ATTENTION HEAD SPECIALIZATION: PHASE TRANSITIONS IN SAMPLE COMPLEXITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention heads specialize into distinct functions during training, often emerging in stages together with sharp changes in model behavior. While this phenomenon has mostly been studied as a function of training time, we ask how head specialization depends on the amount of available training data. Empirically, we find that heads specialized to syntactic relations in BERT emerge as the dataset size increases. We then study this phenomenon in a solvable high-dimensional model of multi-head attention and establish that, in this model, head specialization is a genuine phase transition in sample complexity, corresponding to a non-analyticity in the test error and other observables. Depending on data structure and regularization, specialization can follow uninformative head differentiation or occur through a cascade of phase transitions that progressively recover weaker latent features, providing a statistical form of simplicity bias. Finally, we relate these statistical transitions to the stagewise emergence of specialized heads during online learning, disentangling statistical from dynamical origins of head specialization.

Multi-head attention is a central architectural ingredient of modern transformer models, enabling multiple attention patterns to coexist within a single layer. Different heads learn to perform distinct functions, while their collaboration allows implementing circuits that no individual head captures alone. Empirical studies have identified such specialized heads across multiple settings, including induction heads involved in in-context learning (ICL) (Olsson et al., 2022; Yin & Steinhardt, 2025), heads specializing to different syntactic relations (Chen et al., 2024a), and heads progressively differentiating into distinct functional roles during training (Wang et al., 2025). Head specialization has predominantly been studied as a function of training time: specialized mechanisms emerge at distinct stages of training, sometimes coinciding with sharp improvements in performance or drops in training loss (Hoogland et al., 2025; Aoyama et al., 2026).

Much less is known about specialization as a function of the amount of available training data. Rather than asking *when* during training different heads specialize, one can ask *how many samples* are needed for distinct specialized heads to emerge. This distinction separates a statistical question about the information available in the dataset from the particular optimization dynamics used to learn it. We investigate this question in BERT masked language models (MLMs), where syntactic attention structure (SAS), characterized by heads specializing to different syntactic relations (Clark et al., 2019), emerges during training alongside a sharp drop in loss (Chen et al., 2024a). Empirically, training $\mathrm{BERT}_{\mathrm{Small}}$ models from random initialization on datasets of N sequences from BookCorpus and English Wikipedia reveals a clear phenomenology shown in Fig. 1: as N increases, SAS emerges sharply over a characteristic range of dataset sizes, accompanied by abrupt changes in the learning curves. For each dataset size, we fix the sequences and masking patterns throughout training and report metrics at the checkpoint with lowest validation MLM loss. Following (Chen et al., 2024a), we quantify SAS on a separate syntactically annotated corpus as the fraction of parent-child pairs recovered by each head's maximum word-level attention. Aggregating accuracy across relations for the best head gives the unlabeled attachment score (UAS), which measures alignment between attention and syntactic structure enabled by specialization of the heads.

Sample complexity provides a natural control parameter for asking whether such emergence corresponds to a genuine phase transition. In the statistical physics of learning, one considers a proportional high-dimensional limit in which the number of trainable degrees of freedom and training samples grow together, $D, N \to \infty$ at fixed $\alpha = N/D$. This framework has a long history in neural-network theory, where increasing sample complexity can induce phase transitions from per-

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.

mutation symmetric to specialized solutions in the committee machine (Schwarze & Hertz, 1992; 1993; Aubin et al., 2018). In this limit, changes between qualitatively different learning regimes can become sharp, with nonanalyticities of macroscopic observables at critical values of α. For multi-head attention, this raises a precise question: can the emergence of specialized heads be understood as a genuine phase transition controlled by sample complexity?

We answer this question in a tractable multi-feature variant of the single-location regression (SLR) model (Marion et al., 2025; Sagitova et al., 2026), whose head-specialization dynamics under online learning were characterized in Sagitova et al. (2026). The task is to retrieve one token carrying a low-dimensional latent signal from a sequence of otherwise uninformative tokens, using multiple attention heads that can specialize to different latent features. Here, instead of following training dynamics, we study the associated empirical risk minimization (ERM) problem as a function of sample complexity. Fig. 1 shows that this model exhibits behavior qualitatively analogous to that observed in the $\mathrm{BERT}_{\mathrm{Small}}$ experiment. Moreover, we show that, in this model, head specialization occurs through phase transitions in sample complexity, characterized by non-analyticities in the order parameters and prediction error at critical values of α.

The resulting phase diagram reveals a richer structure than a single specialization threshold. At low regularization and low sample complexity, specialization is preceded by an unspecialized phase with broken head-permutation symmetry, in which heads differ by fitting sample-specific fluctuations without recovering distinct latent feature. This illustrates that distinct attention patterns alone do not imply specialization that improves generalization. For heterogeneous signals, increasing sample complexity instead produces a cascade of specialization transitions: heads successively recover latent features from strongest to weakest, so that increasingly subtle structure requires increasingly more data. This provides a statistical counterpart of stagewise learning and simplicity bias, arising here from the structure of the ERM solutions rather than from a particular training dynamics.

Finally, we connect these equilibrium transitions back to the training-time phenomenology from which we started (Olsson et al., 2022; Wang et al., 2025; Aoyama et al., 2026; Chen et al., 2024a). In online learning, training time and the amount of observed data increase together, entangling the statistical transition with optimization dynamics. In the solvable model, we show how the critical sample complexities map onto the emergence of specialized heads along training. This perspective suggests that some of the abrupt or stagewise emergence observed during transformer training may reflect underlying statistical phase transitions, while distinguishing these from phenomena that are genuinely dynamical and helping clarify how such transitions manifest in learning dynamics.

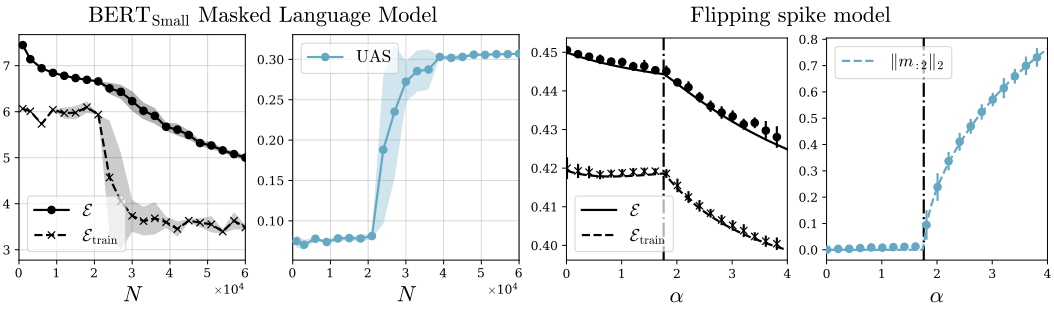

Figure 1: **Head specialization as a function of sample complexity. Left:** $\mathrm{BERT}_{\mathrm{Small}}$ models trained independently on fixed datasets of N sequences through minimization of masked-language-modeling (MLM) loss. **(first panel)** Validation error and training MLM loss (104). **(second panel)** Unlabeled attachment score (UAS)(107), measuring the fraction of dependency parents correctly recovered from word-level attention by the best head, layer, and attention direction. Curves show averages over five runs, with shaded bands indicating seed variability. Reported metrics are evaluated at early-stopped checkpoints selected by the lowest validation loss along each training trajectory. Experimental details are provided in App. E.2. **Right:** Qualitative comparison with the solvable flipping spike model (Sec. 1) as the sample complexity $\alpha = N/D$ increases, with N the number of sequences and D the token dimension. Solid lines correspond to theoretical predictions and markers to numerical experiments with $D = 2000$, $\nu_1 = 0.35$, $\nu_2 = 2.0$, $L = 10$, $H = 2$, $\lambda = 0.2$. **(third panel)** Test and training errors exhibit non-analyticity at the specialization threshold. **(fourth panel)** Head–feature alignment reveals specialization to the latent feature.

Main Contributions The main empirical result of the paper is the characterization of head specialization in the number of training samples. In a tractable model of multi-head attention, we then show that head specialization is a statistical phase transition in sample complexity, and connect the equilibrium transition to specialization during training. Our main technical contributions are:

- We derive a tight asymptotic characterization of the regularized ERM problem in the proportional high-dimensional limit $N, D \to \infty$ at fixed $\alpha = N/D$, through closed self-consistent equations for the errors and order parameters.
- We show that head specialization is a genuine phase transition in sample complexity. Depending on regularization and signal structure, we uncover an unspecialized phase with broken head-permutation symmetry, as well as a cascade of specialization transitions in which increasingly weaker latent features are recovered with increasing sample complexity, yielding a statistical form of simplicity bias.
- We establish a controlled correspondence between these equilibrium phase transitions and the emergence of specialized heads during online learning, relating critical sample complexities to training-time specialization and disentangling statistical from optimization-dependent effects.

Notations. We write $[n] = \{1, \dots, n\}$, δ_{ij} for the Kronecker delta, and $\mathbb{I}_n$ for the identity matrix. For a matrix A, A_i and $A_{:j}$ denote its i-th row and j-th column. We write $\|\cdot\|_2$ for the Euclidean norm, $\|\cdot\|_F$ for the Frobenius norm, and $v^{\otimes 2} = vv^\top$. We denote by $\mathcal{N}(\mu, \Sigma)$ the Gaussian distribution with mean μ and covariance Σ.

RELATED WORKS

Empirical emergence and head specialization. Specialized attention heads have been observed across architectures and tasks, including induction heads associated with in-context learning (Olsson et al., 2022; Yin & Steinhardt, 2025), syntactic heads in masked language models (Chen et al., 2024a), and circuits studied in controlled synthetic settings (Singh et al., 2024). A complementary literature studies when such capabilities and internal mechanisms emerge during training, including abrupt changes in performance (Wei et al., 2022), stagewise development and head differentiation (Hoogland et al., 2025; Wang et al., 2025), and the emergence time of induction heads (Aoyama et al., 2026). Such behavior can depend strongly on data, initialization, and training trajectory: emergent capabilities can be transient (Singh et al., 2023), vary across random seeds (Juneja et al., 2023; Zhao et al., 2025), and depend on the structure of the training data (Qin et al., 2025). While these works predominantly follow a model through training, our BERT experiments instead hold dataset size fixed for each training run, and study how syntactic head specialization changes across datasets of different sizes.

Theory of attention and head specialization. Recent solvable models characterize learning in attention architectures, including single-location regression (Marion et al., 2025; Duranthon et al., 2026), sequence multi-index models (Troiani et al., 2025), or attention-indexed models (Boncoraglio et al., 2026). Multi-head specialization has primarily been studied through training dynamics (Chen et al., 2024b; Singh et al., 2023; Zhang et al., 2025; Sagitova et al., 2026; Yüksel et al., 2026). Here we instead study specialization at equilibrium as a function of sample complexity.

Phase transitions and specialization in neural networks. Specialization as a sample-complexity phase transition has a long history in committee machines, where increasing $\alpha = N/D$ can induce transitions to solutions in which different hidden units align with distinct features of the data (Schwarze & Hertz, 1992; 1993; Aubin et al., 2018; Pérez Ortiz et al., 2026). Phase transitions have also been studied in solvable attention models (Cui et al., 2024); most closely, Lavie et al. (2026) establish a sample-complexity transition for the emergence of a single copy head. We instead study specialization of attention heads and the associated emerging generalization capabilities.

1 SINGLE-LOCATION REGRESSION MODEL

We consider the multi-feature spiked single-location regression (SLR) model of Sagitova et al. (2026).The task is to retrieve one randomly located relevant token from a sequence of otherwise uninformative tokens. The informative token is structured by a low rank planted signal over F latent directions. Learning this structure with multiple low-rank heads provides a tractable setting for studying head specialization.

Data model We consider an embedding dimension D. We first draw F independent hidden spikes $k_f^* \sim \mathcal{N}(0, \mathbb{I}_D)$ defining the latent structure that remains fixed across all sequences. We denote $K^* = \{k_f^*\}_{f=1}^F$ the matrix with rows $k_f^{*\top}$. A sequence $X \in \mathbb{R}^{L\times D}$ consists of L tokens $\{X_\ell\}_{1\le\ell\le L}$. A hidden index $\epsilon \in [L]$ identifies the relevant token X_ϵ containing a planted spike $\hat{k}$ generated from the F latent directions, while remaining tokens $X_{\ell\neq\epsilon}$ are independent isotropic Gaussian noise. For each sequence, we independently draw $\epsilon \sim \text{Unif}([L])$ and the feature weights $\theta \sim P_\theta$ with $\theta \in \mathbb{R}^F$, defining the sequence-dependent spike $\hat{k}$, such that the tokens X_ℓ are independently distributed conditionally on ϵ and θ as

$$\hat{k} = \frac{K^{*\top}\theta}{\sqrt{D}} \in \mathbb{R}^D, \qquad X_\ell \sim \mathcal{N}(\delta_{\ell,\epsilon}\hat{k}, \mathbb{I}_D), \quad \ell \in [L]. \tag{1}$$

The associated label is the relevant token $y = X_\epsilon$. As the latent ϵ and θ are sampled at every sequence/context, deducing them can also be seen as a simple case of in-context learning. In the following, we will consider two choices for the distribution P_θ. In the **flipping spike model**, a latent direction is kept constant while another flips sign randomly across sequences: $F = 2$, $\theta_1 = \sqrt{\nu_1}$, and $\theta_2 \sim \text{Unif}(\{-\sqrt{\nu_2}, \sqrt{\nu_2}\})$, where we define signal strengths $\nu_1 > 0$, $\nu_2 > 0$. In the **non-isotropic Gaussian model**, all components of the planted spike fluctuate independently with different covariances: $\theta_f \sim \mathcal{N}(0, \nu_f)$, where we define signal strengths $\nu_1 \ge ... \ge \nu_F \ge 0$.

