_1), \cdots, x_k, f_w(x_k), x_q)$, where $\{(x_i, f_w(x_i)\}$ are the in-context demonstrations and x_q is the query input. For the sake of evaluation, the ground truth function f_w is randomly sampled from F, and $\{x_i\}, x_q$ are still sampled from $N(0, I_d)$. Note that for better illustration and alignment with the original experiments conducted in [14], we apply the pre-trained model[4] to conduct all inference-only experiments (*i.e.*, standard training without label noises in the training set) in Section 5 and 7, which is larger than our trained toy models. We also tried to apply the same model size and training epochs but failed due to the computational limitations.

Note that in this paper, we may both consider the cases in which noises exist in the labels $f_w(x_i) + \epsilon$ for both train or inference phases, which are specified in the corresponding sections.

4 Train transformers with noisy labels

In this section, we explore the robustness of transformers against label noises during training for in-context learning ability.

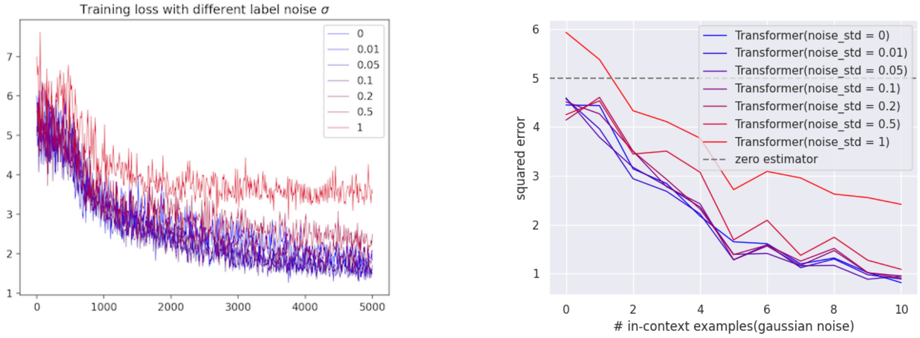
[3] The complete basic configurations for our experiments are listed at `https://github.com/InezYu0928/in-context-learning-1/blob/main/src/conf/toy.yaml`.

[4] Checkpoint available at `https://github.com/dtsip/in-context-learning/releases/download/initial/models.zip`.

4.1 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 $\epsilon_i \sim N(0, \sigma^2)$. In this experiment, we compare $\sigma \in \{0.01, 0.05, 0.1, 0.2, 0.5, 1\}$. 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. Note that in this set of experiments, the labels of in-context inference for evaluation are without noise.

4.2 Training convergence with label noises



(a) Train loss curve comparison. (b) In-context learning performance comparison.

Figure 1: Train loss curve and in-context learning performance comparison for models trained with different label noise magnitudes σ.

For each σ, we train the model with corresponding label noise distribution and plot the training loss curve in Figure 1(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. This shows that the training of transformers is fairly robust against noisy demonstrations.[5]

4.3 In-context learning performance evaluation

With these trained models, we further compare their in-context learning performance with the standard trained model, as shown in Figure 1(b). We also provide a clearer numerical comparison in Table 1.

Table 1: Final training loss and squared loss of 10 in-context demonstrations comparison for noisy label training with different σ.

σ	0	0.01	0.05	0.1	0.2	0.5	1
Final training loss	1.83	1.84	1.87	1.91	2.08	2.60	3.53
Squared error (w/10 ICL examples)	0.89	0.98	0.98	0.99	0.99	1.08	2.50

Similar to the observation above, the squared errors of in-context learning of the models trained over $\sigma < 0.5$ are almost the same as the standard one, showing again their robustness against such label noises in the training set. Besides, the model with $\sigma = 0.5$ performs comparable with the standard model, while the model with $\sigma = 1$ exhibits worse yet still has non-trivial in-context learning ability.

Overall, our findings in this section suggest that even though noises exist in the labels of the demonstrations, the transformers can still generalize well to perform in-context inference.

[5] Our trained checkpoints are available at https://drive.google.com/drive/folders/1-Z2-1JMQ8QjQIVaV0eJdDP1QtBxUpRec

5 In-context inference with noisy labels

We now turn our attention to the inference stage of in-context learning with different distributions of label noises, including Gaussian, Uniform, Exponential, and Poisson distributions. The overall experiment pipeline can be described as producing noise with different distributions, adding it to each demonstration in the prompts, and evaluating the pre-trained model's [14] performance.

