DIFFERENTIAL LEARNING KINETICS GOVERN THE TRANSITION FROM MEMORIZATION TO GENERALIZATION DURING IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformers exhibit in-context learning (ICL): the ability to use novel information presented in the context without additional weight updates. Recent work shows that ICL emerges when models are trained on a sufficiently diverse set of tasks and the transition from memorization to generalization is sharp with increasing task diversity. One interpretation is that a network's limited capacity to memorize favors generalization. Here, we examine the mechanistic underpinnings of this transition using a small transformer applied to a synthetic ICL task. Using theory and experiment, we show that the sub-circuits that memorize and generalize can be viewed as largely independent. The relative *rates* at which these sub-circuits learn explains the transition from memorization to generalization, rather than capacity constraints. We uncover a memorization scaling law, which determines the task diversity threshold at which the network generalizes. The theory quantitatively explains a variety of other ICL-related phenomena, including the long-tailed distribution of when ICL is acquired, the bimodal behavior of solutions close to the task diversity threshold, the influence of contextual and data distributional statistics on ICL, and the transient nature of ICL.

1 INTRODUCTION

Large transformer models trained to predict the next token exhibit powerful generalization capabilities. One signature of such generalization capabilities is in-context learning (ICL): the ability to solve a task based on new information presented in the context without additional weight updates (Brown et al. (2020); Dai et al. (2022); Dong et al. (2022); Garg et al. (2022); Xie et al. (2021); Olsson et al. (2022)). Arguably, the ability to interpret novel inputs on-the-fly is a core feature of any intelligent system. However, updating synaptic weights on rapid behavioral timescales is challenging, both for natural and artificial systems. The emergence of ICL in large language models (LLMs) shows that finding network states that learn on-the-fly is indeed possible. Understanding how ICL emerges in LLMs promises insights into how such algorithms may be implemented in the brain and how the data distribution, training objective and network architecture interact to enable ICL acquisition at scale.

Various methods have been used to probe the ICL capabilities of LLMs (Brown et al. (2020); Dong et al. (2022); Pan (2023); Min et al. (2022); Olsson et al. (2022)). A common ICL paradigm is to present exemplars as a sequence of item-label pairs, and measure the network's response to a target item (Chan et al. (2022); Kirsch et al. (2022); Garg et al. (2022); Akyürek et al. (2022); Von Oswald et al. (2023); Raventós et al. (2023); Bai et al. (2023)). While LLMs display remarkable capabilities on such ICL tasks, interpreting the underlying network mechanisms that give rise to these capabilities remains challenging (but see Wang et al. (2022)). Recent work has approached this challenge by examining how small transformer models solve synthetic ICL tasks (Reddy (2023); Bietti et al. (2024); Akyürek et al. (2022); Ahn et al. (2023); Von Oswald et al. (2023); Edelman et al. (2024)). We highlight two notable aspects of ICL phenomenology relevant for our current work: the influence of task diversity on whether the network memorizes a finite dataset or acquires ICL (i.e., generalizes), and how ICL is acquired (and lost) during training.

First, data distributional properties (such as task diversity and their rank-frequency distribution) influence whether the network acquires ICL or encodes the response to queries seen during training within its weights (Kirsch et al. (2022); Chan et al. (2022); Raventós et al. (2023). Following previous work, we refer to such memorization as in-weights learning (IWL). Notably, the transition from memorization (IWL) to generalization (ICL) is sharp with respect to task diversity. A curious feature of this transition is that solutions close to the task diversity threshold are bimodal (Kirsch et al. (2022)). That is, across different random number seeds, solutions either show IWL or acquire ICL but intermediate solutions are unlikely.

Second, ICL is often implemented by multi-layer computations involving nonlinear attention heads and MLPs (Olsson et al. (2022); Von Oswald et al. (2023); Reddy (2023); Bietti et al. (2024)). The rugged loss landscape induced by such multi-layer, nonlinear operations leads to long plateaus followed by a sharp drop in loss (Reddy (2023)). Finally, ICL is seemingly transient, i.e., the network gradually loses the ICL capability if it is trained for sufficiently long (Singh et al. (2023)).

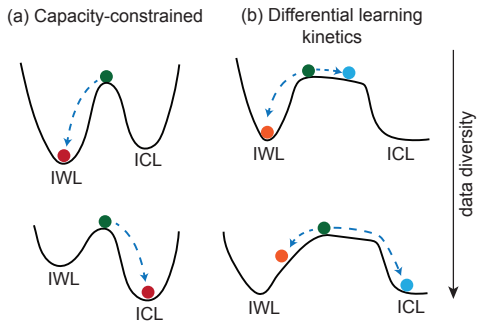

Figure 1: (a) In the capacity-constrained model, the network's limited capacity to memorize favors ICL acquisition with increasing task diversity. (b) In the differential learning kinetics model, independent sub-circuits contribute towards IWL and ICL. IWL is slower for greater task diversity. The network acquires ICL before the network can significantly memorize the training set. IWL is significantly slowed down as ICL explains most of the loss, but does eventually memorize the training set. The network subsequently loses the ICL capability due to regularization.

It is unclear what leads to the transition from IWL to ICL acquisition with increasing task diversity. The loss is minimized when a finite training dataset is perfectly memorized. Thus, a naive hypothesis would suggest that as the task diversity increases, the network's limited capacity to memorize favors the ICL solution (Figure 1a). Our goal is to test this hypothesis by deriving a precise quantitative description of the transition, and thus we set out to identify a minimal setting that captures the phenomenon.

Contributions and outline. We first identify a one-layer transformer model trained on an in-context classification task that recapitulates the sharp transition from memorization to generalization. Despite its simplicity, the one-layer model displays surprisingly rich phenomenology, including abrupt ICL learning dynamics, transience and bimodal solutions close to the task diversity threshold.

Next, we derive an analytical framework that quantitatively characterizes ICL acquisition and its competition with IWL. We show that the transition from IWL to ICL for our network is governed by a *dynamical competition* between memorization and generalization (Figure 1b). However, whether the network is capacity-constrained or rate-determined depends on the network architecture, and we derive a quantitative measure to determine which of these constraints is at play. The theory predicts that the number of iterations before ICL is acquired is exponentially sensitive to the initial parameters, which in turn explains the bimodal behavior of solutions close to the task diversity threshold. The task diversity threshold follows a power law whose exponent has a non-trivial relationship with another novel memorization scaling law. ICL transience naturally follows from the theory under standard L_2 regularization. Finally, we validate our theory by empirically verifying these predictions using our transformer model.

2 Task formulation

We consider a simplified version of an ICL task proposed by Chan et al. (2022), which allows for disentangling ICL and IWL performance (Figure 2a). Before training, we generate a dataset $\mathcal{D}$ that contains K item-label pairs, $\mathcal{D} = \{(x_1, \ell_1), (x_2, \ell_2), \ldots, (x_K, \ell_K)\}$. Each item x_i is a D-

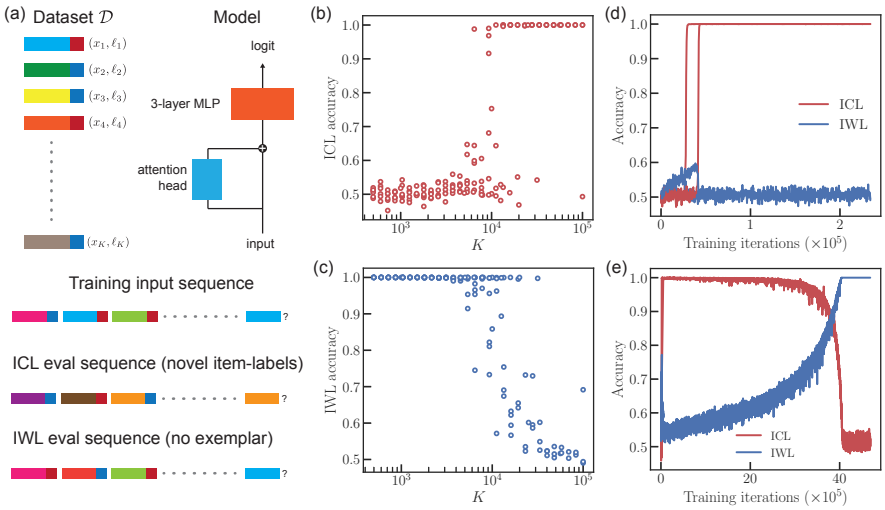

Figure 2: (a) Data generation process: We create a dataset $\mathcal{D}$ consisting of K item-label (x_i, ℓ_i) pairs where each $x_i \sim \mathcal{N}(0, 1/D)$ and ℓ_i is randomly sampled from $\{-1, +1\}$. The network receives a sequence of $N+1$ tokens. Each of the first N tokens is a concatenation $x_i \oplus \ell_i$ of an (x_i, ℓ_i) pair sampled uniformly from $\mathcal{D}$ (details in main-text). The final $N+1$ target token consists of only the item x_i as its label component is zero-ed out. The network is trained to correctly predict the label of the last item. Model architecture: The input is normalized using LayerNorm, then fed to our network, which consists of a one-layer attention network followed by a 3-layer MLP. (b) ICL performance demonstrates a sharp transition as a function of data diversity K. Further, at the transition threshold $K^* \approx 10^4$, we observe bimodality where the model either memorizes or generalizes. (c) IWL accuracy vs K. IWL and ICL accuracies follow opposite trends. (d) ICL accuracy curves show that ICL performance plateaus at the beginning of training but undergoes a rapid transition as ICL is acquired. (e) ICL is transient, i.e., ICL accuracy gradually decreases to chance levels when the parameters in the attention head are heavily regularized.

dimensional random vector with components drawn i.i.d from $\mathcal{N}(0, 1/D)$ and is randomly assigned one of two labels, $\ell_i \in \{-1, +1\}$.