Multi-head attention We consider learning of the stated task with a single layer of low-rank attention with H heads. This architecture is well suited for retrieval of low rank structure, with different heads attending to different features. Specifically, we consider the estimator $f_K : \mathbb{R}^{L\times D} \to \mathbb{R}^D$ given by rank-one attention heads with trainable weights $K = \{k_h\}_{h=1}^H \in \mathbb{R}^{H\times D}$:

$$f_K(X) = \frac{1}{H}\sum_{h=1}^{H} \text{softmax}\left(\frac{Xk_h}{\sqrt{D}}\right)^\top X, \tag{2}$$

where attention vectors are uniformly averaged across heads. This model can be viewed as a simplification of cross-attention with fixed sequence-independent query embedding and identity value matrix, as discussed in Marion et al. (2025) and Duranthon et al. (2026). Note that, while considering larger (extensive) rank attention is technically possible (Boncoraglio et al., 2026), this is left for future work. Similarly, training the value matrix with a fixed number of parameters could be accommodated by an extended parametrization. The current model then constitutes a relevant minimal estimator for the study of head specialization in the designed task. As mentioned in Sagitova et al. (2026), this type of attention layer is also used in practice (Carion et al., 2020).

Empirical risk minimization We study learning from a finite dataset $\mathcal{D} = \{X_\mu, y_\mu\}_{\mu=1}^N$ of N independently sampled sequences, and their associated labels. An attention layer $f_{\hat{K}}$ is learned by regularized empirical risk minimization (ERM):

$$\hat{K} = \underset{K\in\mathbb{R}^{H\times D}}{\text{argmin}}\, \mathcal{L}_{\text{train}}(K), \qquad \mathcal{L}_{\text{train}}(K) = \sum_{\mu=1}^{N} \frac{1}{2D}\|y_\mu - f_K(X_\mu)\|_2^2 + \frac{\lambda}{2}\|K\|_F^2 \tag{3}$$

where $\lambda > 0$ is the ℓ_2 regularization strength. The performance of the estimator (3) is evaluated through the training error $\mathcal{E}_{\text{train}}$ associated with the objective (3), and the test error $\mathcal{E}$:

$$\mathcal{E}_{\text{train}} = \frac{1}{N}\sum_{\mu=1}^{N} \frac{1}{2D}\|y_\mu - f_K(X_\mu)\|_2^2, \qquad \mathcal{E} = \mathbb{E}_{X,y}\left[\frac{1}{2D}\|y - f_{\hat{K}}(X)\|_2^2\right]. \tag{4}$$

High-dimensional limit We analyze the ERM problem (3) in the proportional limit where $D, N \to \infty$ while the ratio $\alpha = N/D$, the sample complexity, remains of order $\Theta(1)$. In order to study head specialization, we further assume the number of heads H, the number of features F, the sequence length L, and the regularization strength λ to be $\Theta(1)$. This proportional limit has been considered in a stream of previous works (e.g. Cui et al. (2024); Troiani et al. (2025); Cui & Zdeborová (2024); Duranthon et al. (2026)), and allows deriving a closed-form characterization of the ERM problem which we present in the next section. As the number of trainable degrees of freedom grows linearly with D, taking $N = \Theta(D)$ retains nontrivial interactions between the dataset

and the architecture, and sample complexity α appears as a natural control parameter to study head-specialization phase transitions. This particular scaling is, however, setting-specific: other well-defined control parameters may arise in different high-dimensional limits (e.g. for extensive-rank architectures Gerace et al. (2020); Erba et al. (2025); Boncoraglio et al. (2026)), or be defined through properties of the data.

2 Asymptotic characterization of the empirical risk

The main technical result of this work is a closed-form characterization of the minimizers of the regularized empirical risk (3) obtained in the high-dimensional limit for the model (2) through the test error (4), training error (4) and additional summary statistics as functions of sample complexity. Defined from the high-dimensional weights, these summary statistics concentrate in the high-dimensional limit and fully describe the model's macroscopic state: the head–feature overlaps m_{hf} quantify the alignment between each head's learned attention vector k_h and each planted feature direction k_f^*, while the self-overlaps $q_{hh'}$ quantify the alignment between learned attention vectors k_h and $k_{h'}$, with diagonal entries encoding their normalized squared norms.

Result 2.1. *Under the assumption of replica symmetry, in the limit $N, D \to \infty$, with $H, F, L, \|\theta\|_2, \lambda, N/D = \Theta_D(1)$, the summary statistics*

$$m = \frac{KK^{*\top}}{D} \in \mathbb{R}^{H\times F}, \qquad q = \frac{KK^{\top}}{D} \in \mathbb{R}^{H\times H} \tag{5}$$

concentrate in probability, and are solutions of the set of finite-dimensional self-consistent equations

$$\begin{cases} m = (\lambda \mathbb{I}_H + \hat{V})^{-1}\hat{m} \\ q = (\lambda \mathbb{I}_H + \hat{V})^{-1}(\hat{m}\hat{m}^{\top} + \hat{q})(\lambda \mathbb{I}_H + \hat{V})^{-1} \\ V = (\lambda \mathbb{I}_H + \hat{V})^{-1} \end{cases} \tag{6}$$

$$\begin{cases} \hat{m} = \alpha \mathbb{E}_{\theta,\Xi}\left[V^{-1}\left(\mathrm{prox}(\theta,\Xi)_1 - \omega_1\right)\theta^{\top}\right] \\ \hat{q} = \alpha \mathbb{E}_{\theta,\Xi}\left[V^{-1}\left(\sum_{\ell=1}^{L}\left(\mathrm{prox}(\theta,\Xi)_\ell - \omega_\ell\right)^{\otimes 2}\right)V^{-1}\right] \\ \hat{V} = -\alpha \mathbb{E}_{\theta,\Xi}\left[V^{-1}\left(\sum_{\ell=1}^{L}\left(\mathrm{prox}(\theta,\Xi)_\ell - \omega_\ell\right)\xi_\ell^{\top}\right)q^{-1/2}\right], \end{cases} \tag{7}$$

with the summary statistics $\{m, q, V\}$ and their conjugates $\{\hat{m}, \hat{q}, \hat{V}\}$. In (6-7), we define the Gaussian random variables $\Xi = \{\xi_\ell\}_{\ell=1}^{L}$ with $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_H)$ and the associated local fields $\Omega = \{\omega_\ell\}_{\ell=1}^{L}$ with $\omega_\ell := \delta_{\ell,1} m\theta + q^{1/2}\xi_\ell$. The proximals $\{\mathrm{prox}(\theta,\Xi)_\ell\}_{\ell=1}^{L}$ are defined as minimizers of the Moreau envelope

$$\mathcal{M}(\theta,\Xi) = \inf_{Z\in\mathbb{R}^{L\times H}}\left\{\frac{1}{2}\sum_{\ell=1}^{L}\left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H}\mathrm{softmax}(Z_{:h})_\ell\right)^2 + \frac{1}{2}\sum_{\ell=1}^{L}\left(Z_\ell - \omega_\ell\right)^{\top} V^{-1}\left(Z_\ell - \omega_\ell\right)\right\}. \tag{8}$$

In the same limit, the test error (4) converges in probability to

$$\mathcal{E} = \frac{1}{2}\mathbb{E}_{\theta,\Xi}\left[\sum_{\ell=1}^{L}\left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H}\mathrm{softmax}(\Omega_{:h})_\ell\right)^2\right]. \tag{9}$$

Finally, the training error at the ERM minimizer converges in probability to

$$\mathcal{E}_{\mathrm{train}} = \mathbb{E}_{\theta,\Xi}\left[\mathcal{M}\left(\theta,\Xi\right)\right] - \frac{1}{2\alpha}\mathrm{Tr}\left(\hat{q}V\right). \tag{10}$$

Result 2.1 provides a tight asymptotic characterization of the test error $\mathcal{E}$ (9) and the training error $\mathcal{E}_{\mathrm{train}}$ (10) as a function of finite-dimensional summary statistics $\{m, q, V\}$ characterized in closed form by the set of equations (6-7), providing a finite-dimensional description of the high dimensional learning problem (3). Solutions of the self-consistent equations (6-7) asymptotically describes the global minimum of (3) as well as saddles and local minima. Multiple solutions can coexist and describe competing minima of the training loss. Thus, changing the control parameters such as sample complexity can change the stability of minima or selection of the global minimum, and produce phase transitions.

The derivation of Result 2.1, detailed in App. A.1, relies on the replica method under a replica-symmetric (RS) ansatz. While replica-symmetric predictions have been rigorously established for convex ERM problems (see, e.g., Aubin et al. (2018)) and under specific conditions for non-convex ones (Vilucchio et al., 2025), their validity in our non-convex setting remains an assumption. The resulting characterization is supported by numerical experiments: in the regimes studied, training the attention layer (2) on the ERM objective (3) by gradient-based optimization from random initializations (App. E) yields solutions whose summary statistics, training error and test error agree with those of the predicted global minima. Finally, we mention that while Result 2.1 is presented specifically with softmax activation and ℓ_2 regularization of the training loss for clarity, similar results can be reached for generic convex regularizer and activation function.

3 Head-specialization phase transition

Phases and phase transitions We characterize the phase transitions that emerge in the high-dimensional limit, through the non-analyticities of the order parameters and the test error in the sample complexity. The empirical risk (3) is invariant under permutation of the heads, while the self-consistent equations (6-7) additionally inherit the symmetries of the data distribution. We call a phase specialized when the heads have distinct overlaps with the planted features, and unspecialized otherwise. At finite sample complexity, however, head weights can grow and correlate while being orthogonal to the planted features, allowing for the existence of an unspecialized phase breaking head-permutation symmetry in the self-overlap q.

For the flipping-spike model with $H = 2$, $F = 2$, the data distribution is invariant under sign reversal of the second feature. We distinguish the permutation symmetric (PS), permutation-symmetry-broken (PSB), and specialized (SPEC) phases using the head-feature overlap m and antisymmetric self-overlap $q_\perp \equiv (q_{11} + q_{22} - 2q_{12})/2$, which measures the squared distance between the attention-head vectors. The three phases satisfy

$$\text{PS}: \quad m_{11} = m_{21} > 0, \quad m_{12} = m_{22} = 0, \quad q_\perp = 0, \tag{11}$$

$$\text{PSB}: \quad m_{11} = m_{21} > 0, \quad m_{12} = m_{22} = 0, \quad q_\perp > 0, \tag{12}$$

$$\text{SPEC}: \quad m_{11} = m_{21} > 0, \quad m_{12} = -m_{22} \neq 0, \quad q_\perp > 0. \tag{13}$$

Thus, PSB distinguishes the heads without aligning their difference with the flipping feature, whereas SPEC develops opposite nonzero overlaps with this feature.

For the non-isotropic Gaussian model with distinct feature strengths ordered in decreasing order, the solutions studied below exhibit successive specialized phases (r-SPEC) indexed by $r \in \{0, \dots, \min(F, \lfloor H/2 \rfloor)\}$ and characterized by

$$r\text{-SPEC}: \quad \begin{cases} \|m_{:f}\|_2 > 0, & 1 \leq f \leq r, \\ \|m_{:f}\|_2 = 0, & r < f \leq F. \end{cases} \tag{14}$$

Here, $r = 0$ corresponds to the unspecialized permutation symmetric solution. PSB phases can also occur at low regularization; for simplicity, we restrict their analysis to the flipping-spike model.

Head-specialization phase transition in the flipping spike model We first turn to the flipping spike model defined in Sec. 1, where a constant latent direction is learnable without head specialization while a flipping direction requires specialized heads to be used in the retrieval task. At sufficiently large regularization, $\lambda \geq \lambda_{\text{tric}}$, we find two solutions:

- PS solution for $\alpha < \alpha_c^{\text{spec}}(\lambda)$: both heads implement the same attention pattern, and generalization improves with sample complexity through recovery of the constant direction ($\|m_{:1}\|_2 > 0$), while the flipping feature remains unrecovered ($\|m_{:2}\|_2 = 0$).
- SPEC solution for $\alpha > \alpha_c^{\text{spec}}(\lambda)$: the heads implement distinct attention patterns, each adapted to one sign of the flipping feature ($\|m_{:2}\|_2 > 0$). Recovery of this additional direction improves retrieval, with a gain in error depending on the relative signal strengths.

The multi-head attention model displays a phase transition in sample complexity from the PS solution to the SPEC solution at critical sample complexity α_c^{spec}. The transition is of second order, marked by non-analyticities in the overlaps and errors.

In Fig. 2, we compare different metrics for the lowest training loss solutions of the self-consistent equations as a function of sample complexity with the corresponding experimental results. Gradient

based optimization on the ERM objective from random initializations reaches the predicted global minima, with errors and overlaps matching theoretical predictions. The transition is visible through the continuous splitting of the head-feature overlaps, the associated non-analyticity in the errors, and the onset of flipping-feature recovery at α_c^{spec}.

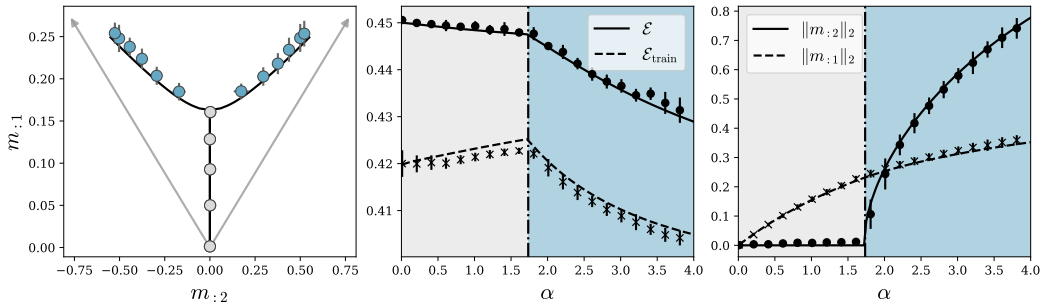

Figure 2: **Head-specialization phase transition in the flipping spike model:** $\nu_1 = 0.1$, $\nu_2 = 2.0$, $L = 10$, $H = 2$, $\lambda = 0.2$, $D = 2000$. Here and in Fig. 3 and 4, black solid lines correspond to theoretical predictions (Res. 2.1). Markers with error bars correspond to training the model (Sec. 1) with L-BFGS from random initialization in dimension D across ten seeds. **Left:** Head-feature overlaps $(m_{:2}, m_{:1})$ parametrized by sample complexity. Grey arrows indicate the directions of the two possible draws of the planted spike $\hat{k}$. **Center:** Test error $\mathcal{E}$ and training error $\mathcal{E}_{\text{train}}$. **Right:** Norm of head-feature overlaps $\|m_{:1}\|_2$, $\|m_{:2}\|_2$. Grey and blue identify the permutation symmetric (PS) and specialized (SPEC) phases. Vertical lines mark the specialization threshold α_c^{spec}.

