IN-CONTEXT LEARNING BEFORE SPECIALISATION: THEORY OF LEARNING IN A SOLVABLE TOPIC MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Sequence models can generalise in fundamentally distinct ways: extrapolating structures seen in training, or using their prompt to adapt to novel tasks. The prevailing strategy determines how the model behaves beyond its training distribution, controlling potential capabilities and risks. Recent work has characterised many different strategies theoretically, and empirically described the transition between them as the quality and quantity of pretraining data vary. However, there is no learning theory for this competition, hence no fundamental understanding of *why* one solution might be favoured over another. Here we provide such a theory in a solvable topic model—documents drawn from a finite mixture of latent topics, mirroring the heterogeneity of modern pretraining data. The dataset comes with two distinct solutions: a specialised predictor, that stores the finite topic mixture, and an in-context one, optimal for topics never seen during training. Our theory predicts their emergence at distinct sample complexities: the in-context predictor is available early, while the specialised one emerges only after the training tokens outnumber the parameters of the data-generating process. Trained transformers reproduce both scales and the separating crossover in remarkable agreement with the theory, with no fitted parameters.

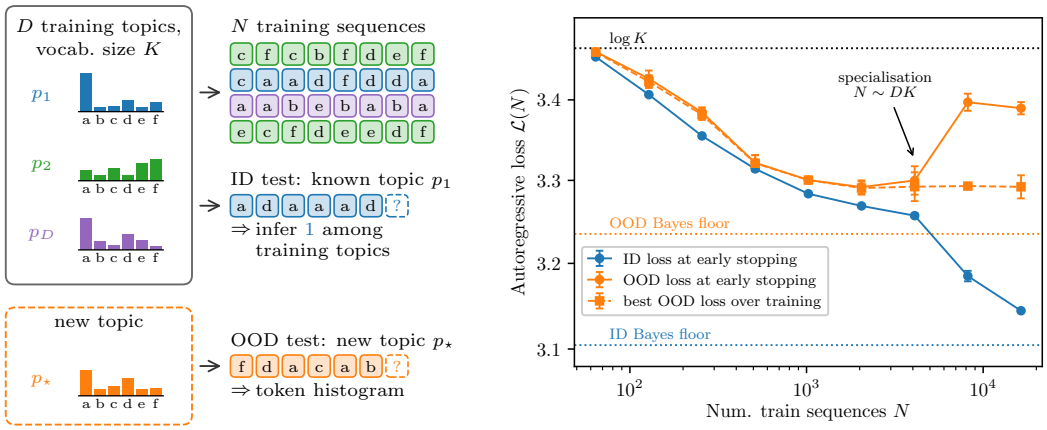


Figure 1: **In-context learning and specialisation in a random topic model. Left:** A random topic model consists of D topics, i.e. unigram probabilities over a vocabulary of size K, $\mathbf{p}_z \in \Delta^{K-1}$ for $z = 1, \ldots, D$. Training sequences are obtained by first sampling a topic z uniformly from the D available, then sampling T tokens i.i.d. from $\mathbf{p}_z$. **Right:** Learning curve of a one-layer transformer trained for next-token prediction on a random topic model with $(K, D, T) = (32, 512, 128)$. Initially, both In-Distribution (ID, on the D training topics, blue) and Out-Of-Distribution (OOD, on a new topic $\mathbf{p}_\star$, orange) descend with N towards the entropy of the Bayes-optimal OOD predictor (orange dotted): the add-one estimator $r_t(a \mid X_{\leq t}) = (\#\{i \leq t \mid X_i = a\} + 1)/(t + K)$. Later, the predictor converges to Bayes-optimal inference of the training topic, resulting in convergence of the ID loss to the ID Bayes-optimal floor (blue dotted), while the OOD loss rises as in overfitting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Large language models generalise in ways that were not anticipated in training. In particular, novel behaviours can be elicited simply via prompting, including few-shot learning of supervised tasks (Brown et al., 2020) or adopting different roles depending on a set of instructions (Shanahan et al., 2023). This In-Context Learning (ICL) mechanism is not captured by the classical learning-theoretic paradigm, where a model is evaluated on fresh samples from the distribution it was trained on. Nevertheless, ICL is a critical source of both generality and risk (Anwar et al., 2024) in modern AI systems, calling for a theory of what trained models do beyond their training distribution.

An early theoretical attempt framed ICL as implicit Bayesian inference of the latent concept behind the prompt (Xie et al., 2022). This framework has found extensive support: trained transformers behave as Bayes-optimal estimators for many families of tasks (Garg et al., 2022), and how such estimators are represented and acquired during training is understood in simple settings (Nichani et al., 2024; Edelman et al., 2024; Lu et al., 2025; Zhang et al., 2025; Bondaschi et al., 2026). However, other settings admit distinct classes of predictors, and models transition between them as the training conditions vary (Akyürek et al., 2023). Raventos et al. (2023) identified task diversity as the key control parameter: When trained at low diversity, the transformer can only memorise the training tasks, acquiring general ICL abilities only above a certain task-diversity threshold.

The role of task diversity has been explored in synthetic data, revealing that the competition between the two behaviours extends along training time (Singh et al., 2023; Reddy, 2024; Nguyen & Reddy, 2025; Park et al., 2025), culminating in a unifying hierarchical Bayesian framework that fits the model behaviour to a weighted mixture of the memorising and generalising predictors (Carroll et al., 2025; Wurgaft et al., 2025). Still missing is the *why*: a theory that predicts the favoured solution from data properties alone, and how much data it takes for one solution to give way to the other.

Our Contribution Here we provide this theory in a synthetic dataset containing the minimal required ingredients: a task diversity parameter D, a vocabulary of size K, a document length T and a corpus of N training documents. Task diversity enters as a mixture of D latent topics from which documents are generated (Fig. 1, left). In this setting, we provide:

- **A solvable model of strategy competition.** We characterise analytically two distinct Bayes-optimal predictors: a *specialised* one, relying on knowledge of the D topics in the data-generating process, and an *in-context* one, equivalent to the add-one estimator from statistics under our data model (Sec. 2).
- **A first-principles learning-curve theory for specialisation.** We derive the specialisation learning curve in a Bayesian learning setting: the excess loss over the asymptotic floor is $DK/(2NT)$ at leading order—a function of the number of training tokens per parameter of the data-generating process (Sec. 3).
- **A statistical mechanism for the in-context phase.** The in-context predictor is linear in the token counts, and a learner that estimates this map from data acquires it at a sample complexity set by its K^2 parameters and by the information that a context carries about the next token: independent of D, and growing as K^3 at fixed context length. Comparison with the specialisation scale gives a task-diversity threshold above which (see Fig. 1) *the in-context solution emerges first, it remains stable in a window that widens with task diversity, then it gives way to the specialised solution* (Sec. 4).
- **Parameter-free agreement with trained transformers.** Across vocabularies, task diversities, sequence lengths and corpus sizes, transformer in-distribution losses collapse as functions of NT/DK, converging to the specialised solution at the predicted rate. Non-trivial ICL performance develops at the predicted, D-independent sample complexity $NT = O(K^3)$, then vanishes again with specialisation (Sec. 5).
- **Separation from classical spectral methods.** Reconstructing the topics from low-order moments requires resolving a Marchenko–Pastur bulk of eigenvalues, which takes $N \sim D^2K$ samples (Hsu et al., 2012; Anandkumar et al., 2014): a different, worse scaling with task diversity than the $NT \sim DK$ of the Bayes-optimal learner, which the trained transformers follow (Sec. 5.1).

1.1 Related work

Mixtures of tasks as ICL testbeds. Finite task mixtures underlie most theoretical studies of ICL: linear regression (Garg et al., 2022; Akyürek et al., 2023; Raventos et al., 2023; Lu et al., 2025; Zhang et al., 2025; Bordelon et al., 2026), classification (Reddy, 2024; Nguyen & Reddy, 2025), and mixtures of Markov (Edelman et al., 2024; Park et al., 2025; Bondaschi et al., 2026) and hidden Markov (Riechers et al., 2025; Ray et al., 2026) models. The "Balls & Urns" setting of Wurgaft et al. (2025) is precisely the random topic model studied here, and their memorising and generalising predictors are our specialised and in-context ones. Our work complements their phenomenological Bayesian framework: where they assume a scaling law and a low-Kolmogorov-complexity bias, we derive sample complexities and interpolation between the two predictors from first principles. Thus, our approach does not require estimating (intractable) Kolmogorov complexities and fitting parameters, but rather prescribes a simple picture where the functional simplicity of the ICL solution promotes its emergence as long as it leads to a noticeable loss gain.

Representations, Belief-state geometry and task vectors. Due to next-token pretraining, transformers linearly represent the belief state over latent structure (Shai et al., 2024). In particular, mixtures of independent components, or *non-ergodic* sources, give rise to a telescopic geometry, which sparsifies along the context (Ray et al., 2026). The random topic model is the simplest non-ergodic source, and our results predict the sample dependence of the resulting belief geometry. Beyond controlled data structures, ICL has been linked to task vectors encoding the inferred task (Hendel et al., 2023), whose geometry distinguishes dual modes of task inference corresponding to our two predictors (Yan et al., 2026). Overall, these approaches describe *how* a trained network implements the chosen strategies; ours predicts *when* the data favours each.

Learning theories of ICL. Most analytical studies of learning focus on the infinite-data and infinite-diversity limit. Lu et al. (2025) is closest to us: they solve linear attention trained on Gaussian regression in a proportional asymptotic limit, obtaining sharp learning curves and a genuine transition with increasing task diversity. However, since their linear model cannot fit the specialised predictors exactly, they did not observe the crossover from in-context to specialised predictor as the number of data increases. Furthermore, our setup can be extended to mixtures of more complex hidden Markov models (Riechers et al., 2025) and hierarchical structures (Cagnetta et al., 2024).

2 Data model and optimal predictors

This section introduces our synthetic model of mixture data and the two reference predictors representing the *specialised* and *in-context solutions*. The solutions are distinguished by their next-token-prediction loss profiles. As shown in Fig. 2, trained transformers behave like the specialised predictor at low data diversity, while converging to the in-context one at high diversity. The rest of the paper explains why, and on which scale of the corpus size N each solution is selected.

Random Topic Model A dataset is specified by a vocabulary size K, a number of topics D and a document length T. Each topic is a probability distribution over the vocabulary, $\mathbf{p}_z \in \Delta^{K-1}$ with entries $p_z(a)$ for $a = 1, \dots, K$, drawn independently and uniformly from the simplex, $\mathbf{p}_z \sim \mathrm{Dir}_K(\mathbf{1})$ for $z = 1, \dots, D$, and then held fixed. We collect the topics as the columns of the topic matrix $P = [\mathbf{p}_1, \dots, \mathbf{p}_D] \in \mathbb{R}^{K \times D}$. A document is generated by drawing a topic index Z uniformly from $\{1, \dots, D\}$, then T tokens $X_1, \dots, X_T \in \{1, \dots, K\}$ independently from $\mathbf{p}_Z$ (Fig. 1, left). Given the topic, the likelihood of a context $X_{\le t} = (X_1, \dots, X_t)$ depends only on its histogram, the count vector

$$C(X_{\le t}) \in \mathbb{N}^K, \qquad C_a(X_{\le t}) := \#\{s \le t : X_s = a\}. \tag{1}$$

The training corpus consists of N documents, all generated with the same topic matrix P.

Data versus topic distribution. Two kinds of averages appear throughout. $\mathbb{E}[\cdot]$ denotes the expectation over documents at fixed topic matrix P, i.e. over the topic index Z and the tokens: this is the average that defines the loss of any predictor, including a trained transformer. $\bar{\cdot}$ denotes the further average over the draw of the topics. Unless stated otherwise, the curves in the figures are averaged over independent realisations of the topic matrix P.

Next-token prediction, In- and Out-Of-Distribution. We consider two test distributions. In-distribution (ID) documents are generated as the training ones, from one of the D topics of P. Out-of-distribution (OOD) documents are generated from a fresh topic $\mathbf{p}_\star \sim \mathrm{Dir}_K(\mathbf{1})$, drawn from the same prior as the training topics but independently, and hence almost surely different from all of them. A next-token predictor $\mathbf{f}$ maps each context $X_{\leq t}$ to a distribution $\mathbf{f}_t(X_{\leq t}) \in \Delta^{K-1}$ over X_{t+1}. We measure ID performance by the next-token loss at context length t, or t-gram entropy

$$\mathcal{L}_t[\mathbf{f}] := \mathbb{E}\left[-\log f_t(X_{t+1} \mid X_{\leq t})\right], \qquad t = 0, \ldots, T-1, \tag{2}$$

where the expectation is over fresh documents generated from the D training topics. The same expectation over a fresh topic $\mathbf{p}_\star$ defines the OOD loss $\mathcal{L}_t^\star[\mathbf{f}]$. The usual autoregressive loss $\mathcal{L}_{\mathrm{AR}}$ is the average of $\mathcal{L}_t$ over positions.