The data is presented to the network as a sequence of $N+1$ tokens, where each token t_j is the item concatenated with its label, $t_j = x_j \oplus \ell_j$. The first N tokens in the sequence are drawn uniformly from $\mathcal{D}$. The $N+1$th token (the target token) has an empty label vector. The target token is chosen uniformly randomly from the N items in the context, so that there is an exemplar always present in the context. Given an input sequence, the network is trained to predict the label of the target token using a binary cross-entropy loss. Since the total number of item-label pairs (K) is finite, the network can either memorize each item's label (IWL), or it can learn to use the exemplar within the context to predict the correct label (ICL).

To measure ICL, we construct a test dataset $\mathcal{D}_{\text{test}}$ consisting of novel item-label pairs (sampled like $\mathcal{D}$) and evaluate the network's accuracy on sequences sampled from $\mathcal{D}_{\text{test}}$ using the previously described procedure. To measure IWL, we evaluate the network on sequences sampled from $\mathcal{D}$, except that the target has no corresponding exemplar in the sequence. In this case, the context has no useful information, and the network must rely on the label's information encoded within its weights.

3 Results

3.1 A one-layer transformer model recapitulates ICL phenomenology

We begin with a one-layer attention-based network followed by a multi-layer perceptron (MLP). Given tokens $t_1, t_2, \ldots, t_{N+1}$, we first apply a LayerNorm operation to obtain $t'_i = \text{LayerNorm}(t_i)$.

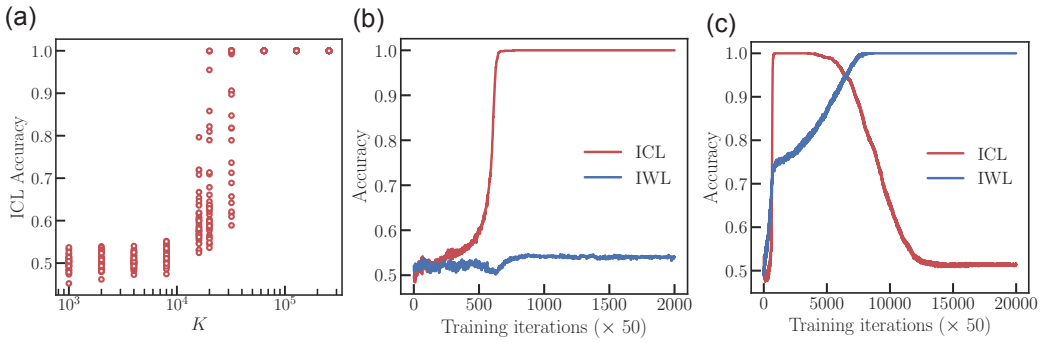

Figure 3: Phenomenology of the minimal model. (a) ICL performance in the minimal model demonstrates a sharp transition as a function of data diversity K. (b) ICL acquisition is abrupt during training. (c) ICL is transient when w is exclusively regularized.

The attention operation then produces the output

$$u = t'_{N+1} + \sum_{j=1}^{N+1} \frac{e^{t_j'^T K^T Q t'_{N+1}}}{\sum_k e^{t_k'^T K^T Q t'_{N+1}}} V t'_j, \tag{1}$$

where Q, K, V are $(D+2) \times (D+2)$ query, key and value matrices (ℓ_j's are one-hot vectors for the one-layer model, and we use $\ell_j = \pm 1$ elsewhere). The MLP ϕ takes input u of dimension $D+2$ and produces logits $\phi(u)$. If the target's true label is $(1, 0)$, the loss is $-\log \sigma(\phi(u))$, and if the target's true label is $(0, 1)$, the loss is $-\log \sigma(-\phi(u))$, where σ is the logistic function. The MLP is a three-layer ReLU network with hidden dimension d. We use stochastic gradient descent (SGD) with batch size 128, learning rate 0.01, weight-decay 10^{-10}, $D = 63$, $N = 100$ and $d = 512$, unless otherwise specified.

We examine how ICL accuracy scales with the number of item-label pairs, K. For $K \ll K^*$, where $K^* \sim 10^4$, we find that the network shows IWL but fails to acquire ICL (Figure 2b,c). Conversely, for $K \gg K^*$, the network consistently acquires ICL but does not show IWL. The transition from IWL to ICL is sharp. Solutions in the vicinity of $K = K^*$ are bimodal (highlighted in Figure 2b and quantified in greater detail later). Examining the ICL accuracy during learning shows an abrupt transition from chance-level to perfect accuracy (Figure 2d). Previous work has also shown that ICL accuracy gradually decays to zero if training is continued for a sufficiently large number of iterations (Singh et al. (2023)). Transience is recapitulated when we significantly extend the number of training iterations (Figure 2e), provided the parameters within the attention head are regularized more heavily than those in the MLP. This skewed regularization scheme hints at a potential explanation for what causes transience (as noted in Singh et al. (2023)), which we will later explain quantitatively using our theoretical framework.

3.2 Disentangling ICL and IWL in a minimal model

Figure 2(c-f) shows that, despite its simplicity, the one-layer transformer model captures the core features of the memorization to generalization transition and ICL training dynamics observed in more complex models. However, a mechanistic analysis is still challenging due to the nonlinearities in the attention head, the MLP and how these two operations interact. To make progress, we further reduce our one-layer transformer model into a disentangled model (which we refer to as the "minimal model" hereafter) by proposing two ansatz. We will show empirically that the minimal model also reproduces the phenomena in Figure 2(c-f). This minimal model is amenable to a theoretical analysis and leads to specific quantitative predictions. We then validate our ansatz by empirically testing these predictions using our original transformer model.

To motivate the ansatz, we observe that ICL in this task involves a simple match-to-sample operation implemented by the attention head. The attention paid by the target token is determined by its dot-product similarity with the content (the first D dimensions) of the tokens in the context. The value matrix reads the labels of the tokens weighted by the attention paid to those tokens and passes it on

to the MLP. Put together, these observations suggest that the relevant operations performed by the query-key product and the value matrix are captured by

$$K^TQ = \begin{pmatrix} \beta I_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} \end{pmatrix}, \quad V = \begin{pmatrix} 0_{D\times D} & 0 \\ 0 & w \end{pmatrix} \tag{2}$$

where $I_{D\times D}$ is a D-dimensional identity matrix, β, w are (learnable) scalars and the rest of the components are zeroes.

From equation 1, the MLP receives the sum of the target token t_{N+1} and the output of the attention head as its input. Note that all the information required by the MLP to memorize the label of the target is present in t_{N+1}. That is, IWL does not require the output of the attention head. Similarly, all the information required to predict the target's label using ICL is contained in the output of the attention head. Based on these observations, we posit that the final logit used to predict the target's label is the *sum* of logits generated independently by an MLP (which takes x_{N+1} as input) and the attention head (using the simplified K^TQ and V matrices in equation 2). Specifically, given an input sequence $t_1, t_2, \ldots, t_{N+1}$, we assume the MLP ϕ produces a logit $z_{\text{MLP}} = \phi(x_{N+1})$ and the attention head produces the logit

$$z_{\text{ATT}} = \sum_{j=1}^{N} \frac{e^{\beta x_j^T x_{N+1}}}{\sum_{k=1}^{N} e^{\beta x_k^T x_{N+1}}} w\ell_j, \tag{3}$$

Put together, the predicted probability that the target's label is $+1$ is given by $\sigma(z_{\text{MLP}} + z_{\text{ATT}})$. The binary cross-entropy loss for a given input sequence is then $-\log\sigma(\ell_t(z_{\text{MLP}} + z_{\text{ATT}}))$, where $\ell_t = \pm 1$ is the true label of the target.

In summary, we assume an *independence* ansatz, where the attention head and MLP perform ICL and IWL respectively, and additively contribute towards the prediction of the target's label. Further, we assume that the majority of the ICL learning dynamics is captured by reducing the K^TQ and V matrices to two "order parameters", β and w. That the *strengths* of the relevant attention operations determine ICL acquisition is our second ansatz.

The minimal model is parameterized by β, w and the parameters of the MLP (a three-layer ReLU network of hidden dimension 512). The model is trained and evaluated using the same procedures used on the transformer model. Figure 3(a-c) show that the three phenomena of interest, that is, the transition from memorization to generalization, abrupt ICL learning and ICL transience, are reproduced by the minimal model.

3.3 The loss landscape of the minimal model

We now use the minimal model to derive an approximate expression for the loss. We consider the asymptotic limit $K \gg N \gg 1$ and the infinite-dimensional limit $D \to \infty$ (recall, $K > 10^3, N = 10^2, D = 63$ in our experiments). From equation 3, ICL is acquired when $w, \beta \gg 1$. Our goal is compute the time taken for the network to acquire ICL starting from $w = w_0, \beta = \beta_0$ with $|w_0|, |\beta_0| \ll 1$.

