Adapting to Context: A Case Study on In-Context Learning of Decision Tree Algorithms by Large Language Models

Abdullah Azhar
UC Berkeley / South Hall, Berkeley CA, 94704
abdullah_azhar@berkeley.edu

Abstract

The advent of large language models such as GPT-3 has introduced in-context learning to natural language processing. This method leverages a sequence of task-specific input-output examples alongside a new query to elicit a response from the model during inference without modifying its parameters. Despite being a nascent field, it is underexplored, raising questions about the models' reliance on inductive reasoning versus their extensive pretraining datasets. Building on recent publications that explore the in-context learning capability of transformers for linear and non-linear functions without noisy training, this project extends these studies by training transformer architectures from scratch on non-linear functions derived from decision tree classes, incorporating distribution shifts during both training and inference. The objective is to assess whether transformers can effectively learn and generalize across these distribution shifts, thereby enhancing our understanding of how in-context learning functions without the need for fine-tuning. Inference testing included standard prompting, random quadrant shifts, and overlapping train-test prompts, along with training on varying levels of Gaussian noise. Although model performance degrades with increased noise and out-of-distribution shifts as measured by root-mean-squared error, transformer performance remains comparable, sometimes exceeding that of conventional decision tree algorithms like XGBoost. These results are promising, highlighting the model's robustness in learning beyond simple linear class functions. The full code repository with model checkpoints is available at https://github.com/azhara001/in-context-learning-trees/tree/main.

1 Introduction

In-context learning in large language models (LLMs) represents a significant shift in how these models leverage existing knowledge to generate outputs. Originating from the insights presented in (Brown et al., 2020), in-context learning enables a model to generate responses based on a sequence of input-output examples provided directly in the prompt, without any gradient updates or retraining. This approach allows LLMs like GPT-3 to perform tasks based solely on the context of the prompt.

However, the extent to which these models truly 'learn' new tasks from in-context examples remains an open question. Critics, such as (Xie et al., 2022), suggest that the models might not be learning in a conventional sense but are instead effectively retrieving information related to known tasks embedded in their training data (latent space). This perspective is supported by observations made in recent studies, such as (Min et al., 2022), which propose that LLMs may primarily index a vast set of tasks acquired during their initial training.

Efforts to understand and demystify the mechanisms of in-context learning have led to targeted studies exploring the boundaries of this capability. One such investigation, detailed in (Garg et al., 2023), examined the ability of transformer models to adapt and generalize across in-context learning from a set of function classes including linear functions, sparse linear functions, and other complex functions like decision trees and 2-layer neural networks.

This project will delve into the capabilities of transformer models to engage in in-context learning with complex non-linear functions, specifically focusing on decision trees label noise adjustments during inference and training. Building on foundational studies such as (Garg et al., 2023), the investigation will aim to enhance our understanding of how these advanced models adapt to different types of data and assess their ability to encode effective learning algorithms under varied conditions.

1.0.1 Benchmarking with Decision Trees

Following the procedures outlined in (Garg et al., 2023), this study will involve training a decision tree using in-context learning under a fixed configuration with a dimension $d = 8$ and a tree depth of 4. This setup will serve as a benchmark case, providing a baseline for evaluating the performance of transformer models in their ability to learn in-context when faced with non-linearly separable data. While (Garg et al., 2023) performed training with $d = 20$ and 101 in-context examples, we have used dimension size $d = 8$ with 40 in-context examples owing to compute constraints.

A key objective of this project is to empirically demonstrate that standard transformer models, trained from scratch, can effectively learn and apply the principles of in-context learning to non-linear functions, particularly those represented by decision trees. The models will be trained using an input distribution D_X, characterized by an i.i.d isotropic Gaussian in 8 dimensions, and D_F, the distribution over non-linear functions with varying gaussian label variance. Success will be measured by the model's ability to achieve root-mean-squared error rates comparable to those trained by decision tree regressors such as the XGBOOST model.