2.1 The specialised predictor

A learner that knows the topic matrix P still does not observe the topic Z of a given document, and must infer it from the context. The Bayes-optimal next-token distribution is the posterior mixture of the topics,

$$\mathbf{q}_t(X_{\leq t}) := \sum_{z=1}^{D} \gamma_t(z)\, \mathbf{p}_z, \qquad \gamma_t(z) := \mathbb{P}\left(Z = z \mid X_{\leq t}\right) = \frac{\prod_{a=1}^{K} p_z(a)^{C_a}}{\sum_{y=1}^{D} \prod_{a=1}^{K} p_y(a)^{C_a}}, \tag{3}$$

where the posterior over the topic $\gamma_t(z)$ depends on the context only through the counts. $\mathbf{q}_t$ is the true conditional distribution of X_{t+1} given the context, hence it minimises $\mathcal{L}_t$ over all predictors (App. B). We call it *specialised*, as it is built from the D topics realised in training.

The cross-entropy loss splits into an entropy plus a Kullback–Leibler divergence,

$$\mathcal{L}_t[\mathbf{q}] = \frac{1}{D} \sum_{z=1}^{D} H(\mathbf{p}_z) + \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbf{p}_Z \,\|\, \mathbf{q}_t\right)\right], \tag{4}$$

both averaged over the unknown topic. The first term is the loss of an oracle that observes the topic, and no predictor can go below it: it is the *entropy rate* of the data-generating process. For sufficiently large D, the entropy rate converges to the average over draws of the topic matrix $\overline{H(\mathbf{p})}$, which is $\log K - 1 + \gamma_{\mathrm{E}} + O(1/K)$ for large K (App. A), with $\gamma_{\mathrm{E}} \simeq 0.57721$ Euler's constant. The second term is the excess loss due to the uncertainty on the topic and, by Eq. 3, it is controlled by the posterior mass on the wrong topics $y \neq z$. As shown in App. B, the mass on a wrong topic y decays exponentially in t, at a rate given by the divergence $D_{\mathrm{KL}}(\mathbf{p}_z \| \mathbf{p}_y)$. For random topics this divergence is of order one (App. A), so that discriminating the true topic from the other $D-1$ requires a context of length $t \sim \log D$ (App. B). This scale is visible in the left panel of Fig. 2.

2.2 The in-context predictor

Consider now an OOD document, generated from a topic $\mathbf{p}_\star$ that does not belong to P. A learner that knows the prior $\mathrm{Dir}_K(\mathbf{1})$ but not $\mathbf{p}_\star$ must infer from the context the topic itself, rather than its index within a known dictionary. By Dirichlet categorical conjugacy, the posterior over $\mathbf{p}_\star$ given $X_{\leq t}$ is $\mathrm{Dir}_K(C_1{+}1, \ldots, C_K{+}1)$, and the Bayes-optimal next-token distribution is its mean, the add-one smoothing of the histogram (App. C)

$$r_t(a \mid X_{\leq t}) := \mathbb{P}\left(X_{t+1} = a \mid X_{\leq t}\right) = \frac{C_a(X_{\leq t}) + 1}{t + K}. \tag{5}$$

We call it *in-context* as it does not depend on the topic matrix P, and thus applies unchanged to documents from seen or unseen topics.

The OOD loss of $\mathbf{r}_t$ decomposes as in Eq. 4, with the unseen topic in place of $\mathbf{p}_Z$,

$$\mathcal{L}_t^\star[\mathbf{r}] = \overline{H(\mathbf{p})} + \mathbb{E}\left[D_{\mathrm{KL}}\left(\mathbf{p}_\star \,\|\, \mathbf{r}_t\right)\right] = \overline{H(\mathbf{p})} + \frac{K-1}{2t} + O\left(\frac{\log t}{t^2}\right), \tag{6}$$

where the expansion holds at fixed K for $t \gg K$ (App. C). Simply put, the histogram needs to include at least order K tokens to depart from a trivial prediction $1/K$. Note that, since $\mathbf{p}_\star$ is drawn

from the same prior as the topics of P, the entropy floor is the same for the two predictors, while the approach is significantly different (see Fig. 2, left). Identifying one of D known topics requires $\log D$ nats of evidence, which the context delivers at a constant rate per token, hence the exponential convergence of Sec. 2.1. Estimating a new topic is instead a $(K-1)$-dimensional problem, and the excess in Eq. 6 is the classical half of the number of parameters per observation behaviour.

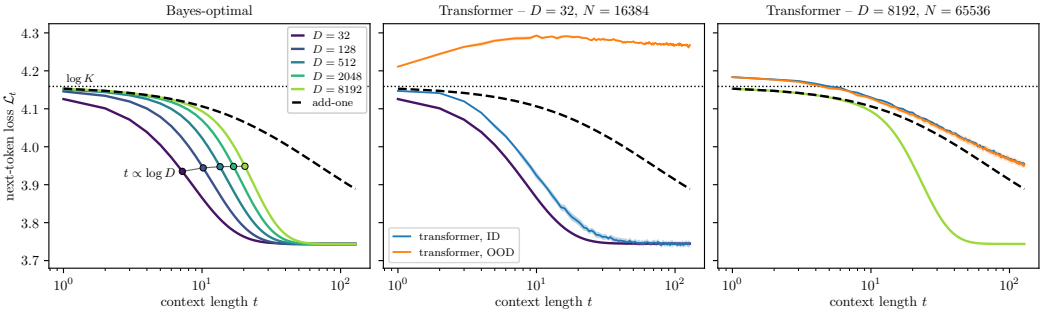


Figure 2: **Next-token loss profiles of the two Bayes-optimal predictors and of trained transformers**, for $K = 64$ and $T = 128$. **Left:** ID loss $\overline{\mathcal{L}_t[\mathbf{q}]}$ of the specialised predictor for D from 32 to 8192 (colour), averaged over realisations of the topic matrix, and OOD loss $\mathcal{L}_t^\star[\mathbf{r}]$ of the in-context predictor (black dashed), which does not depend on D. Dotted line: $\log K$. Markers: context length at which each specialised profile has covered half of its descent, growing as $\log D$ (grey guide; Sec. 2.1). **Middle and right:** the same two references at $D = 32$ and $D = 8192$, with the ID (blue) and OOD (orange) loss profiles of a one-layer transformer trained on N documents at that D, mean $\pm$ s.e. over four runs with independent topic matrix, corpus and initialisation, at the checkpoint of minimal validation loss. At $D = 32$ the transformer matches the specialised predictor in distribution and stays above $\log K$ out of distribution; at $D = 8192$ both profiles follow the in-context predictor.

3 Learning-curve theory of specialisation

The training corpus contains only the D topics in the matrix P. Hence, a sufficiently expressive, well-optimised machine learning model *must* converge to the specialised solution Eq. 3 in the infinite-data and training time limit. But *how many training documents does it take to specialise?*

Bayesian learning We answer this question with a Bayesian learner that uses the training corpus optimally: It knows the model class, including D and K, but not the realised topic matrix P, and observes a collection $\mathcal{C}_N$ of N length-T training documents, together with the test context. The training corpus informs a posterior $\pi(dP|\mathcal{C}_N)$ over the topic matrix P, proportional, by Bayes' rule, to prior $\mu(dP)$ times the joint probability of the corpus under P, $\mathbb{P}\{\mathcal{C}_N\}$. The Bayes-optimal learner is the posterior mean of the P-dependent specialised predictor from Eq. 3,

$$\hat{\mathbf{q}}_t(X_{\leq t}) \propto \int \mathbf{q}_t(X_{\leq t}; P)\, \pi(dP|\mathcal{C}_N, X_{\leq t}), \tag{7}$$

where the posterior also includes the test context $X_{\leq t}$ (see App. D for details). Note that, after averaging over the draw of the topics, no predictor built from the same data has a lower loss than $\hat{\mathbf{q}}$. Its loss is the statistical limit of specialisation from N training documents, hence a natural reference against which a transformer trained on the same corpus can be measured.

Learning curve and mutual information. The learning curve of specialisation is the excess loss of the Bayes-optimal learner over the specialised predictor, averaged over topic matrices,

$$\varepsilon_t(N) := \overline{\mathcal{L}_t[\hat{\mathbf{q}}] - \mathcal{L}_t[\mathbf{q}]}\,, \tag{8}$$

Averaging over the topics gives it an information-theoretic meaning: $\varepsilon_t(N)$ is the conditional mutual information between P and the next token, given corpus and context (App. D),

$$\varepsilon_t(N) = I(P; X_{t+1} \mid X_{\leq t}, \mathcal{C}_N) \Rightarrow \sum_{t=0}^{T-1} \varepsilon_t(N) = I(P; \mathcal{C}_{N+1}) - I(P; \mathcal{C}_N), \tag{9}$$

where $\mathcal{C}_{N+1}$ adds the test document to the corpus. The excess loss of the learner on the test document equals the information that this document adds about the topic matrix. Thus, the learning curve quantifies how fast the corpus reveals P.

The DK/NT law. The topic matrix has $D(K-1)$ free parameters, and a corpus of N documents reveals them at the rate of parametric statistics: under identifiability and regularity, $I(P;\mathcal{C}_N) = (D(K-1)/2)\log N + O(1)$ (Clarke & Barron, 1990; 1994). The increment gives the leading term of the learning curve at fixed D and K (App. D),

$$\frac{1}{T}\sum_{t=0}^{T-1}\varepsilon_t(N) = \frac{D(K-1)}{2NT} + o\big(N^{-1}\big)\,. \tag{10}$$

Specialisation thus requires $NT \gg DK$, and the learning curve is a function of the number of training tokens per parameter, $x := NT/(DK)$, decaying as x^{-1}. As shown in Sec. 5, trained transformers specialise on the same sample-complexity scale, and with the same decay. Two remarks. First, identifiability, i.e. that the law of a document determines P up to a relabelling of the topics, requires a sufficiently long document and a *generic* topic matrix. For $D \leq K$, $T \geq 3$ suffices: the columns of P are linearly independent almost surely under the topics draw, and the third-order moment of the tokens has a unique decomposition with the topics as factors (Kruskal, 1977; Anandkumar et al., 2014). For larger D the same holds for the lifted topics $\mathbf{p}_z^{\otimes m}$, which are generic in the same sense, i.e. linearly independent, as long as $D \leq \binom{K+m-1}{m}$, the dimension of the space of symmetric tensors, at the price of longer documents, $T \geq 2m{+}1$ (App. D). For instance, pairs of tokens cover $K < D \leq K(K{+}1)/2$ with $T \geq 5$. Secondly, the learning curve is an average over topic matrices, whereas an experiment runs on individual realisations. Comparing the two assumes that the excess loss of a given realisation is close to its average, i.e. typicality of the random topic matrix. An empirical check of this typicality is the spread across independent realisations in Sec. 5.

The same law at fixed, large position. Equation 10 concerns the loss averaged over positions; at a fixed position t the same law follows from a simple picture. Beyond the resolution scale $\log D$ of Sec. 2.1, the learner can identify the topic of the test document, and likewise the topics of the training documents if $T \gg \log D$. It is then left with estimating the token distribution of the test topic from the training documents of that topic, about N/D of them, i.e. from $m \approx NT/D$ tokens: this is the problem solved by the in-context predictor of Sec. 2.2, with m in place of t, whose excess $(K-1)/(2m)$ becomes $D(K-1)/(2NT)$. The t-gram loss at large t therefore obeys the same law as the average, and App. D makes this rigorous at fixed P, under conditions on t, T and N; Fig. 10 compares it with the T-gram learning curves of trained transformers.

4 When does ICL emerge?

Simplicity of the in-context predictor. If any learner eventually converges to the specialised predictor, *when does ICL emerge?* Both predictors of Sec. 2 are functions of the counts, but of very different complexity. The in-context predictor is linear up to normalisation,

$$\mathbf{r}_t(X_{\leq t}) \propto \mathbf{C}(X_{\leq t}) + \mathbf{1}, \tag{11}$$

whereas the specialised predictor is exponential in the counts (Eq. 3). Then, simplicity bias suggests that the in-context predictor is easier to learn. $\mathbf{r}_t$ represents a point in the family of K^2 linear maps from counts to next-token probability, whereas random topic models P lie in $D(K-1)$ dimensions. The estimation error follows from the same argument as Sec. 3, in the form appropriate to a fitted parametric family rather than a Bayesian posterior (App. E). Simply put, a learner estimating the K^2 parameters of the linear map from a corpus of NT tokens pays an excess $K^2/(2NT)$ over the optimal in-context solution, again proportional to the number of parameters per training token.

Estimation penalty versus loss gain. This excess must be compared with the gain of the add-one rule over the uniform prediction, equal to the mean mutual information between the next token and its context (App. E),

$$G(K,T) := \log K - \frac{1}{T}\sum_{t<T}\mathcal{L}_t^\star[\mathbf{r}] = \frac{1}{T}\sum_{t=0}^{T-1} I\left(X_{t+1}; X_{\leq t}\right). \tag{12}$$

The new solution can only appear once the error due to finite data $K^2/(2NT)$ is smaller than the gain, i.e. for

$$N \gtrsim N_{\rm icl} := \frac{K^2}{2T\,G(K,T)}, \tag{13}$$

independently of the task diversity D.

Task-diversity threshold. By Eq. 10, the specialised solution emerges at $N_{\rm spec} = DK/T$. Hence, if

$$D \gg \frac{K}{2\,G(K,T)}, \tag{14}$$

then the in-context solution emerges before the specialised one, and persists in the window $(N_{\rm icl}, N_{\rm spec})$. The resulting picture is that of Fig. 1, right. In the $(N_{\rm icl}, N_{\rm spec})$ window the learner applies the same rule to seen and unseen topics alike, thus OOD and ID loss decay together towards the in-context level $\mathcal{L}_t^\star[\mathbf{r}]$. At $N_{\rm spec}$, the learner resolves the dictionary: the ID loss descends to the entropy rate, and the OOD loss rises towards $\mathcal{L}_t^\star[\mathbf{q}]$—OOD loss of the specialised predictor.