In the limit $D \to \infty$, the dot product $x_j^T x_{N+1}$ is 1 if x_j is a copy of the target and 0 otherwise. It is unlikely there is more than one copy of the target in the context when $K \gg N$. Let c denote the index of this copy. From equation 3, we have

$$z_{\text{ATT}} \approx w\left(\frac{e^\beta}{e^\beta + N - 1}\ell_c + \frac{1}{e^\beta + N - 1}\left(2n_+ - N - \ell_c\right)\right), \tag{4}$$

where n_+ is the number of tokens with label $+1$ amongst the N tokens in the context. Next, the MLP's contribution to the average loss appears only through the *distribution* of logits obtained by applying the MLP to each of the K items in $\mathcal{D}$. In particular, denote P^+ as the distribution of MLP logits of the items with a $+1$ label. Since the two labels are symmetric, the average loss $\mathcal{L}$ can be written as

$$\mathcal{L} \approx -2\sum_{n_+=0}^{N} \frac{n_+}{N} B(n_+) \left\langle \log\sigma\left(\phi^+ + w\left(\frac{e^\beta}{e^\beta + N - 1} + \frac{1}{e^\beta + N - 1}\left(2n_+ - N - 1\right)\right)\right)\right\rangle_{\phi^+\sim P^+}$$

where $B(n_+)$ is the distribution of the number of $+1$ labels in the context (a binomial distribution in our experiments) and the prefactor n_+/N is the probability that the target is $+1$ given n_+. The angled brackets denote an expectation over the MLP logits ϕ^+ drawn from P^+. When $N \gg 1$, n_+/N is well-approximated by a normal random variable, $n_+/N \approx 1/2 + \eta/2\sqrt{N}$, where $\eta \sim \mathcal{N}(0,1)$. We are left with

$$\mathcal{L} \approx -\left\langle \left(1 + \frac{\eta}{\sqrt{N}}\right) \log \sigma \left(\phi^+ + w\left(\frac{e^\beta - 1}{e^\beta + N - 1} + \frac{\eta\sqrt{N}}{e^\beta + N - 1}\right)\right)\right\rangle_{\eta\sim\mathcal{N}(0,1),\phi^+\sim P^+} . \quad (5)$$

The loss landscape throughout training is thus specified by P^+, w and β. Retaining the fluctuations in η is necessary to accurately describe ICL acquisition.

3.4 THE DYNAMICS OF ICL ACQUISITION

To examine the dynamics of ICL acquisition, we find an expression for the loss when $|w| \ll \sqrt{N}$ and $e^\beta - 1 \ll N$ (for arbitrary P^+). Both these conditions are satisfied at initialization ($|w_0|, |\beta_0| \ll 1$). From equation 5, a few steps of simplification leads to

$$\mathcal{L} \approx \left\langle \left(1 + \frac{\eta}{\sqrt{N}}\right) \log \left(1 + e^{-\phi^+ - w\left(\frac{e^\beta - 1}{N} + \frac{\eta}{\sqrt{N}}\right)}\right)\right\rangle_{\eta,\phi^+} , \quad (6)$$

$$\approx \left\langle \log(1 + e^{-\phi^+})\right\rangle_{\phi^+} - \frac{c_1}{N}\left(e^\beta w - \frac{c_2 w^2}{2}\right), \quad (7)$$

where $c_1 \equiv \langle \sigma(-\phi^+)\rangle_{\phi^+}$ and $c_2 \equiv 1 - \langle \sigma(-\phi^+)^2\rangle_{\phi^+}/c_1$. Here, we used $|w| \ll \sqrt{N}$ and $e^\beta - 1 \ll N$ to Taylor expand the $\log$ term in equation 6 and retained terms to order $1/N$.

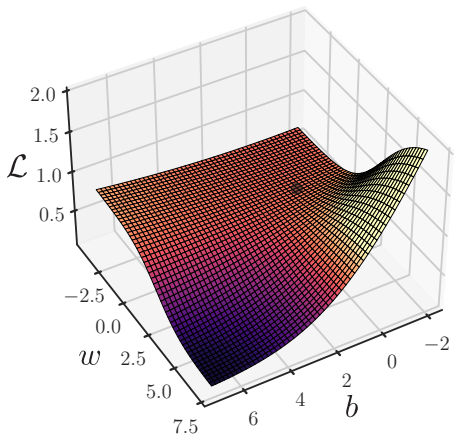

Figure 4: The approximate loss landscape $\mathcal{L}$ in the minimal model as a function of the key parameters β, w exhibits a nearly flat region close to initialization, but the dynamics always leads to ICL acquisition ($w, \beta \gg 1$).

Equation 7 allows us to make several important inferences. The first term in the r.h.s of equation 7 is the loss incurred by the MLP. It does not involve w, β and thus does not affect ICL learning dynamics. Since the second term in the r.h.s of equation 7 is small at initialization, the rate at which the MLP memorizes is not affected by ICL learning. That is, IWL proceeds without any competition from ICL until ICL is acquired (which happens abruptly).

The scalar variables c_1 and c_2 depend on P^+ and thus depend on the time t since training began. Their evolution in general depends on multiple factors, including MLP architecture, initialization scheme and the number of tasks K to be memorized. Importantly, IWL influences ICL acquisition only through $c_1(t)$ and $c_2(t)$, which in turn depend only on how the MLP memorizes class labels. We proceed with our analysis by retaining $c_1(t)$ and $c_2(t)$ as yet-to-be-determined time-dependent variables, keeping in mind that their dependence on t and K will play an important role in our analysis further below.

Gradient descent dynamics over the loss in equation 7 gives

$$\frac{dw}{dt} = \frac{c_1}{N}\left(e^\beta - c_2 w\right), \quad (8)$$

$$\frac{d\beta}{dt} = \frac{c_1}{N}\left(w e^\beta\right). \quad (9)$$

Learning initially proceeds at a slow rate c_1/N (since $N \gg 1$ and $0 < c_1 < 1$). Since ϕ^+ on average increases as the MLP memorizes, c_1 decreases and slows down ICL acquisition. If the MLP (near) perfectly memorizes the K item-label pairs before ICL is acquired, then $c_1 \approx 0$ and ICL is never

acquired. In other words, the loss "explained away" due to MLP memorization creates an effective competition between IWL and ICL acquisition despite the additive contributions of the MLP and the attention head to the logit.

Since $0 < c_2 < 1$, equation 8 shows that w eventually converges from its initial value to a positive value $w = e^{\beta}/c_2$. β increases monotonically when w is positive until ICL is acquired. Thus, equation 8 and equation 9 imply that ICL will always be acquired, however slowly, if the MLP is unable to perfectly memorize the K item-label pairs (i.e., $c_1(\infty) > 0$). However, the choice of label statistics in the context (i.e., $B(n_+)$) matters. For example, consider the case when N is even and there are exactly $N/2$ tokens with $+1$ and -1 labels in the context. To compute the mean loss $\mathcal{L}'$ in this scenario, we take $\eta = 0$ in equation 6 and Taylor expand w.r.t $(e^{\beta} - 1)/N$ to get

$$\mathcal{L}' \approx \left\langle \log(1 + e^{-\phi^+}) \right\rangle_{\phi^+} - \frac{c_1}{N} w \left(e^{\beta} - 1\right). \tag{10}$$

The origin is a saddle point. The parameters either flow to the ICL solution ($w, \beta > 0$) or to a suboptimal solution ($w, \beta < 0$) depending on the initial values w_0, β_0. That is, ICL acquisition is not guaranteed.

3.5 Exponential dependence of t_{ICL} on initial conditions

Equation 8 and equation 9 allow us to estimate the number of iterations t_{ICL} it takes to acquire ICL. We fix $w_0 = 0$. We distinguish three cases: β_0 large ($e^{\beta_0} \gg 1$), $-\beta_0$ large ($e^{-\beta_0} \gg 1$) and $|\beta_0|$ small ($|\beta_0| \ll 1$). The first case is trivial as ICL is acquired immediately.

When $|\beta_0| \ll 1$, since $e^{\beta} \approx 1$, the w term in equation 8 can be ignored, $dw/dt \approx (c_1/N)e^{\beta}$. Dividing this equation by equation 9 and integrating, we get the conservation equation $w^2/2 - \beta = w_0^2/2 - \beta_0 = -\beta_0$, so that $w = \sqrt{2(\beta - \beta_0)}$. Substituting this expression for w into equation 9, we have

$$\frac{d\beta}{dt} \approx \frac{c_1\sqrt{2(\beta - \beta_0)}}{N} e^{\beta}. \tag{11}$$

Re-arranging and integrating from $\beta = \beta_0$ to $\beta = \infty$, we get

$$N\sqrt{\frac{\pi}{2}} e^{-\beta_0} \approx I_K(t_{\text{ICL}}), \quad \text{where } I_K(t) = \int_0^t c_1(t')dt'. \tag{12}$$

The subscript in I_K is used to highlight that c_1 depends on K. When $K = \infty$, the MLP cannot memorize and ϕ_+ is distributed symmetrically around 0 so that $c_1(t), c_2(t) \approx 1/2$ throughout training. In that case, we obtain a linear dependence on N and exponential dependence on β_0,

$$t_{\text{ICL}} \approx N\sqrt{2\pi} e^{-\beta_0}. \tag{13}$$

The dynamics are qualitatively different when $-\beta_0$ is large and $e^{\beta_0} \ll 1$. In this case, we obtain $t_{\text{ICL}} \approx Ne^{-2\beta_0}$ as $K \to \infty$ (see Appendix A.3). We numerically verify the exponential dependence of t_{ICL} on the initial values of β_0 (Supplementary Figure A.1a). A consequence of this exponential dependence is that normal-distributed values of β_0 will lead to a long-tailed distribution of t_{ICL}. In pictorial terms, due to the nearly flat loss landscape close to initialization (Figure 4), small variation in the initial parameters leads to large variation in when ICL is acquired.

3.6 Memorization scaling laws and the transition from memorization to generalization

Equation 12 shows that the behavior of an MLP-specific quantity, $c_1(t)$ (via I_K), determines when ICL is acquired. A corollary to equation 12 is that there are two possibilities: either $I_K(t)$ diverges as $t \to \infty$ so that ICL is always acquired irrespective of K, or $I_K(\infty)$ has a finite value. In the latter case, ICL is acquired when $I_K(\infty) > N\sqrt{\pi/2}e^{-\beta_0}$.

