In-Context Learning under Distribution Shift: Optimal Attention Temperature for Transformers

Anonymous Author(s)
Affiliation
Address
email

Abstract

Pretrained Transformers exhibit strong in-context learning (ICL) capabilities, enabling them to perform new tasks from a few examples without parameter updates. However, their ICL performance often deteriorates under distribution shifts between pretraining and test-time data. Recent empirical work suggests that adjusting the attention temperature—a scaling factor in the softmax—can improve the performance of Transformers under such distribution shifts, yet its theoretical role remains poorly understood. In this work, we provide the first theoretical analysis of attention temperature in the context of ICL with pretrained Transformers. Focusing on a simplified setting with "linearized softmax" attention, we derive closed-form expressions for the generalization error under distribution shifts. Our analysis reveals that distributional changes in input covariance or label noise can significantly impair ICL, and that an optimal attention temperature exists which provably minimizes this error. We validate our theory through simulations on linear regression tasks and experiments with LLaMA2-7B on question-answering benchmarks. Our results establish attention temperature as a critical lever for robust in-context learning, offering both theoretical insight and practical guidance for tuning pretrained Transformers under distribution shift.

1 Introduction

Transformers [27] have become the cornerstone of modern AI systems, powering state-of-the-art models such as ChatGPT, Gemini, and DeepSeek. A key capability underlying their success is *in-context learning* (ICL)—the ability to adapt to new tasks directly from prompts, without modifying internal weights [4]. This emergent behavior has sparked significant interest in understanding the mechanisms behind ICL [2, 29], as well as how factors such as task diversity and model scale influence performance [30, 33].

Despite its promise, ICL remains sensitive to distribution shifts between pretraining and downstream tasks. Empirical and theoretical studies have shown that such shifts can degrade performance [35], raising critical questions about the robustness and adaptability of pretrained Transformers.

At the heart of the Transformer architecture lies the self-attention mechanism, formally expressed as

$$\text{Attention}(\boldsymbol{Z}) := \boldsymbol{V}\boldsymbol{Z} \cdot \text{softmax}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \tag{1}$$

where $\boldsymbol{Z}$ is the input, and $\boldsymbol{Q}$, $\boldsymbol{K}$, and $\boldsymbol{V}$ are the query, key, and value weight matrices, respectively. The parameter $\tau > 0$, known as the *attention temperature*, modulates the sharpness of the softmax distribution. While the original Transformer set $\tau = \sqrt{d_k}$ [27] where d_k is the dimension of the key matrix, later works in both NLP and vision have found that tuning or learning attention temperature can improve performance [16, 36, 21, 13, 5, 37].

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

Temperature controls how sharply attention weights focus on certain inputs—a property that could play a critical role under distribution shift. Surprisingly, despite its operational importance, the effect of temperature on the ICL behavior of pretrained Transformers has received little theoretical attention. This gap is particularly relevant in practice, where distribution mismatch between training and deployment is the norm.

This work — In this paper, we present a theoretical and empirical investigation of the attention temperature in the context of ICL. Our main focus is on how tuning temperature can improve the generalization performance of pretrained Transformers under distribution shifts. We study this question in the setting of linear regression tasks, which serve as a tractable framework for understanding ICL [9, 35]. Departing from prior work that considers linear attention, we analyze a Transformer with *linearized softmax* attention, which retains the essential temperature-dependent behavior of standard attention while allowing for mathematical tractability.

Our analysis identifies a closed-form expression for the *optimal temperature*—the value of τ that minimizes generalization error during inference. We show that this optimal temperature depends explicitly on the nature of the distribution shift, and that setting it appropriately can recover or even surpass baseline ICL performance. We validate our theoretical predictions through extensive experiments on both synthetic (linear regression) and real-world (question answering with LLMs) tasks, demonstrating that temperature tuning offers a simple yet powerful mechanism to improve robustness.

Contributions — Our work makes the following contributions:

1. We theoretically characterize the optimal attention temperature for pretrained Transformers with linearized softmax attention in in-context learning tasks.
2. We analyze the generalization behavior of such models under a broad range of distribution shifts, using a relaxed set of assumptions compared to prior work.
3. We establish a clear theoretical and empirical link between distribution shifts and temperature, showing that tuning temperature significantly enhances ICL performance across tasks.

Taken together, our results offer new insights into the interplay between temperature, distribution shift, and generalization in in-context learning, with implications for both theory and practice in the deployment of pretrained Transformers.

2 Related work

In-context learning — The ICL capability of Transformers was first brought to prominence by [4], leading to a surge of empirical and theoretical investigations. Several works have demonstrated that ICL performance improves with model scale [30, 19, 25], underscoring its importance in modern AI systems.

To better understand this phenomenon, synthetic tasks such as linear regression have served as controlled testbeds for analyzing ICL in Transformers [9, 35, 24]. A prevailing hypothesis in recent theoretical work is that Transformers implicitly learn algorithms during pretraining, which they subsequently execute during inference [3, 14, 2, 1, 29, 18, 7, 35, 15, 20]. However, there remains ongoing debate over the precise nature of these learned procedures.

In this context, simplified Transformer variants—particularly those using linear attention—have proven useful for gaining analytical insights. Notably, [35] showed that linear Transformers approximate Bayes-optimal inference in linear regression tasks, even under distribution shift.