5.1 Experiment set-up

Prompt design. In this study, we still 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. Similar to the construction of the noisy training set (Section 4), we sample $x_1, \cdots, x_k, x_{query}$ from the isotropic Gaussian distribution $N(0, I_d)$ and the *i.i.d.* noise $\epsilon_1, \cdots, \epsilon_k$ from various distributions with different noise magnitudes. Subsequently, each $y_i = w^T x_i + \epsilon_i$ is computed, and the prompt P is constructed as $P = (x_1, y_1, x_2, y_2, \cdots, x_k, y_k, x_{query})$.

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 $P(\cdot)$.

Baselines. To provide context for evaluating our trained model, we compare its performance with other simple learning algorithms, similar to [14]. 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.

Robustness Evaluation. 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.

5.2 Experiment results

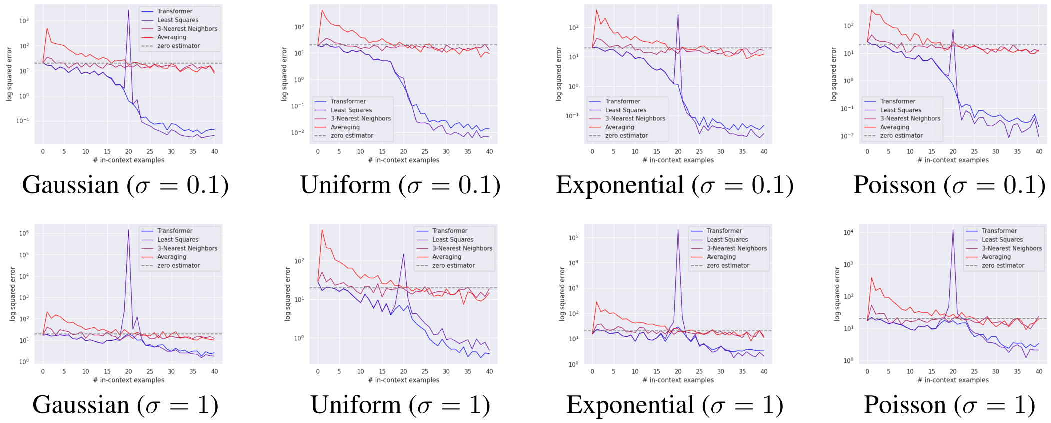
In this section, we employ the loss functions introduced above to evaluate the transformer model and baseline models' robustness across different noise types and varying levels of label noise.

Robustness threshold of label noise with different noise distribution. In Table 2, we report some examples of test accuracy under label noise, considering various distributions and standard deviations. We report the complete results over various numbers of in-context samples and noise distributions in Table 4 in Appendix A. In addition, we plot the representative comparison of the transformer and other baselines under these settings in Figure 2 and leave the complete figures for all σ in Figure 5 in Appendix A.

Table 2: Average Squared Error with Different Number of ICL Examples. The results that show satisfactory accuracy are shown in **bold**. Complete results are shown in Table 4.

# 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

Figure 2: Robustness Comparison under Different Noise Types and $\sigma = \{0.1, 1\}$. The X-axis represents the number of in-context examples. A complete comparison for all σ is shown in Figure 5.



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 ($\sigma = 1$)

Specifically, the robustness is examined across four distinct noise distributions at moderate levels. Notably, our finding indicates that under most error levels, loss functions associated with symmetrical distributions exhibit swifter learning in comparison to those with asymmetrical distributions. However, 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.

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 (detailed in Appendix A), we estimate such thresholds for different noise distributions and summarize them in Table 3.