The divergence of I_K as $t \to \infty$ may occur either because the MLP has limited capacity to memorize the K samples or because of the data distribution. For example, if the rank-frequency distribution of item-label pairs follows a Zipf's law $p(f) \sim f^{-\alpha}$ with exponent $\alpha \leq 1$, then the MLP's loss is dominated by rare item-label pairs that are not memorized and lead to I_K divergence. While previous work has shown that such skewed data distributions favor ICL acquisition (Chan et al. (2022)), our theory suggests that such data dependencies can be examined by simply measuring I_K.

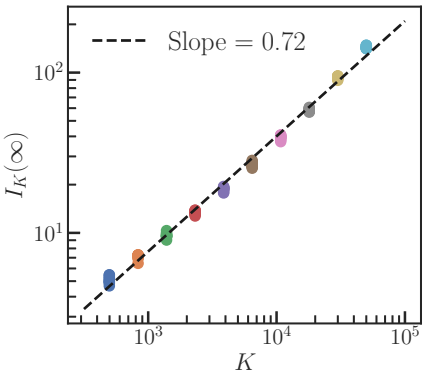

Figure 5: $I_K(\infty)$ shows a power-law scaling with K with exponent $\nu \approx 0.72$.

We examine the behavior of I_K for a uniform distribution over item-label pairs. To our knowledge, current deep learning theory does not inform how the distribution of logits (or summary statistics such as c_1) scales with t and K for typical MLP architectures. To make progress, we empirically measure $c_1(t), c_2(t)$ for different K using an independent set of MLP experiments (Supplementary Figure A.2). We find that $c_1(t)$ has a universal tail behavior $c_1(t) \sim t^{-\mu}$, where $\mu \approx 1.25$ is independent of K for K between 10^3 and 10^5 (See Supplementary Figure A.3a). The limit $I_K(\infty)$ is thus finite and our MLP is not capacity constrained. Further, we uncover a scaling law, $I_K(\infty) \sim K^\nu$, where $\nu \approx 0.72$ (Figure 5). From equation 12, this scaling law leads to an estimate for the task diversity threshold K^* for the transition from memorization to generalization,

$$K^* \sim N^{1/\nu} e^{-\beta_0/\nu}, \tag{14}$$

up to a constant prefactor. Note the exponential dependence on β_0, whose random initialization implies that there is a range of K for which the network will either show memorization or generalization. The theory further predicts a non-trivial power-law relation between the task diversity threshold and the context length.

3.7 Slow IWL explains transient ICL

Our theory naturally explains ICL transience once we introduce regularization into our model. In practice, this regularization may be explicit or implicit. For simplicity, we assume an L_2 regularization on w. We return to equation 5 for the loss, which applies throughout training. Since $w, \beta \gg 1$ after ICL is acquired, we can ignore the terms involving η in equation 5. We have

$$\mathcal{L} \approx -\left\langle \log \sigma \left(\phi^+ + w \left(\frac{e^\beta - 1}{e^\beta + N - 1} \right) \right) \right\rangle_{\phi^+} + \frac{\lambda_w w^2}{2}. \tag{15}$$

$$\approx \left\langle e^{-\phi^+} \right\rangle_{\phi^+} e^{-w} + \frac{\lambda_w w^2}{2} \tag{16}$$

where λ_w is the L_2 regularization parameter and we have used $\frac{e^\beta - 1}{e^\beta + N - 1} \approx 1$, $e^{-(\phi^+ + w)} \ll 1$ and $\log(1+x) \approx x$ for $x \ll 1$. Once ICL is acquired, memorization slows down drastically due to the small factor e^{-w}. However, this slow IWL gradually decreases w until ICL fades away entirely. Equation 16 implies that w after ICL acquisition tracks w_{tr}, where

$$w_{\text{tr}}(t) \approx W\left(c_3(t)/\lambda_w\right), \quad c_3(t) = \left\langle e^{-\phi^+} \right\rangle_{\phi^+}. \tag{17}$$

The Lambert W function $W(x)$ is monotonic in x when x is positive. Thus, w_{tr} decreases with IWL, which in turn decreases c_3 until w_{tr} decays to zero when the dataset is perfectly memorized ($c_3 = 0$). Thus, the analysis suggests that extremely slow IWL after the network acquires ICL leads to slow decay of ICL due to regularization. We note however that in more complex models the effects of a global regularization parameter on different sub-circuits are hard to disentangle.

Equation 17 hints at a relationship between the ICL loss $\mathcal{L}_{\text{ICL}}$ and the IWL loss $\mathcal{L}_{\text{IWL}}$ after ICL is acquired. From our independence ansatz, $\mathcal{L}_{\text{IWL}}$ is given by $\mathcal{L}_{\text{IWL}} = -\left\langle \log \sigma\left(\phi^+ + \psi\right)\right\rangle_{\phi^+}$. After ICL acquisition, $|\psi| \gg |\phi^+|$ and $|\psi| \gg 1$. In the IWL testing paradigm, since there is no copy of the target in the context, the network attends to a random token and gets the target's label right with probability 1/2. This implies that IWL loss is large when ψ has opposite sign as the target's label and is negligible when it has the same sign. We can then approximate

$$\mathcal{L}_{\text{IWL}} \approx \frac{1}{2}\log(1 + e^{|\psi|}) \approx \frac{1}{2}|\psi| \approx -\frac{1}{2}\log(\mathcal{L}_{\text{ICL}}), \tag{18}$$

whenever $\mathcal{L}_{\text{ICL}} \ll 1$. A similar argument shows that whenever $\mathcal{L}_{\text{IWL}} \ll 1$, we have $\mathcal{L}_{\text{ICL}} \approx -\frac{1}{2}\log(\mathcal{L}_{\text{IWL}})$. These approximate relations between $\mathcal{L}_{\text{ICL}}$ and $\mathcal{L}_{\text{IWL}}$ are generic consequences of the independence ansatz. If the independence ansatz is valid, the theory predicts that these relations should hold from the moment ICL is acquired until it fades due to gradual IWL.

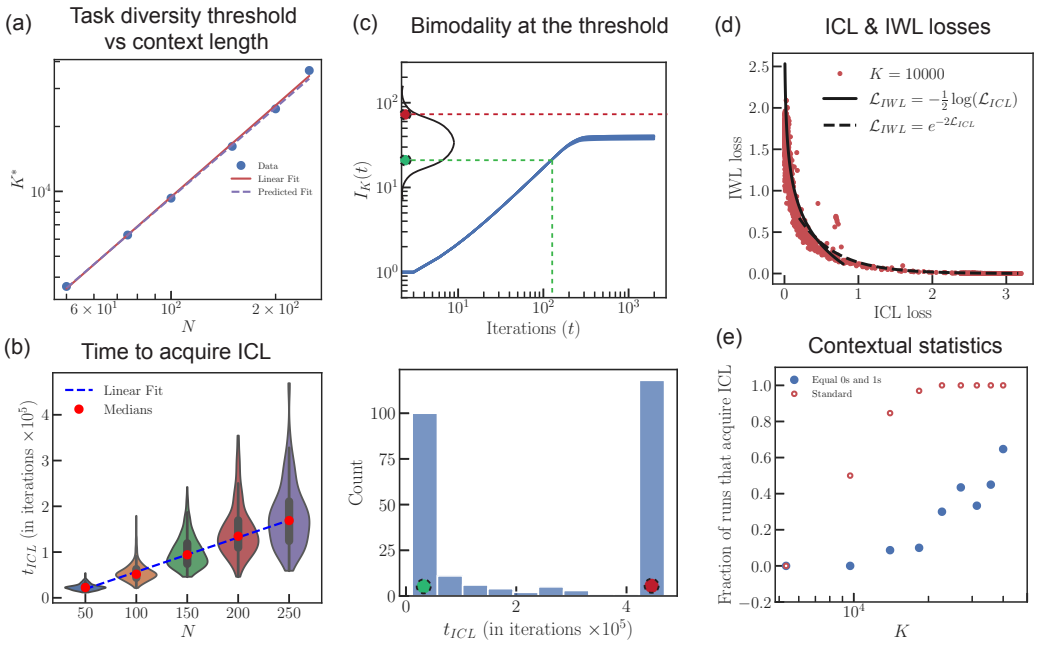

Figure 6: (a) The critical task diversity threshold K^* exhibits a power law relationship with respect to N. The experimentally determined critical exponent (linear fit) closely matches our predicted critical exponent $1/\nu$ where $\nu \approx 0.72$ (predicted fit). (b) The median time taken to acquire ICL (t_{ICL}) scales linearly as a function of N and the distribution of t_{ICL} is long-tailed. For clearer visualization of the full distribution, see Supplementary Figure A.4. (c) Close to $K = K^*$, depending on the initialization seed (red and green dots), the network will either generalize (green) or not (red). Different seeds are predicted to show either small t_{ICL} or do not acquire ICL, with few intermediates. A histogram of t_{ICL} for $K = 9666$ confirms this prediction at the critical task diversity threshold K^*. Note that the maximum number of iterations is $\approx 4.5 \times 10^5$. (d) In runs that exhibit transience, we observe a relationship between ICL loss and IWL loss after the model has acquired ICL, closely matching our predicted functional relationship. (e) We measured the fraction of solutions that acquired ICL as we vary K, in a new dataset $\mathcal{D}'$ where every sequence is constrained to contain exactly $N/2$ items of each label. We observe that the critical task diversity threshold K^* is greatly increased and that many more solutions fail to achieve ICL compared to models trained on our standard dataset $\mathcal{D}$.

4 Empirical Validation

The theory makes a number of new quantitative predictions related to ICL acquisition. We empirically test six nontrivial predictions that span various aspects of ICL phenomenology using the original transformer model in equation 1.

