Rethinking In-Context Learning in Large Language Models as Gradient Descent

Tomer Bar Natan, Gilad Deutch, Nadav Magar
Supervised by Guy Dar
Tel-Aviv University

Abstract

In-context learning (ICL) has shown impressive results in few-shot learning tasks, yet its underlying mechanism is still not fully understood. Recent works suggest that ICL could be thought of as a gradient descent (GD) based meta-optimization process. While promising, these results mainly focus on simplified settings of ICL and provide only a preliminary evaluation of the similarities between the two methods. In this work, we revisit the comparison between ICL and GD-based finetuning and study what properties of ICL an equivalent process must follow. We begin at the model prediction level and reexamine how ICL and finetuning's prediction updates differ. Next, we address the differences in layer causality between ICL and standard finetuning. To study how this dissimilarity affects the model's behavior we propose a causally aware finetuning process and compare it with previous results. The code implementation for our experiments is available at: https://github.com/GiilDe/ft-vs-icl

1 Introduction

In recent years, large language models (Brown et al., 2020) have shown strong emergent in-context learning ability (Wei et al., 2022) where a pretrained model's performance significantly improves on various downstream tasks by simply conditioning on a few input-label pairs (demonstrations). Despite its success and avid research regarding ICL abilities, the inner workings behind the process are still not fully understood. In-context learning operates in a seemingly different approach to few-shot and meta-learning which require additional parameter updates. Nevertheless, a series of recent works show significant similarities between ICL and gradient descent-based optimization (Irie et al., 2022; Von Oswald et al., 2023; Akyürek et al., 2023).

In this paper, we study the results of (Dai et al., 2023) which empirically show connections between ICL and standard finetuning on large GPT models and language classification tasks. We identify both empirical and theoretical considerations not accounted for by their analysis:

- **Prediction Alignment**: From the perspective of model prediction, we show that ICL's and finetuning's predictions are poorly aligned on most tasks. Our results show that both methods yield unique, different prediction updates, especially when taking erroneous changes into account. Motivated by these observations we propose the use of the Jaccard index between relative prediction changes.

- **Layer Causality**: an intrinsic difference in the information flow of attention output updates between standard finetuning and ICL. Think of the update induced by each method to the output of the l-th attention layer in the model. In ICL this update is autoregressive, in the sense that it depends on the previous (lower) layer's output only. In standard finetuning, on the other hand, the update to the attention output of the l-th layer depends on the output of all others - as it is derived from the gradient with respect to all model parameters.

2 Background and Preliminaries

2.1 Dual Form Between Attention and Linear Layers Optimized by Gradient Descent

The view of language models as meta-optimizers originates from the presentation of the dual and primal forms of the perceptron (Aizerman et al., 2019). This notion was later expressed in terms of key-value-query attention operation by (Irie et al., 2022; Dai et al., 2023; Von Oswald et al., 2023) which applied it in the modern context of deep neural networks. They show that linear layers optimized by gradient descent have a dual representation as linear attention.

Let $W \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$ be the weight matrix of a linear layer initialized at W_0, and let $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_n \in \mathbb{R}^{d_{\text{in}}}$ be the input and training examples representation respectively. One step of gradient descent on the loss function $\mathcal{L}$ with learning rate η yields the weight update ΔW. This update can be written as the outer products of the training examples $\mathbf{x}_1, \ldots, \mathbf{x}_n$ and the gradients of their corresponding outputs $\mathbf{e}_i = -\eta \nabla_{W_0 x_i} \mathcal{L}$

$$\Delta W = \sum_i \mathbf{e}_i \otimes \mathbf{x}_i'^T \tag{1}$$

Thus the computation of the optimized linear layer can be formulated as

$$\begin{aligned} \mathcal{F}(\mathbf{x}) &= (W_0 + \Delta W)\,\mathbf{x} \\ &= W_0\mathbf{x} + \Delta W \mathbf{x} \\ &= W_0\mathbf{x} + \sum_i (\mathbf{e}_i \otimes \mathbf{x}_i)\,\mathbf{x} \\ &= W_0\mathbf{x} + \sum_i \mathbf{e}_i \left(\mathbf{x}_i^T \mathbf{x}\right) \\ &= W_0\mathbf{x} + \text{LinearAttn}\,(E, X, \mathbf{x})\,, \end{aligned} \tag{2}$$