Low regularization strengths can favor fitting sample fluctuations without feature recovery. In the regime $\lambda < \lambda_{\text{tric}}$, we find an additional solution to the asymptotic characterization:

- PSB solution for $\alpha_c^{\text{psb}}(\lambda) < \alpha < \alpha_c^{\text{spec}}(\lambda)$: The two heads differs in components orthogonal to the planted directions. They have identical feature overlaps, and therefore do not recover the flipping spike.

When this solution becomes global, specialization in sample complexity is delayed, producing a re-entrant specialization phase boundary at low regularization. Figure 3 shows the resulting phase diagram and a low-regularization cut, where the PSB solution is visible in the head self-overlaps before the onset of flipping-feature recovery. A similar type of solution, noted "memorization", has been observed in the ERM of committee-machines (Pérez Ortiz et al., 2026). Here we stress that, at low regularization, the validity of the RS ansatz requires particular scrutiny. In Fig. 3 the validity of RS assumption in this regime is supported trough agreement with simulations. In general, the possible need for a replica-symmetry-breaking (RSB) needs to be investigated.

Head-specialization cascade and simplicity bias in the non-isotropic Gaussian model We now consider the non-isotropic Gaussian model, with zero-mean feature weights of strengths $\{\nu_f\}_{f=1}^F$. For $F = 2$, we parameterize the latent structure by its total signal ν_{tot} and anisotropy δ:

$$\nu_{\text{tot}} = \nu_1 + \nu_2, \qquad \delta = \frac{\nu_1 - \nu_2}{\nu_1 + \nu_2}, \tag{15}$$

where $\delta = 0$ corresponds to isotropic strengths and increasing δ favors one latent direction. We take $H = 2F$, allowing one pair of heads per feature. At sufficiently large regularization we find:

- PS solution for $\alpha < \alpha_c^{(1)}$: no planted feature is recovered. In the absence of an easy direction, specialization is necessary to acquire the retrieval capability. The test error initially increases as the model fits uninformative sample fluctuations.
- r-SPEC solution for $\alpha_c^{(r)} < \alpha < \alpha_c^{(r+1)}$: the heads recover the r strongest latent features, from largest to smallest variance. Each additional specialization transition therefore recovers a weaker feature and incrementally improves retrieval.

Increasing sample complexity therefore produces a cascade of second-order specialization phase transitions at critical sample complexities $\{\alpha_c^{(r)}\}_{r=1}^F$, marked by the recovery of successive latent

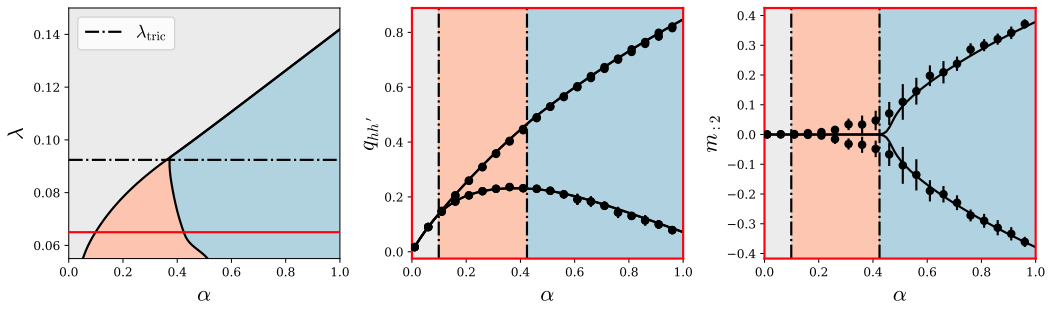

Figure 3: **Head differentiation without specialization in the flipping spike model:** $\nu_1 = 0.5, \nu_2 = 2.0$, $L = 10$, $H = 2$, $D = 6000$. **Left:** Phase diagram in sample complexity α and regularization λ. Phase boundaries are obtained by stability analysis of the different solutions (App. C). The red horizontal line at $\lambda = 0.065$ selects the cut shown in the other panels. **Center:** Head self-overlaps $q_{11} = q_{22}$ and mutual overlaps $q_{12} = q_{21}$. **Right:** Head-feature overlaps m_{12} and m_{22}. Grey, orange and blue identify the permutation symmetric (PS), permutation symmetry broken (PSB) and specialized (SPEC) phases, respectively. Vertical lines mark their boundaries α_c^{psb} and α_c^{spec}.

features and associated non-analyticities in the errors. With stronger features recovered at lower sample complexity, this hierarchy constitutes a statistical form of simplicity bias in the global ERM solutions, independent of the training dynamics. In Fig. 4, the (α, δ) phase diagram illustrates this simplicity bias. At fixed anisotropy δ, we compare the sample-complexity evolution of the errors and head-feature overlaps predicted by Result 2.1 with numerical experiments. Gradient-based optimization from random initialization reaches the predicted global minima, with errors and overlaps matching our theoretical predictions.

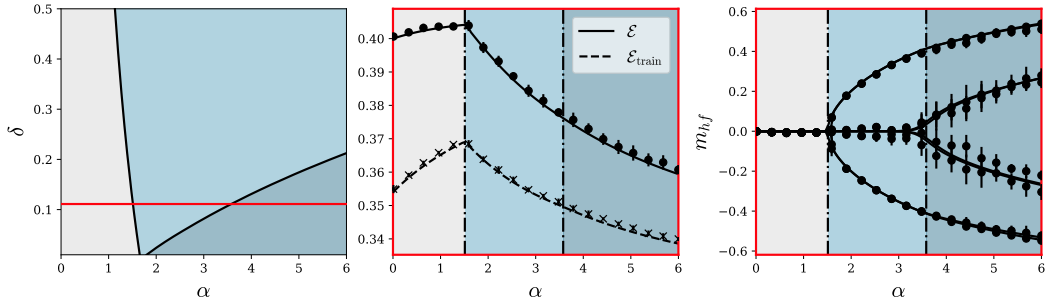

Figure 4: **Head-specialization cascade in the non-isotropic Gaussian model:** $F = 2$, $\nu_{\text{tot}} = 9.0$, $L = 5$, $H = 4$, $\lambda = 0.2$, $D = 4000$. **Left:** Phase diagram in sample complexity α and anisotropy δ. Phase boundaries are obtained by stability analysis of the different solutions (App. C). The red horizontal line at $\delta = 1/9$ corresponding to $(\nu_1, \nu_2) = (5.0, 4.0)$ selects the cut shown in the other panels. **Center:** Test error $\mathcal{E}$ and training error $\mathcal{E}_{\text{train}}$. **Right:** Head-feature overlaps m_{hf} for all heads and both latent features. Grey, light blue and darker blue identify the permutation symmetric (PS) and successively specialized (1-SPEC, 2-SPEC) phases, respectively. Vertical dash-dotted lines mark the specialization thresholds $\alpha_c^{(1)}$ and $\alpha_c^{(2)}$ along the cut.

4 Link between statistical transitions and dynamical emergence

We now connect the phase transitions in sample complexity to specialization during online learning. We consider the non-isotropic Gaussian model where $(m, q) = (0, 0)$ is a saddle point of the loss, and focus on the first specialization, assumed to occur only along k_1^*. In the right limits, we show that the sample complexity and the time for specialization are governed by the largest curvature $C \in \mathbb{R}^+$ of minus the population loss $\mathcal{E}_{\text{pop}}$ along this feature, defined as:

$$C = D^{-2}\lambda_{\max}\left[-k_1^{*\top}(\nabla^2_{K,K}\mathcal{E}_{\text{pop}})k_1^*\right], \qquad \mathcal{E}_{\text{pop}}(K) = \mathbb{E}_{X,y}\left[\frac{1}{2D}\|y - f_K(X)\|_2^2\right], \tag{16}$$

where $\lambda_{\max}[\cdot]$ denotes the largest eigenvalue and $\nabla^2_{K,K}$ denotes the Hessian.

To describe the dynamics, we consider the gradient-flow limit of unregularized online SGD,

$$\frac{\mathrm{d}K}{\mathrm{d}\tau} = -D\nabla_K \mathcal{E}_{\mathrm{pop}}(K) \tag{17}$$

with independent initial weights $k_h(\tau = 0) \sim \mathcal{N}(0, \eta^2 \mathbb{I}_D)$. We take the limit of vanishing initialization $\eta \to 0$ after $D \to \infty$, often referred to as the feature-learning limit (e.g. Wu et al. (2025)). Sagitova et al. (2026) derives closed evolution equations for the order parameters and shows that specialization is governed by an exponential escape from the saddle of the head-feature overlap m. A self-contained derivation is provided in App. D.1. We compare these dynamics with ERM (Res. 2.1) in the limit of large regularization limit $\lambda \gg 1$ in which the the ERM landscape is well approximated by the population loss. A linear stability analysis (App. D.2) then relates the critical sample complexity to the specialization time through the Hessian (16). Defining τ^{spec} as the first time at which the specializing component reaches a fixed small amplitude $\|m_{:1}\| > \varepsilon$, for small enough $\varepsilon > 0$ independent of D and η, we obtain:

$$\lim_{\eta \to 0^+} \lim_{D \to \infty} \frac{\tau^{\mathrm{spec}}}{\log(\sqrt{D}/\eta)} = \lim_{\lambda \to \infty} \frac{\alpha_c^{\mathrm{spec}}}{\lambda} = \frac{1}{C}. \tag{18}$$

In the considered limits the rescaled static and dynamical specialization transitions are governed by the same instability. The rescaled time is asymptotically deterministic and independent of the initialization and of ε. As in online learning training, time and number of observed sequences increase together, this correspondence illustrates how a statistical phase transition can manifest itself dynamically through sudden drops in the loss at critical training times.

5 Discussion and conclusion

The theoretical picture developed in this work suggests that abrupt and stagewise changes observed during training of multi-head attention layers can be dynamical manifestations of underlying statistical phase transitions. Solvable models make this perspective precise by identifying a limit in which specialization phase transitions are well defined, with sample complexity as a control parameter for studying the minima of the ERM objective. In our setting, this description distinguishes the statistical origins of specialization, encoded in the structure of the ERM loss landscape, from its dynamical origins. This theoretical perspective helps to clarify a possible statistical interpretation of the drops in loss that accompany abrupt changes in capabilities during the training of language models.

MLM experiments with BERT models provide strong empirical support for this statistical interpretation. Syntactic attention structure associated with head specialization emerges sharply with increasing dataset sizes, coherently with the already observed drops in the loss in training time, as suggested by the connections between ERM and online learning. While, by nature, finite-dimensional experiments do not by themselves establish the asymptotic non-analyticities that define phase transitions, these observations parallel the specialization phenomenology of the solvable model and strongly suggest a statistical phase transition interpretation of the emergence of specialized heads and SAS in BERT. For head specialization, sample complexity appears as the principled choice of control parameter to study these phase transitions. More generally, other capability changes could also have purely dynamical origins and not have a corresponding statistical characterization.

Our results motivate further experiments to characterize specialization transitions yet more systematically. For example, the enhanced variability across seeds near the emergence of SAS observed in Fig. 1 suggests that fluctuations could help locate candidate transitions in sample complexity. This approach is motivated by the diverging sensitivity of macroscopic observables to perturbations near second-order phase transitions, a familiar phenomenon in physics. Related considerations about variability across runs are beginning to be explored in mechanistic interpretability (Juneja et al., 2023; Zhao et al., 2025). A further test of the sample-complexity interpretation would extend our BERT experiments by varying training procedures with regularized objective and training controlled to convergence. For protocols reaching comparable ERM minima, one could examine whether SAS emerges at the same characteristic dataset size. Repeating this comparison across model sizes could help identify an effective scaling of sample complexity, while extending the analysis to individual syntactic relations could reveal distinct thresholds for the emergence of statistically more complex capabilities.

AI USE STATEMENT

In this work, we used generative AI tools to help formulate mathematical claims, implement methods, and polish writing. We have not used generative AI tools to generate synthetic data sets, help develop theoretical models or conceptual frameworks, propose or refine hypotheses, design or provide feedback on research methodology or experiments, assist with translation, clean and reformat dataset, support qualitative and thematic data analysis, interpret results. We have reviewed all AI-assisted work. We used LLMs to sketch extensions of our hand-derived linear stability analysis to other solutions and to adapt our stability-based numerical search for critical sample complexity to other phase boundaries. We also used LLMs to help implement and reproduce a previously designed experiment using a reference paper's repository, including identifying a word-indexing bug affecting the WSJ validation data. Finally, we used LLMs for language editing (grammar, spelling, and word choice). We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

A complete derivation of all the claims is included in the appendix with references provided in the main text. For numerical simulations, source code is submitted as supplementary material. For the experiments setup, details are provided in App. E.

REFERENCES

Tatsuya Aoyama, Ethan Gotlieb Wilcox, and Nathan Schneider. Predicting the emergence of induction heads in language model pretraining. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=50E1hqm6qk.

Benjamin Aubin, Antoine Maillard, Jean Barbier, Florent Krzakala, Nicolas Macris, and Lenka Zdeborová. The committee machine: Computational to statistical gaps in learning a two-layers neural network. *Advances in Neural Information Processing Systems*, 31, 2018.

Fabrizio Boncoraglio, Vittorio Erba, Emanuele Troiani, Yizhou Xu, Florent Krzakala, and Lenka Zdeborová. Single-head attention in high dimensions: A theory of generalization, weights spectra, and scaling laws. *International Conference on Machine Learning*, 2026. preprint arXiv:2509.24914.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Angelica Chen, Ravid Shwartz-Ziv, Kyunghyun Cho, Matthew Leavitt, and Naomi Saphra. Sudden drops in the loss: Syntax acquisition, phase transitions, and simplicity bias in mlms. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 14479–14510, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/3e8df255a03b383f30abe09203770213-Paper-Conference.pdf.

Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. In *Conference on Learning Theory (COLT)*, 2024b. arXiv:2402.19442.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. What does BERT look at? an analysis of BERT's attention. In Tal Linzen, Grzegorz Chrupała, Yonatan Belinkov, and Dieuwke Hupkes (eds.), *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 276–286, Florence, Italy, August 2019. Association for Computational Linguistics. doi: 10.18653/v1/W19-4828. URL https://aclanthology.org/W19-4828/.