Our work builds on this line of research but focuses specifically on the role of the temperature parameter in attention. Unlike [35], we (i) employ linearized softmax attention to isolate the influence of temperature, (ii) study how temperature adjustments can mitigate the effects of distribution shifts, and (iii) derive and evaluate the optimal temperature for improving ICL performance. These contributions extend prior analyses and provide a deeper understanding of how tuning temperature can enhance Transformer generalization under distributional variations.

Linear vs. softmax attention — A parallel thread of research investigates the comparative efficacy of linear and softmax attention mechanisms, which is directly relevant to our study since temperature is traditionally associated with softmax attention. While linear attention has gained popularity for its computational efficiency, it is often outperformed by softmax-based counterparts, prompting efforts to close this performance gap [6, 22].

A key development in this area is the work of [11], who demonstrated that a linearized variant of softmax attention can closely match the performance of standard softmax attention. Motivated by this finding, we adopt the "linearized softmax" formulation, allowing for tractable theoretical analysis while preserving the critical role of temperature. This approach facilitates a principled investigation of how temperature tuning impacts ICL in pretrained Transformers.

Temperature — Despite its central role in attention mechanisms, the temperature parameter remains underexplored in the context of ICL. Recent work by [28] proposes adaptive temperature as a means to sharpen softmax outputs, and temperature adjustments are sometimes reported in empirical studies of pretrained LLMs [26]. However, a systematic analysis of temperature's effect on ICL—particularly under distribution shift—has been lacking.

To address this gap, we provide a theoretical and empirical investigation of temperature within Transformers using "linearized softmax" attention. Our results clarify how the optimal temperature depends on the data distribution and how it can be tuned to reduce generalization error in in-context learning scenarios.

3 Setting

We describe the setup for analyzing ICL in linear regression using pretrained Transformers, covering the data model, linearized attention with reparameterization, evaluation metrics, and the Bayes-optimal benchmark.

Notation — We follow standard notation from [10]. The spectral norm of matrix $\boldsymbol{M}$ is denoted by $\|\boldsymbol{M}\|$, and the trace by $\mathrm{Tr}(\boldsymbol{M})$. Matrix entries and slices are denoted as $M_{i,j}$, $\boldsymbol{M}_{:,j}$, and $\boldsymbol{M}_{i,:}$.

3.1 Problem setup: In-context learning for linear regression

We study the ICL abilities of pretrained Transformers on linear regression tasks. Given a sequence of tokens, i.e., input-label pairs, $\{\boldsymbol{x}_1, y_1, \boldsymbol{x}_2, y_2, \ldots, \boldsymbol{x}_{l-1}, y_{l-1}, \boldsymbol{x}_l\}$, where each input vector $\boldsymbol{x}_i \in \mathbb{R}^d$ and corresponding label $y_i \in \mathbb{R}$ are independently sampled from an unknown joint distribution, the model must predict y_l using only the context $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{l-1}$ and the query $\boldsymbol{x}_l$, where $l-1$ is referred as the "context length". Each $(\boldsymbol{x}_i, y_i)$ pair is sampled i.i.d. from a joint distribution defined by:

$$\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \quad y_i = \boldsymbol{w}^T \boldsymbol{x}_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma^2), \tag{2}$$

where the task vector $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$ is fixed within a context but varies across tasks.

Assumption 3.1 (Well-Behaved Data Distributions)**.** There exist constants $c_1, c_2, c_3 > 0$ such that:

$$\|\boldsymbol{\mu}_x\|, \|\boldsymbol{\mu}_w\| \le c_1, \quad \lambda_{\min}(\boldsymbol{\Sigma}_x), \lambda_{\min}(\boldsymbol{\Sigma}_w) \ge c_2, \quad \lambda_{\max}(\boldsymbol{\Sigma}_x), \lambda_{\max}(\boldsymbol{\Sigma}_w) \le c_3.$$

This assumption ensures well-behaved distributions by bounding the means and covariances of input and task vectors, offering greater flexibility than the more restrictive setup in [35].

Assumption 3.2 (High-Dimensional Regime)**.** The context length l and input dimension d diverge jointly: $l, d \to \infty$.

This assumption reflects realistic settings where both context length and input dimension grow, aligning with modern ML trends and enabling analysis of generalization in high-dimensional regimes.

Under this set of assumptions, we define ICL for linear regression tasks as follows:

Definition 3.3 (In-Context Learning (ICL))**.** A model succeeds at ICL for linear regression if its generalization error nearly matches that of the Bayes-optimal linear model (defined in Section 3.6).

3.2 Modeling attention with transformers

Following the convention established in [35], we embed the input sequence into an embedding matrix:

$$\boldsymbol{Z} := \begin{bmatrix} \boldsymbol{x}_1 & \cdots & \boldsymbol{x}_{l-1} & \boldsymbol{x}_l \\ y_1 & \cdots & y_{l-1} & 0 \end{bmatrix} \in \mathbb{R}^{(d+1)\times l}, \tag{3}$$

where the last column corresponds to the query input with no label.

Using the embedding matrix, the softmax self-attention output is given by:

$$\boldsymbol{S} := \boldsymbol{Z} + \boldsymbol{V}\boldsymbol{Z} \cdot \text{softmax}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \tag{4}$$

where $\boldsymbol{K}$, $\boldsymbol{Q}$, and $\boldsymbol{V}$ are the key, query, and value matrices, respectively, and τ is the temperature. Here, we denote the model's prediction as $S_{d+1,l}$ — the last element in the final row.