where $\text{LinearAttn}(V, K, \mathbf{q})$ denotes the linear attention operation. From the perspective of attention, we regard training examples X as keys, their corresponding gradients as values, and the current input $\mathbf{x}$ as the query.

2.2 Understanding Transformer Attention as Meta-Optimization

In this section, we explain the simplified mathematical view of in-context learning as a process of meta-optimization presented in (Dai et al., 2023). For analysis, it is useful to view the change to the output induced by attention to the demonstration tokens as equivalent parameter update ΔW_{ICL} that takes effect on the original attention parameters.

Let $\mathbf{x} \in \mathbb{R}^d$ be the input representation of a query token t, and $\mathbf{q} = W_Q\mathbf{x} \in \mathbb{R}^{d'}$ be the attention query vector. We use the relaxed linear attention model, whereby the softmax operation and the scaling factor are omitted:

$$\begin{aligned} \mathcal{F}_{\text{ICL}}(\mathbf{q}) &= \text{LinearAttn}(V, K, \mathbf{q}) \\ &= W_V[X'; X]\left(W_K[X'; X]\right)^T \mathbf{q} \end{aligned} \tag{3}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d' \times d}$ are the projection matrices for computing the attention queries, keys, and values, respectively; X denotes the input representations of query tokens before t; X' denotes the input representations of the demonstration tokens; and $[X'; X]$ denotes the matrix concatenation.

They define $W_{\text{ZSL}} = W_V X \left(W_K X\right)^T$ as the initial parameters of a linear layer that is updated by attention to in-context demonstrations. To see this, note that W_{ZSL} is the attention result in the zero-shot learning setting where no demonstrations are given (Equation 3). Following the reverse direction of Equation (2), you arrive at the dual form of the Transformer attention:

$$\begin{aligned} \mathcal{F}_{\text{ICL}}(\mathbf{q}) &= W_{\text{ZSL}}\mathbf{q} + \text{LinearAttn}\left(W_V X', W_K X', \mathbf{q}\right) \\ &= W_{\text{ZSL}}\mathbf{q} + \sum_i W_V \mathbf{x}_i' \left(\left(W_K \mathbf{x}_i'\right)^T \mathbf{q}\right) \\ &= W_{\text{ZSL}}\mathbf{q} + \sum_i \left(W_V \mathbf{x}_i' \otimes \left(W_K \mathbf{x}_i'\right)\right) \mathbf{q} \\ &= W_{\text{ZSL}}\mathbf{q} + \Delta W_{\text{ICL}}\mathbf{q} \\ &= \left(W_{\text{ZSL}} + \Delta W_{\text{ICL}}\right)\mathbf{q}. \end{aligned} \tag{4}$$

By analogy with Equation (2), we can regard $W_K \mathbf{x}_i'$ as the training examples and $W_V X'$ as their corresponding meta-gradients.

3 Experiments

3.1 Evaluation Datasets and Models

We evaluated our experiments on six datasets. **SST2** (Socher et al., 2013) **SST5** (Socher et al., 2013), **MR** (Pang and Lee, 2005) and **Subj** (Pang and Lee, 2004) are four datasets for sentiment classification; **AGNews** (Zhang et al., 2015) is a topic classification dataset; and **CB** (de Marneffe et al., 2019) is used for natural language inference. In our experiments, we use the same GPT-like pre-trained language models used by (Dai et al., 2023) with 1.3B released by fairseq[1].

3.2 Evaluation Metrics

The following sections describe the evaluation metrics adopted from (Dai et al., 2023) to compare the behavior of ICL and finetuning.