5 Transformer experiments

This section collects empirical learning curves of one-block transformers trained with Adam on data generated from random topic models of varying K and D. Experiments with varying T are reported in App. H. Since we are testing the scaling of performance with the number of training tokens, we conducted extensive hyperparameter search, looking for the best configuration of model size, learning rate, batch size and training time at fixed depth and optimiser. Further details of architectures and training algorithms are reported in App. G.

Fig. 3 shows that transformers learn random topic models at the same rate as the Bayesian learner of Sec. 3.

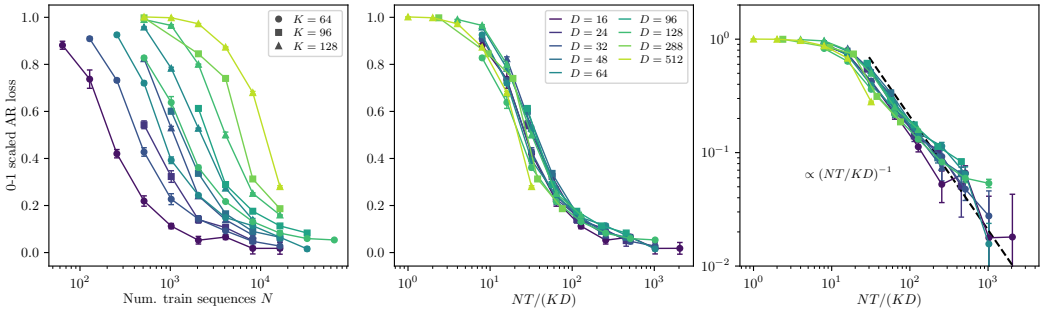


Figure 3: **Transformers match the Bayes-optimal specialisation learning curve. Left:** Scaled ID autoregressive loss $(\mathcal{L}_{\rm AR}(N) - \mathcal{L}_{\rm AR}(\infty))/(\log K - \mathcal{L}_{\rm AR}(\infty))$ versus number of training sequences for one-block transformers trained on random topic model data with varying (K, D), $T = 128$. The scaled loss is 1 at the uniform predictor $1/K$, 0 at the specialised predictor. **Middle:** Rescaling the x-axis as the number of tokens per parameter of the data-generating process collapses all the curves onto a single functional form. **Right:** The decay towards 0 (black dashed) shows remarkable agreement with the x^{-1} rate of the Bayes-optimal learning curve.

Even above the task-diversity threshold Eq. 14, specialisation kicks in as soon as the number of tokens NT surpasses the number of parameters DK. This is clearly shown by Fig. 4: the large-N decay of the ID loss scales with $(NT/(DK))$, while the early-N decay develops a D-independent behaviour. Correspondingly, the OOD losses at early stopping (dashed lines) show an initial descent, followed by a steep ascent around $N \sim (DK)/T$. This result shows that the transience of ICL is inevitable, unless the learning model is regularised.

Fig. 4 shows OOD learning curves in the high-diversity regime: the initial D-independent descent scales like $K^2/G(K,T)$ as predicted in Eq. 13. Since specialisation occurs at $NT \sim DK$, the

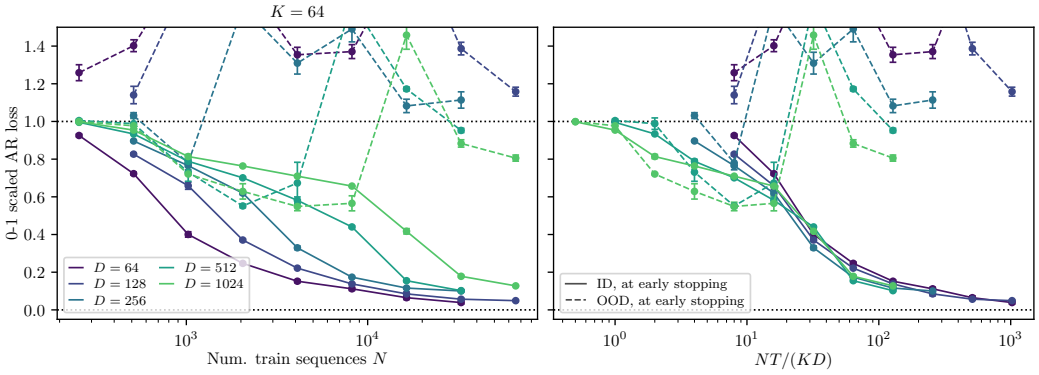


Figure 4: **In-context learning gives way to specialisation at the parametric scale $NT \sim DK$. Left:** Scaled ID (solid) and OOD (dashed) autoregressive losses, both subtracting the ID loss floor for comparison. **Right:** Rescaling the x-axis as the number of tokens per parameter of the data-generating process collapses both the descent of ID losses and the ascent of OOD losses.

in-context solution remains stable for a window of training set sizes linear in D. Fig. 11 of App. H shows the same scaling behaviour for the initial decay of ID learning curves.

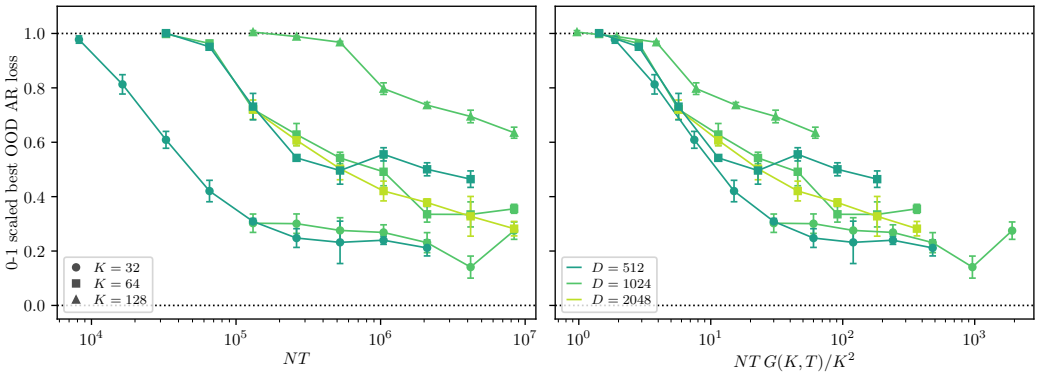


Figure 5: **Empirical OOD learning curves decay at the predicted sample scale. Left:** Scaled OOD autoregressive loss $(\mathcal{L}^\star_{\mathrm{AR}}(N) - \mathcal{L}^\star_{\mathrm{AR}}(\infty))/(\log K - \mathcal{L}^\star_{\mathrm{AR}}(\infty))$ versus number of training sequences for one-block transformers trained on random topic model data with varying (K, D), $T = 128$. These losses are not recorded at early stopping (training step with the lowest ID validation loss), but at the training step with the lowest OOD validation loss. Hence, the OOD losses shown in this figure do not rise at $N \sim DK/T$ as in Fig. 4. **Right:** The early-N departure from the trivial loss $\log K$ does not depend on D and collapse after dividing the number of training sequences by $K^2/G(K,T)$, in perfect agreement with the theory of Eq. 13.

5.1 Comparison with spectral baselines

Latent-variable models like the random topic model are classically learnt by spectral methods based on the moments of the tokens (Hsu et al., 2012; Balle et al., 2014; Anandkumar et al., 2014). The first step is recovery of the D-dimensional subspace spanned by the topics from a low-order moment: the pair moment $M_2 = \frac{1}{D}\sum_z \mathbf{p}_z \mathbf{p}_z^\top$, of rank D, when $D \leq K$, and the moment of $2m$ tokens when $D \leq \binom{K+m-1}{m}$, the number of independent products of m topic entries. The topics themselves then follow from the moment with one more token, which is the identifiability step of Sec. 3 (App. F). The subspace is resolved once the empirical moment, estimated from N independent $2m$-tuples, is close to the population moment. The corresponding sample size is a lower bound for the method to work.

The error on the empirical moment is of order $(K^m N)^{-1/2}$ by the matrix Bernstein inequality (App. F, verified numerically in Fig. 8). The scales of the population moments, instead, follow from random matrix theory, in the limit $D, K \to \infty$. With $D/K = \alpha < 1$ and $m = 1$, the bulk of M_2's spectrum converges to a Marchenko–Pastur distribution with eigenvalues of order $1/(DK)$ (Fig. 6). Comparison with the noise gives $N \gtrsim D^2 K$. With $D/(K(K+1)/2) = \beta < 1$ and $m = 2$, the relevant bulk of M_4 is of order $1/(DK^2)$ (Fig. 7). Comparison with the noise gives $N \gtrsim D^2 K^2$. This analysis is carried out in a different limit from the rest of the paper, and it counts independent tuples rather than documents of T tokens, so we only compare the dependence on task diversity: quadratic in D for the spectral baseline, against the linear $NT \sim DK$ of the Bayes-optimal learner, which the trained transformers follow. Furthermore, the order of K grows with the order of the required moment.

6 Conclusions

We have studied a solvable model of a mixture of D topics, where two strategies compete: in-distribution specialisation to the finite mixture, or out-of-distribution generalisation to topics never seen in training. Specialisation requires more training tokens than there are parameters in the data-generating process—a statistical limit that no learner can beat. The in-context solution is available at a D-independent scale, hence it comes first whenever the number of topics is large enough. Trained transformers follow both scales. It follows that the transience of ICL (Singh et al., 2023) is inevitable as the corpus grows, unless a constraint on parameters or training budget keeps the learner in the in-context window.

The specialised solution was called *memorising* in previous studies of the same competition (Raventos et al., 2023; Nguyen & Reddy, 2025; Carroll et al., 2025; Wurgaft et al., 2025). Indeed, its placement after the in-context phase mirrors the training dynamics of diffusion models, where generalisation is acquired before the memorisation of the training set (Favero et al., 2025). We find the label misleading: the specialised predictor stores the D topics of the population, not the N documents of the corpus—memorisation happens in *feature* space. This is a form of in-distribution generalisation, distinct from the adaptation to new topics of the in-context predictor, and we name it after the specialisation of teacher-student networks, where the hidden units of the student align with those of the teacher (Goldt et al., 2019; Barbier et al., 2025).

Several directions are open. The first is the training dynamics of transformers on the random topic model, which would connect our sample complexities with the phenomenological description of Wurgaft et al. (2025). More importantly, the dynamics determine which representations are learned in each phase and the corresponding geometry of the residual stream (Shai et al., 2024; Ray et al., 2026). Clarifying the distinction between representations in the two phases could help understand the emergence and role of task vectors (Hendel et al., 2023), including scenarios where multiple solutions coexist (Yan et al., 2026). The second is the competition between task diversity and other structures of the data. The tools developed here can address this problem in mixtures of general non-ergodic hidden Markov models (Riechers et al., 2025), of which the random topic model is the simplest instance. Such mixtures are natural models of large heterogeneous corpora. Solvable models like the one proposed here could help understand how heterogeneity contributes to data statistics and scaling laws (Cagnetta et al., 2026), and how improvements on the individual components of a mixture interact.

AI USE STATEMENT

In this work, we used generative AI tools for: assistance in the writing of proofs; implementation of numerical methods. We have not used generative AI tools for: Generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, proposing or refining hypotheses, designing feedback on research methodology or experiments, interpreting results, and assisting with translation, cleaning and reformatting dataset, supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for: modifying scientific figures, drafting parts of a research paper, identifying relevant literature. We have reviewed all AI-assisted work. We checked LLM-generated proofs; LLM-generated numerical implementations were verified and tested for correctness by 2 authors; We checked LLM-generated descriptions of additional related works. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

The theoretical results are stated with their assumptions and proved in the appendices. Appendices A–C prove the statements of Sec. 2: the averages over the topic prior, the optimality and loss decomposition of the two predictors, and the exact finite-t loss of the in-context predictor, Eq. 28, from which the in-context reference, the gain $G(K, T)$ and the sample scale N_{icl} used to rescale the axes of the figures are computed. App. D states the identifiability and regularity assumptions behind the DK/NT law, Eq. 10, gives sufficient conditions on (D, K, T) under which they hold almost surely, proves the law and gives its fixed-position version. App. E lists the conditions under which the estimation excess of a fitted family equals its parameter count, behind Eq. 13, and App. F states the joint limit in which the spectral scales hold. Arguments that we do not prove are identified as such where they appear: the concentration of pairwise divergences behind the $\log D$ constant, the signal-to-noise criterion and the block approximation of the spectral baseline, and the applicability of the estimation argument to a trained transformer.

The experiments are fully specified in App. G: the data-generating process, with the topic prior, the sizes of the validation and OOD sets and the number of fresh OOD topics; the architecture, initialisation and parameter count; the optimiser, learning-rate schedule and batching; the evaluation protocol, the definition of early stopping and the two readings of the OOD loss; and the hyperparameter grid with the selection rule. Every point of a learning curve is the mean over four runs with independent topic matrix, corpus and initialisation, with standard errors as error bars, and every run stores its full configuration, including the seeds of the draws, so that any point can be regenerated exactly. The theoretical references in the figures are evaluated from closed-form expressions or by Monte Carlo estimates over topic realisations, as stated in the captions. The code for data generation, training, evaluation and the theoretical curves will be made public upon de-anonymisation.

