Why Larger Language Models Do In-context Learning Differently?

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, Yingyu Liang
University of Wisconsin, Madison
{zhmeishi,yliang}@cs.wisc.edu, {jwei53,zhuoyan.xu}@wisc.edu

Abstract

Large language models (LLM) have emerged as a powerful tool for many AI problems and are deeply involved in many aspects of human activity. One important emergent ability is *in-context learning* (ICL), where LLM can perform well on unseen tasks based on a brief series of task examples without necessitating any adjustments to the model's parameters. Many works trying to study ICL and one recent interesting counter-intuitive observation is that different scale language models may have different ICL behaviors. Despite the tremendous success made by ICL, why different ICL behaviors remains a mystery. In this work, we are trying to answer this question. As a limited understanding of the ICL mechanism, we study a simplified setting, one-layer single-head linear self-attention network pretrained on linear regression in-context task. We characterize language model scale as the rank of key and query matrix in attention. We show that smaller language models are more robust to noise, while larger language models are easily distracted, leading to different ICL behaviors. We also conduct ICL experiments utilizing the LLaMA model families. The results are consistent with previous work and our analysis.

1 Introduction

As large language models (LLM), e.g., ChatGPT [42] and GPT4 [43], are profoundly changing human society and development, it is critical to understand its mechanism for safe and efficient deployment. One important emergent ability [62], which makes LLM successful, is *in-context learning* (ICL), where models are given a few exemplars of input–label pairs as part of the prompt before performing the evaluation on some new input. More specifically, ICL is a few-shot [9] evaluation method without updating parameters in LLM. Surprisingly, people find that, through ICL, LLM can perform well on tasks that have never been seen before, even without any fine-tuning. It means LLM can adapt to wide-ranging downstream tasks under efficient sample and computation complexity. The mechanism of in-context learning is different from traditional machine learning, such as supervised learning, unsupervised learning, and self/semi-supervised learning. For example, in neural networks, learning usually occurs in gradient updates, whereas there is only a forward inference in ICL and no gradient updates. Several recent works, trying to answer why LLM can learn in-context, argue that LLM secretly performs gradient descent as meta-optimizers with just a forward pass during in-context learning empirically [15, 36, 58] and theoretically [2, 35].

However, recently, there have been some important observations [39, 45, 49, 65] that cannot be explained by existing studies. In particular, [49] finds that LLM is not robust during ICL and can be easily distracted by an irrelevant context. Furthermore, [65] shows that when we inject noise into the prompts, the larger language models may have a worse ICL ability than the small language models, and conjectures that the larger language models may overfit into the prompts and forget the prior knowledge from pretraining, while small models tend to follow the prior knowledge. On the other hand, [39, 45] demonstrate that injecting noise does not affect the in-context learning that much for

R0-FoMo: Workshop on Robustness of Few-shot and Zero-shot Learning in Foundation Models at NeurIPS 2023.

smaller models, which have a more strong pretraining knowledge bias. To understand the mechanism of ICL and to use ICL efficiently and safely, we are interested in the following question:

Why do larger language models do in-context learning differently?

To answer this question, we study a simplified setting, one-layer single-head linear self-attention network [2, 3, 35, 48, 58, 70] pretrained on linear regression in-context task [2, 3, 6, 24, 32, 35, 47, 58, 70]. We characterize language model scale as rank of key and query matrix in attention. Then, we show that smaller language models are more robust to label noise and input noise during evaluation, while larger language models may easily be distracted by such noises, so larger language models may have a worse ICL ability than a smaller language model. We also conduct in-context learning experiments on five prevalent NLP tasks utilizing various sizes of the LLaMA model families [55, 56], whose results are consistent with previous work [39, 45, 65] and our analysis.

2 Related Work

Large language model. Transformer-based [57] neural networks have rapidly emerged as the primary machine learning architecture for tasks in natural language processing. Pretrained transformers with billions of parameters on broad and varied datasets are called large language models (LLM) or foundation models [8], e.g., BERT [17], PaLM [12], LLaMA[55], ChatGPT [42], GPT4 [43] and so on. LLM has shown powerful general intelligence [10] in various downstream tasks. To better use the LLM for a specific downstream task, there are many adaptation methods, such as adaptor [21, 25, 50, 69], calibration [71, 72], multitask finetuning [22, 58, 68], prompt tuning [23, 29], instruction tuning [13, 31, 40], symbol tuning [64], black-box tuning [52], chain-of-thoughts [28, 63], scratchpad [41], reinforcement learning from human feedback (RLHF) [44] and many so on.

In-context learning. One important emergent ability [62] from LLM is in-context learning (ICL) [9]. Specifically, when presented with a brief series of input-output pairings (known as a prompt) related to a certain task, they can generate predictions for test scenarios without necessitating any adjustments to the model's parameters. ICL is widely used in broad scenarios, e.g., reasoning [73], negotiation [20], self-correction [46], machine translation [1] and so on. Many works trying to improve the ICL and zero-shot ability of LLM [26, 38, 60, 61]. There is a line of insightful works to study the mechanism of transformer learning [4, 5, 27, 30, 33, 34, 53, 54] and in-context learning [2, 3, 6, 15, 24, 32, 35, 45, 47, 58, 67, 70] empirically and theoretically. On the basis of these works, our analysis takes a step forward to show the ICL behavior difference under different scales of language models.

3 Preliminary Setup

Notation. We follow the setup and notation of the problem in [2, 35, 70]. We denote $[n] := \{1, 2, \ldots, n\}$. For a positive semidefinite matrix A, we denote $\|x\|_A^2 := x^\top Ax$ as the norm induced by a positive definite matrix A. We denote $\|\cdot\|_F$ as the Frobenius norm.

In-context learning. We consider the linear regression task for in-context learning which is widely studied empirically [3, 6, 24, 47, 58] and theoretically [2, 32, 35, 70]. During learning ICL (pretraining), for each prompt, we have an embedding matrix E_τ which is formed using a d-dimension task weights $w_\tau \in \mathbb{R}^d$ and N examples $(x_{\tau,1}, y_{\tau,1}), \ldots (x_{\tau,N}, y_{\tau,N})$ and a query token $x_{\tau,q}$ for prediction, where for any $i \in [N]$ we have $y_{\tau,i} = \langle w_\tau, x_{\tau,i}\rangle \in \mathbb{R}$ and $x_{\tau,i}, x_{\tau,q} \in \mathbb{R}^d$. The form of E_τ is,

$$E_\tau := \begin{pmatrix} x_{\tau,1} & x_{\tau,2} & \ldots & x_{\tau,N} & x_{\tau,q} \\ y_{\tau,1} & y_{\tau,2} & \ldots & y_{\tau,N} & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(N+1)}. \tag{1}$$

We assume the task weights $w_\tau \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d\times d})$ and for any $i \in [N]$ tokens $x_{\tau,i}, x_{\tau,q} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda)$, where Λ is the covariance matrix of the token. We have a network f which has the parameter θ. Then the network prediction is $\widehat{y}_{\tau,q} := f_\theta(E_\tau)$. We consider the mean square error (MSE) loss so that the empirical risk over independent B prompts is defined as

$$\widehat{\mathcal{L}}(f_\theta) := \frac{1}{2B}\sum_{\tau=1}^{B}\left(\widehat{y}_{\tau,q} - \langle w_\tau, x_{\tau,q}\rangle\right)^2. \tag{2}$$

Remark 1. *For simplicity, we consider a fixed embedding method so that there are no embedding parameters in the network. Also, we do not consider noise in labels during pretraining, while we may consider noise in labels during evaluation.*

In fact, our pre-training period is called learning to learn in-context [38] or in-context training warmup [18], where the network needs to pretrain on some related in-context learning prompts and then evaluate on a new task, e.g., a new w above, which may never be seen in pertaining. The learning to learn in-context is the first step to understanding the mechanism of ICL in LLM.