Attention Output Direction (SimAOU) This metric quantifies the similarity between two updates to the attention output of a layer with respect to the zero-shot setting. For a given query example, let $h_X^{(l)}$ represent the normalized output representation of the last token at the l-th attention layer within setting X. The updates induced by ICL and finetuning are given by $h_{\text{ICL}}^{(l)} - h_{\text{ZSL}}^{(l)}$

[1]https://github.com/facebookresearch/fairseq

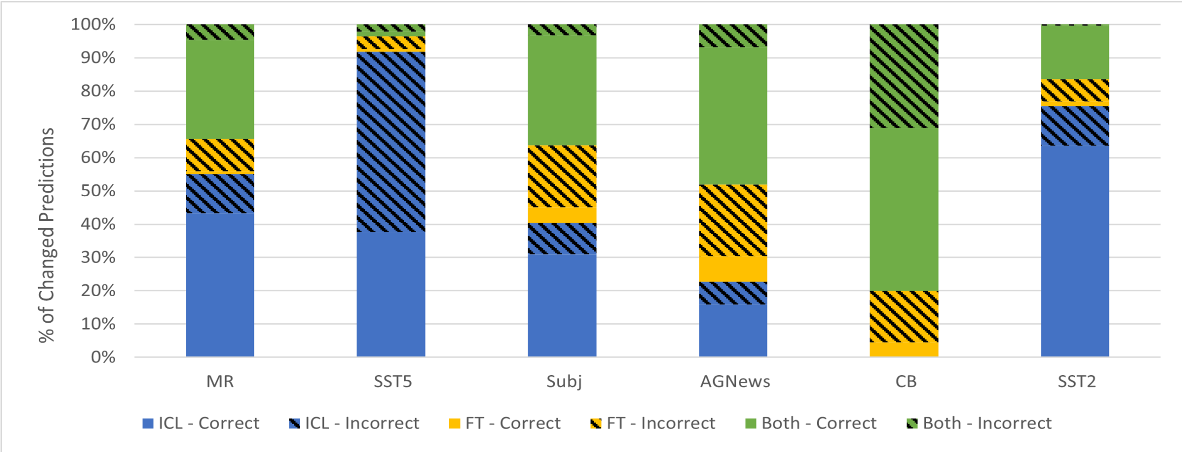

Figure 1: Partition of prediction changes with regards to the zero-shot setting induced by ICL and FT. For each task we evaluate both methods on the same validation set. Correct changes are examples that are misclassified in the ZSL setting and are correctly classified following the prediction update. incorrect changes are examples whose prediction is changed from the ZSL setting into an incorrect label.

and $h_{\text{FT}}^{(l)} - h_{\text{ZSL}}^{(l)}$, respectively. The attention output similarity (**SimAOU**) is defined to be the cosine similarity between these updates averaged across all layers. A higher SimAOU score indicates that ICL is more inclined to adjust the attention output in the same direction as finetuning. For the sake of comparison, this score is compared with a baseline of SimAOU with random attention output updates: $h_{\text{rand}}^{(l)} - h_{\text{ZSL}}^{(l)}$ where $h_{\text{rand}}^{(l)}$ is sampled uniformly.

Attention Map Similarity We use SimAM to measure the similarity between attention maps and query tokens for ICL and finetuning. For a query example, let $m_X^{(l,h)}$ represent the attention weights before softmax in the h-th head of the l-th layer for setting X. In ICL, we focus solely on query token attention weights, excluding demonstration tokens. Initially, before finetuning, we compute the cosine similarity between $m_{\text{ICL}}^{(l,h)}$ and $m_{\text{ZSL}}^{(l,h)}$, averaging it across attention heads to obtain SimAM (Before Finetuning) for each layer. Similarly, after finetuning, we calculate the cosine similarity between $m_{\text{ICL}}^{(l,h)}$ and $m_{\text{FT}}^{(l,h)}$ to obtain SimAM (After FT). A higher SimAM (After FT) relative to SimAM (Before FT) indicates that ICL's attention behavior aligns more with a finetuned model than a non-finetuned one.