REFERENCES

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. URL https://openreview.net/forum?id=0g0X4H8yN4I.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009. URL http://www.jstor.org/stable/25662188.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.*, 15(1):2773–2832, 2014. doi: 10.5555/2627435.2697055. URL https://dl.acm.org/doi/10.5555/2627435.2697055.

Usman Anwar, Abulhair Saparov, Javier Rando, Daniel Paleka, Miles Turpin, Peter Hase, Ekdeep Singh Lubana, Erik Jenner, Stephen Casper, Oliver Sourbut, Benjamin L. Edelman, Zhaowei Zhang, Mario Günther, Anton Korinek, Jose Hernandez-Orallo, Lewis Hammond, Eric J Bigelow, Alexander Pan, Lauro Langosco, Tomasz Korbak, Heidi Chenyu Zhang, Ruiqi Zhong, Sean O hEigeartaigh, Gabriel Recchia, Giulio Corsi, Alan Chan, Markus Anderljung, Lilian Edwards, Aleksandar Petrov, Christian Schroeder de Witt, Sumeet Ramesh Motwani, Yoshua Bengio, Danqi Chen, Philip Torr, Samuel Albanie, Tegan Maharaj, Jakob Nicolaus Foerster, Florian Tramèr, He He, Atoosa Kasirzadeh, Yejin Choi, and David Krueger. Foundational challenges in assuring alignment and safety of large language models. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. URL https://openreview.net/forum?id=oVTkOs8Pka. Survey Certification, Expert Certification.

Borja Balle, William L. Hamilton, and Joelle Pineau. Methods of moments for learning stochastic languages: unified presentation and empirical comparison. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pp. 1386–1394. PMLR, 2014.

Jean Barbier, Francesco Camilli, Minh-Toan Nguyen, Mauro Pastore, and Rudy Skerk. Optimal generalisation and learning transition in extensive-width shallow neural networks near interpolation. *arXiv preprint arXiv:2501.18530*, 2025.

Marco Bondaschi, Nived Rajaraman, Xiuying Wei, Razvan Pascanu, Caglar Gulcehre, Michael Gastpar, and Ashok Vardhan Makkuva. From markov to laplace: How mamba in-context learns markov chains. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=kmK3WSCOCT.

Blake Bordelon, Mary Letey, and Cengiz Pehlevan. Theory of scaling laws for in-context regression: Depth, width, context and time. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=qA42mWsnbl.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/1457c0d6bfcb4967418bfb8ac142f64a-Abstract.html.

Francesco Cagnetta, Leonardo Petrini, Umberto M. Tomasini, Alessandro Favero, and Matthieu Wyart. How deep neural networks learn compositional data: The random hierarchy model. *Physical Review X*, 14:031001, 2024. doi: 10.1103/PhysRevX.14.031001. URL https://doi.org/10.1103/PhysRevX.14.031001.

Francesco Cagnetta, Allan Raventos, Surya Ganguli, and Matthieu Wyart. Deriving neural scaling laws from the statistics of natural language. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=VXlniX0C5y.

Liam Carroll, Jesse Hoogland, Matthew Farrugia-Roberts, and Daniel Murfet. Dynamics of transient structure in in-context linear regression transformers. *arXiv preprint arXiv:2501.17745*, 2025.

Djalil Chafaï and Konstantin Tikhomirov. On the convergence of the extremal eigenvalues of empirical covariance matrices with dependence. *Probability Theory and Related Fields*, 170:847–889, 2018. doi: 10.1007/s00440-017-0778-9. URL https://arxiv.org/abs/1509.02231.

Bertrand S. Clarke and Andrew R. Barron. Information-theoretic asymptotics of Bayes methods. *IEEE Transactions on Information Theory*, 36(3):453–471, 1990. doi: 10.1109/18.54897. URL https://bertrandsclarke.weebly.com/uploads/2/7/3/6/2736392/clarkebarron1990.pdf.

Bertrand S. Clarke and Andrew R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference*, 41(1):37–60, 1994. doi: 10.1016/0378-3758(94)90153-8. URL https://doi.org/10.1016/0378-3758(94)90153-8.

Ezra Edelman, Nikolaos Tsilivis, Benjamin L. Edelman, eran malach, and Surbhi Goel. The evolution of statistical induction heads: In-context learning markov chains. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=qaRT6QTIqJ.

Alessandro Favero, Antonio Sclocchi, and Matthieu Wyart. Bigger isn't always memorizing: Early stopping overparameterized diffusion models. *arXiv preprint arXiv:2505.16959*, 2025.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=flNZJ2eOet.

Sebastian Goldt, Madhu Advani, Andrew M. Saxe, Florent Krzakala, and Lenka Zdeborová. Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023. URL https://openreview.net/forum?id=QYvFUlF19n.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012.

Joseph B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear algebra and its applications*, 18(2):95–138, 1977. doi: 10.1016/0024-3795(77)90069-6. URL https://linkinghub.elsevier.com/retrieve/pii/0024379577900696.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122. URL https://www.pnas.org/doi/abs/10.1073/pnas.2502599122.

Alex Nguyen and Gautam Reddy. Differential learning kinetics govern the transition from memorization to generalization during in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=INyi7qUdjZ.

Eshaan Nichani, Alex Damian, and Jason D. Lee. How transformers learn causal structure with gradient descent. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=jNM4imlHZv.

Alain Pajor and Leonid Pastur. On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution. *Studia Mathematica*, 195(1):11–29, 2009. doi: 10.4064/sm195-1-2. URL https://arxiv.org/abs/0710.1346.

Core Francisco Park, Ekdeep Singh Lubana, and Hidenori Tanaka. Competition dynamics shape algorithmic phases of in-context learning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=XgH1wfHSX8.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BtAz4a5xDg.

Kyle J. Ray, Paul M. Riechers, and Adam S. Shai. The geometry of nonergodic composition. *Belief Updates (Simplex Blog)*, 2026. URL https://belief-updates.pub/nonergodic-geometry/.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

Paul M Riechers, Henry R Bigelow, Eric A Alt, and Adam Shai. Next-token pretraining implies in-context learning. *arXiv preprint arXiv:2505.18373*, 2025.

Adam Shai, Paul M. Riechers, Lucas Teixeira, Alexander Gietelink Oldenziel, and Sarah Marzen. Transformers represent belief state geometry in their residual stream. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIB7REL8UC.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. Role play with large language models. *Nature*, 623(7987):493–498, 2023.

Aaditya K Singh, Stephanie C.Y. Chan, Ted Moskovitz, Erin Grant, Andrew M Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=Of0GBzow8P.

Joel A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012. doi: 10.1007/s10208-011-9099-z. URL https://link.springer.com/content/pdf/10.1007/s10208-011-9099-z.pdf.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah Goodman. In-context learning strategies emerge rationally. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=bBUUOQI0N6.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Hao Yan, Haolin Yang, and Yiqiao Zhong. Task vector geometry underlies dual modes of task inference in transformers. *arXiv preprint arXiv:2605.03780*, 2026.

Yedi Zhang, Aaditya K Singh, Peter E. Latham, and Andrew M Saxe. Training dynamics of in-context learning in linear attention. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=aFNq67ilos.

A STATISTICS OVER RANDOM TOPIC MODELS

This appendix collects the averages over the topic prior used in the main text. Let $\mathbf{p}, \mathbf{p}' \sim \mathrm{Dir}_K(\mathbf{1})$ be independent, i.e. uniform on the simplex Δ^{K-1}, and let $\psi = \Gamma'/\Gamma$ denote the digamma function. Each entry is marginally $p_a \sim \mathrm{Beta}(1, K-1)$, with moments

$$\overline{p_a^r} = \frac{\Gamma(K)\,\Gamma(r+1)}{\Gamma(K+r)}, \qquad \overline{\log p_a} = \psi(1) - \psi(K), \qquad \overline{p_a \log p_a} = \frac{1}{K}\left[\psi(2) - \psi(K+1)\right], \tag{15}$$

where the last two follow by differentiating the first with respect to r at $r = 0$ and $r = 1$.

Entropy of a topic. By symmetry of the entries,

$$\overline{H(\mathbf{p})} = -K\,\overline{p_a \log p_a} = \psi(K+1) - \psi(2) = \log K - 1 + \gamma_{\mathrm{E}} + \frac{1}{2K} + O\big(K^{-2}\big), \tag{16}$$

using $\psi(2) = 1 - \gamma_{\mathrm{E}}$ and $\psi(K+1) = \log K + 1/(2K) + O(K^{-2})$. A random topic thus has entropy $1 - \gamma_{\mathrm{E}} - 1/(2K) \approx 0.42$ nats below the uniform value $\log K$, at leading orders: this gap is the height of the descent of every profile in Fig. 2.

Divergence between two topics. Expanding the divergence and using the independence of $\mathbf{p}$ and $\mathbf{p}'$,

$$\overline{D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{p}')} = K\,\overline{p_a \log p_a} - K\,\overline{p_a}\;\overline{\log p'_a} = \left[\psi(2) - \psi(K+1)\right] - \left[\psi(1) - \psi(K)\right] = 1 - \frac{1}{K}, \tag{17}$$

by the recurrence $\psi(x+1) = \psi(x) + 1/x$. Two random topics are thus separated by an order-one divergence for every vocabulary size, which sets the rate of topic resolution in App. B.

Bhattacharyya coefficient. The non-asymptotic resolution bound of App. B involves $\rho(\mathbf{p}, \mathbf{p}') := \sum_a \sqrt{p_a p'_a}$, whose average follows from Eq. 15 at $r = 1/2$,

$$\overline{\rho(\mathbf{p}, \mathbf{p}')} = K\left(\overline{\sqrt{p_a}}\right)^2 = \frac{\pi K}{4}\left[\frac{\Gamma(K)}{\Gamma(K+\frac{1}{2})}\right]^2 = \frac{\pi}{4}\left(1 + \frac{1}{4K} + O\big(K^{-2}\big)\right). \tag{18}$$

B THE SPECIALISED PREDICTOR

This appendix proves the statements of Sec. 2.1: the optimality of $\mathbf{q}_t$, the loss decomposition Eq. 4, and the resolution of the topic on the context scale $\log D$. Throughout, the topic matrix P is fixed, with distinct columns, and all expectations are over documents at fixed P.

Optimality. Conditional on the context, X_{t+1} is distributed as $\mathbf{q}_t(X_{\le t})$: by construction $q_t(a \mid X_{\le t}) = \sum_z \mathbb{P}(Z = z \mid X_{\le t})\,\mathbb{P}(X_{t+1} = a \mid Z = z)$, using that X_{t+1} is independent of the context given Z. For any predictor $\mathbf{f}$, therefore,

$$\mathbb{E}\left[-\log f_t(X_{t+1} \mid X_{\le t}) \,|\, X_{\le t}\right] = H(\mathbf{q}_t) + D_{\mathrm{KL}}(\mathbf{q}_t\|\mathbf{f}_t) \ge H(\mathbf{q}_t), \tag{19}$$

with equality if and only if $\mathbf{f}_t = \mathbf{q}_t$. Averaging over contexts, $\mathcal{L}_t[\mathbf{f}] \ge \mathcal{L}_t[\mathbf{q}]$.

Loss decomposition. Conditioning instead on the topic and the context, $X_{t+1} \sim \mathbf{p}_Z$, and the same identity gives $\mathbb{E}[-\log q_t(X_{t+1} \mid X_{\le t}) \mid Z, X_{\le t}] = H(\mathbf{p}_Z) + D_{\mathrm{KL}}(\mathbf{p}_Z\|\mathbf{q}_t)$; averaging over the topic, uniform on $\{1, \dots, D\}$, and over the context yields Eq. 4. The floor is the loss of the oracle $\mathbf{f}_t = \mathbf{p}_Z$, which is not a function of the context and hence not a predictor in the sense of Eq. 2; by the optimality of $\mathbf{q}_t$, every predictor pays at least the excess $\mathbb{E}[D_{\mathrm{KL}}(\mathbf{p}_Z\|\mathbf{q}_t)]$ above it.

Excess loss and wrong-topic mass. Write the posterior in terms of the likelihood ratios of the wrong topics against the true one,

$$R_y^{(t)} := \frac{\gamma_t(y)}{\gamma_t(Z)} = \prod_{s=1}^{t} \frac{p_y(X_s)}{p_Z(X_s)}, \qquad \gamma_t(Z) = \Big(1 + \sum_{y \ne Z} R_y^{(t)}\Big)^{-1}. \tag{20}$$

Isolating the true topic in Eq. 3, $\mathbf{q}_t = \gamma_t(Z)\,\mathbf{p}_Z + \sum_{y\neq Z}\gamma_t(y)\,\mathbf{p}_y \geq \gamma_t(Z)\,\mathbf{p}_Z$ entrywise, hence

$$D_{\mathrm{KL}}(\mathbf{p}_Z\|\mathbf{q}_t) \leq \sum_{a=1}^{K} p_Z(a)\log\frac{p_Z(a)}{\gamma_t(Z)\,p_Z(a)} = -\log\gamma_t(Z) = \log\Big(1+\sum_{y\neq Z} R_y^{(t)}\Big). \tag{21}$$

Conditional on Z, the tokens are i.i.d. from $\mathbf{p}_Z$, so by the strong law of large numbers

$$\frac{1}{t}\log R_y^{(t)} = \frac{1}{t}\sum_{s=1}^{t}\log\frac{p_y(X_s)}{p_Z(X_s)} \xrightarrow[t\to\infty]{\text{a.s.}} -D_{\mathrm{KL}}(\mathbf{p}_Z\|\mathbf{p}_y) < 0, \tag{22}$$

i.e. each wrong topic is suppressed exponentially in t, at a rate equal to its divergence from the true topic. This is the mechanism described in the main text; it is asymptotic in t at fixed P.

