# SIMPLE PERMUTATIONS CAN FOOL LLAMA: PERMUTATION ATTACK AND DEFENCE FOR LARGE LANGUAGE MODELS

**Liang Chen**[♠] **Yatao Bian**[♡] **Li Shen**[♣] **Kam-Fai Wong**[♠]
♠ The Chinese University of Hong Kong
♡ Tencent AI Lab
♣ Sun Yat-sen University
{lchen,kfwong}@se.cuhk.hk
{yatao.bian,mathshenli}@gmail.com

## ABSTRACT

In-context learning (ICL) enables Large Language Models (LLMs) to undertake challenging tasks through given examples. However, it is prone to instability: different orderings of input examples can significantly influence predictions. Current mitigation strategies, focused on post-processing, fail to enhance the model's inherent robustness. This paper extensively investigates this issue of LLMs and uncovers a natural, permutation-based attack that can nearly achieve 100% success rates on LLMs, while remaining imperceptible to humans. To address this vulnerability, we propose a distributionally robust optimization (DRO)-based tuning method as a defence, explicitly optimizing the model's performance against worst-case permutations to bolster robustness. Our framework comprises two modules: the Permutation Proposal network (P-Net) and LLM. The P-Net formulates the identification of the most challenging permutation as an optimal transport problem, solved using the Sinkhorn algorithm. Through adversarial training, the P-Net progressively enhances the LLM's robustness against permutation instability. Experiments with a synthetic task and ICL tuning task demonstrate that our methodology effectively mitigates permutation attacks and enhances overall performance.

## 1 INTRODUCTION

Human intelligence excels at learning new tasks from limited examples, a trait increasingly emulated by LLMs through few-shot, in-context learning (Brown et al., 2020; Chowdhery et al., 2023), with performance further enhanced by targeted fine-tuning (OpenAI et al., 2023; Min et al., 2022; Wei et al., 2023). This capability has been harnessed to achieve success in a broad range of NLP tasks, such as dialogue generation (Deng et al., 2023; Wang et al., 2023) and question answering (Chen et al., 2023a). However, despite these advances, the ICL capabilities of LLMs can react unpredictably to minor prompt perturbations, notably the permutation of input examples (Lu et al., 2022; Zhao et al., 2021). This fragility underscores a significant gap in achieving human-like adaptability.

Our investigations have identified a critical susceptibility in LLMs to adversarial attacks via strategic reordering of ICL examples. This method achieves near-universal success in deceiving the advanced open-source LLM, LLaMA-2-7B, on 11 public datasets without altering prompt semantics, remaining undetected by humans but significantly impairing LLM performance.

Current countermeasures have emerged in two primary forms: output calibration, effective for classification but limited in generative tasks (Zhao et al., 2021); permutation order optimization, constrained by its exponential complexity (Lu et al., 2022). These strategies fail to fundamentally fortify LLMs against the subtle yet impactful manipulations of example ordering.

To counteract this vulnerability, we introduce a defence mechanism based on distributionally robust optimization (DRO) (Ben-Tal et al., 2011). Rather than viewing each training instance merely in terms of its permutations observed during training, our method conceptualizes each instance as part of a broader distribution that includes all conceivable permutations. This comprehensive set of

permutations is termed the ambiguity set. By persistently optimizing against the worst-case scenarios within this ambiguity set, our strategy substantially enhances the robustness of LLMs.

Our DRO mechanism operates as a two-player game involving a Permutation Proposal Network (P-Net) and the LLM. P-Net, acting as the adversary, strives to find the most challenging permutation of ICL examples to maximize the LLM's loss. The LLM, in response, aims to minimize the loss despite P-Net's interventions. P-Net formulates the search for the toughest ICL permutation as an optimal transport (OT) problem (Monge, 1781), iteratively addressed with the Sinkhorn algorithm. This adversarial training allows P-Net to challenge the LLM with increasingly difficult permutations, compelling the LLM to improve its defence against potential attacks. Our framework explicitly targets the LLM's worst-case performance, enhancing its defence against permutation attacks.

Our empirical studies on fitting linear functions and in-context tuning reveal that our DRO tuning framework effectively counters permutation attacks and improves LLM performance. These results confirm the method's capacity to strengthen LLM robustness and advance adaptability.