1.1 Generalization to Out-of-Distribution Prompts

This research further aims to explore how well the trained model can generalize its learning capabilities to out-of-distribution prompts, a crucial aspect of real-world applicability in assessing the robustness of the model. The study will analyze the model's performance against two primary types of distribution shifts:

- Within distribution at inference time regarding the in-context labeled examples and x_{query}.
- Overlapping in-context and x_{query} distribution at inference time.
- Random quadrant distribution between in-context labeled examples and x_{query} distribution.

These prompting strategies were observed for both pre-trained model checkpoints from (Garg et al., 2023) as well as for the noisy labeled training done in-house for this project. Due to limited compute power, we performed noisy labeled training for $\sigma = 0, 1,$ and 3.

2 Related Work

Transformers have revolutionized the field of natural language processing and have become foundational to the development of large language models (LLMs). Introduced by Vaswani et al. in their seminal work (Vaswani et al., 2023), the transformer architecture has facilitated significant advances in model performance across a range of NLP tasks due to its efficient handling of sequence-to-sequence tasks and its scalability. The architecture leverages self-attention mechanisms that directly compute interactions between all tokens in a sequence, regardless of their positions, enabling more dynamic representations and understanding.

In-context learning, a capability prominently featured in models like GPT-3, represents a burgeoning area within machine learning. This approach involves the model using provided examples within a prompt to generate responses for new, similar tasks without any gradient updates or retraining, showcasing an impressive degree of flexibility and applicability. Brown et al. (Brown et al., 2020) highlighted the potential of in-context learning to handle a diverse array of tasks using a single model, suggesting a shift towards more adaptive and efficient learning paradigms in artificial intelligence.

Despite its emerging success, the nature of what models learn through in-context examples and whether they genuinely acquire the ability to perform inductive reasoning remains under scrutiny. Xie et al. (Xie et al., 2022) introduced the concept of implicit Bayesian inference in the latent concept space as a potential explanation for the success of in-context learning. This theory posits that models may be implicitly averaging over learned representations by the concept of learning a latent concept space, allowing them to handle new instances more effectively. Similarly, Razzleghi et al. (**?**) explored how the frequency and nature of training data impact the performance of models on tasks like simple arithmetic, suggesting that exposure to certain types of data can significantly influence model behavior and capabilities, thereby raising additional concerns around the capabilities of models to learn in-context.

Further investigations by Rong et al. (Rong, 2021) and Liu et al. (Liu et al., 2021) have examined the limits of what transformers can learn in-context, particularly when faced with tasks that require extrapolation beyond the training data. These studies reveal that while transformers exhibit a

high degree of proficiency within their training distribution, their performance can degrade when tasked with out-of-distribution examples. Song et al. (Song et al., 2022) expanded on this by demonstrating the potential for transformers to learn complex function mappings in-context, indicating that the models might not only be recalling information but also developing a form of procedural memory.

Olsson et al. (Olsson et al., 2022) provide a comprehensive review of in-context learning, discussing its implications for the future of machine learning models and their deployment in real-world applications. Their work underscores the importance of understanding the mechanisms behind in-context learning to leverage its full potential effectively.

These pieces of literature collectively underscore the dynamic nature of in-context learning and its potential to redefine the boundaries of machine learning. As this field continues to evolve, it is crucial to understand both the capabilities and limitations of models trained under this paradigm to harness their potential responsibly and effectively.

As evident, work is being conducted within the space of in-context learning, including experimentation with methods such as few-shot and chain of thought reasoning (without gradient updates) (Wei et al., 2023), as well as through approaches focused on pretraining (Peng et al., 2023). However, research into the robustness of these models against noisy label inferences remains comparatively sparse. Studies like Garg (2021) (Garg et al., 2021) explore handling label noise within neural networks, yet few have bridged this with in-context learning strategies (Cheng et al., 2024). This gap at the intersection of handling noisy data in in-context learning frameworks forms the primary motivation behind projects like (Garg et al., 2023). A recent preprint has begun to address this niche, suggesting the onset of a new era in the study of robust in-context learning (Cheng et al., 2023). This emerging line of research promises to deepen our understanding of how large language models can maintain performance even when trained or queried with imperfect data, thereby enhancing their practicality for real-world applications.