Resolution on the scale $\log D$**.** A bound valid at every t follows from a fractional moment of the likelihood ratios. Using $\log(1+\sum_y R_y) \leq \sum_y \log(1+R_y) \leq \sum_y \sqrt{R_y}$ in Eq. 21, then the independence of the tokens given Z,

$$\begin{aligned}\mathbb{E}\left[D_{\mathrm{KL}}(\mathbf{p}_Z\|\mathbf{q}_t)\right] \leq \mathbb{E}\left[-\log\gamma_t(Z)\right] &\leq \frac{1}{D}\sum_{z=1}^{D}\sum_{y\neq z}\mathbb{E}\Big[\sqrt{R_y^{(t)}}\;\Big|\; Z=z\Big] \\ &= \frac{1}{D}\sum_{z=1}^{D}\sum_{y\neq z}\rho_{zy}^{t} \leq (D-1)\,\rho_\star^{t},\end{aligned} \tag{23}$$

where $\rho_{zy} := \sum_a \sqrt{p_z(a)\,p_y(a)}$ is the Bhattacharyya coefficient of topics z and y, and $\rho_\star := \max_{z\neq y}\rho_{zy} < 1$ by the Cauchy–Schwarz inequality and the distinctness of the topics. Hence the excess loss of the specialised predictor is below any $\epsilon > 0$ as soon as

$$t \geq \frac{\log\left[(D-1)/\epsilon\right]}{\log(1/\rho_\star)}, \tag{24}$$

which is the $\log D$ scale of Sec. 2.1, with a rate set by the closest pair of topics. The guaranteed rate $\log(1/\rho_{zy})$ is at most half of the almost-sure rate in Eq. 22, since $\log\rho_{zy} \geq -\frac{1}{2}D_{\mathrm{KL}}(\mathbf{p}_z\|\mathbf{p}_y)$ by Jensen's inequality, and $\rho_\star$ grows with D as the closest pair among D random topics gets closer: Eq. 24 is a statement about the dependence on t at fixed P, not a sharp constant. For the constant, the typical pair is the relevant one: two random topics have $\overline{D_{\mathrm{KL}}(\mathbf{p}\|\mathbf{p}')} = 1 - 1/K$ and $\overline{\rho(\mathbf{p},\mathbf{p}')} \simeq \pi/4$ (App. A), and if the $D-1$ divergences from the true topic concentrate near their mean, the odds of a wrong topic in Eq. 21 are approximately $(D-1)\,e^{-(1-1/K)\,t}$, so that the floor is reached for $t \approx \log(D-1)$. This last estimate is heuristic, as it assumes a concentration of the pairwise divergences that we do not prove; it is consistent with the half-descent points in Fig. 2, which grow linearly in $\log D$.

C THE IN-CONTEXT PREDICTOR

This appendix proves the statements of Sec. 2.2: the form and optimality of $\mathbf{r}_t$, the loss decomposition Eq. 6, and the exact finite-t risk behind its expansion. Throughout, the document is OOD: $\mathbf{p}_\star \sim \mathrm{Dir}_K(\mathbf{1})$ and, given $\mathbf{p}_\star$, the tokens are i.i.d. from $\mathbf{p}_\star$.

Posterior and add-one rule. The prior density is constant on the simplex and the likelihood of a context is $\prod_a p_\star(a)^{C_a}$, so the posterior density is proportional to $\prod_a p_\star(a)^{C_a}$: this is $\mathrm{Dir}_K(C_1{+}1,\ldots,C_K{+}1)$, whose density is proportional to $\prod_a p_a^{\alpha_a-1}$ with $\alpha_a = C_a{+}1$. Given $\mathbf{p}_\star$ and the context, $X_{t+1} \sim \mathbf{p}_\star$, and the mean of $\mathrm{Dir}_K(\boldsymbol{\alpha})$ is $\alpha_a/\sum_b \alpha_b$, so the posterior predictive distribution is

$$r_t(a \mid X_{\leq t}) = \mathbb{E}\left[p_\star(a) \mid X_{\leq t}\right] = \frac{C_a+1}{t+K}, \tag{25}$$

which is Eq. 5.

Optimality. As in App. B, $\mathbf{r}_t$ is the conditional distribution of X_{t+1} given the context, so for any predictor $\mathbf{f}$

$$\mathbb{E}\left[-\log f_t(X_{t+1} \mid X_{\le t}) \,|\, X_{\le t}\right] = H(\mathbf{r}_t) + D_{\mathrm{KL}}(\mathbf{r}_t \| \mathbf{f}_t) \ge H(\mathbf{r}_t), \tag{26}$$

with equality if and only if $\mathbf{f}_t = \mathbf{r}_t$; averaging over contexts, $\mathcal{L}_t^\star[\mathbf{f}] \ge \mathcal{L}_t^\star[\mathbf{r}]$. Unlike $\mathbf{q}_t$, the predictor $\mathbf{r}_t$ does not depend on the topic matrix: on a document generated from any topic $\mathbf{p}$, its loss is a function $R_t(\mathbf{p}) := \mathbb{E}[-\log r_t(X_{t+1} \mid X_{\le t}) \mid \mathbf{p}]$ of that topic alone. Its ID loss at fixed P is $\mathcal{L}_t[\mathbf{r}] = \frac{1}{D}\sum_z R_t(\mathbf{p}_z)$ and its OOD loss is $\mathcal{L}_t^\star[\mathbf{r}] = \overline{R_t(\mathbf{p})}$, whence $\overline{\mathcal{L}_t[\mathbf{r}]} = \mathcal{L}_t^\star[\mathbf{r}]$.

Loss decomposition. Conditioning on $\mathbf{p}_\star$ and the context, $X_{t+1} \sim \mathbf{p}_\star$, so $\mathbb{E}[-\log r_t(X_{t+1} \mid X_{\le t}) \mid \mathbf{p}_\star, X_{\le t}] = H(\mathbf{p}_\star) + D_{\mathrm{KL}}(\mathbf{p}_\star \| \mathbf{r}_t)$; averaging over $\mathbf{p}_\star$ and the context gives the first equality in Eq. 6, with $\mathbb{E}[H(\mathbf{p}_\star)] = \overline{H(\mathbf{p})}$ because $\mathbf{p}_\star$ has the law of the training topics. Moreover, by Eq. 26 at $\mathbf{f} = \mathbf{r}$, $\mathcal{L}_t^\star[\mathbf{r}] = H(X_{t+1} \mid X_{\le t})$ is a conditional entropy; since the tokens are exchangeable given $\mathbf{p}_\star$, $H(X_{t+2} \mid X_{\le t}) = H(X_{t+1} \mid X_{\le t})$, and conditioning on one more token shows that $\mathcal{L}_t^\star[\mathbf{r}]$ is non-increasing in t.

Exact finite-t risk. Integrating the likelihood against the uniform prior, the count vector of a context is uniform over the $\binom{t+K-1}{K-1}$ weak compositions of t into K parts,

$$\mathbb{P}\left(C(X_{\le t}) = \mathbf{n}\right) = \frac{t!}{\prod_a n_a!}\,\frac{\Gamma(K)\prod_a \Gamma(n_a{+}1)}{\Gamma(t{+}K)} = \binom{t+K-1}{K-1}^{-1}, \qquad \textstyle\sum_a n_a = t. \tag{27}$$

Hence, by Eq. 25 and the symmetry of the entries, for $K \ge 2$

$$\begin{aligned}\mathcal{L}_t^\star[\mathbf{r}] &= \log(t+K) - \mathbb{E}\left[\sum_a r_t(a \mid X_{\le t}) \log(C_a + 1)\right] \\ &= \log(t+K) - \frac{K}{(t+K)\binom{t+K-1}{K-1}} \sum_{n=0}^{t} (n+1)\log(n+1)\binom{t-n+K-2}{K-2}, \end{aligned} \tag{28}$$

where $\binom{t-n+K-2}{K-2}$ counts the compositions with $C_1 = n$. In particular $\mathcal{L}_0^\star[\mathbf{r}] = \log K$ and $\mathcal{L}_t^\star[\mathbf{r}] \to \overline{H(\mathbf{p})}$ as $t \to \infty$. Expanding at large t and fixed K gives the second equality in Eq. 6,

$$\mathcal{L}_t^\star[\mathbf{r}] = \psi(K{+}1) - \psi(2) + \frac{K-1}{2t} + O_K\!\left(\frac{\log t}{t^2}\right), \tag{29}$$

where the logarithm in the remainder comes from the mass that the uniform prior puts near the boundary of the simplex. The leading correction is one half of the number of parameters of a topic, $K-1$, per observation, and the expansion requires $t \gg K$: at $K = 64$ and $t = 128$, Eq. 28 gives an excess of 0.15 nats over the floor, against 0.25 for the leading term alone; the former is the gap between the dashed line and the floor at the end of the profiles in Fig. 2.

D The Bayesian learner

This appendix proves the statements of Sec. 3: the form and optimality of the Bayesian learner, the information identity, the DK/NT law and its fixed-position version.

Posterior and predictor. Let $\mathcal{C}_N = (X_{1:T}^{(n)})_{n=1}^N$ be the corpus, with count vectors $C^{(n)}$, and write $\ell_C(\mathbf{p}) := \prod_a p(a)^{C_a}$ for the likelihood of a count vector under a topic. Given a candidate topic matrix $\tilde{P}$, integrating out the latent topic of each document gives the likelihood of corpus and context,

$$L(\tilde{P}) := \prod_{n=1}^{N}\Big[\frac{1}{D}\sum_{z=1}^{D}\ell_{C^{(n)}}(\tilde{\mathbf{p}}_z)\Big] \times \frac{1}{D}\sum_{z=1}^{D}\ell_{C(X_{\le t})}(\tilde{\mathbf{p}}_z), \tag{30}$$

up to multinomial factors that do not depend on $\tilde{P}$. The posterior of P is $\pi(\mathrm{d}\tilde{P}) \propto L(\tilde{P})\,\mu(\mathrm{d}\tilde{P})$, with μ the product of D uniform priors on the simplex, and Eq. 7 reads $\hat{\mathbf{q}}_t = \int \mathbf{q}_{t,\tilde{P}}(X_{\le t})\,\pi(\mathrm{d}\tilde{P})$, where $\mathbf{q}_{t,\tilde{P}}$ is the specialised predictor of Eq. 3 computed with $\tilde{P}$. By construction $\hat{\mathbf{q}}_t$ is the conditional law of X_{t+1} given $(\mathcal{C}_N, X_{\le t})$ under the joint law of topics, corpus and test document, so, as in App. B, every predictor $\mathbf{f}$ built from corpus and context satisfies $\overline{\mathcal{L}_t[\mathbf{f}]} \ge \overline{\mathcal{L}_t[\hat{\mathbf{q}}]}$, with equality if and only if $\mathbf{f}_t = \hat{\mathbf{q}}_t$.

Excess and mutual information. At fixed P, conditioning on the context, $X_{t+1} \sim \mathbf{q}_t$ and the corpus is independent of the test document, so $\mathcal{L}_t[\hat{\mathbf{q}}] - \mathcal{L}_t[\mathbf{q}] = \mathbb{E}[D_{\mathrm{KL}}(\mathbf{q}_t \| \hat{\mathbf{q}}_t)] \geq 0$. Averaged over P, the two losses are the conditional entropies $H(X_{t+1} \mid X_{\leq t}, \mathcal{C}_N)$ and $H(X_{t+1} \mid X_{\leq t}, \mathcal{C}_N, P)$, whence

$$\varepsilon_t(N) = I(P; X_{t+1} \mid X_{\leq t}, \mathcal{C}_N). \tag{31}$$

The second entropy does not change if one more document $X_{1:T}^{(N+1)}$ is added to the corpus, since given P the corpus is irrelevant, so $\varepsilon_t(N) - \varepsilon_t(N+1) = I(X_{t+1}; X_{1:T}^{(N+1)} \mid X_{\leq t}, \mathcal{C}_N) \geq 0$: the learning curve is non-increasing in N. Summing Eq. 31 over positions with the chain rule, and using that the test document has the law of $X_{1:T}^{(N+1)}$ and is independent of the corpus given P,

$$\sum_{t=0}^{T-1} \varepsilon_t(N) = I\big(P; X_{1:T}^{(N+1)} \,\big|\, \mathcal{C}_N\big) = I(P; \mathcal{C}_{N+1}) - I(P; \mathcal{C}_N), \tag{32}$$

which is Eq. 9, the document-level form of the cumulative predictive risk identity of Clarke & Barron (1990).