## 2 Related Work

**In-Context Learning** Large Language models have demonstrated in-context learning through exemplar-based adaptation, a capability pioneered by (Brown et al., 2020). Subsequent studies have attributed this to the transformer's standard learning algorithms (Akyürek et al., 2023; von Oswald et al., 2022; Dai et al., 2023). Further research (Garg et al., 2022; Min et al., 2022; Wei et al., 2023) has shown that transformers can significantly enhance ICL when explicitly fine-tuned with in-context objectives. Nevertheless, the robustness of LLMs to permutations of input examples (Brown et al., 2020; Zhao et al., 2021) remains an unresolved challenge. Contemporary approaches have centred on post-processing techniques, such as model calibration (Zhao et al., 2021) or exhaustively seeking the optimal sequence of ICL samples (Lu et al., 2022)—a process with prohibitive combinatorial complexity. However, these methods fail to fundamentally strengthen the LLMs' robustness against varying input orders. Our contribution is an improved ICL tuning algorithm that bolsters LLMs' resistance to suboptimal permutations, thereby directly enhancing robustness to permutation variability.

**Distributionally Robust Optimization** In distributionally robust optimization (DRO), ambiguity sets are often defined as divergence balls centred on the empirical distribution of data pairs $(x, y)$, which act as regularizers for small radii (Ben-Tal et al., 2013; Lam & Zhou, 2015; Duchi et al., 2016; Miyato et al., 2018). However, larger radii can result in excessively conservative sets. Our methodology deviates from this radius-centric paradigm, constructing ambiguity sets via permutations of samples that are semantically invariant. Prior applications of DRO have addressed distributional shifts, including label (Hu et al., 2018) and data source shifts (Oren et al., 2019). In contrast, our work employs DRO in the context of overparameterized LLMs, which are prone to suboptimal worst-case generalization, a departure from the bulk of DRO research that focuses on traditional, underparameterized models (Namkoong & Duchi, 2017; Duchi et al., 2019).

**Optimal Transport** Optimal transport (OT), a foundational mathematical discipline established by (Monge, 1781; Kantorovich, 1942), provides a metric for measuring distances between distributions, commonly known as the Wasserstein distance or Earth Mover Distance. It has been applied as a tool for manipulating probability distributions. In our study, the Permutation Proposal Network (P-Net) is designed to act as a conduit for transportation between two discrete measures, leveraging entropy-constrained OT (Cuturi, 2013), also referred to as the Sinkhorn distance, to enable the derivation of a differentiable loss (Genevay et al., 2018). Our work extends the concept of learning permutation structures through neural networks, as explored in (Mena et al., 2018) for learning to sort numbers or solve jigsaw puzzles. However, we apply this concept to the more intricate domain of NLP, where P-Net learns the most challenging permutations of ICL samples in a meta-learning manner, aiming to enhance the robustness of LLMs.

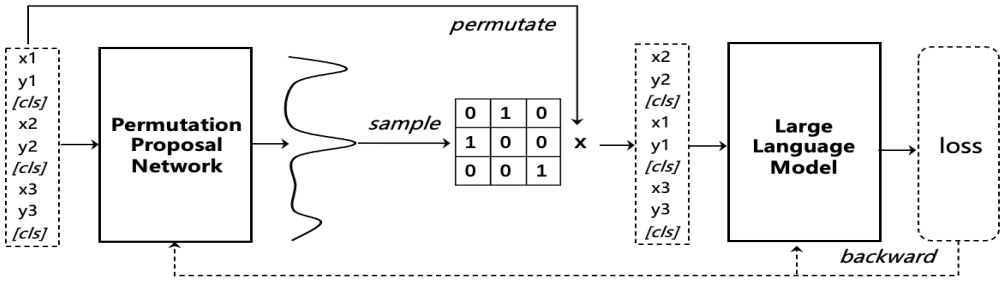


Figure 1: An overview of the adversarial training framework.