3 Methodology and Training

3.1 Dataset Generation

The dataset for this study mirrors the structure used in (Garg et al., 2023), where the function f applied to an input x is evaluated by traversing a decision tree starting from the root node. Navigation through the tree involves moving to the right child if the coordinate of the current node is positive and to the left child if negative, with each node's threshold set to zero. The function value $f(x)$ corresponds to the value at the leaf node reached at the end of this traversal. To generate a random prompt $P = (x_1, f(x_1), \ldots, x_k, f(x_k), x_{\text{query}})$, prompt inputs x_{is} and x_{query} are drawn from a normal distribution $N(0, I_d)$, while f is determined by a tree whose non-leaf node coordinates are uniformly selected at random from the set $\{1, 2, \ldots, \text{d}\}$ and leaf node values are drawn from $N(0, 1)$. The training data was drawn randomly for each batch, and for this study, we followed a similar paradigm to (Garg et al., 2023) with a batch size of 64. Depending on the prompting strategy—standard, train-test overlap, and random quadrant—the data and labels were randomly drawn for each batch, and the model was then trained. Additionally, given computational constraints, a GPT-2 model with 22 million parameters was employed for training and inference tasks. This choice was driven by practical limitations on available computing resources, making it impractical to utilize larger models while still aiming to achieve robust and insightful results. Moreover, given the compute constraints, noisy label training was conducted on a single NVIDIA L4 Tensor Core GPU with 40 in-context examples. However, for noise-free training, we used pretrained model checkpoints as those models were trained over 500,001 steps. For this study, we trained the models over 10,001 steps. For both configurations, the same three prompting strategies were used.

Evaluation Strategy The primary evaluation metric used is the root-mean-squared error (RMSE), a standard metric for assessing accuracy. This project does not employ a tokenizer, allowing us to leverage traditional evaluation metrics effectively. We established benchmarks for the final model results as done by (Garg et al., 2023), evaluating the GPT-2 model alongside "Least Squares," "3-Nearest Neighbors," "Greedy Tree Learning," "XGBoost," and "Averaging" models. The overall goal was to assess the model's performance across different modeling approaches, given our knowledge of the actual data distribution. Our comparative analysis extended across three configurations of noisy labels to understand the model's resilience.

Notably, in scenarios involving noise-free prompting and training, our model's performance proved comparable to that of XGBoost. This study also explores how adjustments in the number of in-context examples and input data dimensions affect model performance, particularly using a non-gradient update approach through few-shot prompting.