Table 3: 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

5.3 Further observations

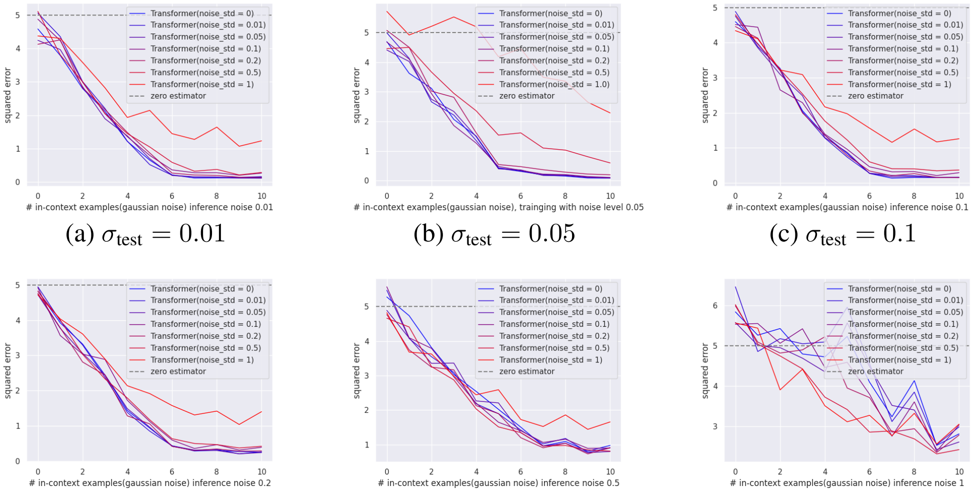
As shown in Figure 2, under the condition of label noise falling below a predefined threshold (except the manifest fact that the ICL model always outperforms 3-Nearest Neighbors and average method), we highlight several noteworthy observations:

Inadequate in-context examples. When the quantity of in-context examples is below the input dimension of $d = 20$, the loss of the trained model closely mirrors that of the least squares estimator.

Near input-dimension number of examples. As the number of in-context examples approaches the input dimension of $d = 20$, the least-square estimator exhibits significant errors, while the ICL model continues to improve its accuracy.

Sufficient in-context examples. As the number of in-context examples surpasses $d = 20$, the performance of the ICL model shows a rapid and notable improvement. Furthermore, in scenarios where label noise conforms to uniform distributions, the performance of the ICL model even surpasses that of the least-square estimator.

These nuanced findings contribute to a comprehensive understanding of the ICL's robustness under various conditions.



(a) $\sigma_{\text{test}} = 0.01$

(b) $\sigma_{\text{test}} = 0.05$

(c) $\sigma_{\text{test}} = 0.1$

(d) $\sigma_{\text{test}} = 0.2$

(e) $\sigma_{\text{test}} = 0.5$

(f) $\sigma_{\text{test}} = 1$

Figure 3: Noisy in-context learning performance comparison for models trained with different label noise magnitudes σ. Each figure represents an inference noise level, each line represents a model. The X-axis represents the number of in-context examples.

6 Noisy inference with noisy training

In this section, we study whether inducing some moderate label noises in the training set can enhance the robustness of transformers during in-context inference. Motivated by the fact that adversarial training methods [26, 51] which adds adversarial noises in the training sample can enhance the model's adversarial robustness [6, 2], we wonder if the robustness of transformer against in-context label noises can be improved by adding the same noises in the training demonstrations, which can be viewed as a type of data augmentation. To be more specific, we conduct ICL experiments on the linear function task with several inference noise levels for different noisy-trained models.

6.1 Experiment set-up

Similar to experiments in the sections above, we conduct in-context inference with different inference Gaussian noise $\sigma \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1\}$ (the same as what we used to train our model) using models separately trained in Section 4. 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. We construct prompts as what we do in section 5, *i.e.* adding Gaussian noise into the output of each demonstration pair, and the noise levels are selected in $\sigma \in \{0, 0.01, 0.05, 0.1, 0.2, 0.5, 1\}$.

6.2 Experiment results

In this experiment, we conduct in-context inference with label noises across different training models. The results are shown in Figure 3, where we observe that for smaller inference noises $\sigma (\leq 0.2)$, the model's performance remains relatively consistent, *i.e.* the models trained with label noises cannot enhance such robustness over standard training. The only exception is for large noises $\sigma_{\text{test}} = \{0.5, 1\}$, where the models trained within these noise levels can slightly outperform the baselines in terms of these large noises (Figure 3(e)(f)). However, as discussed in Section 4, training with such large noise levels cannot converge well and decreases the ICL ability for the clean prompts. Furthermore, as shown in other sub-figures, these two models do not exhibit better robustness in terms of other noises $(\sigma_{\text{test}} \leq 0.2)$. Therefore, we suggest that adding extremely large noises to enhance the model's robustness against label noises is not a practical solution, and adding moderate noises in the training set cannot effectively improve such robustness.

6.3 Discussion

As discussed above, a counter-intuitive finding of our study is that introducing moderate label noises in the training demonstrations **can not** effectively enhance the robustness of transformers during in-context inference. This outcome appears to be quite contradictory to existing viewpoints. For instance, in basic computer vision tasks, adding noise to training images (such as adversarial perturbations) typically results in improved robustness against similar noise distributions during inference. Contrarily, in our study, incorporating the same Gaussian noise distribution into input data does not yield increased robustness, and in some cases, even diminishes it.