Linear self-attention networks. We study a one-layer single-head linear self-attention network (LSA). The linear self-attention module is widely studied [2, 3, 35, 48, 58, 70]. It is defined as

$$f_{\text{LSA},\theta}(E) = \left[E + W^{PV} E \cdot \frac{E^\top W^{KQ} E}{\rho}\right]_{(d+1),(N+1)} \tag{3}$$

where $\theta = (W^{PV}, W^{KQ})$, $E \in \mathbb{R}^{(d+1)\times(\rho+1)}$ being an embedding matrix, and ρ is a normalization factor (guarantee the linear self-attention having similar behavior as the softmax-attention), being as the length of examples, i.e., $\rho = N$ during pretraining. Similar to existing work, for simplicity, we merge the projection matrix and the value matrix into W^{PV}, and we merge the key matrix and the query matrix in attention into W^{KQ}. We also have a residual connection in our LSA network. The prediction of the network for the token x_q will be the bottom right entry of the matrix output, that is, the entry $(d+1), (N+1)$, while other entries are features we may ignore. Thus, there are some parameters irrelevant to our loss. To see how, let us denote

$$W^{PV} = \begin{pmatrix} W_{11}^{PV} & w_{12}^{PV} \\ (w_{21}^{PV})^\top & w_{22}^{PV} \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)}, \quad W^{KQ} = \begin{pmatrix} W_{11}^{KQ} & w_{12}^{KQ} \\ (w_{21}^{KQ})^\top & w_{22}^{KQ} \end{pmatrix} \in \mathbb{R}^{(d+1)\times(d+1)},$$

where $W_{11}^{PV}, W_{11}^{KQ} \in \mathbb{R}^{d\times d}$ and $w_{12}^{PV}, w_{21}^{PV}, w_{12}^{KQ}, w_{21}^{KQ} \in \mathbb{R}^d$ and $w_{22}^{PV}, w_{22}^{KQ} \in \mathbb{R}$. Then,

$$\widehat{y}_q = f_{\text{LSA},\theta}(E) = \begin{pmatrix} (w_{21}^{PV})^\top & w_{22}^{PV} \end{pmatrix} \left(\frac{EE^\top}{\rho}\right) \begin{pmatrix} W_{11}^{KQ} \\ (w_{21}^{KQ})^\top \end{pmatrix} x_q. \tag{4}$$

Remark 2. *In practical pertaining, transformers use a sliding window of stride 1 (moving 1 data point/token forward each time), while our setting can be viewed as moving a large stride sliding window so that no overlapping among windows, i.e., all windows are independent with each other.*

Measure model scale by rank. Before introducing our measurement, we first introduce a lemma from previous work which simplifies MSE loss to a quadratic function so that we can easily calculate the optimal solution. For notation simplicity, we denote $U = W_{11}^{KQ}, u = w_{22}^{PV}$.

Lemma 3 (Lemma A.1 in [70])**.** *Let* $\Gamma := \left(1 + \frac{1}{N}\right)\Lambda + \frac{1}{N}\operatorname{tr}(\Lambda) I_{d\times d} \in \mathbb{R}^{d\times d}$*. Let*

$$\mathcal{L}(f_{\text{LSA},\theta}) = \lim_{B\to\infty} \widehat{\mathcal{L}}(f_{\text{LSA},\theta}) = \frac{1}{2}\mathbb{E}_{w_\tau, x_{\tau,1},\ldots,x_{\tau,N},x_{\tau,q}}\left[(\widehat{y}_{\tau,q} - \langle w_\tau, x_{\tau,q}\rangle)^2\right], \tag{5}$$

$$\tilde{\ell}(U, u) = \operatorname{tr}\left[\frac{1}{2}u^2 \Gamma\Lambda U \Lambda U^\top - u\Lambda^2 U^\top\right], \tag{6}$$

we have $\mathcal{L}(f_{\text{LSA},\theta}) = \tilde{\ell}(U, u) + C$*, where* C *is a constant independent with* θ*.*

Lemma 3 tells us that the loss only depends on uU where u is a scalar. If we consider non-zero u, w.l.o.g, letting $u = 1$, then we can see that the loss only depends on $U \in \mathbb{R}^{d\times d}$, non-strictly,

$$\mathcal{L}(f_{\text{LSA},\theta}) = \operatorname{tr}\left[\frac{1}{2}\Gamma\Lambda U \Lambda U^\top - \Lambda^2 U^\top\right]. \tag{7}$$

Note that $U = W_{11}^{KQ}$, then it is natural to measure the size of the language model by rank of U. Recall that we merge the key matrix and the query matrix in attention together, i.e., $W^{KQ} = (W^K)^\top W^Q$. Thus, a low-rank U is equivalent to the constraint $W^K, W^Q \in \mathbb{R}^{r\times d}$ where $r \ll d$. The low-rank key and query matrix are practical and have been widely studied [7, 11, 16, 19, 25, 51]. In this work, we use $r = \operatorname{rank}(U)$ to measure the scale of language models, i.e., larger r representing larger language models. Thus, to study the behavior difference under different scale language models, we will analyze when U under different rank constraints.

4 Theoretical Results

Now, we are ready to present our theoretical results. In Section 4.1, we study the optimal rank-r solution of $f_{\text{LSA},\theta}$, and show that the optimal rank-r solution indeed is the truncated version of the optimal full-rank solution. On the basis of that, we can show the behavior difference in Section 4.2. In short, as a small language model is a truncated version of a large language model, the small model is able to rule out additional label noise and input noise so that it may have a better ICL ability.

4.1 Low Rank Optimal Solution

As Λ is a covariance matrix, Λ is a positive semidefinite symmetric matrix. Thus, we have Λ is diagonalizable, where its eigenvalues are real and non-negative, and its eigenvectors are orthogonal. We have eigendecomposition $\Lambda = QDQ^\top$, where Q is an orthonormal matrix containing eigenvectors of Λ and D is a sorted diagonal matrix with non-negative entries containing eigenvalues of Λ, denoting as $D = \text{diag}([\lambda_1, \dots, \lambda_d])$, where $\lambda_1 \geq \dots \geq \lambda_d \geq 0$. Then, we have the following theorem.

Theorem 4.1 (Optimal rank-r solution)**.** *Recall the loss function $\tilde{\ell}$ in Lemma 3. Let*

$$U^*, u^* = \underset{U\in\mathbb{R}^{d\times d}, \text{rank}(U)\leq r, u\in\mathbb{R}}{\text{argmin}} \tilde{\ell}(U, u). \tag{8}$$

Then $U^* = cQV^*Q^\top, u = \frac{1}{c}$, where c is any non-zero constant and $V^* = \text{diag}([v_1^*, \dots, v_d^*])$ is satisfying for any $i \leq r, v_i^* = \frac{N}{(N+1)\lambda_i + \text{tr}(D)}$ and for any $i > r, v_i^* = 0$.