Hugo Cui and Lenka Zdeborová. High-dimensional asymptotics of denoising autoencoders. *Journal of Statistical Mechanics: Theory and Experiment*, 2024(10):104018, oct 2024. doi: 10.1088/1742-5468/ad65e1. URL https://doi.org/10.1088/1742-5468/ad65e1.

Hugo Cui, Freya Behrens, Florent Krzakala, and Lenka Zdeborová. A phase transition between positional and semantic learning in a solvable model of dot-product attention. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 36342–36389. Curran Associates, Inc., 2024. doi: 10.52202/079017-1146. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3fefebc2d4e3c1c6ee9b892bd293117d-Paper-Conference.pdf.

O Duranthon, Pierre Marion, Claire Boyer, Bruno Loureiro, and Lenka Zdeborova. Statistical advantage of softmax attention: Insights from single-location regression. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 87068–87100, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8c21b98a47fd3b3783feba9784feffda-Paper-Conference.pdf.

Vittorio Erba, Emanuele Troiani, Lenka Zdeborová, and Florent Krzakala. The nuclear route: Sharp asymptotics of erm in overparameterized quadratic networks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 88862–88901. Curran Associates, Inc., 2025. doi: 10.52202/085713-2971. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/8039ca1e9860daab3a79e45d010d5398-Paper-Conference.pdf.

Federica Gerace, Bruno Loureiro, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Generalisation error in learning with random features and the hidden manifold model. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 3452–3462. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/gerace20a.html.

Jesse Hoogland, George Wang, Matthew Farrugia-Roberts, Liam Carroll, Susan Wei, and Daniel Murfet. Loss landscape degeneracy and stagewise development in transformers. *Transactions on Machine Learning Research*, 2025. arXiv preprint arXiv:2402.02364.

Jeevesh Juneja, Rachit Bansal, Kyunghyun Cho, João Sedoc, and Naomi Saphra. Linear connectivity reveals generalization strategies. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=hY6M0JHl3uL.

Itay Lavie, Kirsten Fischer, Andrey Lekov, Frederic Van Maele, Zohar Ringel, and Moritz Helias. Phase transitions in attention: A bayesian theory of copy head emergence, 2026. URL https://arxiv.org/abs/2606.12058.

Mitchell P. Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz, and Ann Taylor. Treebank-3, 1999. URL https://catalog.ldc.upenn.edu/LDC99T42. LDC Catalog No. LDC99T42.

Pierre Marion, Raphaël Berthier, Gérard Biau, and Claire Boyer. Attention layers provably solve single-location regression. In *International conference on learning representations*, volume 2025, pp. 88447–88488, 2025.

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

Rodrigo Pérez Ortiz, Gibbs Nwemadji, Jean Barbier, Federica Gerace, Alessandro Ingrosso, Clarissa Lauditi, and Enrico M Malatesta. Generalization performance of narrow shallow neural networks in the teacher–student setting. *Journal of Statistical Mechanics: Theory and Experiment*, 2026 (4):043405, 2026.

Tian Qin, Naomi Saphra, and David Alvarez-Melis. Data drives unstable hierarchical generalization in LMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11722–11740, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.593. URL https://aclanthology.org/2025.emnlp-main.593/.

M. Sagitova, O. Duranthon, and L. Zdeborová. Specialization of softmax attention heads: insights from the high-dimensional single-location model, 2026. URL https://arxiv.org/abs/2603.03993.

Sebastian Schuster and Christopher D. Manning. Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 2371–2378, Portorož, Slovenia, May 2016. European Language Resources Association (ELRA). URL https://aclanthology.org/L16-1376/.

H. Schwarze and J. Hertz. Generalization in a large committee machine. *EPL (Europhysics Letters)*, 20(4):375–380, 1992.

H. Schwarze and J. Hertz. Generalization in fully connected committee machines. *EPL (Europhysics Letters)*, 21(7):785–790, 1993.

Aaditya Singh, Stephanie Chan, Ted Moskovitz, Erin Grant, Andrew Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 27801–27819. Curran Associates, Inc., 2023. doi: 10.52202/075280-1208. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/58692a1701314e09cbd7a5f5f3871cc9-Paper-Conference.pdf.

Aaditya K Singh, Ted Moskovitz, Felix Hill, Stephanie C.Y. Chan, and Andrew M Saxe. What needs to go right for an induction head? A mechanistic study of in-context learning circuits and their formation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 45637–45662. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/singh24c.html.

Emanuele Troiani, Hugo Cui, Yatin Dandi, Florent Krzakala, and Lenka Zdeborová. Fundamental limits of learning in sequence multi-index models and deep attention networks: High-dimensional asymptotics and sharp thresholds. *Proceedings of the 42nd International Conference on Machine Learning*, 2025. arXiv preprint arXiv:2502.00901.

Matteo Vilucchio, Yatin Dandi, Matéo Pirio Rossignol, Cedric Gerbelot, and Florent Krzakala. Asymptotics of non-convex generalized linear models in high-dimensions: A proof of the replica formula. *arXiv preprint arXiv:2502.20003*, 2025.

George Wang, Jesse Hoogland, Stan van Wingerden, Zach Furman, and Daniel Murfet. Differentiation and specialization of attention heads via the refined local learning coefficient. In *International Conference on Learning Representations*, volume 2025, pp. 101037–101082, 2025.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=yzkSU5zdwD. Survey Certification.

Frank Zhengqing Wu, Berfin Simsek, and Francois Gaston Ged. Loss landscape of shallow relu-like neural networks: Stationary points, saddle escape, and network embedding. In *International Conference on Learning Representations*, 2025. arxiv:2402.05626.

Kayo Yin and Jacob Steinhardt. Which attention heads matter for in-context learning? In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 72428–72461. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yin25e.html.

Oğuz Kaan Yüksel, Rodrigo Alvarez Lucendo, and Nicolas Flammarion. Incremental learning of sparse attention patterns in transformers. *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2602.19143. preprint arxiv:2602.19143.

Yedi Zhang, Aaditya K. Singh, Peter E. Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. In *Proceedings of the 42th International Conference on Machine Learning (ICML)*, 2025. arXiv:2501.16265.

Rosie Zhao, Tian Qin, David Alvarez-Melis, Sham Kakade, and Naomi Saphra. Random scaling of emergent capabilities. *arXiv preprint arXiv:2502.17356*, 2025.

APPENDIX: TABLE OF CONTENTS

A Derivations for the exact characterization of the ERM **15**

A.1 Derivation of result 2.1 with the replica method 15

A.2 Model-specific saddle-point equations 19

B Model details and supplementary figures **20**

B.1 Flipping spike model 20

B.2 Non-isotropic Gaussian model 21

C Local stability and phase boundaries **21**

C.1 Flipping-spike model 22

C.2 Non-isotropic Gaussian model 23

D Link between statistical transitions and dynamical emergence **23**

D.1 Online gradient flow 23

D.2 ERM 24

E Implementation details **25**

E.1 Experiments with solvable models 25

E.2 Experiments with BERT MLMs 26

A Derivations for the exact characterization of the ERM

A.1 Derivation of result 2.1 with the replica method

This appendix provides a derivation of result 2.1 based on the replica method of the statistical physics of disordered systems.

A.1.1 Setting and notation

Recall that our goal is to characterize the minimizers of the regularized empirical risk (3) for the multi-head attention estimator (2) in the high-dimensional setting of Sec. 1. We take $N, D \to \infty$ at fixed $\alpha = N/D$, H, F, and L, such that $D^{-1}K^*K^{*\top} \to \mathbb{I}_F$. Let $\hat{K}$ be the minimizer of the empirical risk:

$$\mathcal{L}(K) = \sum_{\mu=1}^{N} \ell\big(y_\mu, f_K(X_\mu)\big) + \frac{\lambda}{2}\|K\|_F^2, \qquad \ell(y,\hat{y}) = \frac{1}{2D}\|y-\hat{y}\|^2, \tag{19}$$

where f_K is the estimator (2), and we assume $\lambda > 0$. Both error and regularization contributions to the loss are extensive; we define the unregularized training error associated with $\hat{K}$

$$\mathcal{E}_{\text{train}} = \frac{1}{N}\sum_{\mu=1}^{N} \ell\left(y^\mu, f_{\hat{K}}(X^\mu)\right). \tag{20}$$

We first derive the results for generic feature weights distribution P_θ, and later evaluate the remaining feature average for the flipping spike and non isotropic Gaussian models. We also consider generic activation function $\sigma : \mathbb{R}^L \to \mathbb{R}^L$, and specialize to softmax at the end of the derivation.

A.1.2 Gibbs formulation of the minimization problem

To begin with, we define the Gibbs measure $d\mu_\beta$ over weights $K \in \mathbb{R}^{H\times D}$, and recall that for any test function $\phi(\hat{K})$ of the minimizer $\hat{K}$ of (3),

$$\phi(\hat{K}) = \lim_{\beta\to\infty} \mathbb{E}_{\mathcal{D}}\left[\int \phi(K) d\mu_\beta\right], \qquad d\mu_\beta = \frac{1}{\mathcal{Z}_\beta} e^{-\beta\mathcal{L}(K)} dK, \tag{21}$$

where $\mathcal{Z}_\beta$ is the normalization factor also known as partition function in statistical physics:

$$\mathcal{Z}_\beta = \int dK\; e^{-\beta\mathcal{L}(K)}. \tag{22}$$

The goal of the replica method is to compute the generating function associated to the Gibbs measure (21), the free entropy density

$$\Phi_\beta = \lim_{D\to\infty} \frac{1}{D}\,\mathbb{E}_{\mathcal{D}}\left[\log \mathcal{Z}_\beta\right], \tag{23}$$

building on the identity

$$\mathbb{E}_{\mathcal{D}}\left[\log \mathcal{Z}_\beta\right] = \lim_{n\to 0} \frac{\mathbb{E}_{\mathcal{D}}[\mathcal{Z}_\beta^n] - 1}{n}. \tag{24}$$

Therefore, the whole problem reduces to computing the average replicated partition function:

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{Z}_\beta^n\right] = \mathbb{E}_{\mathcal{D}}\left[\int \prod_{a=1}^{n} dK^a\; e^{-\beta\sum_{a=1}^{n}\left[\sum_{\mu=1}^{N}\ell(y_\mu, f_{K^a}(X_\mu)) + \frac{\lambda}{2}\|K^a\|_F^2\right]}\right]. \tag{25}$$

A.1.3 Replicated partition function

With explicit dataset definitions, the partition function reads

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{Z}_\beta^n\right] = \int dP_{K^*} \prod_{a=1}^{n} dK^a\, e^{-\frac{\beta\lambda}{2}\sum_{a=1}^{n}\|K^a\|_F^2} \prod_{\mu=1}^{N} \mathbb{E}_{\epsilon_\mu,\theta_\mu,X_\mu,y_\mu|K^*}\left[e^{-\beta\sum_{a=1}^{n}\ell\left(y_\mu, f_{K^a}(X_\mu)\right)}\right]. \tag{26}$$

We introduce the local fields

$$Z_\mu^a = \frac{X_\mu K^{a\top}}{\sqrt{D}} \in \mathbb{R}^{L\times H}, \qquad Z_\mu^* = \frac{X_\mu K^{*\top}}{\sqrt{D}} \in \mathbb{R}^{L\times F}, \tag{27}$$

following jointly Gaussian statistics with means

$$\mathbb{E}_X\left[Z_\ell^a\right] = \delta_{\ell\epsilon} m^a \theta, \qquad \mathbb{E}_X\left[Z_\ell^*\right] = \delta_{\ell\epsilon}\theta, \tag{28}$$

and covariances

$$\mathrm{Cov}\left(Z_\ell^a, Z_{\ell'}^{b\top}\right) = \delta_{\ell\ell'}\frac{K^a K^{b\top}}{D}, \quad \mathrm{Cov}\left(Z_\ell^a, Z_{\ell'}^*\right) = \delta_{\ell\ell'}\frac{K^a K^{*\top}}{D}, \quad \mathrm{Cov}\left(Z_\ell^*, Z_{\ell'}^*\right) = \delta_{\ell\ell'}\mathbb{I}_F. \tag{29}$$