This unexpected result may be attributed to the fundamental differences between the mechanisms underlying supervised learning and ICL. In supervised learning, the objective is to directly optimize performance on tasks using the data itself, such as in classification and regression tasks. In contrast, ICL demands more complex reasoning about underlying rules within the data, often with minimal guidance. For instance, ICL involves identifying mapping classes, as explored in this research. In such scenarios, noise injection as data augmentation may effectively improve such robustness. Instead, it could act as a disruption, obscuring the underlying rules and broadening the problem domain, potentially leading to confusion in the model's training process.

7 Extrapolating beyond *i.i.d.* noises

Finally, 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.

7.1 Experiment set-up

In this experiment, we focus on the class of noisy linear functions $F = \{f_w | f_w(x) = w^T x + \nVdash_{outlier} * \epsilon, w \in R^d\}$ in $d = 20$ dimensions, where noises are only injected into demonstrations of outliers. Similar to the construction of the noisy training set (Section 4), 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 $y_i = w^T x_i + \nVdash_{i \in S} * \epsilon_i$ 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.

7.2 Experiment results

We present the results with small multiples in Figure 4 (detailed in Appendix B). Overall, the transformer model shows superior performance than other baselines, which is similar to the results of *i.i.d.* noises presented in Section 5 and shows its consistent robustness against such noises, which we elaborate on below.

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 [14], 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.

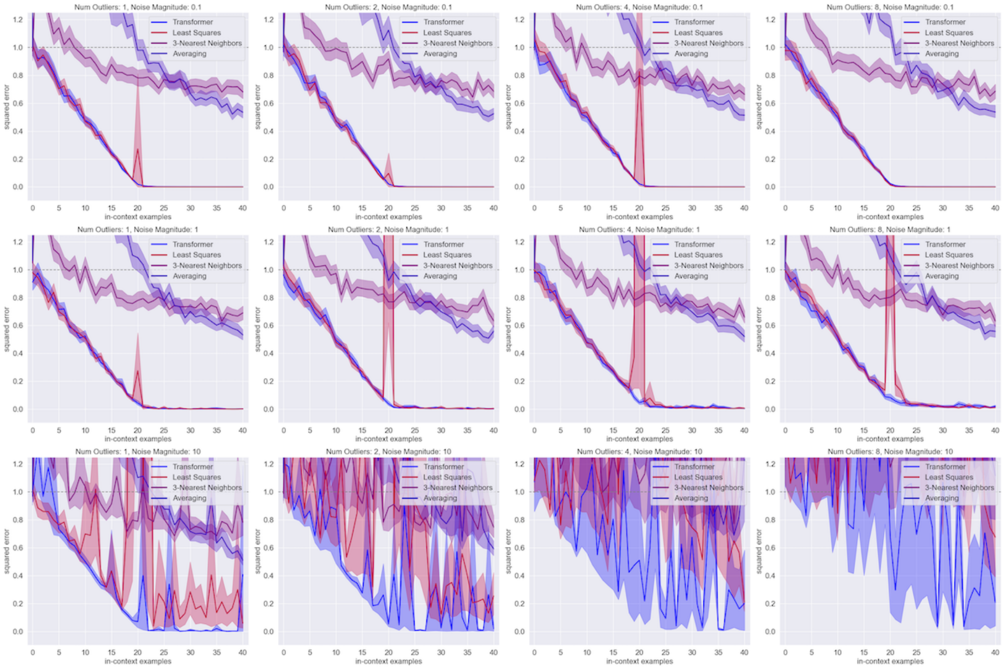


Figure 4: In-context learning on prompts with outliers. We evaluate the trained model on prompts with outliers in two cases. (1) On the rows, we explored the effect of the number of outliers on performance, and (2) on the columns, we explored the effect of the magnitude of outliers on performance.

8 Further discussions and limitations

This paper uncovers some intriguing properties of transformers in terms of the intersection of in-context learning and label noises. Our distilled findings reveal a dual nature. On the one hand, Transformers demonstrate commendable robustness during both training and inference phases against various types of label noises. On the other hand, our study indicates a pessimistic aspect where introducing label noise as a data augmentation strategy during training does not necessarily enhance inference robustness. These observations uncover the fundamental particular property of in-context learning with transformers, warranting further in-depth theoretical and empirical exploration. However, we acknowledge several limitations of this work. First, The experiments were confined to linear regression tasks, which is relatively simple. Moreover, incorporating a theoretical analysis would provide a stronger foundation for our findings on in-context learning robustness. Additionally, further trials and methodologies to enhance transformer robustness could be encompassed, particularly by examining the impact of expanded model sizes.