Proof sketch of Theorem 4.1. We defer the full proof to Appendix A.1. The proof idea is that

$$\underset{U\in\mathbb{R}^{d\times d}, \text{rank}(U)\leq r, u\in\mathbb{R}}{\text{argmin}} \tilde{\ell}(U, u) = \underset{U\in\mathbb{R}^{d\times d}, \text{rank}(U)\leq r, u\in\mathbb{R}}{\text{argmin}} \left(\tilde{\ell}(U, u) - \min_{U\in\mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U, u)\right). \tag{9}$$

Denote $D' = \left(1 + \frac{1}{N}\right) D + \frac{1}{N}\text{tr}(D)I_{d\times d}$. We can see $\Lambda^{\frac{1}{2}} = QD^{\frac{1}{2}}Q^\top, \Gamma^{\frac{1}{2}} = QD'^{\frac{1}{2}}Q^\top$, and $\Gamma^{-1} = QD'^{-1}Q^\top$. We can show $\tilde{\ell}(U, u) - \min_{U\in\mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U, u) = \frac{1}{2}\left\|D'^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - D'^{-1}\right)D^{\frac{1}{2}}\right\|_F^2$. We denote $V^* = \text{argmin}_{V\in\mathbb{R}^{d\times d}, \text{rank}(V)\leq r}\left\|D'^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - D'^{-1}\right)D^{\frac{1}{2}}\right\|_F^2$. We can see that V^* is a diagonal matrix. Denote $D' = \text{diag}([\lambda'_1, \dots, \lambda'_d])$ and $V^* = \text{diag}([v_1^*, \dots, v_d^*])$. Then, we have $\left\|D'^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - D'^{-1}\right)D^{\frac{1}{2}}\right\|_F^2 = \sum_{i=1}^d \left(\left(1 + \frac{1}{N}\right)\lambda_i + \frac{\text{tr}(D)}{N}\right)\lambda_i^2 \left(v_i^* - \frac{1}{\left(1+\frac{1}{N}\right)\lambda_i + \frac{\text{tr}(D)}{N}}\right)^2$. We have that $v_i^* \geq 0$ for any $i \in [d]$ and if $v_i^* > 0$, we have $v_i^* = \frac{1}{\left(1+\frac{1}{N}\right)\lambda_i + \frac{\text{tr}(D)}{N}}$. Denote $g(x) = x^2\left(\frac{1}{\left(1+\frac{1}{N}\right)x + \frac{\text{tr}(D)}{N}}\right)$. We get the conclusion by $g(x)$ is an increasing function on $[0, \infty)$. ☐

We denote U^*, u^* above and corresponding $f_{\text{LSA},\theta}$ as the optimal rank-r solution. In detail, the optimal rank-r solution $f_{\text{LSA},\theta}$ satisfies

$$W^{*PV} = \begin{pmatrix} 0_{d\times d} & 0_d \\ 0_d^\top & u \end{pmatrix}, W^{*KQ} = \begin{pmatrix} U^* & 0_d \\ 0_d^\top & 0 \end{pmatrix}. \tag{10}$$

Theorem 4.1 shows that the optimal rank-r solution indeed is the truncated version of the optimal full-rank solution. In detail, (1) for the optimal full-rank solution, we have for any $i \in [d], v_i^* = \frac{N}{(N+1)\lambda_i + \text{tr}(D)}$; (2) for the optimal rank-r solution, we have for any $i \leq r, v_i^* = \frac{N}{(N+1)\lambda_i + \text{tr}(D)}$ and for any $i > r, v_i^* = 0$. Thus, as a small language model is a truncated version of a large language model, the small language model may ignore less important features (noise) but still keep the most important ones (signal) so that it has a smaller evaluation loss and better ICL ability.

4.2 Behavior Difference

We formalize the previous intuition here, where we can see that the different scale language models may have different behaviors. We consider the evaluation prompt to have M examples (may not be equal to N examples during pretraining for a general evaluation setting) with noise in labels (our results can extend to the noiseless case when $\sigma = 0$). Formally, the evaluation prompt is

$$\widehat{E} := \begin{pmatrix} x_1 & x_2 & \dots & x_M & x_q \\ y_1 & y_2 & \dots & y_M & 0 \end{pmatrix} \tag{11}$$

$$= \begin{pmatrix} x_1 & x_2 & \dots & x_M & x_q \\ \langle w, x_1 \rangle + \epsilon_1 & \langle w, x_2 \rangle + \epsilon_2 & \dots & \langle w, x_M \rangle + \epsilon_M & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(M+1)}, \tag{12}$$

where for any $i \in [M]$, $\epsilon_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$.

Remark 4. *Here, we consider task shifts as defined in [70]. We have $x_{\tau,i} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \Lambda)$ for any $i \in [M]$, i.e., no covariate shifts defined in [70], because the transformer will fail in this case. See detailed discussion in [24, 70].*

Recall Q is eigenvectors of Λ, i.e., $\Lambda = QDQ^\top$ and $D = \text{diag}([\lambda_1, \dots, \lambda_d])$. In practice, we can see the large variance part in the inputs x as a useful signal (like words "positive", "negative"), e.g., top r direction in Q, and the small variance part in x as the noise information (like words "even", "just"), e.g., bottom $d - r$ direction in Q. Based on such intuition, we can decompose w accordingly.

Let $s \in \mathbb{R}^d$ be a truncated vector whose non-zero entry can only be in the first r dimensions, i.e., for any $r < i \le d, s_i = 0$. Let $\xi \in \mathbb{R}^d$ be a residual vector whose non-zero entry can only be in the last $d - r$ dimensions, i.e., for any $1 \le i \le r, \xi_i = 0$. Then, we can decompose any task $w = Q(s + \xi)$, where Qs corresponds to inputs useful signal and $Q\xi$ corresponds to inputs noise information. Indeed, the way we decompose w can be viewed as using prior knowledge Λ from pretraining inputs x. If $w = Qs$, the task is related to the attitude (signal), e.g., "positive", "negative". If $w = Q\xi$, the task is related to some minor information (input noise), e.g., "even", "just". On the other hand, as the pertaining data may be from noisy resources, e.g., websites, we suppose $w \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_{d\times d})$ for pertaining. However, for evaluation, we probably focus on the useful signal rather than the noise information of inputs. Thus, our w decomposition captures this intuition. Now, we can decompose our evaluation MSE loss accordingly in the following theorem.

Theorem 4.2 (Evaluation loss)**.** *Recall $s, \xi \in \mathbb{R}^d$ are truncated and residual vectors respectively. Let $w = Q(s + \xi) \in \mathbb{R}^d$. Then for the optimal rank-r solution $f_{\text{LSA},\theta}$ and V^* in Theorem 4.1 , we have evaluation population MSE loss*

$$\mathcal{L}(f_{\text{LSA},\theta}; \widehat{E}) := \mathbb{E}_{x_1,\epsilon_1,\dots,x_M,\epsilon_M,x_q} \left(f_{\text{LSA},\theta}(\widehat{E}) - \langle w, x_q \rangle \right)^2 \tag{13}$$

$$= \frac{1}{M}\|s\|^2_{(V^*)^2D^3} + \frac{1}{M}\left(\|s+\xi\|^2_D + \sigma^2\right) \text{tr}\left((V^*)^2D^2\right) + \|\xi\|^2_D + \sum_{i\in[r]} s_i^2 \lambda_i \left(\lambda_i v_i^* - 1\right)^2. \tag{14}$$

Proof sketch of Theorem 4.2. We defer the full proof to Appendix A.2. The proof idea is the following. Denote $\widehat{\Lambda} := \frac{1}{M}\sum_{i=1}^M x_i x_i^\top$ and $U^* = QV^*Q^\top$. Note the fact that U^* and Λ commute. By Theorem 4.1, we have $\widehat{y}_q = \left(w^\top \widehat{\Lambda} + \frac{1}{M}\sum_{i=1}^M \epsilon_i x_i^\top\right) U^* x_q$. Then, we have

$$\mathbb{E}_{x_1,\epsilon_1,\dots,x_M,\epsilon_M,x_q} \left(\widehat{y}_q - \langle w, x_q \rangle\right)^2 = \underbrace{\mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q - w^\top x_q\right)^2\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\left(\frac{1}{M}\sum_{i=1}^M \epsilon_i x_i^\top U^* x_q\right)^2\right]}_{\text{(II)}}.$$