Expanding the squared loss then yields the reparametrization

$$\ell(y_\mu, f_{K^a}(X_\mu)) \to \ell(\epsilon_\mu, Z_\mu^a) \coloneqq \frac{1}{2}\sum_{\ell=1}^{L}\left(\delta_{\ell\epsilon_\mu} - \frac{1}{H}\sum_{h=1}^{H}\sigma\left(Z_{\mu,:h}^a\right)_\ell\right)^2, \tag{30}$$

where we used the fact that $X_\ell^\top X_{\ell'}/D \to \delta_{\ell\ell'}$ in the considered limit. Enforcing the definitions of local fields with Dirac deltas in the replicated partition function gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\mathcal{Z}_\beta^n\right] \propto \int dP_{K^*} \prod_{a=1}^{n} dK^a\, e^{-\frac{\beta\lambda}{2}\sum_{a=1}^{n}\|K^a\|_F^2} \\
\prod_{\mu=1}^{N} dZ_\mu^* d\hat{Z}_\mu^* \prod_{a=1}^{n} dZ_\mu^a d\hat{Z}_\mu^a\, \mathbb{E}_{\epsilon_\mu,\theta_\mu}\left\{e^{-\beta\sum_{a=1}^{n}\ell(\epsilon_\mu, Z_\mu^a)}\right. \\
\left.\mathbb{E}_{X_\mu|\epsilon_\mu,\theta_\mu,K^*}\left[e^{i\sum_{\ell=1}^{L}\hat{Z}_{\mu,\ell}^{*\top}\left(Z_{\mu,\ell}^* - \frac{K^* X_{\mu,\ell}}{\sqrt{D}}\right) + i\sum_{\ell=1}^{L}\sum_{a=1}^{n}\hat{Z}_{\mu,\ell}^{a\top}\left(Z_{\mu,\ell}^a - \frac{K^a X_{\mu,\ell}}{\sqrt{D}}\right)}\right]\right\} \qquad (31) \\
\propto \int dP_{K^*} \prod_{a=1}^{n} dK^a\, e^{-\frac{\beta\lambda}{2}\sum_{a=1}^{n}\|K^a\|_F^2} \\
\prod_{\mu=1}^{N} dZ_\mu^* d\hat{Z}_\mu^* \prod_{a=1}^{n} dZ_\mu^a d\hat{Z}_\mu^a\, \mathbb{E}_{\epsilon_\mu,\theta_\mu}\left\{e^{-\beta\sum_{a=1}^{n}\ell(\epsilon_\mu, Z_\mu^a) + \theta_\mu^\top Z_{\mu,\epsilon_\mu}^* - \frac{1}{2}\theta_\mu^\top\theta_\mu}\right. \\
\left.\mathbb{E}_{X_\mu\sim\mathcal{N}(0,\mathbb{I}_D)}\left[e^{i\sum_{\ell=1}^{L}\hat{Z}_{\mu,\ell}^{*\top}\left(Z_{\mu,\ell}^* - \frac{K^* X_{\mu,\ell}}{\sqrt{D}}\right) + i\sum_{\ell=1}^{L}\sum_{a=1}^{n}\hat{Z}_{\mu,\ell}^{a\top}\left(Z_{\mu,\ell}^a - \frac{K^a X_{\mu,\ell}}{\sqrt{D}}\right)}\right]\right\} \qquad (32) \\
\propto \int dP_{K^*} \prod_{a=1}^{n} dK^a\, e^{-\frac{\beta\lambda}{2}\sum_{a=1}^{n}\|K^a\|_F^2} \prod_{\mu=1}^{N} dZ_\mu^* \prod_{a=1}^{n} dZ_\mu^a \\
\mathbb{E}_{\epsilon_\mu,\theta_\mu}\left\{e^{-\beta\sum_{a=1}^{n}\ell(\epsilon_\mu, Z_\mu^a) + \theta_\mu^\top Z_{\mu,\epsilon_\mu}^* - \frac{1}{2}\theta_\mu^\top\theta_\mu}\prod_{\ell=1}^{L}\mathcal{N}\left(\begin{pmatrix}Z_{\mu,\ell}^* \\ Z_{\mu,\ell}\end{pmatrix}; 0, \begin{pmatrix}\mathbb{I}_F & m^\top \\ m & Q\end{pmatrix}\right)\right\}
\end{aligned} \tag{33}$$

where we introduced the summary statistics

$$m^a = \frac{K^a K^{*\top}}{D} \in \mathbb{R}^{H\times F}, \qquad Q^{ab} = \frac{K^a K^{b\top}}{D} \in \mathbb{R}^{H\times H}. \tag{34}$$

with $Z_{\mu,\ell} = \{Z^a_{\mu,\ell}\}_{a=1}^n$, $m = \{m^a\}_{a=1}^n$, $Q = \{Q^{ab}\}_{a,b=1}^n$. Enforcing the definitions of m^a and Q^{ab} with Dirac functions brings the replicated partition function in the following form:

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{Z}_\beta^n\right] \propto \int \prod_{a=1}^n dm^a d\hat{m}^a \prod_{a,b=1}^n dQ^{ab} d\hat{Q}^{ab} \underbrace{e^{-D\sum_{a=1}^n \mathrm{Tr}(\hat{m}^{a\top} m^a) + \frac{D}{2}\sum_{a,b=1}^n \mathrm{Tr}(\hat{Q}^{ab\top} Q^{ab})}}_{e^{nD\Psi_t}}$$
$$\underbrace{\int dP_{K^*} \prod_{a=1}^n dK^a e^{-\frac{\beta\lambda}{2}\sum_{a=1}^n \|K^a\|_F^2 + \sum_{a=1}^n \mathrm{Tr}(\hat{m}^{a\top} K^a K^{*\top}) - \frac{1}{2}\sum_{a,b=1}^n \mathrm{Tr}(\hat{Q}^{ab} K^b K^{a\top})}}_{e^{nD\Psi_K}}$$
$$\underbrace{\mathbb{E}_{\epsilon,\theta}\left\{\mathbb{E}_{Z^*,\{Z^a\}_{a=1}^n}\left[e^{-\beta\sum_{a=1}^n \ell(\epsilon, Z^a) + \theta^\top Z^*_\epsilon - \frac{1}{2}\theta^\top\theta}\right]\right\}^{\alpha D}}_{e^{n\alpha D\Psi_y}}. \tag{35}$$

We decomposed the replicated free entropy into the trace Ψ_t, entropic Ψ_K and energetic Ψ_y potentials. In the high-dimensional limit where $D \to \infty$ while $\alpha = N/D$ stay finite, the integral (35), written explicitly as

$$\mathbb{E}_{\mathcal{D}}\left[\mathcal{Z}_\beta^n\right] \propto \int \prod_{a=1}^n dm^a d\hat{m}^a \prod_{a,b=1}^n dQ^{ab} d\hat{Q}^{ab} e^{D\Phi_\beta^{(n)}\left(m^a, \hat{m}^a, Q^{ab}, \hat{Q}^{ab}\right)}, \tag{36}$$

concentrate around the values of the overlaps that extremize the finite temperature free entropy $\Phi_\beta^{(n)}(m^a, \hat{m}^a, Q^{ab}, \hat{Q}^{ab}) = n\left(\Psi_t + \Psi_K + \alpha\Psi_y\right)$, and can be evaluated using a Laplace saddle-point approximation.

A.1.4 Replica-symmetric ansatz

We have thus rephrased the analysis of the measure (21) as a optimization problem over the order parameters $\{m^a, Q^{ab}\}$, and the associated conjugate variables. However, these still represent $2nHF + 2n^2H^2$ variables with $n \to 0$. In order to make progress, we assume that the saddle point is of replica-symmetric (RS) form

$$m^a = m \in \mathbb{R}^{H\times F}, \qquad \hat{m}^a = \hat{m} \in \mathbb{R}^{H\times F}, \tag{37}$$
$$Q^{ab} = q\delta_{a\neq b} + Q\delta_{ab} \in \mathbb{R}^{H\times H}, \qquad \hat{Q}^{ab} = -\hat{q}\delta_{a\neq b} + \hat{Q}\delta_{ab} \in \mathbb{R}^{H\times H}. \tag{38}$$

The RS ansatz holds in a number of machine learning settings, notably for convex problems and Bayes-optimal settings. In the present setting, since the empirical loss (19) is non-convex, the RS ansatz constitutes a heuristic technical assumption of our analysis.

Under this ansatz, we can factorize the replicas by introducing random Gaussian variables $\Xi = \{\xi_\ell\}_{\ell=1}^L$ with $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_H)$. With quadratic regularization, we can perform the integration over the weights K in the entropic term. Introducing the variance order parameter

$$V = Q - q, \qquad \hat{V} = \hat{Q} + \hat{q}, \tag{39}$$

and taking the limit $n \to 0$, the free entropy can then be expressed as

$$\Phi_\beta = \lim_{n\to 0}\frac{1}{n}\underset{m,\hat{m},q,\hat{q},V,\hat{V}}{\mathrm{extr}}\left\{\Phi_\beta^{(n)}(m^a, \hat{m}^a, Q^{ab}, \hat{Q}^{ab})\right\} = \underset{m,\hat{m},q,\hat{q},V,\hat{V}}{\mathrm{extr}}\left\{\Psi_t + \Psi_K + \alpha\Psi_y\right\} \tag{40}$$

with

$$\Psi_t = -\mathrm{Tr}(\hat{m}^\top m) + \frac{1}{2}\mathrm{Tr}\left[\hat{V}(q+V) - \hat{q}V\right] \tag{41}$$
$$\Psi_K = -\frac{1}{2}\log\left[\det(\beta\lambda\mathbb{I}_H + \hat{V})\right] + \frac{1}{2}\mathrm{Tr}\left[(\beta\lambda\mathbb{I}_H + \hat{V})^{-1}(\hat{q} + \hat{m}\hat{m}^\top)\right] \tag{42}$$
$$\Psi_y = \mathbb{E}_{\Xi,\epsilon,\theta}\left\{\log\left[\int dZ e^{-\beta\ell(\epsilon,Z)}\frac{e^{-\frac{1}{2}\sum_{\ell=1}^L (Z_\ell - \omega_\ell)^\top V^{-1}(Z_\ell - \omega_\ell)}}{\det(2\pi V)^{L/2}}\right]\right\} \tag{43}$$

where we introduced the local fields $\Omega = \{\omega_\ell\}_{\ell=1}^L$ with $\omega_\ell \coloneqq \delta_{\ell\epsilon} m\theta + q^{1/2}\xi_\ell$ and we omit additive constants in the potentials. Note that these expressions do not depend on K^* anymore; the planted directions enter only through $K^*K^{*\top}/D \to \mathbb{I}_F$, such that no further specification or averaging of their distribution is required.

A.1.5 ZERO TEMPERATURE LIMIT

We now take the limit $\beta \to \infty$, rescaling

$$\hat{m} \to \beta \hat{m}, \qquad \hat{q} \to \beta^2 \hat{q}, \qquad \hat{V} \to \beta \hat{V}, \qquad V \to \beta^{-1} V. \tag{44}$$

The potentials and free entropy reduce to

$$\Phi = \lim_{\beta \to \infty} \frac{1}{\beta} \Phi_\beta = \underset{m, \hat{m}, q, \hat{q}, V, \hat{V}}{\operatorname{extr}} \{ \Psi_t + \Psi_K + \alpha \Psi_y \} \tag{45}$$

with

$$\Psi_t = -\mathrm{Tr}(\hat{m}^\top m) + \frac{1}{2} \mathrm{Tr}(\hat{V} q - \hat{q} V) \tag{46}$$

$$\Psi_K = \frac{1}{2} \mathrm{Tr} \left[(\lambda \mathbb{I}_H + \hat{V})^{-1} (\hat{q} + \hat{m} \hat{m}^\top) \right] \tag{47}$$

$$\Psi_y = -\mathbb{E}_{\Xi, \epsilon, \theta} \left[\mathcal{M}(\epsilon, \theta, \Xi) \right] \tag{48}$$

where the Moreau envelope is defined as

$$\mathcal{M}(\epsilon, \theta, \Xi) = \inf_{Z \in \mathbb{R}^{L \times H}} \left\{ \ell(\epsilon, Z) + \frac{1}{2} \sum_{\ell=1}^{L} (Z_\ell - \omega_\ell)^\top V^{-1} (Z_\ell - \omega_\ell) \right\}. \tag{49}$$

A.1.6 REPLICA FREE ENTROPY

Up to this point, the derivation can be carried generally with respect to the activation function and low dimensional distributions of the data model. One already reached a low dimensional characterization and can in principle state the final result generally. For clarity, we state the results for the particular setting studied. Specifying the distribution of ϵ and using permutation equivariance of softmax with respect to the tokens to fix $\epsilon = 1$, we can simplify further the equations (45-49). One finally reaches an expression for the replica free entropy:

$$\Phi = \underset{m, \hat{m}, q, \hat{q}, V, \hat{V}}{\operatorname{extr}} \left\{ \frac{1}{2} \mathrm{Tr}(\hat{V} q - \hat{q} V) - \mathrm{Tr}(\hat{m}^\top m) + \frac{1}{2} \mathrm{Tr} \left[(\lambda \mathbb{I}_H + \hat{V})^{-1} (\hat{q} + \hat{m} \hat{m}^\top) \right] - \alpha \mathbb{E}_{\Xi, \theta} \left[\mathcal{M}(\theta, \Xi) \right] \right\} \tag{50}$$

where we defined the local fields $\omega_\ell = \delta_{\ell 1} m \theta + q^{1/2} \xi_\ell$ with $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_H)$ and the Moreau envelope

$$\mathcal{M}(\theta, \Xi) = \inf_{Z \in \mathbb{R}^{L \times H}} \left\{ \frac{1}{2} \sum_{\ell=1}^{L} \left(\delta_{\ell 1} - \frac{1}{H} \sum_{h=1}^{H} \mathrm{softmax}(Z_{:h})_\ell \right)^2 + \frac{1}{2} \sum_{\ell=1}^{L} (Z_\ell - \omega_\ell)^\top V^{-1} (Z_\ell - \omega_\ell) \right\}. \tag{51}$$

A.1.7 SADDLE-POINT EQUATIONS

The extremization for the free entropy (50) obtained by setting the gradient of its potential to zero yields the following system of self-consistent equations on the summary statistics:

$$\begin{cases} m = (\lambda \mathbb{I}_H + \hat{V})^{-1} \hat{m} \\ q = (\lambda \mathbb{I}_H + \hat{V})^{-1} (\hat{m} \hat{m}^\top + \hat{q}) (\lambda \mathbb{I}_H + \hat{V})^{-1} \\ V = (\lambda \mathbb{I}_H + \hat{V})^{-1} \end{cases} \tag{52}$$

$$\begin{cases} \hat{m} = \alpha \mathbb{E}_{\theta, \Xi} \left[V^{-1} \left(\mathrm{prox}(\theta, \Xi)_1 - \omega_1 \right) \theta^\top \right] \\ \hat{q} = \alpha \mathbb{E}_{\theta, \Xi} \left[V^{-1} \left(\sum_{\ell=1}^{L} \left(\mathrm{prox}(\theta, \Xi)_\ell - \omega_\ell \right)^{\otimes 2} \right) V^{-1} \right] \\ \hat{V} = -\alpha \mathbb{E}_{\theta, \Xi} \left[V^{-1} \left(\sum_{\ell=1}^{L} \left(\mathrm{prox}(\theta, \Xi)_\ell - \omega_\ell \right) \xi_\ell^\top \right) q^{-1/2} \right] \end{cases} \tag{53}$$

where the proximals $\{\mathrm{prox}(\theta, \Xi)_\ell\}_{\ell=1}^{L}$ refer to the minimizers of the Moreau envelope $\mathcal{M}(\theta, \Xi)$ (51).

Note that the update equation for $\hat{V}$ contains a term $q^{-1/2}$ that could become singular, in particular on the PS branch. This apparent singularity is removable using integration by parts

$$\hat{V} = -\alpha \, \mathbb{E}_{\theta, \Xi} \left[\sum_{\ell=1}^{L} V^{-1} \left(\frac{\partial \, \mathrm{prox}(\theta, \Xi)_\ell}{\partial \omega_\ell} - \mathbb{I}_H \right) \right], \tag{54}$$

and we use this formulation to improve the stability of the numerical implementation of state evolution for the non-isotropic Gaussian model.