Power-law scaling of the task diversity threshold with context length. Equation 14 predicts a highly non-trivial power law relationship between the task diversity threshold K^* and N. To test this prediction, we train our original transformer model (equation 1) at varying N and K. At each N, we observe a sharp transition from memorization to generalization as K increases (Supplementary Figure A.5. For each N, we determine K^* by fitting a sigmoidal curve to ICL performance as a function of K (See Supplementary Figure A.5). As $\nu \approx 0.72$, Equation 14 predicts an exponent of ≈ 1.39, closely matching our measured exponent ≈ 1.41 (Figure 6a).

Linear scaling of the time taken to acquire ICL with context length. Equation 13 predicts that t_{ICL} scales linearly with N. To test this, we train our original transformer model (equation 1) at varying N and take the limit $K \to \infty$, by resampling our dataset $\mathcal{D}$ at every training iteration. We then determine t_{ICL} (time taken to acquire ICL) as the epoch at which ICL performance exceeds 95%. We train approximately 100 seeds for each N to obtain the full distribution of t_{ICL}. Figure 6b, confirms a linear relationship between the median t_{ICL} as a function of N.

MLP memorization governs the dependence of ICL on data distributional properties. Bimodality in solutions at the transition threshold can be explained by the fact that the threshold $I_K(\infty)$ required for ICL acquisition exhibits an exponential dependence on (random) β_0. Then, in solutions where $I_K(\infty)$ converges to a value below the threshold $I_K(\infty)$ required for ICL, the network fails to acquire ICL (Figure 6c).

Long-tailed distribution of the time taken to acquire ICL. Equation 39 predicts a long-tailed distribution of t_{ICL}. We verify this prediction in histograms of t_{ICL} for each N, using the same data as in Figure 6b (See Supplementary Figure A.4).

Interdependence of the ICL loss and IWL loss after ICL acquisition. Equation 18 predicts a non-trivial relationship between the ICL loss $\mathcal{L}_{\text{ICL}}$ and the IWL loss $\mathcal{L}_{\text{IWL}}$ after ICL is acquired. In runs in which we observed transience, we plot IWL performance as a function of ICL performance and observe a close match between equation 18 and data (Figure 6d).

Failure of ICL acquisition when every sequence contains $N/2$ tokens of each label. Equation 10 predicts that the network is more likely to fail to acquire ICL when there are exactly $N/2$ tokens with $+1$ and -1 labels in the context. We test this prediction in a dataset $\mathcal{D}'$ with $N = 100$, where each training sequence (including the target token) viewed by the one-layer transformer has $N/2$ tokens of each class in its context. For each K, we trained the model using at least 20 seeds. We measure the probability of acquiring ICL as the fraction of seeds whose final ICL performance exceeds 75%. We observe much lower probability of acquiring ICL for models trained on $\mathcal{D}'$ compared to those trained on our standard $\mathcal{D}$ (Figure 6e). Moreover, the loss curves across seeds are diverse, with many seeds not acquiring ICL and some saturating at sub-optimal solutions (See Supplementary Figure A.6).

5 Conclusion

Here, we propose a theory based on the ansatz that the network contains sub-circuits that are independently involved in memorization and generalization. A trade-off arises simply because the rate at which one sub-circuit is optimized depends on how much loss is already explained by the other sub-circuit(s). Building on this theory, we show that the transition from memorization to generalization in our model is determined by the relative rates at which these sub-circuits memorize and generalize. However, the theory does not rule out the possibility that capacity constraints play a role in other scenarios.

This ansatz, despite being cast in the context of a simplified one-layer model, explains a surprising variety of ICL-related phenomena observed with much larger models. These include a long-tailed distribution in when ICL is acquired, an MLP memorization scaling law, the bimodality of solutions at the task diversity threshold, the transient nature of ICL, amongst other novel quantitative relations that our theory identifies. The two most striking predictions are (1) the non-trivial relationship between an MLP-specific memorization scaling law ($I_K(\infty) \sim K^\nu$) and a task diversity threshold scaling law w.r.t context length ($K^* \sim N^{1/\nu}$), and (2) the long-tailed distribution of when ICL is acquired and its linear scaling with context length ($t_{\text{ICL}} \sim N$). Both these predictions have been validated in our experiments.

Our results offer some hope that seemingly intractable phenomena observed in large models can be reproduced and analyzed using simpler, tractable models through careful experimental design. However, further work is necessary to examine to what extent such insights provided by small models remain valid for larger models (that potentially contain many sub-circuits) and for more naturalistic tasks (where a clear distinction between memorization and generalization cannot be made) (Min et al. (2022); Wei et al. (2023); Pan (2023); Shi et al. (2024)).

References

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.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

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.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36, 2024.

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.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems*, 35:18878–18891, 2022.

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.

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.

Benjamin L Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

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.

Louis Kirsch, James Harrison, Jascha Sohl-Dickstein, and Luke Metz. General-purpose in-context learning by meta-learning transformers. *arXiv preprint arXiv:2212.04458*, 2022.

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? *arXiv preprint arXiv:2202.12837*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Jane Pan. What in-context learning "learns" in-context: Disentangling task recognition and task learning. Master's thesis, Princeton University, 2023.

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.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. *arXiv preprint arXiv:2312.03002*, 2023.

Zhenmei Shi, Junyi Wei, Zhuoyan Xu, and Yingyu Liang. Why larger language models do in-context learning differently?, 2024. URL `https://arxiv.org/abs/2405.19592`.

Aaditya K Singh, Stephanie CY Chan, Ted Moskovitz, Erin Grant, Andrew M Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. *arXiv preprint arXiv:2311.08360*, 2023.

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*, pp. 35151–35174. PMLR, 2023.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*, 2022.

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.

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*, 2021.

A Appendix

We present a detailed analysis of the minimal model outlined in the main text. Suppose the $(D+1)$-dimensional input tokens are $t_1, t_2, \ldots, t_{N+1}$, where $t_i = (x_i, \ell_i)$ for $i \leq N$ and $t_{N+1} = (x_{N+1}, 0)$. In the minimal model, we consider two logits $z_{\text{MLP}} = \phi(x_{N+1})$ and

$$z_{\text{ATT}} = \sum_{j=1}^{N} \frac{e^{\beta x_j^T x_{N+1}}}{\sum_{k=1}^{N} e^{\beta x_k^T x_{N+1}}} w\ell_j. \tag{19}$$

ϕ is a multi-layer perceptron (MLP), which throughout our paper is a three-layer ReLU network with hidden dimension 512. The final logit z used for classification is the sum of the contributions from the MLP and the attention head, $z = z_{\text{MLP}} + z_{\text{ATT}}$. The minimal model can be obtained from the one-layer transformer model by (1) removing the LayerNorm operation, (2) the interaction strength ansatz, i.e., by assuming the query, key and value matrices in the attention head have the form

$$K^T Q = \begin{pmatrix} \beta I_{D\times D} & 0_{D\times 1} \\ 0_{1\times D} & 0_{1\times 1} \end{pmatrix}, \quad V = \begin{pmatrix} 0_{D\times D} & 0 \\ 0 & w \end{pmatrix}, \tag{20}$$

and (3) the independence ansatz, where the residual term x_{N+1} is processed by the MLP to produce z_{MLP}, the output of the attention head is z_{ATT} and these two logits are summed to produce the final logit z. However, we stress that the minimal model serves as a phenomenological model and is not derived from the one-layer transformer model.

To reproduce the phenomenology shown in Figure 3, we optimize β, w and the parameters of the MLP ϕ using the same procedure used to train the full model. In particular, we use stochastic gradient descent (SGD) with batch size 128, learning rate 0.01, weight-decay 10^{-10}, $D = 63, N = 100$ and MLP hidden dimension $d = 512$. To reproduce transience in Figure 3c with a fewer number of iterations, we increase the weight-decay parameter on w and β to 10^{-3}.

A.1 Theoretical analysis of the minimal model

To derive an analytical expression for the loss landscape of the minimal model, we assume (1) that only one copy (say x_t) of the target x_{N+1} is present in the context, (2) that $x_{N+1}.x_i = 1$ if $i = t$ and 0 otherwise, and (3) that the distributions of logits obtained when the MLP is applied to all the items in $\mathcal{D}$ with labels $+1$ and -1, (say, $P^+(\phi)$ and $P^-(\phi)$ respectively) are identical except for the sign: $P^+(\phi) = P^-(-\phi)$.

These three assumptions are justified when $K \gg N \gg 1$ and $D \to \infty$. Note that in the simulations presented in Figure 3, we use $K > 10^3$, $N = 100$ and $D = 63$. Assumption (1) is justified when $K \gg N$ as it is unlikely that more than one copy of the target is sampled from the dataset $\mathcal{D}$ ($|\mathcal{D}| = K$) in a single context of length N. Assumption (3) is justified when $K \gg 1$ as all the items in $\mathcal{D}$ are statistically identical irrespective of the label. Assumption (2) is justified when $D \to \infty$ based on our sampling process, which will ensure that in this limit the norm of each item is one and different items are orthogonal. For finite D, our theory will break down when N is sufficiently large, though the precise scaling relationship between D and N for when the theory will break down remains to be examined.

Let t be the index of the target's copy in the context. Under the assumptions stated above,

$$\begin{aligned} z_{\text{ATT}} &= w\left(\frac{e^\beta}{e^\beta+N-1}\ell_t + \frac{1}{e^\beta+N-1}\sum_{j\neq t}\ell_j\right) \\ &= w\left(\frac{e^\beta}{e^\beta+N-1}\ell_t + \frac{1}{e^\beta+N-1}\left(2n_+ - N - \ell_t\right)\right), \end{aligned} \tag{21}$$

where n_+ is the number of tokens which have label $+1$ amongst the N tokens in the context. Suppose we split the dataset $\mathcal{D}$ into two datasets $\mathcal{D}^+$ and $\mathcal{D}^-$ containing items with $+1$ and -1 labels respectively. The MLP (at any particular stage of training) when applied to the items from $\mathcal{D}^\pm$ produces two histograms of logits $P^\pm(\phi)$.