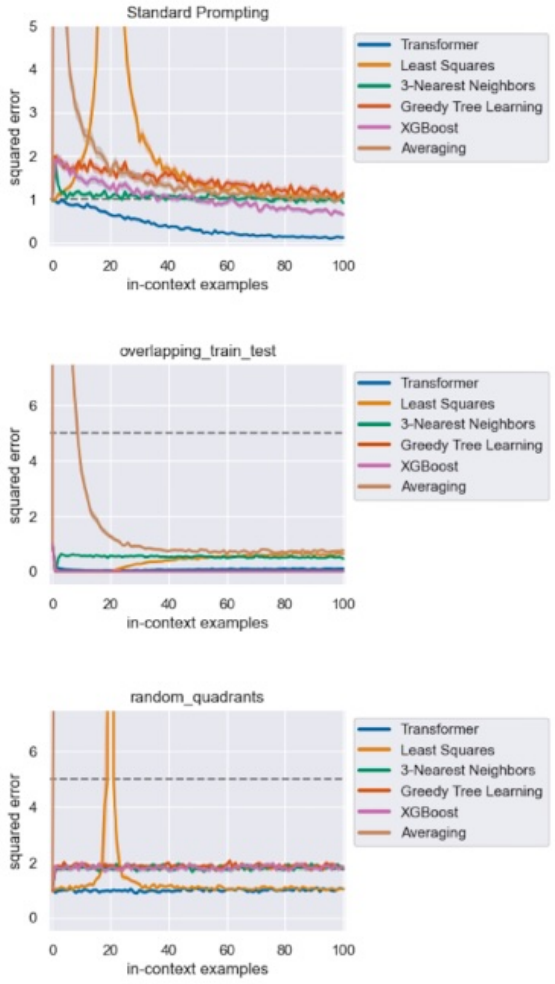

Figure 1: Model Performance on noise-free training

4 Results

We initiated our investigation by analyzing the model's performance on noise-free training, employing original checkpoints from the (Garg et al., 2023) paper. As depicted in Figure 2, we evaluated the performance under various prompting strategies. Initially, we found that our model's performance either matched or surpassed that of XGBOOST during standard prompting scenarios, as emphasized by (Garg et al., 2023). This indicates a robust baseline capability in standard operational settings.

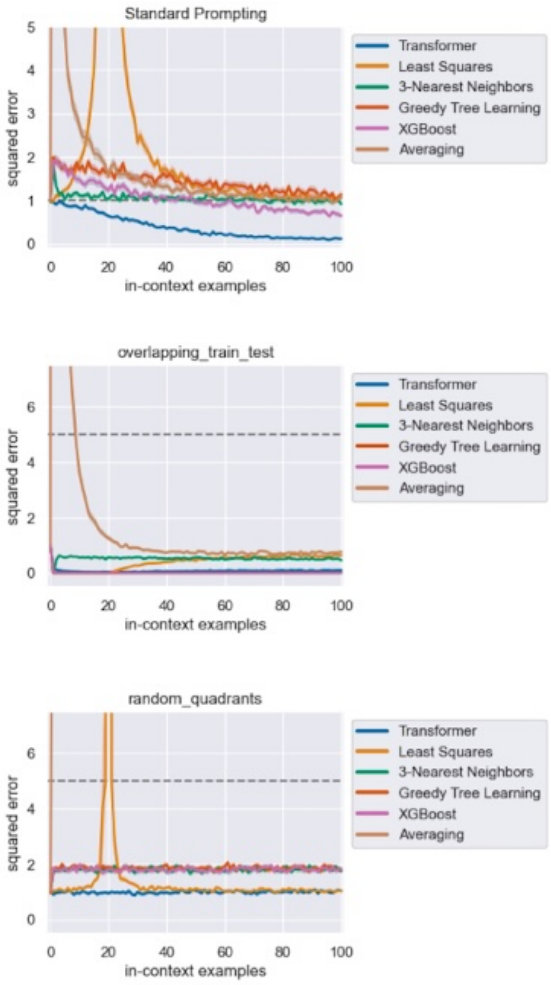

Figure 2: Model Performance on noise-free training

When examining the performance under overlapping distributions, all baseline models showed a near-zero RMSE, suggesting a high degree of accuracy and efficiency comparable to the XGBOOST model. This was a significant finding, reinforcing the model's reliability in consistent or predictable environments.

However, our analysis took an intriguing turn when we considered scenarios involving random quadrant distributions. Here, despite increasing the number of in-context examples—a strategy typically beneficial in enhancing model performance—the improvement plateaued, indicating that our model's adaptability might be limited by the randomness and unpredictability of the data distribution. This finding underscores a critical limitation: while increasing in-context examples generally improves performance, it is not univer-

sally effective across all distribution types.

Main Findings: Within the same distribution settings, the incremental addition of in-context examples notably improves model performance. Conversely, in random quadrants, the performance remains relatively stable regardless of the increase in in-context examples, suggesting a saturation point beyond which additional context does not equate to better performance.

Next, our study progressed to include models trained under noisy conditions. As illustrated in Figure 3, the performance was similar to that of the baseline models under noise-free conditions, signifying robustness against standard noise levels.