A.1.8 TEST ERROR

The generalization performance is measured by the test error

$$\mathcal{E} = \lim_{D\to\infty} \mathbb{E}_{\mathcal{D}} \left\{ \mathbb{E}_{X,y|K^*} \left[\frac{1}{2D} \|y - f_{\hat{K}}(X)\|^2 \right] \right\}, \tag{55}$$

where the test samples $\{X, y\}$ are independent of the training set $\mathcal{D}$, conditionally on the planted spikes K^*. Introducing the local fields $X\hat{K}^\top/\sqrt{D}$ and expliciting the expression for the data model, the test error admits an asymptotic reparameterization in terms of the summary statistics characterized by (52):

$$\mathcal{E} = \frac{1}{2}\mathbb{E}_{\theta,\Xi} \left[\sum_{\ell=1}^{L} \left(\delta_{\ell 1} - \frac{1}{H} \sum_{h=1}^{H} \mathrm{softmax}(\Omega_{:h})_\ell \right)^2 \right] \tag{56}$$

where, as before, $\theta \sim P_\theta$, $\Xi = \{\xi_\ell\}_{\ell=1}^L$ with $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_H)$, $\Omega = \{\omega_\ell\}_{\ell=1}^L$ $\omega_\ell = \delta_{\ell 1} m\theta + q^{1/2}\xi_\ell$, and we used token permutation equivariance of the softmax activation.

A.1.9 TRAINING ERROR

As noted in Cui et al. (2024), the unregularized training error (20) can be obtained through the free energy $-\Phi$ at zero temperature:

$$\mathcal{E}_{\mathrm{train}} = -\frac{1}{\alpha}\Phi - \frac{\lambda}{2\alpha}\mathrm{Tr}(q) = \mathbb{E}_{\theta,\Xi}\left[\mathcal{M}(\theta, \Xi)\right] - \frac{1}{2\alpha}\mathrm{Tr}(\hat{q}V), \tag{57}$$

where again, $\Xi = \{\xi_\ell\}_{\ell=1}^L$ with $\xi_\ell \sim \mathcal{N}(0, \mathbb{I}_H)$, and $\mathcal{M}(\theta, \Xi)$ denotes the Moreau envelope (51). This can also be obtained directly from the relation $\mathcal{L}_{\mathrm{train}} = -\lim_{\beta\to\infty} \partial_\beta \log \mathcal{Z}_\beta / D$, where the derivative is taken on the finite temperature free entropy (40) before the limit.

A.2 MODEL-SPECIFIC SADDLE-POINT EQUATIONS

We now evaluate the remaining feature averages for the two choices of P_θ defining the models of Sec. 1.

A.2.1 FLIPPING SPIKE MODEL

In the case of the flipping spike distribution where $\theta_1 = \sqrt{\nu_1}$ and $\theta_2 \sim \mathrm{Unif}(\{-\sqrt{\nu_2}, \sqrt{\nu_2}\})$ we can define

$$P_\theta = \frac{1}{2}(\delta_{\theta_+} + \delta_{\theta_-}), \qquad \theta_\pm = \begin{pmatrix} \sqrt{\nu_1} \\ \pm\sqrt{\nu_2} \end{pmatrix}. \tag{58}$$

The random variable is discrete and averaging over θ leads to the saddle-point equations:

$$\begin{cases} m = (\lambda \mathbb{I}_H + \hat{V})^{-1}\hat{m} \\ q = (\lambda \mathbb{I}_H + \hat{V})^{-1}(\hat{m}\hat{m}^\top + \hat{q})(\lambda \mathbb{I}_H + \hat{V})^{-1} \\ V = (\lambda \mathbb{I}_H + \hat{V})^{-1} \end{cases} \tag{59}$$

$$\begin{cases} \hat{m} = \frac{\alpha}{2}\sum_{s=\pm 1} \mathbb{E}_\Xi \left[V^{-1} \left(\mathrm{prox}(\Xi)_1^{(s)} - \omega_1^{(s)} \right) \theta_s^\top \right] \\ \hat{q} = \frac{\alpha}{2}\sum_{s=\pm 1} \mathbb{E}_\Xi \left[V^{-1} \left(\sum_{\ell=1}^L \left(\mathrm{prox}(\Xi)_\ell^{(s)} - \omega_\ell^{(s)} \right)^{\otimes 2} \right) V^{-1} \right] \\ \hat{V} = -\frac{\alpha}{2}\sum_{s=\pm 1} \mathbb{E}_\Xi \left[V^{-1} \left(\sum_{\ell=1}^L \left(\mathrm{prox}(\Xi)_\ell^{(s)} - \omega_\ell^{(s)} \right) \xi_\ell^\top \right) q^{-1/2} \right] \end{cases} \tag{60}$$

where we define the local fields $\Omega^{(s)} = \{\omega_\ell^{(s)}\}_{\ell=1}^L$ as $\omega_\ell^{(s)} := \delta_{\ell 1} m\theta_s + q^{1/2}\xi_\ell$ and with $\xi \sim \mathcal{N}(0, \mathbb{I}_H)$ the proximals $\{\mathrm{prox}(\Xi)_\ell^{(s)}\}_{\ell=1}^L$ refer to the minimizers of the Moreau envelope

$$\mathcal{M}(\Xi)^{(s)} = \inf_{Z\in\mathbb{R}^{L\times H}} \left\{ \frac{1}{2}\sum_{\ell=1}^{L} \left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H} \mathrm{softmax}(Z_{:h})_\ell \right)^2 + \frac{1}{2}\sum_{\ell=1}^{L} \left(Z_\ell - \omega_\ell^{(s)} \right)^\top V^{-1} \left(Z_\ell - \omega_\ell^{(s)} \right) \right\}. \tag{61}$$

The test error then becomes

$$\mathcal{E} = \frac{1}{4}\sum_{s=\pm 1}\mathbb{E}_{\Xi}\left[\sum_{\ell=1}^{L}\left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H}\mathrm{softmax}(\Omega_{:h}^{(s)})_{\ell}\right)^2\right], \tag{62}$$

and the train error

$$\mathcal{E}_{\mathrm{train}} = \frac{1}{2}\sum_{s=\pm 1}\mathbb{E}_{\Xi}\left[\mathcal{M}(\Xi)^{(s)}\right] - \frac{1}{2\alpha}\mathrm{Tr}(\hat{q}V). \tag{63}$$

A.2.2 Non-isotropic Gaussian model

In the case of the non-isotropic Gaussian distribution we have

$$\theta \sim \mathcal{N}(0,\Lambda), \qquad \Lambda = \mathrm{diag}(\nu_1,\dots,\nu_F). \tag{64}$$

Dependencies on θ enter the proximals only through the local fields $\Omega = \{\omega_\ell\}_{\ell=1}^{L}$ and we have

$$\mathbb{E}[\theta \mid \Omega] = \Lambda m^\top (q + m\Lambda m^\top)^{-1}\omega_1, \qquad \mathbb{E}[\xi_\ell \mid \Omega] = q^{1/2}(q+\delta_{\ell 1} m\Lambda m^\top)^{-1}\omega_\ell. \tag{65}$$

Averaging over (θ,Ξ) conditionally on the local fields then leads to the saddle-point equations:

$$\begin{cases} m = (\lambda \mathbb{I}_H + \hat{V})^{-1}\hat{m} \\ q = (\lambda \mathbb{I}_H + \hat{V})^{-1}(\hat{m}\hat{m}^\top + \hat{q})(\lambda \mathbb{I}_H + \hat{V})^{-1} \\ V = (\lambda \mathbb{I}_H + \hat{V})^{-1} \end{cases} \tag{66}$$

$$\begin{cases} \hat{m} = \alpha \mathbb{E}_{\Omega}\left[V^{-1}\left(\mathrm{prox}(\Omega)_1 - \omega_1\right)\omega_1^\top\right](q + m\Lambda m^\top)^{-1} m\Lambda \\ \hat{q} = \alpha \mathbb{E}_{\Omega}\left[V^{-1}\left(\sum_{\ell=1}^{L}\left(\mathrm{prox}(\Omega)_\ell - \omega_\ell\right)^{\otimes 2}\right)V^{-1}\right] \\ \hat{V} = -\alpha \mathbb{E}_{\Omega}\left[V^{-1}\left(\sum_{\ell=1}^{L}\left(\mathrm{prox}(\Omega)_\ell - \omega_\ell\right)\omega_\ell^\top (q+\delta_{\ell 1} m\Lambda m^\top)^{-1}\right)\right] \end{cases} \tag{67}$$

where $\Omega = \{\omega_\ell\}_{\ell=1}^{L}$ with $\omega_\ell \sim \mathcal{N}\left(0, q + \delta_{\ell 1} m\Lambda m^\top\right)$ and $\{\mathrm{prox}(\Omega)_\ell\}_{\ell=1}^{L}$ refer to the minimizers of the Moreau envelope

$$\mathcal{M}(\Omega) = \inf_{Z\in\mathbb{R}^{L\times H}}\left\{\frac{1}{2}\sum_{\ell=1}^{L}\left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H}\mathrm{softmax}(Z_{:h})_\ell\right)^2 + \frac{1}{2}\sum_{\ell=1}^{L}(Z_\ell - \omega_\ell)^\top V^{-1}(Z_\ell - \omega_\ell)\right\}. \tag{68}$$

The test error then becomes

$$\mathcal{E} = \frac{1}{2}\mathbb{E}_{\Omega}\left[\sum_{\ell=1}^{L}\left(\delta_{\ell 1} - \frac{1}{H}\sum_{h=1}^{H}\mathrm{softmax}(\Omega_{:h})_\ell\right)^2\right], \tag{69}$$

and the train error

$$\mathcal{E}_{\mathrm{train}} = \mathbb{E}_{\Omega}\left[\mathcal{M}(\Omega)\right] - \frac{1}{2\alpha}\mathrm{Tr}(\hat{q}V). \tag{70}$$

B Model details and supplementary figures

B.1 Flipping spike model

In the flipping spike model defined in Sec. 1, a constant latent direction is learnable without head specialization while a flipping direction requires specialized heads to be used in the retrieval task. The signal strengths (ν_1,ν_2) can be parametrized as

$$\nu_{\mathrm{tot}} = \nu_1 + \nu_2, \qquad \delta = \frac{\nu_2}{\nu_1+\nu_2}, \tag{71}$$

to control total signal strength ν_{tot} and flipping ratio δ. More total signal facilitates the retrieval task, while larger flipping ratio increases the importance of specialization. For $\nu_2 = 0$, task requires no head specialization, and no phase transition occurs. For $\nu_1 = 0$, there is no constant direction and specialization is necessary for generalization, as in the non-isotropic Gaussian model. In Fig. 5 we present the effect of the relative importance of the two features on specialization and performances.

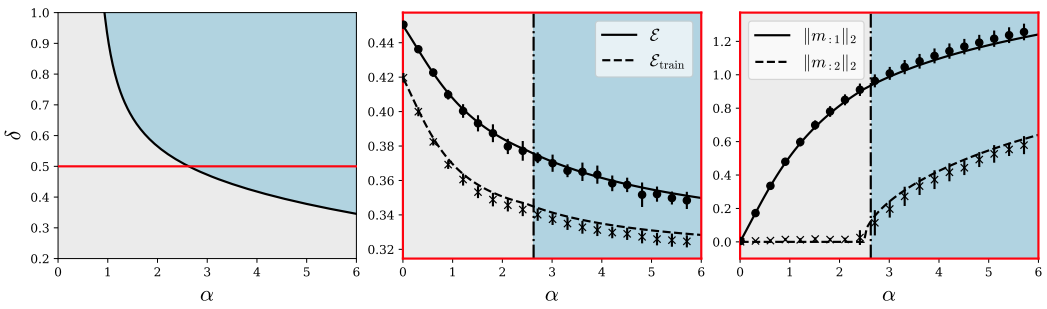

Figure 5: **Influence of signal composition on head specialization in the flipping spike model:** $\nu_{\text{tot}} = 4$, $L = 10$, $H = 2$, $\lambda = 0.2$. Black solid lines correspond to theoretical predictions (Res 2.1). Markers with error bars correspond to training the model (Sec. 1) with L-BFGS from random initialization in dimension $D = 2000$ across ten seeds. **Left:** Phase diagram in sample complexity α and flipping ratio δ. Phase boundaries are obtained by stability analysis of the different solutions (App. C). The red horizontal line at $\delta = 0.5$ corresponding to $\nu_1 = \nu_2 = 2$, selects the cut shown in the other panels. **Center:** Test error $\mathcal{E}$ and training error $\mathcal{E}_{\text{train}}$. **Right:** Norms of the head–feature overlaps $\|m_{:1}\|_2$ and $\|m_{:2}\|_2$. Vertical dash-dotted lines mark the specialization threshold α_c^{spec} along the selected cut.

B.2 Non-isotropic Gaussian model

In Fig. 6 we present additional figures for the non-isotropic Gaussian model, notably parametric curves for the head–feature overlaps. In Fig. 7 we present the same results for isotropic Gaussian distribution of the weights ($\nu_1 = \nu_2$).

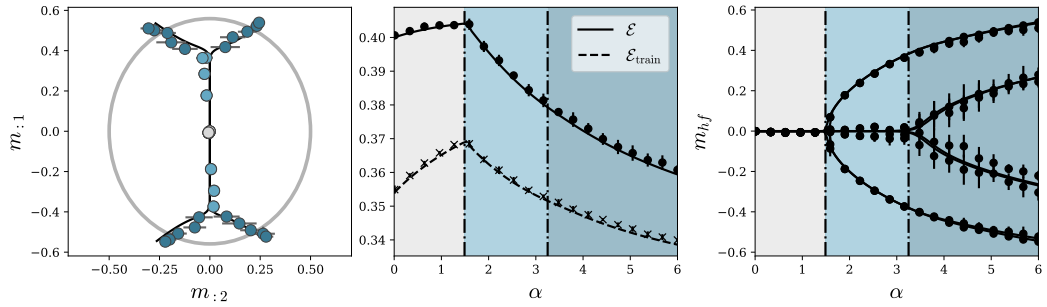

Figure 6: **Head-specialization phase transitions in the non-isotropic Gaussian model:** $\nu_1 = 5.0$, $\nu_2 = 4.0$, $L = 5$, $H = 4$, $\lambda = 0.2$. Black solid lines correspond to theoretical predictions (Res. 2.1). Markers with error bars correspond to training the model (Sec. 1) with L-BFGS from random initialization in dimension $D = 4000$ across ten seeds. **Left:** Head-feature overlaps $(m_{:2}, m_{:1})$ parametrized by sample complexity. Grey ellipse indicate the distribution anisotropy of possible draws of the planted spike $\hat{k}$. **Center:** Test error $\mathcal{E}$ and training error $\mathcal{E}_{\text{train}}$. **Right:** Head-feature overlaps m_{hf} for all heads and both latent features. Grey, light blue and darker blue identify the permutation-symmetric (PS) and successive specialized (1-SPEC, 2-SPEC) phases, respectively. Vertical dash-dotted lines mark the specialization thresholds $\alpha_c^{(1)}$ and $\alpha_c^{(2)}$.