# 3 METHODOLOGY

## 3.1 ICL TUNING VIA DISTRIBUTIONALLY ROBUST OPTIMIZATION

Training LLMs for the ICL problem involves predicting outputs $y \in \mathcal{Y}$ from input examples $x \in \mathcal{X}$ and few-shot prompts $p \in \mathcal{P}$. The standard goal, given a language model parameter space $\Theta$ and a loss function $\ell : \Theta \times (\mathcal{P} \times \mathcal{X} \times \mathcal{Y}) \to \mathbb{R}_+$, is to find a model $\theta \in \Theta$ that minimizes the expected loss $\mathbb{E}_P[\ell(\theta; (p, x, y))]$ over the underlying data distribution $P$. Typically, this minimization is approached via empirical risk minimization (ERM):

$$\hat{\theta}_{\text{ERM}} \arg\min_{\theta \in \Theta} \mathbb{E}_{(p,x,y)\sim\hat{P}}[\ell(\theta; (p, x, y))] \tag{1}$$

where $\hat{P}$ denotes the empirical distribution over the training data. However, given that $\hat{P}$ covers only a subset of all possible permutations of the ICL prompt demonstrations, the model might face a variety of permutations during testing, where performance is not guaranteed and could significantly deteriorate. Training exhaustively on all permutations is not computationally viable, as it is an NP-hard challenge due to the combinatorial explosion of possible inputs.

To overcome this challenge, our approach seeks to identify the worst-case scenarios within the combinatorial inputs and optimize model performance accordingly. For this purpose, we employ a distributionally robust optimization (DRO) strategy:

$$\hat{\theta}_{\text{DRO}} = \arg\min_{\theta \in \Theta} \Big\{ \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(p,x,y)\sim Q}[\ell(\theta; (p, x, y))] \Big\} \tag{2}$$

The $\mathcal{Q}$ denotes the ambiguity set, it represents the combinatorial space of inputs that the model aims to be robust against. This considers training data not merely as individual points but as a distribution subject to perturbations, thereby enhancing the model's generalization capabilities. Moreover, by optimizing a lower bound on $\mathcal{Q}$, we render this process computationally feasible. To this end, we endeavour to spot a realistic set of potential worst-case permutations, enhancing the model's robustness to such variations.

## 3.2 CRAFTING WORST LATENT PERMUTATIONS VIA SINKHORN OPERATOR

The NP-hard nature of finding challenging permutations for language models in in-context learning necessitates a computationally feasible alternative to brute force searches. We introduce an approach that frames the problem within the scope of entropy-constrained optimal transport, resolved by the Sinkhorn operator (Sinkhorn, 1966; Mena et al., 2018).

Our Permutation Proposal Network (P-Net) is designed to generate challenging permutations for ICL examples. The P-Net processes examples $\{(x_i^p, y_i^p)\}$, each initiated with a [cls] token. This yields a latent matrix $H$, which is then projected into a permutation matrix $X$ through a transformation $W$ and nonlinear activations $g$, where $X$ is given by $X = g((HW)H^T)$.

To refine $X$ towards a permutation matrix, the iterative Sinkhorn normalization is applied until convergence to a doubly stochastic matrix is achieved (Sinkhorn, 1966):

$$S(X) = \lim_{l \to \infty} \mathcal{T}_c(\mathcal{T}_r(\exp(X))) \tag{3}$$

$$\mathcal{T}_r(X) = X \oslash (X \mathbf{1}_N \mathbf{1}_N^\top) \tag{4}$$

$$\mathcal{T}_c(X) = X \oslash (\mathbf{1}_N \mathbf{1}_N^\top X) \tag{5}$$

where $\mathcal{T}_r(X)$ and $\mathcal{T}_c(X)$ represent the row and column normalization operators of a matrix, respectively, with $\oslash$ indicating element-wise division, and $\mathbf{1}_N$ a column vector of ones. As established by Sinkhorn (1966), the Sinkhorn operator $S(X)$ converges to the set of doubly stochastic matrices as the number of iterations $l$ approaches infinity.

For further convergence to a discrete permutation matrix, we utilize a Gumbel distribution with a temperature parameter $\tau$, as $\tau$ approaches zero (Gumbel, 1954):

$$\Pi = \lim_{\tau \to 0^+} S\left((X + u)/\tau\right), \quad u \sim -\log(-\log(U(0,1))) \tag{6}$$

### 3.3 Adversarial Optimization

In our learning framework, the LLM and the P-Net engage in a co-optimization process. P-Net is designed to generate permutations that convert ICL demonstrations into prompts that are challenging for the LLM. Subsequently, the LLM predicts outputs for these transformed instances (Figure 1).

This adversarial relationship features duality: P-Net progressively escalates the complexity of permutations to challenge the LLM. In turn, the LLM responds to these permutations and provides feedback to refine P-Net's permutation strategy. The iterative process persists until P-Net can uniformly simulate permutation distributions and the LLM can effectively interpret complex permutations.