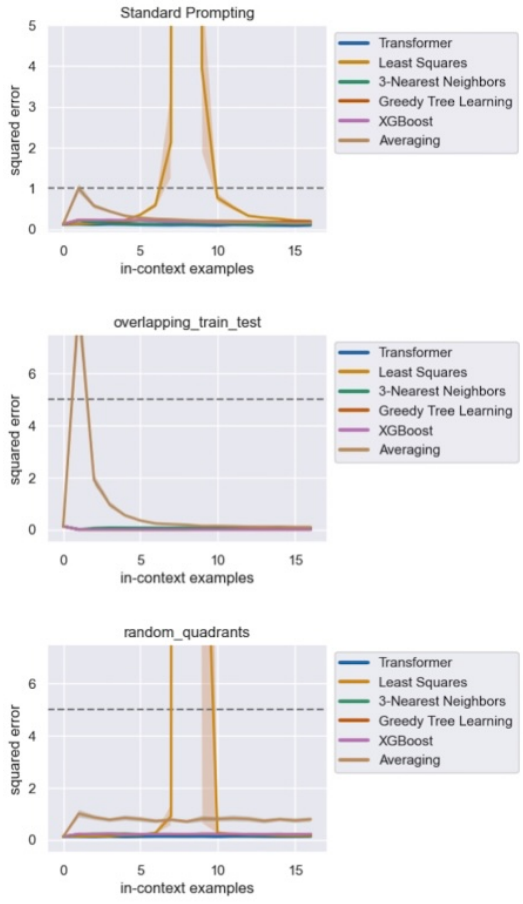

Figure 3: Model Performance on noisy training with std = 0 (baseline)

As we introduced a moderate level of noise (standard deviation of 1), as shown in Figure 4, our model continued to perform comparably to other baseline models. This performance parity was particularly noteworthy given the challenges posed by noise.

In pushing the boundaries of our analysis to a higher noise level (standard deviation of 3), we observed a noticeable increase in the RMSE across all baseline models, as displayed in Figure 5. Remarkably, the transformer model not only outperformed other models but also maintained a consistent level of performance despite the increased number of in-context examples. This resistance to performance degradation under heightened noise levels offers promising implications for its application in real-world scenarios where data imperfections are common.

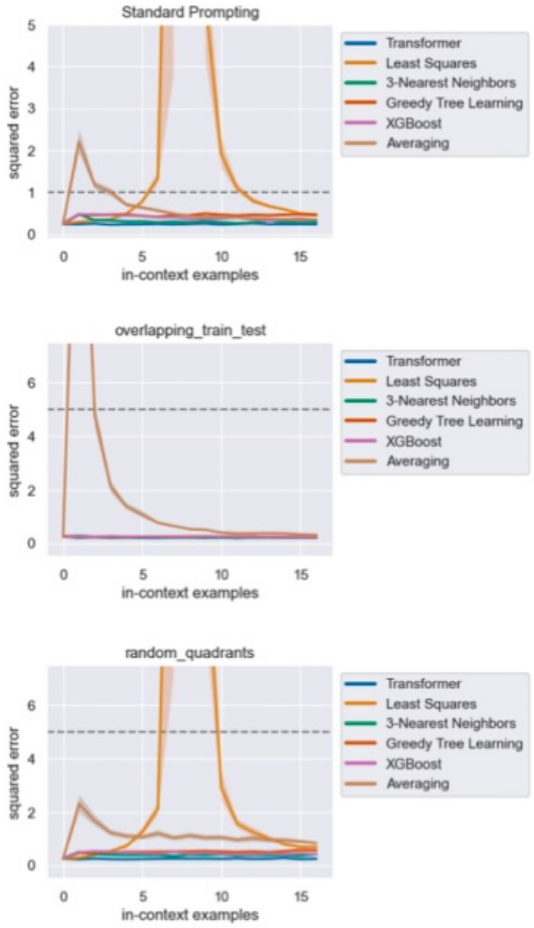

Figure 4: Model Performance on noisy training with std = 1

These insights collectively highlight the necessity for more exhaustive experimentation, especially with an increased number of dimensions, to further explore the limits of in-context learning and its practical effectiveness across varying conditions.