C Local stability and phase boundaries

We derive the local stability conditions for the solutions of the self-consistent equations (6) by tracking small perturbations under the state-evolution (SE) map. These conditions are used to compute the phase boundaries shown in Figs. 3, 4, and 5.

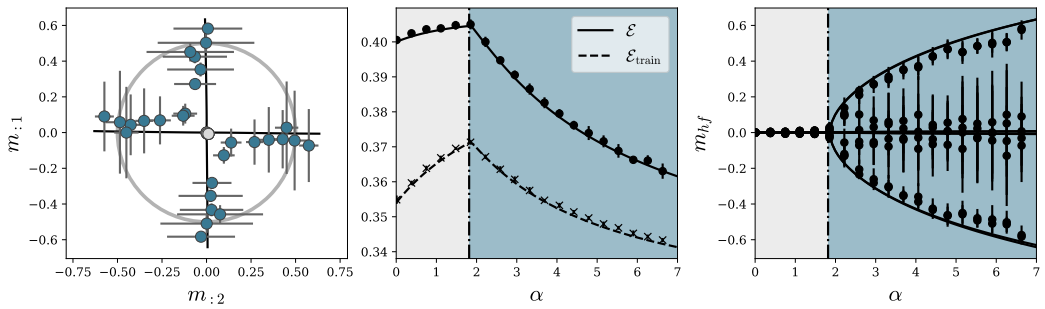

Figure 7: **Head-specialization phase transitions for isotropic Gaussian latent spikes:** $\nu_1 = 4.0$, $\nu_2 = 4.0$, $L = 5$, $H = 4$, $\lambda = 0.2$. Black solid lines correspond to theoretical predictions (Res. 2.1). Markers with error bars correspond to training the model (Sec. 1) with L-BFGS from random initialization in dimension $D = 2000$ across ten seeds. **Left:** Head-feature overlaps $(m_{:2}, m_{:1})$ parametrized by sample complexity. Grey circle indicates the distribution isotropy of possible draws of the planted spike $\hat{k}$. **Center:** Test error $\mathcal{E}$ and training error $\mathcal{E}_{\text{train}}$. **Right:** Head-feature overlaps m_{hf} for all heads and both latent features. Grey and darker blue identify the permutation-symmetric (PS) and specialized (2-SPEC) phases, respectively. Vertical dash-dotted lines mark the specialization thresholds $\alpha_c^{(1)} = \alpha_c^{(2)}$.

C.1 Flipping-spike model

Stability against specialization Both PS and PSB satisfy $m_{:2} = 0$, a condition preserved by the SE map. Denote an unperturbed solution by (m_0, q_0, V_0), its local fields by Ω^0, and its proximal by Z'^0, with $V_0 = (\lambda \mathbb{I}_H + \hat{V}_0)^{-1}$. We perturb $m_{:2} = \delta m$ and denote the corresponding conjugate perturbation by $\delta\hat{m}$. By symmetry under $\theta_2 \mapsto -\theta_2$, the remaining order parameters have no first-order variation. Writing $\theta_2 = s\sqrt{\nu_2}$, with $s = \pm 1$, the relevant linearized updates are

$$\begin{cases} \delta m = V_0 \delta\hat{m}, \\ \delta\hat{m} = \dfrac{\alpha}{2} \displaystyle\sum_{s=\pm 1} \mathbb{E}_\Xi \left[V_0^{-1} \left(\delta Z_1'^{(s)} - \delta\omega_1^{(s)} \right) s\sqrt{\nu_2} \right]. \end{cases} \tag{72}$$

We therefore need the first-order variations of the local fields and proximal. The fields are already linear in the perturbation:

$$\omega_\ell^{(s)} = \omega_\ell^0 + \delta_{\ell 1} s\sqrt{\nu_2}\, \delta m, \qquad \omega_\ell^0 = \delta_{\ell 1} m_{:1}^0 \sqrt{\nu_1} + q_0^{1/2} \xi_\ell. \tag{73}$$

Thus $\delta\omega_1^{(s)} = s\sqrt{\nu_2}\, \delta m$, while the other fields remain unchanged. To obtain $\delta Z'^{(s)}$, we expand the condition defining the proximal. Let $\ell(Z)$ denote the loss term in the Moreau minimization. Writing $Z'^{(s)} = Z'^0 + \delta Z'^{(s)}$, its stationarity condition is

$$0 = \nabla_Z \ell(Z'^0 + \delta Z'^{(s)}) + (\mathbb{I}_L \otimes V_0^{-1}) \left(Z'^0 + \delta Z'^{(s)} - \Omega^0 - \delta\Omega^{(s)} \right). \tag{74}$$

The zeroth-order terms cancel because Z'^0 satisfies the unperturbed stationarity condition. Taylor-expanding the gradient therefore gives, to first order,

$$\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1} \right] \delta Z'^{(s)} = (\mathbb{I}_L \otimes V_0^{-1}) \delta\Omega^{(s)}. \tag{75}$$

Since only ω_1 is perturbed, we obtain

$$\delta Z_1'^{(s)} = (\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1} \right]^{-1})_{11} V_0^{-1} \delta\omega_1^{(s)} \tag{76}$$

Therefore, at first order,

$$\begin{aligned} \delta\hat{m} &= \frac{\alpha}{2} \sum_{s=\pm 1} \mathbb{E}_\Xi \left[V_0^{-1} \left((\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1} \right]^{-1})_{11} V_0^{-1} - \mathbb{I}_H \right) \delta\omega_1^{(s)} s\sqrt{\nu_2} \right] \\ &= \alpha\nu_2 V_0^{-1} \mathbb{E}_\Xi \left[(\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1} \right]^{-1})_{11} V_0^{-1} - \mathbb{I}_H \right] \delta m. \end{aligned} \tag{77}$$

Finally, using $\delta m = V_0 \delta \hat{m}$, we get the SE map for the perturbation

$$\delta m^{t+1} = J_{\rm spec}(\alpha)\delta m^t, \quad J_{\rm spec}(\alpha) = \alpha\nu_2 \mathbb{E}_\Xi \left[(\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1}\right]^{-1})_{11} V_0^{-1} - \mathbb{I}_H \right]. \quad (78)$$

The unspecialized solution is stable if the spectral radius of the linear operator $\delta m^t \to \delta m^{t+1} = J(\alpha)\delta m^t$ is less than one. Finding the critical sample complexity for the specialization transition then comes down to studying the largest eigenvalue of the matrix $J(\alpha)$ with the condition

$$\lambda_{\max}[J_{\rm spec}(\alpha_c^{\rm spec})] = 1, \quad (79)$$

where we noted $\lambda_{\max}[\cdot]$ the largest eigenvalue. Evaluating this condition on the PS or PSB solutions gives the PS–SPEC or PSB–SPEC stability respectively.

Stability against permutation symmetry breaking At the PS solution with $m_{:2} = 0$, we perturb towards PSB with $q = q_0 + \delta q$, $\delta q = q_\perp$. Denoting $e_- = \frac{1}{\sqrt{2}}(1, -1)^\top$ and following similar steps as before leads to the SE map for the perturbation

$$\delta q^{t+1} = J_{\rm psb}(\alpha)\delta q^t, \quad (80)$$

$$J_{\rm psb}(\alpha) = \alpha \mathbb{E}_\Xi \left[\sum_{\ell,k=1}^{L} \left(e_-^\top \left[(\left[\nabla_Z^2 \ell(Z'^0) + \mathbb{I}_L \otimes V_0^{-1}\right]^{-1})_{\ell k} V_0^{-1} - \delta_{\ell k} \mathbb{I}_H \right] e_- \right)^2 \right]. \quad (81)$$

Again, the critical sample complexity is recovered through the condition

$$\lambda_{\max}[J_{\rm psb}(\alpha_c^{\rm psb})] = 1. \quad (82)$$

C.2 Non-isotropic Gaussian model

In the case of the non-isotropic Gaussian model, we study the stability against specialization to the next largest signal spike. Consider an r-SPEC solution (m_r, q_r, V_r), for which $m_{r,:j} = 0$ for all $j > r$. We perturb towards alignment with the next feature with $m_{:r+1} = \delta m_{:r+1}$. Since $\theta_{r+1} \sim \mathcal{N}(0, \nu_{r+1})$ is centered and independent of the recovered features, the remaining order parameters have no first-order variation. Denoting by Z_r' the proximal evaluated at (m_r, q_r, V_r); following similar steps as before leads to the SE map for the perturbation

$$\delta m_{:r+1}^{t+1} = J_r(\alpha)\, \delta m_{:r+1}^t, \quad (83)$$

$$J_r(\alpha) = \alpha\nu_{r+1}\, \mathbb{E}_{\theta,\Xi} \left[\left(\left[\nabla_Z^2 \ell(Z_r') + \mathbb{I}_L \otimes V_r^{-1}\right]^{-1} \right)_{11} V_r^{-1} - \mathbb{I}_H \right]. \quad (84)$$

the critical sample complexity is recovered through the condition

$$\lambda_{\max}\left[J_r(\alpha_c^{(r+1)}) \right] = 1. \quad (85)$$

D Link between statistical transitions and dynamical emergence

We consider the case of the non-isotropic Gaussian model where $m = 0, q = 0$ is a saddle point of the loss, by symmetry $K \mapsto -K$. We focus on the first specialization and assume it happens only along the first feature k_1^*. We introduce the largest curvature $C \in \mathbb{R}^+$ of minus the population loss $\mathcal{E}_{\rm pop}$ in the direction of the first feature k_1^*, defined as

$$C = D^{-2}\lambda_{\max}\left[-(k_1^*)^\top (\nabla_{K,K}^2 \mathcal{E}_{\rm pop}) k_1^* \right], \qquad \mathcal{E}_{\rm pop}(K) = \mathbb{E}_{X,y}\left[\frac{1}{2D} \| y - f_K(X) \|_2^2 \right] \quad (86)$$

where $\lambda_{\max}[\cdot]$ is the largest eigenvalue and $\nabla_{K,K}^2$ the Hessian.

D.1 Online gradient flow

We consider the gradient-flow limit of unregularized online SGD. At time τ the update is

$$\frac{{\rm d}K}{{\rm d}\tau} = -D\nabla_K \mathcal{E}_{\rm pop}(K), \quad (87)$$

with the weights being independently initialized as $k_h(\tau = 0) \sim \mathcal{N}(0, \eta^2 \mathbb{I}_D)$ for any h and for $\eta > 0$. The normalization D in eq. 87 insures that the time variation of the overlap m are of order $\mathcal{O}_D(1)$.

We follow Sagitova et al. (2026). They show that the alignment m follows the evolution

$$\frac{\mathrm{d}m}{\mathrm{d}\tau} = -\nabla_m \mathcal{E}_{\text{pop}} \ , \quad m_h(\tau = 0) \sim \mathcal{N}(0, D^{-1}\eta^2 \mathbb{I}_D) \tag{88}$$

and that in the limit $\eta \to 0$ the weights stay in the span of the features $k_h \in \text{span}(k_f^*)_{f=1}^F$ with $q = mm^\top$. $m = 0$ is an unstable fixed-point; we perform a linear stability analysis around it. We project m onto the direction corresponding to the fastest descent direction, defined as the eigenvector of the maximal eigenvalue in eq. 86. We write $\bar{m} \in \mathbb{R}$ its projection. We have the following exponential escape

$$\partial_\tau m(\tau) = -\nabla_m \mathcal{E}_{\text{pop}}(m(\tau)) \tag{89}$$

$$= -\nabla^2_{m,m} \mathcal{E}_{\text{pop}}(0) m(\tau) \tag{90}$$

$$= -D^{-2} m(\tau) K^* \nabla^2_{K,K} \mathcal{E}_{\text{pop}}(0) (K^*)^\top \tag{91}$$

$$\bar{m}(\tau) = \bar{m}(0) e^{C\tau} \tag{92}$$

For any small enough threshold $\varepsilon > 0$, independent of η, we define τ^{spec} the specialization time to be the first τ such that $|\bar{m}| = \varepsilon$. We have

$$\varepsilon = |\bar{m}(0)| e^{C\tau^{\text{spec}}} \tag{93}$$

$$\tau^{\text{spec}} = \frac{1}{C} \left(\log \varepsilon - \log |\bar{m}(0)| \right) \tag{94}$$

We define the rescaled specialization time $\tilde{\tau}^{\text{spec}}$ as

$$\tilde{\tau}^{\text{spec}} = \frac{\tau^{\text{spec}}}{\log \sqrt{D}/\eta} \tag{95}$$

We have $|\bar{m}(0)| = \eta \bar{\bar{m}}/\sqrt{D}$ with $\bar{\bar{m}}$ a random variable independent of η and D. In the limits $D \to \infty$ and $\eta \to 0$ we obtain that

$$\tilde{\tau}^{\text{spec}} = \frac{1}{C} \left(\frac{\log \varepsilon}{\log \sqrt{D}/\eta} - \frac{\log \bar{\bar{m}}}{\log \sqrt{D}/\eta} + 1 \right) \tag{96}$$

$$\to 1/C \tag{97}$$

that is to say the specialization time is only controlled by the Hessian of the loss at initialization. The random fluctuations due to the initialization and the effect of ε vanish.