We see that the label noise can only have an effect in the second term. For the term (I) we have,

$$\text{(I)} = \underbrace{\mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q - w^\top \Lambda U^* x_q\right)^2\right]}_{\text{(III)}} + \underbrace{\mathbb{E}\left[\left(w^\top \Lambda U^* x_q - w^\top x_q\right)^2\right]}_{\text{(IV)}}. \tag{15}$$

We inject $w = Q(s + \xi)$. For the (III) term, by the property of trace and Lemma 6, we have (III) $= \frac{1}{M}\|s\|^2_{(V^*)^2D^3} + \frac{1}{M}\|s + \xi\|^2_D \,\text{tr}\left((V^*)^2D^2\right)$. Similarly, for the term (IV) and term (II), we have (IV) $= \|\xi\|^2_D + \sum_{i\in[r]} s_i^2 \lambda_i \left(\lambda_i v_i^* - 1\right)^2$, and (II) $= \frac{\sigma^2}{M} \text{tr}\left((V^*)^2D^2\right)$. We can conclude by combining four terms. □

In Theorem 4.2, if we have N that is large enough so that $N\lambda_r \gg \operatorname{tr}(D)$, which is practical as we usually pretrain networks on super long text, then we have

$$\mathcal{L}(f_{\text{LSA},\theta}; \widehat{E}) \approx \|\xi\|_D^2 + \frac{1}{M}\left((r+1)\|s\|_D^2 + r\|\xi\|_D^2 + r\sigma^2\right) + \frac{1}{N^2}\|s\|_D^2, \tag{16}$$

where $\frac{1}{N^2}(\cdot)$ is small comparing to the other two terms. For the other two terms: (1) The $\|\xi\|_D^2$ term is due to the approximation power of the network, e.g., $\|\xi\|_D^2 = 0$ for the full-rank optimal solution. On the other hand, if the main component of our evaluation task w is from Qs, i.e., the task w focuses on the useful signal of inputs rather than the noise information, we will have small $\|\xi\|_D^2$ and small evaluation loss. (2) The $\frac{1}{M}(\cdot)$ term will vanish to zero if we have a large sample complexity in the evaluation prompt. However, we mostly only have limited examples in evaluation, e.g. $N \gg M = 8$, so this term will be dominant. On the other hand, if we assume $\|\xi\|_D^2$ is small, we will have roughly $\frac{r}{M}(\|s\|_D^2 + \sigma^2)$ in Equation (16), which means that larger models (optimal solutions of higher rank) have larger evaluation loss, the so-called different size language models doing ICL differently. We formalize the above insight into the following theorem.

Theorem 4.3 (Behavior difference)**.** *Suppose $0 \le r_1 \le r_2 \le d$ and $w = Qs$ where s is r_1-dimension truncated vector. Denote the optimal rank-r_1 solution as f_1 and the optimal rank-r_2 solution as f_2. Then, we have*

$$\mathcal{L}(f_2; \widehat{E}) - \mathcal{L}(f_1; \widehat{E}) = \frac{1}{M}\left(\|s\|_D^2 + \sigma^2\right)\left(\sum_{i=r_1+1}^{r_2}\left(\frac{N\lambda_i}{(N+1)\lambda_i + \operatorname{tr}(D)}\right)^2\right). \tag{17}$$

Proof of Theorem 4.3. Let $V^* = \operatorname{diag}([v_1^*, \ldots, v_d^*])$ satisfying for any $i \le r_1, v_i^* = \frac{N}{(N+1)\lambda_i + \operatorname{tr}(D)}$ and for any $i > r_1, v_i^* = 0$. Let $V'^* = \operatorname{diag}([v'^*_1, \ldots, v'^*_d])$ be satisfied for any $i \le r_2, v'^*_i = \frac{N}{(N+1)\lambda_i + \operatorname{tr}(D)}$ and for any $i > r_2, v'^*_i = 0$. Note that V^* is a truncated diagonal matrix of V'^*. By Theorem 4.1 and Theorem 4.2, we have

$$\begin{aligned}
&\mathcal{L}(f_2; \widehat{E}) - \mathcal{L}(f_1; \widehat{E}) && (18)\\
=&\left(\frac{1}{M}\|s\|^2_{(V'^*)^2D^3} + \frac{1}{M}\left(\|s\|_D^2 + \sigma^2\right)\operatorname{tr}\left((V'^*)^2D^2\right) + \sum_{i\in[r_2]} s_i^2\lambda_i\left(\lambda_i v'^*_i - 1\right)^2\right) && (19)\\
&-\left(\frac{1}{M}\|s\|^2_{(V^*)^2D^3} + \frac{1}{M}\left(\|s\|_D^2 + \sigma^2\right)\operatorname{tr}\left((V^*)^2D^2\right) + \sum_{i\in[r_1]} s_i^2\lambda_i\left(\lambda_i v^*_i - 1\right)^2\right) && (20)\\
=&\frac{1}{M}\left(\|s\|_D^2 + \sigma^2\right)\left(\operatorname{tr}\left((V'^*)^2D^2\right) - \operatorname{tr}\left((V^*)^2D^2\right)\right) && (21)\\
=&\frac{1}{M}\left(\|s\|_D^2 + \sigma^2\right)\left(\sum_{i=r_1+1}^{r_2}\left(\frac{N\lambda_i}{(N+1)\lambda_i + \operatorname{tr}(D)}\right)^2\right). && (22)
\end{aligned}$$

□

Main intuition. By Theorem 4.3, when task w only focuses on useful signal,

$$\mathcal{L}(f_2; \widehat{E}) - \mathcal{L}(f_1; \widehat{E}) \approx \underbrace{\frac{r_2 - r_1}{M}\|s\|_D^2}_{\text{input noise}} + \underbrace{\frac{r_2 - r_1}{M}\sigma^2}_{\text{label noise}}. \tag{23}$$

We can decompose Equation (23) to label noise and input noise, and we know that $\|s\|_D^2 + \sigma^2$ only depends on the intrinsic property of evaluation data and is independent of the model size. When we have a larger language model (larger r_2), we will have a larger evaluation loss gap between the large and small models. It means larger language models may be easily affected by the label noise and input noise and may have worse in-context learning ability, while smaller language models may be more robust to these noises. Moreover, if we increase the label noise scale on purpose, the larger language models will be more sensitive to the injected label noise. This main intuition is consistent with the observation in [49, 65] and our experimental results in Section 5.

5 Experiments

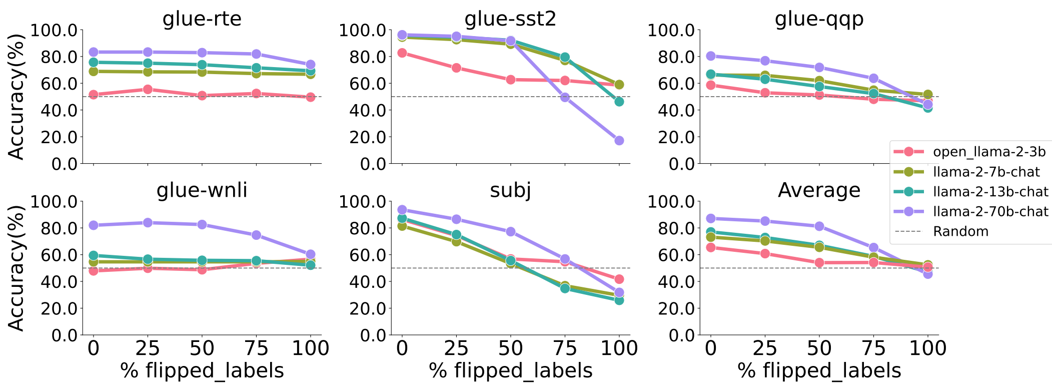

Figure 1: Larger models are easier to override semantic meanings when presented with flipped labels than smaller models for many datasets and model families. Accuracy is calculated over 1000 evaluations prompts per dataset with $M = 16$ in-context exemplars.