3.3 Linearized attention approximation

To analytically study the effect of temperature on ICL, we adopt a linearized approximation of softmax attention (see Appendix B for the derivation):

$$\boldsymbol{E} := \boldsymbol{Z} + \frac{1}{l}\boldsymbol{V}\boldsymbol{Z}\left(\frac{(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau} + \mathbf{1} - \frac{\mathbf{1}}{l}\sum_{j=1}^{l}\frac{(\boldsymbol{K}\boldsymbol{Z}_{:,j})^T(\boldsymbol{Q}\boldsymbol{Z})}{\tau}\right), \tag{5}$$

where $\hat{y} := E_{d+1,l}$ is the predicted label. Unlike traditional linear attention (e.g., [35]):

$$\boldsymbol{Z} + \frac{1}{l}\boldsymbol{V}\boldsymbol{Z}(\boldsymbol{K}\boldsymbol{Z})^T(\boldsymbol{Q}\boldsymbol{Z}), \tag{6}$$

our linearized version preserves normalization properties, improving interpretability and robustness.

Remark 3.4 (Linear vs. Linearized Attention). Linearized attention preserves row-wise normalization, making it more robust to variations in input means — a key failure mode of linear attention in ICL. Appendix C illustrates this distinction.

3.4 Reparametrization of linearized attention

To streamline analysis, we reparametrize the matrices $\boldsymbol{V}$ and $\boldsymbol{M} := \boldsymbol{K}^T\boldsymbol{Q}$ as:

$$\boldsymbol{V} = \begin{bmatrix} * & * \\ \boldsymbol{v}_{21}^T & v_{22} \end{bmatrix}, \quad \boldsymbol{M} = \begin{bmatrix} \boldsymbol{M}_{11} & * \\ \boldsymbol{m}_{21}^T & * \end{bmatrix}, \tag{7}$$

where only $\boldsymbol{v}_{21}$, v_{22}, $\boldsymbol{m}_{21}$, and $\boldsymbol{M}_{11}$ influence the prediction $\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M})$. The remaining terms are denoted by $*$ as they are not relevant for predicting y_l in this context. The prediction from the linearized attention model can thus be expressed as a function of $\boldsymbol{M}$ and $\boldsymbol{V}$, i.e., $\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) := E_{d+1,l}$. This form parallels the approach in [35], allowing for tractable theoretical analysis.

By analyzing this reparameterization, we gain a deeper understanding of how the model parameters interact with the data to address the ICL problem effectively. This foundational insight will provide the necessary basis for discussing the pretraining of these parameters in Section 4.1.

3.5 Evaluating generalization performance

We focus on evaluating the performance of our attention model by assessing its generalization error. For a given set of parameters $(\boldsymbol{V}, \boldsymbol{M})$, the model's generalization (ICL) error is:

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) := \mathop{\mathbb{E}}_{(\boldsymbol{Z}, y_l)\sim\mathcal{D}^{test}}\left[(y_l - \hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}))^2\right], \tag{8}$$

where $\mathcal{D}^{test}$ denotes the distribution of the test set, which includes input-output pairs that the model has not encountered during training. In this context, the ICL task assesses the genuine ICL capabilities of the linearized attention module. Here, the task vectors in the test set differ from those encountered during training, requiring the model to infer these new vectors based solely on the provided context.

3.6 Bayes-optimal ridge estimator

The Bayes-optimal ridge estimator provides a robust framework for estimating the task vector $\boldsymbol{w}$ given a prior distribution and a set of $l-1$ samples. It is defined as:

$$\hat{\boldsymbol{w}}_{Bayes} = \left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{X}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\left(\frac{\bar{\boldsymbol{X}}^T\bar{\boldsymbol{y}}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\boldsymbol{\mu}_w\right), \tag{9}$$

where $\bar{\boldsymbol{X}}$ is the centered input matrix and $\bar{\boldsymbol{y}}$ is the centered label vector. This estimator integrates data information while incorporating prior beliefs about the distribution of $\boldsymbol{w}$, effectively balancing bias and variance, hence serves as the gold standard against which we compare model predictions. The terms including $\boldsymbol{\Sigma}_w^{-1}$ introduce a regularization effect, which is especially beneficial in high-dimensional settings.

The derivation of this estimator, detailed in Appendix A, illustrates how Bayesian principles can inform regression techniques by combining observed data with prior distributions to yield more reliable predictions. In our context, the inputs and labels originate from the prompt matrix $\boldsymbol{Z}$, and the prediction of the Bayes-optimal linear model for any input $\boldsymbol{x}$ is given by $\hat{\boldsymbol{w}}_{Bayes}^T \boldsymbol{x}$.

4 Theoretical results

In this section, we present our main theoretical results on the behavior of the linearized attention model in the context of ICL. We begin by showing how to pretrain the model to approximate the Bayes-optimal linear predictor, thereby grounding its predictive performance. We then identify specific conditions under which the model fails to generalize under distribution shifts at test time, revealing key limitations of linearized attention in ICL. Following this, we provide a detailed characterization of its generalization error, offering a principled framework for analyzing performance. Finally, we investigate the role of the temperature parameter and demonstrate that tuning it appropriately can substantially improve generalization—especially in cases where the model initially fails to perform effective in-context learning.