The combined loss function for the LLM, which includes a maximum likelihood loss and a Lipschitz penalty term as proposed by (Arjovsky et al., 2017), is given by:

$$L(\theta)_{LM} = \mathbb{E}_{(p,x,y)\sim\hat{P},\Pi\sim G(\phi;p)}\left[\ell(\theta;(\Pi\cdot p,x,y)) + \alpha\left(\|\nabla_{\hat{p}}\ell(\theta;\hat{p},x,y)\|_2 - 1\right)^2\right] \tag{7}$$

where $\hat{p} = \beta \cdot p + (1-\beta)\cdot\Pi\cdot p$, and $\beta$ is sampled from a uniform distribution $\mathcal{U}(0,1)$.

P-Net's loss function, aiming to maximize the LLM's difficulty level, can be expressed as:

$$L(\phi)_{P-Net} = \mathbb{E}_{(p,x,y)\sim\hat{P},\Pi\sim G(\phi;p)}[b - \ell(\theta;(\Pi\cdot p,x,y))] \tag{8}$$

The hyperparameter $\lambda$ is introduced to balance the optimization of both networks. The aggregate optimization objective, minimizing the combination of the LLM's and P-Net's losses, is thus:

$$\min_{\theta,\phi}\left\{L(\theta)_{LM} + \lambda L(\phi)_{P-Net}\right\} \tag{9}$$

The comprehensive training algorithm of this co-optimization is presented in Algorithm 1.

**Algorithm 1** Training Procedure for LM with P-Net

**Input**: Corpus $\hat{P}$, LM parameters $\theta$, P-Net parameters $\phi$, the number of LM iterations per generator iteration $n$, Difficulty coefficient $\lambda$.
**repeat**
  **for** $t = 1, \ldots, n$ **do**
    Sample an instance $(p, x, y)$ from $\hat{P}$.
    Generate a permutation matrix $\Pi \sim G(\phi; p)$.
    Update $\theta$ by ascending its gradient $\nabla_\theta L(\theta)_D$.
  **end for**
  Update $\phi$ by ascending its gradient $\lambda\nabla_\phi L(\phi)_G$.
**until** the loss function converges or a predefined number of iterations is reached
**Output**: Optimized parameters $\theta$, $\phi$.

## 4 Experiments

We validate the effectiveness of our method on fitting linear functions and in-context tuning tasks. For each test sample, we use randomized ICL demonstrations and assess all possible permutations. The statistical metrics across these permutations were reported.

| Model | ICL # | Mean | Var. | Worst. |
|---|---|---|---|---|
| $\text{GPT}_{CL}$ | 3 | 1.45 | 2.13 | 2.67 |
| | 4 | 1.20 | 2.02 | 3.34 |
| | 5 | 1.28 | 2.90 | 5.03 |
| $\text{GPT}_{DRO}$ | 3 | **0.86** | **0.01** | **0.92** |
| | 4 | **0.79** | **0.07** | **1.11** |
| | 5 | **0.87** | **0.12** | **1.33** |

Table 1: Impact of Permutation on ICL.

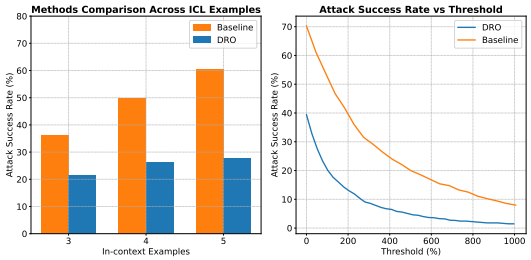


Figure 2: Comparison of Attack Success Rates.

Experimental setup details are presented in Appendix A.