9 Conclusion

In this paper, we thoroughly investigate the robustness of transformers' in-context learning capability when confronted with noisy labels across a spectrum of scenarios. Our study reveals that transformers display a significant degree of robustness against label noise within training datasets, thereby maintaining their in-context learning efficiency. Furthermore, through systematic experiments detailed in the previous section, we demonstrate that transformer models consistently uphold their robustness across a variety of label noise types and magnitudes during in-context inference. Intriguingly, our findings also uncover that introducing similar noises into the training set does not enhance this robustness, thereby presenting a remarkable observation. Finally, we extend the scope of our research by examining scenarios that deviate from the *i.i.d.* label noise paradigm. The outcomes in non-*i.i.d.* scenarios indicate that Transformer models consistently persist in demonstrating robustness amidst such varied noise distributions. This comprehensive investigation significantly contributes to a nuanced understanding of Transformer models' capacity to navigate and adapt to label noise, underscoring their robustness in diverse operational contexts.

References

[1] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. 2

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018. 3, 7

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023. 1, 2

[4] 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. 3

[5] 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, 2

[6] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods, 2017. 2, 3, 7

[7] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks, 2017. 3

[8] Canyu Chen and Kai Shu. Combating misinformation in the age of llms: Opportunities and challenges. *arXiv preprint arXiv:2311.05656*, 2023. 3

[9] Huanran Chen, Yichi Zhang, Yinpeng Dong, and Jun Zhu. Rethinking model ensemble in transfer-based adversarial attacks. *arXiv preprint arXiv:2303.09105*, 2023. 3

[10] 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. 2

[11] 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. 1, 2

[12] 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, 2023. 1, 2

[13] 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*, 2023. 3

[14] 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, 5, 8

[15] Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. Towards robustness to label noise in text classification via noise modeling. In *CIKM*. ACM, October 2021. 1, 2

[16] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 3

[17] 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. 2

[18] 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. 3

[19] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey, 2021. 3

[20] 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. 3

[21] 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. 2

[22] Ang Li, Yifei Wang, Yiwen Guo, and Yisen Wang. Adversarial examples are not real features. *arXiv preprint arXiv:2310.18936*, 2023. 3

[23] 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. 3

[24] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack, 2022. 3

[25] Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. Are emergent abilities in large language models just in-context learning?, 2023. 2

[26] 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. 2, 3, 7

[27] Naresh Manwani and Senior Member Ieee P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43:1146–1151, 2011. 2

[28] 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. 2

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

[30] Giorgio Patrini, Alessandro Rozza, Aditya Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: a loss correction approach, 2017. 2

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

[32] Yao Qiang, Xiangyu Zhou, and Dongxiao Zhu. Hijacking large language models via adversarial in-context learning, 2023. 3

[33] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019. 3

[34] 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. 3

[35] 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. 2

[36] 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

[37] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018. 3

[38] 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, 2

[39] 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. 1, 2

[40] Jiongxiao Wang, Zichen Liu, Keun Hee Park, Muhao Chen, and Chaowei Xiao. Adversarial demonstration attacks on large language models, 2023. 3

[41] 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. 3

[42] 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, 2023. 2

[43] 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, 2

[44] 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. 2

[45] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail?, 2023. 3

[46] Zeming Wei, Yifei Wang, and Yisen Wang. Jailbreak and guard aligned language models with only few in-context demonstrations. *arXiv preprint arXiv:2310.06387*, 2023. 1, 3

[47] Noam Wies, Yoav Levine, and Amnon Shashua. The learnability of in-context learning, 2023. 2

[48] 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

[49] 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

[50] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022. 2

[51] 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. 3, 7

[52] 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

[53] 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, 2

[54] Andy Zou, Zifan Wang, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models, 2023. 3

A Complete results in Section 5

We provide detailed complementary experiment results in the following Table 4 and Figure 5.