4.1 Model pretraining

We begin our pretraining analysis by observing that the prediction generated by the linearized attention model can be reduced to the following form (see Appendix D for the derivation):

$$\hat{y}(\boldsymbol{Z}; \boldsymbol{V}, \boldsymbol{M}) := E_{d+1,l} = \frac{1}{\tau} \hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V})^T \boldsymbol{x}_l + b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}), \tag{10}$$

where $\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}, C_{yy}; \boldsymbol{M}, \boldsymbol{V}) \in \mathbb{R}^d$ and $b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) \in \mathbb{R}$. $\boldsymbol{s}_x$ and s_y denote the sample means of the input $\boldsymbol{x}$ and the label y, respectively, and $\boldsymbol{C}_{xx}$ and $\boldsymbol{C}_{xy}$ are the sample covariances corresponding to $\mathrm{Cov}(\boldsymbol{x})$ and $\mathrm{Cov}(\boldsymbol{x}, y)$. These statistics are computed from the prompt matrix $\boldsymbol{Z}$.

For pretraining, we optimize the parameters $\boldsymbol{V}$ and $\boldsymbol{M}$ using m samples of $(\boldsymbol{Z}, y_l)$ drawn from the distribution $\mathcal{D}^{train}$, where each $\boldsymbol{Z}$ contains $l-1$ $(\boldsymbol{x}, y)$ pairs intended for ICL. Building upon prior work that connects ICL in linear regression to the Bayes-optimal ridge estimator [35, 24], we configure $\boldsymbol{M}$ and $\boldsymbol{V}$ to emulate Bayes-optimal ridge regression. Specifically, we aim for $\hat{\boldsymbol{w}}_{Att}(\boldsymbol{C}_{xx}, \boldsymbol{C}_{xy}; \boldsymbol{M}, \boldsymbol{V}) \approx \hat{\boldsymbol{w}}_{Bayes}$ and $b_{Att}(\boldsymbol{s}_x, s_y; \boldsymbol{V}) \approx 0$.

Lemma 4.1 (Pretrained Parameters)**.** *When the temperature parameter is set to $\tau = 1$ during pretraining, the following parameter configuration approximates the Bayes-optimal estimator in (9):*

$$\boldsymbol{M}_{11} = d \left(\frac{\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}}}{ml} + \frac{\sigma^2}{l} \boldsymbol{\Sigma}_w^{-1} \right)^{-1}, \qquad \boldsymbol{m}_{21} = \boldsymbol{0}, \tag{11}$$

$$\boldsymbol{v}_{21} = \frac{\sigma^2}{dl} \left(\frac{\hat{\boldsymbol{X}}^T \hat{\boldsymbol{X}}}{ml} \right)^{-1} \boldsymbol{\Sigma}_w^{-1} \boldsymbol{\mu}_w, \qquad v_{22} = \frac{1}{d},$$

where $\hat{\boldsymbol{X}} \in \mathbb{R}^{ml \times d}$ is the centered input matrix formed from ml samples of $\boldsymbol{x}$. This configuration aligns the linearized attention model with Bayes-optimal ridge regression. The quantities $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_w$ can be estimated from the pretraining data. A detailed derivation is provided in Appendix E.

This lemma establishes a theoretical connection between the pretrained parameters and the Bayes-optimal estimator, reinforcing the foundation of our approach.

Moreover, specific instances of Lemma 4.1 recover settings explored in prior studies. For example, under the assumptions $\boldsymbol{\Sigma}_x = \boldsymbol{\Sigma}_w = \boldsymbol{I}$, $\boldsymbol{\mu}_w = \boldsymbol{0}$, and $\sigma = 0$, [29] employ $\boldsymbol{M}_{11} = \mathrm{Cov}(x)^{-1}$ and $\boldsymbol{v}_{21} = \boldsymbol{0}$ within a linear attention framework. Our formulation generalizes this by allowing $\boldsymbol{v}_{21} \neq \boldsymbol{0}$, which reflects our assumption that $\boldsymbol{\mu}_w \neq \boldsymbol{0}$—a departure from earlier works. In our self-attention-based analysis, $\boldsymbol{v}_{21}$ encodes task vector bias. Additionally, our choice of $\boldsymbol{M}_{11}$ explicitly

accounts for label noise (σ^2), thereby enhancing the model's adaptability and maintaining a Bayesian interpretation.

We further comment on task diversity and parameter optimality in the following two remarks:

Remark 4.2. A high degree of task diversity (i.e., the number of distinct tasks) is crucial for enabling in-context learning [33]. In our framework, task diversity significantly affects the accuracy of estimating $\boldsymbol{\mu}_w$ and $\boldsymbol{\Sigma}_w$ during pretraining.

Remark 4.3. Although the pretrained parameters specified in Lemma 4.1 may not be optimal in all scenarios, they are analytically valuable for understanding the effects of distribution shifts and the influence of the temperature parameter in ICL. Notably, our characterization of ICL performance and temperature optimality does not rely on these specific parameter choices.

Based on Lemma 4.1, we arrive at the following corollary:

Corollary 4.4. *Suppose there is no distribution shift between training and inference. Then, under the parameter configuration of Lemma 4.1, the linearized attention model emulates the Bayes-optimal linear model, implying that it is capable of in-context learning according to Definition 3.3.*

Since the pretrained model succeeds in ICL for $\mathcal{D}^{test} = \mathcal{D}^{train}$, we next investigate how distribution shifts affect its ICL performance.