The DK/NT law. Up to a relabelling of the topics, the law $Q_P := \frac{1}{D}\sum_z \mathbf{p}_z^{\otimes T}$ of a document has $d = D(K-1)$ parameters. Assume that $P \mapsto Q_P$ is identifiable up to relabelling and regular: twice continuously differentiable in local coordinates, with positive definite Fisher information, and with a prior density that is positive and continuous. Then, for fixed D, K, T and $N \to \infty$, Theorem 1 of Clarke & Barron (1994) gives

$$I(P; \mathcal{C}_N) = \frac{d}{2}\log N + c + o(1), \tag{33}$$

with a constant c that depends on prior and document model but not on N. The increment in Eq. 32 does not follow by direct subtraction, since an $o(1)$ term need not have $o(N^{-1})$ increments, but it does follow from monotonicity. Let $a_N := \sum_t \varepsilon_t(N)$, non-increasing in N, and for $\lambda > 1$ let $m = \lfloor \lambda N \rfloor$ and $l = \lfloor N/\lambda \rfloor$. Summing a_n over $N \leq n < m$ and over $l \leq n < N$ and using monotonicity,

$$\frac{I(P; \mathcal{C}_m) - I(P; \mathcal{C}_N)}{m - N} \;\leq\; a_N \;\leq\; \frac{I(P; \mathcal{C}_N) - I(P; \mathcal{C}_l)}{N - l}, \tag{34}$$

where both numerators equal $\frac{d}{2}\log\lambda + o(1)$ by Eq. 33. Letting $N \to \infty$ and then $\lambda \downarrow 1$ squeezes $N a_N$ between $\frac{d}{2}\frac{\log\lambda}{\lambda - 1}$ and $\frac{d}{2}\frac{\lambda\log\lambda}{\lambda - 1}$, both of which tend to $d/2$; dividing by T gives Eq. 10.

Identifiability. For $T \geq 3$ the third-order moment of the tokens, $M_3 = \frac{1}{D}\sum_z \mathbf{p}_z^{\otimes 3}$, is a marginal of Q_P, and by Kruskal's theorem (Kruskal, 1977; Allman et al., 2009) its decomposition into D rank-one terms is unique up to permutation when $3\,\mathrm{rank_K}(P) \geq 2D + 2$, where the Kruskal rank $\mathrm{rank_K}(P)$ is the largest k such that any k columns of P are linearly independent. For topics drawn from the simplex, $\mathrm{rank_K}(P) = \min\{K, D\}$ almost surely, so the condition holds for $D \leq K$ (and up to $D \leq 3K/2 - 1$), and the topics can be recovered constructively: the second moment $M_2 = \frac{1}{D}\sum_z \mathbf{p}_z \mathbf{p}_z^\top$ has rank D and whitens M_3 into an orthogonally decomposable tensor whose factors are the topics (Anandkumar et al., 2014). For larger D, lift the topics to $\mathbf{v}_z = \mathbf{p}_z^{\otimes m}$, which lie in the space of symmetric tensors, of dimension $s_m = \binom{K+m-1}{m}$, and span it as $\mathbf{p}_z$ ranges over the simplex, since a homogeneous polynomial vanishing on the simplex vanishes identically; some $s_m \times s_m$ minor of $[\mathbf{v}_1, \ldots, \mathbf{v}_D]$ is then a nonzero polynomial in the topic entries, so the lifted topics are linearly independent almost surely whenever $D \leq s_m$. In that case the $2m$-token moment $\frac{1}{D}\sum_z \mathbf{v}_z \mathbf{v}_z^\top$ has rank D, the rank of the Hankel matrix of the token process, and together with the $(2m{+}1)$-token moment it determines the law of documents of every length, hence P up to relabelling (Balle et al., 2014) (App. F). Moreover, for $T \geq 2m + 1$ the Fisher information of a document is nonsingular: a perturbation δP that leaves the law of a document unchanged to first order leaves the $(2m{+}1)$-token marginal $\frac{1}{D}\sum_z \mathbf{v}_z \otimes \mathbf{v}_z \otimes \mathbf{p}_z$ unchanged, and contracting its variation with $a_z \otimes a_z \otimes \mathrm{Id}$, where the functionals a_z satisfy $a_z(\mathbf{v}_{z'}) = \delta_{zz'}$, gives $\delta\mathbf{p}_z + 2a_z(\delta\mathbf{v}_z)\,\mathbf{p}_z = 0$, whose sum over coordinates forces $a_z(\delta\mathbf{v}_z) = 0$ and then $\delta\mathbf{p}_z = 0$. Hence $D \leq \binom{K+m-1}{m}$ and $T \geq 2m + 1$ suffice almost surely: $D \leq K$ with $T \geq 3$, $D \leq K(K{+}1)/2$ with $T \geq 5$, and so on; these conditions are sufficient, not necessary, and they are the only place where the genericity of the topic matrix is used. Regularity additionally excludes the sets, of prior measure zero, where two topics coincide or a topic lies on the boundary of the simplex.

Fixed position: the resolved learner. Reveal to the learner the topic labels of the training documents and of the test document. Conditional on the labels, by the conjugacy of App. C, the Bayesian predictor is the add-one rule on the pooled counts of the test topic,

$$\hat{q}_t^{\mathrm{res}}(a \mid X_{\le t}) = \frac{M_{Z,a} + C_a + 1}{N_Z T + t + K}, \tag{35}$$

where N_Z is the number of training documents of topic Z and $M_{Z,a}$ their total count of token a: the in-context predictor of Eq. 5 applied to $m = N_Z T + t$ tokens. Averaged over topics, its excess over the entropy rate $\frac{1}{D}\sum_z H(\mathbf{p}_z)$ is therefore $\mathbb{E}[g_K(N_Z T + t)]$, with $g_K(m) := \mathcal{L}_m^\star[\mathbf{r}] - \overline{H(\mathbf{p})}$ the excess of the in-context predictor in Eq. 6 and $N_Z \sim \mathrm{Bin}(N, 1/D)$. If $N/D \to \infty$ then $N_Z/(N/D) \to 1$ in probability, and if moreover $tD/(NT) \to 0$ and $\log T = o(N/D)$, so that the event that the test topic is absent from the corpus is negligible on the $1/(NT)$ scale, Eq. 6 gives

$$\varepsilon_t^{\mathrm{res}}(N) := \mathbb{E}\left[g_K(N_Z T + t)\right] = \frac{D(K-1)}{2NT}\big(1 + o(1)\big). \tag{36}$$

The resolved learner is not the Bayesian learner, which must also infer the labels and the topic matrix. For a fixed P with distinct topics, and assuming identifiability from a marginal of $k \le T$ tokens, the two excesses differ by at most

$$C(P)\left(e^{-c(P)\,t} + N e^{-c(P)\,T}\right) + \exp\big\{-a(P)\,N + O_P\big(\log[N(T+t)]\big)\big\}, \tag{37}$$

with $c(P) = \log(1/\rho_\star)$ from Eq. 23 and $a(P) > 0$. The first two terms are the costs of unresolved test and training labels, by the bound of App. B applied to the test context and to each of the N training documents; the last is the posterior mass on topic matrices outside a fixed neighbourhood of P, exponentially small in N by a likelihood-ratio test on the identifying marginal, against the polynomial prior mass of the neighbourhood. Hence the Bayesian learner obeys Eq. 36 at fixed P whenever t and T grow like $A \log N$ with A large enough for Eq. 37 to be $o(1/(NT))$: the fixed-position form of the DK/NT law. The statement is conditional on P; averaging it over the prior would require uniform control over nearly coinciding topics, for which $c(P)$ is arbitrarily small.

E THE IN-CONTEXT SOLUTION

This appendix supports Sec. 4: the estimation excess of a fitted parametric family, the count-linear family that contains the in-context predictor, and the gain $G(K,T)$ with its interpretation and expansions.

Estimation excess of a fitted family. Let $\{f_\theta : \theta \in \Theta \subset \mathbb{R}^d\}$ be a family of distributions of a document $Y = X_{1:T}$, with all d coordinates of θ free, and let $\ell(\theta; Y) := -\log f_\theta(Y)$ and $\mathcal{R}(\theta) := \mathbb{E}[\ell(\theta; Y)]$ be its log loss and its population risk under the true law of Y. Let $\theta_\star$ be an interior minimiser of $\mathcal{R}$, isolated up to finite symmetries, around which $f_\theta(y) > 0$ is three times continuously differentiable, and define the risk curvature and the gradient covariance

$$\mathsf{H} := \nabla^2 \mathcal{R}(\theta_\star) \succ 0, \qquad \mathsf{J} := \mathbb{E}\left[\nabla \ell(\theta_\star; Y)\, \nabla \ell(\theta_\star; Y)^\top\right]; \tag{38}$$

the gradient has zero mean because $\theta_\star$ is a minimiser. Let $\hat{\theta}_N$ be a consistent fit on N independent documents whose empirical gradient is $o_{\mathbb{P}}(N^{-1/2})$, as for an interior minimiser of the empirical risk $\widehat{\mathcal{R}}_N$, and let $g_N := \nabla \widehat{\mathcal{R}}_N(\theta_\star)$, which has mean zero and covariance J/N by independence. Expanding the empirical gradient and the population risk around $\theta_\star$,

$$\hat{\theta}_N - \theta_\star = -\mathsf{H}^{-1} g_N + u_N, \qquad \mathcal{R}(\hat{\theta}_N) - \mathcal{R}(\theta_\star) = \tfrac{1}{2}(\hat{\theta}_N - \theta_\star)^\top \mathsf{H}\,(\hat{\theta}_N - \theta_\star) + v_N, \tag{39}$$

with $u_N = o_{\mathbb{P}}(N^{-1/2})$ and $v_N = o_{\mathbb{P}}(N^{-1})$. If the remainders are also small in expectation, $\mathbb{E}\|u_N\|^2 = o(N^{-1})$ and $\mathbb{E}|v_N| = o(N^{-1})$, then substituting and using $\mathbb{E}[g^\top A g] = \mathrm{tr}(A \,\mathrm{Cov}\, g)$ for centred g,

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_N) - \mathcal{R}(\theta_\star)\right] = \frac{d_{\mathrm{eff}}}{2N} + o(N^{-1}), \qquad d_{\mathrm{eff}} := \mathrm{tr}\left(\mathsf{J}\mathsf{H}^{-1}\right). \tag{40}$$

If the family is well specified, i.e. the true law is $f_{\theta_\star}$, differentiating $\sum_y f_\theta(y) = 1$ twice gives $\mathsf{H} = \mathsf{J}$, the Fisher information of one document, and $d_{\mathrm{eff}} = d$ whenever it is nonsingular. The remainder

bounds hold, for instance, for an empirical-risk minimiser over a compact set that contains $\theta_\star$ in its interior and on which probabilities are bounded below; they are not automatic for an unrestricted fit, nor for a transformer. Divided by T, Eq. 40 is a per-token excess $d/(2NT)$. For the topic matrix itself, $d = D(K-1)$, and the Fisher information is nonsingular almost surely for $D \le K$ and $T \ge 3$ (the Jacobian of the three-token marginal is injective when the topics are linearly independent), so a fitted dictionary recovers Eq. 10 by an argument on a fitted parameter rather than on the posterior; for the count map below, $d \simeq K^2$.

The count-linear family. The predictors that are linear in the counts up to normalisation,

$$f_{A,\mathbf{b}}(a \mid X_{\le t}) := \frac{\big(A\,C(X_{\le t}) + \mathbf{b}\big)_a}{\mathbf{1}^\top\big(A\,C(X_{\le t}) + \mathbf{b}\big)}, \qquad A \in \mathbb{R}^{K\times K},\ \mathbf{b} \in \mathbb{R}^K, \tag{41}$$

contain the in-context predictor at $(A, \mathbf{b}) = (I, \mathbf{1})$. Counts are bounded by T, so the family is positive and smooth on a neighbourhood of this point, and it has $d = K^2 + K - 1$ free parameters, the overall scale of $(A, \mathbf{b})$ being immaterial. On OOD documents the family is well specified, since the law of an OOD document is the product of its add-one predictors (App. C), so $\theta_\star = (I, \mathbf{1})$ and Eq. 40 gives the excess $d/(2NT) \simeq K^2/(2NT)$ of Sec. 4 for a learner that fits this family. On the training distribution the family is misspecified at finite D, its optimum being a fit to $\mathbf{q}_t$, a ratio of exponentials of the counts. The misspecification vanishes as $D \to \infty$: writing Eq. 3 as a ratio of averages over the dictionary,

$$q_t(a \mid X_{\le t}) = \frac{\frac{1}{D}\sum_z p_z(a)\prod_b p_z(b)^{C_b}}{\frac{1}{D}\sum_z \prod_b p_z(b)^{C_b}} \xrightarrow[D\to\infty]{\text{a.s.}} \frac{\overline{p(a)\prod_b p(b)^{C_b}}}{\overline{\prod_b p(b)^{C_b}}} = \frac{C_a + 1}{t + K} = r_t(a \mid X_{\le t}), \tag{42}$$

by the strong law of large numbers applied to numerator and denominator at fixed K and count vector, and the Dirichlet moment $\overline{\prod_b p(b)^{n_b}} = \Gamma(K)\prod_b \Gamma(n_b{+}1)/\Gamma(K{+}\sum_b n_b)$, of which Eq. 15 is the one-entry case. The convergence is pointwise in the count vector, and the dictionary size needed for a given accuracy grows with t, since the weights in Eq. 42 concentrate on a neighbourhood of the empirical frequencies that shrinks with t. Note that the family cannot be reached through the specialised predictor of a small dictionary. $\mathbf{q}_t$ is a convex combination of the D topics, so its Jacobian with respect to the counts has rank at most $D - 1$, whereas that of the add-one map is $K - 1$: a learner whose predictor tracks the in-distribution optimum can realise the count-linear rule only if $D \ge K$.