Table 2: Impact of Permutation on Performance of Llama-2.

| Model | ICL # | Origin | Mean | Variance | Best Case | Worst Case |
|---|---|---|---|---|---|---|
| $\text{Llama}_{MetaICL}$ | 3 | 0.190 | 0.112 | 0.039 | 0.230 | 0.02 |
| | 4 | 0.280 | 0.120 | 0.054 | 0.390 | 0.002 |
| | 5 | 0.210 | 0.064 | 0.033 | **0.570** | 0.001 |
| $\text{Llama}_{DRO}$ | 3 | 0.190 | **0.183** | 0.007 | **0.250** | **0.170** |
| | 4 | **0.300** | **0.272** | 0.015 | **0.400** | **0.240** |
| | 5 | **0.220** | **0.200** | 0.007 | 0.500 | **0.180** |

**Fitting Linear Functions** The permutation of training examples significantly affects the learning outcomes of GPT2, as evidenced by the data presented in Table 1. Despite the observed average performance enhancement with an increase in the number of incremental ICL samples, this trend is accompanied by greater variability in performance. Specifically, altering the sequence of training examples can lead to a performance drop, characterized by a fourfold increase in mean squared error (MSE) metrics. The left chart in Figure 2 illustrates the success rates of two methodologies across varying ICL sample sizes, with the baseline threshold set at a 50% increase in MSE. The right chart, conversely, depicts the relationship between attack success rates and different thresholds, based on an analysis with four ICL samples. The implementation of our proposed algorithm effectively reduces both the most significant performance declines and the attack success rates by more than 50%.

**ICL Tuning** Table 2 reveals that perturbations significantly compromised accuracy, reducing it to almost zero for scenarios with three, four, and five shots. This reduction implies an attack success rate nearing 100%. In contrast, the deployment of our defence strategy successfully limited performance declines to within 20%, effectively curtailing the susceptibility of the LLama2-7b model to permutation attacks.

## 5 Conclusion

We introduced a DRO tuning approach to enhance the robustness of LLMs against malicious permutations. This approach employs a Permutation Proposal Network (P-Net) that utilizes the Sinkhorn algorithm to generate challenging permutations, combined with adversarial training to systematically improve LLM performance. Through empirical evaluations in both synthetic and in-context learning tuning tasks, our framework has proven effective in mitigating attacks and enhancing the adaptability of LLMs. This research addresses a significant vulnerability in LLMs, setting a foundation for the development of more resilient future language models.

## Limitations

While our framework is designed to be a universal solution for mitigating order sensitivity in NLP fine-tuning, the scope of our empirical validation has been focused on linear function fitting and in-context learning with LLMs. This concentrated approach has not allowed us to explore the framework's applicability across all NLP tasks, including those that involve complex dialogue sequence management in conversational systems or the nuanced ordering of information in knowledge-based documents (Chen et al., 2023b). To ensure the comprehensive efficacy and adaptability of our framework, future research will be directed towards incorporating a wider array of NLP scenarios.

## REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2211.15661.

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 214–223. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/arjovsky17a.html.

A. Ben-Tal, D. den Hertog, A. D. Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59:341–357, 2013.

Aharon Ben-Tal, Dick den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Advanced Risk & Portfolio Management® Research Paper Series*, 2011. URL https://api.semanticscholar.org/CorpusID:761793.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020.

Liang Chen, Yang Deng, Yatao Bian, Zeyu Qin, Bingzhe Wu, Tat-Seng Chua, and Kam-Fai Wong. Beyond factuality: A comprehensive evaluation of large language models as knowledge generators. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 6325–6341, Singapore, December 2023a. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.390. URL https://aclanthology.org/2023.emnlp-main.390.

Liang Chen, Hongru Wang, Yang Deng, Wai Chung Kwan, Zezhong Wang, and Kam-Fai Wong. Towards robust personalized dialogue generation via order-insensitive representation regularization. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 7337–7345, Toronto, Canada, July 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.462. URL https://aclanthology.org/2023.findings-acl.462.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. URL http://jmlr.org/papers/v24/22-1144.html.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In C.J. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? Language models secretly perform gradient descent as meta-optimizers. In *Workshop on Understanding Foundation Models at the International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2212.10559.

Yang Deng, Lizi Liao, Liang Chen, Hongru Wang, Wenqiang Lei, and Tat-Seng Chua. Prompting and evaluating large language models for proactive dialogues: Clarification, target-guided, and non-collaboration. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10602–10621, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.711. URL https://aclanthology.org/2023.findings-emnlp.711.

J. Duchi, P. Glynn, and H. Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv*, 2016.

J. Duchi, T. Hashimoto, and H. Namkoong. Distributionally robust losses against mixture covariate shifts. https://cs.stanford.edu/~thashim/assets/publications/condrisk.pdf, 2019.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30583–30598. Curran Associates, Inc., 2022.

Aude Genevay, Gabriel Peyre, and Marco Cuturi. Learning generative models with sinkhorn divergences. In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1608–1617. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/genevay18a.html.