4.2 Effect of distribution shift

In this section, we explore scenarios where $\mathcal{D}^{test} \neq \mathcal{D}^{train}$, indicating a shift in the input, task, or noise distribution after pretraining the linearized attention model. We consider three cases: (1) a shift in the input distribution (altered mean or covariance), (2) a shift in the task distribution, and (3) a change in the noise levels.

To evaluate the impact of these distribution shifts on ICL performance, we assess whether adjustments to $\boldsymbol{M}$ and/or $\boldsymbol{V}$ are necessary to match the Bayes-optimal linear model under the new distribution. If so, the model is considered sensitive to the shift. Otherwise, it is deemed robust.

Case I: Shift in input distribution — Recall that inputs are drawn as $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, as defined in (2). Let $\boldsymbol{\mu}_x^{train}, \boldsymbol{\Sigma}_x^{train}$ and $\boldsymbol{\mu}_x^{test}, \boldsymbol{\Sigma}_x^{test}$ denote the input means and covariances for pretraining and testing, respectively. We consider two subcases:

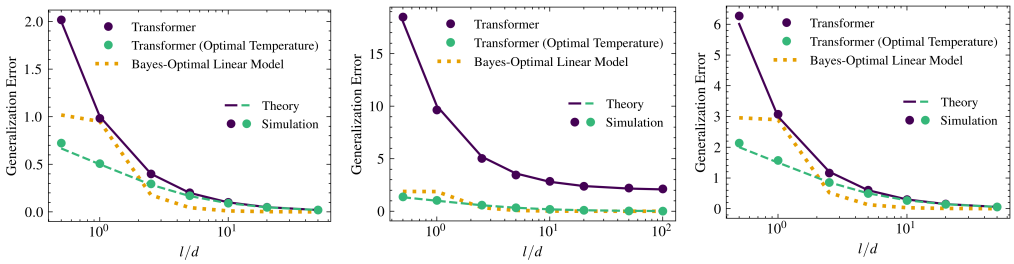
(i) **Shift in mean ($\boldsymbol{\mu}_x^{train} \neq \boldsymbol{\mu}_x^{test}$)**: The mean shift does not affect the linearized attention model since it uses centered inputs. However, this impacts the linear attention model, which operates on uncentered inputs, as discussed in Remark 3.4.

(ii) **Shift in covariance ($\boldsymbol{\Sigma}_x^{train} \neq \boldsymbol{\Sigma}_x^{test}$)**: A covariance shift necessitates retraining, as $\boldsymbol{M}_{11}$ is tailored to the pretraining input covariance. A mismatch leads to a significant deviation from the Bayes-optimal estimator, consistent with findings in prior work on linear attention [35].

Case II: Shift in Task Distribution — The task vectors follow $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$. Let $\boldsymbol{\mu}_w^{train}, \boldsymbol{\Sigma}_w^{train}$ and $\boldsymbol{\mu}_w^{test}, \boldsymbol{\Sigma}_w^{test}$ be the mean and covariance of the task distribution during pretraining and testing, respectively. The linearized attention model can incorporate $\boldsymbol{\mu}_w^{train}$ and $\boldsymbol{\Sigma}_w^{train}$ via the pretrained parameters $\boldsymbol{M}_{11}$ and $\boldsymbol{v}_{21}$ (see Lemma 4.1). However, as the context length l increases, the model's dependence on the task distribution diminishes. Thus, shifts in the task distribution primarily affect ICL performance for small l.

Case III: Shift in noise distribution — Finally, consider a change in the noise distribution: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, with σ_{train}^2 and σ_{test}^2 denoting pretraining and testing noise variances. If $\sigma_{train}^2 \neq \sigma_{test}^2$, the parameters $\boldsymbol{M}_{11}$ and $\boldsymbol{v}_{21}$ become suboptimal relative to the Bayes-optimal linear model. However, as with the task distribution, the impact of noise shift diminishes as $l \to \infty$.

Summary — The linearized attention model is robust to shifts in input mean but sensitive to input covariance changes. Shifts in task or noise distribution reduce ICL performance at small l, though increasing l mitigates these effects. In Section 4.4, we explore optimal temperature selection as a way to enhance robustness. First, we analyze the generalization error of the linearized attention model in the next section.



(a) Without Distribution Shift ($\mathcal{D}^{train} = \mathcal{D}^{test}$)

(b) With Shift in Input Covariance ($\boldsymbol{\Sigma}_{test} = 2\boldsymbol{\Sigma}_{train}$)

(c) With Shift in Task Distribution ($\boldsymbol{\mu}_w^{test} = \mathbf{1}/\sqrt{d}$, $\boldsymbol{\Sigma}_w^{test} = 3\boldsymbol{\Sigma}_w^{train}$)

Figure 1: Experiments with Transformer (Linearized Attention) on ICL under distribution shifts. Parameters are set using (11). Here, $d = 50$, $m = 5000$ (a new task per sample), $\sigma = 0.1$, $\boldsymbol{\mu}_x^{train} = \boldsymbol{\mu}_w^{train} = \mathbf{0}$, and $\boldsymbol{\Sigma}_x^{train} = \boldsymbol{\Sigma}_w^{train} = \boldsymbol{I}$.