Figure 5: Complete Robustness Comparison under Different Noise Types for Figure 2

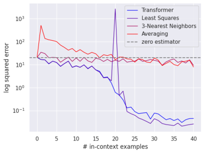
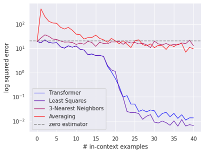
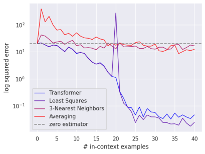
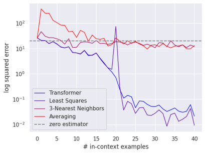
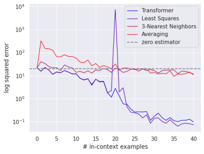
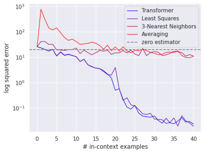
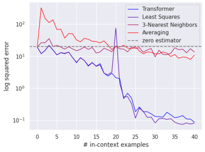
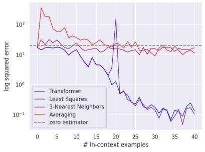
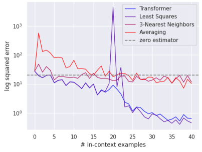
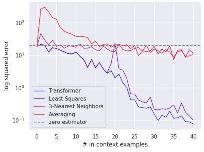
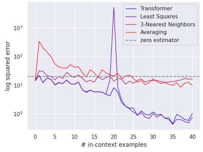
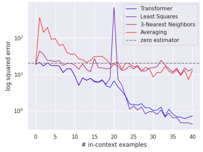
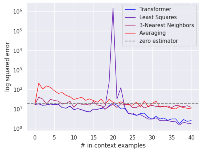
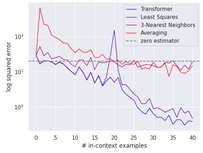
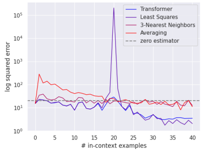
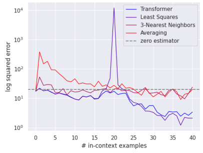
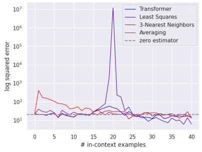
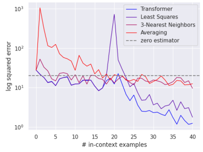
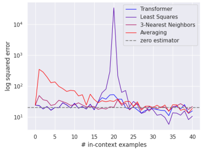
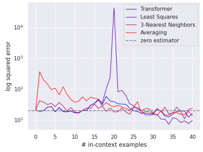
type σ	Gaussian	Uniform	Exponential	Poisson
0.1				
0.2				
0.5				
1.0				
2.0				