Recall that the binary cross-entropy loss for a given input sequence is $-\log\sigma(\ell_t(z_{\text{MLP}}+z_{\text{ATT}}))$. The average loss is obtained by averaging over the cases when the target is drawn from $\mathcal{D}^+$ and $\mathcal{D}^-$, and by taking an expectation over n_+ (which follows a binomial distribution, $B(n_+)$). Given that there are n_+ items with label $+1$ in the sequence, the target's label is $\ell_t = \pm 1$ with probability n_+/N and $1-n_+/N$ respectively. The average binary cross-entropy loss is then

$$\begin{aligned} \mathcal{L} = &-\sum_{n_+=0}^{N}\frac{n_+}{N}B(n_+)\left\langle \log\sigma\left(\phi^+ + w\left(\frac{e^\beta}{e^\beta+N-1}+\frac{1}{e^\beta+N-1}\left(2n_+-N-1\right)\right)\right)\right\rangle_{\phi^+\sim P^+} \\ &-\sum_{n_+=0}^{N}\frac{N-n_+}{N}B(n_+)\left\langle \log\sigma\left(-\phi^- + w\left(\frac{e^\beta}{e^\beta+N-1}+\frac{1}{e^\beta+N-1}\left(N-2n_+-1\right)\right)\right)\right\rangle_{\phi^-\sim P^-}, \end{aligned} \tag{22}$$

where the two sums consider the cases when the target's label is $+1$ and -1. Given the symmetry between positive and negative labels, the contribution to the loss when $\ell_t = -1$ (the second sum) is approximately equal to the loss when $\ell_t = +1$ (the first sum). Formally, using $P^+(\phi) = P^-(-\phi)$, replacing n_+ with $N - n_+$ in the second sum and using the property $B(n_+) = B(N-n_+)$, we get

$$\mathcal{L} \approx -2\sum_{n_+=0}^{N}\frac{n_+}{N}B(n_+)\left\langle \log\sigma\left(\phi^+ + w\left(\frac{e^\beta}{e^\beta+N-1}+\frac{1}{e^\beta+N-1}\left(2n_+-N-1\right)\right)\right)\right\rangle_{\phi^+\sim P^+}. \tag{23}$$

When $N \gg 1$, n_+/N is approximately Gaussian-distributed: $n_+/N \approx 1/2 + \eta/2\sqrt{N}$, where $\eta \sim \mathcal{N}(0,1)$. This gives

$$\mathcal{L} \approx -\left\langle\left(1+\frac{\eta}{\sqrt{N}}\right)\log\sigma\left(\phi^+ + w\left(\frac{e^\beta-1}{e^\beta+N-1}+\frac{\eta\sqrt{N}}{e^\beta+N-1}\right)\right)\right\rangle_{\eta\sim\mathcal{N}(0,1),\phi^+\sim P^+}, \tag{24}$$

which is presented in the main text (equation 5).

A.2 Expanding the loss before ICL acquisition

w, β are initialized at w_0, β_0 where $|w_0|, |\beta_0| \ll 1$. Define $\gamma \equiv e^\beta - 1$. Close to initialization, w, β satisfy $\gamma/N \ll 1$ and $|w|/\sqrt{N} \ll 1$. We compute the time it takes for the network to acquire the in-context solution, that is, when $w, \beta \gg 1$. From equation 24, since $\gamma \ll N$, we have

$$\mathcal{L} \approx -\left\langle\left(1+\frac{\eta}{\sqrt{N}}\right)\log\sigma\left(\phi^+ + w\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)\right)\right\rangle_{\eta,\phi^+}, \tag{25}$$

where the distributions from which η and ϕ^+ are drawn has been dropped for notational convenience. Since $\gamma/N \ll 1$ and $|w|/\sqrt{N} \ll 1$, we have $w\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right) \ll 1$. Using $\sigma(x) = 1/(1+e^{-x})$ and

two Taylor expansions, we get

$$\mathcal{L} \approx \left\langle \left(1+\frac{\eta}{\sqrt{N}}\right) \log\left(1+e^{-\phi^+ - w\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)}\right)\right\rangle_{\eta,\phi^+}, \tag{26}$$

$$\approx \left\langle \left(1+\frac{\eta}{\sqrt{N}}\right) \log\left(1+e^{-\phi^+}\left(1-w\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)-\frac{w^2}{2}\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)^2\right)\right)\right\rangle_{\eta,\phi^+}, \tag{27}$$

$$\approx \left\langle \log(1+e^{-\phi^+})\right\rangle_{\phi^+} + \left\langle \left(1+\frac{\eta}{\sqrt{N}}\right) \log\left(1-\sigma(-\phi^+)\left(w\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)+\frac{w^2}{2}\left(\frac{\gamma}{N}+\frac{\eta}{\sqrt{N}}\right)^2\right)\right)\right\rangle_{\eta,\phi^+}, \tag{28}$$

$$\approx \left\langle \log(1+e^{-\phi^+})\right\rangle_{\phi^+} - \frac{c_1}{N}\left((1+\gamma)w-\frac{c_2 w^2}{2}\right), \tag{29}$$

$$= \left\langle \log(1+e^{-\phi^+})\right\rangle_{\phi^+} - \frac{c_1}{N}\left(e^{\beta}w-\frac{c_2 w^2}{2}\right) \tag{30}$$

where we define the two MLP-specific dynamical variables,

$$c_1 \equiv \langle \sigma(-\phi^+)\rangle_{\phi^+}, \tag{31}$$

$$c_2 \equiv 1 - \langle \sigma(-\phi^+)^2\rangle_{\phi^+}/c_1. \tag{32}$$

In the second to last step, we have expanded $\log(1+x)$ for small x, computed the expectation over η and ignored higher order terms as the $1/N$ term is the dominant contribution (the $1/\sqrt{N}$ contribution vanishes in expectation). Equation 30 is presented in the main text (equation 7).

Since $\sigma(-x) = 1-\sigma(x)$, c_1 is one minus the probability of a correct classification, averaged over the dataset. At initialization, the logits are centered around $\phi^+ \approx 0$ so that $\sigma(\phi^+) \approx 1/2$ and therefore $c_1, c_2 \approx 1/2$. ϕ^+ increases on average during training, which means c_1 decreases and c_2 increases during training. $c_1 = 0$ and $c_2 = 1$ in the idealized scenario where the dataset is perfectly memorized. We train an MLP in independent experiments to memorize $\mathcal{D}$ with the same hyperparameters used for training the one-layer transformer. We verify empirically that our MLP can near-perfectly memorize the dataset for dataset size K up to 10^4 (Figure A.2 and A.3a).

A.3 ICL ACQUISITION

We consider ICL to be acquired when $w, \beta \gg 1$. To compute the time taken to acquire ICL, we consider gradient descent dynamics of w and β close to initialization. From equation 30, we get

$$\frac{dw}{dt} = \frac{c_1}{N}\left(e^{\beta} - c_2 w\right), \tag{33}$$

$$\frac{d\beta}{dt} = \frac{c_1}{N}\left(we^{\beta}\right). \tag{34}$$

Numerical simulations (Figure A.1) with c_1, c_2 fixed at $1/2$ reveal qualitatively different dynamics depending on whether $|\beta_0| \ll 1$ or $-\beta_0 \gg 1$. Note that $e^{\beta_0} - 1 \ll N$ in both these cases, so that our approximation equation 30 is still valid. We first examine the relevant case $|\beta_0| \ll 1$. The analysis of the second case $-\beta_0 \gg 1$ is presented further below for completeness.

Equations 33 and 34 cannot be solved exactly and we resort to approximations. When $|\beta_0| \ll 1$ and $|w_0| \ll 1$, we have $e^{\beta} \approx 1$ and $|w| \ll 1$ close to initialization. By definition, $c_2 < 1$. Put together, the w term on the right hand side of equation 33 can be ignored close to initialization. Then, $dw/dt \approx (c_1/N)e^{\beta}$. Dividing this by equation 34 and integrating, we get the conservation equation $w^2/2 - \beta = w_0^2/2 - \beta_0$, so that $w = \sqrt{2(\beta - \beta_0 + w_0^2/2)}$. Substituting this expression for w into equation 34, we have

$$\frac{d\beta}{dt} \approx \frac{c_1\sqrt{2(\beta-\beta_0+w_0^2/2)}}{N}e^{\beta}. \tag{35}$$

Integrating this equation from $\beta = \beta_0$ to $\beta = \infty$ allows us to estimate the time t_{ICL} it takes to acquire ICL. Performing this integration, we get

$$\frac{N\Gamma(1/2, w_0^2/2)}{\sqrt{2}} e^{-\beta_0 + w_0^2/2} \approx I_K(t_{\text{ICL}}) \tag{36}$$

where $\Gamma(.,.)$ is the upper incomplete gamma function and we have defined $I_K(t) \equiv \int_0^t c_1(t')dt'$. The subscript K is introduced to highlight c_1's (and thus I_K's) dependence on K. In the main text, we examine the case when $w_0 = 0$ so that t_{ICL} satisfies

$$N\sqrt{\pi/2}e^{-\beta_0} \approx I_K(t_{\text{ICL}}). \tag{37}$$

When $-\beta_0 \gg 1$ and $|w_0| \ll 1$, the dynamics can be split into two parts: 1) w converges from its initial value to the nullcline, $dw/dt = 0$, where $w = e^{\beta}/c_2$, 2) w and β gradually increase along this nullcline until β is large enough that the exponential dependence in equation 9 leads to abrupt ICL acquisition. The first regime is shorter and t_{ICL} is dominated by the duration of the latter regime. In the latter regime, since $w = e^{\beta}/c_2$, we get

$$\frac{d\beta}{dt} = \frac{c_1}{Nc_2} e^{2\beta}. \tag{38}$$

Integrating this equation, we get

$$Ne^{-2\beta_0} \approx I'_K(t_{\text{ICL}}), \quad \text{where } I'_K(t) = \int_0^t \frac{c_1(t')}{c_2(t')} dt'. \tag{39}$$

The $e^{-\beta_0}$ and $e^{-2\beta_0}$ scalings of t_{ICL} for $|\beta_0| \ll 1$ and $-\beta_0 \gg 1$, respectively, are consistent with those obtained when equations 33 and 34 are numerically integrated beginning from $w_0 = 0$ (Figure A.1). In these numerical simulations, we fix $c_1 = 1/2, c_2 = 1/2$, which is equivalent to setting $z_{\text{MLP}} = 0$ (i.e., the MLP does not contribute to the logit).