4.3 In-context learning performance

We analyze the in-context learning (ICL) performance of the linearized attention model by evaluating the generalization error defined in (8). To establish a general setting for the subsequent results, we impose the following assumption on the pretrained parameters:

Assumption 4.5. There exists a constant $c > 0$ such that

$$\|\boldsymbol{M}_{11}\| \le cd, \quad \|\boldsymbol{m}_{21}\| = 0, \quad \|\boldsymbol{v}_{21}\| \le \frac{c}{dl}, \quad |v_{22}| \le \frac{c}{d}.$$

Note that the pretrained parameters obtained in Lemma 4.1 satisfy Assumption 4.5 with high probability under Assumptions 3.1–3.2. However, the generalization error result stated below holds for any parameters $\boldsymbol{M}, \boldsymbol{V}$ that satisfy Assumption 4.5.

Theorem 4.6 (Generalization error for ICL)**.** *Suppose Assumptions 3.1–3.2 and 4.5 hold. At test time, assume the input, task, and noise distributions are given by $\mathcal{N}(\boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x)$, $\mathcal{N}(\boldsymbol{\mu}_w, \boldsymbol{\Sigma}_w)$, and $\mathcal{N}(0, \sigma^2)$, respectively. Define*

$$\boldsymbol{A} := \boldsymbol{\Sigma}_x + \boldsymbol{\mu}_x\boldsymbol{\mu}_x^T, \quad \boldsymbol{B} := \boldsymbol{\Sigma}_w + \boldsymbol{\mu}_w\boldsymbol{\mu}_w^T.$$

Then, the generalization error is

$$\mathcal{G}(\boldsymbol{V}, \boldsymbol{M}) = \frac{1}{\tau^2}\operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right) - \frac{1}{\tau}\operatorname{Tr}\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right) + \operatorname{Tr}\left(\boldsymbol{A}\boldsymbol{B}\right) + \sigma^2, \quad (12)$$

where

$$\boldsymbol{F}_1 := \left(\boldsymbol{\Sigma}_x\hat{\boldsymbol{B}} + \frac{1}{l}\left(v_{22}^2\sigma^2 + \operatorname{Tr}(\hat{\boldsymbol{B}}\boldsymbol{\Sigma}_x)\right)\boldsymbol{I}\right)\boldsymbol{\Sigma}_x, \quad (13)$$

$$\boldsymbol{F}_2 := (\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{B})\boldsymbol{\Sigma}_x, \quad (14)$$

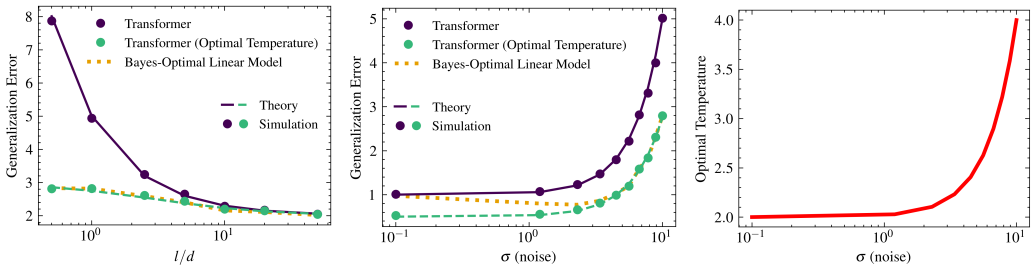
and

$$\hat{\boldsymbol{B}} := v_{22}\boldsymbol{\mu}_w\boldsymbol{v}_{21}^T + v_{22}\boldsymbol{v}_{21}\boldsymbol{\mu}_w^T + v_{22}^2\boldsymbol{B}.$$

Proof. The generalization error is derived using Isserlis' theorem [12] to compute higher-order moments. See Appendix F for the full derivation. □

Theorem 4.6 illustrates how the parameters $\boldsymbol{M}$, $\boldsymbol{V}$, and the test-time data distribution affect the generalization error. Notably, the temperature parameter τ plays a critical role.

Although temperature can be implicitly encoded in $\boldsymbol{M}$ during pretraining, it becomes especially important under distribution shifts that the model is not equipped to handle. In such cases, one can optimize generalization performance by choosing the temperature τ_{optimal} that minimizes the generalization error, as discussed next.



(a) Effect of l/d when $\sigma_{test} = 10$ (b) Effect of σ_{test} when $l/d = 1$ (c) Optimal temperature

Figure 2: Effect of noise shift on Transformer (Linearized Attention). The pretraining noise is $\sigma_{train} = 0.1$, while σ_{test} varies across plots. Panels (b) and (c) show generalization error and optimal temperature, respectively, as informed by Theorem 4.7. This setting matches Figure 1a, except for changes in test-time noise σ_{test}.

4.4 Optimal attention temperature improves performance

To address distribution shifts, we define the optimal attention temperature as follows:

Theorem 4.7 (Optimal attention temperature)**.** *Suppose Assumptions 3.1, 3.2, and 4.5 hold. To minimize the generalization error, the optimal attention temperature for inference is given by*

$$\tau_{optimal} = \frac{2Tr\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right)}{Tr\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right)}, \tag{15}$$

provided that $Tr\left(\boldsymbol{A}\left(\boldsymbol{F}_2\boldsymbol{M}_{11} + \boldsymbol{M}_{11}^T\boldsymbol{F}_2^T\right)\right) > 0$ *and* $Tr\left(\boldsymbol{A}\boldsymbol{M}_{11}^T\boldsymbol{F}_1\boldsymbol{M}_{11}\right) > 0$.