3.3 ICL Predictions Poorly Align with Finetuning

In this section, we focus on the perspective of model predictions, regarding both ICL and finetuning as black-box updates to the original zero-shot prediction. While this analysis provides less insight into the inner workings of ICL, prediction alignment is easily interpretable and seems necessary for downstream applications of such comparisons.

Revisiting the results of (Dai et al., 2023), the authors find that ICL achieves high recall to finetuning (Rec2FTP) scores across multiple tasks, which means it covers most of the correct predictions of finetuning. However, their results show a discrepancy between the accuracy of the finetuned model and the ICL setting, average difference of **19.38%** relative to the original zero-shot accuracy (Table 1).

We argue that Rec2FTP is insufficient to quantify prediction alignment in this setting because: (1) it does not the difference between the number of changes induced by each method; (2) it does not account for incorrect prediction changes. Figure 1 shows the number of prediction updates induced by each method. The results show that overall ICL is more inclined to change the ZSL prediction. Note that although ICL covers almost all FT correct prediction changes (high Rec2FTP), in most tasks unique FT and ICL predictions constitute the majority of all updates. This observation shows the importance of measuring incorrect prediction changes as well.

Instead, we measure the **Jaccard Index** of both methods' prediction changes, to quantify prediction alignment in this setting. Given a validation set, we denote the subset of examples whose prediction is changed with regards to the zero-shot setting by ICL or FT by D_{ICL} and D_{FT} respectively. The

	CB	SST2	SST5	Subj	MR	AGNews
ZSL Accuracy	37.5	70.5	39.3	72.6	65.9	46.2
FT Accuracy	57.1	74	39.4	77.8	72.6	66.7
ICL Accuracy	50	92.7	45.0	90.0	89.1	79.2
Jaccard Index (%)	80.0	16.4	3.4	36.2	34.0	48.0
Jaccard - Correct (%)	91.7	19.8	3.2	48.0	40.2	63.5
Jaccard - Incorrect (%)	66.7	1.8	3.5	10.2	17.7	19.2

Table 1: Validation accuracy and Jaccard Index for ZSL, finetuning, and ICL settings on all six classification datasets.

Jaccard index between the updates is given by:

$$\mathcal{J}\left(D_{\mathrm{ICL}}, D_{\mathrm{FT}}\right) = \frac{|D_{\mathrm{ICL}} \cap D_{\mathrm{FT}}|}{|D_{\mathrm{ICL}}| + |D_{\mathrm{FT}}| - |D_{\mathrm{ICL}} \cap D_{\mathrm{FT}}|}$$

We report the indexes across all tasks in Table 1. For each, we also compute the Jaccard index computed over only prediction changes to correct labels, and those for incorrect labels separately.

3.4 Layer Causality in Finetuning

In this section, we highlight an intrinsic difference in the information flow of attention output updates between standard finetuning and ICL.

1. **Layer Causality**: In ICL the update to the output of the l-th attention layer is dependent only on the output of previous (lower) layers. In contrast, the update to the l-th attention output induced by finetuning is determined by the gradient of the entire model's trainable parameters.

2. **Sequential Update**: Equation 4 shows that in ICL each attention layer's output is updated frequently throughout the model's depth axis. However, in GD-based finetuning, all layer's parameters are updated in a single concurrent step.

Motivated by these observations we propose a layer causality-aware finetuning method where each layer is updated individually. Specifically, we project the output of each layer into the label space using the pertained projection head and compute the cross-entropy loss of this prediction. We update each layer sequentially, regarding the output of the previous layer as constant.