W. Hu, G. Niu, I. Sato, and M. Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning (ICML)*, 2018.

Leonid Kantorovich. On the transfer of masses. In *Doklady Akademii Nauk*, volume 37, pp. 227–229, 1942.

H. Lam and E. Zhou. Quantifying input uncertainty in stochastic optimization. In *2015 Winter Simulation Conference*, 2015.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. Datasets: A community library for natural language processing. In *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2021. URL https://arxiv.org/abs/2109.02846.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556.

Gonzalo Mena, David Belanger, Scott Linderman, and Jasper Snoek. Learning latent permutations with gumbel-sinkhorn networks. In *International Conference on Learning Representations*, 2018.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. MetaICL: Learning to learn in context. In Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (eds.), *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2791–2809, Seattle, United States, July 2022. Association for Computational Linguistics.

T. Miyato, S. Maeda, S. Ishii, and M. Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

Gaspard Monge. Memoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*, 1781.

H. Namkoong and J. Duchi. Variance regularization with convex objectives. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2023.

Y. Oren, S. Sagawa, T. Hashimoto, and P. Liang. Distributionally robust language modeling. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

Richard Sinkhorn. A relationship between arbitrary positive matrices and stochastic matrices. *Canadian Journal of Mathematics*, 18:303–306, 1966. doi: 10.4153/CJM-1966-033-9.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2022. URL https://arxiv.org/abs/2212.07677.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *BlackboxNLP Workshop at the Conference on Empirical Methods in Natural Language Processing*, 2018. URL https://arxiv.org/abs/1804.07461.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Conference on Neural Information Processing Systems*, 2019. URL https://arxiv.org/abs/1905.00537.

Hongru Wang, Lingzhi Wang, Yiming Du, Liang Chen, Jingyan Zhou, Yufei Wang, and Kam-Fai Wong. A survey of the evolution of language model-based dialogue systems, 2023.

Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, and Quoc Le. Symbol tuning improves in-context learning in language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 968–979, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.61. URL https://aclanthology.org/2023.emnlp-main.61.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang (eds.), *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/zhao21c.html.

# APPENDIX

## A EXPERIMENTAL SETUP DETAILS

### A.1 FITTING LINEAR FUNCTIONS

In this section, we detail the process of training a transformer model, specifically to in-context learn a defined class of functions. We concentrate on a straightforward function class—linear functions—and examine the model's robustness to different demonstration permutations.

**Data Construction.** We focus on the class of linear functions, denoted as $\mathcal{F} = \{f \mid f(x) = w^\top x, w \in \mathbb{R}^d\}$, in a $d$-dimensional space where $d = 5$. We independently draw samples $x_1, \ldots, x_k$, $x_{\text{query}}$, and $w$ from the isotropic Gaussian distribution $N(0, I_d)$. Subsequently, we calculate $y_i = w^\top x_i$ for each $i$ and construct the prompt as $p = (x_1, y_1, x_2, y_2, \ldots, x_k, y_k, x_{\text{query}})$. The model underwent training on a dataset comprising 40,000 linear functions. During testing, novel functions were sampled to evaluate the model's capability to learn the new weight $w$ through given in-context demonstrations.

**Baselines.** In alignment with (Garg et al., 2022), we adopt a curriculum learning (CL) strategy for training a GPT-2 model, using squared error loss as the optimization criterion. The model is initially exposed to 3 demonstrations, with the complexity gradually increasing to 5 demonstrations.

### A.2 IN-CONTEXT LEARNING

**Tuning Tasks & Prompt Construction** In alignment with (Wei et al., 2023), our study involves 22 publicly accessible NLP datasets from HuggingFace (Lhoest et al., 2021), acknowledged widely

in research (Wang et al., 2018; 2019). These datasets cover a broad spectrum of NLP tasks, which we divide into seven categories. For ICL tuning, we generate prompts using training split examples, incorporating 2 to 10 in-context exemplars per class, chosen randomly.

**Evaluation Tasks** To evaluate model performance on unfamiliar tasks, we avoid datasets used in the ICL and instruction tuning stages. We selected 11 unique NLP datasets from HuggingFace, ensuring none were involved in any finetuning phase.

**Baselines** We compare our approach against MetaICL (Min et al., 2022), which refines the base model through explicit ICL fine-tuning on a multitude of tasks.