Proof. We minimize the generalization error from Theorem 4.6 with respect to τ (Appendix H). □

Consider the optimal temperature τ_{optimal} from Theorem 4.7. When $\tau_{\text{optimal}} \neq 1$, using an unadjusted temperature leads to suboptimal generalization error. Thus, incorporating the optimal temperature improves generalization in in-context learning under distribution shift.

A natural question is whether the optimal temperature can completely mitigate the adverse effects of distribution shifts. This depends on both the pretraining and test distributions. In some settings, the adjustment can entirely compensate for the shift. For example, if the task distribution is fixed as $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, the noise variance is $\sigma = 0$, and the input distribution changes from $\boldsymbol{x}_{\text{train}} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$ to $\boldsymbol{x}_{\text{test}} \sim \mathcal{N}(\boldsymbol{0}, c\boldsymbol{I})$, then the optimal temperature $\tau_{\text{optimal}} = c$ fully counteracts the shift. In more complex scenarios, it may only partially mitigate the impact, yet still yields improved ICL.

5 Experimental results

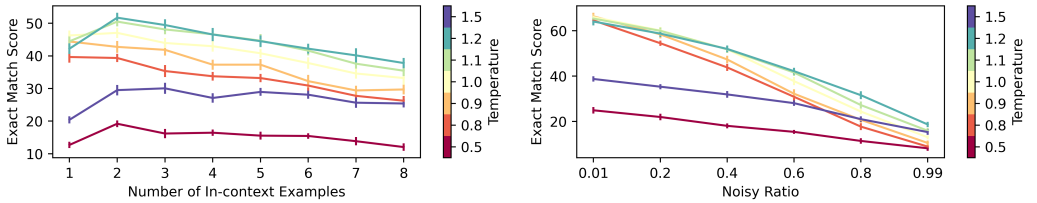
In this section, we empirically validate our theoretical predictions and demonstrate the impact of the optimal attention temperature on generalization. We begin with controlled experiments on linear regression tasks and progress to evaluating large-scale pretrained models on real-world datasets.

We experiment with two model classes on the linear regression tasks: (i) the linearized attention model, and (ii) the GPT-2 model [23], which incorporates multi-head softmax attention and MLP layers [1]. These experiments show that our theoretical insights generalize from simplified models to more expressive architectures. Finally, we examine the role of temperature in large language models (LLMs), using the Llama2-7B [26] on in-context learning tasks derived from SCIQ dataset [31].

5.1 Experiments on linear regression tasks

We consider a Transformer architecture with linearized attention and no MLP layers, as analyzed in our theoretical development. Figures 1 and 2 illustrate its behavior on linear regression tasks (2). Theoretical predictions closely match empirical performance across a range of conditions, confirming the robustness of our analysis. In Figure 1, we compare the ICL performance of the model with and without applying the optimal temperature. As context length l increases (Figure 1a), the

[1] Due to the space limitation, we provide the results with GPT-2 in the appendix.



(a) Effect of context length

(b) Effect of noisy ratio

Figure 3: LLM Experiments: The effect of the attention temperature on the ICL performance of the Llama2-7B model [26] using the SCIQ dataset [31]. A distribution shift is introduced by corrupting in-context demonstrations with noisy labels, selected as "relevant" but not necessarily correct answers following [8]. In (a), the noisy ratio is fixed at 0.6; in (b), the number of in-context examples is fixed at 6. Results are averaged over 20 Monte Carlo runs, with error bars indicating 0.33 standard deviations. Attention temperature in all layers is set to $\tau\sqrt{d_k}$, where d_k is the key dimension, to make τ values dimension-independent. Full details are provided in Appendix I.

model's predictions converge to those of the Bayes-optimal linear model, validating its ICL capability. Figure 1b shows that under an input covariance shift, model performance degrades—but applying the optimal temperature restores alignment with the Bayes-optimal solution. Additionally, Figure 1c shows that the influence of task distribution shift decreases as l increases.

We further evaluate robustness to label noise in Figure 2. In Figure 2a, we observe that noise effects diminish as the context length increases, consistent with our theoretical predictions. However, at small l, temperature adjustment becomes critical. In Figure 2b (for $l = d$), the Transformer increasingly diverges from the Bayes-optimal model as noise grows, yet optimal temperature correction closes this gap. Figure 2c shows that the optimal temperature increases with noise level, indicating a principled relationship between noise and temperature under limited context.

5.2 Experiments with LLMs for in-context question answering tasks

To assess the practical relevance of our theoretical framework, we investigate how attention temperature impacts the ICL behavior of LLMs. Following prior work [8], we use the SCIQ dataset [31] to create ICL tasks that incorporate distribution shift via noisy labels in the demonstrations. Examples of ICL prompts and the design of noisy labels are provided with full experimental details in Appendix I. We employ the Llama2-7B model [26], evaluating its ICL performance using the exact match score.

Figure 3 presents our results. In Figure 3a, we plot performance as a function of the number of in-context examples under a fixed noisy ratio. Due to the label noise, the performance curve exhibits non-monotonic behavior—highlighting the trade-off between additional context and accumulated noise. Figure 3b shows that as the proportion of noisy demonstrations increases, the optimal temperature also increases, aligning precisely with our theoretical expectations (cf. Figure 2c).