We evaluate our method in comparison to standard finetuning using the experiments proposed by (Dai et al., 2023). Table 2 shows the SimAOU and

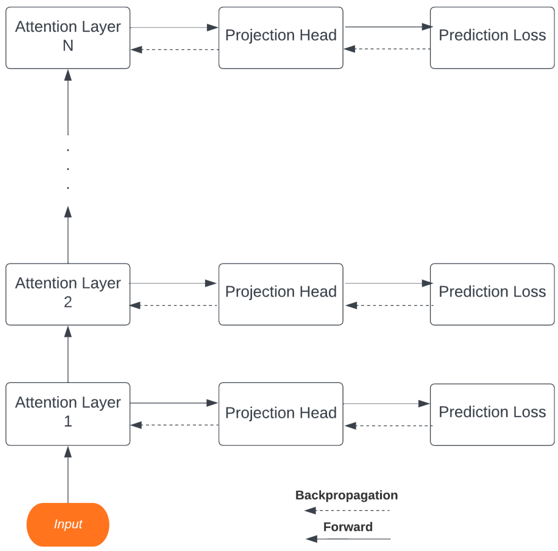

Figure 2: **Layer-Causal Finetuning**: The output of each layer is projected to the label space and used as an intermediate prediction. We compute the prediction loss of each intermediate layer sequentially.

SimAM measurements of both methods in comparison to ICL. Layer causal FT achieves higher SimAOU in 5 out of 6 tasks, but its SimAM is significantly lower.

We highlight two fallacies of our proposed method, and how they might have contributed to lower performance and SimAM scores.

- Lower layers in the models are not trained to directly produce predictions, thus our FT method causes significant drift from the original models' learned weights.

- The gradients applied to lower layers have a bigger magnitude in layer causal FT. During pertaining the gradient of early layers is propagated throughout the model which usually dampens its norm.

	SST2	SST5	MR	Subj	AGNews	CB	Average
SimAOU (Random)	0.0016	0.0025	0.0008	0.0022	0.0021	0.0037	0.0021
SimAOU (FT)	0.1091	**0.113**	0.2190	0.1932	0.3053	0.2013	0.1901
SimAOU (Layer Casual FT)	**0.2297**	0.1065	**0.3299**	**0.3439**	**0.3213**	**0.3435**	**0.2791**
SimAM (ZSL)	0.5546	0.3913	0.3979	0.3786	0.1518	0.1524	0.3377
SimAM (Standard FT)	**0.5850**	**0.4047**	**0.4980**	**0.4870**	**0.4944**	**0.1875**	**0.4427**
SimAM (Layer Causal FT)	0.5774	0.4039	0.2919	0.2844	0.1201	0.0293	0.2845

Table 2: SimAOU and SimAM comparison of standard FT and layer causal FT across six classification datasets. Layer causal FT achieves higher SimAOU across 5 out of 6 tasks, yet its SimAM is significantly lower.

We verify this hypothesis by comparing the norm of each attention layer's gradient during the standard FT process and layer causal FT in Figure 3.

Following this finding, we attempt to apply gradient clipping in the layer causal process with limited success. We report preliminary results using a manually selected clip value on the Subj task in Table 3.

4 Discussion

4.1 Differences Between Theoretical Analysis and Proposed Methods

Most works connecting ICL with gradient-based optimization are motivated by theoretical intuition or even include rigorous analysis for simplified settings (Von Oswald et al., 2023; Akyürek et al., 2023) (see section 5). Our work aims to demonstrate such connections empirically, guided by intuition provided in section 2.1. It is important to note the differences between this analysis and practical settings: (1) It assumes linear attention is used (2) The analysis applies to the update of a single layer (3) Starting from the ICL dual view, we get a partial-differential equation for the underlying loss function whose true value seems intractable. While the empirical results provide direction for future research, we believe further analysis is needed in order to gain more insight from such comparisons.

4.2 Limitations and Future Directions

The method and results shown in section 3.4 are inconclusive. We address its limitation using quantitative analysis, which suggests that further modifications may yield better similarity with ICL. We leave such work to future research.