Table 4: Complete results for Table 2

Type		Gaussian					Uniform					Exponential					Poisson				
Std		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
# of Examples	0	22.65	23.45	25.26	22.11	24.19	22.13	21.79	15.47	20.45	20.96	19.12	25.17	19.77	20.18	12.76	19.12	25.17	19.77	20.18	12.76
	1	18.02	19.97	20.11	18.47	18.37	22.37	20.38	15.97	20.78	15.70	15.18	20.24	18.87	22.61	24.89	15.18	20.24	18.87	22.61	24.89
	2	16.05	18.21	13.91	19.05	12.22	16.02	22.82	21.59	16.92	19.73	17.05	24.02	19.62	21.79	15.70	17.05	24.02	19.62	21.79	15.70
	3	12.38	14.98	23.37	24.02	17.73	12.41	15.25	17.87	22.16	26.48	15.85	11.22	16.79	18.97	14.66	15.85	11.22	16.79	18.97	14.66
	4	13.31	13.95	14.65	12.44	15.02	15.96	17.58	23.04	26.61	19.72	18.76	18.69	16.16	15.09	24.15	18.76	18.69	16.16	15.09	24.15
	5	8.77	11.16	12.94	9.92	10.68	13.72	16.69	14.95	14.97	13.08	14.85	16.74	18.03	15.34	17.45	14.85	16.74	18.03	15.34	17.45
	6	16.55	13.31	11.76	12.04	9.93	19.49	14.83	13.10	14.70	18.94	12.98	10.48	13.09	14.82	16.47	12.98	10.48	13.09	14.82	16.47
	7	14.89	15.52	12.18	15.26	14.50	8.06	18.81	16.87	14.01	13.74	17.52	14.85	19.81	16.84	16.92	17.52	14.85	19.81	16.84	16.92
	8	12.45	12.24	11.01	12.49	15.63	13.67	8.19	19.63	12.16	13.23	9.91	11.23	12.77	13.27	11.37	9.91	11.23	12.77	13.27	11.37
	9	12.65	12.49	11.15	12.79	13.93	13.91	11.23	12.20	11.55	17.98	13.22	12.38	7.27	12.94	13.64	13.22	12.38	7.27	12.94	13.64
	10	12.68	9.51	10.69	11.00	12.74	13.52	13.93	11.93	10.73	11.31	12.04	10.06	12.16	8.76	8.96	12.04	10.06	12.16	8.76	8.96
	11	9.58	12.73	12.77	12.15	13.63	8.47	8.39	15.66	10.94	10.46	11.26	11.82	7.75	8.46	12.83	11.26	11.82	7.75	8.46	12.83
	12	7.90	6.21	14.45	9.84	12.28	7.47	8.94	9.48	11.31	8.07	10.57	6.55	6.81	9.33	11.38	10.57	6.55	6.81	9.33	11.38
	13	9.72	7.27	6.77	7.09	10.99	7.59	5.56	6.68	6.78	7.34	8.06	7.61	7.03	8.35	7.92	8.06	7.61	7.03	8.35	7.92
	14	6.61	7.34	8.27	5.15	7.24	8.09	6.33	7.53	7.69	6.79	5.98	8.47	4.96	5.99	5.32	5.98	8.47	4.96	5.99	5.32
	15	4.94	5.62	6.17	5.26	9.28	5.72	7.96	8.66	8.20	5.31	5.68	5.24	5.56	4.73	5.11	5.68	5.24	5.56	4.73	5.11
	16	4.22	5.25	4.68	5.87	8.78	4.81	5.14	7.79	7.90	9.85	3.20	3.75	5.49	7.25	5.92	3.20	3.75	5.49	7.25	5.92
	17	3.76	6.25	4.51	8.98	7.89	6.14	6.13	5.87	6.34	9.07	3.34	3.13	5.12	5.85	3.73	3.34	3.13	5.12	5.85	3.73
	18	4.04	4.51	5.01	7.73	10.31	5.58	8.00	6.24	7.56	8.50	3.14	4.03	4.04	7.48	4.51	3.14	4.03	4.04	7.48	4.51
	19	4.08	4.82	6.62	14.81	17.94	7.39	9.30	6.87	7.39	12.60	2.09	2.54	3.56	7.61	4.88	2.09	2.54	3.56	7.61	4.88
	20	6.27	7.25	6.66	10.17	11.42	10.43	10.69	10.29	14.43	15.20	1.37	2.30	4.71	4.65	7.61	1.37	2.30	4.71	4.65	7.61
	21	3.17	6.03	7.60	9.99	13.80	5.47	6.06	6.53	6.68	8.03	0.56	0.90	2.16	3.01	5.24	0.56	0.90	2.16	3.01	5.24
	22	2.06	3.18	6.28	8.90	8.89	2.62	3.31	3.58	3.89	4.33	0.60	1.27	4.62	6.87	5.34	0.60	1.27	4.62	6.87	5.34
	23	3.13	3.46	4.73	4.86	6.86	2.96	3.98	3.15	4.70	4.07	1.28	1.22	3.94	2.62	3.99	1.28	1.22	3.94	2.62	3.99
	24	1.64	1.54	2.57	5.67	4.52	1.25	1.55	2.77	2.76	2.88	**0.24**	1.08	1.40	1.27	3.08	**0.24**	1.08	1.40	1.27	3.08
	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
	26	0.69	1.65	1.94	1.90	2.99	1.22	1.30	1.52	2.12	2.02	**0.23**	**0.44**	1.77	1.16	1.86	**0.23**	**0.44**	1.77	1.16	1.86
	27	0.86	1.48	1.84	2.40	2.29	0.75	1.31	2.22	2.12	1.98	**0.23**	**0.28**	0.74	0.92	1.53	**0.23**	**0.28**	0.74	0.92	1.53
	28	0.63	1.56	2.06	2.27	2.67	0.90	1.01	1.24	1.08	1.18	**0.25**	**0.45**	0.84	1.14	1.57	**0.25**	**0.45**	0.84	1.14	1.57
	29	0.68	1.21	1.33	1.79	2.50	0.76	0.77	0.74	0.82	0.89	**0.19**	**0.36**	1.00	0.81	1.81	**0.19**	**0.36**	1.00	0.81	1.81
	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
	31	0.48	0.73	1.21	1.53	2.16	0.48	0.77	0.95	0.66	1.29	**0.13**	**0.31**	0.55	0.96	0.82	**0.13**	**0.31**	0.55	0.96	0.82
	32	0.72	1.23	1.17	1.73	2.30	0.52	0.78	0.58	0.70	0.99	**0.14**	**0.27**	0.66	0.75	1.17	**0.14**	**0.27**	0.66	0.75	1.17
	33	0.43	0.62	1.15	1.13	2.14	**0.42**	0.49	0.85	1.02	1.10	**0.13**	**0.36**	0.70	0.79	1.27	**0.13**	**0.36**	0.70	0.79	1.27
	34	0.58	0.77	1.58	2.10	1.20	**0.42**	0.35	0.55	0.73	0.75	**0.09**	**0.20**	0.60	0.76	1.25	**0.09**	**0.20**	0.60	0.76	1.25
	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
	36	0.53	0.74	1.66	1.70	1.90	**0.48**	0.50	0.59	0.61	0.71	**0.16**	**0.18**	0.66	0.65	0.86	**0.16**	**0.18**	0.66	0.65	0.86
	37	**0.46**	0.78	1.04	1.67	1.33	**0.42**	0.45	0.67	0.80	0.88	**0.10**	**0.16**	**0.33**	0.79	0.98	**0.10**	**0.16**	**0.33**	0.79	0.98
	38	**0.46**	0.57	0.69	1.57	1.78	**0.35**	0.51	0.43	0.72	0.67	**0.10**	**0.26**	**0.40**	0.66	0.74	**0.10**	**0.26**	**0.40**	0.66	0.74
	39	**0.41**	0.93	1.02	1.49	1.79	**0.32**	0.52	0.60	0.63	0.66	**0.11**	**0.17**	**0.29**	0.83	0.88	**0.11**	**0.17**	**0.29**	0.83	0.88
	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