Experimental setup. Following the experimental protocols in [39, 65], we conduct experiments on five prevalent NLP tasks, leveraging datasets from **GLUE** [59] tasks and **Subj** [14]. Our experiments utilize various sizes of the LLaMA model families [55, 56] specifically: 3B, 7B, 13B, 70B. We follow the prior literature on in-context learning [65] and use $M = 16$ in-context exemplars. We aim to assess the models' ability to prioritize input-label correlations presented in-context over inherent semantic biases from pretraining. As part of this experiment, we introduce variability by inverting an escalating percentage of in-context example labels. To illustrate, a 100% label inversion for the SST-2 dataset implies that every "positive" exemplar is now labeled "negative". However, while we manipulate the in-context example labels, the evaluation sample labels remain consistent.

Results. Figure 1 shows the result of model performance across all datasets with respect to the proportion of labels that are flipped. As 0% label flips, we see larger language models have better in-context abilities. On the other hand, we observe that the override performance is more significant for language models with a larger scale. As the percentage of label alterations increases, which can be viewed as increasing label noise σ^2, the performance of small models remains flat and seldom is worse than random guessing while large models are easily affected by the noise, corresponding to a larger gap in Equation (23). These results indicate that large models can override their pretraining biases in-context input-label correlations, while small models may not and are robust to label noise. This observation aligns with the findings in [65] and our analysis in Section 4.2.

For tasks **RTE** and **WNLI**, whose patterns are less pronounced, we do not see a significant decrease curve for large models. A possible reason could be the inherent complexity of these tasks, which require predictions about sentence entailments, a challenge distinct from simpler sentiment classification or semantic equivalence tasks.

6 Conclusion

In this work, we answer our research question: why do larger language models do in-context learning differently? Our theoretical study shows that larger language models are easily overfitted to input noise and label noise during in-context learning, while smaller models are robust to noise, leading to different behaviors. Our empirical results support our claim and are consistent with previous work.

Acknowledgements

The work is partially supported by Air Force Grant FA9550-18-1-0166, the National Science Foundation (NSF) Grants 2008559-IIS, 2023239-DMS, and CCF-2046710.

References

[1] Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. In-context examples selection for machine translation. *arXiv preprint arXiv:2212.02437*, 2022.

[2] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*, 2023.

[3] Ekin Akyurek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*, 2023.

[4] Zeyuan Allen-Zhu and Yuanzhi Li. Physics of language models: Part 1, context-free grammar. *arXiv preprint arXiv:2305.13673*, 2023.

[5] Sanjeev Arora and Anirudh Goyal. A theory for emergence of complex skills in language models. *arXiv preprint arXiv:2307.15936*, 2023.

[6] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

[7] Srinadh Bhojanapalli, Chulhee Yun, Ankit Singh Rawat, Sashank Reddi, and Sanjiv Kumar. Low-rank bottleneck in multi-head attention models. In *International conference on machine learning*. PMLR, 2020.

[8] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[9] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 2020.

[10] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[11] Beidi Chen, Tri Dao, Eric Winsor, Zhao Song, Atri Rudra, and Christopher Ré. Scatterbrain: Unifying sparse and low-rank attention. *Advances in Neural Information Processing Systems*, 2021.

[12] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[13] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

[14] Alexis Conneau and Douwe Kiela. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), 2018.

[15] 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. *arXiv preprint arXiv:2212.10559*, 2022.

[16] Jyotikrishna Dass, Shang Wu, Huihong Shi, Chaojian Li, Zhifan Ye, Zhongfeng Wang, and Yingyan Lin. Vitality: Unifying low-rank and sparse approximation for vision transformer acceleration with a linear taylor attention. In *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)*. IEEE, 2023.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 2019.

[18] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, and Zhifang Sui. A survey for in-context learning. *arXiv preprint arXiv:2301.00234*, 2022.

[19] Xinyan Fan, Zheng Liu, Jianxun Lian, Wayne Xin Zhao, Xing Xie, and Ji-Rong Wen. Lighter and better: low-rank decomposed self-attention networks for next-item recommendation. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2021.

[20] Yao Fu, Hao Peng, Tushar Khot, and Mirella Lapata. Improving language model negotiation with self-play and in-context learning from ai feedback. *arXiv preprint arXiv:2305.10142*, 2023.

[21] Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, et al. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*, 2023.

[22] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[23] Tianyu Gao, Adam Fisch, and Danqi Chen. Making pre-trained language models better few-shot learners. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, 2021.

[24] Shivam Garg, Dimitris Tsipras, Percy S Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 2022.

[25] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[26] Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, et al. Opt-iml: Scaling language model instruction meta learning through the lens of generalization. *arXiv preprint arXiv:2212.12017*, 2022.

[27] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 2022.

[28] Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp. *arXiv preprint arXiv:2212.14024*, 2022.

[29] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2021.

[30] Hongkang Li, Meng Wang, Sijia Liu, and Pin-Yu Chen. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. In *The Eleventh International Conference on Learning Representations*, 2023.

[31] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*. Association for Computational Linguistics, 2021.

[32] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

[33] Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *Proceedings of the 40th International Conference on Machine Learning*, Proceedings of Machine Learning Research. PMLR, 2023.

[34] Zeping Luo, Shiyou Wu, Cindy Weng, Mo Zhou, and Rong Ge. Understanding the robustness of self-supervised learning through topic modeling. In *The Eleventh International Conference on Learning Representations*, 2023.

[35] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

[36] Sadhika Malladi, Tianyu Gao, Eshaan Nichani, Alex Damian, Jason D Lee, Danqi Chen, and Sanjeev Arora. Fine-tuning language models with just forward passes. *Advances in Neural Information Processing Systems*, 2023.

[37] JV Michalowicz, JM Nichols, F Bucholtz, and CC Olson. An isserlis' theorem for mixed gaussian variables: Application to the auto-bispectral density. *Journal of Statistical Physics*, 2009.

[38] Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. Metaicl: Learning to learn in context. *arXiv preprint arXiv:2110.15943*, 2021.

[39] 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? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022.

[40] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[41] Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*, 2021.

[42] OpenAI. Introducing ChatGPT. https://openai.com/blog/chatgpt, 2022. Accessed: 2023-09-10.

[43] OpenAI. GPT-4 technical report. *arXiv preprint arxiv:2303.08774*, 2023.

[44] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 2022.

[45] Jane Pan, Tianyu Gao, Howard Chen, and Danqi Chen. What in-context learning 'learns' in-context: Disentangling task recognition and task learning. In *Findings of Association for Computational Linguistics (ACL)*, 2023.

[46] Mohammadreza Pourreza and Davood Rafiei. Din-sql: Decomposed in-context learning of text-to-sql with self-correction. *arXiv preprint arXiv:2304.11015*, 2023.

[47] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *arXiv preprint arXiv:2306.15063*, 2023.

[48] Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*. PMLR, 2021.

[49] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. Large language models can be easily distracted by irrelevant context. In *International Conference on Machine Learning*. PMLR, 2023.