A.4 Transience

We examine the influence of applying L_2 regularization to the parameters of the attention head. For simplicity, we assume regularization (with L_2 regularization parameter λ_w) is applied only to w. Our goal in this section is to show that such a regularization parameter (however small) is necessary to induce transience, and to delineate the values of λ_w for which transience is recapitulated in the minimal model. A sufficiently large regularization parameter will of course also affect ICL acquisition. Further analysis could delineate the range of values of λ_w for which regularization will have a significant effect on ICL acquisition; this analysis is beyond the scope of this paper.

We write down the dynamics of w *after* ICL is acquired. Re-writing the expression for the loss in equation 24,

$$\mathcal{L} \approx -\left\langle \left(1 + \frac{\eta}{\sqrt{N}}\right) \log \sigma \left(\phi^+ + w \left(\frac{e^{\beta} - 1}{e^{\beta} + N - 1} + \frac{\eta\sqrt{N}}{e^{\beta} + N - 1}\right)\right)\right\rangle_{\eta, \phi^+}. \tag{40}$$

Note that the terms involving η inside the logarithm are at most of order $1/\sqrt{N}$. Since $w, \beta \gg 1$ after ICL is acquired, the first term $(e^{\beta} - 1)/(e^{\beta} + N - 1) \approx 1$ will be much larger than the term involving η. This leads to

$$\mathcal{L} \approx -\left\langle \left(1 + \frac{\eta}{\sqrt{N}}\right) \log \sigma \left(\phi^+ + w\right)\right\rangle_{\phi^+}. \tag{41}$$

Since $\langle \eta \rangle = 0$, we have

$$\mathcal{L} \approx \left\langle -\log \sigma \left(\phi^+ + w\right)\right\rangle_{\phi^+}, \tag{42}$$

$$= \left\langle \log \left(1 + e^{-\left(\phi^+ + w\right)}\right)\right\rangle_{\phi^+}, \tag{43}$$

$$\approx \left\langle e^{-\phi^+}\right\rangle_{\phi^+} e^{-w}. \tag{44}$$

where we have used $e^{-(\phi^+ + w)} \ll 1$ and $\log(1+x) \approx x$ for $x \ll 1$. Define $c_3 \equiv \left\langle e^{-\phi^+} \right\rangle_{\phi^+}$, which is our third MLP-specific dynamical variable. c_3 will decrease while the MLP memorizes the dataset. Equation 44 implies that, without any regularization, w will continue to increase (albeit at a decreasing rate) and ICL will not be transient.

Introducing an L_2 regularization term, we have

$$\mathcal{L} \approx c_3 e^{-w} + \frac{\lambda_w w^2}{2}. \tag{45}$$

The gradient descent dynamics of w is given by

$$\frac{dw}{dt} = c_3 e^{-w} - \lambda_w w. \tag{46}$$

With regularization, once the network sufficiently memorizes the dataset (i.e., c_3 is sufficiently small), w reaches a steady state, denoted w_{tr}, which satisfies $w_{\text{tr}} e^{w_{\text{tr}}} = c_3/\lambda_w$. Re-expressing this relationship in terms of the Lambert W function, we have

$$w_{\text{tr}} \approx W\left(c_3/\lambda_w\right). \tag{47}$$

Since $c_3 = \left\langle e^{-\phi^+} \right\rangle_{\phi^+}$, we have $c_3 \approx 1$ at initialization. c_3 will continue to decrease (however slowly) after ICL is acquired while the loss in equation 45 is minimized. λ_w is typically chosen to be a small value (the largest value across all our simulations is $\lambda_w = 10^{-3}$). Due to slow memorization (declining c_3), w_{tr} slowly declines while tracking equation 47. ICL fades when w_{tr} is of order one; since $W(x) = 1$ when $x \approx 3$, i.e., some number of order one, ICL fades when $c_3 \approx \lambda_w$.

This analysis suggests that ICL is transient in our minimal model when the network sufficiently memorizes the dataset (c_3 is sufficiently small). Importantly, whether ICL is transient or not depends both on the extent to which the network is able to memorize the dataset and how much regularization is applied to each of the network elements. In transformer networks, how regularization may affect each sub-circuit is hard to disentangle. For example, it is possible that the output of the attention head passes through the MLP before producing the final logit. Increasing the regularization on MLP parameters may decrease the rates at which the network memorizes *and* generalizes.

A.5 The relationship between ICL and IWL loss after ICL acquisition

Recall that IWL is measured on sequences where the item-label pairs are drawn from $\mathcal{D}$ but a copy of the target item is (most likely) not in the sequence. This allows us to define an IWL loss $\mathcal{L}_{\text{IWL}}$, which corresponds to the binary cross-entropy loss of the network on such sequences. Similarly, we define an ICL loss $\mathcal{L}_{\text{ICL}}$, which is the binary cross-entropy loss on input sequences that contain N novel item-label pairs and where the target is one of the N items in the sequence. Equation 47 hints at a relationship between $\mathcal{L}_{\text{ICL}}$ and $\mathcal{L}_{\text{IWL}}$ after ICL is acquired. We now provide a heuristic argument to derive this relationship.

Immediately after ICL is acquired, the ICL loss is small, $\mathcal{L}_{\text{ICL}} \ll 1$. Our arguments in the previous section show that ICL will be transient if the MLP that can sufficiently memorize the dataset and appropriate regularization is applied. Once ICL fades and the network near-perfectly memorizes the dataset, we expect $\mathcal{L}_{\text{IWL}} \ll 1$.

We first consider the case when $\mathcal{L}_{\text{ICL}} \ll 1$, that is, after ICL is acquired but before it fades. ICL acquisition corresponds to $w, \beta \gg 1$. Recall from the independence ansatz that the final logit is $z_{\text{MLP}} + z_{\text{ATT}}$. IWL sequences do not contain a copy of the target in the sequence. Instead, since $\beta \gg 1$, the target x_{N+1} will pay attention to a random item, say x_i, in the input sequence that maximizes the dot product $x_{N+1}.x_i$. Item i's label ℓ_i matches the target's label ℓ_t with probability half. In the case that it does match, $z_{\text{ATT}} \approx w\ell_i = w\ell_t$. Since the MLP has not yet memorized $\mathcal{D}$, the typical logit from the MLP is small compared to w, $|z_{\text{MLP}}| \ll w$. The IWL loss when $\ell_i = \ell_t$ is $-\log\sigma(\ell_t(z_{\text{MLP}} + z_{\text{ATT}})) \approx -\log\sigma(\ell_t(w\ell_i)) = -\log\sigma(w)$, which is negligible since $w \gg 1$. In the alternative case, $\ell_i = -\ell_t$, the IWL loss is $-\log\sigma(w\ell_i\ell_t) = -\log\sigma(-w) = \log(1+e^w) \approx w$, where we used $w \gg 1$ in the final step. Since $\ell_i = \pm\ell_t$ with equal probability of $1/2$, the expected IWL loss is $\mathcal{L}_{\text{IWL}} \approx w/2$.

On the other hand, ICL sequences do indeed contain a copy of the target in the sequence. In this case, $z_{\text{ATT}} = w\ell_t$. Since $|z_{\text{MLP}}| \ll w$, the ICL loss is $\mathcal{L}_{\text{ICL}} = -\log\sigma(\ell_t(z_{\text{MLP}} + z_{\text{ATT}})) \approx -\log\sigma(\ell_t(w\ell_t)) \approx \log(1+e^{-w}) \approx e^{-w}$, where we used $\log(1+x) \approx x$ for $x \ll 1$ in the final step. Putting the expressions for $\mathcal{L}_{\text{IWL}}$ and $\mathcal{L}_{\text{ICL}}$ together, we get

$$\mathcal{L}_{\text{IWL}} \approx -\frac{1}{2}\log\mathcal{L}_{\text{ICL}} \tag{48}$$

after ICL acquisition and before ICL fades away.

We now consider $\mathcal{L}_{\text{IWL}} \ll 1$, that is, after ICL fades away and the network significantly memorizes the dataset. In this case, since $|w|, |\beta| \ll 1$, we have $|z_{\text{ATT}}| \ll |z_{\text{MLP}}|$. The loss on ICL sequences is dominated by the logit produced by the MLP on novel items. Suppose that the typical magnitude of the logit produced by the MLP is ξ. The IWL loss is $\mathcal{L}_{\text{IWL}} \approx -\log\sigma(\xi) = \log(1+e^{-\xi})$. $\mathcal{L}_{\text{IWL}} \ll 1$ implies $\xi \gg 1$ and thus $\mathcal{L}_{\text{IWL}} \approx e^{-\xi}$. For novel items, the MLP will produce either $\pm\xi$ irrespective of the true label of the target. The ICL loss is either negligible or $\log(1+e^{\xi})$, depending on whether the sign of z_{MLP} matches the target's label or not. Since $\xi \gg 1$, the ICL loss averaged over these two cases is $\mathcal{L}_{\text{ICL}} \approx \xi/2$. We thus get

$$\mathcal{L}_{\text{ICL}} \approx -\frac{1}{2}\log\mathcal{L}_{\text{IWL}} \tag{49}$$

after ICL fades away and the network significantly memorizes the dataset.