B Complete results in Section 7

We provide detailed complementary experiment results in the following Figure 6.

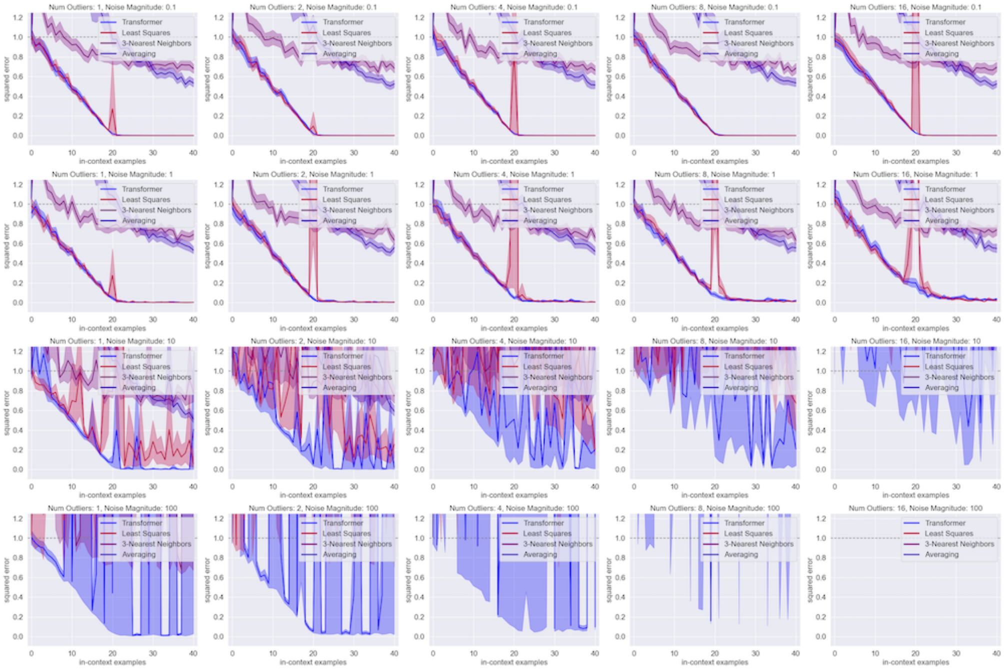


Figure 6: Complete experiment results of in-context learning on prompts with outliers. We evaluate the trained model on prompts with outliers in two cases. (1) On the rows, we explored the effect of the number of outliers on performance, and (2) on the columns, we explored the effect of the magnitude of outliers on performance.