The gain of the in-context predictor. On an OOD document every token is marginally uniform, $\mathbb{P}(X_{t+1}{=}a) = \overline{p(a)} = 1/K$, so $H(X_{t+1}) = \log K$, and $\mathcal{L}_t^\star[\mathbf{r}] = H(X_{t+1} \mid X_{\le t})$ by App. C; hence

$$\log K - \mathcal{L}_t^\star[\mathbf{r}] = I(X_{t+1}; X_{\le t}), \tag{43}$$

$$T\,G(K,T) = T\log K - H(X_{1:T}) = T\left[\log K - \overline{H(\mathbf{p})}\right] - I(\mathbf{p}_\star; X_{1:T}), \tag{44}$$

by the chain rule, and $H(X_{1:T}) = H(X_{1:T} \mid \mathbf{p}_\star) + I(\mathbf{p}_\star; X_{1:T})$ with $H(X_{1:T} \mid \mathbf{p}_\star) = T\,\overline{H(\mathbf{p})}$. The gain per token is thus the entropy gap between the uniform distribution and a random topic, $1 - \gamma_{\mathrm{E}} - 1/(2K) + O(K^{-2})$ by Eq. 16, minus the information about its own topic that the document leaves unrevealed. Its exact value follows from Eq. 28, and its two regimes are

$$\begin{aligned} G(K,T) &= \log K - \overline{H(\mathbf{p})} - \frac{K-1}{2}\,\frac{\log T}{T} + O_K\!\left(\frac{1}{T}\right) && (T \gg K),\\ G(K,T) &= (2\log 2 - 1)\,\frac{T-1}{2K} + O\!\left(\frac{T^2}{K^2}\right) && (K \gg T). \end{aligned} \tag{45}$$

The first follows by summing Eq. 29 over positions, which gives $I(\mathbf{p}_\star; X_{1:T}) = \frac{K-1}{2}\log T + O_K(1)$: a long context reveals its topic up to the usual parametric cost. The second follows from Eq. 28 at $t \ll K$: a given token appears once in the context with probability $\mathbb{P}(C_a{=}1) = (K-1)\,t/[(t+K-1)(t+K-2)] \simeq t/K$ and more than once with probability $O(t^2/K^2)$, so that $\mathcal{L}_t^\star[\mathbf{r}] = \log(t{+}K) - \frac{K}{t+K}\,2\log 2\;\mathbb{P}(C_a{=}1) + O(t^2/K^2) = \log K - (2\log 2{-}1)\,t/K + O(t^2/K^2)$; in a short context most tokens are new, and the gain comes from the few repeated ones.

F THE SPECTRAL BASELINE

This appendix supports Sec. 5.1: the Hankel formulation of spectral learning, the spectra of the pair and four-token moments under the random topic model, the noise of their empirical estimates, and the resulting sample complexities. Its asymptotic statements are in the joint limit $D, K \to \infty$ with D/K, respectively D/K^2, fixed.

Hankel matrix and subspace. For the token process, the Hankel matrix has entries $H(u, v) := \mathbb{P}(X_{1:|u|+|v|} = uv)$, indexed by a prefix string u and a suffix string v; since tokens are i.i.d. given the topic, $H(u, v) = \frac{1}{D}\sum_z g_z(u)\, g_z(v)$ with $g_z(u) = \prod_s p_z(u_s)$, so that $\operatorname{rank} H \le D$ (Balle et al., 2014). Spectral learning first recovers the column space of H from a complete sub-block, i.e. one of rank D, then the observation operators from the block with one more symbol; for the random topic model the hidden state is constant within a document, the transition matrix is the identity and the observation matrix is P (Hsu et al., 2012). The block on prefixes and suffixes of m tokens is $M_{2m} = \frac{1}{D}\sum_z (\mathbf{p}_z^{\otimes m})(\mathbf{p}_z^{\otimes m})^\top$, whose nonzero spectrum is that of the $D \times D$ Gram matrix $\frac{1}{D}[(\mathbf{p}_i^\top \mathbf{p}_j)^m]_{ij}$. The lifted topics $\mathbf{p}_z^{\otimes m}$ lie in the space of symmetric tensors, of dimension $s_m = \binom{K+m-1}{m}$, and span it as $\mathbf{p}_z$ ranges over the simplex, since a homogeneous polynomial vanishing on the simplex vanishes identically; some $s_m \times s_m$ minor of the lifted topic matrix is therefore a nonzero polynomial in the topic entries, and $\operatorname{rank} M_{2m} = \min\{D, s_m\}$ almost surely for topics drawn from the simplex. The single-token block M_2 is thus complete for $D \le K$, the two-token block M_4 for $D \le K(K+1)/2$, and so on; the topics follow from M_3, M_5, …, for $D \le K$ through the whitening of M_3 by M_2 and the decomposition of the resulting orthogonal tensor (Anandkumar et al., 2014), as recalled in App. D.

Spectrum of M_2. By Eq. 15, $\overline{p_a p_b} = 1/(K(K+1))$ for $a \ne b$ and $\overline{p_a^2} = 2/(K(K+1))$, so the mean of M_2 over the topics is $(I + \mathbf{1}\mathbf{1}^\top)/(K(K+1))$, with one eigenvalue $1/K$ along $\mathbf{1}$ and $1/(K(K+1))$ on its complement. For a realisation, write $\mathbf{p}_z = \mathbf{1}/K + \boldsymbol{\xi}_z$ with $\boldsymbol{\xi}_z$ centred, orthogonal to $\mathbf{1}$, of covariance $\frac{1}{K(K+1)}(I - \mathbf{1}\mathbf{1}^\top/K)$ and independent across topics; the nonzero spectrum of $M_2 = \frac{1}{D}PP^\top$ is that of $\frac{1}{D}P^\top P = \frac{1}{D}\left[\frac{1}{K}\mathbf{1}\mathbf{1}^\top + \Xi^\top\Xi\right]$ with $\Xi = [\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_D]$. In the $(K-1)$-dimensional complement of $\mathbf{1}$, the rescaled columns $\sqrt{K(K+1)}\,\boldsymbol{\xi}_z$ are independent, isotropic and log-concave, being uniform on a convex set, so that as $D, K \to \infty$ with $D/(K-1) \to \alpha \in (0, 1)$ the eigenvalues of $\frac{K(K+1)}{K-1}\Xi^\top\Xi$ follow the Marchenko–Pastur law of ratio α, with extreme eigenvalues converging to the edges $(1 \pm \sqrt{\alpha})^2$ (Pajor & Pastur, 2009; Chafaï & Tikhomirov, 2018). The rank-one term removes at most one eigenvalue from this bulk, by interlacing, and produces the outlier D/K; dividing by D,

$$\lambda_1(M_2) = \frac{1}{K}\,(1 + o_{\mathbb{P}}(1)), \qquad \lambda_2(M_2),\ \lambda_D(M_2) = \frac{K-1}{DK(K+1)}\Big(1 \pm \sqrt{\tfrac{D}{K-1}}\Big)^2 (1 + o_{\mathbb{P}}(1)), \tag{46}$$

and the remaining $K - D$ eigenvalues vanish: one spike at $1/K$ and a bulk of $D - 1$ eigenvalues at scale $1/(DK)$, as in Fig. 6.

Spectrum of M_4. For $K < D \le K(K+1)/2$ the complete block is M_4, whose nonzero spectrum is that of $\frac{1}{D}[(\mathbf{p}_i^\top \mathbf{p}_j)^2]_{ij}$. Writing $\mathbf{p}^{\otimes 2} = \mathbf{u}^{\otimes 2} + (\mathbf{u} \otimes \boldsymbol{\xi} + \boldsymbol{\xi} \otimes \mathbf{u}) + \boldsymbol{\xi}^{\otimes 2}$ with $\mathbf{u} = \mathbf{1}/K$ separates three scales. The mean feature has squared norm $(K+3)/(K(K+1)^2) \simeq K^{-2}$ and gives one spike at K^{-2}; the linear features span $K - 1$ directions with covariance eigenvalue $2/(K^2(K+1))$; the centred quadratic features span the remaining $r_2 = K(K-1)/2$ directions, with average covariance eigenvalue $2K^{-4}(1 + O(K^{-1}))$. Neglecting the coupling between the blocks, a Marchenko–Pastur approximation of each predicts $K - 1$ eigenvalues of order K^{-3}, with edges $\frac{2}{K^2(K+1)}(1 \pm \sqrt{(K-1)/D})^2$, and $D - K$ eigenvalues of order $1/(DK^2)$, with edges $\frac{1+O(K^{-1})}{DK^2}(1 \pm \sqrt{D/r_2})^2$ when $D/r_2 \to \beta \in (0, 1)$. This block picture is heuristic, since the quadratic features are neither isotropic nor log-concave and the coupling is neglected; Fig. 7 compares it with the spectrum of one realisation at $D = K(K+1)/4$, where both bulks follow the predicted curves and their Frobenius norms agree with the predictions within 7% for the linear block and 2% for the quadratic one, for K from 64 to 192.

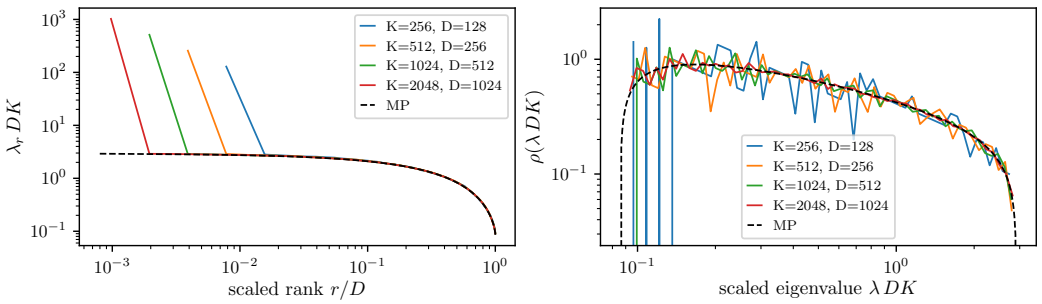


Figure 6: **Population spectrum of M_2 for $D \le K$. Left:** ranked eigenvalues $\lambda_r\, DK$ against the scaled rank r/D at $\alpha = D/K = 1/2$, for $K \in \{256, 512, 1024, 2048\}$ and one topic realisation per (K, D); dashed: the ranked Marchenko–Pastur law; the spike is the outlier at $\lambda_1 DK \approx D$. **Right:** density of the bulk eigenvalues, spike removed, on the $\lambda\, DK$ scale, against the Marchenko–Pastur density (dashed).

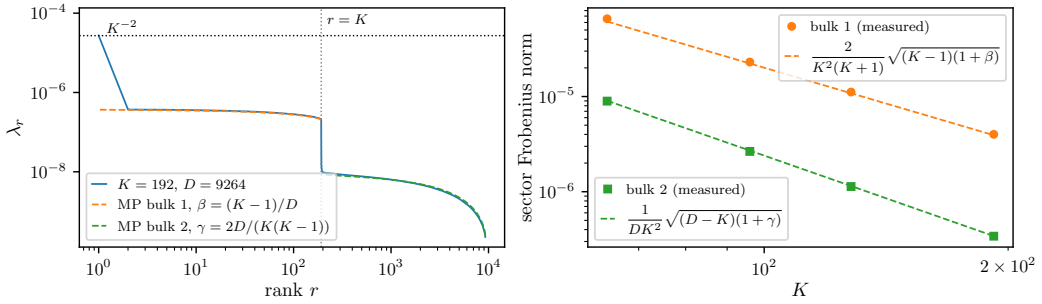


Figure 7: **Population spectrum of M_4 for $K < D \le K(K+1)/2$. Left:** ranked eigenvalues of M_4 for one realisation at $K = 192$, $D = K(K+1)/4 = 9264$: the spike at K^{-2} (dotted), the $K-1$ eigenvalues of the linear block and, beyond $r = K$ (dotted), the $D-K$ eigenvalues of the quadratic block, against the Marchenko–Pastur predictions of the two blocks (dashed). **Right:** Frobenius norm of each block against K, for $D = K(K+1)/4$, with the parameter-free predictions (dashed).