A.6 Supplementary Figures

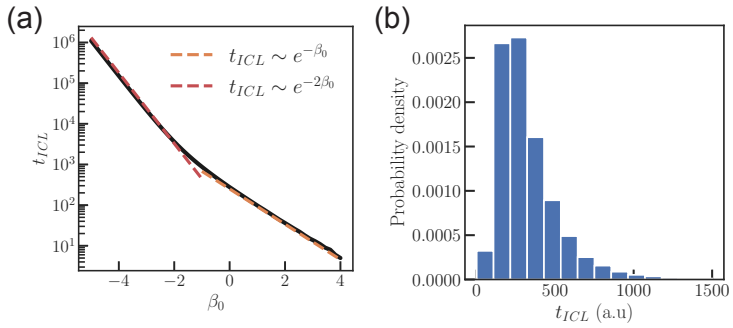

Figure A.1: (a) t_{ICL} exhibits exponential dependence on the initial values of β_0. (b) Histogram of t_{ICL} in the minimal model given normally distributed values of β_0. The distribution of t_{ICL} is long-tailed due to its exponential dependence on β_0.

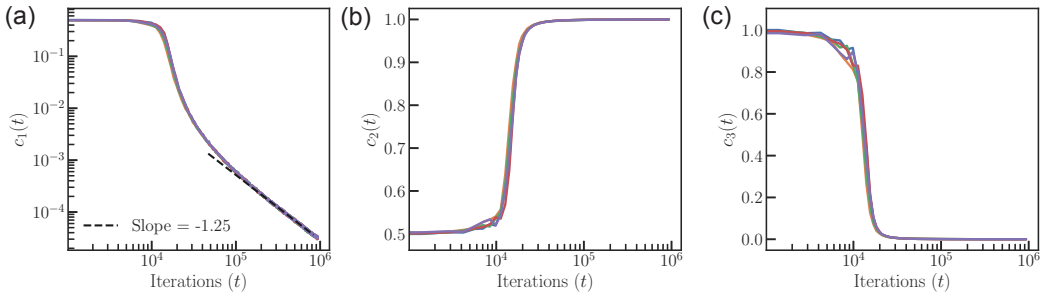

Figure A.2: Time evolution of summary statistics of the MLP logit distribution. (a) $c_1 \equiv \langle\sigma(-\phi^+)\rangle_{\phi^+}$ exhibits a scaling law $c_1(t) \sim t^{-\mu}$, where $\mu \approx 1.25$. (b) $c_2 \equiv 1 - \langle\sigma(-\phi^+)^2\rangle_{\phi^+}/c_1$ (c) $c_3(t) = \left\langle e^{-\phi^+}\right\rangle_{\phi^+}$.

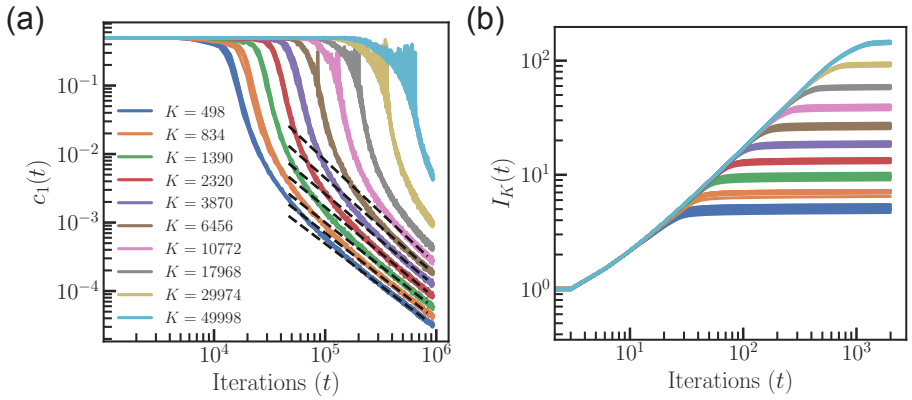

Figure A.3: (a) c_1 curve satisfies a scaling law $c_1(t) \sim t^{-\mu}$, where $\mu \approx 1.25$, which is largely independent of K. (b) Time evolution of $I_K(t) = \int_0^t c_1(t')dt'$.

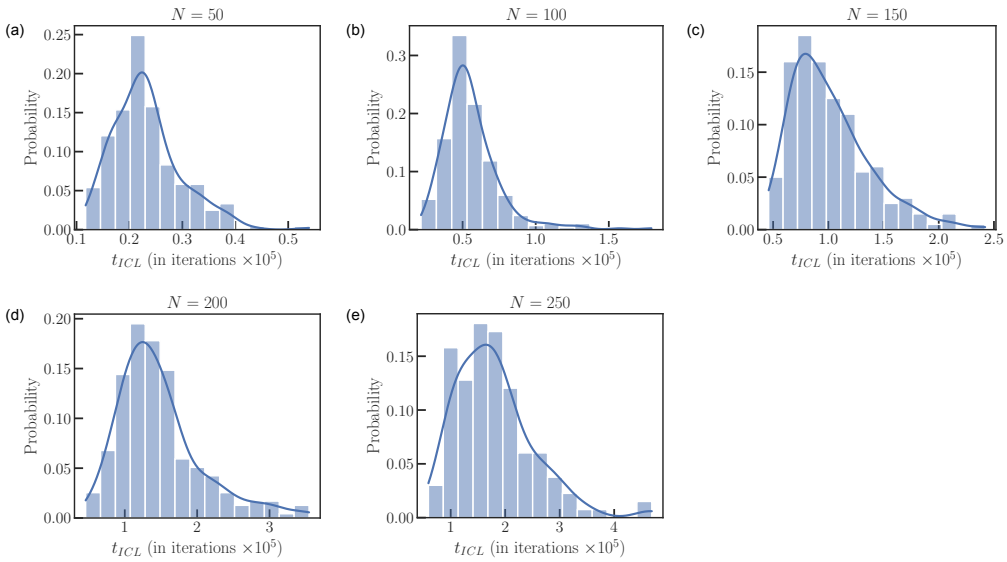

Figure A.4: For each N, histograms of t_{ICL} show a long-tail.

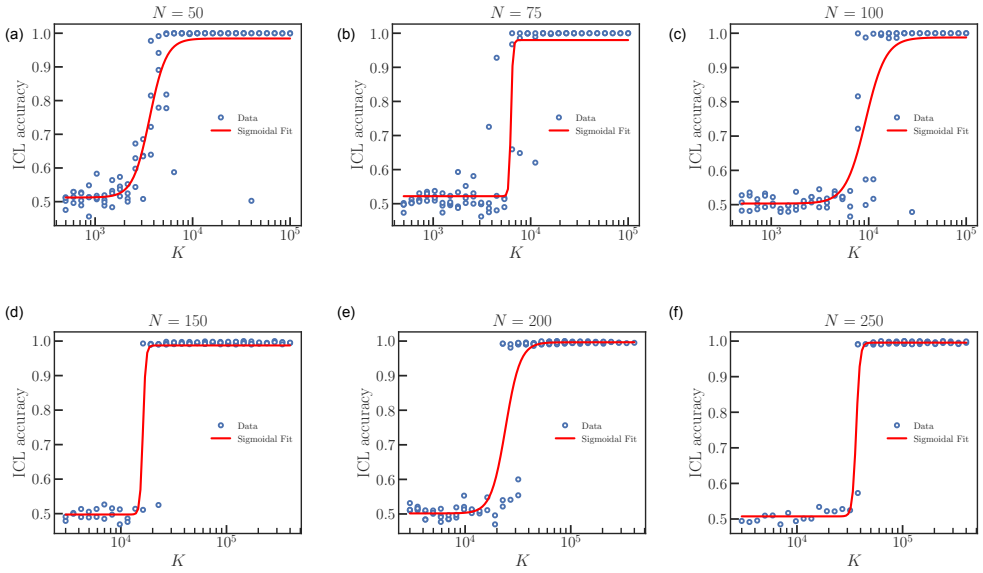

Figure A.5: ICL performance is measured for various N and K. For all N, we see a sharp transition from memorization to generalization as K increases. We fit a sigmoid to the data to determine the critical task diversity threshold K^* for each N. (a) $N = 50$. (b) $N = 75$. (c) $N = 100$. (d) $N = 150$. (e) $N = 200$. (e) $N = 250$.

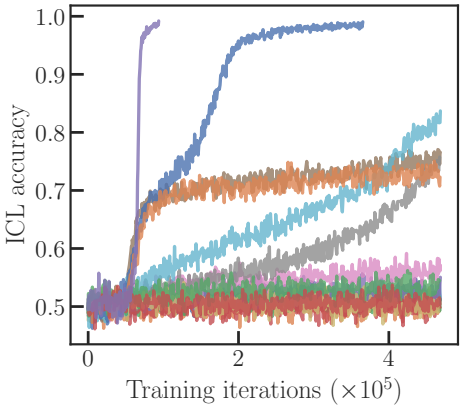

Figure A.6: ICL performance curves (for different random number seeds) for models trained on our modified $\mathcal{D}'$, where every sequence contains exactly $N/2$ tokens with $+1$ and -1 labels for $K = 31336$. We see novel intermediate suboptimal solutions, and many of the runs never acquire ICL.