5 Related Works

A series of recent works explores the similarities between ICL and gradient descent-based optimization. (Akyürek et al., 2023) show that Transformer-based in-context learners can implement standard optimization algorithms on linear models implicitly. (Von Oswald et al., 2023) provide a construction for linear attention-only Transformers models that implicitly perform gradient descent like procedure. (Irie et al., 2022) rewrite the dual form of a linear perceptron in terms of query-key-value attention, and use it to analyze how a trained model is affected by its training samples.

Different from these works, we base our study on (Dai et al., 2023) which studies large GPT transformers on structured language classification tasks. We study how different aspects of ICL can affect the results of the comparison made in (Dai et al., 2023). On the prediction level, we introduce the RPA metric for the evaluation of prediction level alignment. Furthermore, while (Dai et al., 2023) compares standard GD-based finetuning, we test a novel layer causality-aware finetuning process.

6 Conclusion

Inspired by recent works, we attempted to further explore the relationship between in-context learning and gradient descent-based finetuning in practical settings. We revisit existing work and show that ICL and FT predictions are misaligned and propose a better measure to quantify this difference. Finally, we address a fundamental difference in information flow between the methods and suggest a novel FT method that respects layer causality. Our results show potential for a more plausible explanation of ICL and may suggest exciting possible practical applications for embedding context into a model's weights.

7 Acknowledgements

We would like to express appreciation to Guy Dar who supervised this project, for his help with for-

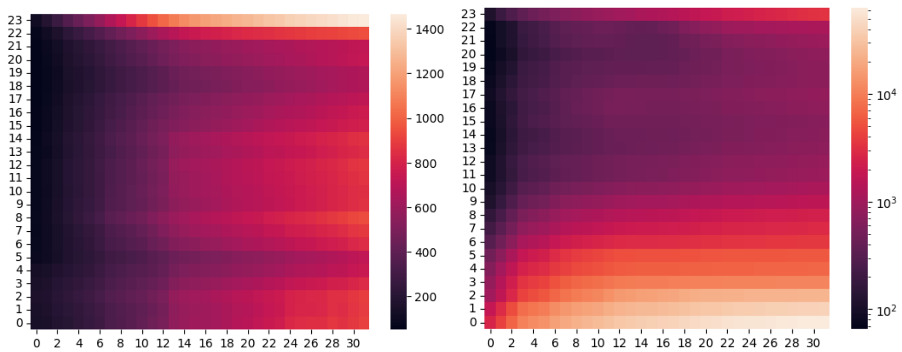

Figure 3: Heatmap of ℓ_2 norms of the gradients computed during finetuning on the Subj task. Note the different scales of magnitude. **Horizontal Axis**: training demonstration index. **Vertical Axis**: layer index in ascending order (from input to network output). **Left**: Standard FT. **Right**: Layer-Causal FT (norm magnitude in logarithmic scale).

	Subj
SimAUO (Standard FT)	0.1932
SimAUO (LC-FT Clipped)	**0.3480**
SimAM (ZSL)	0.3786
SimAM (Standard FT)	**0.4870**
SimAM (LC-FT Clipped)	0.4227

Table 3: Comparison of layer-causal finetuning with gradient norm clipping (clipped to 12.0 in ℓ_∞ norm). This results show that even arbitrary clipping may resolve the drop in SimAM shown in table 2.

mulating the research question and methodology, helpful insights and continuous feedback.

References

Mark A. Aizerman, E. M. Braverman, and Lev I. Rozonoer. 2019. Theoretical foundation of potential functions method in pattern recognition.

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

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. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. 2023. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse.

Kazuki Irie, Róbert Csordás, and Jürgen Schmidhuber. 2022. The dual form of neural networks revisited: Connecting test time predictions to training patterns via spotlights of attention. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 9639–9659. PMLR.

Bo Pang and Lillian Lee. 2004. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 271–es, USA. Association for Computational Linguistics.

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, ACL '05, page 115–124, USA. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank.

In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. 2023. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.