Noise and sample complexity. Let $\widehat{M}_2$ be the empirical pair moment from N independent pairs, so that $\widehat{M}_2 - M_2 = \frac{1}{N}\sum_\mu Y_\mu$ with $Y_\mu = \mathbf{x}_1^{(\mu)}\mathbf{x}_2^{(\mu)\top} - M_2$ centred, $\|Y_\mu\|_{\mathrm{op}} \le 1 + 1/K$ and $\mathbb{E}[Y_\mu Y_\mu^\top] \preceq \mathrm{diag}(\bar{\mathbf{p}})$, with $\bar{\mathbf{p}} = \frac{1}{D}\sum_z \mathbf{p}_z$ the unigram marginal, of entries close to $1/K$. The matrix Bernstein inequality (Tropp, 2012) then gives, with probability at least $1-\delta$,

$$\|\widehat{M}_2 - M_2\|_{\mathrm{op}} \le \sqrt{\frac{2\log(2K/\delta)}{KN}} + \frac{2\log(2K/\delta)}{3N}, \tag{47}$$

a noise level of order $(NK)^{-1/2}$; the measured error follows $1.5/\sqrt{NK}$ across eleven (K, D) configurations (Fig. 8, left). Treating the noise as a generic perturbation, an eigenvalue λ of M_2 is resolved once $\lambda \gtrsim 1.5/\sqrt{NK}$: the spike at $N \sim K$, and the lower edge of the bulk, $\frac{K-1}{DK(K+1)}(1-\sqrt{\alpha})^2$, at $N^* = 2.25\, D^2K/(1-\sqrt{\alpha})^4$, i.e. at $N \sim D^2K$ up to an α-dependent constant. This criterion is heuristic, but the scale is confirmed by the eigenvectors: the overlap between the bulk eigenvector subspaces of $\widehat{M}_2$ and M_2 rises from its random value $(D-1)/(K-1)$ to one around $N/(D^2K) = O(1)$ (Fig. 8, right). For the four-token moment the same bound has a variance proxy of order K^{-2} and gives a noise level of order $(NK^2)^{-1/2}$, so the spike at K^{-2} is resolved at $N \sim K^2$, the linear block at $N \sim K^4$ and the quadratic block, of scale $1/(DK^2)$, at $N \sim D^2K^2$; extrapolating, the bulk of M_{2m} sits at $1/(DK^m)$ against a noise of $(NK^m)^{-1/2}$, and its resolution takes $N \sim D^2K^m$.

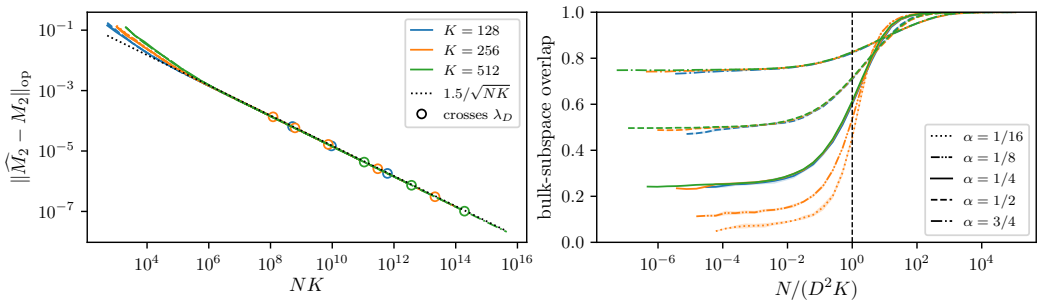


Figure 8: **Convergence of the empirical pair moment for $D \leq K$. Left:** operator-norm error $\|\widehat{M}_2 - M_2\|_{\mathrm{op}}$, geometric mean over eight replicas, against NK for eleven (K, D) configurations, $\alpha = D/K \in \{1/4, 1/2, 3/4\}$ times $K \in \{128, 256, 512\}$ plus $K = 256$ with $D \in \{16, 32, 64, 128, 192\}$; colour encodes K, line style α; all curves collapse on the dotted $1.5/\sqrt{NK}$; circles mark where each curve crosses its own population λ_D, beyond which the lower bulk edge is resolved. **Right:** overlap $\frac{1}{D-1}\|U_{2:D}^{\top}\widehat{U}_{2:D}\|_F^2$ between the bulk eigenvector subspaces of M_2 and $\widehat{M}_2$, mean $\pm$ one standard deviation over replicas, against $N/(D^2K)$; each curve rises from the random-subspace value $(D-1)/(K-1)$ towards one around $N/(D^2K) = O(1)$ (dashed line at one). One topic realisation per (K, D), uniform topic prior.

G Experimental details

This appendix specifies the transformer experiments of Sec. 5.

Data. Every run draws its own topic matrix P, with D topics from $\mathrm{Dir}_K(\mathbf{1})$ and uniform topic weights, and from it N training documents and 32 768 validation documents; a further 32 768 OOD documents are generated from 16 fresh topics, drawn independently of P from the same prior. Each point of a learning curve averages four runs with independent topic matrices, test ID and OOD corpora and initialisation, and error bars are standard errors over the four. Documents carry $T+1$ tokens: the model reads the first T and is trained to predict tokens 2 to $T+1$, so that its loss profile covers context lengths 1 to T, the T-gram loss is the loss at context length T, and the autoregressive loss is the average of the profile over the T positions.

Model. A one-block decoder-only transformer: token embeddings of dimension d_{emb} with no positional embedding, so that the causal mask is the only source of order, as befits documents that are exchangeable given their topic; one causal self-attention head; a ReLU multilayer perceptron with one hidden layer of width $4d_{\mathrm{emb}}$; both sub-layers pre-normalised and residual; a final layer normalisation and an untied linear readout to the K logits. Attention and perceptron carry no biases and there is no dropout. The model has $12\, d_{\mathrm{emb}}^2 + 2Kd_{\mathrm{emb}} + O(d_{\mathrm{emb}})$ parameters, from 5×10^4 at $(K, d_{\mathrm{emb}}) = (32, 64)$ to 3×10^6 at $(128, 512)$. Weights are initialised as standard normals and each linear map is scaled by the inverse square root of its input dimension, the readout by $1/d_{\mathrm{emb}}$.

Training. The loss is the cross-entropy of the next token, averaged over the T positions and the documents of a batch. Batches of b documents are taken cyclically from the corpus and the parameters are updated with Adam, moment parameters $(0.9, 0.999)$ and no weight decay, at a learning rate η raised linearly from zero over the first 8 steps and constant thereafter, for a fixed number of passes over the corpus.

Evaluation and early stopping. The validation autoregressive loss is measured at log-spaced steps throughout training. The checkpoint of lowest validation loss defines early stopping, and its loss is the value that enters the ID learning curves (Fig. 1, Fig. 3, Fig. 9). At a coarser log-spaced subset of the steps the full loss profiles on the validation and on the OOD documents are recorded as well. The T-gram ID loss of a run (Fig. 10) is the minimum over these recorded checkpoints of the loss at context length T; the OOD losses are read either at the recorded checkpoint of lowest

validation loss, "at early stopping" (Fig. 1, Fig. 4), or minimised over the recorded checkpoints, "best over training" (Fig. 1, Fig. 5), as stated in each caption.

Hyperparameter selection. The embedding dimension, batch size, learning rate and training budget are selected separately for every (K, D, T, N) on the validation loss at early stopping, so that each learning curve reports the best model of the family rather than a fixed architecture, and the model is not the bottleneck of the comparison with the Bayes-optimal learning curves. The grid is d_{emb} from K to 512 in powers of two, $b \in \{16, 32, 64, 128\}$ and $\eta \in \{10^{-2}, 3\times10^{-2}, 10^{-1}\}$; larger dictionaries require larger d_{emb}, and the selected width grows with D at fixed K. The number of passes is extended until the early-stopping checkpoint falls well inside the budget, or until a doubling of the budget improves the validation loss by less than the spread over seeds. A candidate combination replaces the incumbent only when its early-stopping validation loss, over four fresh seeds, is lower beyond the standard errors, or tied with a clearly earlier early-stopping step. The combination selected on the ID validation loss serves for every quantity reported at that (K, D, T, N), including the OOD losses, which play no role in the selection.

H ADDITIONAL FIGURES

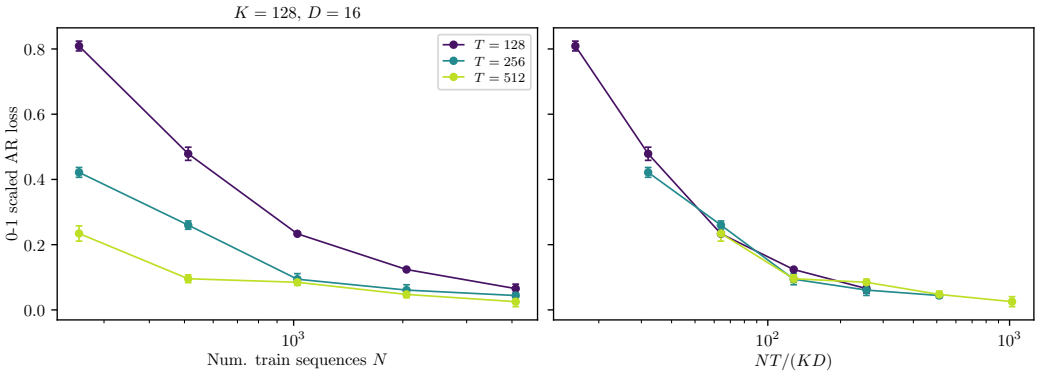


Figure 9: **Learning curves at different document lengths collapse on the number of tokens per parameter. Left:** Scaled ID autoregressive loss $(\mathcal{L}_{\mathrm{AR}}(N) - \mathcal{L}_{\mathrm{AR}}(\infty))/(\log K - \mathcal{L}_{\mathrm{AR}}(\infty))$ versus number of training sequences for one-block transformers trained on random topic model data with $K = 128$, $D = 16$ and $T = 128, 256, 512$ (colour). $\mathcal{L}_{\mathrm{AR}}(\infty)$ is the autoregressive loss of the specialised predictor at each T, averaged over 32 realisations of the topic matrix. **Right:** Against $NT/(DK)$ the three curves coincide: at fixed N, longer documents bring the loss closer to its floor in proportion to the number of tokens they contribute.

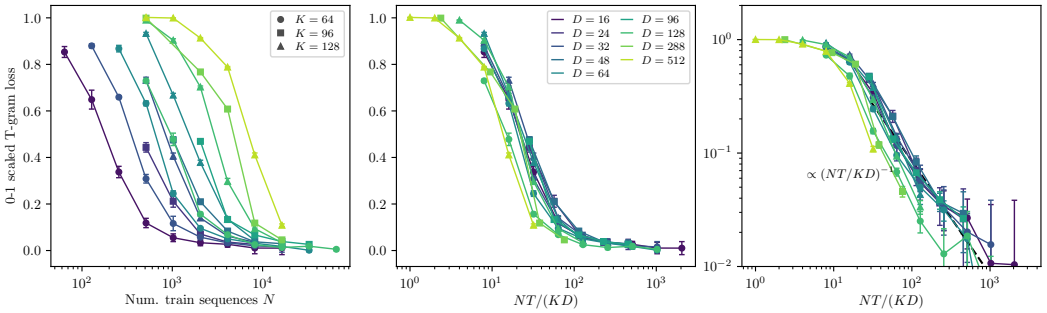


Figure 10: **Transformers match the Bayes-optimal specialisation learning curve at the longest context too.** T-gram counterpart of Fig. 3, for the same twelve (K, D) at $T = 128$. **Left:** Scaled ID T-gram loss $(\mathcal{L}_T(N) - \mathcal{L}_T(\infty))/(\log K - \mathcal{L}_T(\infty))$ versus number of training sequences, where $\mathcal{L}_T(\infty)$ is the mean entropy $\frac{1}{D}\sum_z H(\mathbf{p}_z)$ of the topics of each run, the large-context floor of the specialised predictor. **Middle:** Rescaling the x-axis as the number of tokens per parameter of the data-generating process collapses all the curves onto a single functional form, more tightly than for the autoregressive loss. **Right:** The decay towards 0 (black dashed) follows the x^{-1} rate of the Bayes-optimal learning curve.

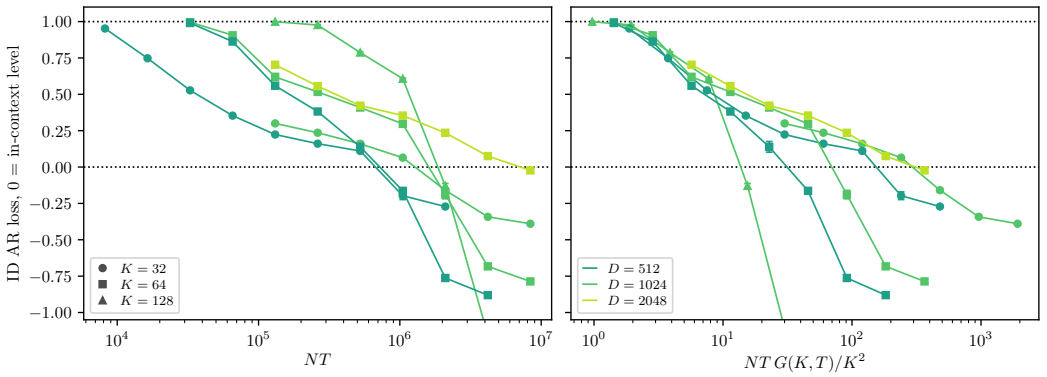


Figure 11: **The ID loss descends to the in-context level at the predicted sample scale. Left:** ID autoregressive loss at early stopping, scaled as $(\mathcal{L}_{\mathrm{AR}}(N) - \mathcal{L}^{\star}_{\mathrm{AR}}[\mathbf{r}])/(\log K - \mathcal{L}^{\star}_{\mathrm{AR}}[\mathbf{r}])$, where $\mathcal{L}^{\star}_{\mathrm{AR}}[\mathbf{r}] = \log K - G(K,T)$ is the autoregressive loss of the in-context predictor, the same on ID and OOD documents (App. C). This scaled loss is 1 at the uniform prediction, 0 at the in-context solution, and negative at specialisation. Versus number of training tokens NT, for the (K, D) of Fig. 5. **Right:** When plotted against $NT\,G(K,T)/K^2 = N/(2N_{\mathrm{icl}})$ (Eq. 13) the descents to the in-context level coincide for all K and D, with no adjustable constant. The subsequent departure below zero is specialisation.