[50] Zhenmei Shi, Jiefeng Chen, Kunyang Li, Jayaram Raghuram, Xi Wu, Yingyu Liang, and Somesh Jha. The trade-off between universality and label efficiency of representations from contrastive learning. In *The Eleventh International Conference on Learning Representations*, 2023.

[51] Zhenmei Shi, Yifei Ming, Ying Fan, Frederic Sala, and Yingyu Liang. Domain generalization with nuclear norm regularization. In *NeurIPS 2022 Workshop on Distribution Shifts: Connecting Methods and Applications*, 2022.

[52] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *International Conference on Machine Learning*. PMLR, 2022.

[53] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 2023.

[54] Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Du. Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention, 2023.

[55] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[56] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[57] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 2017.

[58] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*. PMLR, 2023.

[59] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Association for Computational Linguistics, 2018.

[60] Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, et al. Super-naturalinstructions: Generalization via declarative instructions on 1600+ nlp tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, 2022.

[61] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2022.

[62] 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. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022.

[63] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 2022.

[64] Jerry Wei, Le Hou, Andrew Lampinen, Xiangning Chen, Da Huang, Yi Tay, Xinyun Chen, Yifeng Lu, Denny Zhou, Tengyu Ma, et al. Symbol tuning improves in-context learning in language models. *arXiv preprint arXiv:2305.08298*, 2023.

[65] Jerry Wei, Jason Wei, Yi Tay, Dustin Tran, Albert Webson, Yifeng Lu, Xinyun Chen, Hanxiao Liu, Da Huang, Denny Zhou, et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

[66] Gian-Carlo Wick. The evaluation of the collision matrix. *Physical review*, 1950.

[67] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022.

[68] Zhuoyan Xu, Zhenmei Shi, Junyi Wei, Yin Li, and Yingyu Liang. Improving foundation models for few-shot learning via multitask finetuning. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

[69] Renrui Zhang, Jiaming Han, Aojun Zhou, Xiangfei Hu, Shilin Yan, Pan Lu, Hongsheng Li, Peng Gao, and Yu Qiao. Llama-adapter: Efficient fine-tuning of language models with zero-init attention. *arXiv preprint arXiv:2303.16199*, 2023.

[70] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

[71] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*. PMLR, 2021.

[72] Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*, 2023.

[73] Hattie Zhou, Azade Nova, Hugo Larochelle, Aaron Courville, Behnam Neyshabur, and Hanie Sedghi. Teaching algorithmic reasoning via in-context learning. *arXiv preprint arXiv:2211.09066*, 2022.

Appendix

A Deferred Proof

A.1 Proof of Theorem 4.1

Here, we provide the proof of Theorem 4.1.

Theorem 4.1 (Optimal rank-r solution)**.** *Recall the loss function $\tilde{\ell}$ in Lemma 3. Let*

$$U^*, u^* = \underset{U \in \mathbb{R}^{d\times d}, \mathrm{rank}(U)\le r, u\in\mathbb{R}}{\operatorname{argmin}} \tilde{\ell}(U, u). \tag{8}$$

Then $U^* = cQV^*Q^\top, u = \frac{1}{c}$, where c is any non-zero constant and $V^* = \mathrm{diag}([v_1^*, \ldots, v_d^*])$ is satisfying for any $i \le r, v_i^* = \frac{N}{(N+1)\lambda_i + \mathrm{tr}(D)}$ and for any $i > r, v_i^* = 0$.

Proof of Theorem 4.1. Note that,

$$\begin{aligned}\underset{U \in \mathbb{R}^{d\times d}, \mathrm{rank}(U)\le r, u\in\mathbb{R}}{\operatorname{argmin}} \tilde{\ell}(U, u) &= \underset{U \in \mathbb{R}^{d\times d}, \mathrm{rank}(U)\le r, u\in\mathbb{R}}{\operatorname{argmin}} \tilde{\ell}(U, u) - \min_{U \in \mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U, u) &(24)\\ &= \underset{U \in \mathbb{R}^{d\times d}, \mathrm{rank}(U)\le r, u\in\mathbb{R}}{\operatorname{argmin}} \left(\tilde{\ell}(U, u) - \min_{U \in \mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U, u)\right). &(25)\end{aligned}$$

Thus, we may consider Equation (65) in Lemma 5 only. On the other hand, we have

$$\begin{aligned}\Gamma &= \left(1 + \frac{1}{N}\right)\Lambda + \frac{1}{N}\mathrm{tr}(\Lambda) I_{d\times d} &(26)\\ &= \left(1 + \frac{1}{N}\right) QDQ^\top + \frac{1}{N}\mathrm{tr}(D) Q I_{d\times d} Q^\top &(27)\\ &= Q\left(\left(1 + \frac{1}{N}\right) D + \frac{1}{N}\mathrm{tr}(D) I_{d\times d}\right) Q^\top. &(28)\end{aligned}$$

We denote $D' = \left(1 + \frac{1}{N}\right) D + \frac{1}{N}\mathrm{tr}(D) I_{d\times d}$. We can see $\Lambda^{\frac{1}{2}} = QD^{\frac{1}{2}}Q^\top$, $\Gamma^{\frac{1}{2}} = Q{D'}^{\frac{1}{2}}Q^\top$, and $\Gamma^{-1} = Q{D'}^{-1}Q^\top$. We denote $V = uQ^\top UQ$. Since Γ and Λ are commutable and the Frobenius norm (F-norm) of a matrix does not change after multiplying it by an orthonormal matrix, we have Equation (65) as

$$\begin{aligned}\tilde{\ell}(U, u) - \min_{U \in \mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U, u) &= \frac{1}{2}\left\|\Gamma^{\frac{1}{2}}\left(u\Lambda^{\frac{1}{2}}U\Lambda^{\frac{1}{2}} - \Lambda\Gamma^{-1}\right)\right\|_F^2 &(29)\\ &= \frac{1}{2}\left\|\Gamma^{\frac{1}{2}}\Lambda^{\frac{1}{2}}\left(uU - \Gamma^{-1}\right)\Lambda^{\frac{1}{2}}\right\|_F^2 &(30)\\ &= \frac{1}{2}\left\|{D'}^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - {D'}^{-1}\right)D^{\frac{1}{2}}\right\|_F^2. &(31)\end{aligned}$$

As W^{KQ} is a matrix whose rank is at most r, we have V is also at most rank r. Then, we denote $V^* = \operatorname{argmin}_{V \in \mathbb{R}^{d\times d}, \mathrm{rank}(V)\le r} \left\|{D'}^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - {D'}^{-1}\right)D^{\frac{1}{2}}\right\|_F^2$. We can see that V^* is a diagonal matrix. Denote $D' = \mathrm{diag}([\lambda_1', \ldots, \lambda_d'])$ and $V^* = \mathrm{diag}([v_1^*, \ldots, v_d^*])$. Then, we have

$$\begin{aligned}&\left\|{D'}^{\frac{1}{2}}D^{\frac{1}{2}}\left(V - {D'}^{-1}\right)D^{\frac{1}{2}}\right\|_F^2 &(32)\\ =&\sum_{i=1}^d \left({\lambda_i'}^{\frac{1}{2}}\lambda_i\left(v_i^* - \frac{1}{\lambda_i'}\right)\right)^2 &(33)\\ =&\sum_{i=1}^d \left(\left(1 + \frac{1}{N}\right)\lambda_i + \frac{\mathrm{tr}(D)}{N}\right)\lambda_i^2\left(v_i^* - \frac{1}{\left(1 + \frac{1}{N}\right)\lambda_i + \frac{\mathrm{tr}(D)}{N}}\right)^2. &(34)\end{aligned}$$