These results affirm that even for highly overparameterized practical models such as Llama2-7B, tuning the attention temperature serves as a principled and effective mechanism to mitigate the negative effects of distribution shifts on in-context learning.

6 Conclusion

This work provides a theoretical and empirical foundation for understanding the role of attention temperature in the in-context learning (ICL) capabilities of pretrained Transformers under distribution shifts. By introducing a simplified yet expressive framework based on linearized softmax attention, we analytically characterized how shifts in input covariance and label noise degrade ICL performance. Crucially, we identified and derived an optimal temperature that provably minimizes generalization error in these settings. Our theoretical predictions are validated through extensive experiments on both synthetic linear regression tasks and real-world benchmarks using GPT-2 and LLaMA2 models. Together, our findings offer actionable insights: tuning attention temperature is not merely a heuristic but a principled lever to enhance the robustness of ICL in pretrained Transformers. This advances our understanding of Transformer behavior under distribution shift and opens new directions for improving adaptability in large-scale models.

References

[1] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[2] Ekin Akyürek, 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.

[3] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[4] 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*, 33:1877–1901, 2020.

[5] Xiangyu Chen, Qinghao Hu, Kaidong Li, Cuncong Zhong, and Guanghui Wang. Accumulated trivial attention matters in vision transformers on small datasets. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 3984–3992, 2023.

[6] Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, David Benjamin Belanger, Lucy J Colwell, and Adrian Weller. Rethinking attention with performers. In *International Conference on Learning Representations*, 2021.

[7] Deqing Fu, Tianqi CHEN, Robin Jia, and Vatsal Sharan. Transformers learn higher-order optimization methods for in-context learning: A study with linear models, 2024.

[8] Hongfu Gao, Feipeng Zhang, Wenyu Jiang, Jun Shu, Feng Zheng, and Hongxin Wei. On the noise robustness of in-context learning for text generation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[9] 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*, 35:30583–30598, 2022.

[10] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[11] Dongchen Han, Yifan Pu, Zhuofan Xia, Yizeng Han, Xuran Pan, Xiu Li, Jiwen Lu, Shiji Song, and Gao Huang. Bridging the divide: Reconsidering softmax and linear attention. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[12] L. Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[13] Seung Hoon Lee, Seunghyun Lee, and Byung Cheol Song. Vision transformer for small-size datasets. *arXiv preprint arXiv:2112.13492*, 2021.

[14] Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *International Conference on Machine Learning*, pages 19565–19594. PMLR, 2023.

[15] Yingcong Li, Ankit Singh Rawat, and Samet Oymak. Fine-grained analysis of in-context linear estimation: Data, architecture, and beyond. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[16] Junyang Lin, Xu Sun, Xuancheng Ren, Muyu Li, and Qi Su. Learning when to concentrate or divert attention: Self-adaptive attention temperature for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2985–2990, 2018.

[17] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

[18] Arvind V. Mahankali, Tatsunori Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *The Twelfth International Conference on Learning Representations*, 2024.

[19] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

[20] Core Francisco Park, Ekdeep Singh Lubana, Itamar Pres, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. *arXiv preprint arXiv:2412.01003*, 2024.

[21] Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. YaRN: Efficient context window extension of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

[22] Zhen Qin, Weixuan Sun, Hui Deng, Dongxu Li, Yunshen Wei, Baohong Lv, Junjie Yan, Lingpeng Kong, and Yiran Zhong. cosformer: Rethinking softmax in attention. In *International Conference on Learning Representations*, 2022.

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

[24] Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. *Advances in Neural Information Processing Systems*, 36, 2024.

[25] Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[26] 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.

[27] 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.

[28] Petar Veličković, Christos Perivolaropoulos, Federico Barbero, and Razvan Pascanu. softmax is not enough (for sharp out-of-distribution). *arXiv preprint arXiv:2410.01104*, 2024.

[29] Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 23–29 Jul 2023.

[30] 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.

[31] Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pages 94–106, 2017.

[32] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics.

[33] Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *The Twelfth International Conference on Learning Representations*, 2024.

[34] Zhenyu Wu, YaoXiang Wang, Jiacheng Ye, Jiangtao Feng, Jingjing Xu, Yu Qiao, and Zhiyong Wu. Openicl: An open-source framework for in-context learning. *arXiv preprint arXiv:2303.02913*, 2023.

[35] Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[36] Shengqiang Zhang, Xingxing Zhang, Hangbo Bao, and Furu Wei. Attention temperature matters in abstractive summarization distillation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 127–141, 2022.

[37] Yixiong Zou, Ran Ma, Yuhua Li, and Ruixuan Li. Attention temperature matters in vit-based cross-domain few-shot learning. In *Neural Information Processing Systems*, 2024.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed when introducing the setting.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: All assumptions are given when introducing the setting, while all the proofs are given in the appendix.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The experimental results are reproducible as all the details are provided.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code will be released with the camera-ready version of the paper.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The experimental details for each result are given in the corresponding captions.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We report error bars when applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: While most of our experiments can be conducted on any modern computer, the LLM experiments (Figure 3) should be run on a GPU, and the relevant computer resources are mentioned in the appendix when providing the details for the experiment.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research conducted in the paper conforms with the Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: As a theory paper, it is not expected to have any direct societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not provide new data or new model.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The creators/owners of the used assets are properly credited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.