D.2 ERM

We consider the static specialization in sample complexity. We consider the limit of large regularization $\lambda \gg 1$. In this limit, the model does not overfit data and $\mathcal{E}_{\text{train}} = \mathcal{E}_{\text{pop}}$. We also have $q = mm^\top$ and $D^{-1}||K||_F^2 = \text{Tr} q = ||m||_F^2$, that is to say the weights K are solely in the space of the ground truth K^* and do not have components that fit noise.

We expand the training objective to the 2nd order in m.

$$\frac{1}{D}\mathcal{L}_{\text{train}} = \alpha \mathcal{E}_{\text{pop}} + \frac{\lambda}{2} ||m||_F^2 \tag{98}$$

$$= \alpha \mathcal{E}_{\text{pop}}(0) + \frac{1}{2} \left(\alpha \nabla^2_{m,m} \mathcal{E}_{\text{pop}}(0) + \lambda \mathbb{I}_H \otimes \mathbb{I}_F \right) \cdot (m, m) \tag{99}$$

$$= \alpha \mathcal{E}_{\text{pop}}(0) + \frac{1}{2} \left(\alpha D^{-2} K^* \nabla^2_{K,K} \mathcal{E}_{\text{pop}}(0) (K^*)^\top + \lambda \mathbb{I}_H \otimes \mathbb{I}_F \right) \cdot (m, m) \tag{100}$$

We define α^{spec} the smallest sampling ratio such that $m = 0$ is no longer a local minimum of $\mathcal{L}_{\text{train}}$. We only consider the specialization along k_1^* and project m onto k_1^*. The stability of the fixed-point

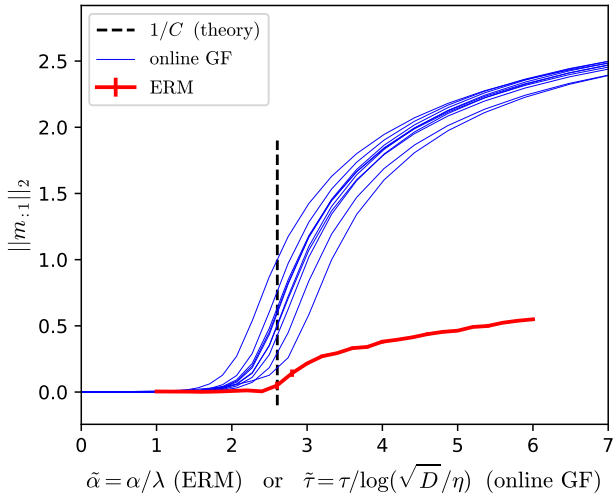

Figure 8: **Correspondence between α_c^{spec} and τ_c^{spec}:** head–feature overlap norm $\|m_{:1}\|_2$ for online gradient flow as a function of rescaled time $\tilde{\tau}$ (ten independent runs) and for empirical risk minimizers as a function of rescaled sample complexity $\tilde{\alpha}$ (averaged over five instances). $1/C$ is computed analytically in Sagitova et al. (2026). $D = 5 \times 10^3$, $\eta = 10^{-3}$, $\lambda = 3$. $F = 1$, Gaussian model with $\nu_1 = 8$, $H = 2$, $L = 5$.

$m = 0$ is controlled by whether the smallest eigenvalue of the quadratic form eq. 100 is negative. It reads $-\alpha C + \lambda$. This gives the critical sample complexity

$$\alpha^{\text{spec}} = \frac{\lambda}{C} \tag{101}$$

Last we rescale α as

$$\tilde{\alpha}^{\text{spec}} = \frac{\alpha^{\text{spec}}}{\lambda} \tag{102}$$

and obtain that

$$\tilde{\alpha}^{\text{spec}} = \tilde{\tau}^{\text{spec}} = 1/C. \tag{103}$$

The extension to more realistic scenarios, including online SGD and finite regularization λ, could be approached by a perturbation of this analysis; we expect the values of $\tilde{\tau}^{\text{spec}}$ spec and $\tilde{\alpha}_c^{\text{spec}}$ to be different but the underlying mechanism would stay the same. We provide a numerical comparison of the mapping between the ERM and the online gradient flow in Figure 8 at finite large D. We observe that specialization happens around the same $\tilde{\tau}^{\text{spec}}$ and $\tilde{\alpha}^{\text{spec}}$, whose values are well predicted by $1/C$.

E Implementation details

E.1 Experiments with solvable models

Resolution of the self-consistent equations We obtain solutions for the set of self-consistent equations (6) by iterating the associated state-evolution until convergence. The low-dimensional expectations over random variables (θ, Ξ) are evaluated with Monte-Carlo, with the proximal minimization performed numerically for each sample. Different solutions are first identified from uninformed initializations. Once their structure is known, branches can be followed using planted initialization, or by restricting the state-evolution updates on each solution. Their regularized training loss (or zero-temperature free entropy) identifies the global ERM minimum.

Numerical simulations For numerical simulations, we train the attention layer (2) on the regularized ERM objective (3), using synthetic datasets drawn from the model of Sec. 1. For each realization, we sample the planted directions and generate $N = \lfloor \alpha D \rfloor$ training sequences, keeping both fixed throughout optimization. Training is implemented in PyTorch using full-batch L-BFGS with learning rate 0.5, history size of 10 steps, until convergence, starting from random initialization of the weights orthogonal to the planted directions. Weights are sampled from an isotropic Gaussian distribution and projected onto the orthogonal complement of the planted-feature span, removing finite-size initial alignments with the planted directions; the effect of this projection vanishes at large D. Final overlaps and training losses are computed directly, while test errors are evaluated on 1000 separately generated sequences sharing the same latent directions. Each parameter setting is repeated over ten seeds, varying both the dataset and initialization. Before averaging empirical head–feature overlaps across seeds, we order symmetry-equivalent solutions consistently.

E.2 Experiments with BERT MLMs

Fixed dataset experiment We adapt the BERT masked-language-modeling experiments from Chen et al. (2024a), which studied the emergence of syntactic attention structure (SAS) in training time, close to the online learning regime. To investigate an equivalent analysis in sample complexity, we train models on fixed datasets of N sequences until the overfitting regime. To avoid data augmentation considerations, we also remove dynamic masking. Effective regularization is obtained through early stopping, evaluating the different metrics on checkpoint models with the lowest validation error along training trajectory. Repeating this procedure across dataset sizes and optimization seeds reveals the correspondence between emergence of specialized heads as a function of the number of samples and along training.

Training dataset Training sequences are sampled from the preprocessed BookCorpus and English Wikipedia dataset of the reference experiment. Training datasets are nested subsets of N sequences of 128 tokens randomly sampled from the original $\sim 9\mathrm{M}$ sequences dataset, with 2048 disjoint validation sequences shared across runs. Masking selects 15% of tokens, following the original $80/10/10$ replacement rule.

Model and training We train $\mathrm{BERT}_{\mathrm{Small}}$ from random initialization, with 4 layers, 8 heads per layer, hidden dimension 512, feed-forward dimension 2048, and dropout probability 0.1. Optimization uses AdamW with learning rate 10^{-4}, weight decay 0.1, and an effective batch size of 256 sequences, obtained by accumulating four minibatches of 64. The learning rate increases linearly over 1000 warmup steps and then decreases linearly on a fixed 10^6-step horizon. We train for 10000 optimizer steps, saving and evaluating checkpoints every 250 steps. Dataset sizes are $N \in \{1000, 3000, 6000, 9000, 12000, 15000, ..., 60000\}$ and seeds $\in \{0, 1, 2, 3, 4\}$. Seeds vary initialization, minibatch order, and dropout. Given a sequence $\tilde{X}_\mu$ corrupted at positions M_μ from an original sequence X_μ, the MLM loss is

$$\mathcal{L}_{\mathrm{MLM}}(\Theta) = -\frac{1}{\sum_{\mu=1}^{N} |M_\mu|} \sum_{\mu=1}^{N} \sum_{i \in M_\mu} \log p_\Theta(x_{\mu i} \mid \tilde{X}_\mu), \tag{104}$$

where $p_\Theta(x_{\mu i} \mid \tilde{X}_\mu)$ is the vocabulary distribution at position i predicted by the model with parameters Θ. The training error $\mathcal{E}_{\mathrm{train}}$ therefore corresponds to the latest logged MLM loss in training, while the validation error $\mathcal{E}$ is the MLM loss computed on the validation dataset.

Syntactic attention structure SAS is evaluated on a separate dataset of 1000 fixed Wall Street Journal sentences from the Penn Treebank (Marcus et al., 1999), with dependency parents and relation labels supplied in Stanford Dependencies format (Schuster & Manning, 2016). We follow the implementation from the repository of Chen et al. (2024a). Token attention is first translated to word attention. Then, for head h in layer l and attention direction $d \in \{\rightarrow, \leftarrow\}$, the head-specific probe $\hat{j}_{h,l,d}$ for a target word x_i is the other word with maximum attention in that direction:

$$\hat{j}_{h,l,d}(x_i) = \arg\max_{j \neq i} \alpha_{ij}^{h,l,d}. \tag{105}$$

Here $\alpha^{h,l}$ denotes the word-level attention matrix of head h in layer l, with $\alpha_{ij}^{h,l,\rightarrow} = \alpha_{ij}^{h,l}$ and $\alpha_{ij}^{h,l,\leftarrow} = \alpha_{ji}^{h,l}$. Let $\mathcal{R}$ denote the set of syntactic relations present in the probe dataset, and $\mathcal{D}_R$ the collection of annotated child–parent pairs (x_i, x_j) carrying relation R. The relation-specific UAS is

$$\mathrm{UAS}_R = \max_{h,l,d} \frac{1}{|\mathcal{D}_R|} \sum_{(x_i,x_j)\in\mathcal{D}_R} \mathbf{1}\{\hat{j}_{h,l,d}(x_i) = j\}. \tag{106}$$

This measures how accurately best individual heads recover relation R. Different relations can select different heads, layers, and attention directions, allowing their specialization to be studied separately. The overall UAS selects the best head, layer, and attention direction across all dependencies:

$$\mathrm{UAS} = \max_{h,l,d} \frac{1}{\sum_{R\in\mathcal{R}} |\mathcal{D}_R|} \sum_{R\in\mathcal{R}} \sum_{(x_i,x_j)\in\mathcal{D}_R} \mathbf{1}\{\hat{j}_{h,l,d}(x_i) = j\}. \tag{107}$$

This measures the fraction of correctly recovered parents using a single head, layer, and attention direction. For relation-specific scores, relation labels select the probes, while correctness depends only on predicting the parent. These scores serve as benchmark proxies relating to head–feature overlaps; quantifying alignment between attention structure and syntactic relations structure, a structural latent feature of language datasets.

In Fig. 9 we present the training dynamics for $\mathrm{BERT}_{\mathrm{Small}}$ at different dataset sizes. We observe the presence of drops in the error corresponding to the acquisition of SAS only at large enough dataset sizes while models trained on small datasets overfit the data early. Fig. 10 details relation-refined UAS_R for the statistically most present relations in the dataset, exhibiting a same critical dataset size for emergence of SAS suggesting a coordinated acquisition of syntactic structure. It would be interesting to explore if different capabilities, maybe different types of syntactic rules, are obtained at different sample complexities.

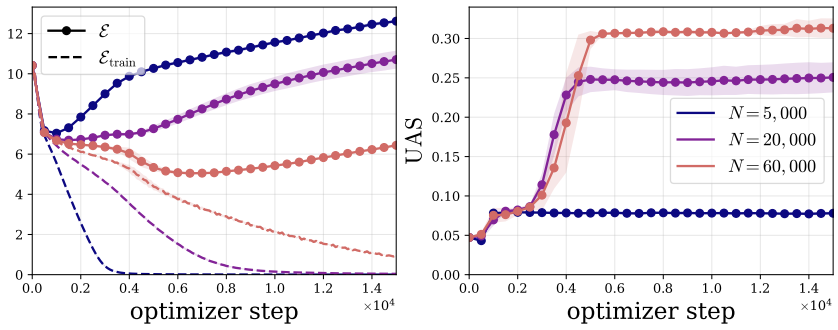

Figure 9: **Training dynamics on fixed datasets:** Separate runs of $\text{BERT}_{\text{Small}}$ models trained on datasets of $N \in \{5000, 20000, 60000\}$ sequences, following the training procedure described in App. E.2 with 15000 optimization steps and checkpoints every 500 steps. **Left:** Validation MLM loss $\mathcal{E}$ at registered checkpoints and logged training loss $\mathcal{E}_{\text{train}}$ along training. **Right:** Evolution of syntactic attention structure, measured by UAS. Larger datasets allow an abrupt emergence of SAS during training. Curves show averages across optimization seeds, with shaded bands indicating one standard deviation.

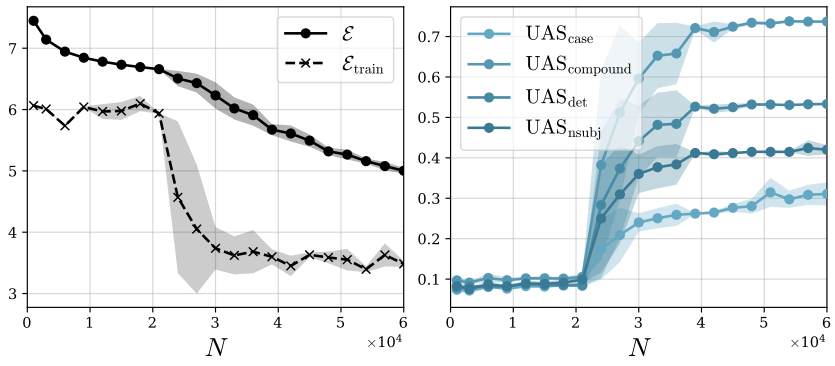

Figure 10: **Acquisition of different syntactic relations with increasing dataset size:** $\text{BERT}_{\text{Small}}$ models are trained on fixed datasets of increasing size N following the training procedure described in App. E.2. **Left:** Validation MLM loss $\mathcal{E}$ at registered checkpoints and logged training loss $\mathcal{E}_{\text{train}}$. **Right:** Relation-specific attachment scores for the four most frequent relations in the validation dataset: case, compound, det, and nsubj. Validation loss and attachment scores are evaluated with early stopping. Curves show seed averages, with shaded bands indicating one standard deviation.