As V^* is the minimum rank r solution, we have that $v_i^* \geq 0$ for any $i \in [d]$ and if $v_i^* > 0$, we have $v_i^* = \frac{1}{(1+\frac{1}{N})\lambda_i + \frac{\mathrm{tr}(D)}{N}}$. Denote $g(x) = \left(\left(1+\frac{1}{N}\right)x + \frac{\mathrm{tr}(D)}{N}\right)x^2\left(\frac{1}{(1+\frac{1}{N})x+\frac{\mathrm{tr}(D)}{N}}\right)^2 = x^2\left(\frac{1}{(1+\frac{1}{N})x+\frac{\mathrm{tr}(D)}{N}}\right)$. It is easy to see that $g(x)$ is an increasing function on $[0, \infty)$. Now, we use contradiction to show that V^* only has non-zero entries in the first r diagonal entries. Suppose $i > r$, such that $v_i^* > 0$, then we must have $j \leq r$ such that $v_j^* = 0$ as V^* is a rank r solution. We find that if we set $v_i^* = 0, v_j^* = \frac{1}{(1+\frac{1}{N})\lambda_j + \frac{\mathrm{tr}(D)}{N}}$ and all other values remain the same, Equation (34) will strictly decrease as $g(x)$ is an increasing function on $[0, \infty)$. Thus, here is a contradiction. We finish the proof by $V^* = uQ^\top U^* Q$. □

A.2 Proof of Theorem 4.2

Here, we provide the proof of Theorem 4.2.

Theorem 4.2 (Evaluation loss)**.** *Recall $s, \xi \in \mathbb{R}^d$ are truncated and residual vectors respectively. Let $w = Q(s+\xi) \in \mathbb{R}^d$. Then for the optimal rank-r solution $f_{\mathrm{LSA},\theta}$ and V^* in Theorem 4.1 , we have evaluation population MSE loss*

$$\mathcal{L}(f_{\mathrm{LSA},\theta}; \widehat{E}) := \mathbb{E}_{x_1,\epsilon_1,\ldots,x_M,\epsilon_M,x_q}\left(f_{\mathrm{LSA},\theta}(\widehat{E}) - \langle w, x_q\rangle\right)^2 \tag{13}$$

$$= \frac{1}{M}\|s\|^2_{(V^*)^2D^3} + \frac{1}{M}\left(\|s+\xi\|_D^2 + \sigma^2\right)\mathrm{tr}\left((V^*)^2D^2\right) + \|\xi\|_D^2 + \sum_{i\in[r]} s_i^2\lambda_i\left(\lambda_i v_i^* - 1\right)^2. \tag{14}$$

Proof of Theorem 4.2. By Theorem 4.1, w.l.o.g, letting $c = 1$, the optimal rank-r solution $f_{\mathrm{LSA},\theta}$ satisfies $\theta = (W^{PV}, W^{KQ})$, and

$$W^{*PV} = \begin{pmatrix} 0_{d\times d} & 0_d \\ 0_d^\top & 1 \end{pmatrix}, W^{*KQ} = \begin{pmatrix} U^* & 0_d \\ 0_d^\top & 0 \end{pmatrix}, \tag{35}$$

where $U^* = QV^*Q^\top$.

We can see that U^* and Λ commute. Denote $\widehat{\Lambda} := \frac{1}{M}\sum_{i=1}^M x_i x_i^\top$. Note that we have

$$\widehat{y}_q = f_{\mathrm{LSA},\theta}(\widehat{E}) \tag{36}$$

$$= \begin{pmatrix} 0_{d\times d} & 0_d \\ 0_d^\top & 1 \end{pmatrix}\left(\frac{\widehat{E}\widehat{E}^\top}{M}\right)\begin{pmatrix} U^* & 0_d \\ 0_d^\top & 0 \end{pmatrix} x_q \tag{37}$$

$$= \begin{pmatrix} 0_{d\times d} & 0_d \\ 0_d^\top & 1 \end{pmatrix}\begin{pmatrix} \frac{1}{M}\left(x_q x_q^\top + \sum_{i=1}^M x_i x_i^\top\right) & \frac{1}{M}\left(\sum_{i=1}^M x_i x_i^\top w + \sum_{i=1}^M \epsilon_i x_i\right) \\ \frac{1}{M}\left(\sum_{i=1}^M w^\top x_i x_i^\top + \sum_{i=1}^M \epsilon_i x_i^\top\right) & \frac{1}{M}\sum_{i=1}^M (w^\top x_i + \epsilon_i)^2 \end{pmatrix}$$

$$\cdot \begin{pmatrix} U^* & 0_d \\ 0_d^\top & 0 \end{pmatrix} x_q \tag{38}$$

$$= \left(w^\top \widehat{\Lambda} + \frac{1}{M}\sum_{i=1}^M \epsilon_i x_i^\top\right) U^* x_q. \tag{39}$$

Then, we have

$$\mathbb{E}_{x_1,\epsilon_1,\ldots,x_M,\epsilon_M,x_q}\left(\widehat{y}_q - \langle w, x_q\rangle\right)^2 \tag{40}$$

$$= \mathbb{E}_{x_1,\epsilon_1,\ldots,x_M,\epsilon_M,x_q}\left(w^\top\widehat{\Lambda}U^*x_q + \frac{1}{M}\sum_{i=1}^M \epsilon_i x_i^\top U^* x_q - w^\top x_q\right)^2 \tag{41}$$

$$= \underbrace{\mathbb{E}\left[\left(w^\top\widehat{\Lambda}U^*x_q - w^\top x_q\right)^2\right]}_{\text{(I)}} + \underbrace{\mathbb{E}\left[\left(\frac{1}{M}\sum_{i=1}^M \epsilon_i x_i^\top U^* x_q\right)^2\right]}_{\text{(II)}}, \tag{42}$$

where the last equality is due to i.i.d. of ϵ_i. We see that the label noise can only have an effect in the second term. For the term (I) we have,

$$\text{(I)} = \mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q - w^\top \Lambda U^* x_q + w^\top \Lambda U^* x_q - w^\top x_q\right)^2\right] \tag{43}$$

$$= \underbrace{\mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q - w^\top \Lambda U^* x_q\right)^2\right]}_{\text{(III)}} + \underbrace{\mathbb{E}\left[\left(w^\top \Lambda U^* x_q - w^\top x_q\right)^2\right]}_{\text{(IV)}}, \tag{44}$$

where the last equality is due to $\mathbb{E}[\widehat{\Lambda}] = \Lambda$ and $\widehat{\Lambda}$ is independent with x_q. Note the fact that U^* and Λ commute. For the (III) term, we have

$$\text{(III)} = \mathbb{E}\left[\mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q\right)^2 + \left(w^\top \Lambda U^* x_q\right)^2 - 2\left(w^\top \widehat{\Lambda} U^* x_q\right)\left(w^\top \Lambda U^* x_q\right)\right]\Big| x_q\right] \tag{45}$$

$$= \mathbb{E}\left[\left(w^\top \widehat{\Lambda} U^* x_q\right)^2 - \left(w^\top \Lambda U^* x_q\right)^2\right]. \tag{46}$$