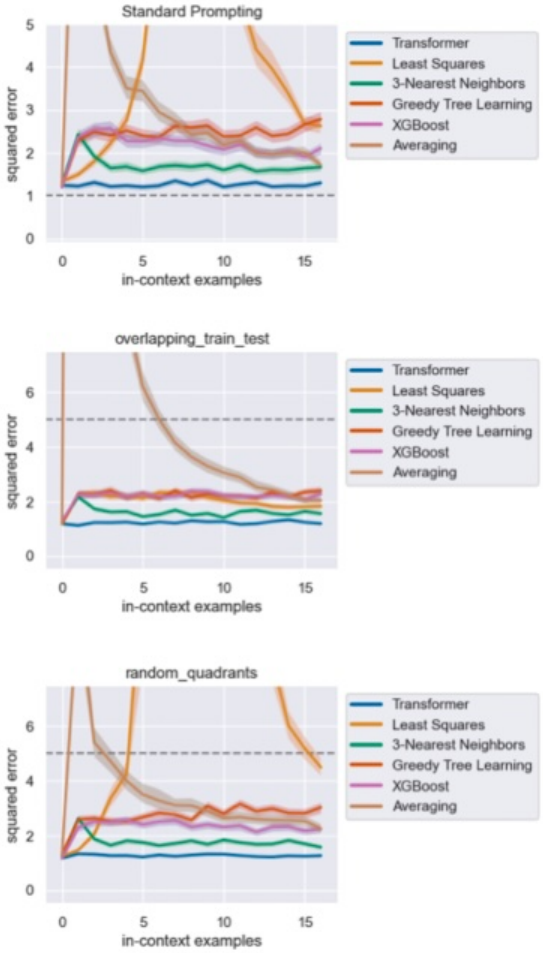

Figure 5: Model Performance on noisy training with std = 3

5 Future Directions

An interesting dimension for this study required comparing our model's performance with a tokenizer-based pretrained architecture like GPT-2. This approach necessitated an evaluation of a model that was trained to learn in-context versus a model trained in a self-supervised manner with a corpus of text. This comparison led to bottlenecks in two main areas. Firstly, as we aimed to keep the model performances comparable in terms of parameters, we encountered limitations due to the maximum token length of the GPT tokenizer. In fact, for our custom configuration, the tokenizer was found to max out on token length. To address this issue, we considered employing larger architectures, such as GPT-3.5 and even GPT-4.

However, an interesting observation with GPT-4 was that the model attempted to fit the data points to a linear regression model, i.e., it tapped into its training data and tried to fit the data based on that model. This behavior is indicative that the model does try to select a latent space from the training data and uses that in conjunction with the in-context examples. To summarize, the intricate relationship between training data and in-context examples is very nuanced, which requires further experimentation. This study serves as a means to contribute to this novel field, where we explored how non-linear decision tree algorithms can be learned in-context.

References

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. 2020. Language models are few-shot learners.

Chen Cheng, Haodong Wen, Xinzhi Yu, and Zeming Wei. 2023. On the robustness of in-context learning with noisy labels: Train, inference, and beyond.

Chen Cheng, Xinzhi Yu, Haodong Wen, Jingsong Sun, Guanzhang Yue, Yihao Zhang, and Zeming Wei. 2024. Exploring the robustness of in-context learning with noisy labels.

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

Siddhant Garg, Goutham Ramakrishnan, and Varun Thumbe. 2021. Towards robustness to label noise in text classification via noise modeling. In *Proceedings of the 30th ACM International Conference on Information amp; Knowledge Management*, CIKM '21. ACM.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

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

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds,

Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2022. In-context learning and induction heads.

Qiyao Peng, Hongtao Liu, Zhepeng Lv, Qing Yang, and Wenjun Wang. 2023. Contrastive pre-training for personalized expert finding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15797–15806, Singapore. Association for Computational Linguistics.

Frieda Rong. 2021. Extrapolating to unnatural language processing with gpt-3's in-context learning: The good, the bad, and the mysterious.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

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