By the property of trace, we have,

$$\text{(III)} = \mathbb{E}\left[\operatorname{tr}\left(\widehat{\Lambda} w w^\top \widehat{\Lambda} (U^*)^2 \Lambda\right)\right] - \|w\|^2_{(U^*)^2\Lambda^3} \tag{47}$$

$$= \mathbb{E}\left[\frac{1}{M^2} \operatorname{tr}\left(\left(\sum_{i=1}^M x_i x_i^\top\right) w w^\top \left(\sum_{i=1}^M x_i x_i^\top\right) (U^*)^2 \Lambda\right)\right] - \|w\|^2_{(U^*)^2\Lambda^3} \tag{48}$$

$$= \mathbb{E}\left[\frac{M-1}{M} \operatorname{tr}\left(\Lambda w w^\top \Lambda (U^*)^2 \Lambda\right) + \frac{1}{M} \operatorname{tr}\left(x_1 x_1^\top w w^\top x_1 x_1^\top (U^*)^2 \Lambda\right)\right] - \|w\|^2_{(U^*)^2\Lambda^3} \tag{49}$$

$$= -\frac{1}{M}\|w\|^2_{(U^*)^2\Lambda^3} + \frac{1}{M}\mathbb{E}\left[\operatorname{tr}\left(x_1 x_1^\top w w^\top x_1 x_1^\top (U^*)^2 \Lambda\right)\right] \tag{50}$$

$$= -\frac{1}{M}\|w\|^2_{(U^*)^2\Lambda^3} + \frac{1}{M}\mathbb{E}\left[\operatorname{tr}\left(\left(\|w\|^2_\Lambda \Lambda + 2\Lambda w^\top w \Lambda\right)(U^*)^2 \Lambda\right)\right] \tag{51}$$

$$= \frac{1}{M}\|w\|^2_{(U^*)^2\Lambda^3} + \frac{1}{M}\|w\|^2_\Lambda \operatorname{tr}\left((U^*)^2 \Lambda^2\right), \tag{52}$$

where the third last equality is by Lemma 6. Furthermore, injecting $w = Q(s + \xi)$, as $\xi^\top V^*$ is a zero vector, we have

$$\text{(III)} = \frac{1}{M}\|s+\xi\|^2_{(V^*)^2 D^3} + \frac{1}{M}\|s+\xi\|^2_D \operatorname{tr}\left((V^*)^2 D^2\right) \tag{53}$$

$$= \frac{1}{M}\|s\|^2_{(V^*)^2 D^3} + \frac{1}{M}\|s+\xi\|^2_D \operatorname{tr}\left((V^*)^2 D^2\right). \tag{54}$$

Similarly, for the term (IV), we have

$$\text{(IV)} = \mathbb{E}\left[\left((s+\xi)^\top Q^\top \Lambda U^* x_q - (s+\xi)^\top Q^\top x_q\right)^2\right] \tag{55}$$

$$= \mathbb{E}\left[\left(s^\top D V^* Q^\top x_q - s^\top Q^\top x_q - \xi^\top Q^\top x_q\right)^2\right] \tag{56}$$

$$= s^\top (V^*)^2 D^3 s + s^\top D s + \xi^\top D \xi - 2 s^\top V^* D^2 s \tag{57}$$

$$= \xi^\top D \xi + \sum_{i\in[r]} s_i^2 \lambda_i \left(\lambda_i^2 (v_i^*)^2 - 2\lambda_i v_i^* + 1\right) \tag{58}$$

$$= \|\xi\|^2_D + \sum_{i\in[r]} s_i^2 \lambda_i \left(\lambda_i v_i^* - 1\right)^2, \tag{59}$$

where the third equality is due to $s^\top A \xi = 0$ for any diagonal matrix $A \in \mathbb{R}^{d\times d}$.

Now, we analyze the label noise term. By U^* and Λ being commutable, for the term (II), we have

$$\text{(II)} = \frac{\sigma^2}{M^2}\mathbb{E}\left[\left(\sum_{i=1}^{M} x_i^\top U^* x_q\right)^2\right] \tag{60}$$

$$= \frac{\sigma^2}{M^2}\mathbb{E}\left[\operatorname{tr}\left(\left(\sum_{i=1}^{M} x_i\right)^\top U^*\Lambda U^* \left(\sum_{i=1}^{M} x_i\right)\right)\right] \tag{61}$$

$$= \frac{\sigma^2}{M}\mathbb{E}\left[\operatorname{tr}\left(x_1^\top U^*\Lambda U^* x_1\right)\right] \tag{62}$$

$$= \frac{\sigma^2}{M}\operatorname{tr}\left((V^*)^2 D^2\right), \tag{63}$$

where all cross terms vanish in the second equality. We conclude by combining four terms. □

A.3 Auxiliary Lemma

Lemma 5 provides the structure of the quadratic form of our MSE loss.

Lemma 5 (Corollary A.2 in [70])**.** *The loss function $\tilde{\ell}$ in Lemma 3 satisfies*

$$\min_{U\in\mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U,u) = -\frac{1}{2}\operatorname{tr}[\Lambda^2\Gamma^{-1}], \tag{64}$$

where $U = c\Gamma^{-1}, u = \frac{1}{c}$ for any non-zero constant c are minimum solution. We also have

$$\tilde{\ell}(U,u) - \min_{U\in\mathbb{R}^{d\times d}, u\in\mathbb{R}} \tilde{\ell}(U,u) = \frac{1}{2}\left\|\Gamma^{\frac{1}{2}}\left(u\Lambda^{\frac{1}{2}}U\Lambda^{\frac{1}{2}} - \Lambda\Gamma^{-1}\right)\right\|_F^2. \tag{65}$$

Lemma 6. *Let $x \sim \mathcal{N}(0,\Lambda), \epsilon \sim \mathcal{N}(0,\sigma^2)$ and $y = \langle w, x\rangle + \epsilon$, where $w \in \mathbb{R}^d$ is a fixed vector. Then we have*

$$\mathbb{E}\left[y^2 x x^\top\right] = \sigma^2\Lambda + \|w\|_\Lambda^2\Lambda + 2\Lambda w^\top w\Lambda, \tag{66}$$

$$\mathbb{E}(yx)\mathbb{E}(yx)^\top = \Lambda^\top w w^\top \Lambda, \tag{67}$$

$$\mathbb{E}\left[(yx - \mathbb{E}(yx))(yx - \mathbb{E}(yx))^\top\right] = \sigma^2\Lambda + \|w\|_\Lambda^2\Lambda + \Lambda w^\top w\Lambda. \tag{68}$$

Proof of Lemma 6. As y is a zero mean Gaussian, by Isserlis' theorem [37, 66], for any $i, j \in [d]$ we have

$$\mathbb{E}[y^2 x_i x_j] = \mathbb{E}[y^2]\mathbb{E}[x_i x_j] + 2\mathbb{E}[y x_i]\mathbb{E}[y x_j] \tag{69}$$

$$= \left(\sigma^2 + w^\top\Lambda w\right)\Lambda_{i,j} + 2\Lambda_i^\top w w^\top \Lambda_j. \tag{70}$$

Thus, we have $\mathbb{E}\left[y^2 x x^\top\right] = \left(\sigma^2 + w^\top\Lambda w\right)\Lambda + 2\Lambda w^\top w\Lambda$. Similarly, we also have $\mathbb{E}(yx)\mathbb{E}(yx)^\top = \Lambda^\top w w^\top \Lambda$. Thus, we have

$$\mathbb{E}\left[(yx - \mathbb{E}(yx))(yx - \mathbb{E}(yx))^\top\right] \tag{71}$$

$$= \mathbb{E}\left[y^2 x x^\top - yx\mathbb{E}(yx)^\top - \mathbb{E}(yx) y x^\top + \mathbb{E}(yx)\mathbb{E}(yx)^\top\right] \tag{72}$$

$$= \mathbb{E}\left[y^2 x x^\top\right] - \mathbb{E}(yx)\mathbb{E}(yx)^\top \tag{73}$$

$$= \left(\sigma^2 + w^\top\Lambda w\right)\Lambda + \Lambda w^\top w\Lambda. \tag{74}$$

□